Asian Journal of Probability and Statistics

Volume 27, Issue 2, Page 57-72, 2025; Article no.AJPAS.129937
ISSN: 2582-0230

Uncovering Risk Factors for Heart Disease and Predicting Outcomes Using Machine Learning Approaches

Shohel Mahmud [a*], Salma Akter Tania [a], Tanzila Tamanna [b], Md Habibur Rahman [b] and Saiful Islam [b]

[a] Department of Statistics, Noakhali Science and Technology University, Noakhali, Bangladesh.
[b] Department of Statistics and Data Science, Jahangirnagar University, Savar, Dhaka, Bangladesh.

Authors' contributions

This work was carried out in collaboration among all authors. All authors read and approved the final manuscript.

Article Information

DOI: https://doi.org/10.9734/ajpas/2025/v27i2713

Open Peer Review History:
This journal follows the Advanced Open Peer Review policy. Identity of the Reviewers, Editor(s) and additional Reviewers, peer review comments, different versions of the manuscript, comments of the editors, etc are available here: https://www.sdiarticle5.com/review-history/129937

Original Research Article

Received: 27/11/2024
Accepted: 29/01/2025
Published: 31/01/2025

Abstract

Aims: This study aims to create a robust machine learning model capable of accurately discerning the presence of heart-related disorders. The aim of this study is to find the best machine learning classification model that is most suitable for predicting risk factors related to heart disease.
Study Design: Analytical cross-sectional study.
Place and Duration of Study: Department of Statistics at the Noakhali Science and Technology University, and three tertiary level hospitals of Bangladesh (Noakhali General Hospital, Chittagong Medical College Hospital, and the National Institute of Cardiovascular Diseases, Dhaka), from June 2022 to August 2023.
Methodology: The conceptual framework underlying this study proposes a descriptive methodology in which study data are collected from hospital admitted patients who have heart disease symptoms and equal size of patients who have no heart related disease. Primary data were obtained using self-designed

**Corresponding author: Email: smahmud.stat@nstu.edu.bd;*

Cite as: *Mahmud, Shohel, Salma Akter Tania, Tanzila Tamanna, Md Habibur Rahman, and Saiful Islam. 2025. "Uncovering Risk Factors for Heart Disease and Predicting Outcomes Using Machine Learning Approaches". Asian Journal of Probability and Statistics 27 (2):57-72. https://doi.org/10.9734/ajpas/2025/v27i2713.*

questionnaire which were administered by the researchers. The sample size for the study is 340 comprising of 247 males and 93 females, who were selected by convenient sample method.
Results: Evaluating simulation models reveals the Decision Tree as the most compelling choice due to its high accuracy, interpretability, and statistical significance. The outcomes of real data analysis that the Decision Tree model emerges as the preeminent candidate, showcasing extraordinary predictive proficiencies in discerning the risk quotient associated with heart disease, achieving an accuracy of 91%, a sensitivity of 88%, and a specificity of 91%.
Conclusion: The results highlight the most effective machine learning algorithms for classification in the context of heart-related disease risk factors predictions. However, future research endeavors could enhance this study by incorporating additional clinical, demographic, and social determinants.

Keywords: Machine learning; model comparison; accuracy; heart disease prediction; Bangladesh.

1 Introduction

A machine learning algorithm is a diverse set of statistical, probabilistic, and optimization techniques designed to learn from past experiences, extracting valuable models from extensive, unstructured, and complex datasets (Uddin, 2019). These algorithms and techniques fall under the broader process known as knowledge discovery in databases or data mining (Beunza, 2019). Machine learning algorithm are applied in various field such as medical image detection, disease prediction, network intrusion detection and email-filtering (Lin, 2023, Asare, 2023, Mashaleh, 2022, Kabir, 2023). The ability of ML to handle enormous volumes of medical data enables the identification of patterns and the prediction of disease outcomes, leading to improvements in healthcare procedures (Ahsan, 2022). Machine learning algorithm plays a vital role in identifying instances of heart disease. By predicting these conditions in advance, doctors can acquire essential information that greatly facilitates the diagnosis and treatment of patients (Chang, 2022).

Machine learning operates by discerning intricate patterns in data to make informed inferences. For instance, global fatalities due to heart-related diseases, numbering about 17.9 million yearly, underscore their lethal impact, they account for 31% of global deaths, prompting the integration of machine learning-based diagnostic models in clinical decision support systems. These models, like the cardiovascular-specific predictive model developed, aid in disease determination based on risk factors. Cardiovascular diseases are recognized by the World Health Organization as a major global cause of death, with specific proportions reported by various sources like the European Public Health Alliance (Ozcan, 2023, Ramalingam, 2018, Sugendran, 2023, Iscra, 2022, Howlader, 2017). Cardiovascular disease (CVD), constituting over 30% of global fatalities, remains a significant global health concern, especially in nations like Bangladesh (Nagavelli, 2022). Though cardiac disease prediction has achieved some accuracy, cutting-edge machine learning techniques are pivotal to address its complexity (Chowdhury, 2021). Utilizing data from the Bangladesh Demographic and Health Survey (BDHS) highlight the prevalence of anemia in young Bangladeshi children. Bangladesh employs machine learning techniques across a spectrum of health concerns including coronary artery disease, cardiovascular disease, anemia, diabetes, cancer, liver disease (Khan, 2019, Islam, 2013, Islam, 2020, Nipa, 2023, Behera, 2023). The nation faces a pressing challenge of cardiovascular disease, involving established and population-specific risk factors, as well as hereditary influences (Ahmed, 2018). Alarmingly, significant portions of the Bangladeshi population, irrespective of gender, are susceptible to early-life cardiovascular diseases. The projected rise in heart disease-related fatalities by 2030 by the World Health Organization emphasizes the need for proactive interventions, especially in low-income countries. The potential of machine learning algorithms to automate triaging through pattern recognition among patients with varying symptoms accelerates healthcare processes, enhancing overall efficacy (Khanna, 2023, Hajian-Tilaki, 2013, Hossain, 2024).

