International Journal of Innovative Technology and Interdisciplinary Sciences
https://ijitis.org/index.php/IJITIS
ISSN: 2613-7305
Volume 6, Issue 4
DOI: doi.org/10.15157/IJITIS.2023.6.4.1244-1254
Received: 01.11.2023; Accepted: 20.12.2023

Historical Data-Based Heart Disease Analysis Using Machine Learning Techniques

Shruti Mishra[1*], Ajanta Das[1*], Bishal Kumar[1*]

[1] Information Technology, Amity University, Kolkata, India
***mishra.shruti1913@gmail.com, ajanta.desarkar@gmail.com, bishal4409@gmail.com**

Abstract

Healthcare solutions can be provided to every human being with the advancement of machine learning techniques, irrespective of age. Utilizing classification and clustering techniques, diseases can be predicted using a dataset of that specific disease, thereby reducing costs. Due to a lack of knowledge and skills to provide first aid to heart patients, emergency fatalities may occur. This research studies various datasets to identify different features or characteristics causing heart disease. Analysis of these features or the interrelationships between these features can play a vital role in the prediction of heart disease using machine learning algorithms and data mining techniques. The research aims to develop an accurate predictive model that can effectively identify individuals at high risk of developing heart disease. The study utilizes a diverse dataset consisting of various clinical and demographic features, including age, gender, blood pressure, cholesterol levels, diabetes, thalassemia, electrocardiogram readings, etc. The objective of this paper is to propose an integrated framework for pre-processing (as and when required), mining, training, and testing. This research implements three classification algorithms to analyze various historical datasets to make accurate predictions. The classifiers K-Nearest Neighbour, Naïve Bayes, and Decision Tree are employed to train and evaluate predictive models. The dataset is pre-processed, including handling missing values, normalizing features, and addressing class imbalances if present. In order to compare the accuracy of various datasets, a range of evaluation metrics, such as accuracy, precision, recall, and F1-score, are measured, and a performance evaluation confusion matrix is prepared. The results of the study demonstrate that a decision tree classifier with the selected features in the chosen dataset can be used to effectively predict heart disease. The novelty of this research is to select important features causing heart disease with the highest probability.

Keywords*:* Features; Classification; Prediction; Accuracy; Machine Learning Algorithms

INTRODUCTION

The rapid advancements in machine learning techniques have paved the way for their integration into various industries, including healthcare. This paper aims to explore the selection of different features from various datasets and then study the impact of these features in the prediction of heart disease using different classifiers. Certain health parameters, such as high cholesterol, obesity, high blood pressure, and high blood sugar, increase the risk of heart disease [1]. All these symptoms are reminiscent of many diseases that occur in adults. This makes it difficult to make an accurate diagnosis, which can lead to death in the future. This paper utilizes datasets from open sources for accessing patient data to execute machine learning algorithms to accurately diagnose patients and prevent deadly diseases. Various machine learning and deep learning models can be used to diagnose diseases and classify or predict outcomes [2]. According to Melillo et al., an automated classification system for heart failure identifies high-risk and low-risk patients [3]. They used a machine learning algorithm called classification and regression (CART), which scored 93% sensitivity and 63.5% specificity.

The Random Forest and CART achieve an accuracy of 87.6%, exceeding the accuracy achieved for classification. In [4], the support vector machine (SVM) technique used by Parthiban and Srivatsa to identify patients with a history of diabetes and predict cardiovascular disease was 94% successful, with 60% correct predictions. This research considers the main health parameters, such as blood glucose, patient age, and blood pressure data.

This research exploits various open-source datasets, balanced and imbalanced, from Kaggle and selects the important or pertaining features for prediction. In this research work, feature selection plays a vital role in dealing with variable or high volumes of data. The combination of non-parametric data analysis and principal component analysis allows the selection of the best features to achieve better results. This paper presents a brief overview of three chosen classifiers in machine learning. Based on the type of dataset, balanced or imbalanced, this paper proceeds with prediction. In this paper, a framework is proposed with a detailed methodology. We have used pre-processing techniques to address missing values and outliers while ensuring data integrity. This carefully curated dataset was used to train and evaluate the algorithm, allowing for robust performance assessments. The metrics for performance evaluation are accuracy, recall, precision, and F1 (score). The algorithm's improved accuracy, interpretability, and potential for early detection and prevention of heart diseases highlight its significance and potential impact on healthcare.

