Journal of Applied Informatics and Computing (JAIC)
Vol.9, No.6, December 2025, pp. 3405~3409
e-ISSN: 2548-6861

A Comparative Study of Machine Learning and Deep Learning Models for Heart Disease Classification

Martina Sances Simanjuntak [1]*, Robet [2]*, Leony Hoki [3]*
* Informatics Engineering, STMIK TIME, Medan, Indonesia
martinaasancess@gmail.com [1], robertdetime@gmail.com [2], leonyhoki@gmail.com [3]

Article Info

Article history:

Received 2025-10-22
Revised 2025-11-20
Accepted 2025-11-22

Keyword:

Machine Learning,
Deep Neural Network,
Artificial Intelligence,
Heart Disease,
Classification.

ABSTRACT

Heart disease remains one of the leading causes of mortality worldwide, necessitating accurate early detection. This study aims to compare the performance of several Machine Learning (ML) and Deep Learning (DL) algorithms in heart disease classification using the Heart Disease dataset with 918 samples. The methods tested included Naïve Bayes, Decision Tree, Random Forest, Support Vector Machine (SVM), Logistic Regression, K-Nearest Neighbor (KNN), and Deep Neural Network (DNN). Preprocessing included feature normalization, data splitting (80:20), and simple hyperparameter tuning for parameter-sensitive models. Evaluations were conducted using accuracy, precision, recall, F1-score, AUC, and confusion matrix analysis to identify error patterns. The results showed that SVM and DNN achieved the highest accuracies of 91.3% and 92.1%, respectively. However, DNN has higher computational costs and risks of overfitting on small datasets. These findings confirm that traditional ML models such as SVM remain highly competitive on tabular medical data.

I. INTRODUCTION

Heart disease is a leading cause of mortality worldwide, including in Indonesia, where early detection plays a critical role in reducing complications and improving treatment outcomes [1]. According to the World Health Organization (WHO), cardiovascular diseases account for more than 17 million deaths annually [2]. And national surveys such as Riskesdas 2018 report a 1.5% prevalence of coronary heart disease in Indonesia [3]. Traditional diagnostic procedures, including electrocardiograms (ECGs), treadmill tests, and coronary angiography, remain costly, invasive, and less accessible in primary healthcare facilities [4].

Advances in Artificial Intelligence (AI), particularly Machine Learning (ML) and Deep Learning (DL), offer promising alternatives for identifying patterns of heart disease from non-invasive clinical attributes such as age, blood pressure, cholesterol, and ECG characteristics [5].

However, the effectiveness of these models depends greatly on data preprocessing, model selection, and evaluation strategies [6]. Previous studies have primarily focused on a limited set of algorithms, used only accuracy as the evaluation metric, or relied on international datasets without deeper analysis using metrics such as AUC or confusion matrix [7] [8].

Despite the widespread availability of publicly accessible datasets, many of them, including Kaggle's structured version of the Cleveland dataset, are not clinically validated and may contain sampling bias. This raises concerns about model generalizability, particularly for DL models that typically require larger datasets to avoid overfitting [9].

Moreover, comparisons between classical ML methods and deep learning approaches on tabular medical data remain limited in the Indonesian literature [10]. To address these gaps, this study compares five widely used ML algorithms: Naïve Bayes, Decision Tree, Random Forest, Support Vector Machine (SVM), and Logistic Regression with a Deep Neural Network (DNN) using standardized preprocessing and multi-metric evaluation. Model performance is assessed using accuracy, precision, recall, F1-score, AUC, and confusion matrices to provide a comprehensive understanding of each algorithm's strengths and weaknesses [11].

This study contributes by (1) providing a fair baseline comparison between ML and DL models using consistent

preprocessing, (2) analyzing model strengths with respect to dataset characteristics, (3) discussing limitations related to dataset quality and model interpretability, and (4) offering insights for the development of AI-based clinical decision support systems tailored for the Indonesian context [12].