Machine learning's transformative role in healthcare is evident, aiding in the detection of cardiac diseases for early intervention and lowered mortality rates. However, challenges in early disease detection persist, encompassing precision, accuracy, and temporal complexity. Addressing these challenges, this study introduces a machine learning-based risk factor predict for disease. Given the complexity of disease mechanisms and symptoms, traditional diagnosis methods are time-consuming and resource-intensive, often reliant on human capacity. The primary contributions of this work are twofold: first, to enhance accurate disease diagnosis, particularly in relation to heart-related conditions; and second, to determine the most crucial causes and traits associated with these conditions. Additionally, the study undertakes an in-depth analysis and comparison of the

effectiveness of different machine learning algorithms for disease detection and categorization. Moreover, it delves into investigating the potential of machine learning methods in improving the early diagnosis of heart-related disorders. Focusing on heart related disease risk factor prediction; this study employs curated patient data from diverse medical sources to identify crucial risk factors. By bridging the gap in efficient illness diagnosis, this work aligns with the global imperative for enhanced healthcare outcomes.

2 Materials and Methods

2.1 Participants

We conducted a retrospective selection of patients admitted to Noakhali General Hospital, Chittagong Medical College Hospital, and the National Institute of Cardiovascular Diseases (NICVD). Data collection was conducted involving two distinct groups: one comprised individuals with heart-related diseases, while the other encompassed individuals with different medical conditions. The dataset comprises a total of 340 samples, evenly divided between 170 individuals diagnosed with heart disease and 170 individuals without any heart-related conditions. This study aimed to discern patterns and characteristics that differentiate between these two groups, contributing to a better understanding of the distinctive features associated with heart-related diseases.

2.2 Study design and procedure

A comprehensive and informative survey has been created with 21 questions divided into two parts. The first part focuses on demographic information, including gender, age, and BMI. The second part gathers background data that can be utilized to predict factors associated with heart-related diseases. Following the face-to-face collection of survey responses, the data was entered into Microsoft Excel and subsequently converted into CSV format. This allows for the utilization of the CSV data file in our R-programming. Afterward, irrelevant portions of the dataset, such as participant names and phone numbers, were removed to streamline the data for prediction purposes. Following that, we undertook the statistical analysis of the gathered data.

2.3 Statistical analysis

This study categorized the features and conducted several statistical analyses. Firstly, the researchers proceeded to determine the descriptive statistics and determine the associations between dependent and independent variables. And then, machine learning algorithms were employed to efficiently compare machine learning models. Researchers split the dataset into two sets when it was finished: a training set and a testing set. Eighty percent of the data was allocated for training the machine learning model, while the remaining 20% was reserved for evaluating the final model post-training. Finally, these partitioned datasets were used for implementing machine learning (ML) using the R programming language. As the experiment's outcomes are categorical, yielding values such as yes or no, this study opted for classification algorithms within supervised machine learning techniques, as opposed to regression. The algorithms are as follows: decision tree (DT), naïve bayes (NB), k-nearest-neighbor (KNN), random forest (RF), support vector machine (SVM), and generalized linear model (GLM). The most successful prediction model for illness detection and the risk variables that go along with it were found by comparing the categorization models. The flowchart of the research is presented in the Fig. 1. At the end, the study used an illustration of the receiver operating characteristic (ROC) curve to show how the best model will perform and forecast it. The data analysis for this study was carried out using a variety of software programs, including MS Excel and R version 4.2.3 were utilized to expedite the data analysis procedure.

3 Results

The research was divided into two clear groups: the initial one included individuals diagnosed with heart disease, and the second consisted of participants who did not have heart disease. Equal-sized samples were collected from both groups, ensuring a balanced representation in the study.

3.1 Respondents characteristics information

Table 1 outlines the patient’s characteristics of 340 participants, with 73.5% of the heart disease group and 71.8% of the non-heart disease group being male. Senior adults constitute a significant portion, comprising

51.2% of those with heart disease and 34.7% without. Middle-aged adults represent 39.4% and 27.6% in the heart and non-heart disease groups, respectively. Notably, a majority in both groups have a normal BMI (67.6% for heart disease, 72.9% for non-heart disease), while 23.5% with heart disease and 10.6% without are overweight. The highest numbers in both groups exhibit normal BMI. Concerning smoking habits, 82.9% with heart disease and 78.2% without are non-smokers, emphasizing a prevalent non-smoking trend in both groups. Additionally, 8.8% with heart disease and 16.5% without are classified as underweight.

