Journal of Artificial Intelligence, Machine Learning and Neural Network
Vol: 03, No. 02, July-Dec 2023
ISSN: 2799-1172, DOI: 10.55529/jaimlnn.32.78.85

Research Paper

Machine learning for heart disease prediction a comparison analysis

Nishat Tasnim[1], Kazi Tanvir[2*], Sanjid Bin Karim Sezan[3]
[1,2*,3]Department of Computer Science, American International University-Bangladesh (AIUB), Kuratoli, Dhaka, Bangladesh.

Article Info

Article History:
Received: 24 May 2023
Revised: 03 August 2023
Accepted: 10 August 2023
Published: 27 September 2023

Keywords:
Heart Diseases
Machine Learning Algorithms
Logistic Regression
Decision Tree
SVM
Naive Bayes

ABSTRACT

Predicting cardiac conditions remains one of the most formidable tasks within the medical field today, with heart disease claiming a life every minute in the contemporary landscape. The data-rich healthcare industry necessitates the application of data science for efficient data processing. Given the intricate nature of prognosticating heart-related disorders, the automation of this process becomes a necessity, aiming to mitigate potential risks and offer timely alerts to patients. In this research endeavor, the heart disease dataset extracted from the UCI machine learning repository is employed. The proposed study embraces an array of data mining strategies, encompassing Logistic Regression, Decision Tree, Support Vector Machine (SVM), and Naive Bayes algorithm, to anticipate the likelihood of Heart Disease and stratify patient risk levels. This article undertakes a comparative analysis of various machine learning algorithms to assess their effectiveness. The trial outcomes indicate that, compared to other utilized ML algorithms, Support Vector Machine (SVM) emerges with the highest accuracy, registering at 90.48%.

Corresponding Author:
Kazi Tanvir
Department of Computer Science, American International University-Bangladesh (AIUB), Kuratoli, Dhaka, Bangladesh.
Email: kazitanvir.ai@gmail.com

1. INTRODUCTION

The heart stands as the vital cornerstone of the human body, holding the utmost significance. The functionality of the heart stands as the linchpin on which human existence hinges. Optimal heart function

paves the way for a life imbued with well-being. However, in the contemporary setting, cardiovascular ailments have risen to prominence, contributing significantly to mortality rates among both men and women. The advent of the Covid-19 pandemic has introduced a slew of physical complications, with the virus being linked to adverse cardiac effects. Cardiac inflammation induced by the coronavirus is accountable for instances of heart failure. Notwithstanding the presence of respiratory indications, empirical investigations reveal that a significant proportion of patients, specifically 20%, exhibited cardiac harm resulting from the impact of the coronavirus.

Amongst the various forms of cardiovascular ailments, coronary heart disease prevails as the most common. Approximately 25% of all mortalities, equating to roughly 630,000 cases, stem from heart-related conditions. The evaluation of a patient's medical background, familial records, a thorough physical assessment, and the outcomes of medical tests frequently constitute the fundamental basis for diagnosing instances of cardiac failure [1]. Diagnosing heart disease can prove to be complex due to a multitude of risk elements, encompassing diabetes, elevated blood pressure, increased cholesterol levels, an erratic heart rate, and various additional medical conditions [2]. Given its widespread prevalence, timely and accurate detection of heart disease becomes a pressing necessity to safeguard numerous lives. While an array of scanning techniques exists for identifying cardiac ailments, the potential to anticipate a heart condition prior to its overt manifestation holds the promise of rescuing a significant number of individuals. Through the utilization of a tool facilitating visual evaluation of patient data, we furnish supplementary information to healthcare administrators, enhancing their insights. Detecting and analyzing the presence of arrhythmia at the earliest opportunity is essential to prevent the onset of cardiac problems in individuals [3].

In numerous scenarios, the presence of minimal levels of cardiac rhythm can be attributed to the occurrence of stroke or heart failure. The healthcare domain holds significant potential for harnessing machine learning to support health systems in evaluating and diagnosing ailments by leveraging comprehensive data mining, including habits and elevated cholesterol levels, to enhance this latter group. The domain of data mining within machine learning, which adeptly handles extensive and well-structured datasets, accomplishes this task effectively. Machine learning possesses the capacity to contribute to medical practice by identifying, detecting, and prognosticating a diverse array of medical conditions. The primary aim of this paper centres on providing healthcare practitioners with a tool for the early detection of heart disease, thus enabling timely intervention and averting severe consequences. The indispensability of machine learning (ML) in discerning concealed discrete patterns and comprehensively analysing the furnished data holds immense significance. Subsequent to meticulous data scrutiny, machine learning methodologies facilitate the prompt identification and anticipation of cardiac disorders. This study, geared towards pre-emptively identifying cardiac ailments, evaluates the effectiveness of various ML techniques, encompassing Naive Bayes, Decision Trees, Logistic Regression, and Support Vector Machines (SVM).

