Journal of Artificial Intelligence and Engineering Applications

Website: https://ioinformatic.org/

15th June 2025. Vol. 4. No. 3; e-ISSN: 2808-4519

Optimization of Supervised Learning Algorithms for Early Prediction of Heart Attack Risk

Angge Firizkiansah[1]*, Imron Rizki Maulana[2], Ali Muhmmad[3], Aliyah Kurniasih[4]

[1,2,3]*Universitas Sains Indonesia*
[4]*Universitas Ary Ginanjar*
angge.firizkiansah@lecturer.sains.ac.id[1*]

Abstract

Cardiovascular disease, particularly heart attacks, persists as a primary global cause of mortality. Heart attacks arise from an abrupt obstruction of oxygenated blood flow to a segment of the cardiac muscle, resulting in inadequate oxygen supply to the heart. This obstruction may stem from modifiable risk factors, including suboptimal dietary habits, physical inactivity, obesity, and tobacco consumption, alongside non-modifiable factors such as age, sex, and familial predisposition. Contemporary research increasingly focuses on preemptive strategies against heart attacks to mitigate associated mortality rates. One such strategy involves the application of artificial intelligence for predictive modeling of heart attack risk. These models may utilize machine learning algorithms, such as logistic regression, support vector machine, k-nearest neighbors, and random forest, all categorized under supervised learning paradigms. This study undertakes a thorough examination and optimization of diverse supervised learning algorithms for the prospective prediction of heart attack risk. Findings suggest that machine learning algorithms possess utility in predicting heart attack risk, with the random forest model demonstrating a peak accuracy of 64%. Nevertheless, the model's efficacy is constrained by high feature dimensionality, suggesting avenues for refinement via feature dimension reduction techniques and meticulous hyperparameter optimization across the employed machine learning algorithms.

Keywords: *Early Prediction; Heart Attack Risk; Optimization; Random Forest, Supervised Learning*

1. Introduction

Cardiovascular diseases, particularly heart attacks, remain a leading cause of mortality worldwide. According to WHO data, total heart attacks account for approximately 17.9 million deaths annually, representing nearly one-third of all global mortalities [1]. Heart attack occurs when the flow of oxygen-rich blood to a section of heart muscle suddenly becomes blocked and the heart can't get enough oxygen [2]. Some causes of blockage of blood flow containing oxygen to the heart include modifiable risk factors such as unhealthy diet, physical inactivity, obesity, and tobacco use, as well as non-modifiable factors like age, sex, and family history [3]. To prevent or minimize heart attack, traditional approach such as changing lifestyle and pharmacological interventions can be employed to mitigate risk factors and manage the progression of the disease; therefore, early and accurate risk assessment is critical for timely intervention and improved patient outcomes [4]. In addition, health checks using electrocardiograms, echocardiography, and cardiac CT scans are commonly deployed, providing detailed insights into cardiac electrophysiology, myocardial structure and function, and coronary artery anatomy, respectively, thereby facilitating more informed clinical decision-making [5]. But this approach costs a lot in terms of money and may not be easily accessible to all individuals, especially in underserved communities with limited access to advanced medical facilities.

Nowadays, more researchers are studying early preventive measures against the risk of heart attack to reduce the rate of heart attack fatalities [6][7]. The ability to accurately forecast the likelihood of a heart attack can significantly improve patient outcomes through timely lifestyle modifications, medical treatments, and preventive measures [8]. The early and accurate prediction of heart disease is paramount, as it remains a leading cause of mortality globally, emphasizing the critical need for advanced diagnostic strategies in cardiology [9][10]. To address the risk of heart attacks early, the latest developments in information technology greatly support its application. In particular, the development of artificial intelligence in supporting the application of technology in the health sector, such as diagnosing and treating cardiovascular diseases, showcases the potential of advanced computational techniques in improving healthcare outcomes [11]. The application of artificial intelligence models to support early detection of a disease, one of which is using machine learning. Machine learning, a subset of artificial intelligence, offers promising avenues for predicting individual heart attack risk by leveraging large datasets of clinical and demographic information [12]. Supervised learning algorithms, trained on labeled datasets of individuals with and without a history of heart attacks, can identify complex patterns and relationships between risk factors and disease occurrence [13].

