SSRG International Journal of Electrical and Electronics Engineering
ISSN: 2348-8379/ https://doi.org/10.14445/23488379/IJEEE-V11I10P117

Volume 11 Issue 10, 168-173, October 2024

Original Article

Integrating Artificial Intelligence with Decision Tree Classifiers for Superior Heart Disease Detection

Hardik Prajapati[1], Dushyantsinh B. Rathod[2]

[1]*Faculty of Engineering and Technology, Sankalchand Patel University, Gujarat, India.*
[2]*Ahmedabad Institute of Technology, Gujarat, India.*

[1]*Corresponding Author : hardikjp2707@gmail.com*

Received: 11 August 2024 Revised: 10 September 2024 Accepted: 12 October 2024 Published: 30 October 2024

Abstract - *Due to the heart's crucial function as one of the most essential systems in the human body, it needs concentrated care. Given the correlation between several illnesses and cardiovascular well-being, it is important to possess precise data for forecasting such ailments. An investigation that compares different aspects of this field is essential for this objective. Many people nowadays suffer from illnesses often discovered at a late stage, mostly because of the imprecise nature of diagnostic methods. Hence, it is crucial to determine the most important data for illness prediction. The use of machine learning, an exceptionally efficient testing technique, is very pertinent in this context. Artificial intelligence operates by the use of iterative testing and training procedures. One of its subfields, called machine learning, is instructing robots to imitate human capabilities. The integration of these technologies is typically connected with the term "artificial intelligence" since they are trained to recognize and use data. In this study, we use physiological indicators such as cholesterol levels, heart rate, biological sex, and age as test data to compare the accuracy of different machine learning algorithms. Machine learning naturally learns from natural phenomena. This project specifically employs three algorithms: Gaussian Naive Bayes, Support Vector Machine, and Logistic Regression. The first segment of this article provides a comprehensive introduction to artificial intelligence and its association with heart-related concerns. The second portion explores the intricacies of the Data Mining Algorithm. The third portion examines the current body of literature. The architecture under consideration is examined in the fourth part. The fifth part provides a concise overview of the dataset and its features. The last part provides a summary and a concise examination of the future potential of the investigation.*

Keywords - *Supervised, Confusion matrix, Linear regression, Unsupervised, Python, Reinforced.*

1. Introduction

Given the heart's critical role as Among the majority of vital systems in the human anatomy, it requires focused attention. Since many diseases are linked to heart health, it is essential to have accurate information for predicting such conditions. A comparative study in this area is crucial for this purpose. Numerous individuals today experience diseases that are diagnosed too late, frequently due to a lack of precision in diagnostic tools. Therefore, identifying the most valuable data for disease prediction is vital.

Machine learning, a highly effective testing method, is particularly relevant here. Artificial intelligence functions through testing and training processes. One of its subfields, machine learning, involves training machines to replicate human abilities. Such methods are taught to recognize and utilize data, which is why the term "artificial intelligence" is often associated with integrating these technologies. Machine learning inherently learns from natural occurrences, and in this project, we utilize physiological parameters such as cholesterol levels, heart rate, biological sex, age, and more as test data to contrast the precision of various algorithms. Specifically, this project employs three algorithms: Gaussian Naive Bayes, Support Vector Machine and Logistic Regression.

The integration of AI in healthcare allows for automated analysis of Electronic Health Records (EHRs), imaging, and genetic data, which can enhance diagnostic accuracy and provide individualized treatment recommendations. In heart disease prediction, machine learning algorithms such as Decision Trees, Support Vector Machines (SVM), and Logistic Regression have shown potential in identifying high-risk patients, allowing for timely intervention. Ensemble models, which combine the strengths of multiple algorithms, are increasingly used to improve predictive accuracy in complex medical datasets. By leveraging these AI-driven techniques, healthcare providers can make data-driven decisions, ultimately improving patient outcomes and optimizing healthcare resources.

Despite its promise, AI in heart disease prediction still faces challenges, including data quality, algorithm transparency, and model interpretability. However, with ongoing advancements and collaboration between clinicians and data scientists, AI continues to evolve as a critical component in the early detection and management of heart disease, aiding in the global effort to reduce the impact of this pervasive condition.

The first section of this article offers an overview of artificial intelligence and heart-related issues. The second section delves into the data mining algorithm. The third section reviews existing literature. The proposed architecture is discussed in the fourth section. The fifth section briefly outlines the project's dataset and attributes. The final section concludes with a summary and a brief look into the future scope of the study.