2. RELATED WORK

Numerous studies have been conducted utilizing the UCI Machine Learning dataset [4] to forecast heart disease. Diverse data mining methodologies have been employed, yielding varying degrees of accuracy, as elaborated in the subsequent sections. A study led by Avinash Golande and colleagues delves into various machine learning algorithms to classify cardiac disease, examining the accuracy of Decision Tree, KNN, and K-Means classification techniques [1]. The research revealed that Decision Trees exhibited the highest level of accuracy, leading to the conclusion that enhanced effectiveness could potentially be achieved by amalgamating different methodologies and optimizing its parameters. T. Nagamani proposed an integrated system that merges the MapReduce algorithm with data mining methods, resulting in greater accuracy compared to a conventional fuzzy artificial neural network for the 45 instances in the test set; the incorporation of dynamic schema and linear scaling contributed to the enhanced accuracy of the algorithm [2]. Anjan Nikhil Repaka put forward a system that utilizes Naive Bayesian (NB) techniques for dataset categorization and integrates the Advanced Encryption Standard (AES) algorithm for secure data transportation [5]. Upon reviewing the aforementioned research works, the core idea underpinning the proposed system was to develop a predictive model for heart disease by utilizing input data. To identify the

optimal classification algorithm for heart disease prediction, a comparison was conducted among Logistic Regression [6], Decision Tree [7], Support Vector Machine (SVM) [8], and Naive Bayes classification [9] algorithms, evaluating their Accuracy, Precision, Recall, and f-measure metrics.

2.1 Classifications

2.1.1 Logistic Regression

Derived from a provided dataset of unconnected factors, logistic regression computes the probability of an event, like choosing to vote or not. The outcome, presented as a probability, confines the reliant variable within the range of 0 to 1. In logistic regression, a logit transformation is applied to the odds, encompassing the ratio of the probability of success to the probability of failure [6]. This is frequently referred to as the log odds, or the logarithm of odds, and the logistic function is depicted through the subsequent equations:

$$\text{Logit}(\pi) = \frac{1}{1+e^{\pi}} \quad \ldots\ldots (1)$$

$$In\left(\frac{\pi}{1-\pi}\right) = \beta_0 + \beta_1 X_1 + \cdots + \beta_k X_k \quad \ldots\ldots (2)$$

2.1.2 Decision Tree

Functioning as a supervised learning technique, the Decision Tree is adept at handling classification and regression problems, although its primary application lies in classification tasks [7]. It embraces a hierarchical configuration reminiscent of a tree, where inner nodes represent dataset attributes, branches symbolize decision conditions, and every terminal leaf node signifies a final result. The fundamental aim of the decision tree algorithm is to progressively enhance information gain, prioritizing the partition of nodes/attributes that yield the utmost information gain. This can be computed using the subsequent equation:

$$\text{Information Gain} = \text{Entropy}(S) - \lfloor(\text{Weighted Avg}) \times (\text{Entropy(each feature)})\rfloor \ldots\ldots (4)$$

$$\text{Entropy}(S) = -P(yes)log_2P(yes) - P(no)log_2P(no) \ldots\ldots (5)$$

2.1.3 Support Vector Machine (SVM)

Support Vector Machine (SVM), a widely favoured supervised learning algorithm, finds utility in addressing both Classification and Regression challenges, with a primary focus on Machine Learning Classification tasks [8] Diving into specifics, Support Vector Machine (SVM) stands out as a prominent choice for resolving Classification and Regression problems within the realm of supervised learning. While it demonstrates versatility, its predominant application remains cantered on Machine Learning Classification problems. Regarding the Naive Bayes algorithm, it operates as a probabilistic classifier, marked by its foundation on probability models incorporating robust assumptions of independence [10]. These assumptions, although often oversimplified, are dubbed as "naive" due to their idealized nature. Notably, the algorithm finds a niche in text classification scenarios involving intricate high-dimensional training datasets. The ensuing equation is as follows:

$$P(A|\text{B}) = \frac{P(\text{B}|\text{A})\times\text{P(A)}}{\text{P(B)}} \ldots\ldots (5)$$

Where, P (A|B) is Posterior probability: Probability of hypothesis A on the observed event B. P (B|A) is Likelihood probability: Probability of the evidence given that the probability of a hypothesis is true.