Supervised learning algorithms constitute a pivotal class of machine learning techniques that leverage labeled datasets to construct predictive models [14]. These algorithms, including logistic regression, support vector machines, decision trees, and neural networks, are extensively employed in various domains, including medical diagnostics, fraud detection, and financial forecasting [15]. The effectiveness of supervised learning hinges on their capacity to generalize from training data to accurately classify or predict outcomes for novel, unseen instances, thereby enabling proactive interventions and informed decision-making. However, the performance of these algorithms is highly dependent on various factors, including the quality and representativeness of the training data, the selection of relevant features, and the appropriate tuning of model hyperparameters. The optimization of supervised learning algorithms is essential for enhancing their predictive accuracy and clinical utility in forecasting heart attack risk [14].

This study aims to comprehensively investigate and optimize various supervised learning algorithms for early prediction of heart attack risk. By systematically evaluating and refining model parameters, feature selection techniques, and ensemble methods, the research endeavors to develop a high-performing predictive model that can accurately identify individuals at increased risk of experiencing a heart attack. In this research, the supervised learning algorithm will use Support Vector machine, Logistic Regression, K-Nearest Neighbor, and Random Forest Algorithm. The outcome of this study is to prove that supervised learning algorithms can be used to predict heart attack risk.

Previous studies have demonstrated the efficacy of supervised learning algorithms in various applications. Logistic Regression has achieved accuracy levels of 64% [5], 80% [7], and 93% [9] in predicting heart disease risk, as well as 72.5% in classifying text data related to personal mental health comments [16]. The Support Vector Machine algorithm has also been applied to predict heart disease risk, attaining accuracy levels of 64% [5], 81% [6], 80.3% [7], and 92% [9]. Furthermore, the K-Nearest Neighbor algorithm has been employed in classifying heart disease risk data, with accuracy levels of 56% [5], 80% [6], 87% [9], and has also demonstrated an accuracy of 87% in classifying online learning review text data [17]. Additionally, the Random Forest algorithm has been utilized in predicting heart disease risk yielding accuracies of 64% [5], 89% [6], 77% [7], and 87% [9].

2. Research Methods

The research is built upon the meticulous curation and preparation of a comprehensive dataset, which serves as the foundation for training and evaluating supervised learning models. To achieve the expected research outcomes, the study follows a series of steps, including data selection, data cleaning, exploratory data analysis, machine learning modeling, and model evaluation. These stages are illustrated in Figure 1.

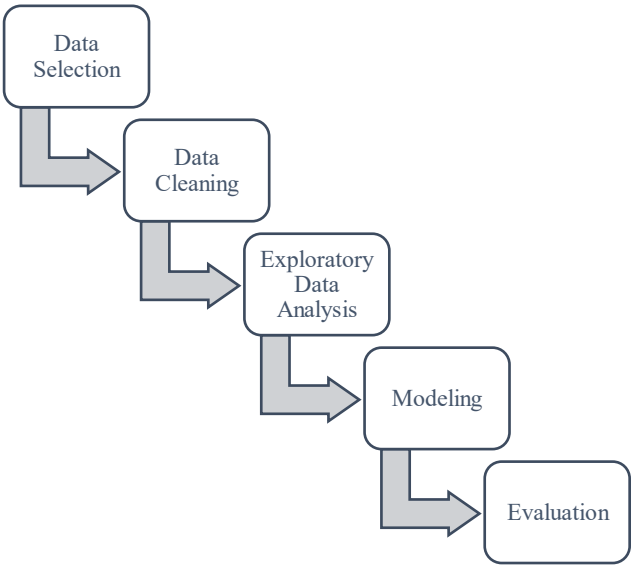

Fig. 1: Research methods

The first step is data selection. To ensure the reliability and generalizability of the findings, the dataset is taken from a trusted repository source kaggle.com. The dataset used in this study is containing demographic features (age, gender), lifestyle factors (smoking, alcohol, diet, exercise), medical conditions (diabetes, obesity, blood sugar), clinical tests (cholesterol, blood pressure, CK-MB, troponin), and target class heart attack risk (binary, 0 = low, 1 = high).