The objective of this paper is to propose an integrated framework for pre-processing (as and when required), mining, training, and testing. This research implements three classification algorithms to analyze various historical datasets to make accurate predictions. This research work considers open-source datasets for a wide variety of datasets. We have implemented K-nearest neighbor, Naive Bayes, and Decision Tree algorithms and executed them for both datasets using varying training and testing datasets in 80:20 and 70:30 ratios simultaneously. The results achieved and presented in this article show that the decision tree algorithm outperforms others, irrespective of dataset type and selected features. The novelty of this research is to select important features based on classifiers that may cause heart disease with the highest probability.

The organization of the paper is as follows: Related work is presented in Section 2 with the motivation of this research. The authors proposed an integrated framework in Section 3. This section includes the details of various datasets, machine learning techniques, and methodologies used in this research. Section 4 discusses the results and analyzes the impact of the different selected features on heart disease. Section 5 concludes the paper.

RELATED WORK

Due to the increasing prevalence of cardiovascular diseases worldwide, prediction of heart disease is necessary and useful to society using machine learning techniques [5,6]. This section studies related research that was conducted to develop accurate models that can predict the risk of heart disease based on patient data such as demographics, clinical history, and laboratory results. In this review of the literature, we present some of the recent studies in this specific domain.

In [7], Krittanawong et al. (2018) used a dataset of 56,770 patients to develop a model for predicting the risk of heart disease using a combination of machine learning algorithms, including logistic regression, decision trees, and random forest. The study found that the random forest algorithm was the most accurate, with an area under the curve (AUC) of 0.903. Mueen et al. (2019) used a dataset of 299 patients to develop a model for predicting the risk of heart disease using support vector machines (SVM) and deep learning algorithms [8]. The study found that the SVM model achieved an accuracy of 93%, while the deep learning model achieved an accuracy of 92%.

In [9], Majumder et al. (2020) developed a model for heart disease prediction using a dataset of 303 patients and presented the performance comparison of various machine learning algorithms, such as decision trees, logistic regression, SVM, and K-nearest neighbors (KNN). Tison et al. (2019) used a dataset of 400,000 patients to develop a deep learning model for predicting the risk of heart disease using electrocardiogram (ECG) signals [10]. The study found that the deep learning model achieved an accuracy of 93%, which was higher than the accuracy of traditional risk factors such as age, sex, and smoking status.

The literature review reveals that different machine learning algorithms can achieve high accuracy in heart disease prediction, and the choice of algorithm may depend on the size and complexity of the dataset and on the specific research need. Authors are motivated to do analysis by leveraging machine learning classification algorithms to identify individuals at high risk of developing heart disease before symptoms manifest [11]. This enables timely interventions, preventive measures, and lifestyle modifications, leading to improved outcomes and saved lives. Machine learning models offer enhanced accuracy and efficiency in analyzing complex datasets, leading to more accurate predictions and informed decision-making for healthcare professionals. This optimizes healthcare resource allocation and reduces unnecessary procedures, hospitalizations, and associated costs. This article also presents various case studies to identify features that tend to cause heart disease. Next section proposes framework for historical data-based heart disease prediction.

PROPOSED FRAMEWORK

This section presents and elaborates on the proposed layered framework in Figure 1. This layered framework is divided into four phases: pre-processing, feature selection, training, and testing. In the pre-processing phase, mostly removing outliers or cleaning data with missing values, null values, etc. Moreover, range values for any attribute are processed as considerations with minimum, maximum, or median. Also, in some cases, yes or no is processed with 1 or 0 to remove duplicate tuples with exactly the same values. Next, the model identifies features that are mostly causing heart disease. After obtaining the clean data, we divided all the data into two parts: the training dataset and the test dataset. The training phase uses training data to build the learned or trained model. A trained model predicts the output based on the given features. These data are divided into two records: Target 0 (having a healthy heart) and Target 1 (having heart disease). Next, the estimated output is compared with the original target. This comparison is used to verify the accuracy of the design and evaluate its performance. As a final phase, testing evaluates the test data for prediction.