3. METHODOLOGY

We evaluated four highly accurate machine learning techniques specific to this predictive model: Logistic Regression, Decision Tree, Support Vector Machine (SVM), and the Naive Bayes classification algorithm. The proposed approach follows this sequence: commencing with data collection, followed by substantial value extraction, and subsequently engaging in data exploration. Data preparation involves addressing missing values, data purification, and standardization as per the adopted techniques. Once the

pre-processed data is ready, the classifier employed in the proposed models is employed to discern the pre-processed information. Subsequently, we subjected the suggested model to testing, evaluating its efficacy and precision using a variety of performance metrics. For testing purposes, half of the complete dataset was employed.

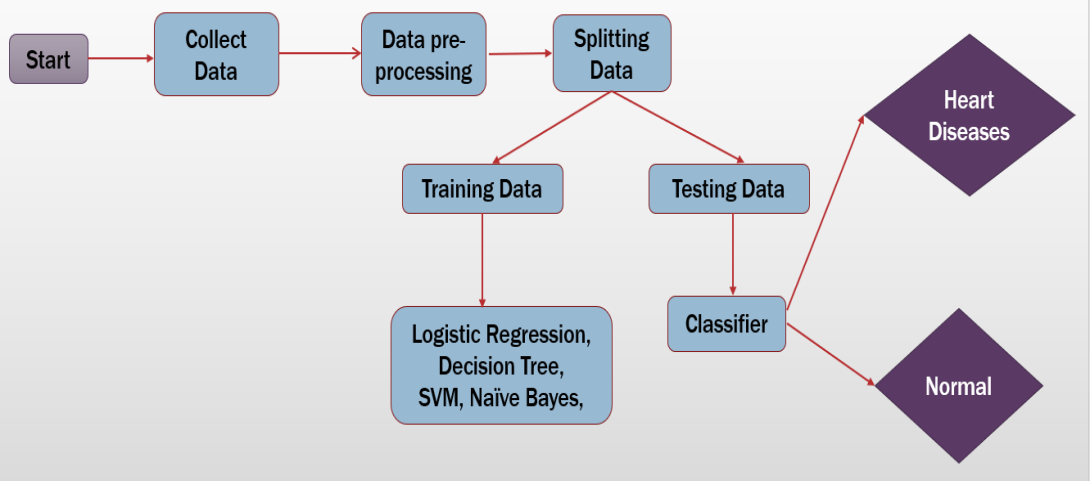

Figure 1. Model of Methods

3.1 Data Source

For model training, data was acquired from The UCI repository's database was utilized [4]. A subset of 14 attributes were utilized instead of the original 76 attributes. The collection contains information on a wide range of people, including their medical histories and histories of heart disease. The dataset comprises of the medical histories of 303 distinct people with a variety of features. The patient's medical features, including age, the type of chest discomfort experienced, blood pressure, sugar levels, angina, and other factors, are well-detailed in this dataset, allowing us to determine whether or not the patient has been given a heart disease diagnosis. The following characteristics are listed:

SL. No	Observation	Description
1.	Age	Age in years
2.	Sex	Sex of patients
3.	CP	Chest pain
4.	Trestbps	Resting blood pressure
5.	Chol	Serum cholesterol
6.	FBS	Fasting blood pressure
7.	Restecg	Resting electrocardiograph results
8.	Thalach	Maximum heart rate achieved
9.	Exang	Exercise-induced angina
10.	Oldpcak	ST depression induced by excercise relative to rest
11.	Slope	the slope of the peak exercise ST segment
12.	Ca	number of major vessels colored by fluoroscopy
13.	thal	Defect type
14.	target	

Figure 2. Dataset Overview

4. RESULTS AND DISCUSSION

This segment presents the results stemming from the utilization of Logistic Regression, Decision Tree, Support Vector Machine (SVM), and Naive Bayes. The performance of these algorithms is assessed

using metrics such as Accuracy score, Precision (P), Recall (R), and F-measure. The precision metric serves as an accurate indicator of positive evaluation, while recall quantifies the number of true positive instances [11]. The F-measure can be seen as a middle ground between recall and precision, achieving high values only when both recall and precision are elevated. It is tantamount to recall when α = 0 and precision when α = 1. The F-measure takes on values within the range of [0, 1] [12].