The next step is data cleaning. Prior to the application of machine learning models, a rigorous data cleaning process is undertaken to rectify inconsistencies, address missing values, and mitigate the impact of outliers, thereby enhancing the integrity and reliability of the dataset [18]. Specifically, the Pandas library in Python facilitates efficient data manipulation, enabling the implementation of appropriate imputation strategies such as mean, median, or mode imputation based on the distribution of each feature, while NumPy offers functionalities for identifying and handling outliers through statistical methods. Following data cleaning, Exploratory Data Analysis is conducted to extract meaningful insights and patterns from the preprocessed data, which encompasses employing statistical techniques, generating visualizations, and formulating hypotheses to guide subsequent modeling endeavors [19].

The next step is machine learning model development. For the machine learning model development phase, the curated dataset is partitioned into training and testing subsets, typically using an 80:20 split, which allows the supervised learning models, like logistic regression, support vector machines, k-nearest neighbor, and random forests, to be trained using the training data. Hyperparameter tuning is conducted

using grid search or randomized search approaches, with the goal of determining the best settings for each model through cross-validation on the training data [20].
The model evaluation phase involves assessing the performance of the trained models on the reserved testing dataset, where various evaluation metrics, including accuracy, precision, recall, and F1-score are rigorously quantified to provide a comprehensive assessment of each model's predictive capabilities [5].

3. Result and Discussion

The research stages begin with data selection. The dataset used in this study is a heart attack risk dataset obtained from kaggle.com. The dataset comprises 27 features, including age, gender, cholesterol, heart rate, diabetes, family history, smoking, obesity, alcohol consumption, exercise hours per week, diet, previous heart problems, medication use, stress level, sedentary hours per day, income, BMI, triglycerides, physical activity days per week, sleep hours per day, heart attack risk (binary), blood sugar, CK-MB, troponin, heart attack risk (text), systolic blood pressure, diastolic blood pressure. The data types of these features are explained in Table 1.

Table 1: Features data type

No	Features	Data Type	Description
1	Age	float	Age of the patient
2	Gender	float	0 = Female, 1 = Male
3	Cholesterol	float	Normalized cholesterol level
4	Heart rate	float	Normalized resting heart rate
5	Diabetes	float	Whether the patient has diabetes (0 = No, 1 = Yes)
6	Family History	float	Heart Disease in the family of patient
7	Smoking	float	Whether the patient smokes (0 = No, 1 = Yes)
8	Obesity	float	Whether the patient is obese (0 = No, 1 = Yes)
9	Alcohol Consumption	float	Frequency of alcohol intake
10	Exercise Hours Per Week	integer	Number of hours spent exercising per week
11	Diet	float	Categorized diet habits
12	Previous Heart Problems	float	Whether the patient had prior heart issues
13	Medication Use	float	Whether the patient is on medication
14	Stress Level	float	Normalized stress level
15	Sedentary Hours Per Day	float	Daily sedentary time in hours
16	Income	float	Normalized income
17	BMI	float	Body Mass Index (Normalized)
18	Triglycerides	float	Normalized triglyceride level
19	Physical Activity Days Per Week	float	Number of days of physical activity
20	Sleep Hours Per Day	float	Daily sleep duration
21	Heart Attack Risk (Binary)	float	Binary value (0 = Low Risk, 1 = High Risk)
22	Blood sugar	float	Normalized blood sugar level
23	CK-MB	float	Creatine Kinase-MB enzyme level
24	Troponin	float	Troponin enzyme level
25	Heart Attack Risk (Text)	object	Binary value (0 = Low Risk, 1 = High Risk)
26	Systolic blood pressure	float	Normalized systolic BP
27	Diastolic blood pressure	float	Normalized diastolic BP

The dataset will serve as both training and testing data for machine learning models designed to predict heart attack risk. Following the acquisition of the dataset, data cleaning will be performed to eliminate errors that could introduce bias during the learning process. The initial assessment of the dataset revealed missing values in several features, necessitating data repair. In this study, missing values will be addressed using the imputation method, a technique that replaces absent entries with plausible estimates [21]. It is important to note that unexpected challenges can arise when algorithms are applied in environments dissimilar to those in which they were developed, particularly those with substantial amounts of missing data or a large number of variables [22]. After the imputation of the missing values in the dataset, the results obtained are presented in Table 2.