Furthermore, a confusion matrix is generated to check the significance of true positive, true negative, false positive, and false negative. Using these four data points, we calculate precision, recall, f-score, and accuracy to measure the performance of the algorithm and, thereby, the model. The flow of the detailed process is presented in Figure 1.

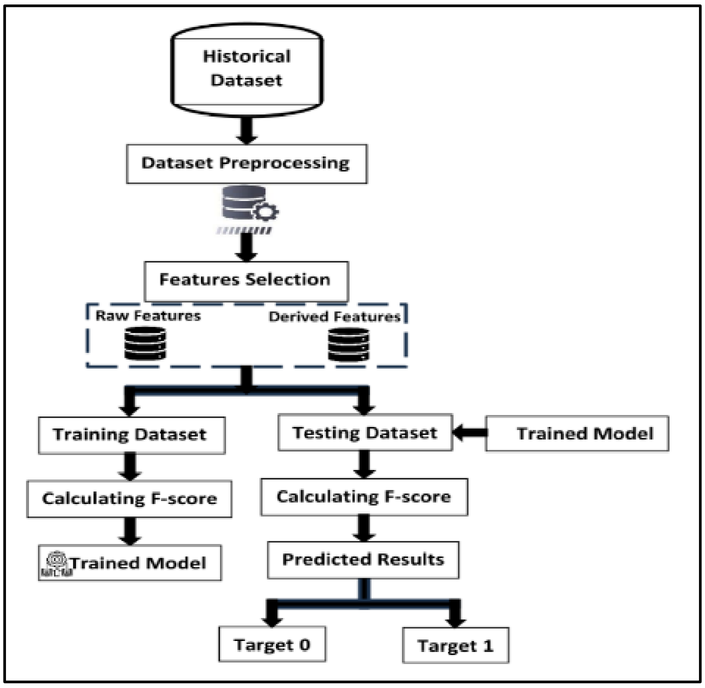

Figure 1. Proposed Framework

Datasets

The proposed framework is implemented and tested with two different types of datasets: balanced and unbalanced. Each dataset is downloaded from the Kaggle web sites [12] and [13]. The dataset downloaded from the Public Health dataset [13] is processed and balanced, termed dataset-1, and another dataset downloaded from [12] is unbalanced, termed dataset-2 in this article.

At first, the input datasets are passed through the pre-processing phase to verify whether they are processed or not. In cases of imbalanced or raw data, pre-processing is necessary. This research exploits oversampling and under sampling of resampling techniques [9, 10] and sets a decision threshold to make the dataset balanced. In reality, most of the datasets are imbalanced, which reflects the skewness between the majority class and the minority class. Resampling techniques are used to remove the bias from an imbalanced training dataset [14]. In cases of oversampling, duplication from the minority class is added, while in cases of under sampling, deletion occurs from the majority class.

Features and data descriptions for each feature are presented in Tables 1 and 2, respectively.

Table 1. Features and Description of Dataset-1

Sl. no	Feature	Description	Data Range/Values
1	age	Age of the patient (numeric)	29-77
2	sex	Gender of the patient (Categorical :0 = female, 1 = male)	0-1
3	cp	Chest pain type (categorical: 0 = typical angina, 1 = atypical angina, 2 = non-anginal pain, 3 = asymptomatic)	0-3
4	trestbps	Resting blood pressure (in mm Hg) (numeric)	94-200
5	chol	Serum cholesterol level (in mg/dL) (numeric)	126-564
6	restecg	Resting electrocardiographic results (categorical: 0 = normal, 1 = having ST-T wave abnormality, 2 = showing probable or definite left ventricular hypertrophy)	0-2
7	thalach	Maximum heart rate achieved (numeric)	71-202
8	oldpeak	ST depression induced by exercise relative to rest (numeric)	0-6.2
9	slope	The slope of the peak exercise ST segment (categorical: 0 = up-sloping, 1 = flat, 2 = down-sloping)	0-2
10	ca	Number of major vessels colored by fluoroscopy (0-3) (numeric)	0-3
11	thal	Thalassemia (categorical: 0 = normal, 1 = fixed defect, 2 = reversible defect)	0-2
12	target	Target: Presence of heart disease (0 = no, 1 = yes)	0-1