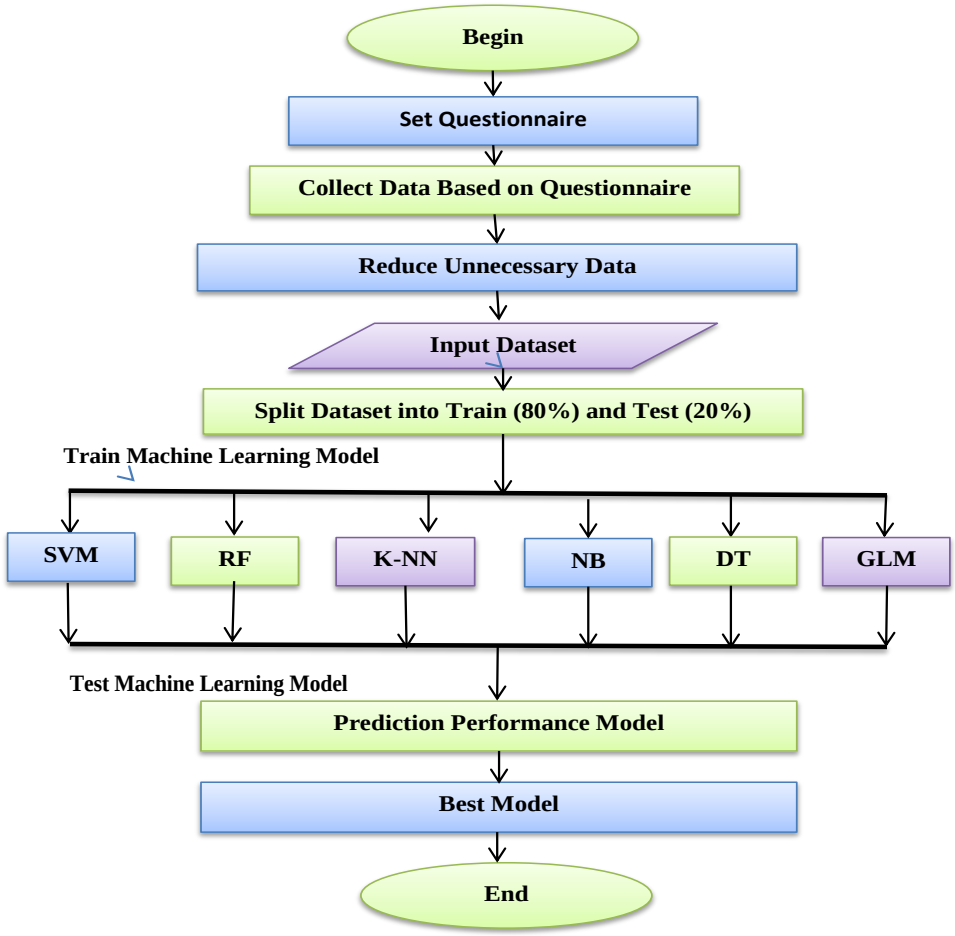

Fig. 1. Flowchart of the proposed research work activities

3.2 Information of risk factors

In addition, Table 2 details the elements of risk of heart disease among the participants in the study. In terms of blood pressure, 45.9% of individuals with heart disease and 56.5% without heart disease exhibited normal systolic blood pressure. Notably, a higher percentage of non-heart disease patients (10.6%) demonstrated normal systolic pressure compared to their heart disease counterparts. Regarding diastolic blood pressure, 54.1% with heart disease and 61.2% without had normal levels. The study observed that hypertension was more prevalent in the heart disease group (14.1%) compared to the non-heart disease group (8.8%). Additionally, it found a higher occurrence of hypertensive crisis cases within the heart disease respondents (17.1%) compared to the non-heart disease group (8.2%). Moving beyond blood pressure, the analysis extended to factors such as heart rate, hemoglobin levels, white blood cell count, platelet levels, and various blood chemistry parameters, offering a detailed insight into the physiological profiles of both groups.

The study delved into a range of hematological and biochemical markers, revealing intriguing patterns between those with heart disease and those without. Notable findings include differences in hemoglobin levels, white

blood cell counts, platelet levels, and Serum Creatinine levels. Moreover, lipid profiles, including LDL and HDL cholesterol levels, exhibited distinct trends between the two groups. The study also explored markers of random blood sugar, sodium and potassium levels, as well as chloride levels, providing a comprehensive overview of the physiological status of individuals with and without heart disease. These detailed insights contribute to a nuanced understanding of the health characteristics and potential risk factors associated with heart disease within the studied population.

Table 1. Respondents background characteristics

Characteristics	Category	Heart Disease Group		Non Heart Disease Group	
		Frequency	Percentage (%)	Frequency	Percentage (%)
Gender	Male	125	73.5	122	71.8
	Female	45	26.5	48	28.2
Age Group	Teenage	00	00	04	2.4
	Adult	16	9.4	60	35.3
	Middle Age Adult	67	39.4	47	27.6
	Senior Adult	87	51.2	59	34.7
BMI	Thinness	15	8.8	28	16.5
	Normal	115	67.6	124	72.9
	Overweight	40	23.5	18	10.6
	Obese	00	00	00	00
Current Smoking	Yes	29	17.1	37	21.8
	No	141	82.9	133	78.2

Table 2. Risk factors of heart diseases

Characteristics	Category	Heart Disease Group		Non Heart Disease Group	
		Frequency	Percentage (%)	Frequency	Percentage (%)
Systolic Blood Pressure	Low level	24	14.1	45	26.5
	Normal	78	45.9	96	56.5
	Elevated	15	8.8	00	0.0
	Hypertension	24	14.1	15	8.8
	Hypertensive level	29	17.1	14	8.2
Diastolic Blood Pressure	Low level	49	28.8	32	18.8
	Normal	92	54.1	104	61.2
	Elevated	19	11.2	27	15.9
	Hypertension	10	5.9	05	2.9
	Hypertensive level	00	00	02	1.2
Heart Rate	Low level	33	19.4	03	1.8
	Normal	99	58.2	159	93.5
	Abnormal	23	22.4	08	4.7
Hemoglobin (Hb^+)	Low level	84	49.4	76	44.7
	Normal	83	48.8	92	54.1
	Alarming	03	1.8	02	1.2
White Blood Cells	Low level	03	1.8	25	14.7
	Normal	105	61.8	85	50.0
	Alarming	62	36.5	60	35.3
Red Blood Cells	Low level	101	59.4	70	41.2
	Normal	65	38.2	92	54.1
	Alarming	04	2.4	08	4.7
Platelets	Low level	11	6.5	64	37.6
	Normal	154	90.6	103	60.6
	Alarming	05	2.9	03	1.8
Neutrophils	Low level	07	4.1	08	4.7

Characteristics	Category	Heart Disease Group		Non Heart Disease Group	
		Frequency	**Percentage (%)**	**Frequency**	**Percentage (%)**
	Normal	32	18.8	44	25.9
	Alarming	131	77.1	118	69.4
Serum Creatinine	Low level	03	1.8	04	2.4
	Normal	127	74.7	101	59.4
	Alarming	40	23.5	65	38.2
LDL	Optimum	35	20.6	14	8.2
	Fairly Good	29	17.1	52	30.6
	High	58	34.1	66	38.8
	Very High	48	28.2	38	22.4
HDL	Very low	128	75.3	93	54.7
	Low level	42	24.7	77	45.3
	Optimal	00	0.0	00	0.0
Total Cholesterol	Optimal	82	48.2	83	48.8
	Elevated	62	36.5	44	25.9
	High	26	15.3	43	25.3
Random Blood Sugar	Normal Blood Sugar	55	32.4	45	26.5
	Prediabetes	48	28.2	71	41.8
	Diabetes	67	39.4	54	31.8
Sodium (Na^+)	Low level	63	37.1	110	64.7
	Normal	106	62.4	59	34.7
	Alarming	01	0.6	01	0.6
Potassium(K^+)	Low level	28	16.5	52	30.6
	Normal	137	80.6	98	57.6
	Alarming	05	2.9	20	11.8
Chloride (Cl^-)	Low level	44	25.9	64	37.6
	Normal	99	58.2	94	55.3
	Alarming	27	15.9	12	7.1

3.3 Relationship between risk factors and heart disease

Table 3, in the cross-tabulation analysis for age groups, the calculated chi-square value signifies a substantial departure from the expected distribution, indicating a strong relationship between predictor variable and the presence of heart disease. This result is further supported by Cramer's V value reflecting a meaningful association predictor variable and the presence of the heart disease. Additionally, the exceedingly low *P*-value indicates a high level of statistical significance, reinforcing the conclusion that predictor variable has a significant impact on the presence of heart Disease.

