Heart Disease Prediction System

Mr. S. S. MHASKE[1], MANGESH I. BONDE[2], RUSHIKESH V. TAKWALE[3], VIRENDRA D. GAWANDE[4], OM V. WARADE[5], KOMAL K. NAHATE[6]
ENTC Dept., MGICOET, Shegaon, Maharashtra, India

Abstract- ***Cardiovascular diseases are a leading cause of death worldwide, with 17.9 million deaths annually. This research explores machine learning algorithms for predicting cardiac abnormalities using ECG analysis and symptom-based data. Four classification algorithms (Random Forest, Decision Tree, Logistic Regression, and K-Nearest Neighbors) were compared using a dataset of 1025 patient records from Kaggle and PhysioNet. The study highlights AI's potential in enhancing diagnosis, particularly in emergency situations. Random Forest achieved high classification accuracy. AI-driven diagnostic tools can improve early detection, especially in resource-limited settings. Future research will focus on refining models and ensuring clinical utility.***

The successful integration of machine learning models into clinical practice can revolutionize cardiovascular care, making early detection and intervention more accessible, especially in underserved areas. Future research will focus on refining the models, improving their interpretability and ensuring that they are clinically useful, ensuring that AI models become an indispensable tool in the diagnosis and management of cardiovascular diseases.

Index Terms- ***Machine Learning, AI, Random Forest, KNN, Decision Tree, Logistic regression, VS Code, Confusion matrix, Heatmap, AUC, Accuracy .***

I. INTRODUCTION

Heart disease remains one of the leading causes of mortality worldwide, posing a significant burden on public health systems and economies. According to the World Health Organization (WHO), an estimated
17.9 million people die each year due to cardiovascular diseases, representing approximately 31% of all global deaths. Among these, coronary artery disease (CAD) is the most common type, accounting for a substantial proportion of cardiovascular-related fatalities.

In high-pressure environments such as emergency rooms or during mass screenings, the probability of diagnostic errors increases, leading to missed or delayed diagnoses. In light of these challenges, there is a growing need for more reliable, efficient and accurate diagnostic tools that can complement human expertise and reduce the likelihood of errors in cardiac diagnostics. In recent years, artificial intelligence (AI) has emerged as a transformative force in healthcare, offering new possibilities for enhancing the diagnostic process. Among the various AI techniques, machine learning has demonstrated particular promise in the realm of medical diagnostics. management of patients' health.

Machine learning (ML) techniques have emerged as promising tools for medical diagnosis and risk assessment. By leveraging algorithms and computational methods, ML models can analyze large datasets, identify patterns, and make predictions based on input features. In the context of heart disease detection, ML algorithms can learn from historical patient data to classify individuals into different risk categories, enabling healthcare providers to prioritize interventions and allocate resources effectively.

The most important organ in the human body is heart. It pumps blood to every part of the body, ensuring that all the organs receive the oxygen and nutrients they need to function properly. If the heart stops working, the brain and other vital organs will stop functioning and a person could die within minutes.

In this project, we propose a Heart Disease Detection and Classification system using machine learning algorithms. Our objective is to develop a robust and accurate predictive model that can assist healthcare professionals in identifying individuals at risk of heart disease. By employing various ML techniques and implementing a user-friendly interface, our system aims to streamline the diagnostic process, enhance patient outcomes, and contribute to the overall improvement of cardiovascular health management strategies.

II. MOTIVATION

Cardiovascular diseases (CVDs) represent a growing global health concern, placing immense pressure on healthcare systems, particularly in developing nations where medical infrastructure and skilled healthcare professionals are in short supply. In such regions, heart-related conditions often go unnoticed until severe symptoms arise, leading to higher rates of morbidity and mortality. This delay in detection significantly contributes to a larger public health issue, as CVDs are among the primary causes of preventable deaths worldwide. Early detection and timely intervention are crucial in reducing fatalities and improving patient outcomes. However, in remote or economically challenged areas, the lack of access to specialized care often prevents individuals from receiving essential screenings for early diagnosis.

This challenge emphasizes the need for innovative, affordable, and scalable solutions. In this regard, machine learning (ML) offers significant promise. ML-based models could be integrated into clinical environments to serve as cost-effective tools for early heart disease detection, especially where traditional healthcare resources are lacking.