Methodology Used

The process of prediction starts with the pre-processing of data, thereby feature selection. Feature selection plays an important role in heart disease prediction. We have selected a group of features, executed classifiers, and observed the target value. However, for various groups of features, the results of target values for different classifiers are very similar. Therefore, it is concluded that age, sex, blood pressure, high cholesterol, and blood glucose are very prominent features of heart disease.

Next, this research uses three different machine learning classification algorithms, K-Nearest Neighbour, Naïve Bayes, and Decision Tree, to prepare the training model. The trained model is used in the next step to predict and compare the performance of the algorithms, thereby comparing the proposed model with the test dataset.

This section briefs three selected classifiers for ease of understanding. Furthermore, to identify the unique contribution of this research, three different case studies of various combinations of features (a selected group of features) were conducted to verify the severe causes of heart disease, thereby predicting the severity of heart disease.

a) K- Nearest Neighbour (KNN):

The K Nearest Neighbour (KNN) class of supervised machine learning algorithms used for classification and regression It is a non-parametric algorithm that classifies new data according to the training samples. In KNN, "K" refers to the number of nearest neighbors to determine the distribution or prediction of new data. For example, if K = 5, the algorithm looks at the 5 nearest neighbors to make a decision. The distance between the new sample and each training sample is calculated using a distance measure, usually the Euclidean or Manhattan distance. To identify a new sample, KNN finds the K nearest neighbors based on their distance from the new sample. KNN is widely used in many fields, including image recognition, recommendation, and bioinformatics. However, it has some limitations, such as the need for training data, sensitivity to K selection, and the cost of measuring distance and calculating the distance between file sizes. It is generally suitable for small or medium data where distance calculation can be efficient.

b) **Naive Bayes:**

Naive Bayes (NB) is a popular class of supervised machine learning algorithms used for classification tasks. The algorithm is called "naive" because it disregards any correlations or dependencies that may exist between them. This assumption simplifies the computation of probabilities and makes the algorithm computationally efficient with conditionally independent variables given the class label and using the Bayes algorithm. Naive Bayes operates by calculating the probability of a new data point belonging to each possible class and then assigning it to the class with the highest probability. It uses prior probabilities (the probabilities of each class occurring in the training data) and conditional probabilities (the probabilities of each feature value given a specific class) to make these predictions. It counts the occurrences of each feature value in each class and calculates the corresponding probabilities. This information is then used to make predictions on new or test data. Naive Bayes is particularly well-suited for text classification tasks, such as spam detection or sentiment analysis, where the features often correspond to the presence or absence of specific words.

c) **Decision Tree:**

Decision tree (DT) is a widespread machine learning algorithm used for classification and regression. This classifier creates a tree with different features from the input properties and dataset. Decision tree algorithms learn from training data by iteratively segmenting data according to the importance of different concepts. It starts with the entire dataset and selects the features that provide the most important segmentation in terms of improving classification or reducing regression errors. At each point in the tree, the decision rule is applied according to the selected features based on information gain value. The value of information gained directs the next point of the path to take in the tree until it reaches the leaf node. Decision trees can handle both categorical and numerical variables and can capture nonlinear relationships between variables. DT is easy to interpret and visualize and can capture nonlinear relationships between variables. Decision trees are valuable for prediction based on some input features. They are especially useful when working with data that has mixed features, missing values, or no correlation. Next section presents results of the two datasets, dataset1 and dataset2.