Additionally, a comparison of patients without and with heart disease have higher related factors depicted in Table 3. The correlation analysis between presence of heart disease and influencing risk factors of heart disease listed in Table 1: age, BMI, systolic blood pressure, diastolic blood pressure, heart rate, red blood cells (RBC), white blood cells (WBC), serum creatinine , platelets, LDL, HDL, total cholesterol, random blood sugar (RBS), sodium, potassium and chloride are positively correlated ($P < .01$) with presence of heart disease and statistically significant. Gender, current smoking, hemoglobin (Hb+) and neutrophils are not statistically significant.

3.4 Model comparison

In the context of Table 4, the study employs various characteristics of given models. Firstly, among the given machine learning models, the decision tree stands out as the most compelling choice due to its remarkable accuracy of 0.91. Decision Trees offer a clear advantage in their ability to capture intricate non-linear relationships within the data, making them particularly adept at handling complex scenarios. Moreover, their intuitive nature facilitates easy interpretation and visualization of decision-making processes. While they can be prone to over fitting on intricate datasets, the competitive accuracy score suggests that the decision tree in

question is well-tuned or pruned appropriately. Despite the strong performances of other models such as the random forest (0.88), the decision tree's blend of accuracy, interpretability, and computational efficiency substantiates its position as the optimal model for this specific task.

Table 3. Association of risk factors and heart disease

Characteristics	Category	Presence of Heart Disease			Pearson Chi-square	Cramer's V	*P*-value
		Yes	No	Total	Value	Value	
Gender	Female	45	48	93	0.133	0.020	.715
	Male	122	125	247			
Age	Teenage	00	04	04			
	Adult	16	60	76			
	Middle Age Adult	67	47	114	38.352	0.336	.000
	Senior Adult	87	59	146			
BMI	Thinness	15	28	43			
	Normal	115	124	239			
	Overweight	40	18	58	12.54	0.192	.002
	Obese	00	00	00			
Systolic Blood Pressure	Low level	24	45	69			
	Normal	78	96	174			
	Elevated	15	00	15	30.56	0.300	.000
	Hypertension	24	15	39			
	Hypertensive level	29	14	43			
Diastolic Blood Pressure	Low level	49	32	81			
	Normal	92	104	196			
	Elevated	19	27	46	9.36	0.166	.053
	Hypertension	10	05	15			
	Hypertensive level	00	02	02			
Heart Rate	Low level	33	03	36			
	Normal	99	159	258	58.51	0.415	.000
	Abnormal	38	08	46			
Current Smoking	Yes	29	37	66	1.20	0.059	.273
	No	141	133	274			
Hemoglobin (Hb+)	Low level	84	76	160			
	Normal	83	92	175	1.06	0.056	.588
	Alarming	03	02	05			
White Blood Cells	Low level	03	25	28			
	Normal	105	85	190	19.42	0.239	.000
	Alarming	62	60	122			
Red Blood Cells	Low level	101	70	171			
	Normal	65	92	157	11.59	0.185	.003
	Alarming	04	08	12			
Platelets	Low level	11	64	75			
	Normal	154	103	257	48.07	0.376	.000
	Alarming	05	03	08			
Neutrophils	Low level	07	08	15			
	Normal	32	44	76	2.64	0.088	.267
	Alarming	131	118	249			
Serum Creatinine	Low level	03	04	07			
	Normal	127	101	228	9.06	0.163	.011
	Alarming	40	65	105			
LDL	Optimum	35	14	49			

Characteristics	**Category**	**Presence of Heart Disease**			**Pearson Chi-square**	**Cramer's V**	***P*-value**
		Yes	**No**	**Total**	**Value**	**Value**	
	Fairly Good	29	52	81	17.21	0.225	.001
	High	58	66	124			
	Very High	48	38	86			
HDL	Very low	128	93	221			
	Low level	42	77	119	15.83	0.216	.000
	Optimal	00	00	00			
Total Cholesterol	Optimal	82	83	165			
	Elevated	62	44	106	7.25	0.146	.027
	High	26	43	69			
Random Blood Sugar	Normal Blood Sugar	55	45	100	6.84	0.142	.033
	Prediabetes	48	71	119			
	Diabetes	67	54	121			
Sodium (Na+)	Low level	63	110	173			
	Normal	106	59	165	26.15	0.277	.000
	Alarming	01	01	02			
Potassium(K+)	Low level	28	52	80			
	Normal	137	98	235	22.62	0.258	.000
	Alarming	05	20	25			
Chloride (Cl-)	Low level	44	64	108			
	Normal	99	94	193	9.60	0.168	.008
	Alarming	27	12	39			

The *P*-value associated with the decision tree ($P < .001$) is impressively small, indicating a substantial and statistically significant improvement in accuracy compared to the baseline expectation. This suggests that the decision tree captures intricate patterns within the data more effectively than the other models, leading to highly accurate predictions. Among the models, the decision tree boasts the highest kappa value of 0.79, indicating strong agreement between its predictions and actual outcomes. This suggests that the decision tree captures the underlying patterns in the data exceptionally well, resulting in reliable predictions. While the random forest also displays a respectable kappa value of 0.68, the decision tree's higher kappa value reaffirms its superior performance in minimizing both false positives and false negatives. Therefore, based on the kappa values alone, the decision tree emerges as the model with the most accurate and consistent predictions.

