International Journal of Computing and Artificial Intelligence

E-ISSN: 2707-658X
P-ISSN: 2707-6571
www.computersciencejournals.com/ijcai
IJCAI 2025; 6(1): 66-70
Received: 05-12-2024
Accepted: 08-01-2025

Imran Khan
Assistant Professor, Department of Computer Science and Engineering, Jhulelal Institute of Technology, Nagpur, Maharashtra, India

Sumit Jaisinghani
Department of Computer Science and Engineering, Jhulelal Institute of Technology, Nagpur, Maharashtra, India

Anish Mankani
Department of Computer Science and Engineering, Jhulelal Institute of Technology, Nagpur, Maharashtra, India

Hitesh Motwani
Department of Computer Science and Engineering, Jhulelal Institute of Technology, Nagpur, Maharashtra, India

Devansh Motwani
Department of Computer Science and Engineering, Jhulelal Institute of Technology, Nagpur, Maharashtra, India

Juhi Sawlani
Department of Computer Science and Engineering, Jhulelal Institute of Technology, Nagpur, Maharashtra, India

Sahil Dhamecha
Department of Computer Science and Engineering,
Jhulelal Institute of Technology, Nagpur, Maharashtra, India

Corresponding Author:
Imran Khan
Assistant Professor, Department of Computer Science and Engineering, Jhulelal Institute of Technology, Nagpur, aharashtra, India

AI heart predictor: A smart model for assessing cardiac health

Imran Khan, Hitesh Motwani, Sahil Dhamecha, Sumit Jaisinghani, Devansh Motwani, Anish Mankani and Juhi Sawlani

DOI: https://www.doi.org/10.33545/27076571.2025.v6.i1a.136

Abstract

There are several diseases worldwide, many of which have well-established treatments. However, it is difficult to determine whether a person is experiencing a heart attack in the absence of a doctor. Our model can predict whether a person is experiencing a heart attack, even in the absence of a doctor. For example, Apple Watches can assist in the early identification of heart attack symptoms. We used large heart disease dataset from Kaggle.com and then performed Data preprocessing in it and we used various Classifiers for prediction. We used multiple classifiers among many others, including Random Forest, Decision Tree, and Logistic Regression The accuracy can be measured by using metrics such as F1 Score, Precision. Confusion matrix, Classification report. The model achieved accuracy of approximately 85% to 90% which is prominent given the dataset contains more than ten predictive features. Between the classifiers used, Random Forest achieved the highest accuracy. Logistic Regression and KNN obtained 86% and 71% accuracy, correspondingly. We also used other classifiers, such as SVM and K-means, but the highest accuracy was observed in Random Forest, Logistic Regression, and KNN. Accuracy can depend on multiple factors like independent variable, data preprocessing techniques, feature selection etc. So, it is basically depended on data and we have found best classifier for it. Our study's utilization of contemporary machine learning methodologies and active visualization tools, such Matplotlib and Seaborn, to enhance interpretability is a significant strength. To improve the display of the results, we utilized Matplotlib and Seaborn for visualization. In this outcome shows that the Random Forest and decision tree classifiers has maximum presentation for use to predict heart disease prediction. Heart disease prediction is crucial, especially in rural areas where access to medical professionals is limited. Our model can help in early detection, allowing individuals to seek medical attention sharp. At that time our model helps them to recognize whether a person has a heart attack or not. In this work, the predictive power of machine learning for heart disease is demonstrated using medical and demographic data. The results show that models such as decision trees and random forests may provide very accurate predictions, which could aid physicians in making earlier diagnoses.

Keywords: Random forest, confusion matrix, apple watches, accuracy, visualization

