**Journal of Management and Service Science, 2024,**
Vol. 04, Iss. 01, S. No. 056, pp. 1-9
ISSN (Online): 2583-1798

# Utilising Exploratory Data Analysis and Machine Learning Algorithms for Heart Disease Analysis and Prediction

**Humra Khan[1], P. Singh[2]**

[1,2]Amity School of Engineering and Technology, Amity University (AUUP), Lucknow, (Uttar Pradesh), India.
[1]khanhumra024@gmail.com, [2]pawansingh51279@gmail.com

**How to cite this paper:** H. Khan and P. Singh, "Utilising Exploratory Data Analysis and Machine Learning Algorithms for Heart Disease Analysis and Prediction," *Journal of Management and Service Science (JMSS)*, Vol. 04, Iss. 01, S. No. 056, pp. 1-9, 2024.

https://doi.org/10.54060/a2zjournals.jmss.56

**Received:** 09/06/2023
**Accepted:** 10/03/2024
**Online First:** 25/04/2024
**Published:** 25/04/2024



## Abstract

*As one of the most common and potentially fatal diseases in the world, heart disease must be detected early for proper treatment. With exploratory data analysis (EDA) and machine learning algorithms for predictive analysis, this research project seeks to thoroughly investigate the different aspects that contribute to heart disease. This will enable prompt diagnosis and risk mitigation. Numerous crucial features affecting the diagnosis of heart disease have been found through in-depth exploratory analysis of data. Among these features, the number of major arteries stained by fluoroscopy, the various forms of chest pain, the maximum heart rate reached, exercise-induced angina, the slope of the peak exercise ST segment, and the ST depression brought on by activity relative to rest stand out as most significant factors. Clinicians can learn a great deal about a patient's risk of developing heart disease by carefully examining these characteristics. In order to put this research's predictive component into practice, machine learning classifiers are built using the UCI heart disease dataset, which contains important variables pertaining to cardiac health. For comparison analysis, six different methods are used: Random Forest (RF), Gradient Boost (GB), K-Nearest Neighbour (KNN), Decision Tree (DT), Support Vector Machine (SVM), and Logistic Regression (LR). After conducting a comprehensive analysis, it has been determined that the Random Forest classifier has the best accuracy rate, attaining a remarkable 85.25%.*

## Keywords

*Exploratory Data Analysis (EDA), Machine Learning, Heart Disease Analysis, Heart Disease Prediction*

## 1. Introduction

Early detection of heart disease is detrimental to lowering heart-related issues and safeguarding it from catastrophic risks. Heart disease symptoms can include physical weakness, breathing problems, chest pain, etc. An expert's symptom analysis report, physical laboratory results, and the patient's medical history are used in conjunction with invasive tests, to diagnose cardiac issues [1, 2]. Exploratory Data Analysis (EDA) aids in recognising patterns and highlights the most important characteristics of the data. EDA is ultimately used to check a theory or validate a claim [3]. Machine Learning (ML) along with exploratory data analysis aim to increase the effectiveness and accuracy of medical professionals' work [4]. ML approaches can help with clinical management in various medical applications, including tumour or cancer cell identification [5], natural language processing, and medical picture analysis [6, 7]. The percentage of mortality from heart disease has decreased because of these machine learning-based expert medical decision-making systems [8].

The objective of this paper is to analyse and visualise the heart disease dataset in order to better understand it and identify any hidden trends and correlations. This will help in building a prediction model for heart disease, which determines a person's risk of developing heart disease based on input of feature variables. The classification techniques such as Logistic Regression(LR), Support Vector Machine (SVM), K-nearest neighbour (KNN), Decision Tree (DT), Gradient Boost (GB) and Random Forest (RF) are used to create the prediction models. Confusion Matrix, Accuracy, Precision, Recall and F1-Score are used as performance metrics for the prediction models. Python is used for the implementation in the Jupiter Notebook computing platform. The following is the structure of the remaining sections of this research paper: Literature Survey, Proposed Methodology, Implementation along with Result Analysis and Conclusion.

## 2. Literature Review

Recently, there has been a lot of interest in the field of health prediction, especially with regard to heart disease, at the nexus of exploratory data analysis (EDA) and machine learning (ML) techniques.

