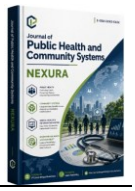
Content lists available at LiteraPublishing
Journal of Public Health and Community Systems
NEXURA
journal homepage : https://journal.literapublishing.id/index.php/nexura



Article

Heart Disease Prediction Using Support Vector Machine (SVM) Classification Based on Clinical Data

Leonita Yulyta Agustin [1*], Nur Alisa Qiroati Sholeha [2], Siti Aisa Nur Apriliana [3] and Niyalatul Muna[4]

[1] Study Program of Health Information Management, Department of Health, Politeknik Negeri Jember, Jember, Indonesia; leonitayulytaagustin@gmail.com
[2] Study Program of Health Information Management, Department of Health, Politeknik Negeri Jember, Jember, Indonesia; shonuralisa@gmail.com
[3] Study Program of Health Information Management, Department of Health, Politeknik Negeri Jember, Jember, Indonesia; sitiaisanurapriliana64@gmail.com
[4] Study Program of Health Information Management, Department of Health, Politeknik Negeri Jember, Jember, Indonesia; niyalatul@polije.ac.id

* Correspondence: leonitayulytaagustin@gmail.com

Abstract: Heart disease is the leading cause of death globally and requires an accurate early prediction system. This study aimed to develop a heart disease classification model using the Support Vector Machine (SVM) method with a Radial Basis Function (RBF) kernel based on the Heart Disease Dataset, which consists of 100 patient records and 13 clinical attributes. The research stages included data preprocessing, feature standardization using StandardScaler, data splitting with an 80:20 ratio, SVM model training, and evaluation using accuracy, precision, recall, F1-score, and confusion matrix metrics. The evaluation results showed that the SVM model with the RBF kernel achieved an accuracy of 85%, with a precision of 0.83 for the negative class and 1.00 for the positive class, recall of 1.00 for the negative class and 0.40 for the positive class, and F1-score of 0.91 for the negative class and 0.57 for the positive class. The confusion matrix produced TN=15, FP=0, FN=3, and TP=2. The low recall of the positive class indicates the model's limitation in detecting actual heart disease cases (false negatives), mainly caused by class imbalance in the dataset (77 negative : 23 positive) and the limited sample size. Prediction on new clinical data resulted in class 0 (negative), although the data were actually classified as positive, confirming the model's tendency to produce false negatives in borderline cases. This study highlights the potential of SVM as a tool for early heart disease diagnosis while emphasizing the importance of handling class imbalance and hyperparameter optimization to improve model sensitivity.

Keywords: Heart Disease, Support Vector Machine, Classification, Machine Learning, Clinical Prediction

Citation: L. Y. Agustin, N. A. Q. Sholeha, S. A. N. Apriliana and N. Muna, "Heart Disease Prediction Using Support Vector Machine (SVM) Classification Based on Clinical Data", *nexura*, vol. 1, no. 1, pp. 50–65, Jun. 2026.

Received: 04-06-2026
Accepted: 27-06-2026
Published: 30-06-2026

1. Introduction

Cardiovascular disease (CVD) has remained the leading cause of death worldwide for more than the last three decades. According to the latest report from the World Health Organization (WHO), approximately 19.8 million people died from cardiovascular disease in 2022, representing around 32% of all global deaths, of which 85% were caused by heart attacks and strokes [1]. Data from the Global Burden of Disease (GBD) Collaboration published in the *Journal of the American College of Cardiology* revealed that the number of global deaths caused by CVD increased from 13.1 million in 1990 to 19.2 million in 2023, while the prevalence of CVD more than doubled from 311 million cases

to 626 million active cases worldwide [2]. The burden of this disease is not limited to health impacts but also has major economic consequences; projections from the American Heart Association indicate that healthcare costs related to cardiovascular risk factors in the United States are expected to triple from USD 400 billion in 2020 to USD 1.344 trillion by 2050 [3]. More concerningly, over three-quarters of CVD-related deaths occur in low- and middle-income countries, including Indonesia, which face limitations in healthcare infrastructure and specialist cardiology services [1]. These conditions confirm that heart disease is not merely an individual medical issue but has evolved into a global health crisis that requires a systematic and technology-based approach.

Modern healthcare systems should ideally have the capability to predict the risk of heart disease early, accurately, and affordably. Competent healthcare professionals, supported by advanced diagnostic technologies such as electrocardiograms (ECG), echocardiography, blood biomarker examinations, and integrated clinical information systems, should be able to identify high-risk patients long before life-threatening complications occur. With advancements in health information technology, clinical patient data stored in electronic medical records should be optimally utilized to generate accurate and personalized risk predictions. However, the reality in practice remains far from ideal. Delayed diagnosis is still a major issue, mainly because most early symptoms of heart disease are non-specific and difficult to identify without in-depth examination [4]. The limited number of cardiology specialists, particularly in remote areas and developing countries, further worsens this condition. On the other hand, human error in interpreting complex clinical data remains an unavoidable risk in conventional diagnostic processes [5]. Furthermore, although healthcare institutions currently generate large volumes of medical data covering hundreds of clinical variables such as blood pressure, cholesterol levels, ECG results, family history, and patient lifestyle, much of this data has not been optimally utilized to support evidence-based clinical decision-making. This gap between the potential of data and the ability to utilize it effectively has become one of the major challenges in heart disease management today.