2. Data Mining Algorithm

There are so many Data mining algorithms, but the following algorithms are referred to for the research study.

2.1. Gaussian Naive Bayes (GNB)

GNB is a strategy for classifying data using probabilistic machine learning. According to this methodology, each attribute or predictor has a distinct and separate impact on predicting the result variable. Naive Bayes is an artificial intelligence function that is used for classification challenges based on the principles of Bayes probability theorem. It is very efficient in text categorization, particularly when dealing with extensive training datasets. Naive Bayes is used in several domains, such as emotion recognition, spam filtering, and categorization of news items. The approach is well recognized for its high level of efficiency, enabling rapid forecasts and model development. It was one of the initial techniques used to solve text categorization difficulties.

2.2. Support Vector Machine (SVM)

Supervised learning is popular. Machine learning applications like classification employ SVM, which excels in regression and classification. The SVM approach seeks to find the optimum line or decision boundary to split classes in a multidimensional space. This divide will simplify future data categorization. A Support Vector Machine's best decision boundary is a hyperplane. SVM is used to identify support vectors for the hyperplane. Support Vector Machine (SVM) uses support vectors to find the hyperplane.

2.3. Decision Tree

Decision trees and supervised learning algorithms may solve regression and classification problems. Their application is common in categorization. The core nodes indicate dataset properties, the branches represent decision rules, and the leaf nodes represent this hierarchical model's final results or conclusions.

Decision and leaf nodes make up a decision tree. Leaf nodes represent decision results and cannot be enlarged. However, decision nodes generate options and may have several branches. The information set's features inform the tree's decisions.

2.4. Logistic Regression (LOR)

The LOR method is a well-known machine learning methodology within the supervised learning category. It is used in conjunction with many independent variables to predict a categorical dependent variable. Discrete values, such as binary options like true or false, yes or no, 0 or 1, etc., may be used in logistic regression to predict a categorical dependent variable. The probabilistic values obtained from logistic regression analysis range between 0 and 1 rather than being precise binary integers of 0 and 1.

3. Literature Survey

Varun Kumar et al. [1] proposed his architecture with an accuracy of up to 85%, the convolutional neural network method analyses the peril of early heart disease using structured data. In addition, photos and unstructured data can be handled using the CNN technique.

P. K. Gupta et al. [2] tested several machine learning techniques to detect heart diseases using the dataset. Using a modified random forest, they achieved 86.84% accuracy. This approach works well in real-time, and adding additional data with CNN and deep learning algorithms might improve its accuracy.

Brahami et al. [3] used healthcare dataset categorization, a machine learning priority. We explored Logistic Regression, Adaptive Boosting, and Multi-Objective Evolutionary Fuzzy Classifiers. All individually, Majority Voting is 80.20% accurate, Logistic Regression the lowest, and AdaBoostM1 the highest.

Rakesh Kumar et al. [4], in his study, the accuracy of the four different machine learning algorithms was examined; KNN delivered the best outcome, with an accuracy of 87%.

Shamsheela Habib et al. [5], in their current study, we offer a novel machine learning technique as the foundation for our advanced cardiac disease prediction approach. Finding correlation-based features that improve prediction accuracy is its main objective. We utilise the UCI Vascular Cardiovascular Disease Dataset in our work and juxtapose our results with those of an earlier investigation. The accuracy of the model we recommended was 85.43%.

Xu Wenxin et al. [6] developed an innovative solution to predict cardiac disease utilizing SVM, decision tree, and ANN models with 87% accuracy.

Shaicy Shaji et al. [7] the proposed scheme aims to diagnose various cardiac diseases and swiftly set preventive measures in place at an affordable cost. The method employs data mining addresses to forecast cardiac problems by feeding information into the Random Forest, SVM, and KNN classification methods. While KNN obtained an accuracy of 83 percent, SVM and random forest model both reached an accuracy of 85 percent.

4. Proposed Architecture

Figure 1 illustrates the several procedures involved in forecasting heart disease. The diagram presents a machine learning workflow for classification tasks. Initially, the dataset is divided into two parts: one for training and the other for testing. The training dataset is used to develop models using various algorithms, such as Logistic Regression, Gaussian Naive Bayes, Support Vector Machines and Decision Trees.