The purpose of this study is to develop a predictive model for heart disease using various machine learning algorithms. Moreover, the research aims to identify the most effective classification algorithm for determining the likelihood of a patient having heart disease. The study justifies its approach through a comparative analysis of widely used ML classification methods. Despite their popularity, achieving high accuracy in heart disease prediction remains a critical goal. Therefore, the selected algorithms are tested under various evaluation criteria to determine the most reliable method.

This work intends to assist researchers and healthcare professionals by offering insight into the most efficient algorithm for predicting heart conditions. Ensuring accurate diagnosis and appropriate treatment is essential for quality healthcare delivery. Poor diagnostic decisions can result in negative outcomes, which are not acceptable in any clinical setting. One of the major challenges faced by healthcare providers is offering high-quality care at a sustainable cost. Reducing the expense of diagnostic tests can be achieved through the smart use of computer-based technologies.

III. METHODOLOGY

1. System Flow for Prediction of Heart Disease

This chapter outlines the approach used for building and assessing machine learning models aimed at predicting the presence of heart disease. It thoroughly explains each stage of the process, beginning with data preprocessing and progressing through model training, performance evaluation, and optimization. The methodology is carefully structured to maximize the model's precision, stability, and clarity, thereby enabling trustworthy and insightful predictions in heart disease diagnosis.

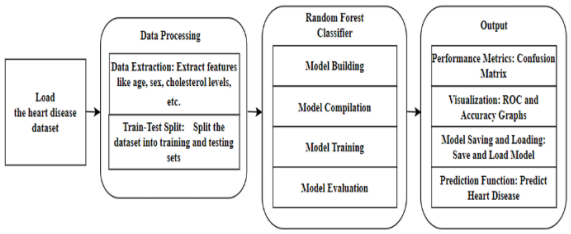

Figure 1: System Flow for Prediction of Heart Disease

The system architecture, as illustrated in the corresponding block diagram, centers around the use of a Random Forest Classifier. The prediction system is divided into four essential phases: dataset loading, data preprocessing, model training and assessment, and finally, result generation. Each component is crucial to ensuring the effective functioning and reliability of the heart disease prediction system.

2. Data Collection

A detailed dataset comprising numerous heart health-related attributes has been gathered from a variety of sources such as hospitals, healthcare organizations, publicly available databases, and previous research work. Emphasis is placed on maintaining high standards of data quality, uniformity, and applicability to the intended population to ensure the dataset is both accurate and relevant for analysis.

3. Data Processing

Once the dataset is loaded, the process moves into the crucial phase of data processing, which focuses on refining and optimizing the dataset for analysis. This involves key steps such as feature selection, where important variables are identified and retained, while irrelevant or redundant features are eliminated to reduce noise and improve model performance. A major part of this phase also

includes addressing missing values to enhance the dataset's reliability. Depending on the situation, missing entries are either filled using statistical measures like the mean or median or removed entirely to prevent skewing the results. Additionally, data cleaning is conducted to correct inconsistencies, outliers, and any abnormal or incorrect entries, thereby preserving data integrity and ensuring the model can make accurate predictions.

Following data preparation, the dataset is split into two parts — a method known as the train-test split, as illustrated in Figure 2. This step is essential for properly evaluating the machine learning algorithm. Typically, 70% of the data is designated for training purposes, while the remaining 30% is set aside for testing. The training subset, which comprises the bulk of the data, is used to train the Random Forest Classifier.

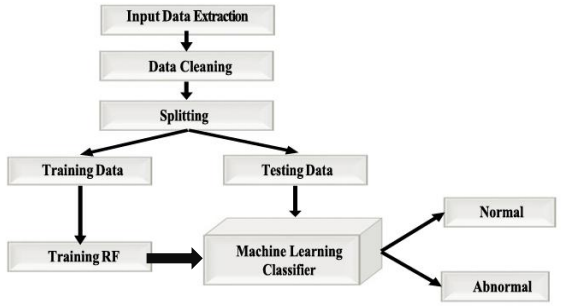

Figure 2: Block Diagram Representing Preprocessing and Classification

As part of the data processing workflow, clinical inputs are gathered through a user-friendly interface, which may be a component of a heart disease prediction application. This interface collects a variety of health-related parameters, with each input acting as a feature for the predictive model. Users are prompted to enter information such as age, gender, blood pressure, cholesterol levels, and electrocardiogram (ECG) results. By using intuitive tools like sliders and dropdown menus, as shown in Figure 5, the system simplifies the data entry process while capturing all essential details. These inputs collectively serve as the model's features, enabling it to analyze individual health metrics and estimate the user's risk of developing heart disease.