Interpreting the sensitivity provided for each machine learning model, the study can evaluate their ability to correctly identify positive instances or true positives. Among the models, both the random forest and the decision tree exhibit the highest sensitivity values of 88%, implying that they are equally adept at capturing true positive cases. This suggests that these models can effectively recognize instances of the positive class, minimizing the risk of false negatives. While SVM, Naïve Bayes, and GLM also show reasonably high Sensitivity values, the consistent performance of the random forest and decision tree sets them apart.

Interpreting the specificity values provided for each machine learning model, we can assess their ability to correctly identify negative instances or true negatives. Among the models, the naïve bayes and the decision tree exhibit the highest specificity values of 91%, indicating their proficiency in accurately recognizing negative cases. This suggests that these models excel in minimizing the occurrence of false positives, thus enhancing their reliability in classifying negative instances. While the random forest also demonstrates a notable specificity value of 80%, the decision tree's performance aligns closely with the naïve bayes model.

The prevalence values for naïve bayes, decision tree, and GLM are identical at 0.53, while the other models have slightly different values. Prevalence reflects the proportion of positive instances in the dataset, and models often perform better on the majority class when prevalence is imbalanced. Considering this, the decision tree's kappa value of 0.79, which is the highest among the models, seems even more impressive as it indicates strong agreement beyond chance in a scenario where the positive class is not significantly overrepresented. This suggests that the decision tree's superior performance isn't solely due to an imbalanced distribution, but rather its capacity to accurately capture patterns across classes. Therefore, based on both prevalence and kappa values, the

decision tree emerges as the most reliable model for this task, demonstrating its ability to generalize well across different class distributions.

Among the models, both the naïve bayes and decision tree display the highest positive predictive value (PPV) values of 91%, indicating their proficiency in correctly identifying positive cases while minimizing false positives. The random forest also demonstrates a noteworthy PPV value of 88%. However, it's important to note that the decision tree's higher kappa value of 0.79 signifies better overall agreement with the actual classifications compared to other models. While naïve bayes and decision tree share the same PPV, the decision tree's superior kappa suggests that its performance is not solely driven by chance or bias toward one class. Therefore, considering both the PPV and kappa values, the decision tree emerges as the most reliable and balanced model.

Among the models, both the random forest and naïve bayes exhibit the highest NPV values of 88%, indicating their proficiency in correctly identifying negative cases while minimizing false negatives. This suggests that these models are effective at avoiding false alarms and classifying instances as negative when they truly belong to the negative class. While the decision tree also shows a competitive NPV value of 83%, the consistently high performance of the random forest and naïve bayes models is noteworthy.

Based on the comprehensive analysis of various evaluation metrics and their corresponding values for each machine learning model, the decision tree model emerges as the most robust and suitable choice for the given task. The decision tree consistently demonstrates exceptional performance across a range of metrics including accuracy, kappa, specificity, positive predictive value, negative predictive value, and sensitivity. Its high accuracy 91%, suggests that it can effectively capture complex relationships within the data, leading to accurate predictions.

Furthermore, its high kappa value 0.79, indicates strong agreement beyond chance with the actual classifications. The decision tree also excels in minimizing false positives and false negatives, as highlighted by its high specificity, positive predictive value, and sensitivity values. The balanced performance across various evaluation aspects, coupled with its capacity to maintain accuracy and generalization even in the presence of varying class distributions, makes the decision tree model the optimal choice for this study.

In the analysis presented in Table 4, the decision tree model demonstrates notable strengths, achieving the highest sensitivity of 88% and accuracy 91%. The model‘s ability to accurately identify positive instances is underscored by its impressive sensitivity, while the random forest model excels in specificity 80% for correctly identifying negative instances. Notably, the decision tree model‘s high area under curve (AUC) value 0.91 affirm its effectiveness in distinguishing between classes across varied thresholds, showcasing a well-balanced performance (Fig. 2). Collectively, these metrics position the decision tree model as a robust performer in accurately classifying instances in the context of the study.

4 Discussion

Numerous machine learning models were suggested to investigate and predict risk factors associated with heart related disease. Drawing from individual studies, several widely utilized machine learning classification models were considered to explore risk factors, allowing for a comparative analysis of these models. This research contributes significantly to the development of an approach for diagnosing heart failure (HF) in symptomatic patients with risk factors. The approach relies on simple clinical data, along with natriuretic peptides and echocardiographic indices (as recommended by ESC guidelines), employing various machine learning techniques (Plati, 2021). Furthermore, this study has significantly advanced the establishment of an approach to diagnose heart disease risk factor, utilizing straightforward clinical data, alongside hematology blood test, biochemistry test, and electrolytes test, by employing diverse machine learning model. In addition, the results for HF diagnosis were quite high in terms of accuracy 91.23%, as well as in terms of sensitivity 93.83% and specificity 89.62%, confirming the classification power of ML approaches. In the contrary, our model achieved 91% accuracy as well as in terms of sensitivity 88% and specificity 91%, confirming the best classification model is decision tree model. Moreover, the study suggests a cloud-based heart disease prediction system that uses machine learning techniques. The algorithm, which was created by analyzing a number of machine learning algorithms on the waikato environment for knowledge analysis (WEKA) platform, successfully identified cardiac illnesses with a high accuracy of 97.53% and excellent levels of sensitivity and specificity

Table 4. Model comparison

Characteristics	**SVM**	**Random forest**	**KNN K=5**	**Naïve Bayes**	**Decision Tree**	**GLM**
Accuracy (C. I.)	0.77 (0.66, 0.87)	0.88 (0.75, 0.90)	0.75 (0.63, 0.84)	0.69 (0.56, 0.79)	0.91 (0.79, 0.95)	0.76 (0.64, 0.85)
Kappa	0.55	0.68	0.49	0.37	0.79	0.52
95% CI	(0.66, 0.87)	(0.75, 0.90)	(0.63, 0.84)	(0.56, 0.79)	(0.79, 0.95)	(0.64, 0.85)
No Information Rate	0.52	0.50	0.52	0.52	0.5	0.52
P-Value	0.000	0.000	0.000	0.004	0.000	0.000
Sensitivity	0.86	0.88	0.77	0.75	0.88	0.88
Specificity	0.68	0.80	0.72	0.62	0.91	0.62
Prevalence	0.52	0.50	0.53	0.53	0.50	0.53
Positive Predictive Value	0.76	0.88	0.78	0.69	0.91	0.73
Negative Predictive Value	0.82	0.88	0.75	0.69	0.88	0.83