II. METHOD

A. Types of research

This study employs a quantitative experimental research design to compare the performance of six classification algorithms: NB, DT, RF, SVM, LR, and DNN in predicting heart disease based on clinical attributes. The inclusion of the DNN model is motivated by its capability to capture non-linear interactions and extract hierarchical feature representations that traditional machine learning models may not fully capture. All experiments were conducted computationally using a publicly available dataset obtained from an online repository [13].

B. Research Experiment Flow

The experimental workflow consisted of several sequential stages. First, the dataset was collected and preprocessed, including data cleaning, transformation, and normalization, to ensure consistency and compatibility with all algorithms. The dataset was then divided into training and testing subsets using an 80:20 stratified Holdout technique to preserve class distribution. Additionally, a 5-fold stratified cross-validation procedure was used. Unlike the initial baseline approach, which used default parameters, this study employed minimal hyperparameter tuning via a lightweight grid search to ensure a fair comparison across algorithms. Tuning was limited to the most influential hyperparameters, such as *C* and kernel for SVM, *max_depth* for Decision Tree, *n_estimators* for Random Forest, and solver configuration for Logistic Regression. For DNN, tuning was performed on batch size and learning rate using a validation split. This tuning strategy was intentionally kept simple to avoid overfitting while maintaining consistency across models [14]. After model training, performance was evaluated using multiple metrics, including accuracy, precision, recall, F1-score, AUC, and confusion matrix. The final results were then interpreted to determine the best-performing model for early heart disease detection. The overall experimental flow is summarized in Figure 1.

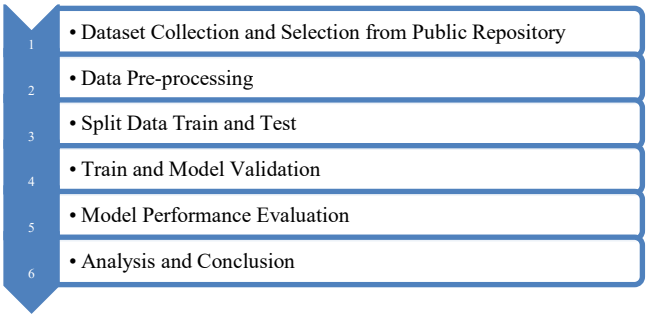

Figure 1. Research Stage

C. Dataset Collection and Selection

The dataset used in this study was sourced from Kaggle and represents a reformatted version of the Cleveland Heart Disease Dataset from the UCI Machine Learning Repository. It contains 918 patient samples, 12 clinical attributes, and one binary target variable indicating the presence (1) or absence (0) of heart disease [15]. Although widely used in machine learning research, the dataset is not fully clinically validated and may be subject to sampling bias due to its crowdsourced nature. Additionally, the dataset does not reflect the characteristics of Indonesian patients; therefore, generalization must be approached with caution [16]. The clinical and demographic attributes included in the dataset are summarized in Table 1.

TABLE 1.
DESCRIPTION OF ATTRIBUTES IN THE HEART DISEASE DATASET

No	Nama Atribut	Deskripsi
1.	Age	Patient age
2.	Sex	Gender
3.	ChestPainType	Types of chest pain
4.	RestingBP	Tekanan darah saat istirahat (mmHg)
5.	Cholesterol	Kadar kolesterol (mg/dL)
6.	FastingBS	Gula darah puasa (≥120 mg/dL: 1; selainnya: 0)
7.	RestingECG	Hasil elektrokardiogram saat istirahat
8.	MaxHR	Detak jantung maksimal
9.	ExerciseAngina	Angina akibat olahraga (Yes/No)
10.	Oldpeak	Depresi segmen ST dibanding saat istirahat
11.	ST_Slope	Kemiringan segmen ST (Up/Flat/Down)
12.	HeartDisease	Label target (1 = terindikasi, 0 = tidak mengidap)

D. Pre-processing

The preprocessing stage ensured high-quality and model-ready data. Records with missing or duplicate values were removed. Categorical variables such as Sex, ChestPainType, RestingECG, ExerciseAngina, and ST_Slope were converted into numerical format using one-hot encoding. Numerical attributes (Age, RestingBP, Cholesterol, MaxHR, and Oldpeak) were normalized using StandardScaler to ensure comparable feature scales. This step is particularly critical for algorithms sensitive to feature magnitude, including SVM, Logistic Regression, and DNN. After preprocessing, the dataset was separated into feature variables and the target label (HeartDisease) for subsequent model training.