4. Feature Selection

In order to improve the accuracy and effectiveness of heart disease detection and classification, various feature selection strategies are employed. Techniques such as correlation analysis, recursive feature elimination, and feature importance ranking are utilized to pinpoint the most significant attributes. By narrowing down the dataset to only the most relevant features, these methods help minimize dimensionality and boost the performance of the model by concentrating on the most impactful data inputs.

5. Model Selection and Training

To detect and classify heart disease, multiple machine learning algorithms are examined, such as Logistic Regression, Support Vector Machines (SVM), K-Nearest Neighbors (KNN), Decision Trees, Random Forest, and Gradient Boosting. Each of these models is trained using the preprocessed data. To enhance their predictive performance, hyperparameters are fine-tuned using optimization strategies like grid search or random search. The dataset is divided into training and testing sets to enable proper performance evaluation. Additionally, cross-validation techniques are employed to assess the stability and generalizability of the models across different data samples.

6. Model Evaluation

The performance of each machine learning model is evaluated using appropriate evaluation metrics, including accuracy, precision, recall, F1 score, and area under the ROC curve (AUC-ROC). Model evaluation results are analyzed to identify the most effective approach for heart disease detection and classification.

7. Model Deployment and Output Generation.

The trained model is integrated into a user-friendly Graphical User Interface (GUI) for healthcare professionals. The GUI collects patient data input and uses the model to generate predictions on heart disease likelihood, offering valuable decision support to healthcare providers through seamless interaction.

8. Result Analysis

The deployed model's output is thoroughly analyzed to assess its accuracy, reliability, and usability in real-world healthcare settings, ensuring it meets clinical standards. A comparative analysis is also conducted to evaluate the model's performance against existing diagnostic methods and alternative

machine learning approaches, highlighting its strengths and areas for improvement.

9. Reporting and Interpretation

The results obtained from the model evaluation and analysis are documented in detail, including performance metrics, findings, and conclusions.

A final research paper or project report is prepared, summarizing the methodology, results, and implications of the study. Recommendations for future research and potential applications of the developed system in clinical practice are also included in the report.

IV. RESULT AND ANALYSIS

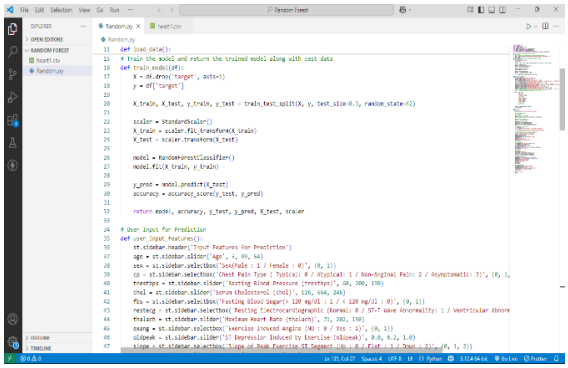

Figure 3: VS_Code Interface Showing Python program of Random Forest Classifier

Figure 3 shows the Interface page of VS code. The program is started to execute by entering the command: streamlit run main.py. Here the streamlit run command is primarily used to run and display a streamlit application from a Python script.

It is useful for turning static Python code into dynamic, interactive web apps with minimal code and effort, allowing easy interaction with data, models and visualizations. It automatically opens a local server and a browser interface for user to interact with the Python script's output, providing a streamlined way to develop data-driven web applications. It is a fundamental command for developing, testing and sharing interactive web applications created using streamlit in Python.

After executing the program, the results for heart disease prediction are displayed, figure 5.16 shows the Random forest model predictions based on the input data features. This figure represents a Heart Disease Prediction dashboard generated using Streamlit, a Python-based web application framework. This dashboard is a tool to predict the possibility f heart disease based on patient input parameters using a machine learning model. The Input Features for Prediction shown on Left Panel are described in table 1.

1. Input Features for Prediction (Left Panel):

The dataset includes various pieces of information about the patients, all of which are essential for making accurate predictions. Each sample in the dataset consists of clinical measurements and demographic information, serving as features or independent variables.