This research gap has encouraged various studies to explore the application of machine learning (ML) approaches in heart disease prediction based on clinical data. Several machine learning methods commonly used include Logistic Regression (LR), K-Nearest Neighbor (KNN), Random Forest (RF), and Support Vector Machine (SVM). Logistic Regression is well known for its simplicity, high interpretability, and computational efficiency; however, this method has limitations in handling non-linear decision boundaries and complex clinical feature patterns [6]. KNN offers ease of implementation without a complicated training phase, but its performance decreases significantly on high-dimensional datasets and it is highly sensitive to imbalanced data distributions commonly found in clinical heart disease data [7]. Random Forest can handle non-linear relationships and complex feature interactions and has been proven to achieve high accuracy, yet it tends to be difficult to interpret, requires higher computational costs, and is prone to overfitting on small datasets [6]. Support Vector Machine (SVM) has emerged as a highly competitive alternative due to its ability to identify the optimal hyperplane that maximizes the margin between positive and negative classes, making it highly effective for high-dimensional datasets with limited sample sizes. SVM is also proven to be more robust against outliers compared to other methods and offers high flexibility through the use of various kernel functions such as linear, polynomial, and Radial Basis Function (RBF) to handle non-linearly separable data [7]. Studies conducted by Natsir et al. [9] and Hasanah & Nurmalitasari [10] demonstrated that the application of SVM to clinical heart disease data can produce competitive prediction accuracy. Similarly, Tino et al. [11] and Nainggolan et al. [12] have demonstrated the potential of SVM in heart disease classification using the UCI Cleveland dataset and local clinical datasets, confirming that SVM consistently outperforms several other ML methods in

certain heart disease prediction scenarios, particularly when the clinical data used are high-dimensional and the sample size is limited.

This study aims to develop and optimize a heart disease prediction model using the Support Vector Machine (SVM) classification method based on standardized clinical data. The study seeks to optimize SVM performance through a systematic hyperparameter tuning process, including the exploration of various kernel functions (linear, polynomial, and RBF) as well as adjustments to the C and gamma parameters to obtain the most optimal classification model. The main contributions of this study include three key aspects: first, producing an optimized SVM model with higher and more stable prediction accuracy compared to the implementation of SVM with default parameters; second, providing a practical contribution in supporting heart disease prediction systems through the optimal utilization of clinical data; and third, presenting a comprehensive comparison between SVM and other machine learning (ML) methods. It is expected that the results of this study will serve as a valuable scientific foundation for the development of artificial intelligence-based clinical decision support systems, ultimately contributing to improved quality and speed of heart disease diagnosis in healthcare facilities.

2. Materials and Methods

This study is a computational quantitative study with an experimental approach aimed at developing a heart disease prediction classification model using the Support Vector Machine (SVM) method. The experimental approach was carried out through a series of computational experiments on patients' clinical data to evaluate the model's ability to classify heart disease based on the patterns of medical attributes possessed by each patient. The study was designed to systematically evaluate the model's performance using several standard classification evaluation metrics in order to obtain a comprehensive understanding of the model's capability in accurately predicting heart disease.

The dataset used in this study was obtained from a public repository on the Kaggle platform in the form of the Heart Disease Dataset with the file name "heart - 100.csv". The dataset is an adaptation of the Cleveland Heart Disease dataset, which has been widely used in machine learning research in the healthcare field due to its relevant and valid clinical data structure for heart disease classification processes. The dataset consists of 100 patient records with 14 clinical attributes covering important characteristics and medical parameters for heart disease identification, namely age, sex, chest pain type (cp), resting blood pressure (trestbps), serum cholesterol level (chol), fasting blood sugar (fbs), resting electrocardiographic results (restecg), maximum heart rate achieved (thalach), exercise-induced angina (exang), ST depression induced by exercise (oldpeak), slope of the peak exercise ST segment (slope), number of major vessels colored by fluoroscopy (ca), thalassemia (thal), and the target variable (target), which indicates the patient's condition, where 0 represents no heart disease and 1 represents heart disease.

Data collection was conducted using a documentation study method on publicly available secondary datasets. The dataset was imported into the programming environment using the Google Colaboratory platform through a manual file upload mechanism utilizing the files module from the google.colab library. Subsequently, the dataset was read using the pd.read_csv() function from the Pandas library in the Python programming language, resulting in a DataFrame structure ready for further analysis stages. The data preprocessing stage was one of the important processes in this study because it aimed to ensure data quality and consistency before being used in the model training process. The initial preprocessing stage was carried out through data exploration and understanding. At this stage, dataset dimensions were inspected using the df.shape attribute to determine the number of rows and columns, several initial records were

displayed using df.head() to obtain an overview of the dataset structure, and df.info() was used to identify the data type of each attribute. In addition, descriptive statistics were calculated using the df.describe() function to obtain information regarding the distribution of data in each feature, including minimum value, maximum value, mean, and standard deviation. After data exploration, the next stage involved checking data quality. Missing value detection was performed using the df.isnull().sum() command to ensure that no empty values existed in each dataset attribute. Duplicate data checking was conducted using df.duplicated().sum() to ensure there were no repeated records so that data integrity remained maintained. Furthermore, the target class distribution was analyzed using df['target'].value_counts() to determine the balance between the number of patients with heart disease and those without heart disease. This stage is important because imbalanced class distribution may affect the performance of the classification model. In addition to data exploration, this study also utilized data visualization using the Matplotlib and Seaborn libraries to help understand data distribution, relationships among variables, and target class distribution before the model training process. Data visualization was used to provide an initial overview of dataset characteristics, thereby facilitating the interpretation and analysis of clinical data.

The next stage involved separating features and the target variable. Features (X) were obtained by removing the target column from the DataFrame using the command df.drop('target', axis=1), while the target variable (y) was taken from the target column. Afterward, a data standardization process was performed using StandardScaler from the Scikit-learn library. Standardization was conducted through the scaler.fit_transform(X) function, which transformed all features to have a mean close to zero and a standard deviation close to one. The standardization process is highly important in the SVM method because this algorithm is sensitive to differences in feature scales. Features with significantly larger value ranges may dominate the hyperplane formation process and affect the model's ability to perform optimal classification. Through standardization, all features have relatively balanced scales, allowing the model training process to run more effectively and produce better classification performance. Dataset splitting was performed using the train_test_split function from the sklearn.model_selection module with an 80:20 proportion, where 80% of the data were used as training data and 20% as testing data. The parameter random_state=42 was used to ensure reproducibility of the data splitting results so that consistent experimental outcomes could be obtained in each program execution. Based on the total of 100 patient records, 80 records were used as training data and 20 records as testing data.