$$\text{Precision} = (TP) / (TP + FP) \ldots\ldots\ldots (6)$$

$$\text{Recall} = (TP) / (TP + FN) \ldots\ldots\ldots (7)$$

$$\text{F- Measure} = \frac{(2 \times \text{Precision} \times \text{Recall})}{\text{Precision} + \text{Recall}} \ldots\ldots\ldots (8)$$

TP (True Positive) signifies cases where the patient has the disease and the test yields a positive result, while FP (False Positive) corresponds to instances where the test produces a positive outcome despite the disease's absence in the patient; TN (True Negative) represents scenarios where the test correctly identifies the lack of disease in patients, and FN (False Negative) indicates situations where the test erroneously indicates a negative result despite the patient being afflicted with the disease [13].

In our result we obtained,

Actual 0 = True negative

Actual 1= True positive

Prediction 0 = False negative

Prediction 1 = False positive

Table 1. Obtaining Accuracy for Different Model

Sr. No	Model	AUC-ROC	Accuracy	Precision	Recall
1	Logistic Regression	0.947232	0.897959	0.9472029	0.855263
2	Decision Tree Classifier	0.950855	0.884354	0.898551	0.861111
3	Support Vector Machine	0.951133	0.904762	0.956522	0.857143

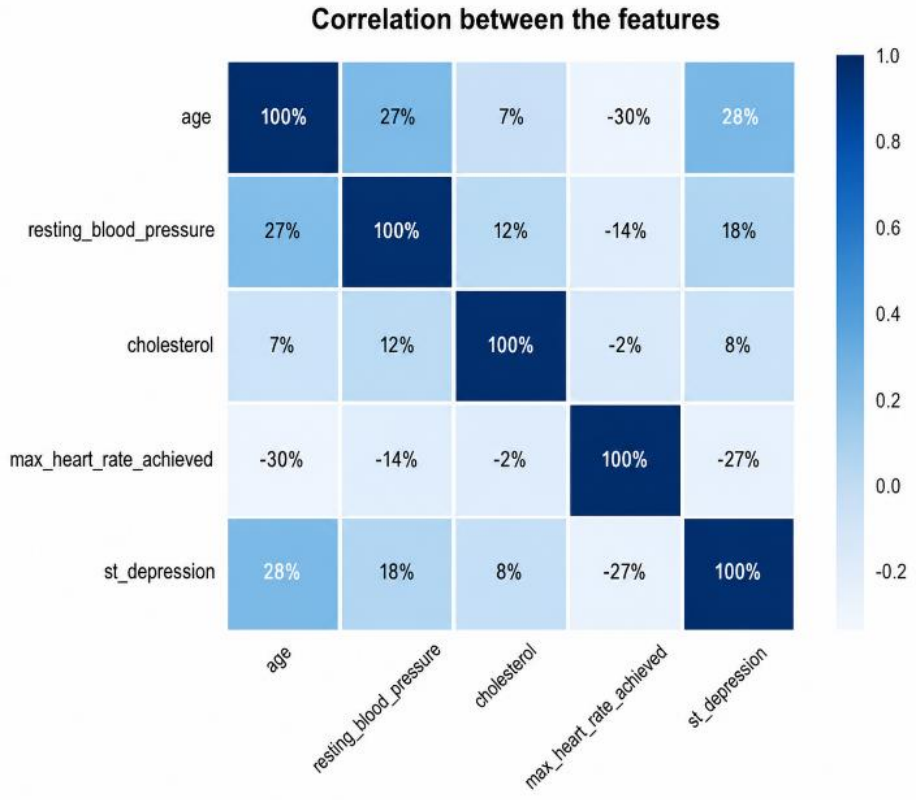

Figure 3. Correlation Matrix of SVM

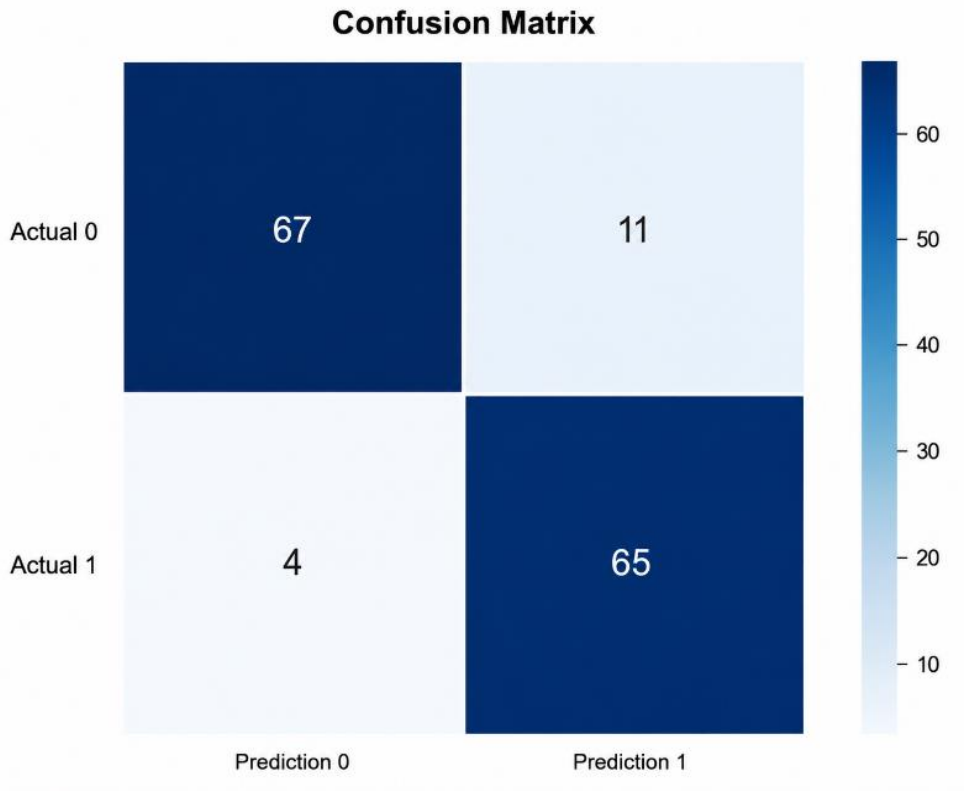

Figure 4. Confusion Matrix of SVM

The indicated performance measures are derived through the utilization of the confusion matrix, which delineates the model's effectiveness. The confusion matrix generated by the suggested model across various algorithms is presented earlier. The accuracy scores achieved for Logistic Regression, Decision Tree, Support Vector Machine (SVM), and Naive Bayes classification methods are depicted in Figure 3 above.

5. CONCLUSION