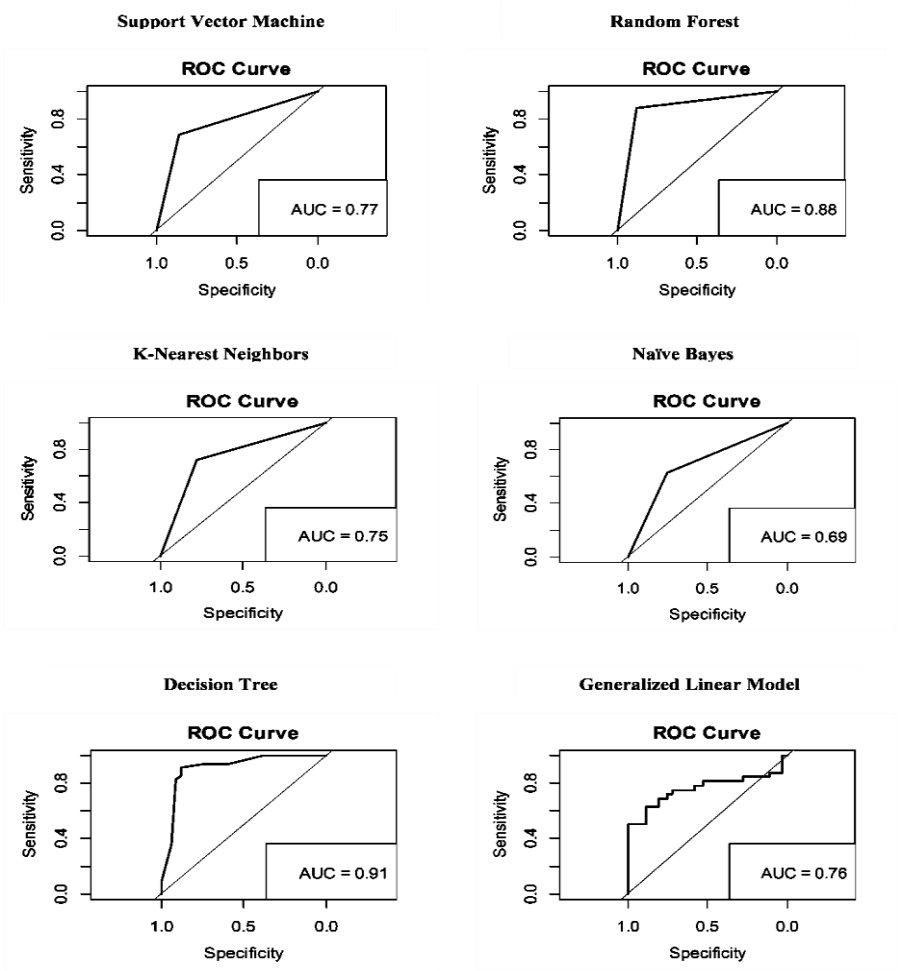

Fig. 2. Receiver Operating Characteristic Curve (ROC) of different ML model

(Nashif, 2018). Another research employed the random forest classifier technique is used by the system to diagnose cardiac illnesses with an accuracy rate of about 83% (Chang, 2022). Another recent study compared interpretable machine learning models for early differential diagnosis of ischemic heart disease (IHD) and dilated cardiomyopathy (DCM). The naive bayes model fared better than other models in that investigation, with a classification accuracy of 73.5% (Islam, 2023).

Moreover, another recent research using several machine learning classification techniques, including logistic regression, random forest classification, and k-nearest neighbors (KNN), constructed a cardiovascular disease detection model. The purpose of the study was to forecast whether a person will have cardiovascular disease based on their medical history, which includes information from a dataset on chest discomfort, blood sugar levels, and blood pressure and so on. With an average accuracy of 87.5%, the project accurately predicts patients who have been diagnosed with heart disease (Jindal, 2021). Also this study compared several machine learning model including support vector machine, random forest, KNN, naïve bayes, decision tree and GLM, to construct the best model to predict the heart related disease based on different risk factors. With 91% accuracy the study predicts the presence of heart disease includes information on blood pressure, random blood sugar, cholesterol and electrolytes test and so on. Furthermore, the recent study employed various supervised machine learning algorithms are compared for their internal validity and accuracy in predicting clinical occurrences (Latha, 2019). On the contrary, this study conducted the best classification model for predict heart related disease. Furthermore, the machine learning model employed previous studies for diagnosis heart disease and machine learning model has been widely used medical field for disease detection (Roy, 2022, Dai, 2022, Agrawal, 2022, Huang, 2022, Learning, 2017).

The study uncovered the interesting association between presence of heart disease and study characteristics. The result showed, decision tree model predict the risk factor with highest accuracy more than other models (SVM, Random Forest, k-NN, Naïve Bayes and GLM). ROC curve illustrate with sensitivity and specificity. Additionally, there have been more recent studies that have employed machine learning (ML) model for heart disease detection for clinical studies and showed impact of the severity of heart disease (Islam, 2020, Huang, 2022). Overall, this study found the accuracy of the decision tree model as 91% among the all ML models included in this study. This accuracy is much better than other study and also it shows that the ML is a reliable predictor model (Alizadeh, 2023).

5 Conclusion

The study involved conducting research on two groups of respondents: one comprising patients with heart disease and the other consisting of individuals without heart disease. Total 340 data was collected from both groups using the same sample size (170) for each. We conducted a comparative analysis of machine learning prediction models aimed at determining the best model in presence of risk factors associated with heart disease in patients. Among the various models assessed, the decision tree exhibited the highest accuracy in predicting the presence of heart disease among the respondents. This study underscores the value of machine learning models and emphasizes the significance of incorporating shared socio-demographic and background characteristics for accurate disease status predictions. Hence, we specifically selected studies that utilized multiple machine learning methods on identical datasets to predict diseases, allowing for direct comparisons.