The implementation of the Support Vector Machine algorithm was carried out using the SVC (Support Vector Classifier) class from the sklearn.svm module. The SVM method was selected in this study because it has strong capability in handling binary classification problems, particularly for small- to medium-sized datasets. In addition, SVM can construct an optimal hyperplane with maximum margin, providing strong generalization capability in distinguishing positive and negative classes. The use of the Radial Basis Function (RBF) kernel also enables the model to handle nonlinear data patterns commonly found in clinical heart disease data. The SVM model was configured using the Radial Basis Function (RBF) kernel, which maps data into a higher-dimensional space, allowing the separation of data that cannot be linearly separated in the original feature space. The model training process was conducted using the model.fit(X_train, y_train) method, in which the model learned patterns from the training data to determine support vectors and the optimal hyperplane capable of maximizing the margin between classes. After the training process was completed, the model was used to make predictions on the testing data through the model.predict(X_test) method, resulting in classification outcomes for each patient record. The main parameters used in the SVM model in this study included kernel='rbf', C=1.0, and gamma='scale'. The parameter C functions as a regularization

parameter that controls the balance between the hyperplane margin and the classification error rate in the training data. A larger C value causes the model to focus more on minimizing classification errors but may increase the risk of overfitting. Meanwhile, the gamma parameter determines the extent of influence of a single training instance on the formation of the model's decision boundary. The use of gamma='scale' allows the gamma value to be calculated automatically based on data variance so that the model can adaptively adjust its complexity.

The software and tools used in this study included Python version 3 programming language executed through the cloud-based Google Colaboratory platform. The main libraries utilized included Pandas for tabular data manipulation, NumPy for numerical computation, Scikit-learn for machine learning algorithm implementation, Matplotlib and Seaborn for data visualization, as well as various model evaluation modules such as accuracy_score, confusion_matrix, and classification_report. The model evaluation stage was conducted using several standard classification evaluation metrics. The first evaluation used accuracy to measure the model's correctness in predicting all testing data. In addition, this study also used precision to measure the model's accuracy in identifying the positive class, recall to measure the model's ability to detect all actual positive cases, and F1-score to evaluate the balance between precision and recall performance. Besides these evaluation metrics, this study also utilized a confusion matrix to present the distribution of classification results in detail. The confusion matrix consists of True Positive (TP), True Negative (TN), False Positive (FP), and False Negative (FN), which were used to analyze the model's capability in distinguishing patients with heart disease from those without heart disease. The confusion matrix visualization was displayed in the form of a heatmap using Seaborn to facilitate interpretation of the classification results. To test the model's generalization capability, this study also performed predictions on new clinical data representing a hypothetical patient record with specific medical attributes. This stage was conducted to evaluate the model's ability to provide predictions on data that had never been used during the training process.

3. Results and Discussions

3.1 Result

This section presents the results of each research stage that was systematically conducted, starting from data exploration and preprocessing to the evaluation stage of the Support Vector Machine classification model performance.

3.1.1 Results of Data Exploration and Preprocessing

The dataset used in this study consisted of 100 patient records and 14 attribute columns, including 13 clinical variables and 1 target variable. Based on the examination results using the df.info() and df.isnull().sum() functions, all attributes in the dataset had numerical data types in the form of int64 and float64, and no missing values were found in any column. These conditions indicate that the dataset was clean and complete, so no missing value handling or data imputation process was required. In addition, the duplicate data examination using df.duplicated().sum() produced a value of 0, indicating that there were no duplicate records in the dataset. Therefore, all available patient data could be optimally utilized in the research process.

Table 1. Initial Display of the Heart Disease Dataset

```
df
```

	age	sex	cp	trestbps	chol	fbs	restecg	thalach	exang	oldpeak	slope	ca	thal	target
0	63	1	3	145	233	1	0	150	0	2.3	0	0	1	1
1	37	1	2	130	250	0	1	187	0	3.5	0	0	2	1
2	41	0	1	130	204	0	0	172	0	1.4	2	0	2	1
3	56	1	1	120	236	0	1	178	0	0.8	2	0	2	1
4	57	0	0	120	354	0	1	163	1	0.6	2	0	2	1
5	57	1	0	140	192	0	1	148	0	0.4	1	0	1	1
6	56	0	1	140	294	0	0	153	0	1.3	1	0	2	1
7	44	1	1	120	263	0	1	173	0	0.0	2	0	3	1
8	52	1	2	172	199	1	1	162	0	0.5	2	0	3	1
9	57	1	2	150	168	0	1	174	0	1.6	2	0	2	1
10	65	0	2	160	360	0	0	151	0	0.8	2	0	2	1
11	51	0	2	140	308	0	0	142	0	1.5	2	1	2	1
12	48	1	1	130	245	0	0	180	0	0.2	1	0	2	1
13	45	1	0	104	208	0	0	148	1	3.0	1	0	2	1
14	53	0	0	130	264	0	0	143	0	0.4	1	0	2	1

The distribution of the target variable shows that the dataset consists of two classes, namely label 0 representing patients without heart disease and label 1 representing patients with heart disease. The distribution of data in each class was visualized using a bar chart to provide an overview of the class distribution within the dataset. This class distribution analysis is important because the balance of data in each class can affect the performance of the classification model, particularly the precision, recall, and F1-score values produced by the Support Vector Machine model.