Table 2: Missing value status before and after imputation

No	Features	Imputation	
		Before	After
1	Diabetes	274	0
2	Family History	274	0
3	Smoking	274	0
4	Obesity	274	0
5	Alcohol Consumption	274	0
6	Previous Heart Problems	274	0

7	Medication Use	274	0
8	Stress Level	274	0
9	Physical Activity Days Per Week	274	0

Next, exploratory data analysis is performed by examining the correlation between features using a heatmap diagram. A heatmap diagram employs a chromatic representation to visualize the correlation matrix, with color gradients indicating the strength and direction of pairwise variable relationships [23]. In this study, the Seaborn library, a Python-based statistical data visualization tool, is used to construct a heatmap diagram, facilitating the generation of a heatmap that effectively reveals the nuanced relationships between variables within the dataset [24]. The results of data visualization using a heatmap diagram are shown in Figure 2.

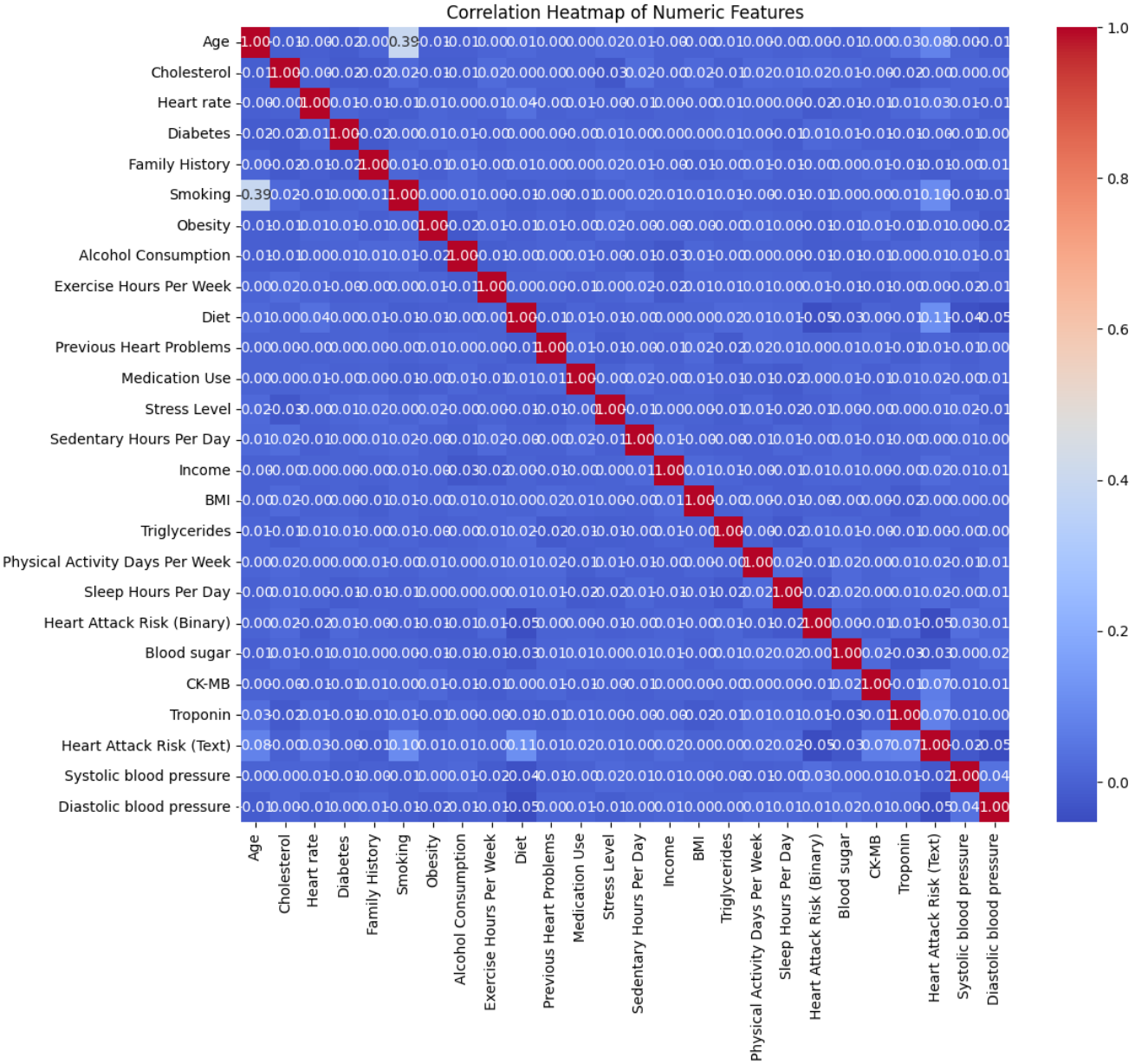

Fig. 2: Heatmap diagram features correlation

Based on the heatmap diagram, the correlation between features appears to be relatively uniform. However, upon closer inspection, the 'Heart Attack Risk' feature is repeated, indicating redundancy. To improve data validity, one of these redundant features will be removed. Additionally, the gender feature is not represented as a numeric type. Therefore, this feature will also be removed to enhance the validity of the data for machine learning modeling.

Next, machine learning models are developed using logistic regression, support vector machines, k-nearest neighbor, and random forest algorithms. The dataset is divided into training and testing sets with an 80:20 ratio. Following the modeling phase, a confusion matrix is used to evaluate the performance of each algorithm, determining accuracy, precision, recall, and F1-score. Accuracy assesses the proportion of correct classifications out of all predictions [16]. Precision measures the proportion of true positive predictions out of all positive predictions [16]. Recall measures the proportion of true positive predictions out of all actual positive instances [16]. Finally, the F1-score provides a balanced measure of precision and recall, offering a comprehensive evaluation of the model's performance [16]. The evaluation results for each machine learning model are presented in Table 3.

