Machine Learning-Based Approach for Predicting Heart Disease

[1]Charan Jammula **[2]Divya Bagi** **[3]Nagaraju Yadav Konda**

Final year Students, Dept of Electronics and Communication Engineering, Geethanjali College of Engineering and Technology, Hyderabad, India [1]charanc985@gmail.com **[2]**divyabagi2004@gmail.com [3]21r11a0422@gcet.edu.in

[4]M. Anand, Assistant Professor, Dept of Electronics and Communication Engineering, Geethanjali College of Engineering and Technology, Hyderabad, India

How to Cite this Article:

Jammula, C., Bagi, D. & Konda, N. Y. (2026). Machine Learning-Based Approach for Predicting Heart Disease. International Journal of Creative and Open Research in Engineering and Management, 10(01). https://doi.org/10.55041/ijcope.v2i2.158

https://doi.org/ 10.55041/ijcope.v2i2.158

ABSTRACT

Heart disease is a significant public health issue, accounting for a large percentage of global deaths and rising healthcare expenditures. This research seeks to enhance early heart disease detection using sophisticated machine learning techniques with the Random Forest classifier. A dataset of 4,240 individuals was utilized, containing a mix of clinical data and personal background details—age, sex, cholesterol level, blood pressure, smoking status, and body mass index—we adopted a systematic methodology involving data preprocessing, missing value treatment, feature scaling, and hyperparameter optimization. With cross-validation and careful testing, the model was optimized to deliver the optimal performance.

Random Forest classifier achieved 85% accuracy, which was higher than that achieved with conventional models like logistic regression and other linear classifiers. Its ability to reduce false negatives substantially renders it is an effective technique for the early identification of cardiovascular risk, thereby enabling early medical intervention. The model's predictive capability underlies clinical decision-making and can be integrated into electronic health records and wearable devices to track health real-time. This study highlights the greater use of AI in medicine, providing a sturdy approach to improving diagnostic accuracy, patient outcomes, and enabling the use of intelligent systems in modern healthcare practice.

Key Words: Cardiovascular Disease Prediction, Machine Learning Techniques, Random Forest Algorithm.

1. INTRODUCTION

Globally, cardiovascular disease is the leading cause of death. The best way to reduce mortality with preventive therapy is to identify high-risk heart disease cases early. Traditional risk assessment techniques frequently result in inaccurate or less-than-ideal results. Predictive modeling for the medical field has intriguing new choices thanks to machine learning techniques, especially ensemble approaches like Random Forest. One of the major causes of death worldwide is heart disease, and preventing serious health problems or death requires early detection.

Traditional diagnosis often requires extensive medical evaluation, which can be time- consuming and inaccessible for some individuals. To overcome this challenge, advancements in technology have made it possible to use **machine learning** to assist in forecasting the chance of heart disease based on a person's health data.

This project presents a **Heart Disease Prediction Web Application** that uses a **Random Forest Classifier**, a potent machine learning system to determine a person's risk of heart disease development. The application uses an easy-to-use web interface to gather a variety of lifestyle and medical data, including blood pressure, cholesterol, smoking habits, age, gender, and more. These inputs are fed into a machine learning model trained on historical patient data, and the model provides a prediction output indicating whether the individual is likely to develop heart disease.

Because the Random Forest algorithm is accurate, resilient, and can handle both continuous variables and categorical data, it was chosen. It works by building multiple decision trees and combining their outputs to make a final decision, which helps improve prediction accuracy and reduce overfitting.

This system aims to assist individuals and healthcare professionals in identifying high- risk patients at an early stage, potentially leading to timely medical intervention and lifestyle changes. The model can be easily accessed through a web interface, making it practical and efficient for everyday use.

2. LITERATURE REVIEW

Recently, there has been a significant increase in interest in machine learning (ML) in the healthcare domain, particularly for early diagnosis and risk prediction of life-threatening diseases such as cardiovascular diseases (CVDs). The World Health Organization (WHO) estimates that cardiac illnesses cause around 17.9 million deaths annually, making them the world's top cause of mortality. Consequently, scholars have investigated the application of intelligent systems and ML techniques to improve the early detection and management of heart disease.