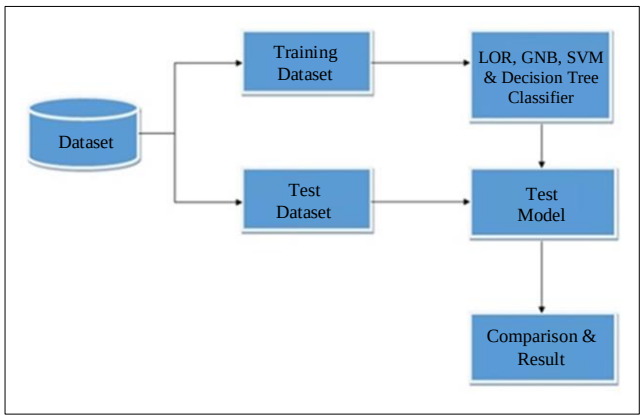


Fig. 1 Proposed architecture

After the models are trained, they are applied to the test dataset to assess their performance. The final step involves comparing the results from these models to evaluate their effectiveness, with the outcomes summarized for comparison and analysis. This process allows for a systematic comparison of different classifiers regarding accuracy and efficiency.

Algorithm :

1. The first phase of the approach is gathering data from many sources, including hospitals, which may be categorized as structured, semi-structured, or unstructured.
2. Once the data is acquired, it goes through a cleansing procedure to remove any missing information and consolidate it into a more refined level of detail. Afterwards, the sanitized data is separated into separate datasets for testing and training.
3. The data is partitioned and then subjected to SMOTE to handle class imbalance. It is then included in several machine learning methods, including Logistic Regression, Gaussian Naive Bayes, SVM, and Decision Tree. This phase is dedicated to training the model to improve its ability to accurately predict outcomes using the provided training data.
4. After training, the model may be tested.
5. The trained model's performance is assessed by testing it on a distinct dataset to verify its correctness and efficacy.
6. Once the necessary degree of anticipated accuracy is reached, the example is deployed.

```
from sklearn.metrics import classification_report
DT_Pred=DTmodel.predict(x_test)
DTreport = classification_report(y_test, DT_Pred)
print(DTreport)
```

	precision	recall	f1-score	support
0	0.91	0.88	0.89	48
1	0.88	0.91	0.90	47
accuracy			0.89	95
macro avg	0.90	0.89	0.89	95
weighted avg	0.90	0.89	0.89	95

Fig. 2 Report of decision tree classification (recall, precision, f1-score)

Figure 2 displays a report that classifies Decision Trees into several categories. The Decision Tree algorithm attained a precision of 0.89. The recall value and f1-score both have a value of 0.89.

4.1. Analysis of Decision Tree Performance Metrics

4.1.1. Precision (0.89)

Precision is the proportion of true positive classifications (correctly identified cases of heart disease) out of all instances classified as positive by the Decision Tree.

A precision of 0.89 indicates that 89% of the cases classified as positive by the Decision Tree are indeed correct predictions. This high precision is beneficial in reducing false positives, meaning fewer healthy individuals are misclassified as having heart disease.

4.1.2. Recall (0.89)

Recall measures the algorithm's ability to identify all true positive cases in the dataset, i.e., how many of the actual cases of heart disease the model correctly identifies.

With a recall of 0.89, the Decision Tree successfully captures 89% of all heart disease cases, which is valuable for ensuring that as few true cases as possible go undetected. High recall is crucial in medical contexts where missing a diagnosis could have serious implications.

4.1.3. F1-Score (0.89)

The F1-score, the harmonic mean of precision and recall, balances these two metrics' trade-offs. It is particularly useful when both precision and recall are equally important. An F1-score of 0.89 indicates that the model achieves a well-rounded

performance in terms of both precision and recall. This score suggests that the Decision Tree is not only accurate in its positive classifications but also reliably identifies most cases of heart disease.