Researchers in [9] applied a variety of ML classifiers, such as Decision Trees (DT), Random Forest (RF), KNN, Multilayer Perceptron, and Naive Bayes, using the Framingham Heart Study dataset. Their investigation confirmed the need of fine-tuning data quality and structure before implementing models, highlighting the role of data pre-processing techniques for enhancing prediction accuracy.

Authors in [10] undertook a thorough analysis of pre-processing methods for heart disease classification, underscoring the significance of preprocessing even further. Their results confirmed the idea that careful data pre-processing is the cornerstone of reliable predictive modelling, highlighting the critical role that data reduction and cleaning play in improving the accuracy of heart disease classifiers.

Using the logistic regression approach, the authors of [11] were able to obtain 77% prediction accuracy on the UCI dataset.

Authors [12] improved their work in their study by comparing global evolutionary computation techniques, and as a result, they saw an increase in prediction accuracy.

Researchers looked into the combined effects of exploratory data analysis and data preprocessing for heart disease on predicted accuracy in [13]. The study investigated ML classifiers, including Random Forest, Support Vector Machine (SVM), and Decision Tree, along with feature scaling techniques and data visualisation tools, using several heart disease datasets from the UCI collection. The study's use of Random Forest yielded an astounding accuracy of 86.41%, highlighting the effectiveness of integrated preprocessing and exploratory analysis approaches in improving predictive performance.

Subsequent research initiatives may concentrate on utilising cutting-edge technologies like ensemble modelling and deep learning to further improve forecast accuracy. Furthermore, investigating innovative methods for data preprocessing and merging various datasets may present fresh opportunities to enhance the resilience and applicability of heart disease predic-

tion models.

## 3. Proposed Methodology

Before In the process of turning raw data into useful insights, each stage is vital. The proposed research methodology summarises a methodical way to extract knowledge and create prediction models from a given dataset.

i. Data Gathering
ii. EDA (exploratory data analysis)
iii. Splitting the Dataset
iv. Using Machine Learning (ML) Classification techniques
v. Assessing the ML Classifiers' Effectiveness
vi. Deployment of the Dashboard and ML Model

## 4. Implementation

### 4.1. Data Gathering

The data is fetched from the UCI repository consisting of four databases. It has 76 fields, however, all published researches only mention using a portion of 14 of these. The "target" field alludes to the patient's having heart illness. 0 means there is no disease, while 1 means there is a disease [14]. The features in the dataset used for Heart Disease Analysis and Prediction include: sex, chest pain types, fasting blood sugar, resting blood pressure, cholesterol, resting electrocardiographic results, maximum heart rate achieved, exercise induced angina, ST depression induced by exercise relative to rest, slope of the peak exercise ST segment, number of major vessels (0-3) coloured by fluoroscopy and thallium scintigraphy.

### 4.2. Exploratory Data Analysis

Initially, the data is cleansed. Following extensive cleaning, 302 rows of data remain in the dataset. The data is visualised after obtaining an overall description. The bar graph in **Figure 1**. illustrates that the target feature has the highest positive correlation with cp-Chest Pain Type (0.43) and thalach-max heart rate (0.42). Conversely, the target feature has the highest negative correlation with exang-Exercise-induced Angina (-0.44), oldpeak-ST-depression (-0.43), and ca-Number of blood vessels coloured (-0.41).

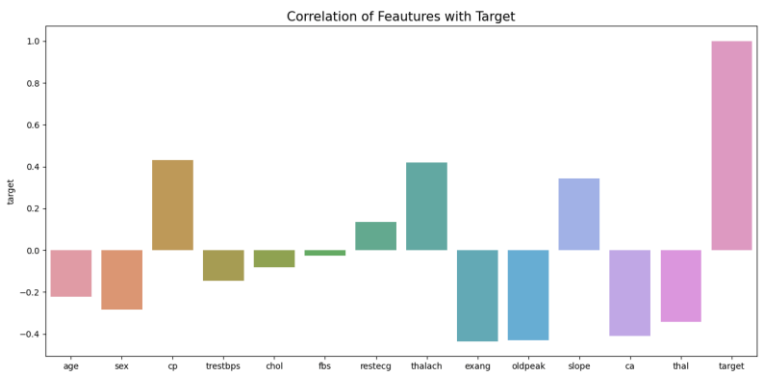


**Figure 1.** Correlation between the target field and all the features

Further results of the Exploratory Data Analysis (EDA) have been summarised in a dynamic and user-friendly Tableau dashboard that

provides a visual depiction of the most important findings regarding the diagnosis of heart disease as seen in **Figure 2**.