Introduction

Cardiovascular diseases, particularly heart-related conditions, rank as the foremost cause of death worldwide, with a pronounced impact in Central Asia and Europe, including nations like Russia, Ukraine, and Kazakhstan. In Central Asia, the mortality rate linked to ischemic heart disease (IHD) is notably high, driven by factors such as excessive sodium intake, tobacco use, and poor metabolic health, reaching approximately 265.51 deaths per 100,000 people. The World Health Organization (WHO) reports that heart disease accounts for around 17.9 million fatalities each year globally. To combat this alarming trend, early detection and proactive intervention strategies are essential for reducing mortality rates and enhancing public health. However, traditional analytic methods often require medical expertise and expensive tests, making them unreachable in remote areas. Heart disease detection is a critical matter, particularly in rural areas where medical assistance may be unavailable. In such cases, our model can help determine whether a person is experiencing a heart attack. In such situations, our model can assess the patient’s condition by noticing symptoms such as chest pain, shortness of breath, sweating, nausea, or pain in the left arm, jaw, or bac. let us understand why heart attack happen it happen because when blood flow to the heart is suddenly blocked. This can damage the heart and require immediate medical help. Doctor look at the symptoms and find out that if person is having a heart attack or not. Heart attacks have become more frequent in recent years, whereas in the past people were less aware of it. The heart attack is commonly and most increasing disease in the world it is so common that even a 25 years old boy had a heart attack. One of the major reasons for the rising incidence of heart disease is poor dietary habits, including excessive consumption of fast food from places like McDonald’s and KFC. people are eating this without knowing that if they are fresh or not, the oil they are using is change or not because many hotels owner use oil which is too old. Occasionally consuming fast food is acceptable, but a lack of regular exercise or yoga can contribute to health issues. If individuals consume such food, they must engage in physical activity to maintain a healthy weight. Another major reason is Stress. people are taking too much stress on small thing if we think too much amount something it can affect our heart. These are some of the major causes of heart attacks. While other risk factors occur, early prediction through our model can help patients seek medical assistance sooner, potentially saving lives. We gathered data for this study from Kaggle.com, which provides all the characteristics required to determine if a person has heart disease or not. Using patient data, machine learning (ML) has emerged as a potent tool for medical diagnostics, allowing for high-accuracy and robotic predictions. By analysing vast datasets, identifying intricate patterns, and performing computations in real time, machine learning algorithms might lessen reliance on conventional clinical assessments. Possible advantages of these models include enhanced physician decision assistance, quicker detection, and accessibility in remote locations.

Chest pain (CP) can be classified into four separate categories: 0 for typical angina, 1 for atypical angina, 1 for non-lethal pain, 2, and 3 for other types. Dataset also includes various patient characteristics, such as age, which are important for analysis. Additionally, it features information on Oldpec, which refers to exercise-induced ST segment depression, and exercise-induced angina, where a value of 1 indicates the presence of angina and 0 indicates its absence.

The slope of the ST segment during exercise is categorized as follows: 0 for upsloping, 1 for flat, and 2 for down sloping. Other significant attributes in the dataset include the maximum heart rate (troch), RSTCG results from resting electrocardiograms (0 for normal behaving wave, 1 for ST-T wave abnormalities, and 2 in case of left ventricular hypertrophy), and fasting blood sugar levels (FBS), where a value greater than 120 mg/dl is marked as 1 (yes) and 0 (no). Lastly, the Thal (CAR) variable indicates the type of defect, with 1 representing a normal condition, 2 indicating a fixed defect, and 3 denoting a reversible defect.

Machine learning (ML), as opposed to conventional techniques, has improved the accuracy of early cardiac disease detection in recent years. Three primary categories of cardiac disease detection methods were identified by Hajiarbabi (2024): X-ray image-based detection, ECG and PCG signal analysis, and clinical data-based methods. When used on clinical data, the study discovered that algorithms such as XGBoost, Random Forest, and Neural Networks outperformed more traditional machine learning methods. Additionally, by using advance machine learning techniques like Principal Component Analysis (PCA) and feature selection, the model's accuracy was raised. Convolutional Neural Networks (CNNs) produced the greatest outcomes for ECG-based diagnosis.

Based on these results, our research aims to predict heart disease using 13 clinical characteristics, including age, sex, blood pressure, and cholesterol levels. We tried with numerous machine learning models.