Table 3: Model evaluation

No	Algoritma	Label	Accuracy	Precision	Recall	F1-score
1	Logistic regression	0,0	66,6%	67%	100%	80%
		1,0		0%	0%	0%
2	Support vector machine	0,0	66,6%	67%	100%	80%
		1,0		0%	0%	0%
3	K-nearest neighbor	0,0	61%	68%	77%	73%
		1,0		39%	29%	33%
4	Random forest	0,0	67,9%	68%	98%	80%
		1,0		67%	7%	13%

Based on the evaluation results in Table 3, the machine learning algorithms demonstrate an average accuracy of 65% in predicting heart disease risk. This level of accuracy suggests that the predictive model is not yet optimal. The precision, recall, and F1-score values indicate a bias towards target label 0.0, implying that the data may still contain imbalances that reduce the accuracy of the machine learning algorithm. Notably, there is an imbalance in the class distribution, with 6320 data points for target class 0.0 and 3331 data points for target class 1.0. To address this, the Synthetic Minority Oversampling Technique is employed to balance the classes. SMOTE eliminates class imbalance by generating synthetic data points based on the minority class, thereby equalizing the amount of data between classes [16]. The evaluation results after applying SMOTE to handle the imbalance classes are presented in Table 4.

Table 4: Model evaluation+SMOTE

No	Algoritma	Label	Accuracy	Precision	Recall	F1-score
1	Logistic regression	0,0	51%	67%	51%	58%
		1,0		34%	50%	41%
2	Support vector machine	0,0	47%	68%	39%	50%
		1,0		34%	63%	44%
3	K-nearest neighbor	0,0	52%	70%	49%	58%
		1,0		36%	57%	44%
4	Random forest	0,0	64%	68%	85%	76%
		1,0		41%	21%	28%

Based on the evaluation results in Table 4, it is evident that employing SMOTE enhances the balance for machine learning algorithms in learning the dataset for prediction. However, when considering the overall performance, the random forest algorithm achieved the highest accuracy, at 64%. This suggests that while the machine learning algorithms can make predictions, they may not represent the optimal model for predicting the risk of heart attacks.