**Figure 2.** Heart Disease Dashboard in Tableau

It is evident from **Figure 3**, that the percentage of individuals with heart disease (138 cases) and those without it (164 cases) are close to each other. The Percentage of females (96 in number) is almost half the percentage of males (206 in number) in the dataset. Among people distressed with heart diseases, a fundamentally higher rate involves females (78%) contrasted with guys (44%). On the other hand, among those without heart diseases, the extent of guys (56%) surpasses that of females (22%). Persons without heart disease have a mean value of around 60 years of age, while people with heart disease have ages more spread-out from 35 to 75 years with the mean value of 55 years.

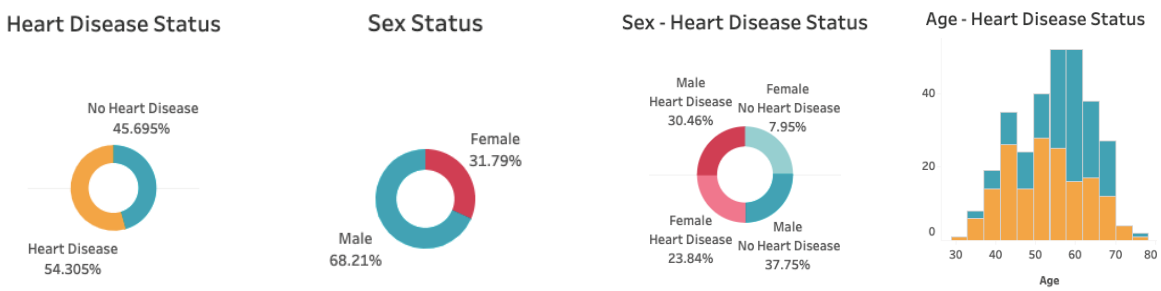


**Figure 3. (a)** Heart Disease Status, **(b)** Sex Status, **(c)** Sex-Heart Disease Status, **(d)** Age – Heart Disease Status

From **Figure 4(a)**, it is clear that the number of false is five times true, i.e., most subjects have healthy fasting blood sugar levels. Astonishingly it is seen that more people with healthy blood sugar levels have heart disease than people with diabetic blood sugar levels. The most common type of chest pain is found to be normal anginal pain, which is followed by non-anginal pain, atypical anginal pain, and pain without symptoms as seen in **Figure 4(b)**. Surprisingly, many people with typical anginal pain do not show signs of heart disease. On the other hand, the correlation between non-anginal pain and heart disease is

clearly evident. From **Figure 4(c)**, it seems to be more strongly correlated with the lack of cardiac disease than with its presence, suggesting that there may be more nuance in the association between symptoms and disease state.

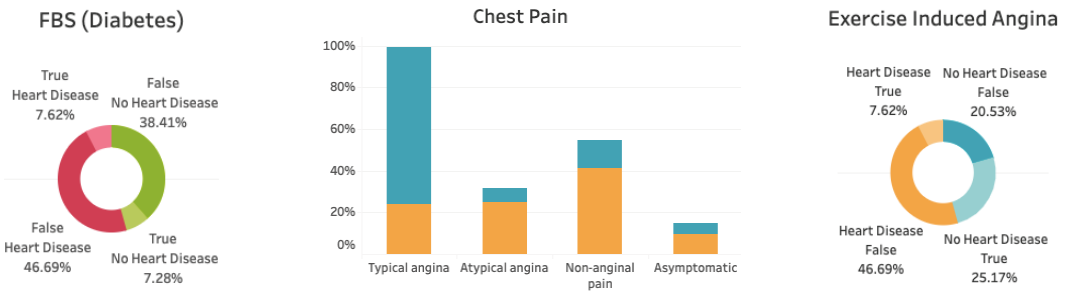


**Figure 4. (a)** FBS Diabetes Donut Chart, **(b)** Chest Pain Bar Chart, **(c)** Exercise Induced Angina Donut Chart

With an emphasis on blood vessel clots, the "Number of Major Vessels" bar chart in **Figure 5(a)** offers a compelling perspective on cardiovascular health. Interestingly, most individuals have non-coloured arteries, which suggests clots are present. In addition, people who have clots are noticeably more likely to have heart disease than people who do not. The distribution of the patients' highest heart rates is shown in **Figure 5(b)**, which shows some intriguing characteristics. Adults' maximum heart rates usually range from 150 to 200 beats per minute (bpm), although some people have lower maximum heart rates than average, which causes the distribution to be left-skewed. There is an interesting association that suggests a higher maximal heart rate is associated with a higher risk of heart disease.