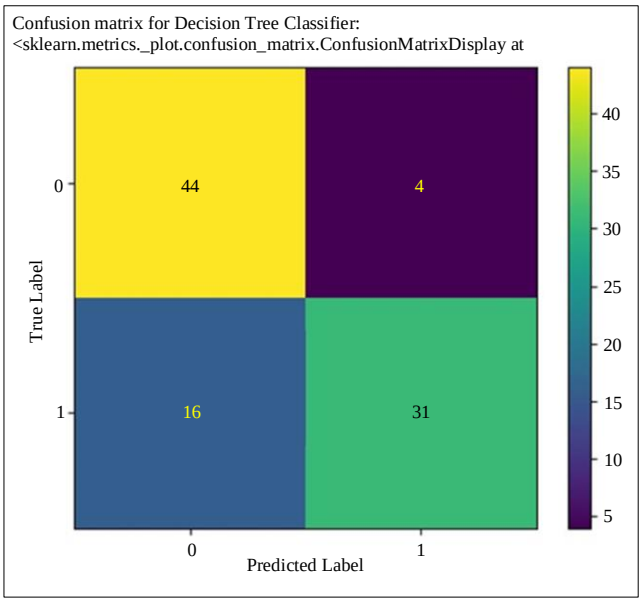


Fig. 3 Decision tree confusion matrix

Figure 3 illustrates a confusion matrix representing the performance of a Decision Tree. The function provides both the confusion matrix and the names of the groups. The heat map tool may be used to graphically depict the confusion matrix.

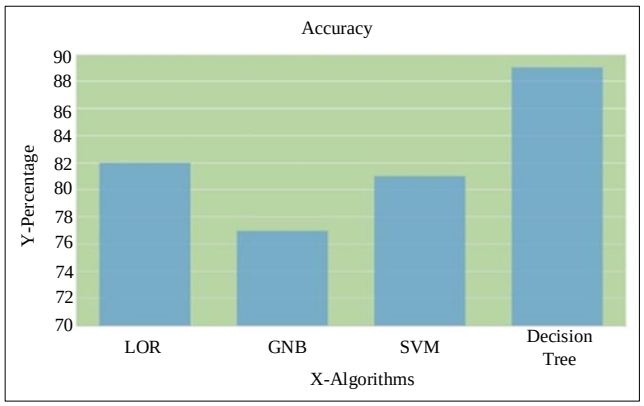


Fig. 4 Accuracy chart of ML algorithms

Table 1. Accuracy table of ML algorithms

Algorithms	Accuracy %
LOR	82
GNB	77
SVM	81
Decision Tree	89

Figure 4 displays a graphic that demonstrates the accuracy of machine learning techniques. The LOR model attains an accuracy rate of 82%, the GNB model attains an accuracy rate of 77%, the SVM model attains an accuracy rate of 81%, and the Decision Tree model attains an accuracy rate of 89%.

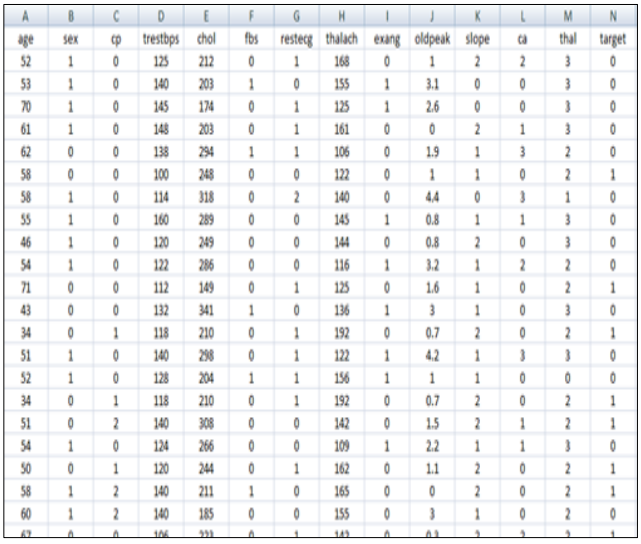
5. Dataset & Model

5.1. Hospital Data

In this paper, we examined medical datasets from hospital records stored in our database [1], encompassing a total of 14 features. They include fundamental patient data like sex, age and cholesterol levels and structured data like laboratory findings, which are all important for diagnosing heart failure. Unstructured data, some of which is outlined in Figure 5, is also considered for future exploration. Our focus has been on predicting heart disease risk with our model. The goal is to establish if a person has heart disease now or is at risk of getting it in the future. The model requires users to input values related to various patient attributes = (x1; x2; xn). This data, which includes general, laboratory, and medical information, is handled by the algorithm, leading to predictions that are more accurate compared to each of the other algorithms that have been studied.