4. Conclusion

Based on the research results, it can be concluded that machine learning algorithms can be utilized for early detection of heart attack risk. The random forest algorithm achieved the highest accuracy value of 64%. However, this value does not represent the best possible machine learning model for predicting heart attack risk. This is partly due to the high feature dimension of the dataset, suggesting that further exploration, such as dimension reduction, is necessary. Additionally, further investigation into hyperparameter tuning is needed to optimize the machine learning algorithms and identify the best predictive model.

References

[1] F. Mohammad and S. Al-Ahmadi, "WT-CNN: A Hybrid Machine Learning Model for Heart Disease Prediction," *Mathematics*, vol. 11, no. 22, p. 4681, Nov. 2023, doi: 10.3390/math11224681.

[2] S. Rafi *et al.*, "Out-of-Hospital Cardiac Arrest Detection by Machine Learning Based on the Phonetic Characteristics of the Caller's Voice," in *Studies in Health Technology and Informatics*, vol. 294, 2022, pp. 445–449. doi: 10.3233/SHTI220498.

[3] R. Spencer, F. Thabtah, N. Abdelhamid, and M. Thompson, "Exploring feature selection and classification methods for predicting heart disease," *Digit. Heal.*, vol. 6, pp. 1–10, Jan. 2020, doi: 10.1177/2055207620914777.

[4] Z. S. Chunawala *et al.*, "Mortality in Patients Hospitalized With Acute Myocardial Infarction Without Standard Modifiable Risk Factors: The ARIC Study Community Surveillance," *J. Am. Heart Assoc.*, vol. 12, no. 13, Jul. 2023, doi: 10.1161/JAHA.122.027851.

[5] I. Rojek, P. Kotlarz, M. Kozielski, M. Jagodziński, and Z. Królikowski, "Development of AI-Based Prediction of Heart Attack Risk as an Element of Preventive Medicine," *Electronics*, vol. 13, no. 2, p. 272, Jan. 2024, doi: 10.3390/electronics13020272.

[6] M. M. Islam, T. N. Tania, S. Akter, and K. H. Shakib, "An Improved Heart Disease Prediction Using Stacked Ensemble Method," in *Lecture Notes of the Institute for Computer Sciences, Social-Informatics and Telecommunications Engineering, LNICST*, vol. 490 LNICST, 2023, pp. 84–97. doi: 10.1007/978-3-031-34619-4_8.

[7] S. Hussain, S. K. Nanda, S. Barigidad, S. Akhtar, M. Suaib, and N. K. Ray, "Novel Deep Learning Architecture for Predicting Heart Disease using CNN," in *2021 19th OITS International Conference on Information Technology (OCIT)*, IEEE, Dec. 2021, pp. 353–357. doi: 10.1109/OCIT53463.2021.00076.

[8] P. Chavda, H. Bhavsar, Y. Pithadia, and R. Kotecha, "Early Detection of Cardiac Disease Using Machine Learning," *SSRN Electron. J.*, 2019, doi: 10.2139/ssrn.3370813.

[9] N. A. M. Zaini and M. K. Awang, "Hybrid Feature Selection Algorithm and Ensemble Stacking for Heart Disease Prediction," *Int. J. Adv. Comput. Sci. Appl.*, vol. 14, no. 2, pp. 158–165, 2023, doi: 10.14569/IJACSA.2023.0140220.

[10] O. S. Hasan and I. A. Saleh, "Development of heart attack prediction model based on ensemble learning," *Eastern-European J. Enterp. Technol.*, vol. 4, no. 2(112), pp. 26–34, Aug. 2021, doi: 10.15587/1729-4061.2021.238528.

[11] N. E. Almansouri *et al.*, "Early Diagnosis of Cardiovascular Diseases in the Era of Artificial Intelligence: An In-Depth Review," *Cureus*, vol. 16, no. 3, pp. 1–18, Mar. 2024, doi: 10.7759/cureus.55869.

[12] P. Iacobescu, V. Marina, C. Anghel, and A.-D. Anghele, "Evaluating Binary Classifiers for Cardiovascular Disease Prediction: Enhancing Early Diagnostic Capabilities," *J. Cardiovasc. Dev. Dis.*, vol. 11, no. 12, p. 396, Dec. 2024, doi: 10.3390/jcdd11120396.