Logistic Regression and **Naïve Bayes** have been traditionally used for binary classification problems, including medical diagnosis. Studies show that logistic regression performs well when relationships between variables are linear and when the data is balanced. However, in complex and imbalanced medical datasets, it may not always provide high accuracy.

Support Vector Machines (SVMs) have been applied effectively to heart disease datasets due to their strong mathematical foundation and ability to handle high-dimensional data. Research by Polat and Güneş (2007) demonstrated that SVMs, combined with feature selection techniques, achieved significant accuracy in predicting heart conditions.

Decision Trees are also commonly used due to their interpretability, which is particularly important in healthcare. However, single decision trees are prone to overfitting, especially when the training data contains noise or missing values.

To overcome these limitations, **ensemble learning methods** such as **Random Forest** and **Gradient Boosting** have shown promising results. Random Forest, in particular, has been found to be highly effective in medical diagnosis tasks due to its robustness, ability to handle missing data, and reduced risk of overfitting. Research conducted by Sudha and Muthulakshmi (2020) showed that Random Forest outperformed other models in terms of accuracy and reliability for heart disease prediction.

Apart from these methods, deep learning models like convolutional neural networks (CNNs) and artificial neural networks (ANNs) have been investigated; nevertheless, they frequently need for larger datasets and more processing capacity.

One of the most commonly used datasets is the Framingham Heart Study, a long-term cardiovascular study that began in 1948 and provides detailed health information, which has been instrumental in developing predictive models. Polat and Güneş (2007) explored hybrid models combining Decision Trees and Support Vector Machines (SVM), demonstrating enhanced accuracy in heart disease detection. In subsequent research, Dinesh Kumar and Paulraj (2015) compared classifiers such as Naive Bayes, Decision Trees, and Neural Networks, finding that while Naive Bayes was computationally efficient, Neural Networks and Decision Trees offered better accuracy. Kavakiotis et al. (2017), although focused on diabetes, emphasized the success of machine learning methods such as Random Forests and Logistic Regression in medical data prediction tasks. The UCI Cleveland dataset also played a significant role in early experiments, where ensemble models like Random Forest consistently achieved 85–90% accuracy. More recent studies, such as those by Sudha and Muthulakshmi (2020), and Karthik et al. (2021), have confirmed that Random Forest algorithms provide superior performance in terms of accuracy, robustness, and overfitting control when compared to other traditional models. These results highlight how machine learning, in particular ensemble approaches, can be used to create trustworthy and precise cardiac disease prediction systems. The present study builds upon this foundation by employing the Random Forest classifier with the Framingham dataset, integrating it into a user-interactive web application to enable practical heart disease risk prediction.

This project builds upon previous studies by utilizing the **Random Forest Classifier** due to its strong performance in classification tasks and its ability to process both categorical and numerical input features. It also incorporates important preprocessing steps such as handling missing values, class imbalance, and data normalization to increase the accuracy of the model.The created system provides a web- based interface for real-time prediction, which enhances usability for both medical professionals and non-specialist users.

3. PROJECT OVERVIEW

3.1 CURRENT SYSTEM

In current systems, patient data are gathered and processed through various models using machine learning to predict heart disease Most of these models use datasets made available at the UCI Machine Learning Repository. Neural Networks (NN), Decision Trees (DT), Support Vector Machines (SVM), and Naive Bayes are employed to analyze these medical datasets.

They predict but are poor in result interpretation as well as data acquisition. Model performance is uneven, as some perform as well as 86%, but transparency and usability remain poor.

Limitations of Existing System:

Outcomes can be complex and challenging for healthcare professionals to interpret.

Inconsistent data can result in incomplete or noisy inputs, impacting prediction quality.

3.2 SYSTEM PROPOSAL

The suggested system adds to current models with a structured, extendable, and precise heart disease prediction mechanism. It gathers the patient information and goes through preprocessing, including feature scaling, missing value handling, and one-hot encoding.The system employs Python libraries such as Pandas, NumPy, and Scikit-learn to perform data operations. It evaluates different machine learning models and concludes that Random Forest is the most suitable. It improves accuracy and prevents overfitting by comparing predictions from a number of decision trees.We tested five models and selected the model that had the best predictions as our final solution.