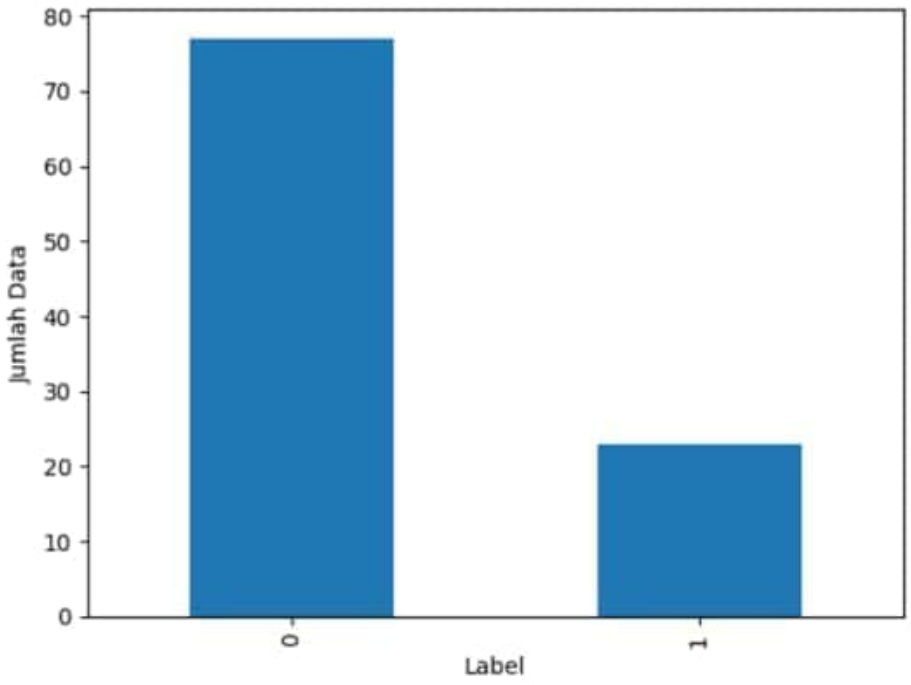


Figure 1. Distribution of Heart Disease Target Classes

The data standardization stage using StandardScaler successfully transformed all features to have a mean close to zero and a standard deviation close to one. The standardization process was performed to ensure that each feature had a balanced scale so that no particular feature dominated the model training process. This step is highly important in the implementation of the SVM method because the algorithm is sensitive to differences in feature value ranges. Through standardization, features with large value ranges such as serum cholesterol level (chol) and resting blood pressure (trestbps) did not dominate other features with smaller or binary values such as sex (sex) and fasting blood sugar (fbs). After the standardization process was completed, the dataset was divided into training data and testing data using the train_test_split function with an 80:20 ratio and the parameter random_state=42 to maintain consistency in the data splitting results. The splitting results showed that the training data consisted of 80 samples, while the testing data consisted of 20 samples. In addition, the resulting data dimensions were X_train of (80, 13), X_test of (20, 13), y_train of (80,), and y_test of (20,), indicating that all 13 features were used in the training and testing processes of the classification model.

```
# ============================================================
# SPLIT DATA TRAINING DAN TESTING
# ============================================================

X_train, X_test, y_train, y_test = train_test_split(
    X,
    y,
    test_size=0.2,
    random_state=42
)
```

Figure 2. Dataset Splitting into Training and Testing Data

The Support Vector Machine classification model with the Radial Basis Function (RBF) kernel was trained using 80 training data samples that had undergone the standardization process. The training process was carried out using the model.fit(X_train, y_train) function, in which the model learned the relationship patterns among clinical features to construct an optimal hyperplane as the decision boundary between classes. During this process, the SVM algorithm internally identified support vectors that played an important role in determining the position of the hyperplane in the high-dimensional feature space generated through the RBF kernel transformation. The use of the RBF kernel enabled the model to handle nonlinear data patterns, allowing the classification process to be performed more effectively. The training results showed that the model was successfully developed without any errors or warnings during the training process. This condition indicates that the quality of the data used was adequate and that the applied model parameters were suitable for the characteristics of the heart disease dataset used in this study.

```
# ============================================================
# MEMBUAT MODEL SVM
# ============================================================

model = SVC(kernel='rbf')
```

Figure 3. Initialization of the SVM Model Using the RBF Kernel

3.1.2 Results of Model Testing and Evaluation

The testing process of the Support Vector Machine model was conducted using 20 testing data samples through the model.predict(X_test) function. The generated predictions were then compared with the actual labels (y_test) to calculate the model performance based on classification evaluation metrics.

Based on the testing results, the SVM model with the Radial Basis Function (RBF) kernel achieved an accuracy of 85%. This value indicates that the model was able to correctly classify 17 out of 20 testing data samples. The obtained accuracy demonstrates that the model has a fairly good classification capability in distinguishing patients diagnosed with heart disease from those not diagnosed with heart disease based on the clinical attributes used in this study.

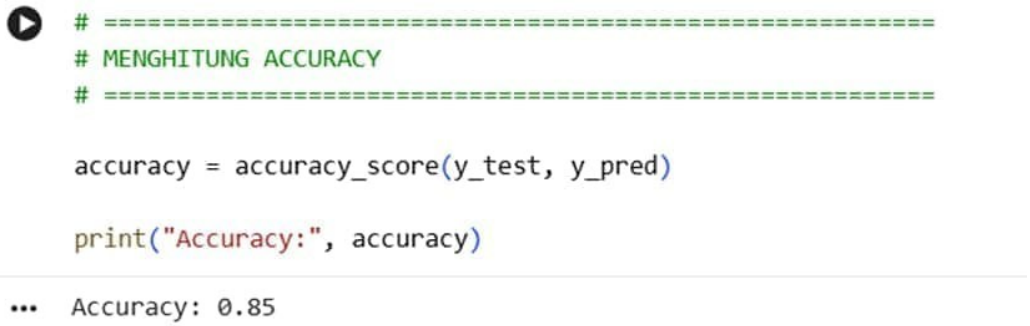


Figure 4. Accuracy Calculation Results of the SVM Model

3.1.3 Confusion Matrix

A confusion matrix is a tabular representation used to evaluate the performance of a classification model by systematically comparing the model predictions with the actual labels. In this study, the confusion matrix was obtained using the confusion_matrix(y_test, y_pred) function to determine the distribution of correct and incorrect predictions generated by the Support Vector Machine model.

The confusion matrix consists of four main components: True Positive (TP), True Negative (TN), False Positive (FP), and False Negative (FN). True Positive represents the number of patients diagnosed with heart disease who were correctly predicted by the model. True Negative represents the number of patients without heart disease who were correctly classified by the model. False Positive represents the number of patients who actually did not have heart disease but were predicted by the model as having heart disease. Meanwhile, False Negative represents the number of patients who actually had heart disease but were predicted as not having heart disease.

The confusion matrix visualization was presented in the form of a heatmap using the Seaborn library through the sns.heatmap(cm, annot=True, fmt='d') function. In this visualization, the horizontal axis (X-axis) represents the predicted labels generated by the model, while the vertical axis (Y-axis) represents the actual labels from the testing data. Each cell in the matrix contains numerical values indicating the number of data samples for a particular combination of actual and predicted classes.