Table 1: Input Features with their Description

Feature	Description
Age	Age of patient in years
Sex	Gender (1 = Male, 0 = Female)
Chest Pain Type	Chest pain types: 0 = typical angina, 1 = atypical angina, 2 = non-anginal pain, 3 = asymptomatic
Resting Blood Pressure	Blood pressure (mm Hg) during rest(94-200)
Serum Cholesterol	Cholesterol level in mg/dl(126-564)
Fasting Blood Sugar	Fasting blood sugar > 120 mg/dl =1, otherwise 0
Resting ECG	Resting electrocardiographic results(Normal-0, ST-T Abnormality-1, Ventricular abnormality-2)
Max Heart Rate (Thalach)	Max heart rate achieved (71-202)
Exercise-Induced Angina	Exercise-induced angina (1 = Yes, 0 = No)
Oldpeak	ST depression induced by exercise relative to rest (0-6.20)
Slope	Slope of the peak exercise ST segment (Up-0, Flat-1, Down-2)

2. Heart Disease Prediction (Main Output Section):

Prediction Output: The result indicates prediction of heart disease like Heart Disease Predicted or Heart disease not predicted as shown in figure 4.

3. Prediction Probability Visualization

A bar chart visualizing the prediction probabilities where red bar represents the probability of having heart disease. Green bar indicates probability of not having heart disease. Model Accuracy of the machine learning model (e.g., 0.98 or 98%) is also displayed after prediction. Figures 5.6 and 5.7 presents an analysis of the Random Forest Model used for heart disease prediction. It includes key evaluation metrics and visualizations.

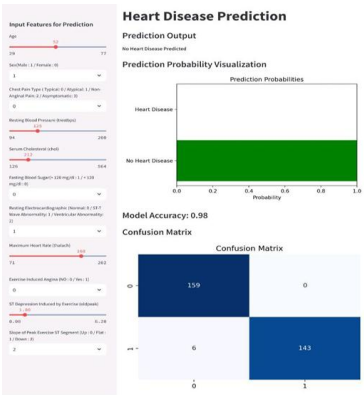

Figure 4: Heart Disease Prediction Dashboards for Random Forest Model

Confusion Matrix:

A Confusion Matrix is a 2x2 matrix used to evaluate the performance of a classification model. It helps to visualize the performance of an algorithm in a straightforward way by showing how many predictions were correct and how many were incorrect. It provides insight into not only the overall accuracy of a model but also into the types of errors made, which is crucial in understanding its effectiveness especially in imbalanced datasets. The confusion matrix for Random Forest Classifier is as given in figure 5.

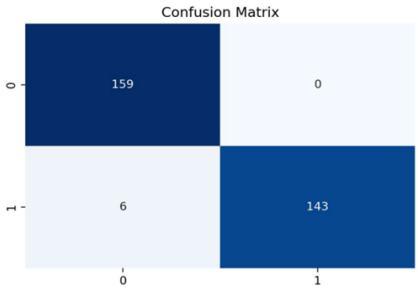

Figure 5: Confusion matrix

True Negatives (159) represented are correctly predicted instances where the patient does not have heart disease. False Positives (0) shows instances incorrectly predicted as having heart disease when they do not. False Negatives (6) represents instances incorrectly predicted as not having heart disease when they actually do. True Positives (143) indicates correctly predicted instances where the patient has heart disease. This shows that the model has a strong classification performance with very few misclassifications.

ROC Curve:

ROC (Receiver Operating Characteristic) Curve: Evaluates the trade-off between the true positive rate(sensitivity) and the false positive rate at various thresholds. AUC (Area under the Curve) = 1.00 indicates that the model has perfect discriminatory ability between classes, a highly desirable result in classification tasks.

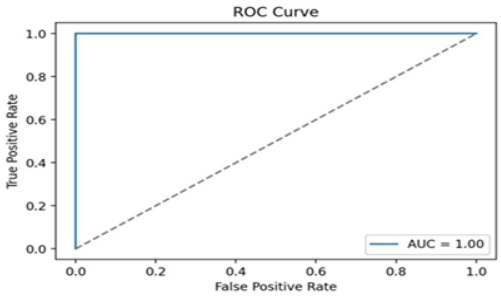

Figure 6: ROC Curve

Feature Correlation Heatmap:

A Feature Correlation Heatmap is a graphical representation used to display the correlation between multiple variables or features in a dataset. It's particularly useful in exploratory data analysis (EDA) for identifying relationships between different features, especially before building a machine learning model.