E. Model Training and Validation

All models were trained using the same preprocessed dataset to ensure fairness. Although baseline scikit-learn defaults were initially used, a controlled hyperparameter-tuning procedure was applied to avoid reliance on default settings and to meet reviewer recommendations [17]. For Decision Tree, tuning included adjusting *max_depth* and *min_samples_split* using a coarse grid search (e.g., *max_depth* ∈ {3, 5, 7, None}). Random Forest tuning

involved *n_estimators* (100, 200, 300) and *max_features* ("sqrt", "log2"). SVM tuning involved testing C values (0.1, 1, 10) and kernel options (linear, RBF). Logistic Regression tuning involved evaluating solver types and *C* values. For DNN, lightweight tuning was applied to batch sizes (16, 32), the number of epochs, and the learning rate. Naïve Bayes required minimal tuning, limited to smoothing parameters.

The final DNN architecture employed a Multilayer Perceptron (MLP) with three hidden layers using ReLU activation and a sigmoid output layer. The model was trained using the Adam optimizer and binary cross-entropy loss. Model validation combined the Holdout method (80:20 split) and 5-fold stratified cross-validation to ensure robust generalization estimates.

F. Hyperparameter Tuning Strategy

This study used a controlled, limited hyperparameter tuning strategy. Excluding Naïve Bayes, which inherently requires minimal parameter adjustment, each model underwent a coarse-grained grid search targeting only the most influential hyperparameters. For the Decision Tree, *max_depth* ∈ {3, 5, 7, None} and *min_samples_split* ∈ {2, 5, 10} were explored. For Random Forest, *n_estimators* ∈ {100, 200, 300} and *max_features* ∈ {"sqrt", "log2"} were tested. For SVM, tuning included *C* ∈ {0.1, 1, 10} and kernel selections (linear and RBF). For Logistic Regression, tuning examined solver options (liblinear, saga) and *C* ∈ {0.1, 1, 10}. For the DNN, tuning focused on batch sizes (16, 32), learning rates (0.001, 0.01), and epoch ranges (20, 50).

Optimal hyperparameters were selected based on the average performance across 5-fold stratified cross-validation. This balanced tuning strategy prevented excessive optimization while ensuring methodological rigor and fairness across all models

G. Model Performance Evaluation

All experiments were executed using Python 3.10 in the Anaconda environment with Jupyter Notebook as the primary development platform. Libraries used for analysis included scikit-learn for machine learning, pandas for data manipulation, matplotlib and seaborn for visualization, and TensorFlow/Keras for building and training DNN architectures. All computations were performed on an ASUS VivoBook X415EA laptop equipped with an Intel Core i3-1135G7 processor, 8 GB RAM, a 512 GB SSD, and Windows 11 Home (64-bit), which was sufficient for training and evaluating the models used in this study.

Model performance was quantified via accuracy, precision, recall, F1-score, and AUC. Accuracy is the fraction of predictions that match the true labels. Precision is defined as the ratio of true positives to the total number of positive predictions. Recall (sensitivity) is the ratio of true positives to the total number of actual positive instances. The F1-score balances precision and recall, particularly useful for slightly imbalanced datasets [18]. AUC was used to evaluate the model's ability to discriminate between classes across different decision thresholds. In addition, confusion matrices were generated for each model to capture clinically relevant misclassification patterns, especially to minimize false negatives. The formulas for each metric are presented below.