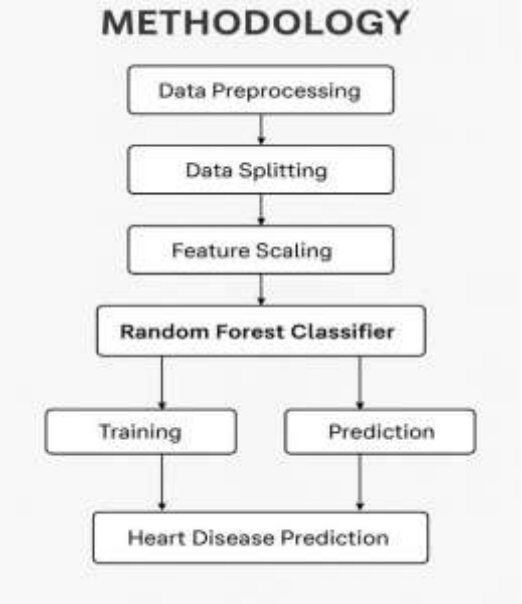
Advantages of Suggested System:

Improved accuracy via model tuning and ensemble learning.

Clinically relevant findings that are user- friendly. Seamlessly integrates with digital health platforms and wearable data.

4. PROJECT DESIGN AND ANALYSIS

4.1 ARCHITECTURE



4.2 DATA PRE-PROCESSING

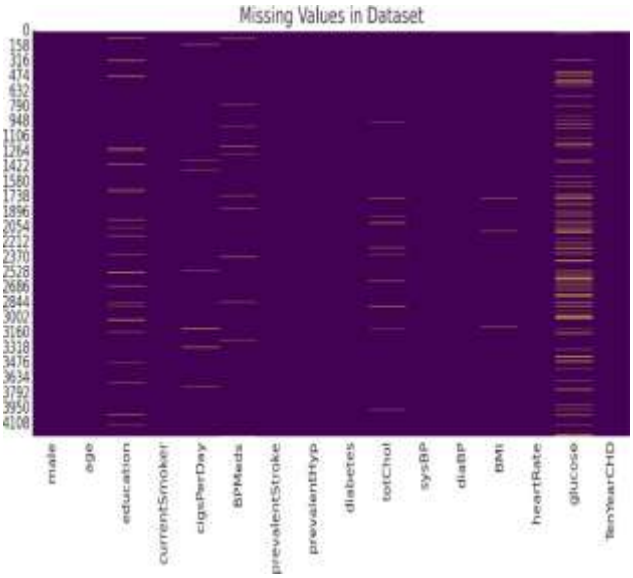
4.2.1 DATA ENCODING

Data encoding is a crucial step in preparing categorical data for machine learning models, which require numerical inputs. In this heart disease prediction project, various input features such as gender, current smoking status, blood pressure medication usage, stroke history, hypertension status, and diabetes status were originally represented as categorical variables (e.g., "male", "female", "yes", "no"). These values were converted into binary numerical format using manual encoding. For instance, "male" was encoded as 1 and "female" as 0, while "yes" and "no" responses were similarly encoded as 1 and 0, respectively. This transformation ensures that the machine learning algorithm can effectively process and analyze the data without misinterpreting textual values.

4.2.2 DATA IMPUTATION

Real-world datasets often contain missing or incomplete values, which can negatively impact the performance of models of machine learning. To deal with this, data imputation was applied to the dataset used in the project. Two strategies were employed based on the type of data. For binary or categorical fields such as gender, diabetes status, or hypertension, the most frequent value (mode) was used to fill in missing entries. For continuous numerical features such as cholesterol, BMI, systolic and diastolic blood pressure, and glucose levels, the median value was used for imputation. This approach helps in maintaining the distribution of the data and prevents bias that could arise from using mean values in skewed datasets.

This heatmap shows where missing values exist in the dataset.