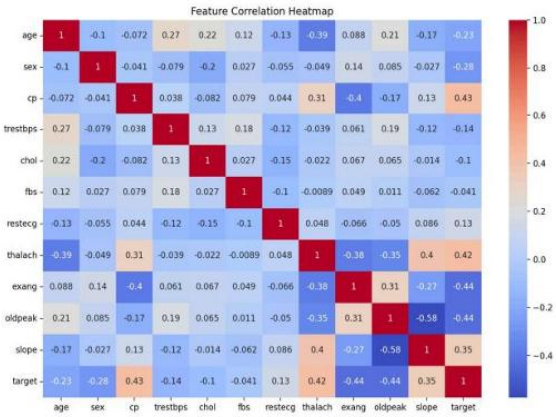

Figure 7: Feature Correlation Heatmap

The heatmap uses color coding to highlight the strength and direction of these correlations, making it easier to spot patterns and trends in the data. The figure 7 shows a heatmap generated for a Randam Forest Model. It highlights the effectiveness of the Randam Forest Model in predicting heart disease with high accuracy and near-perfect ROC performance.

Table2: Comparison of Target and Achieved Outputs

Age	Sex	cp	bps	Chol	fbs	Rest-ecg	thalach	Exang	Old-peak	slope	Target
52	1	0	125	212	0	1	168	0	1	2	0
53	1	0	140	203	1	0	155	1	3.1	0	0
70	1	0	145	174	0	1	125	1	2.6	0	0
61	1	0	148	203	0	1	161	0	0	2	0
62	0	0	138	294	1	1	106	0	1.9	1	0
58	0	0	100	248	0	0	122	0	1	1	1
58	1	0	114	318	0	2	140	0	4.4	0	0
55	1	0	160	289	0	0	145	1	0.8	1	0
46	1	0	120	249	0	0	144	0	0.8	2	0
54	1	0	122	286	0	0	116	1	3.2	1	0
71	0	0	112	149	0	1	125	0	1.6	1	1
43	0	0	132	341	1	0	136	1	3	1	0
34	0	1	118	210	0	1	192	0	0.7	2	1
51	1	0	140	298	0	1	122	1	4.2	1	0
34	0	1	118	210	0	1	192	0	0.7	2	1

Furthermore, we conducted a geographical analysis to assess the availability of medical facilities in various areas, focusing on rural or underserved regions. By integrating this geographical information into our predictive model, we aimed to

customize our approach to address the specific challenges faced by communities with limited access to healthcare.

One of the notable outcomes of our study was the development of a user-friendly Graphical User Interface (GUI) to facilitate interaction with the predictive model. The GUI allows healthcare professionals to input patient data easily and obtain instant predictions regarding the likelihood of heart disease. This intuitive interface enhances accessibility and usability, particularly in settings where technical expertise may be limited.

Overall, our results demonstrate the feasibility and effectiveness of utilizing machine learning techniques to predict CVD risk while considering geographical and socio-economic factors. By tailoring our approach to the unique needs of underserved communities and providing a user friendly interface, we aim to improve early detection and intervention for heart disease, ultimately contributing to better healthcare outcomes for all.

CONCLUSION

This paper explored the application of four machine learning algorithm, K-Nearest Neighbors (KNN), Logistic Regression, Decision Tree and Random Forest in classifying cardiac abnormalities, particularly the presence or absence of heart disease. By utilizing well established datasets, the study aimed to identify which model could provide the most accurate and reliable predictions for diagnosing heart disease based on medical features such as age, cholesterol levels, blood pressure and more. Cardiovascular diseases (CVDs) continue to hold the position as the leading cause of death worldwide, contributing to a significant public health problem.

Random Forest combines multiple decision trees for greater accuracy and can handle both categorical and continuous data. K-Nearest Neighbors (KNN) classifies based on proximity to labeled data points but struggles with larger datasets. Naive Bayes, a probabilistic classifier, is efficient but less accurate than more advanced models. SVMs handle high-dimensional data well but can be computationally expensive. Neural networks, especially deep learning models, are effective in modeling complex relationships but lacks in interpretability.

The dataset also includes information about chest pain labeled as cp. Chest pain type can offer critical information about the patient's condition with different levels signifying various forms of discomfort or pain ranging from typical angina to asymptomatic situations. The dataset includes chol, which records the serum cholesterol levels of the patient. High cholesterol levels are a known risk factor for heart disease, and their inclusion in the dataset aids the model in identifying at-risk individuals.

The comparative analysis of the models highlighted that the Random Forest consistently outperformed the others in terms of confusion matrix which indicates that True prediction are highest while it gives negligible false predictions, whereas it is found that K Nearest Model gives highest number of false predictions which is not desired for cardiac abnormalities predictions. In terms of practical application, the results of this study suggest that machine learning models particularly Random Forest can play a significant role in the timely detection and classification of cardiac abnormalities. As in this research, the patient is classified with Heart disease predicted or Heart disease not predicted, however further work can be carried out with the large dataset using additional techniques but to get large datasets is difficult in medical field.