5.2. Data Pre Processing

As expected, missing data, which can come from numerous places, including human error, can impair the accuracy of predicting. To maintain accuracy, it is essential to address this data loss. The model receives the data after any redundant features are eliminated and missing attributes are filled in. The management of this process occurs in the pre-processing phase when the dataset is separated into test and training sets using randomization. This division allows for the calculation of accuracy, and it gets utilized in assessment the model's performance.



A	B	C	D	E	F	G	H	I	J	K	L	M	N
age	sex	cp	trestbps	chol	fbs	restecg	thalach	exang	oldpeak	slope	ca	thal	target
52	1	0	125	212	0	1	168	0	1	2	2	3	0
53	1	0	140	203	1	0	155	1	3.1	0	0	3	0
70	1	0	145	174	0	1	125	1	2.6	0	0	3	0
61	1	0	148	203	0	1	161	0	0	2	1	3	0
62	0	0	138	294	1	1	106	0	1.9	1	3	2	0
58	0	0	100	248	0	0	122	0	1	1	0	2	1
58	1	0	114	318	0	2	140	0	4.4	0	3	1	0
55	1	0	160	289	0	0	145	1	0.8	1	1	3	0
46	1	0	120	249	0	0	144	0	0.8	2	0	3	0
54	1	0	122	286	0	0	116	1	3.2	1	2	2	0
71	0	0	112	149	0	1	125	0	1.6	1	0	2	1
43	0	0	132	341	1	0	136	1	3	1	0	3	0
34	0	1	118	210	0	1	192	0	0.7	2	0	2	1
51	1	0	140	298	0	1	122	1	4.2	1	3	3	0
52	1	0	128	204	1	1	156	1	1	1	0	0	0
34	0	1	118	210	0	1	192	0	0.7	2	0	2	1
51	0	2	140	308	0	0	142	0	1.5	2	1	2	1
54	1	0	124	266	0	0	109	1	2.2	1	1	3	0
50	0	1	120	244	0	1	162	0	1.1	2	0	2	1
58	1	2	140	211	1	0	165	0	0	2	0	2	1
60	1	2	140	185	0	0	155	0	3	1	0	2	0

Fig. 5 Dataset for heart diseases

6. Conclusion & Future Work

The study used Support Vector Machines, Logistic Regression, Gaussian Naive Bayes and Decision Tree classification models to assess the effectiveness of three supervised data mining approaches in forecasting the likelihood of a patient developing heart disease. All algorithms have been examined. Using the same dataset, we want to determine which strategy yielded the most precise outcomes.

The conclusion section might provide more details on the significance of the research or propose potential applications and future developments. The Logistic Regression model achieved an accuracy of 82%, the Gaussian Naive Bayes model achieved an accuracy of 77%, the Support Vector Machine model had an accuracy of 81%, and the Decision Tree classifier achieved the highest accuracy of 89% in predicting heart disease patients.

In the future, modern technology and machine learning algorithms may be used to predict or identify a diverse array of ailments. Furthermore, the automation of research on cardiac diseases might be improved or broadened by using more sophisticated machine learning techniques. The advanced technology and ML algorithm might be used to forecast or detect a wide range of different illnesses. Furthermore, the automation of research on cardiac diseases might be enhanced or expanded by using other machine-learning methodologies.

References

[1] Viren Viraj Shankar et al., "Heart Disease Prediction Using CNN Algorithm," *SN Computer Science*, vol. 1, 2020. [CrossRef] [Google Scholar] [Publisher Link]

[2] Sarthak Vinayaka, and P.K. Gupta, "Heart Disease Prediction System Using Classification Algorithms," *Advances in Computing and Data Sciences*, pp. 395-404, 2020. [CrossRef] [Google Scholar] [Publisher Link]

[3] Fatma Zahra Abdeldjouad, Menaouer Brahami, and Nada Matta, "A Hybrid Approach for Heart Disease Diagnosis and Prediction Using Machine Learning Techniques," *The Impact of Digital Technologies on Public Health in Developed and Developing Countries*, pp. 299-306, 2020. [CrossRef] [Google Scholar] [Publisher Link]

[4] Archana Singh, and Rakesh Kumar, "Heart Disease Prediction Using Machine Learning Algorithms," *2020 International Conference on Electrical and Electronics Engineering (ICE3)*, Gorakhpur, India, pp. 452-457, 2020. [CrossRef] [Google Scholar] [Publisher Link]