Table 2. Features and Description of Dataset-2

Sl. no	Feature	Description	Data Range/Values
1	HighBP	High blood pressure (0 = no, 1 = yes)	0-1
2	HighChol	High cholesterol (0 = no, 1 = yes)	0-1
3	CholCheck	Cholesterol check (0 = no, 1 = yes)	0-1
4	BMI	Body mass index (a measure of body fat based on height and weight)	126-564
5	Smoker	Smoking status	0-1
6	Stroke	(0 = non-smoker, 1 = smoker)	0-1
7	Diabetes	History of stroke (0 = no, 1 = yes)	0-1
8	Fruits	Diabetes status (0 = no, 1 = yes)	0-1
9	Veggies	Consumption of fruits	0-1
10	HvyAlcoholConsump	Consumption of vegetables	0-1
11	AnyHealthcare	Heavy alcohol consumption	0-1
12	GenHlth	(0 = no, 1 = yes)	1-5
13	MentHlth	Access to healthcare	0-30
14	PhysHlth	(0 = no, 1 = yes)	0-30
15	DiffWalk	General health status	0-1
16	Sex	(e.g., self-rated on a scale)	0-1
17	Age	Mental health status	0-13
18	target	(e.g., self-rated on a scale)	0-1

RESULTS AND DISCUSSION

This research work considers two different datasets, dataset 1 and dataset 2, for prediction accuracy. The source and description of these datasets are already discussed in Section 3. Dataset-1 is processed and contains 1025 records, while Dataset-2 is unbalanced and contains more than 2.5 lacs. However, after feature selection and preprocessing, the data size was reduced to 44441. Next, we have executed three classifiers for both balanced and unbalanced datasets. The same methodology repeats for different ratios of training data and testing data, such as 80:20 and 70:30. Table 3 and Table 4 present the prediction results of datasets 1 and 2, respectively.

Moreover, visualization of Table 3 data is presented in Figures 2 and 3 for varying the ratio of training to testing dataset in the 80:20 and 70:30 ratios, respectively. Next, visualization of Table 4 data is presented in Figures 4 and 5 for varying the ratio of training to testing dataset in the 80:20 and 70:30 ratios, respectively.

Table 3. Prediction of Heart Disease in Dataset-1

Algorithms	80:20		70:30	
	F1-Score (%)	Accuracy (%)	F1-Score (%)	Accuracy (%)
K-Nearest Neighbors	73.42	73.17	71.24	71.42
Decision Tree	98.52	98.53	96.88	97.07
Naïve-Bayes	81.77	80.0	82.35	81.49

Table 4. Prediction of Heart Disease in Dataset-2

Algorithms	80:20		70:30	
	F1-Score (%)	Accuracy (%)	F1-Score (%)	Accuracy (%)
K-Nearest Neighbors	80.66	79.46	80.27	79.08
Decision Tree	85.88	83.83	85.41	83.18
Naïve-Bayes	65.13	70.16	64.59	69.88

Next, this article presents three different case studies varying different features or combinations of features from Dataset 1 in Table 5 and Table 6, varying the ratio between the training and testing datasets (80:20) and (70:30), respectively.

Case Study I: Selected features are sex, age, oldpeak (stress test depression on an ECG induced by exercise stress), and thal (thalassemia).

Case Study II: Selected features are age, slope (stress test segment slope of the ECG), and thalach (maximum heart rate achieved).

Case Study – III: Selected features are *cp* (chest pain), *chol* (serum cholesterol level) *trestbps* (resting blood pressure) and *restecg* (resting electrocardiographic results)

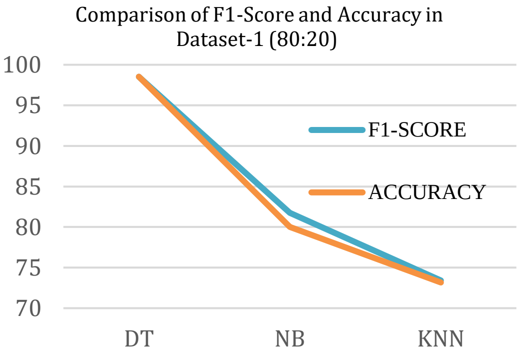

Figure 2. F1-Score and Accuracy in Dataset-1 (80:20)