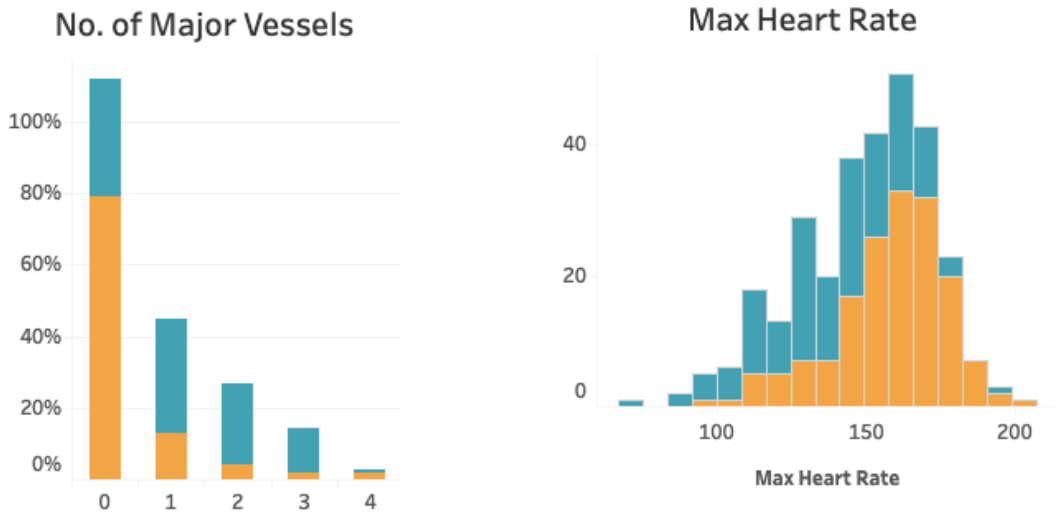


**Figure 5. (a)** No. of Major Vessels Bar Chart, **(b)** Max Heart Rate Histogram

A considerable percentage of results fit into the categories of horizontal and down-sloping ST segment depressions, according to analysis shown in the **Figure 6(a)**. The distribution in **Figure 6(b)** seems to be right-skewed, with most values clustered below 1. This implies that compared to rest, ST depression is generally lower during exercise. A closer look reveals that the distribution of people with heart disease is firmly right-skewed, whereas the distribution of those without heart disease is not widely scattered, suggesting a more consistent pattern.

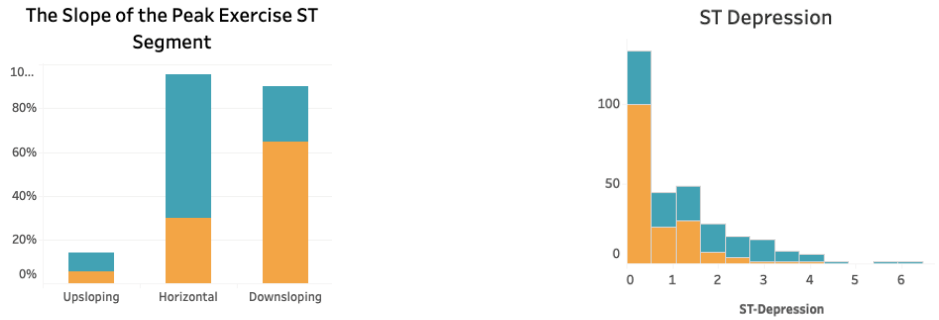


**Figure 6. (a)** Slope of the Peak Exercise ST Segment Bar Chart, **(b)** ST Depression Histogram

### 4.3. Splitting the Dataset

Dataset is then divided into training set and testing set. A known output is part of the training set, and the model is built using this data in order to later generalise it to other data. 80% of the data in this study are used for training

### 4.4. Applying Machine Learning (ML) Classification Algorithms

The classification algorithms proposed for use in this work are Logistic Regression, Support Vector Machine, K- nearest neighbour, Decision Tree, Random Forest and Gradient Boosting.