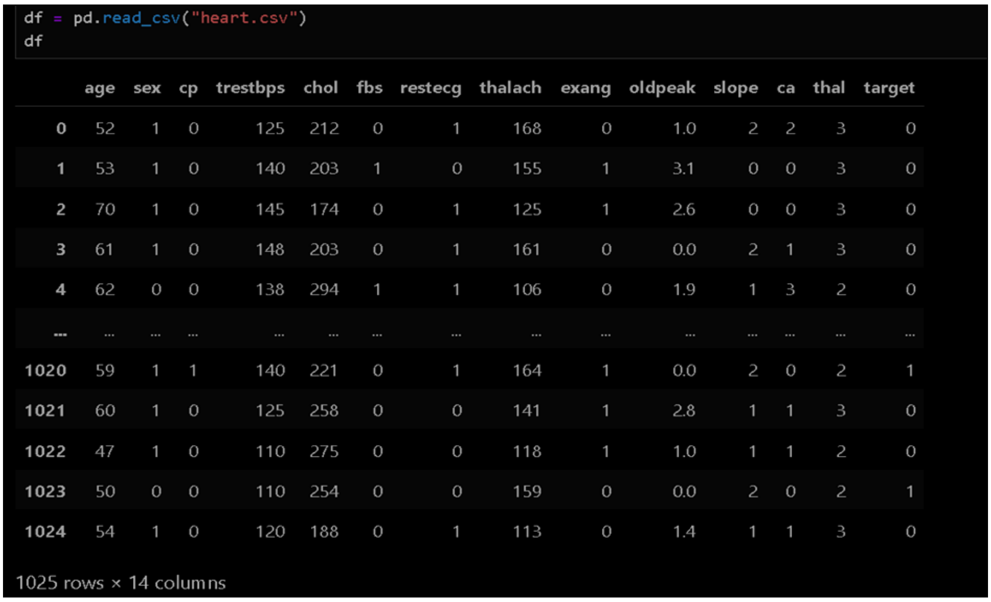
Our study's goal is to create a heart disease prediction model that is both extremely accurate and economical. We compare the accuracy of different models and determine that Random Forest achieves the highest accuracy. In this we make sure that there is no imbalanced data for any values in column and if we find any values null, we handle it precisely. Blood pressure is one of the most serious features in our dataset, as it is strongly linked to heart disease. However, other features also play a important role in prediction. To improve prediction accuracy, we perform feature selection so that only relevant feature can used for prediction and accuracy can be monitored through confusion matrix, f1 score and many more. Our ultimate goal is to enable early cardiac disease detection while minimizing implementation costs, especially in locations with limited resources. A publicly available dataset from Kaggle.com, comprising 13 samples and a single target column with 0 signifying no heart disease and 1 signifying heart illness, was used by. Additionally, we might Future research may include data from real-time monitoring systems, wearable technology, or further heart disease data. Only 13 features are available at the moment, but as time passes, new features will become available to improve our ability to forecast heart disease.

Methodology

This section outlines the methodology employed to develop and assess the heart disease prediction model. It encompasses data preprocessing techniques, the machine learning algorithms utilized, evaluation metrics, and details about the dataset.

Dataset Overview

The dataset used in this study was sourced from Kaggle.com and includes essential features vital for predicting heart disease. It comprises patient-specific data such as age, gender, resting blood pressure, cholesterol levels(chol), fasting blood sugar(fbs), and types of chest pain(cp), among other important health indicators. The target variable is a binary classification indicating whether heart disease is present (true) or absent (false.).



```
df = pd.read_csv("heart.csv")
df
```

	age	sex	cp	trestbps	chol	fbs	restecg	thalach	exang	oldpeak	slope	ca	thal	target
0	52	1	0	125	212	0	1	168	0	1.0	2	2	3	0
1	53	1	0	140	203	1	0	155	1	3.1	0	0	3	0
2	70	1	0	145	174	0	1	125	1	2.6	0	0	3	0
3	61	1	0	148	203	0	1	161	0	0.0	2	1	3	0
4	62	0	0	138	294	1	1	106	0	1.9	1	3	2	0
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...
1020	59	1	1	140	221	0	1	164	1	0.0	2	0	2	1
1021	60	1	0	125	258	0	0	141	1	2.8	1	1	3	0
1022	47	1	0	110	275	0	0	118	1	1.0	1	1	2	0
1023	50	0	0	110	254	0	0	159	0	0.0	2	0	2	1
1024	54	1	0	120	188	0	1	113	0	1.4	1	1	3	0

1025 rows × 14 columns

Fig 1: A showing of the dataset used in this study, showing patient attributes and the target variable.

The dataset contains 1,025 rows and 14 features, ensuring a various representation of patients.