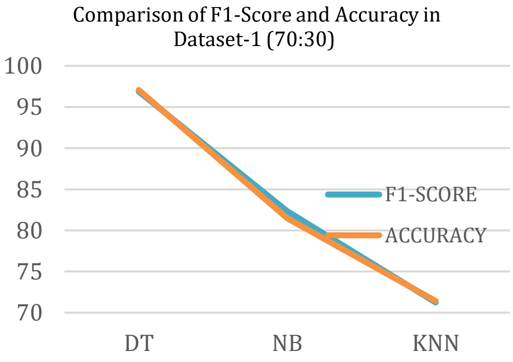

Figure 3. F1-Score and Accuracy in Dataset-1 (70:30)

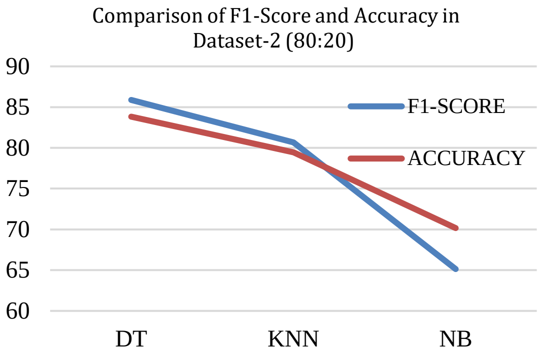

Figure 4. F1-Score and Accuracy in Dataset-2 (80:20)

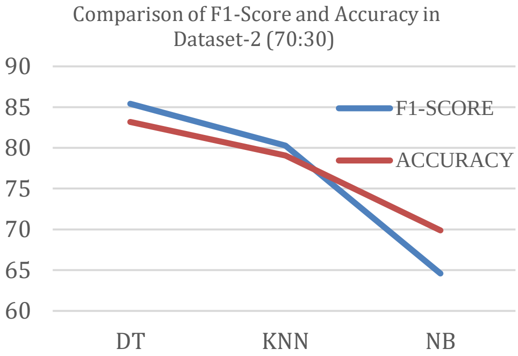

Figure 5. F1-Score and Accuracy in Dataset-2 (70:30)

Table 5. Comparison of Classifiers Using Varying Case Studies of Dataset-1 in 80:20

Algorithms	All Features		Features selected in Case-I		Features selected in Case-II		Features selected in Case-III	
	F1-Score	Accuracy	F1-Score	Accuracy	F1-Score	Accuracy	F1-Score	Accuracy
Decision Tree	98.52	98.53	94.05	94.14	99.01	99.02	97.51	97.56
K-Nearest Neighbors	73.42	73.17	78.09	77.56	69.18	72.19	64.21	66.82
Naïve-Bayes	81.77	80	71	71.7	72.3	71.21	75.12	76.09

Table 6. Comparison of Classifiers Using Varying Case Studies of Dataset-1 in 70:30

Algorithms	All Features		Features selected in Case-I		Features selected in Case-II		Features selected in Case-III	
	F1-Score	Accuracy	F1-Score	Accuracy	F1-Score	Accuracy	F1-Score	Accuracy
Decision Tree	96.88	97.07	91.98	92.53	97.24	97.4	99.32	99.35
K-Nearest Neighbors	71.24	71.42	72.22	79.22	72.99	75.97	66.89	68.5
Naïve-Bayes	82.35	81.49	71.76	72.4	70.66	71.42	75.17	76.62

In summary of the inference, it is evidenced that patients with problems in Case Study I do not have severe heart disease. But patients with difficulties in Case Study II with an 80:20 ratio between training and testing datasets have high chances of heart disease as per the prediction of the DT classifier, which is more than 99%. Finally, patients with difficulties in Case Study III with a 70:30 ratio between training and testing datasets have the highest probability of heart disease as per the prediction of the DT classifier, more than 99.3%.

It is observed that for both datasets, the performance of decision tree classifiers is highest. Likewise, as an inference, we can conclude that the prediction algorithm gives a better result with a processed and balanced dataset compared to a raw and unbalanced dataset.