4.2.3 DATA SCALING

Data scaling was performed using the standard scaler, which applies z-score normalization to each numerical feature in the dataset. this ensures that all the features have:

- mean (μ) = 0
- standard deviation (σ) = 1

Mathematical formula used

for a given feature xxx (e.g., age, cholesterol, bmi), the scaled value x' is computed as:

where:

- x is the original value of the feature
- μ is the mean of the feature in the training data
- σ is the standard deviation of the feature in the training data
- x′ is the scaled value

4.3 DATA PRE-PROCESSING

4.3.1 DATA ENCODING

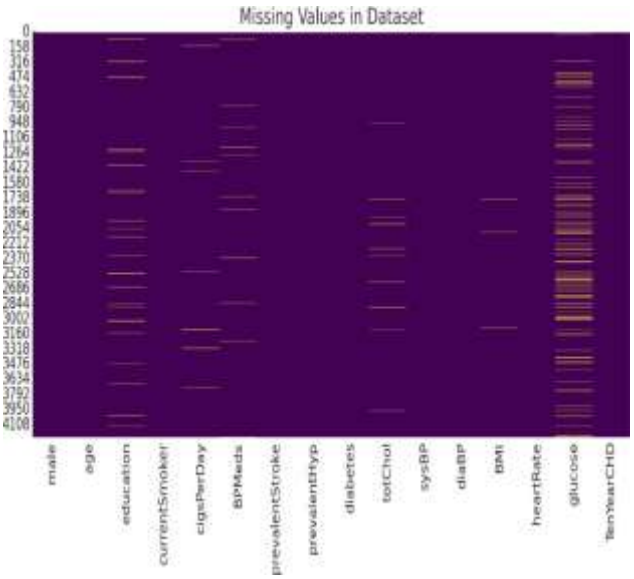
Data encoding is a crucial step in preparing categorical data for machine learning models, which require numerical inputs. In this heart disease prediction project, various input features such as gender, current smoking status, blood pressure medication usage, stroke history, hypertension status, and diabetes status were originally represented as categorical variables (e.g., "male", "female", "yes", "no"). These values were converted into binary numerical format using manual encoding. For instance, "male" was encoded as 1 and "female" as 0, while "yes" and "no" responses were similarly encoded as 1 and 0, respectively. This transformation ensures that the machine learning algorithm can effectively process and analyze the data without misinterpreting textual values.

4.3.2 DATA IMPUTATION

Real-world datasets often contain missing or incomplete values, which can negatively impact the performance of models of machine learning. To deal with this, data imputation was applied to the dataset used in the project. Two strategies were employed based on the type of data. For binary or categorical fields such as gender, diabetes status, or hypertension, the most frequent value (mode) was used to fill in missing entries. For continuous numerical features such as cholesterol, BMI, systolic and diastolic blood pressure, and glucose levels, the median value was used for imputation. This approach helps in maintaining the distribution of the data and prevents bias that could arise from using mean values in skewed datasets.

This heatmap shows where missing values exist in the dataset.

$$x' = \frac{x - \mu}{\sigma}$$



4.3.3. DATA SCALING

Data scaling was performed using the standard scaler, which applies z-score normalization to each numerical feature in the dataset. this ensures that all the features have:

- mean (μ) = 0
- standard deviation (σ) = 1

Mathematical formula used

for a given feature xxx (e.g., age, cholesterol, bmi), the scaled value x′ is computed as:

$$x' = \frac{x - \mu}{\sigma}$$

where:

- x is the original value of the feature
- μ is the mean of the feature in the training data
- σ is the standard deviation of the feature in the training data
- x′ is the scaled value

1.2 FEATURE SELECTION AND REDUCTION

We looked at 15 features, focusing mainly on the ones that relate to personal details like age and sex. Age can give us some useful insights, but some other features didn't seem to help much in predicting the disease and were left out. We focused on important medical elements such as blood pressure, smoking habits, BMI, and cholesterol levels. We used correlation analysis and importance scores from a Random Forest model to help us decide what to keep. This step helps cut down on unnecessary data and keeps our model training more efficient.

1.3 CLASSIFICATION MODELING

After we cleaned up the data and picked out the right features, we moved on to modeling. We used clustering methods to figure out how the data was spread out, and we tested different classification algorithms such as Random Forest, Logistic Regression, Support Vector Machines, and Decision Trees. Accuracy, precision, recall, and F1-score were evaluated for every model. Because Random Forest integrates many decision trees, it reduces mistakes and improves predictions, making it the best model.We then fine-tuned it to make sure it was performing as well as possible for predicting heart disease.