The heatmap visualization simplifies the interpretation process because the prediction distribution can be visually observed through the color intensity in each matrix cell. Darker colors indicate a higher number of data samples in a particular category. Based on the obtained confusion matrix results, the distribution of correct predictions and classification errors made by the model can be identified, thereby supporting the analysis

of the model's performance in distinguishing patients with heart disease from those without heart disease.

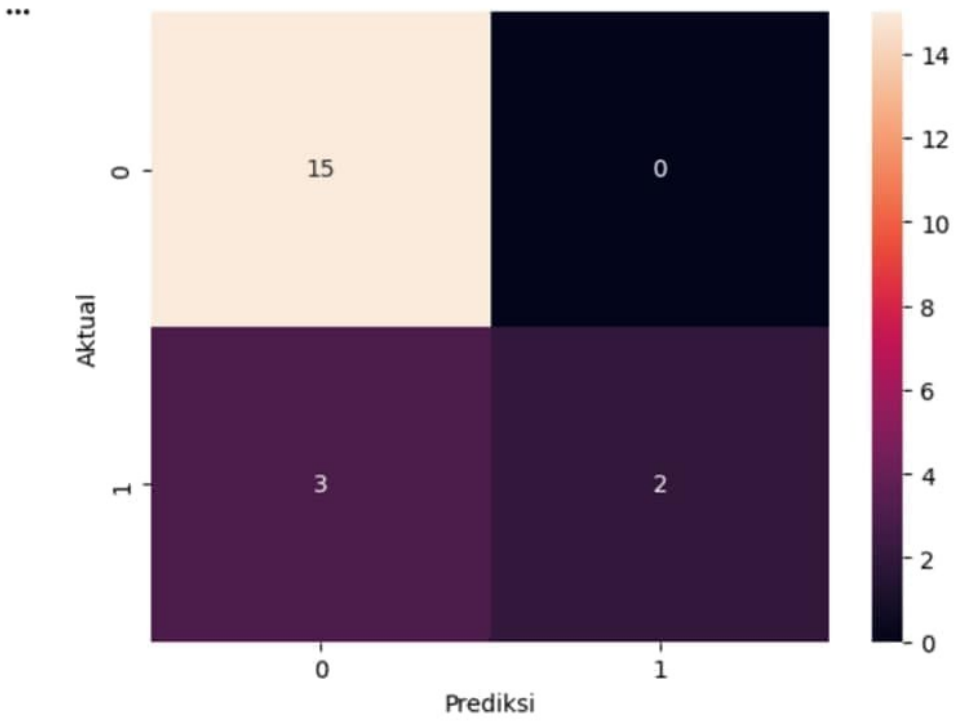


Figure 5. Confusion Matrix Visualization of the SVM Model

3.1.4 Confusion Matrix Table of the SVM Model with RBF Kernel

The confusion matrix table was used to explain the classification results of the Support Vector Machine model in distinguishing patients diagnosed with heart disease from those without heart disease. The confusion matrix presents a comparison between the actual labels and the model prediction results, allowing the number of correct predictions and classification errors produced by the model to be identified. In the confusion matrix table, the rows represent the actual conditions of the data, while the columns represent the prediction results generated by the model. There are four main components in the confusion matrix: True Negative (TN), False Positive (FP), False Negative (FN), and True Positive (TP).

The True Negative (TN) component represents the number of patients who actually did not have heart disease and were correctly predicted by the model as negative. The True Positive (TP) component represents the number of patients with heart disease who were correctly predicted as positive. Meanwhile, False Positive (FP) represents the number of patients who actually did not have heart disease but were predicted by the model as having heart disease. This condition is commonly referred to as a false positive prediction. In contrast, False Negative (FN) represents the number of patients who actually had heart disease but were predicted by the model as not having heart disease. Through the confusion matrix, the model performance can be analyzed in greater detail because it not only shows the number of correct predictions but also illustrates the distribution of classification errors in each class. This information is particularly important in heart disease prediction research because classification errors, especially False Negatives, may affect the identification process of patients who actually have a risk of heart disease but are not detected by the model.

Table 2. Confusion Matrix of the SVM Model with RBF Kernel

	Prediksi: 0 (Negatif)	**Prediksi: 1 (Positif)**
Aktual: 0 (Negatif)	TN	FP
Aktual: 1 (Positif)	FN	TP

3.1.5 Classification Report Results

The model evaluation results were presented through a classification report obtained using the classification_report(y_test, y_pred) function. This output provides information regarding the model's performance in classifying data into each class based on three main metrics: precision, recall, and F1-score. In addition, the report also presents overall average values in the form of macro average and weighted average, which describe the global performance of the model across all classes. Based on the data processing results, these metrics were used to evaluate the accuracy level and the balance of the Support Vector Machine (SVM) model performance in performing classification tasks.

Table 3. Classification Report Evaluation of the SVM Model

	precision	recall	f1-score	support
0	0.83	1.00	0.91	15
1	1.00	0.40	0.57	5
accuracy			0.85	20
macro avg	0.92	0.70	0.74	20
weighted avg	0.88	0.85	0.82	20

The overall accuracy of the SVM model reached 85%, obtained from the comparison between the number of correct predictions and the total number of testing data samples. Precision for the heart disease class indicates the model's ability to avoid false positive predictions, while recall reflects the model's capability to detect all actual heart disease cases by minimizing false negatives. The F1-score represents the harmonic mean of precision and recall, providing a balanced measure of model performance, which is particularly important in the medical context where prediction errors in both directions may have significant clinical consequences.

3.1.6 Results of New Data Prediction

The generalization capability of the model was also tested through prediction on a new clinical dataset representing a hypothetical patient with the following characteristics: 63 years old, male (1), chest pain type 3, resting blood pressure of 145 mmHg, cholesterol level of 233 mg/dL, positive fasting blood sugar (1), normal resting ECG result (0), maximum heart rate of 150 bpm, no exercise-induced angina (0), ST depression of 2.3, flat ST slope (0), no major vessels colored by fluoroscopy (0), and thal type 1. Based on the prediction results generated through the model.predict(data_baru) command, the SVM model produced a class 0 prediction, indicating that the patient with this clinical profile was classified as not having heart disease (negative).