CONCLUSION

Heart prediction using machine learning projects is motivated by the pressing need to improve public health outcomes and reduce the burden of heart diseases. This paper aims to address key challenges in cardiovascular health through early detection of features causing difficulties or discomforts faced by people. In this article, the developed predictive models exhibit promising performance with high accuracy, precision, and recall rates. The research findings highlight the potential of machine learning techniques for early detection and risk assessment of heart disease, enabling early intervention and personalized treatment plans. Furthermore, research and validation on larger datasets are necessary to enhance the accuracy and reliability of the models for clinical implementation. Hence, dataset-2 is used for societal purposes, and its accuracy proves that the decision tree algorithm performs better compared to others.

In summary, the motivation for heart prediction using machine learning projects lies in improving health outcomes, personalized medicine, accuracy, cost reduction, and public health impact. This article proves that the decision tree classifier predicts more accurately than the other two classifiers. This research directive can be followed with preventive measures on a population level. It is necessary to measure predictions using other classifiers for better accuracy, which can be considered a limitation of this research work. However, this article also presents three different case study analyses for a better selection of features causing heart disease. Through early detection, personalized interventions, enhanced accuracy, cost reduction, and population-level impact, machine learning-based heart prediction research holds great promise for improving public health and fostering a healthier future.

CONFLICT OF INTERESTS

The authors confirm that there is no conflict of interests associated with this publication.

REFERENCES

[1]. Rohit Bharti, Aditya Khamparia, Mohammad Shabaz, “Prediction of Heart Disease Using a Combination of Machine Learning” Hindawi, Journals, Volume 2021, https://doi.org/10.1155/2021/8387680, 01 July 2022.

[2]. T. Hastie, R. Tibshirani, and J. Friedman, “The elements of statistical learning,” Data Mining, Inference, and Prediction, Springer, Cham, Switzerland, 2020.

[3]. P. Melillo, N. De Luca, M. Bracale, and L. Pecchia, “Classification tree for risk assessment in patients suffering from congestive heart failure via long-term heart rate variability,” IEEE Journal of Biomedical and Health Informatics, vol. 17, no. 3, pp. 727–733, 2013.

[4]. G. Parthiban and S. K. Srivatsa, “Applying machine learning methods in diagnosing heart disease for diabetic patients,” International Journal of Applied Information Systems, vol. 3, no. 7, pp. 25–30, 2012.

[5]. Krittanawong, C., Zhang, H., Wang, Z., Aydar, M., & Kitai, T. (2018), Artificial intelligence in precision cardiovascular medicine. Journal of the American College of Cardiology, 69(21), 2657-2664.

[6]. Mueen, A., Hossain, M. E., & Alajlan, N. (2019). Deep learning techniques for heart disease diagnosis and prediction: A survey. Information Fusion, 50, 71-79.

[7]. Majumder, A., Saha, S., Banik, S., & Pal, R. (2020), Comparative analysis of various machine learning algorithms for heart disease prediction. Computers in Biology and Medicine, 120, 103715.

[8]. Tison, G. H., Zhang, J., Delling, F. N., Deo, R. C., & Bhavnani, S. P. (2019), Automated deep- learning ECG analysis for the detection of heart rhythm and conduction abnormalities: a validation study. The Lancet, 394(10207), 861-867.

[9]. How to balance a dataset in Python: accessed from https://towardsdatascience.com/how-to-balance-a-dataset-in-python-36dff9d12704, 10-07-2023.

[10]. How to fix Imbalance Dataset: accessed from https://www.kdnuggets.com/2019/05/fix-unbalanced-dataset.html, 10-07-2023

[11]. Han, J., M. Kamber and J. Pei, "Data Mining Concepts and Technique", Third Edition, 2012

[12]. Heart Disease Health Indicator Dataset:https://www.kaggle.com/datasets/alexteboul/heart-disease-health-indicators-dataset

[13]. Public Health Dataset: https://www.kaggle.com/datasets/johnsmith88/heart-disease-dataset

[14]. "Random Oversampling and Under sampling for Imbalanced Classification", by Ja. Brownie, Jan 2021,https://machinelearningmastery.com/random-oversampling-and-undersampling-for-imbalanced-classification/