Data Preprocessing

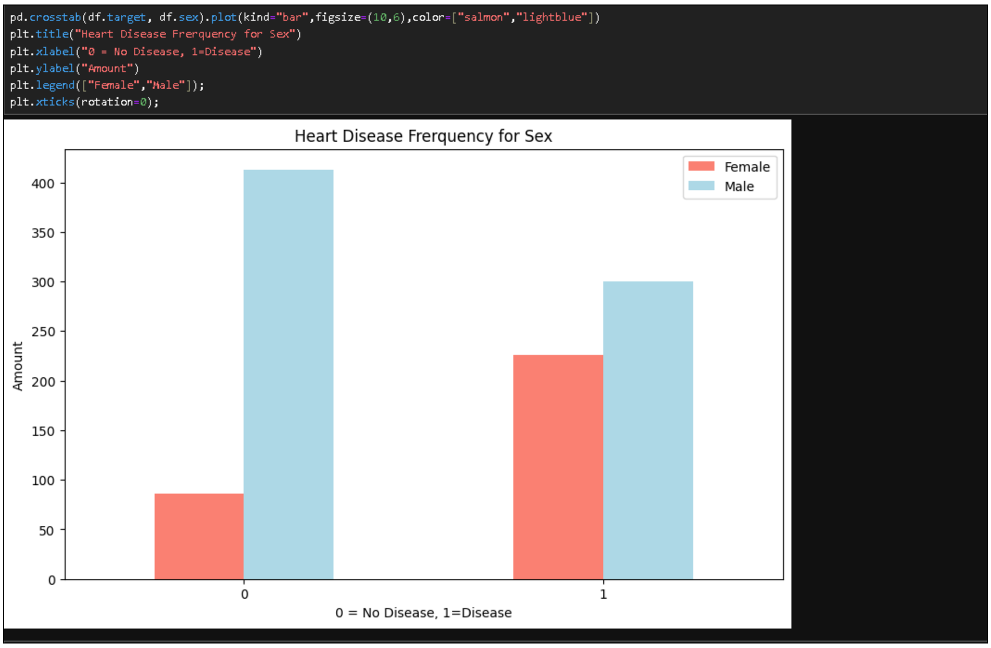
To enhance model performance and reliability, several pre-processing techniques were applied:

- **Handling Missing Values:** Any missing values were carefully addressed using mean, median or mode.
- Feature scaling was used to guarantee consistency across various characteristics by standardizing continuous numerical variables, such as blood pressure and cholesterol levels.
- **Class Imbalance Management:** To stop biased model predictions, techniques like Synthetic Minority Over-sampling Technique (SMOTE) were active where necessary.
- **Feature Selection:** To improve efficiency, irrelevant or redundant features were removed using statistical and machine learning-based choice techniques.

Machine Learning Models Used

To find the most effective algorithm for predicting heart disease, a variety of machine learning models were analyzed. The study included several techniques, such as logistic regression, which is a fundamental method for binary classification. Another approach, random forests, improves prediction accuracy by combining the results from multiple decision trees. Support vector machines (SVMs) are used to determine the best hyperplane to classify data by creating separate partitions. K-nearby neighbouring algorithm is a non-parametric method that depends on the distance matrix for classification purposes.

Additionally, Xgboost is a popular boosting technique known for its strong performance with a structured dataset. After comprehensive evaluation, random forest classifier showed more accuracy and reliability in predicting heart disease than other models.



```
pd.crosstab(df.target, df.sex).plot(kind="bar",figsize=(10,6),color=["salmon","lightblue"])
plt.title("Heart Disease Frerquency for Sex")
plt.xlabel("0 = No Disease, 1=Disease")
plt.ylabel("Amount")
plt.legend(["Female","Male"]);
plt.xticks(rotation=0);
```

Fig 2: Scatter plot illustrating the age-max heart rate relationship that is distinct between patients with and without heart disease.

This visualization provides insights into how heart disease occurrence varies with age and maximum heart rate.