This prediction result indicates that although the patient had several clinical risk parameters such as age of 63 years, blood pressure of 145 mmHg, and cholesterol level of 233 mg/dL, the SVM model considered the overall combination of clinical attributes insufficient to categorize the patient as having heart disease. The model's ability to classify

new unseen data demonstrates the potential application of the SVM model as a tool for early heart disease diagnosis in clinical practice, although the prediction results still need to be confirmed through further medical examination by authorized healthcare professionals.

3.2 Discussion

3.2.1 Analysis of SVM Model Performance

Based on the evaluation results presented in the previous section, the SVM model with the RBF kernel applied to the clinical heart disease dataset consisting of 100 samples achieved an accuracy of 85% on the testing data. This result indicates that the model has a fairly good classification capability in distinguishing patients diagnosed with heart disease from those without heart disease, despite the relatively small dataset size. An accuracy of 85% means that out of 20 testing samples, the model correctly classified 17 samples and made errors on 3 samples. This achievement provides clear evidence of the capability of SVM in handling high-dimensional clinical data classification with limited sample sizes, in accordance with the theoretical advantages of SVM widely discussed in machine learning literature.

The obtained precision value indicates the model's level of accuracy in identifying positive cases (heart disease). A high precision value means that the model rarely classifies healthy patients as patients with heart disease (low false positive rate), which in the clinical context helps reduce unnecessary concern for healthy patients and avoids inefficient use of medical resources for unnecessary follow-up examinations. On the other hand, the recall (sensitivity) value reflects the model's ability to detect all existing heart disease cases. High recall is highly crucial in medical diagnostic applications because failure to detect actual positive cases (false negatives) may result in untreated disease conditions that could threaten the patient's life. The F1-score of 0.57 indicates that although the model achieved very high precision, its ability to identify all positive cases remained limited, resulting in an imbalance between precision and recall.

3.2.2 Interpretation of Accuracy Results and Factors Affecting Model Performance

The accuracy of 85% achieved in this study was influenced by several factors. First, the limited dataset size of 100 samples represents a significant limitation. Although SVM is theoretically effective for high-dimensional datasets with limited sample sizes, a very small number of samples may cause the model to be less capable of capturing the full variation of relevant clinical patterns needed to distinguish between healthy and diseased patients, thereby limiting the model's generalization capability on new data. A larger dataset would provide the model with richer exposure to pattern variations, which in turn is expected to significantly improve accuracy.

Second, the use of default parameter values for C and gamma in the RBF kernel provided a good baseline performance, but these parameters may not necessarily be optimal for the specific characteristics of this dataset. A hyperparameter tuning process using techniques such as Grid Search or Randomized Search Cross-Validation could identify the combination of C and gamma values that produces the best performance. For example, a larger C value would produce a more complex model with a narrower margin but fewer classification errors on the training data, although the risk of overfitting would also increase. Third, the selection of the RBF kernel proved appropriate for this dataset because the kernel is capable of handling data that are not linearly separable in the original feature space.

Fourth, the standardization process using StandardScaler contributed positively to the model's performance. Without standardization, features with large value ranges such as cholesterol level (chol, ranging from 100–600 mg/dL) would dominate the kernel function and obscure the contribution of binary-type features. Standardization ensures that each feature contributes proportionally in the calculation of distances between data points, which is the core mechanism of the SVM kernel. Fifth, the 80:20 data splitting ratio is a reasonable and commonly used choice; however, with a dataset containing only 100 samples, the testing subset consisting of only 20 samples may produce unstable performance estimates. The use of k-fold cross-validation techniques would provide more reliable and representative performance estimates.

3.2.3 Advantages and Limitations of the SVM Method in the Context of This Study

The Support Vector Machine method demonstrated several advantages that are highly relevant to heart disease prediction research. First, SVM proved effective in handling high-dimensional clinical data (14 features) with a limited sample size (100 records). This capability makes SVM particularly suitable for small datasets and high-dimensional data, as reported in previous machine learning studies. Second, the margin maximization mechanism in SVM provides better robustness against outliers compared to distance-based methods such as K-Nearest Neighbor, which is particularly relevant because clinical data often contain extreme values in certain physiological parameters. Third, the flexibility of SVM kernels, particularly the RBF kernel used in this study, allows the model to capture complex nonlinear relationships among clinical features that cannot be modeled using simple linear approaches. Fourth, SVM tends to have a relatively lower risk of overfitting compared to several other methods when applied to small datasets, due to the structural risk minimization principle that forms its theoretical foundation.

Nevertheless, the application of SVM in this study also has several limitations that need to be acknowledged scientifically. First, the interpretability of the SVM model is relatively low (black box) compared to methods such as Decision Tree or Logistic Regression. The ability to explain why the model produces certain predictions (explainability) is highly important for healthcare professionals in building trust toward artificial intelligence-based systems. Second, SVM performance is highly dependent on the selection of appropriate kernels and hyperparameter values. The use of default parameter values without systematic tuning may result in a model that does not achieve its optimal performance. Third, the computational time of SVM can increase significantly as the dataset size grows, although this was not a major issue for the dataset containing 100 samples used in this study. Fourth, class imbalance in the dataset, if present, may asymmetrically affect model performance, where the model tends to be more accurate in predicting the majority class.

3.2.4 Relationship Between Research Results, Theory, and Previous Studies

The results of this study are consistent with the theoretical foundation of SVM, which states that this method is highly effective in classification scenarios involving high-dimensional data and limited sample sizes. Theoretically, SVM works by identifying an optimal hyperplane that maximizes the margin between classes in the feature space; a larger margin is associated with better generalization capability on previously unseen data. Kernelization through the RBF kernel enables SVM to operate implicitly in a very high-dimensional feature space without explicitly computing the transformation, making it computationally efficient while remaining mathematically expressive.

The findings of this study are also consistent with results reported in previous studies. Research conducted by Natsir et al. [9] and Hasanah & Nurmalitasari [10], as cited in the introduction, demonstrated that the application of SVM to clinical heart disease data can achieve competitive prediction accuracy, which is consistent with the 85%

accuracy obtained in this study. Similarly, Tino et al. [11] and Nainggolan et al. [12] also reported favorable results using SVM for heart disease prediction, highlighting its effectiveness in handling clinical datasets with multiple attributes. Although the dataset used in this study was relatively smaller than those used in several previous studies, the overall findings and performance patterns remain consistent with the existing literature, supporting the suitability of SVM for heart disease classification tasks.