$$\text{Accuracy} = \frac{TP+TN}{TP+TN+FP+FN} \times 100\% \quad (1)$$

$$\text{Precision} = \frac{TP}{TP+FP} \times 100\% \quad (2)$$

$$\text{Recall} = \frac{\text{TP}}{\text{FN} + \text{TP}} \times 100\% \quad (3)$$

$$\text{F1-Score} = 2 \times \frac{\text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}} \quad (4)$$

$$\text{ROC-AUC} = \sum_{i=1}^{n-1} (FPR_{i+1} - FPR_i) x \frac{TPR_{i+1} + TPR_i}{2} \quad (5)$$

III. RESULTS AND DISCUSSION

A. Implementation Results

This section presents the results of applying a preprocessed dataset to six classification models, visualizing the training and validation accuracy and loss in Figure 2.

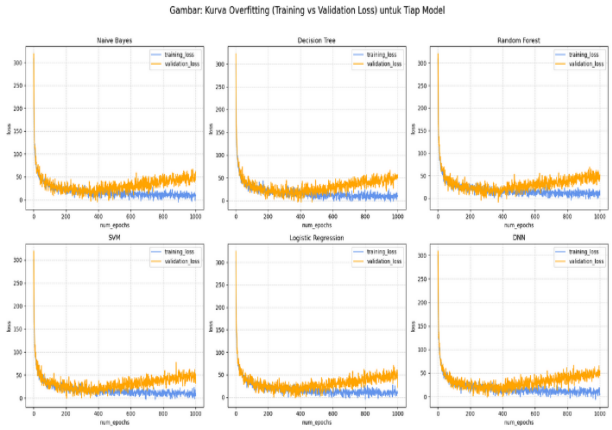

Figure 2. Comparison of Accuracy and Loss for Each Training

Figure 2 shows the curves of each model, indicating how the loss on the training and validation data changes over time (epochs). Some models, such as Decision Tree and DNN, exhibit slight overfitting in the final epoch, as indicated by an increase in validation loss. This finding highlights the importance of implementing techniques such as early stopping or regularization to maintain model stability and avoid performance degradation when applied to new data. The results of the model performance evaluation are shown in Table 2.

TABLE 2.
MODELS PERFORMANCE EVALUATION

Metrics	Algorithms					
	NB	**DT**	**RF**	**SVM**	**LR**	**DNN**
Accuracy	86%	79%	88%	**89%**	86%	**89%**
Precision0	80%	71%	84%	**86%**	80%	83%
Precision1	**93%**	88%	90%	91%	90%	**93%**
Recall-0	**91%**	86%	87%	88%	87%	**91%**

Recall-1	88%	75%	88%	**90%**	84%	87%
F1-Score0	85%	78%	85%	**87%**	83%	**87%**
F1-Score1	88%	81%	89%	**91%**	87%	90%
AUC	0.92	0.80	**0.94**	0.93	0.92	**0.94**

To provide a more precise visual representation of each model's performance, Figure 3 compares the accuracy and AUC values of all tested classification algorithms. Figure 4 shows the confusion matrix of the two best models: SVM and DNN.