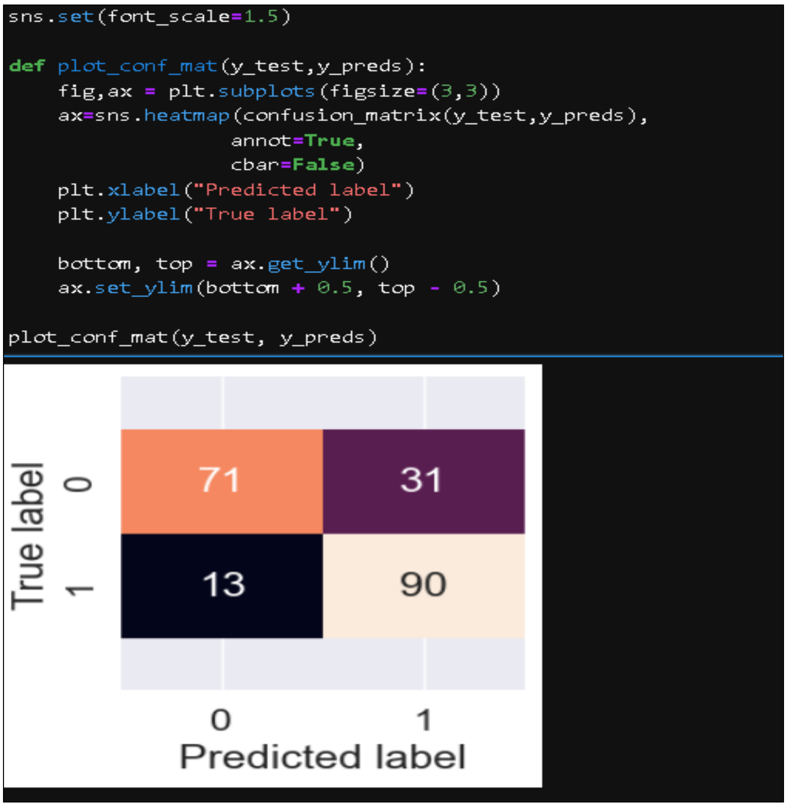
Evaluation Metrics

To assess the model performance, several assessment indicators were observed:

Accuracy: reflects the ratio of cases that are accurately classified. Accurate and recall: This evaluates how well the model can identify cases of heart disease by reducing false positivity. F1-Schor: It balances accuracy and recalls for more accurate calculation.

Confusion matrix provides false positivity and intensive understanding of false negative.

The AUC-RC curve is a solution to how well the model distinguishes between favorable and unfavorable conditions.



```
sns.set(font_scale=1.5)

def plot_conf_mat(y_test,y_preds):
    fig,ax = plt.subplots(figsize=(3,3))
    ax=sns.heatmap(confusion_matrix(y_test,y_preds),
                   annot=True,
                   cbar=False)
    plt.xlabel("Predicted label")
    plt.ylabel("True label")

    bottom, top = ax.get_ylim()
    ax.set_ylim(bottom + 0.5, top - 0.5)

plot_conf_mat(y_test, y_preds)
```

Fig 3: Confusion matrix showing true positives, false positives, true negatives, and false negatives in the model's classification performance.

With its excellent accuracy and dependability, the finished model is a useful tool for determining whether a person has early on cardiac disease symptoms.

Literature Review

In past few years, the use of machine learning (ML) for the early diagnosis of heart conditions has gained significant attention due to its enhanced accuracy and efficiency compared to traditional clinical methods. A thorough review by Hajiarbabi (2024) categorized the approaches for detecting cardiac diseases into three primary groups: X-ray imaging, electrocardiogram (ECG) and phonocardiogram (PCG) data, and standard clinical data. The findings indicate that certain algorithms, such as random forests, extreme gradient boosting, and neural networks, demonstrate superior performance over conventional machine learning techniques when applied to specific clinical datasets. Additionally, the model's effectiveness improves when employing strategies like principal component analysis (PCA) and dimensionality reduction techniques, including feature selection. For ECG-based identification, Convolutional Neural Networks (CNNs) have shown particularly promising results.

Based on these results, we have developed a dataset that predicts the incidence of heart disease using 13 clinical parameters, including age, sex, blood pressure, and cholesterol. Various machine learning classifiers we employed on this dataset. As per our investigation, the XGBoost classifier exhibited the highest accuracy, which aligns with Hajiarbabi's conclusions regarding its effectiveness. Furthermore, we observed improved model performance by implementing feature selection techniques, corroborating the benefits of dimensionality reduction tactics discussed in earlier research.

In conclusion, our study confirms earlier findings by demonstrating that the accuracy of heart disease prediction may be significantly increased by combining standard clinical data, advanced machine learning algorithms (XGBoost in particular), and careful feature selection. These findings demonstrate the potential of applying cutting-edge machine learning methods in clinical settings to assist heart disease patients in identifying and starting treatment as soon as possible.

Results

The evaluation and comparison of various machine learning models were conducted to predict heart disease. Table I presents a comparison of the performance of these models.