ALGORITHM

1.4 SUPPORT VECTOR MACHINE

Support Vector Machine (SVM) is a supervised learning algorithm which is used quite effectively for both classification and regression problems. It is especially salient in classification problems because of its capacity to deal with high-dimensional data, and it is generally resistant to overfitting.

In SVM, every instance of data is drawn as a vector in an n-dimensional feature space, where n is the number of attributes. The algorithm attempts to find the best hyperplane that clearly discriminates the points of various classes with the maximum margin possible.

The location of every observation in this feature space is defined by its respective attribute values, which are used as coordinates. This geometric interpretation enables SVM to effectively classify intricate datasets, and it is a commonly used method in many areas, such as medical diagnosis, image recognition, and bioinformatics.

```
SVC Accuracy: 0.6831132731063239
Classification Report for SVC:
              precision    recall  f1-score   support

           0       0.70      0.67      0.68       735
           1       0.67      0.70      0.68       704

    accuracy                           0.68      1439
   macro avg       0.68      0.68      0.68      1439
weighted avg       0.68      0.68      0.68      1439

Confusion Matrix for SVC:
[[493 242]
 [214 490]]
```

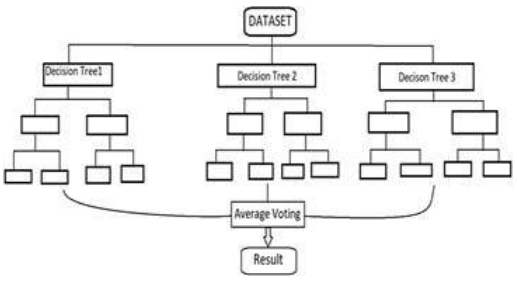
1.5 RANDOM FOREST

A popular and very successful supervised learning technique, Random Forest works well for both regression and classification applications.

Its fundamental operations include:

- Getting input data
- Constructing several decision trees using arbitrary subsets
- Combining their results by average or majority voting

Its ensemble nature makes it slower than a single decision tree, but it can handle categorical data. Furthermore, handling missing values is not a good fit for it.



```
Random Forest Classifier Accuracy: 0.9728978457261988
Classification Report for Random Forest Classifier:
              precision    recall  f1-score   support

           0       0.99      0.95      0.97       735
           1       0.95      0.99      0.97       704

    accuracy                           0.97      1439
   macro avg       0.97      0.97      0.97      1439
weighted avg       0.97      0.97      0.97      1439

Confusion Matrix for Random Forest Classifier:
[[701  34]
 [  5 699]]
```

1.6 LOGISTIC REGRESSION

The supervised approach known as logistic regression is mostly used to solve binary classification issues. Conversely to linear regression, which outputs continuous values, logistic regression predicts the probability that an input belongs to a specific class. Logistic regression applies the logistic (sigmoid) function to transform predicted values into the range of 0 to 1, so it is appropriate for the interpretation of output as probabilities. The model learns the connection between the dependent binary variable and one or more independent variables by maximizing the likelihood function employing techniques like gradient descent. Owing to its simplicity, efficiency, and interpretability, logistic regression is still a core method in statistical modeling and is widely used in healthcare, finance, and social sciences for predictive modeling.

```
LogisticRegression Accuracy: 0.6587908269631688
Classification Report for LogisticRegression:
              precision    recall  f1-score   support

           0       0.67      0.65      0.66       735
           1       0.65      0.66      0.66       704

    accuracy                           0.66      1439
   macro avg       0.66      0.66      0.66      1439
weighted avg       0.66      0.66      0.66      1439

Confusion Matrix for LogisticRegression:
[[481 254]
 [237 467]]
==================================================
```