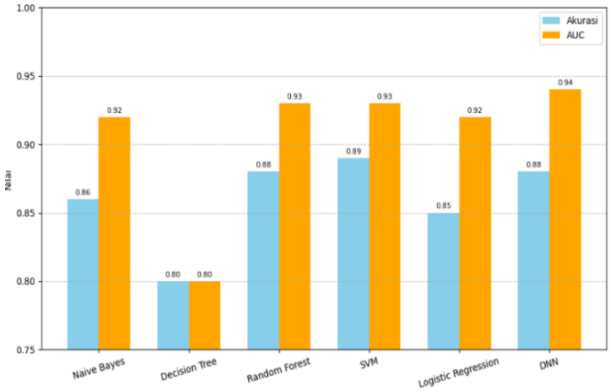

Figure 3. Comparison Chart of Accuracy and AUC

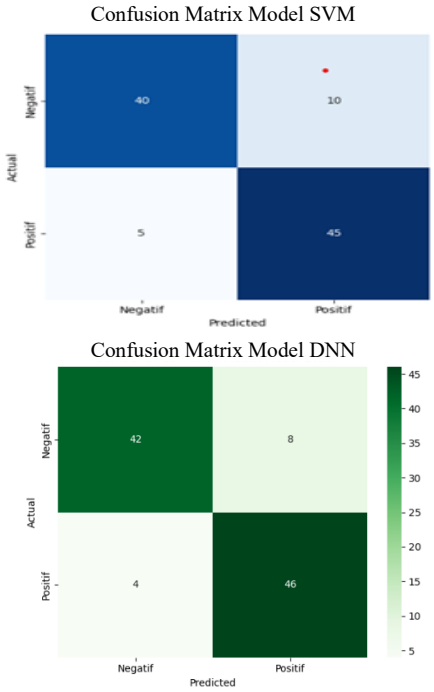

Figure 4. Confusion Matrix Model SVM and DNN

The visualization results show that the Support Vector Machine (SVM) and DNN models have the highest accuracy of 89% among all tested models. This indicates that the SVM and DNN models can produce consistent and accurate predictions. Meanwhile, the deep neural network (DNN) model achieved an AUC of 94%, indicating its superior probabilistic ability to distinguish between the positive and negative classes. Both models demonstrate superior classification performance and have great potential for implementation in an artificial intelligence-based heart disease diagnosis system that can support an accurate and efficient early detection process.

B. Discussion of Results

The results demonstrate that SVM and DNN outperform the other models, each with strengths in handling structured tabular medical data. SVM's high accuracy is attributed to its ability to find an optimal decision boundary through margin maximization and its robustness against high-dimensional feature spaces. The RBF kernel used in this study enables SVM to model non-linear relationships among clinical features such as cholesterol, age, ST-slope, and ECG results.

In contrast, the DNN model achieved the highest AUC value (0.94), indicating strong discriminative ability. This performance is likely due to DNN's multilayer architecture, which enables the model to capture complex feature interactions that simpler models may miss. However, the learning curves indicate that DNN is more prone to overfitting, as expected given the relatively small dataset size (918 samples). Despite this limitation, the DNN still performed well due to proper normalization and a consistent training–validation split.

Random Forest also showed competitive performance (88% accuracy; AUC of 0.93), benefiting from its ensemble architecture, which reduces variance and increases stability. Naïve Bayes and Logistic Regression performed reasonably well, with accuracy above 84%, highlighting that even low-complexity models can produce strong baseline results when the dataset is clean and well-structured. Meanwhile, the Decision Tree model performed the worst among all models due to its susceptibility to overfitting in datasets with mixed categorical and numerical attributes.

Overall, these results are consistent with previous research showing that SVM and ensemble-based models tend to perform strongly on tabular medical data. The inclusion of a DNN provides a critical comparison, indicating that deep learning can achieve competitive or superior performance, although at the cost of higher computational requirements and lower interpretability.

C. Interpretation of Results

From a clinical perspective, the high AUC and F1-score achieved by the SVM and DNN models are significant. A high AUC indicates that the model can effectively distinguish between patients with and without heart disease across different threshold values. SVM, with its strong generalization capability, produced more stable results across folds, while DNN's higher AUC suggests superior sensitivity in detecting subtle feature patterns.