Nevertheless, this research has its limitations. This study was limited due to the fact that it was carried out in only three hospital and other hospitals did not participate in this study. In line with recent data, our findings suggest that short-term outcomes predict the risk factors of heart disease. During the data collection phase, a notable limitation was the restricted availability of patients admitted to the hospital presenting with heart disease symptoms. This limited patient pool could potentially constrain the statistical power of the study and compromise the ability to draw robust conclusions. Collecting data and aligning it with the prevailing conditions in Bangladesh within a short period was indeed challenging. More precisely, the data gathered on heart disease is characterized by a limited sample size, posing a persistent constraint. In this context, future research endeavors could enhance this study by incorporating additional clinical, demographic, and social determinants. This expansion would serve to validate the existing findings and elevate the overall quality of the results obtained.

Disclaimer (Artificial Intelligence)

Author(s) hereby declare that NO generative AI technologies such as Large Language Models (ChatGPT, COPILOT, etc) and text-to-image generators have been used during writing or editing of this manuscript.

Consent

The sample was chosen from our target demographic using a reasonable sampling technique. Prior to completing the survey, patients were informed about the investigation's goal and given assurances about the privacy of their personal data. Two alternative agreement questions (Yes/No) are asked at the beginning of the form in an attempt to get the patients' verbal approval. A tiny subset of patients chose the No option, allowing them to leave after filling out the survey.

Ethical Approval

The study was carried out in accordance with ethical standards (as per The Code of Ethics of the World Medical Association) and approved by the Noakhali Science and Technology University Ethical Committee (NSTUEC). Along with being informed of the goals, benefits, and drawbacks of the study, participants' agreement was obtained legally and ethically.

Acknowledgements

The staffs who assisted in the data collection for this study, as well as all the participants who willingly donated their time and gave empathetic, honest comments, is all appreciated by the authors.

Competing Interests

Authors have declared that no competing interests exist.

References

Agrawal, R. K., Sewani, R. R., Delen, D., & Benjamin, B. (2022). A machine learning approach for classifying healthy and infarcted patients using heart rate variabilities derived vector magnitude. *Healthcare Analytics*, *2*, 100121. https://doi.org/10.1016/j.health.2022.100121

Ahmed, M. R., Mahmud, S. H., Hossin, M. A., Jahan, H., & Noori, S. R. H. (2018). A cloud-based four-tier architecture for early detection of heart disease with machine learning algorithms. In *2018 IEEE 4th International Conference on Computer and Communications (ICCC)* (pp. 1951–1955). IEEE. https://doi.org/10.1109/CompComm.2018.8781022

Ahsan, M. M., Luna, S. A., & Siddique, Z. (2022). Machine-learning-based disease diagnosis: A comprehensive review. *Healthcare*, *10*(3), 541. https://doi.org/10.3390/healthcare10030541

Alizadeh, B., Alibabaei, A., Ahmadi, S., Maroufi, S. F., Ghafouri-Fard, S., & Nateghinia, S. (2023). Designing predictive models for appraisal of outcome of neurosurgery patients using machine learning-based techniques. *Interdisciplinary Neurosurgery*, *31*, 101658. https://doi.org/10.1016/j.inat.2022.101658

Asare, J. W., Appiahene, P., & Donkoh, E. T. (2023). Detection of anaemia using medical images: A comparative study of machine learning algorithms—A systematic literature review. *Inform Med Unlocked*, 101283. https://doi.org/10.1016/j.imu.2023.101283

Behera, M. P., Sarangi, A., Mishra, D., & Sarangi, S. K. (2023). A hybrid machine learning algorithm for heart and liver disease prediction using modified particle swarm optimization with support vector machine. *Procedia Computer Science*, *218*, 818–827. https://doi.org/10.1016/j.procs.2023.01.062

Beunza, J.-J., Puertas, E., García-Ovejero, E., Villalba, G., Condes, E., Koleva, G., Hurtado, C., & Landecho, M. F. (2019). Comparison of machine learning algorithms for clinical event prediction (risk of coronary heart disease). *Journal of Biomedical Informatics*, *97*, 103257. https://doi.org/10.1016/j.jbi.2019.103257

Chang, V., Bhavani, V. R., Xu, A. Q., & Hossain, M. A. (2022). An artificial intelligence model for heart disease detection using machine learning algorithms. *Healthcare Analytics*, *2*, 100016. https://doi.org/10.1016/j.health.2022.100016

Chowdhury, M. N. R., Ahmed, E., Siddik, M. A. D., & Zaman, A. U. (2021). Heart disease prognosis using machine learning classification techniques. In *2021 6th International Conference for Convergence in Technology (I2CT)* (pp. 1–6). IEEE. https://doi.org/10.1109/I2CT51068.2021.9418181

Dai, Q., Sherif, A. A., Jin, C., Chen, Y., Cai, P., & Li, P. (2022). Machine learning predicting mortality in sarcoidosis patients admitted for acute heart failure. *Cardiovascular Digital Health Journal*, *3*, 297–304. https://doi.org/10.1016/j.cvdhj.2022.08.001

Hajian-Tilaki, K. (2013). Receiver operating characteristic (ROC) curve analysis for medical diagnostic test evaluation. *Caspian Journal of Internal Medicine*, *4*, 627.

Hossain, M. R. (2024). Optimizing heart disease prediction: A comparative study of machine learning techniques. *Asian Journal of Cardiology Research*, *7*(1), 348–359. https://doi.org/10.9734/ajcr/2024/v7i1238

Howlader, K. C., Satu, M. S., & Mazumder, A. (2017). Performance analysis of different classification algorithms that predict heart disease severity in Bangladesh. *International Journal of Computer Science and Information Security (IJCSIS)*, *15*, 332–340.