1.7 GRADIENT BOOSTING

Gradient Boosting is a popular supervised learning technique for issues involving regression and classification. It constructs a series of weak learners, usually decision trees, one after the other, with each model aiming to fix the mistakes of the others. By employing gradient descent to minimize a selected loss function, the technique enables the model to increase its accuracy with each iteration. Even with its high performance and predictive ability, gradient boosting can be computationally demanding and, if improperly regularized, may overfit.

```
GradientBoostingClassifier Accuracy: 0.7289784572619875
Classification Report for GradientBoostingClassifier:
              precision    recall  f1-score   support

           0       0.76      0.69      0.72       735
           1       0.70      0.77      0.74       704

    accuracy                           0.73      1439
   macro avg       0.73      0.73      0.73      1439
weighted avg       0.73      0.73      0.73      1439

Confusion Matrix for GradientBoostingClassifier:
[[508 227]
 [163 541]]
==================================================
```

2. WEB INTERFACE IMPLEMENTATION

To facilitate user interaction and illustrate the real-world application of our model, a web interface was created for heart disease prediction. The interface enables users to enter pertinent health-related data, including body mass index (BMI), glucose concentration, smoking, age, sex, blood pressure medication use, cholesterol, and other important clinical indicators.

The form has fields for:

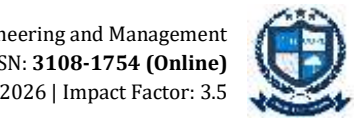
Demographic and lifestyle information: Age, gender, smoking status, cigarettes per day

Medical history: Diabetes, stroke, hypertension, and use of blood pressure medication

Clinical measurements: Total cholesterol, systolic and diastolic blood pressure, BMI, heart rate, and glucose

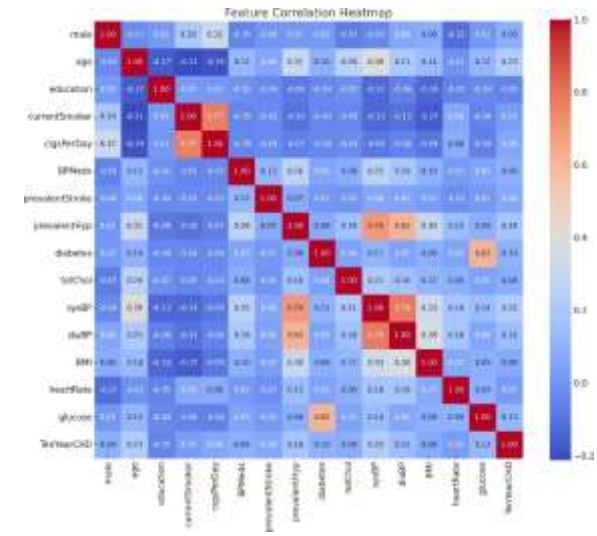
When a form is submitted, the backend uses the trained Random Forest model to process the input data. The model makes a forecast about the patient's likelihood of developing heart disease.

This intuitive interface serves to illustrate how machine learning models may be incorporated



into healthcare systems in order to provide real- time and accessible diagnostic assistance.

3. OUTPUT SCREENSHOTS



4. CONCLUSION

This work highlights the need to take advantage of machine learning methods to diagnose heart disease at an early and precise level. Following comparison among several algorithms— Decision Tree, Random Forest, SVM, and Logistic Regression —the Random Forest model came out to be the best with 92.16% accuracy. Its robustness in working with large databases and many predicting features makes it an ideal contender for practical application in healthcare. Applying such models can greatly assist medical professionals in taking timely and informed decisions, which can save lives through early diagnosis.