1. Logistic regression (LR): To estimate probabilities, LR uses a logistic function, commonly referred to as the sigmoid function. Although it may become overfit in high-dimensional datasets, this model performs particularly well with datasets that can be linearly segregated. Regularisation methods like L1 and L2 regularisation can be used to reduce worries about overfitting [15].

2. Support vector machine (SVM): In order for SVM to function, a hyperplane or group of hyperplanes must be built in this space [16]. Finding the hyperplane that maximises the margin, that is, the distance between the hyperplane and the closest data points from each class is the fundamental idea behind support vector machines (SVM). A higher margin typically leads to a lower generalisation error, which improves the performance of the classifier. As SVM works well in high-dimensional spaces, it's very helpful for problems with a lot of features.

3. K-Nearest Neighbour (KNN): KNN [17] is a supervised, "lazy learning" algorithm that uses "instance-based learning" or non-generalizing learning. Instead of concentrating on creating a broad internal model, it keeps all occurrences that correspond to the training set in an n-dimensional space.

4. Decision Tree (DT): This machine learning technique is non-parametric and supervised. Both the classification and regression problems use DT learning techniques. DT categorises the instances by sorting the tree's leaf nodes from root to leaf. The two most common criteria for splitting are "gini" for Gini impurity and "entropy" for information gain [18].

5. Random Forest (RF): The ensemble classification method known as a random forest classifier applies the "parallel ensembling" methodology to simultaneously fit several decision tree classifiers. As a result, it reduces the over-fitting

issue and improves control and forecast accuracy [19, 20].

6. Gradient Boosting (GB) - Gradient Boosting creates a final model by combining several weak learners, typically decision trees, to construct a stronger predictor, much like Random Forests, another well-liked ensemble method. Gradient Boosting is based on the iterative improvement of the model through loss function minimization.

### 4.5. Evaluating the Effectiveness of ML Algorithms

The performance of all the models is assessed using a confusion matrix and all pertinent metrics, such as, accuracy, precision, recall and F1- score [21] as seen in Table 1.

The random forest model is determined to have the highest accuracy (85.2%), along with the highest F1 score, precision, and recall. It is therefore the most appropriate for predicting heart disease.

**Table 1.** Metrics Evaluation For Various Models.

| | Confusion Matrix | Accuracy | Precision | Recall | F1-Score |
|---|---|---|---|---|---|
| LR | [[24 8]<br>[5 24]] | 78.69% | 75.00% | 82.76% | 78.69% |
| SVM | [[24 8]<br>[4 25]] | 80.33% | 75.76% | 86.21% | 80.65% |
| KNN | [[27 5]<br>[7 22]] | 80.33% | 81.48% | 75.86% | 78.57% |
| DT | [[24 8]<br>[9 20]] | 72.13% | 71.43% | 69.97% | 70.18% |
| RF | [[26 6]<br>[3 26]] | 85.25% | 81.25% | 89.66% | 85.25% |
| GB | [[24 8]<br>[4 25]] | 80.33% | 75.76% | 86.20% | 80.65% |

### 4.6. Loading the ML Model

The random forest model comes out on top with an amazing accuracy rate of 85.2%. Next, we'll concentrate on loading this model with the pickle module. The smooth integration of the model's predictive powers into the app for heart disease prediction, depends on this phase.

### 4.7. Deployment of the Dashboard and ML Model

The study culminates in the release of an intuitive application built with Python and the Streamlit framework disease as seen in **Figure 7.** All of the previously recognised critical fields for heart disease prediction are available for data entry by users.

The software makes predictions about a user's risk of having heart disease in real time using a robust random forest model. Additionally, the software includes an interactive dashboard that examines correlations between different areas to provide users a thorough grasp of the variables affecting the prediction of heart disease.