Huang, Y., Ren, Y., Yang, H., Ding, Y., Liu, Y., Yang, Y., Mao, A., Yang, T., Wang, Y., & Xiao, F. (2022). Using a machine learning-based risk prediction model to analyze the coronary artery calcification score and predict coronary heart disease and risk assessment. *Computers in Biology and Medicine*, *151*, 106297. https://doi.org/10.1016/j.compbiomed.2022.106297

Iscra, K., Miladinović, A., Ajčević, M., Starita, S., Restivo, L., Merlo, M., & Accardo, A. (2022). Interpretable machine learning models to support differential diagnosis between Ischemic Heart Disease and Dilated Cardiomyopathy. *Procedia Computer Science*, *207*, 1378–1387. https://doi.org/10.1016/j.procs.2022.09.194

Islam, A. M., & Majumder, A. A. S. (2013). Coronary artery disease in Bangladesh: A review. *Indian Heart Journal*, *65*, 424–435. https://doi.org/10.1016/j.ihj.2013.06.004

Islam, J. Y., Zaman, M. M., Moniruzzaman, M., Shakoor, S. A., & Hossain, A. E. (2020). Estimation of total cardiovascular risk using the 2019 WHO CVD prediction charts and comparison of population-level costs based on alternative drug therapy guidelines: A population-based study of adults in Bangladesh. *BMJ Open*, *10*, e035842. https://doi.org/10.1136/bmjopen-2019-035842

Islam, M. A., Majumder, M. Z. H., & Hussein, M. A. (2023). Chronic kidney disease prediction based on machine learning algorithms. *Journal of Pathology Informatics*, *14*, 100189. https://doi.org/10.1016/j.jpi.2023.100189

Jindal, H., Agrawal, S., Khera, R., Jain, R., & Nagrath, P. (2021). Heart disease prediction using machine learning algorithms. In *IOP Conference Series: Materials Science and Engineering* (Vol. 1022, p. 012072). IOP Publishing. https://doi.org/10.1088/1757-899X/1022/1/012072

Kabir, M. F., Chen, T., & Ludwig, S. A. (2023). A performance analysis of dimensionality reduction algorithms in machine learning models for cancer prediction. *Healthcare Analytics*, *3*, 100-125. https://doi.org/10.1016/j.health.2022.100125

Khan, J. R., Chowdhury, S., Islam, H., & Raheem, E. (2019). Machine learning algorithms to predict childhood anemia in Bangladesh. *Journal of Data Science*, *17*, 195–218. https://doi.org/10.6339/JDS.201901_17(1).0009

Khanna, V. V., Chadaga, K., Sampathila, N., Prabhu, S., & Chadaga, R. (2023). A machine learning and explainable artificial intelligence triage-prediction system for COVID-19. *Decision Analysis Journal*, 100246. https://doi.org/10.1016/j.dajour.2023.100246

Latha, C. B. C., & Jeeva, S. C. (2019). Improving the accuracy of prediction of heart disease risk based on ensemble classification techniques. *Inform Med Unlocked*, *16*, 100203. https://doi.org/10.1016/j.imu.2019.100203

Learning, M. (2017). Heart disease diagnosis and prediction using machine learning and data mining techniques: A review. *Advances in Computer Science and Technology*, *10*, 2137–2159.

Lin, H., Xue, Q., Feng, J., & Bai, D. (2023). Internet of things intrusion detection model and algorithm based on cloud computing and multi-feature extraction extreme learning machine. *Digital Communications and Networks*, *9*, 111–124. https://doi.org/10.1016/j.dcan.2022.09.021

Mashaleh, A. S., Ibrahim, N. F. B., Al-Betar, M. A., Mustafa, H. M., & Yaseen, Q. M. (2022). Detecting spam email with machine learning optimized with Harris Hawks optimizer (HHO) algorithm. *Procedia Computer Science*, *201*, 659–664. https://doi.org/10.1016/j.procs.2022.03.087

Nagavelli, U., Samanta, D., & Chakraborty, P. (2022). Machine learning technology-based heart disease detection models. *Journal of Healthcare Engineering*, *2022*. https://doi.org/10.1155/2022/7351061

Nashif, S., Raihan, M. R., Islam, M. R., & Imam, M. H. (2018). Heart disease detection by using machine learning algorithms and a real-time cardiovascular health monitoring system. *World Journal of Engineering and Technology*, *6*, 854–873. https://doi.org/10.4236/wjet.2018.64057

Nipa, N., Riyad, M. H., Satu, S., Howlader, K. C., & Moni, M. A. (2023). Clinically adaptable machine learning model to identify early appreciable features of diabetes in Bangladesh. *Intelligent Medicine*. https://doi.org/10.1016/j.imed.2023.01.003

Ozcan, M., & Peker, S. (2023). A classification and regression tree algorithm for heart disease modeling and prediction. *Healthcare Analytics*, *3*, 100130. https://doi.org/10.1016/j.health.2022.100130

Plati, D. K., Tripoliti, E. E., Bechlioulis, A., Rammos, A., Dimou, I., Lakkas, L., Watson, C., McDonald, K., Ledwidge, M., & Pharithi, R. (2021). A machine learning approach for chronic heart failure diagnosis. *Diagnostics*, *11*, 1863. https://doi.org/10.3390/diagnostics11101863

Ramalingam, V. V., Dandapath, A., & Raja, M. K. (2018). Heart disease prediction using machine learning techniques: A survey. *International Journal of Engineering and Technology*, *7*, 684–687. https://doi.org/10.14419/ijet.v7i2.8.10557

Roy, T. S., Roy, J. K., & Mandal, N. (2022). Classifier identification using deep learning and machine learning algorithms for the detection of valvular heart diseases. *Biomedical Engineering Advances*, *3*, 100035. https://doi.org/10.1016/j.bea.2022.100035

Sugendran, G., & Sujatha, S. (2023). Earlier identification of heart disease using enhanced genetic algorithm and fuzzy weight-based support vector machine algorithm. *Measurement and Sensors*, 100814. https://doi.org/10.1016/j.measen.2023.100814

Uddin, S., Khan, A., Hossain, M. E., & Moni, M. A. (2019). Comparing different supervised machine learning algorithms for disease prediction. *BMC Medical Informatics and Decision Making*, *19*(281). https://doi.org/10.1186/s12911-019-1004-8

Peer-review history:
The peer review history for this paper can be accessed here (Please copy paste the total link in your browser address bar)
https://www.sdiarticle5.com/review-history/129937