3.2.5 Analysis of False Negatives and Class Imbalance

One of the critical findings of this study was the occurrence of false negatives in the prediction of new clinical data. The hypothetical patient data used in the testing process (63 years old, blood pressure of 145 mmHg, cholesterol level of 233 mg/dL, ST depression of 2.3, and chest pain type 3) actually had a target label of 1 (positive for heart disease) in the dataset; however, the SVM model predicted it as class 0 (negative). This condition was confirmed by the confusion matrix results, which showed FN = 3 out of 5 positive cases in the testing data, resulting in a positive-class recall of only 0.40 or 40%. There are three main factors that explain why this false negative prediction occurred despite the overall accuracy reaching 85%.

First, class imbalance was the most dominant factor. The dataset distribution consisted of 77 negative samples (77%) and 23 positive samples (23%), resulting in an imbalance ratio of approximately 3.3:1. Consequently, the testing data contained 15 negative samples and only 5 positive samples. Under these conditions, the SVM model tended to be biased toward the majority class (negative) because during training the model was exposed more frequently to negative data patterns. When the model encountered positive cases with attribute patterns different from the majority of the training data, it had difficulty recognizing them as positive cases, resulting in false negatives.

Second, the very limited dataset size (100 samples) meant that only 23 positive samples were available to train the model in recognizing heart disease patterns. With only 80% of the data allocated for training, approximately 18–19 positive samples were used during the training process, which is a very small number for enabling the model to capture the full variation of positive clinical patterns. As a result, the training accuracy reached 97.5%, while the testing accuracy was only 85%, showing a gap of 12.5% that indicates mild overfitting, where the model became too adapted to the training data patterns but lacked sufficient generalization capability on new data.

Third, the use of default parameters (C=1.0 and gamma='scale') without hyperparameter tuning likely produced a decision boundary that was not optimal for handling borderline cases such as patients with ambiguous clinical profiles. For example, a higher C value could produce a narrower margin but greater sensitivity in detecting positive cases, although this would increase the risk of false positives. In the clinical context, false negatives are far more dangerous than false positives because patients who actually have heart disease may go undetected and potentially fail to receive the necessary treatment.

These findings indicate that further model improvement is needed, particularly due to the low sensitivity (positive-class recall) of 40%. To improve performance, future studies should implement class imbalance handling techniques such as SMOTE (Synthetic Minority Over-sampling Technique) or class weighting in SVM parameters, perform systematic hyperparameter tuning, and expand the dataset using more representative clinical data.

One aspect that distinguishes this study from several previous studies is the simplicity and transparency of the end-to-end research workflow implemented using the

cloud-based Google Colaboratory platform, demonstrating the accessibility of SVM implementation for heart disease prediction even under limited computational infrastructure. This approach is particularly relevant in the context of Indonesia, where limitations in healthcare and computational infrastructure may still present challenges. The results suggest that a Python-based SVM model can achieve promising predictive performance with relatively minimal computational resources and may serve as a preliminary foundation for future research on clinical decision-support systems.

5.Conclusion

This study successfully developed a Support Vector Machine (SVM) classification model with a Radial Basis Function (RBF) kernel using a clinical heart disease dataset consisting of 100 samples and 13 clinical attributes. The model achieved an overall testing accuracy of 85%. The confusion matrix results yielded TN=15, FP=0, FN=3, and TP=2, with a precision of 1.00, a recall of 0.40, and an F1-score of 0.57 for the positive class. These findings indicate that the model was highly precise when identifying positive cases but had limited ability to detect all actual heart disease cases. Furthermore, the training accuracy of 97.5% compared with the testing accuracy of 85% revealed a performance gap of 12.5%, suggesting the presence of mild overfitting and indicating that model generalization could be further improved. The observed limitations may be associated with class imbalance (77:23), the relatively small dataset size, and the absence of systematic hyperparameter optimization. Nevertheless, this study demonstrates that SVM implementation using Python on the Google Colaboratory platform can be performed efficiently with minimal computational resources, highlighting its potential as an accessible approach for heart disease prediction. Future research is recommended to address class imbalance through techniques such as SMOTE or the use of the class_weight='balanced' parameter, perform systematic hyperparameter optimization using Grid Search or Randomized Search Cross-Validation, expand the dataset with more diverse clinical records, apply k-fold cross-validation to obtain more robust performance estimates, and compare SVM with other machine learning algorithms, including Random Forest, K-Nearest Neighbor, Logistic Regression, and Gradient Boosting, to further improve predictive performance and model generalization.

References

[1] World Health Organization, "*Cardiovascular diseases (CVDs) — Fact Sheet*," World Health Organization, 2023. [Online]. *Available: https://www.who.int/news-room/fact-sheets/detail/cardiovascular-diseases-%28cvds%29.* [Accessed: Jun. 2, 2026].

[2] G. A. Roth et al., "*Global, Regional, and National Burden of Cardiovascular Diseases and Risk Factors in 204 Countries and Territories, 1990–2023*," Journal of the American College of Cardiology, 2025, doi: 10.1016/j.jacc.2025.08.015.

[3] R. Kumar et al., "*A Comprehensive Review of Machine Learning for Heart Disease Prediction: Challenges, Trends, Ethical Considerations, and Future Directions*," Frontiers in Artificial Intelligence, vol. 8, Art. no. 1583459, 2025.

[4] F. M. Natsir, R. Y. Bakti, and T. Wahyuni, "*Analisis Deteksi Dini Penyakit Jantung dengan Pendekatan Support Vector Machine pada Data Pasien [Early Detection Analysis of Heart Disease Using a Support Vector Machine Approach on Patient Data*]," Arus Jurnal Sains dan Teknologi (AJST), vol. 2, no. 2, pp. 437–446, Oct. 2024.