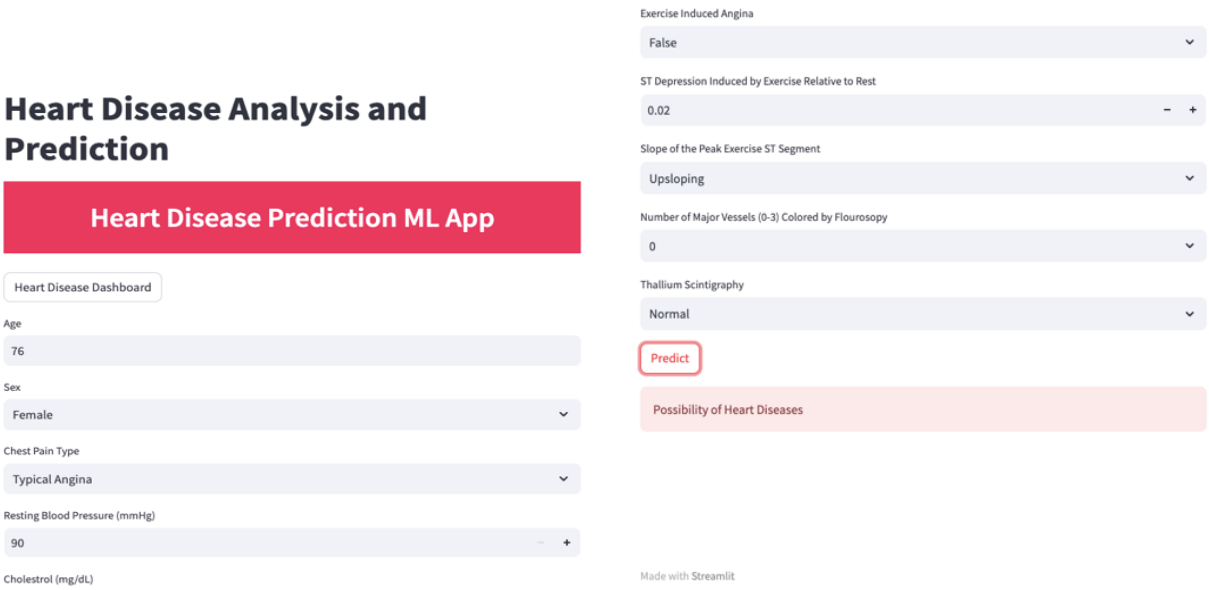


**Figure 7.** Heart Disease Prediction App

## 5. Conclusion

This study aimed to predict heart disease by doing a thorough investigation of a variety of subject features. By examining the distribution and connections between these characteristics using Exploratory Data Analysis (EDA), it became clear that several factors were critical to the diagnosis of heart disease. The most important variables in the diagnosis procedure were, in particular, the forms of chest pain experienced, the highest heart rate attained, exercise-induced angina, the slope of the peak exercise ST segment, the number of major vessels coloured by fluoroscopy, and exercise-induced ST depression. Six machine learning (ML) algorithms were created in the field of predictive modelling Based on performance measures, the Random Forest algorithm performed the best, with an accuracy rate of 85.25% and the greatest performance rate. It also showed excellent recall (89.7%), F1-score (85.2%), and precision (81.2%). This strong performance highlights how well the Random Forest model predicts cardiac disease based on the traits that have been found.

## Acknowledgements

I have endeavored and completed my project; however, it would not have been possible without the kind support and help of many individuals and organizations. I would like to extend my sincere thanks to all of them.

I would like to express my gratitude and thanks to Wg. Cdr. (Dr.) Anil Kumar, Assistant Pro Vice Chancellor and Director of Amity School of Engineering & Technology, Lucknow and Amity University Lucknow for giving me this opportunity of working on the said project. I am highly indebted to my faculty Guide, Dr Pawan Singh for his guidance and constant supervision as well as for providing necessary information regarding the project & also for his support in completing the project.

## References


[1]. A. L. Bui, T. B. Horwich, and G. C. Fonarow, "Epidemiology and risk profile of heart failure," Nat. Rev. Cardiol., vol. 8, no. 1, pp. 30–41, 2011. [Online]. Available: https://doi.org/10.1038/nrcardio.2010.165

[2]. K. Vanisree, "Decision Support System for Congenital Heart Disease Diagnosis based on Signs and Symptoms using Neural Networks," Int. J. Comput. Appl., vol. 19, 2011.

[3]. R. Indrakumari, T. Poongodi, and S. R. Jena, "heart disease prediction using exploratory data analysis," Procedia Comput. Sci., vol. 173, pp. 130–139, 2020. [Online]. Available: https://doi.org/10.1016/j.procs.2020.06.017

[4]. H. C. Koh and G. Tan, "Data mining applications in healthcare," J. Healthcare Inform. Manag., vol. 19, no. 2, pp. 64–72, 2011.