Given the escalating fatality rates linked to heart conditions, it becomes imperative to establish a precise and efficient system for heart disease prediction. The impetus driving this study was to identify the most optimal machine learning algorithm for accurate heart disease diagnosis. Employing the UCI machine learning repository dataset, this research evaluates the accuracy performances of Logistic Regression, Decision Tree, Support Vector Machine (SVM), and Naive Bayes algorithms in heart disease prediction. The study's findings reveal that the Support Vector Machine (SVM) algorithm stands out as the most effective, boasting a noteworthy accuracy rate of 90.48%.

Potential enhancements for future investigations include the development of a web-based application grounded in the Support Vector Machine (SVM) approach and the incorporation of a more extensive dataset compared to the current research. Such endeavors would likely yield more robust insights, assisting medical professionals in achieving precise and efficient cardiac disease predictions.

Acknowledgments

The authors have no specific acknowledgments to make for this research.

Funding Information

This research received no specific grant from any funding agency in the public, commercial, or not-for-profit sectors.

Author Contributions Statement

Name of Author	C	M	So	Va	Fo	I	R	D	O	E	Vi	Su	P	Fu
Nishat Tasnim	✓	✓	✓	✓		✓		✓	✓	✓	✓			
Kazi Tanvir		✓				✓	✓				✓	✓		✓
Sanjid Bin Karim Sezan	✓			✓	✓			✓		✓		✓		

C: **C**onceptualization
M: **M**ethodology
So: **So**ftware
Va: **Va**lidation
Fo: **Fo**rmal analysis
I: **I**nvestigation
R: **R**esources
D: **D**ata Curation
O: Writing- **O**riginal Draft
E: Writing- Review& **E**diting
Vi: **Vi**sualization
Su: **Su**pervision
P: **P**roject administration
Fu: **Fu**nding acquisition

Conflict of Interest Statement

The authors declare that there are no conflicts of interest regarding the publication of this paper.

Informed Consent

All participants were informed about the purpose of the study and their voluntary consent was obtained prior to data collection.

Ethical Approval

The study was conducted in compliance with the ethical principles outlined in the Declaration of Helsinki and approved by the relevant institutional authorities.