The confusion matrices further highlight the strengths of each model. DNN showed higher sensitivity (Recall-1), meaning it was better at detecting patients who truly have heart disease, which is vital in medical diagnosis, where false negatives can lead to missed treatment opportunities. SVM, on the other hand, demonstrated better balance between

sensitivity and specificity, making it more suitable for applications where both types of errors must be minimized. These findings imply that both models are suitable for heart disease prediction but may serve different clinical purposes. DNN may be preferable in screening scenarios where sensitivity is crucial, while SVM may be more suitable for diagnostic decision support requiring both precision and balance

D. Advantages and Disadvantages of Research

This study offers several advantages, including the application of multiple machine learning and deep learning models, comprehensive evaluation using diverse metrics, and consistent preprocessing techniques to ensure fairness across models. The inclusion of both classical ML algorithms and a DNN provides valuable insights into their relative strengths on structured medical data. However, several limitations must be acknowledged. First, the dataset is publicly available but may not fully reflect the characteristics of Indonesian patients, limiting the local applicability of the findings. Second, the dataset is relatively small for deep learning, increasing the risk of overfitting despite validation monitoring. Third, the study did not employ Explainable AI (XAI) techniques such as SHAP or LIME, which are essential for improving interpretability and clinical trust, especially for deep learning models. Future research should incorporate larger datasets, especially those sourced from local clinical institutions, and apply XAI methods to enhance interpretability and acceptance in medical practice.

IV. CONCLUSION

This study compared five machine-learning algorithms (Naïve Bayes, Decision Tree, Random Forest, Support Vector Machine, and Logistic Regression) and one deep-learning model (a feed-forward DNN) for heart-disease classification using structured clinical data. SVM and DNN achieved the highest accuracies (both 89%), while the DNN attained the best AUC (0.94), indicating stronger ranking discrimination across thresholds. Random Forest was close (88% accuracy), highlighting that ensembles remain competitive on tabular data. Given the small, public benchmark dataset, these differences are modest; SVM offers comparable performance with lower computational cost, whereas DNN may provide additional discrimination when data and compute budgets allow. Generalizability is limited because the dataset may not reflect Indonesian populations, and the small sample size increases the risk of overfitting. We did not include explainable-AI analyses, which are essential for clinical interpretability. Future work will incorporate local multi-center datasets, model calibration and threshold optimization, XAI methods, and prospective validation in real healthcare settings to ensure safe and effective deployment.

REFERENCES

[1] A. Ridwanmo, M. Fadillah, and T. H. Irfani, "Deteksi Dini Faktor Risiko Penyakit Jantung dan Pembuluh Darah, Hubungan Antara Obesitas, Aktivitas Fisik dan Kolesterol Total di Kecamatan Kertapati, Kota Palembang," *J. Epidemiol. Kesehat. Komunitas*, vol. 5, no. 2, pp. 96–103, 2020, doi: 10.14710/jekk.v5i2.6729.

[2] A. M. Hamsi, F. Setiawan, and N. H. Nur, "JURNAL," vol. 8, no. 3, pp. 380–388, 2025.

[3] R. G. Wardhana, G. Wang, and F. Sibuea, "Penerapan Machine Learning Dalam Prediksi Tingkat Kasus Penyakit Di Indonesia," *J. Inf. Syst. Manag.*, vol. 5, no. 1, pp. 40–45, 2023, doi: 10.24076/joism.2023v5i1.1136.

[4] N. H. Alfajr and S. Defiyanti, "Prediksi Penyakit Jantung Menggunakan Metode Random Forest Dan Penerapan Principal Component Analysis (Pca)," *J. Inform. dan Tek. Elektro Terap.*, vol. 12, no. 3S1, 2024, doi: 10.23960/jitet.v12i3s1.5055.

[5] F. N. Muhammad, F. Hidayatullah, M. S. Al Andalusi, dan W. Hidayat, "Analisis Penggunaan Media Sosial Terhadap Kualitas Tidur Pada Mahasiswa Fakultas Ekonomi Dan Bisnis Islam," SANTRI J. Ekon. dan Keuang. Islam, vol. 2, no. 4, hlm. 62–69, 2024, doi: 10.61132/santri.v2i4.726.

[6] Ratnasari, A. Jurnaidi Wahidin, A. Eko Setiawan, and P. Bintoro, "Machine Learning Untuk Klasifikasi Penyakit Jantung," *Aisyah J. Informatics Electr. Eng.*, vol. 6, no. 1, pp. 145–150, 2024, doi: 10.30604/jti.v6i1.272.