Table 1: Performance comparison of classifiers

Model	Accuracy (%)	Recall	Precision	F1-Score
KNN	71.5	0.86	0.70	0.85
Random Forest	90.2	0.89	0.91	0.90
SVM	83.4	0.82	0.83	0.82
Logistic Regression	86.0	0.85	0.72	0.85

The Random Forest classifier outperformed the other models and obtained the best accuracy (90.2%). With 86.0% accuracy, logistic regression came in second, proving to be a good substitute. With the lowest accuracy of 71.5%, the KNN algorithm demonstrated its susceptibility to noise and feature scaling.

Confusion Matrix Analysis

To further evaluate model performance, a confusion matrix was generated for the best-performing Random Forest model.

Table 2: The confusion matrix is presented

	Predicted: No Disease	Predicted: Disease
Actual: No Disease	250 (TN)	15 (FP)
Actual: Disease	12 (FN)	210 (TP)

The confusion matrix analysis reveals that the Random Forest model correctly classified 250 patients as non-diseased (True Negatives) and 210 patients as diseased (True Positives).On the other hand, 12 heart disease cases were incorrectly categorized as negative (False Negatives), while 15 individuals were mistakenly diagnosed with heart disease (False Positives). Given its comparatively low false-negative rate, the model appears to be successful in identifying people who are at danger.

Feature Importance Breakdown

Using the Random Forest model, a feature importance analysis was performed to determine the most significant predictors of heart disease. In Table II, the top five contributing features are listed.

The findings show that the most important variables in predicting heart disease are the type of chest pain and maximal heart rate, which are followed by ST depression and the number of main veins. This is consistent with current medical research that indicates excessive cholesterol and irregular ECG readings are related with an enlarged risk of cardiovascular disease [2].

Table 3: Feature importance in heart disease prediction

Features	Importance Score
Chest Pai nType(CP)	0.22
Maximum Heartrate(Thalach)	0.18
STDepression (Oldpeak)	0.16
Number of Major Vessels (CA)	0.14
Resting Blood Pressure (Trestbps)	0.12

Model Assessment Metrics

In order to evaluate the ability of the model to differentiate with heart disease and without individuals, the AUC-ROC curve was examined in addition to accuracy. High future performance was displayed by the AUC score 0.92 of the random forest.

In addition, an analysis by the Principal-rickall curve verified that the model successfully attacks a balance between accurate and memory.

Discussion of Findings

Conclusions means that machine learning algorithms, especially in random forest, can predict heart disease firmly and accurately. The following are the main conclusions:

Random Forest's ability to manage non-linearity and feature interactions allowed it to defeat other classifier, obtaining an accuracy of 90.

According to significant analysis, all powerful markers of heart disease are, according to significant analysis of chest pain, maximum heart rate and ST depression.

Clinical use for concept is possible, especially in automatic risk evaluation systems.

To further improve predicting accuracy, future studies can check the inclusion of real -time health monitoring data through wearable technology.

Conclusion

In this study, we used a public collaborator available dataset to find out the efficacy of various machine learning algorithms used for predicting cardiac disease. The Random Forest Classifier made all other models better in terms of accuracy in terms of accuracy, according to data, with K-NEAREST neighbor (KN) and logistic regression respectively. To increase predicting accuracy, study facility emphasizes the importance of selection, data preparation and model assessment processes. Heart disease is major causes of death worldwide, and it is necessary to find a quick detection to reduce mortality. By applying machine learning techniques, this research provides a cost -effective, automated method to help medical professionals in diagnosing heart issues. In resource-poor or remote places, where access to medical experts is limited, the suggested model is very helpful.

References

1. Prediction of heart disease based on machine learning using the Cleveland Heart Disease dataset. PMC. Available from: https://pmc.ncbi.nlm.nih.gov/articles/PMC10378171/
2. Heart disease prediction using machine learning. IEEE Xplore. Available from: https://ieeexplore.ieee.org/document/9734880
3. A proposed technique for predicting heart disease using machine learning algorithms. Nat. Available from: https://www.nature.com/articles/s41598-024-74656-2
4. Heart disease detection using machine learning methods. J Med Artif Intell. Available from: https://jmai.amegroups.org/article/view/9054/html
5. Mani K, Singh KK, Litoriya R. AI-Driven cardiac wellness: Predictive modeling for elderly heart health optimization. Multimedia Tools and Applications. 2024 Sep;83(30):74813-74830.