[5]. J. Manhas, R. K. Gupta, and P. P. Roy, "A review on automated cancer detection in medical images using machine learning and deep learning based computational techniques: Challenges and opportunities," Arch. Comput. Methods Eng., vol. 29, pp. 2893–2933, 2021. [Online]. Available: https://doi.org/10.1007/s11831-021-09676-6

[6]. A. Barragán-Montero et al., "Artificial intelligence and machine learning for medical imaging: a technology review," Phys. Med., vol. 83, pp. 242–256, 2021.

[7]. S. Nazir, S. Shahzad, S. Mahfooz, and M. Nazir, "Fuzzylogicbased decision support system for component security evaluation," Int. Arab J. Inf. Technol., vol. 15, pp. 224–231, 2018.

[8]. M. I. Hossain et al., "Heart disease prediction using distinct artificial intelligence techniques: performance analysis and comparison," Iran J. Comput. Sci., 2023. [Online]. Available: https://doi.org/10.1007/s42044-023-00148-7

[9]. O. Sami, Y. Elsheikh, and F. Almasalha, "The role of data pre-processing techniques in improving machine learning accuracy for predicting coronary heart disease," Int. J. Adv. Comput. Sci. Appl., vol. 12, 2021. [Online]. Available: https://doi.org/10.14569/ijacsa.2021.0120695

[10]. H. Benhar, A. Idri, and J. L. Fernández-Alemán, "Data preprocessing for heart disease classification: A systematic literature review," Comput. Methods Programs Biomed., vol. 195, p. 105635, 2020. [Online]. Available: https://doi.org/10.1016/j.cmpb.2020.105635

[11]. R. Detrano et al., "International application of a new probability algorithm for the diagnosis of coronary artery disease," Am. J. Cardiol., vol. 64, no. 5, pp. 304–310, 1989. [Online]. Available: https://doi.org/10.1016/0002-9149(89)90524-9

[12]. B. Edmonds, "Using Localised 'Gossip' to Structure Distributed Learning," Struct. Distrib. Learn., 2005.

[13]. K. Mahalakshmi and P. Sujatha, "The role of exploratory data analysis and pre-processing in the machine learning predictive model for heart disease," in 2023 Int. Conf. Adv. Comput. Commun. Appl. Informatics (ACCAI). IEEE, 2023, pp. 1–8. [Online]. Available: https://doi.org/10.1109/ACCAI58221.2023.10199714

[14]. Heart-disease-dataset Homepage, https://www.kaggle.com/datasets/johnsmith88/heart-disease-dataset/data, last accessed 2023/09/27.

[15]. F. Pedregosa et al., "Scikit-learn: Machine Learning in Python," arXiv [cs.LG], 2012. [Online]. Available: http://arxiv.org/abs/1201.0490

[16]. S. S. Keerthi, S. K. Shevade, C. Bhattacharyya, and K. R. K. Murthy, "Improvements to Platt's SMO algorithm for SVM classifier design," Neural Comput., vol. 13, no. 3, pp. 637–649, 2001. [Online]. Available: https://doi.org/10.1162/089976601300014493

[17]. D. W. Aha, D. Kibler, and M. K. Albert, "Instance-based learning algorithms," Mach. Learn., vol. 6, pp. 37–66, 1991. [Online]. Available: https://doi.org/10.1007/bf00153759

[18]. E. Zeinulla, K. Bekbayeva, and A. Yazici, "Effective diagnosis of heart disease imposed by incomplete data based on fuzzy random forest," in 2020 IEEE Int. Conf. Fuzzy Syst. (FUZZ-IEEE).

[19]. I. H. Sarker, P. Watters, and A. Kayes, "Effectiveness analysis of machine learning classification models for predicting personalized context-aware smartphone usage," J. Big Data, vol. 6, pp. 1–28, 2019.

[20]. Y. Amit and D. Geman, "Shape quantization and recognition with randomized trees," Neural Comput., vol. 9, pp. 1545–1588, 1997.

[21]. J. Brownlee, "Tour of Evaluation Metrics for Imbalanced Classification," Mach. Learn. Mastery, 2021.