[5] S. A. Aini, D. Fitriani, M. A. Awwalin, A. A. R. Ramadhan, and I. W. D. Prastya, "K*lasifikasi Penyakit Jantung Berdasarkan Faktor Klinis Menggunakan Algoritma XG Boost, Random Forest, Support Vector Machine (SVM) dan K-Nearest Neighbor (KNN) [Heart Disease Classification Based on Clinical Factors Using XGBoost, Random Forest, Support Vector Machine (SVM), and K-Nearest Neighbor (KNN)]*," JATI (Jurnal Mahasiswa Teknik Informatika), vol. 10, no. 2, pp. 3449–3454, Apr. 2026.

[6] G. A. Pamungkas, R. Rahmadewi, and I. P. Sary, "*Klasifikasi Risiko Penyakit Jantung Menggunakan Algoritma Support Vector Machine (SVM) [Heart Disease Risk Classification Using the Support Vector Machine (SVM) Algorithm]*," JATI (Jurnal Mahasiswa Teknik Informatika), vol. 9, no. 3, pp. 5438–5443, Jun. 2025.

[7] J. K. Nainggolan, F. Sinaga, A. M. Sitorus, A. Khairia, and B. A. Wijaya, *"Analisa Komparasi dengan Algoritma K-Nearest Neighbor (KNN) dan Support Vector Machine (SVM) untuk Prediksi Penyakit Jantung [Comparative Analysis of K-Nearest Neighbor (KNN) and Support Vector Machine (SVM) Algorithms for Heart Disease Prediction]*," DINAMIK, vol. 30, no. 2, pp. 297–306, 2025.

[8] H. Hasanah and Nurmalitasari, "*Perbandingan Tingkat Akurasi Algoritma Support Vector Machines (SVM) dan C4.5 dalam Prediksi Penyakit Jantung [Comparison of the Accuracy Levels of Support Vector Machine (SVM) and C4.5 Algorithms in Heart Disease Prediction]*," in Proceedings of the National Seminar on Technology and Science 2023, Kediri, Indonesia, Jan. 2023, vol. 2, pp. 13–18.

[9] M. D. F. Tino, H. Hasanah, and T. D. Santosa, "*Perbandingan Algoritma Support Vector Machines (SVM) dan Neural Network untuk Klasifikasi Penyakit Jantung [Comparison of Support Vector Machine (SVM) and Neural Network Algorithms for Heart Disease Classification]*," INFOTECH Journal, vol. 3, no. 1, pp. 232–235, May 2023.

[10] D. Setiawan, A. Muhammad, and S. H. F. Dewi, "*Penerapan Algoritma Klasifikasi untuk Deteksi Dini Penyakit Jantung Koroner Berdasarkan Gejala Klinis [Application of Classification Algorithms for Early Detection of Coronary Heart Disease Based on Clinical Symptoms],"* Teknik: Jurnal Ilmu Teknik dan Informatika, vol. 5, no. 1, pp. 18–26, May 2025.

[11] N. L. Fitriyani, M. Syafrudin, G. Alfian, and J. Rhee, "*HDPM: An Effective Heart Disease Prediction Model for a Clinical Decision Support System,*" IEEE Access, vol. 8, pp. 133034–133050, 2020.

[12] J. P. Li, A. U. Haq, S. U. Din, J. Khan, A. Khan, and A. Saboor, *"Heart Disease Identification Method Using Machine Learning Classification in E-Healthcare*," IEEE Access, vol. 8, pp. 107562–107582, Jun. 2020.

[13] S. Mohan, C. Thirumalai, and G. Srivastava, *"Effective Heart Disease Prediction Using Hybrid Machine Learning Techniques,"* IEEE Access, vol. 7, pp. 81542–81554, Jun. 2019.

[14] M. M. Ali, B. K. Paul, K. Ahmed, F. M. Bui, J. M. W. Quinn, and M. A. Moni, *"Heart Disease Prediction Using Supervised Machine Learning Algorithms: Performance Analysis and Comparison,"* Computers in Biology and Medicine, vol. 136, p. 104672, Sep. 2021.

[15] P. Ghosh, S. Azam, M. Jonkman, A. Karim, F. M. J. M. Shamrat, E. Ignatious, S. Shultana, A. R. Beeravolu, and F. De Boer, *"Efficient Prediction of Cardiovascular Disease Using Machine Learning Algorithms With Relief and LASSO Feature Selection Techniques,*" IEEE Access, vol. 9, pp. 19304–19326, Jan. 2021.

[16] R. Katarya and S. K. Meena, *"Machine Learning Techniques for Heart Disease Prediction: A Comparative Study and Analysis,*" Health and Technology, vol. 11, no. 1, pp. 87–97, Jan. 2021, doi: 10.1007/s12553-020-00505-7.

[17] A. Rahim, Y. Rasheed, F. Azam, M. W. Anwar, M. A. Rahim, and A. W. Muzaffar, *"An Integrated Machine Learning Framework for Effective Prediction of Cardiovascular Diseases,"* IEEE Access, vol. 9, pp. 106575–106588, Jul. 2021.

[18] G. N. Ahmad, H. Fatima, S. Ullah, A. S. Saidi, and Imdadullah, *"Efficient Medical Diagnosis of Human Heart Diseases Using Machine Learning Techniques With and Without GridSearchCV,"* IEEE Access, vol. 10, pp. 80151–80173, 2022.

[19] M. Swathy and K. Saruladha, "*A Comparative Study of Classification and Prediction of Cardio-Vascular Diseases (CVD) Using Machine Learning and Deep Learning Techniques,"* ICT Express, vol. 8, no. 1, pp. 109–116, Mar. 2022.

[20] S. Subramani et al., *"Cardiovascular Diseases Prediction by Machine Learning Incorporation with Deep Learning,"* Frontiers in Medicine, vol. 10, p. 1150933, Apr. 2023.

[21] M. Pal, S. Parija, G. Panda, K. Dhama, and R. K. Mohapatra, "*Risk Prediction of Cardiovascular Disease Using Machine Learning Classifiers,"* Open Medicine, vol. 17, no. 1, pp. 1100–1113, Jun. 2022.

[22] E. Owusu, P. Boakye-Sekyerehene, J. K. Appati, and J. Y. Ludu, *"Computer-Aided Diagnostics of Heart Disease Risk Prediction Using Boosting Support Vector Machine,"* Scientific Programming, vol. 2021, Art. no. 3152618, Dec. 2021.