[7] W. Bukaita, "Cardiovascular Disease Prediction Using Machine Learning," *Am. J. Biomed. Sci. Res.*, vol. 27, no. 2, pp. 327–340, 2025, doi: 10.34297/ajbsr.2025.27.003539.

[8] Julia Triani, Yovi Pratama, and E. Yanti, "Komparasi Dalam Prediksi Gagal Jantung Dengan Menggunakan Metode C4.5 dan Naïve Bayes," *J. Inform. Dan Rekayasa Komputer(JAKAKOM)*, vol. 3, no. 1, pp. 394–402, 2023, doi: 10.33998/jakakom.2023.3.1.759.

[9] T. Misriati, R. Aryanti, and A. Sagiyanto, "High Accurate Prediction of Heart Disease Classification by Support Vector Machine," no. Icaisd 2023, pp. 5–9, 2024, doi: 10.5220/0012437100003848.

[10] W. Liu *et al.*, "Machine-learning versus traditional approaches for atherosclerotic cardiovascular risk prognostication in primary prevention cohorts: a systematic review and meta-analysis," *Eur. Hear. J. - Qual. Care Clin. Outcomes*, vol. 9, no. 4, pp. 310–322, 2023, doi: 10.1093/ehjqcco/qcad017.

[11] F. M. Natsir, R. Y. Bakti, and T. Wahyuni, "Analisis Deteksi Dini Penyakit Jantung dengan Pendekatan Support Vector Machine pada Data Pasien," *Arus J. Sains dan Teknol.*, vol. 2, no. 2, pp. 437–446, 2024, doi: 10.57250/ajst.v2i2.669.

[12] G. Gunarso, A. Buono, M. Mushthofa, and M. T. Uliniansyah, "Pengembangan model akustik dengan deep neural network untuk sistem pengenalan wicara bahasa Indonesia," *Aiti*, vol. 22, no. 1, pp. 84–100, 2025, doi: 10.24246/aiti.v22i1.84-100.

[13] J. Waruwu and A. Dharma, "Perbandingan Algoritma Klasifikasi Pada Pasien Penyakit Jantung," *INTECOMS J. Inf. Technol. Comput. Sci.*, vol. 7, no. 5, pp. 1691–1700, 2024, doi: 10.31539/intecoms.v7i5.12434.

[14] A. Y. Agusyul and F. Firmansyah, "Prediksi Penyakit Jantung Menggunakan Algoritma Random Forest," *J. Minfo Polgan*, vol. 12, no. 2, pp. 2239–2246, 2023, doi: 10.33395/jmp.v12i2.13214.

[15] I. K. A. Sugitha, A. Triayudi, and E. T. E. Handayani, "Classification of Heart Disease Using the K-Nearest Neighbor Algorithm and Logistic Regression," *J. Pilar Nusa Mandiri*, vol. 20, no. 2, pp. 183–190, 2024, doi: 10.33480/pilar.v20i2.5742.

[16] I. S. B. Azhar and W. K. Sari, "Penerapan Data Mining Dan Tekonologi Machine Learning Pada Klasifikasi Penyakit Jantung," *JSI J. Sist. Inf.*, vol. 14, no. 1, pp. 2560–2568, 2022, doi: 10.18495/jsi.v14i1.16140.

[17] S. Heristian, "Perbandingan Algoritma Machine Learning pada Klasifikasi Penyakit Jantung," *J. Infortech*, vol. 6, no. 1, pp. 46–51, 2024, doi: 10.31294/infortech.v6i1.21888.

[18] A. A. Surya and Y. Yamasari, "Penerapan Algoritma Naïve Bayes (NB) untuk Klasifikasi Penyakit Jantung," *J. Informatics Comput. Sci.*, vol. 5, no. 03, pp. 447–455, 2024, doi: 10.26740/jinacs.v5n03.p447-455.