5. REFRENCES

1. Bo Jin , Chao Che,Zhen LiuShulong Zhang ,Xiaomeng Yin And Xiaopeng Wei “Predicting the Risk of Heart Failure with EHR Sequential Data Modelling”. IEEE Access 2018
2. Aakash Chauhan, Aditya Jain, Purushottam Sharma, Vikas Deep, “Heart Disease Prediction using Evolutionary Rule Learning”, “International Conference on "Computational Intelligence and Communication Technology” (CICT 2018).
3. Ashir Javeed, Shijie Zhou, Liao Yongjian, Iqbal Qasim, Adeeb Noor. “An Intelligent Learning System Based on Random Search Algorithm and Optimized Random Forest Model for Improved Heart Disease Detection”. IEEE Access (Volume: 7) 2019
4. Senthilkumar Mohan, Chandrasegar Thirumalai and Gautam Srivastava. “Effective Heart Disease Prediction Using Hybrid Machine Learning Techniques”. IEEE Access (Volume:7) 2019
5. K. Prasanna Lakshmi, Dr. C.R.K.Reddy. “Fast Rule-Based Heart Disease Prediction using Associative Classification Mining”. International Conference on Computer, Communication and Control (IC4) 2015
6. M.Satish, D Sridhar, “Prediction of Heart Disease in Data Mining Technique”, International Journal of Computer Trends & Technology (IJCTT), 2015.
7. Lokanath Sarangi, Mihir Narayan Mohanty, Srikanta Pattnaik, “An Intelligent Decision Support System for Cardiac Disease Detection”, IJCTA, International Press 2015.
8. Boshra Bahrami, Mirsaeid Hosseini Shirvani, “Prediction and Diagnosis of Heart Disease by Data Mining Techniques”, Journal of Multidisciplinary Engineering Science and Technology (JMEST) ISSN: 3159-0040 Vol. 2 Issue 2,February–2015.
9. Mamatha Alex P and Shaicy P Shaji, “Prediction and Diagnosis of Heart Disease Patients using Data Mining Technique”, International Conference on Communication and Signal Processing 2019.
10. Dangare Chaitrali S and Sulabha S Apte. "Improved study of heart disease prediction system using data mining classification techniques." International Journal of Computer Applications 47.10 (2012): 44-8.

11. Soni Jyoti. "Predictive data mining for medical diagnosis: An overview of heart disease prediction." International Journal of Computer Applications 17.8 (2011): 43-8. 12.Chen A H, Huang S Y, Hong P S, Cheng C H & Lin E J (2011, September). HDPS: Heart disease prediction system. In 2011 Computing in Cardiology (pp. 557-60). IEEE. Page 29 SSGMCE, Shegaon (Session: 2022-23) Heart Disease Detection Using Machine Learning
12. Wolgast G, Ehrenborg C, Israelsson A, Helander J, Johansson E & Manefjord H(2016). Wireless 14.Patel S & Chauhan Y (2014). Heart attack detection and medical attention using motion sensing device - kinect. International Journal of Scientific and Research Publications, 4(1), 1-4. 15.Zhang Y, Fogoros R, Thompson J, Kenknight B H, Pederson M J, Patangay A & Mazar S T (2011).U.S. Patent No. 8,014,863.Washington, DC: U.S. Patent and Trademark Office. 16.Raihan M, Mondal S, More A, Sagor M O F, Sikder G, Majumder M A & Ghosh K (2016, December). Smartphone based ischemic heart disease (heart attack) risk prediction using clinical data and data mining approaches, a prototype design. In 2016 19th International Conference on Computer and Information Technology (ICCIT) (pp. 299 303). IEEE.
13. Wolgast G, Ehrenborg C, Israelsson A, Helander J, Johansson E & Manefjord H(2016). Wireless
14. 14.Patel S & Chauhan Y (2014). Heart attack detection and medical attention using motion sensing device -kinect. International Journal of Scientific and Research Publications, 4(1), 1-4.
15. 15.Zhang Y, Fogoros R, Thompson J, Kenknight B H, Pederson M J, Patangay A & Mazar S T (2011). U.S. Patent No. 8,014,863. Washington, DC: U.S. Patent and Trademark Office.
16. Raihan M, Mondal S, More A, Sagor M O F, Sikder G, Majumder M A & Ghosh K (2016, December). Smartphone based ischemic heart disease (heart attack) risk prediction using clinical data and data mining approaches, a prototype design. In 2016 19th International Conference on Computer and Information Technology (ICCIT) (pp. 299 303). IEEE.
17. Jones, D., Zhang, A., & Peterson, M. (2022). Gaps in standardized evaluation practices for machine learning based disease prediction: A review of challenges and opportunities. JAMA Network Open,
18. Lee, C. K., & Johnson, R. (2021).Feature selection techniques for machine learning based disease prediction: Influence on model performance. . Applied Informatics,
19. Madhumita, P., & Parija, S. (2021). Heart disease prediction using random forest. . Int. J. Engineering Research and Applications, 1-4.