[5] Muhammad Affan Alim et al., "Robust Heart Disease Prediction: A Novel Approach Based on Significant Feature and Ensemble Learning Model," *2020 3rd International Conference on Computing, Mathematics and Engineering Technologies (iCoMET)*, Sukkur, Pakistan, pp. 1-5, 2020. [CrossRef] [Google Scholar] [Publisher Link]

[6] Xu Wenxin et al., "Heart Disease Prediction Model Based on Model Ensemble," 2020 3rd International Conference on Artificial Intelligence and Big Data (ICAIBD), Chengdu, China, pp. 195-199, 2020. [CrossRef] [Google Scholar] [Publisher Link]

[7] Mamatha Alex P., and Shaicy P. Shaji, "Prediction and Diagnosis of Heart Disease Patients Using Data Mining Technique," *2019 International Conference on Communication and Signal Processing (ICCSP)*, Chennai, India, pp. 0848-0852, 2019. [CrossRef] [Google Scholar] [Publisher Link]

[8] Mohini Chakarverti, Saumya Yadav, and Rajiv Rajan, "Classification Technique for Heart Disease Prediction in Data Mining," *2019 2nd International Conference on Intelligent Computing, Instrumentation and Control Technologies (ICICICT)*, Kannur, India, pp. 1578-1582, 2019. [CrossRef] [Google Scholar] [Publisher Link]

[9] Abhishek Kumar et al., "Comparative Analysis of Data Mining Techniques to Predict Heart Disease for Diabetic Patients," *Advances in Computing and Data Sciences*, pp. 507-518, 2020. [CrossRef] [Google Scholar] [Publisher Link]

[10] M. Anbarasi, E. Anupriya, and N. Iyengar, "Enhanced Prediction of Heart Disease with Feature Subset Selection Using Genetic Algorithm," *International Journal of Engineering Science and Technology*, vol. 2, no. 10, pp. 5370-5376, 2010. [Google Scholar]

[11] Navya Harika, Sita Rama Swamy, and Nilima, "Artificial Intelligence-Based Ensemble Model for Rapid Prediction of Heart Disease," *SN Computer Science*, vol. 2, 2021. [CrossRef] [Google Scholar] [Publisher Link]

[12] Sujata Joshi, and Mydhili K. Nair, "A Risk Assessment Model for Patients Suffering from Coronary Heart Disease Using a Novel Feature Selection Algorithm and Learning Classifiers," *Advances in Artificial Intelligence and Data Engineering*, pp. 237-249, 2020. [CrossRef] [Google Scholar] [Publisher Link]

[13] V. Jothi Prakash, and N.K. Karthikeyan, "Enhanced Evolutionary Feature Selection and Ensemble Method for Cardiovascular Disease Prediction," *Interdisciplinary Sciences: Computational Life Sciences*, vol. 13, pp. 389-412, 2021. [CrossRef] [Google Scholar] [Publisher Link]

[14] Eka Miranda et al., "Detection of Cardiovascular Disease Risk's Level for Adults Using Naive Bayes Classifier," *Healthcare Informatics Research*, vol. 22, no. 3, pp. 196-205, 2016. [CrossRef] [Google Scholar] [Publisher Link]

[15] Abien Fred Agarap, "Deep Learning Using Rectified Linear Units (ReLU)," *arXiv Preprint*, 2019. [CrossRef] [Google Scholar] [Publisher Link]

[16] Vincy Cherian, and Bindu M.S., "Heart Disease Prediction Using Naïve Bayes Algorithm and Laplace Smoothing Technique," *International Journal of Computer Science aprTrends and Technology*, vol. 5, no. 2, pp. 68-73, 2017. [Google Scholar] [Publisher Link]

[17] Uma N. Dulhare, "Prediction System for Heart Disease Using Naive Bayes and Particle Swarm Optimization," *Biomedical Research*, vol. 29, no. 12, pp. 2646-2649, 2018. [CrossRef] [Google Scholar] [Publisher Link]

[18] I. Ketut Agung Enriko, Muhammad Suryanegara, and Dadang Gunawan, "Heart Disease Prediction System Using k-Nearest Neighbor Algorithm with Simplified Patient's Health Parameters," *Journal of Telecommunication, Electronic and Computer Engineering*, vol. 8, no. 12, pp. 59-65, 2016. [Google Scholar] [Publisher Link]