Data Availability

The data that support the findings of this study are available from the corresponding author upon reasonable request.

REFERENCES

[1] A. Golande, 'Heart Disease Prediction Using Effective Machine Learning Techniques', vol. 8, 2019.
[2] T. Nagamani, S. Logeswari, and B. Gomathy, 'Heart Disease Prediction using Data Mining with Mapreduce Algorithm', vol. 8, 2019.
[3] M. Shahreyar, R. Fahhoum, O. Akinseye, S. Bhandari, G. Dang, and R. N. Khouzam, 'Severe sepsis and cardiac arrhythmias', Ann. Transl. Med., vol. 6, no. 1, p. 6, Jan. 2018. doi.org/10.21037/atm.2017.12.26
[4] Andras Janosi, William Steinbrunn, Matthias Pfisterer, Robert Detrano, 'Heart Disease'. UCI Machine Learning Repository, 1989.
[5] "Design And Implementing Heart Disease Prediction Using Naives Bayesian | Semantic Scholar." https://www.semanticscholar.org/paper/Design-And-Implementing-Heart-Disease-Prediction-Repaka-Ravikanti/d1038f406d8662d07b4d95c22ff008f9307043c0 (accessed Aug. 14, 2023).
[6] "What is Logistic regression? | IBM." https://www.ibm.com/topics/logistic-regression (accessed Aug. 14, 2023).
[7] "Decision Tree Algorithm in Machine Learning - Javatpoint." https://www.javatpoint.com/machine-learning-decision-tree-classification-algorithm (accessed Aug. 14, 2023).
[8] "Support Vector Machine (SVM) Algorithm - Javatpoint." https://www.javatpoint.com/machine-learning-support-vector-machine-algorithm (accessed Aug. 14, 2023).
[9] "Naive Bayes Classifier in Machine Learning - Javatpoint." https://www.javatpoint.com/machine-learning-naive-bayes-classifier (accessed Aug. 14, 2023).
[10] F.-J. Yang, 'An implementation of naive Bayes classifier', in 2018 International Conference on Computational Science and Computational Intelligence (CSCI), Las Vegas, NV, USA, 2018. doi.org/10.1109/CSCI46756.2018.00065

[11] "Classification: Precision and Recall | Machine Learning | Google for Developers." https://developers.google.com/machine-learning/crash-course/classification/precision-and-recall (accessed Aug. 14, 2023).
[12] S. Dumais et al., 'F-Measure', in Encyclopedia of Database Systems, Boston, MA: Springer US, 2009, pp. 1147-1147. doi.org/10.1007/978-0-387-39940-9_483
[13] R. Parikh, A. Mathai, S. Parikh, G. C. Sekhar, and R. Thomas, 'Understanding and using sensitivity, specificity and predictive values', Indian J. Ophthalmol, vol. 56, no. 1, pp. 45-50, 2008. doi.org/10.4103/0301-4738.37595

How to Cite: Nishat Tasnim, Kazi Tanvir, Sanjid Bin Karim Sezan. (2023). Machine learning for heart disease prediction a comparison analysis. Journal of Artificial Intelligence, Machine Learning and Neural Network (JAIMLNN), 3(2), 78–85. https://doi.org/10.55529/jaimlnn.35.28.35

BIOGRAPHIES OF AUTHORS

	**Nishat Tasnim,** received her degree in Computer Science from American International University-Bangladesh. Her research interests include machine learning, artificial intelligence, healthcare analytics, and predictive data modelling. She has worked on projects that develop disease prediction systems through the application of supervised learning algorithms. Her academic work concentrates on using machine learning techniques to benefit healthcare operations through their application in early diagnosis and decision-making processes. Email: nishatasnim.aiub@gmail.com
	**Kazi Tanvir,** is affiliated with the Department of Computer Science at American International University-Bangladesh. His research areas include machine learning, data mining, artificial intelligence, and intelligent healthcare systems. He has contributed to several studies that use predictive analytics and classification models. He specializes in creating efficient AI systems that medical professionals can use to diagnose patients while analyzing their healthcare information. Email: kazitanvir.ai@gmail.com
	**Sanjid Bin Karim Sezan,** completed his studies in Computer Science at American International University-Bangladesh. His research interests include data science, machine learning applications, medical data analysis, and software development. He has participated in research about heart disease prediction through the use of classification algorithms and performance analysis between different methods. His work uses intelligent computing methods to create disease prediction systems which medical professionals can use to achieve accurate and trustworthy results. Email: sanjidsezan143@gmail.com