Using algorithms like Random Forest and optimization methods such as the Random Search Algorithm, prediction accuracy and efficiency improve notably. Integration with the Internet of Medical Things enables real-time monitoring and remote diagnosis, expanding their impact. In summary, these efforts represent a shift towards personalized, data-driven healthcare. They offer promise in improving patient outcomes, reducing disparities, and saving lives. Continued research, collaboration, and implementation are crucial to fully harnessing these transformative tools for global health challenge.

REFRENCES

[1] Heart Disease prediction model using Random Forest Algorithm S. S. Mhaske and Dr. C. M. Jadhao. Journal Name-International Journal of Applied Engineering and Technology, Volume 5, December 2023, ISSN 2633-4828

[2] GUI Based Heart Disease Prediction Model using Random Forest Algorithm S. S. Mhaske and Dr. C. M. Jadhao Journal Name-Journal of Electrical Systems Vol 20, 11s (2024),

ISSN 1112-5209

[3] Heart Disease Detection and Classification using Machine Learning S. S. Mhaske, Vishwjeet Tayade The board of International Journal of Innovative Reasearch in Technology, Volume 10(2024), Issue 11, ISSN 2349-6002

[4] Ekta Maini, Bondu Venkateswarlu, Baljeet Maini, Dheeraj Marwaha (2020). Machine learning based heart disease prediction system for Indian population: an exploratory study done in south India. Medical Journal Armed Forces India. https://doi.org/10.1016/j.mjafi.2020.10.013

[5] Ashir Javeed , Shijie Zhou , Liao Yongjian , Redhwan Nour , Samad Wali & Abdul Basit (2019). An Intelligent Learning System based on Random Search Algorithm & Optimized Random Forest Model for Improved Heart Disease Detection . IEEE Access http://dx.doi.org/10.1109/ACCESS.2019.2952107

[6] Chunyan Guo , Jiabing Zhang , Yang Liu , Yaying Xie , Zhiqiang Han , Jianshe Yu (2017). Recursion Enhanced Random Forest With An Improved Linear Model (RERFILM) for HD Detection on the Internet of Medical Things Platform. IEEE Access http://dx.doi.org/10.1109/ACCESS.2020.2981159

[7] Sree, P. K., Prasad, M., PBV, R. R., Ramana, C. V., Murty, P. T. S., Mallesh, A. S., & Raju, P. J. R. S. (2023, October 10). A Comprehensive Analysis on Risk Prediction of Heart Disease using Machine Learning Models. International Journal on Recent and Innovation Trends in Computing and Communication, 11(11s), 605–610. https://doi.org/10.17762/ijritcc.v11i11s.8295

[8] Jummelal, K. (2023, May 29). Chronic Heart Failure Diagnosis from Heart Sounds Using Machine Learning and Full-Stack Deep Learning . https://www.jclmm.com/index.php/journal/article/view/1065

[9] Alkurdi, A. A. H. (2023, January 1). Enhancing Heart Disease Diagnosis Using Machine Learning Classifiers. https://doi.org/10.54216/fpa.130101

[10] Zabeeulla, M., Sharma, C., & Anand, A. (2023, March 1). Early Detection of Heart Disease Using Machine Learning Approach. CARDIOMETRY, 26, 342–347.https://doi.org/10.18137/cardiometry.2023.26.342347

[11] Nayeem, M. J. N., Rana, S., & Islam, M. R. (2022, November 30). Prediction of Heart Disease Using Machine Learning Algorithms. European Journal of Artificial Intelligence and Machine Learning, 1(3), 22 26. https://doi.org/10.24018/ejai.2022.1.3.13

[12] Rindhe, B. U., Ahire, N., Patil, R., Gagare, S., & Darade, M. (2021, May 12). Heart Disease Prediction Using Machine Learning. International Journal of Advanced Research in Science, Communication and Technology, 267–276. 1131 https://doi.org/10.48175/ijarsc

[13] Anusuya, V., & Gomathi, V. (2021, March 25). An Efficient Technique for Disease Prediction by Using Enhanced Machine Learning Algorithms for Categorical Medical Dataset. Information Technology and Control, 50(1), 102–122. https://doi.org/10.5755/j01.itc.50.1.25349