[13] M. S. Pathan, A. Nag, M. M. Pathan, and S. Dev, "Analyzing the impact of feature selection on the accuracy of heart disease prediction," *Healthc. Anal.*, vol. 2, p. 100060, Nov. 2022, doi: 10.1016/j.health.2022.100060.

[14] A. Ogunpola, F. Saeed, S. Basurra, A. M. Albarrak, and S. N. Qasem, "Machine Learning-Based Predictive Models for Detection of Cardiovascular Diseases," *Diagnostics*, vol. 14, no. 2, p. 144, Jan. 2024, doi: 10.3390/diagnostics14020144.

[15] Y. Yazici, "Approaches to Fraud Detection on Credit Card Transactions using Artificial Intelligence Methods," in *Computer Science & Information Technology*, AIRCC Publishing Corporation, Jul. 2020, pp. 235–244. doi: 10.5121/csit.2020.101018.

[16] A. Firizkiansah, A. Muhammad, and I. R. Maulana, "Optimasi Klasifikasi Data Teks Menggunakan Algoritma Logistic Regression dengan TF-IDF dan SMOTE," *JIKOMTI J. Ilm. Ilmu Komput. dan Teknol. Inf.*, vol. 2, no. 1, pp. 29–36, 2025, [Online]. Available: https://ojs.sains.ac.id/index.php/Jikomti/article/view/97/119

[17] A. Firizkiansah, A. Muhammad, and D. Setiawan, "Implementasi Algoritma k-Nearest Neighbor (k-NN) pada Data Ulasan Pelaksanaan Pembelajaran Daring," *JIKOMTI J. Ilm. Ilmu Komput. dan Teknol. Inf.*, vol. 1, no. 1, pp. 16–23, Dec. 2024, Accessed: Mar. 20, 2025. [Online]. Available: https://ojs.sains.ac.id/index.php/Jikomti/article/view/35/35

[18] J. Nourmohammadi-Khiarak, M.-R. Feizi-Derakhshi, K. Behrouzi, S. Mazaheri, Y. Zamani-Harghalani, and R. M. Tayebi, "New hybrid method for heart disease diagnosis utilizing optimization algorithm in feature selection," *Health Technol. (Berl).*, vol. 10, no. 3, pp. 667–678, May 2020, doi: 10.1007/s12553-019-00396-3.

[19] S. Nashif, M. R. Raihan, M. R. Islam, and M. H. Imam, "Heart Disease Detection by Using Machine Learning Algorithms and a Real-Time Cardiovascular Health Monitoring System," *World J. Eng. Technol.*, vol. 06, no. 04, pp. 854–873, 2018, doi: 10.4236/wjet.2018.64057.

[20] I. Abousaber, H. F. Abdallah, and H. El-Ghaish, "Robust predictive framework for diabetes classification using optimized machine learning on imbalanced datasets," *Front. Artif. Intell.*, vol. 7, Jan. 2025, doi: 10.3389/frai.2024.1499530.

[21] S. Harford *et al.*, "A machine learning based model for Out of Hospital cardiac arrest outcome classification and sensitivity analysis," *Resuscitation*, vol. 138, no. March, pp. 134–140, May 2019, doi: 10.1016/j.resuscitation.2019.03.012.

[22] J. A. C. Sterne *et al.*, "Multiple imputation for missing data in epidemiological and clinical research: potential and pitfalls," *BMJ*, vol. 338, no. jun29 1, pp. b2393–b2393, Sep. 2009, doi: 10.1136/bmj.b2393.

[23] D. Toddenroth, T. Ganslandt, I. Castellanos, H.-U. Prokosch, and T. Bürkle, "Employing heat maps to mine associations in structured routine care data," *Artif. Intell. Med.*, vol. 60, no. 2, pp. 79–88, Feb. 2014, doi: 10.1016/j.artmed.2013.12.003.

[24] C. Li, "Preprocessing Methods and Pipelines of Data Mining: An Overview," no. June, pp. 1–7, Jun. 2019, [Online]. Available: http://arxiv.org/abs/1906.08510