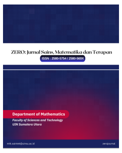
Zero : Jurnal Sains, Matematika, dan Terapan
E-ISSN : 2580-5754; P-ISSN : 2580-569X
Volume 9, Number 2, 2025
DOI: 10.30829/zero.v9i2.26087
Page: 612-619



Prediction of Heart Disease Risk Based on Patient Health History Using the Support Vector Machine (SVM) Algorithm

[1] Septian Simatupang
Department of Software Engineering Technology, Wilmar Bisnis Indonesia Polytechnic, Medan, Indonesia

[2] Rizki Ramadhansyah
Department of Software Engineering Technology, Wilmar Bisnis Indonesia Polytechnic, Medan, Indonesia

[3] Rustianna Tumanggor
Department of Nursing, Universitas Murni Teguh, Medan, Indonesia

[4] Eric Pratama Tan
Department of Software Engineering Technology, Wilmar Bisnis Indonesia Polytechnic, Medan, Indonesia

[5] Syafrizal Amri Fajar
Department of Software Engineering Technology, Wilmar Bisnis Indonesia Polytechnic, Medan, Indonesia

Article Info

Article history:

Accepted, 15 November 2025

Keywords:

Early detection;
Heart disease;
Medical history;
Risk prediction;
Support vector machine

ABSTRACT

Heart disease remains the leading cause of death worldwide, with early detection being critical to improving patient outcomes. This study develops a heart disease risk prediction model using the Support Vector Machine (SVM) algorithm. A dataset of 303 patient records with 14 clinical attributes was used, including age, blood pressure, cholesterol, and chest pain type. Data preprocessing, normalization, and feature selection were performed to optimize the model. Evaluation metrics such as accuracy (92%), precision (90%), recall (96%), and F1-score (93%) demonstrated significant improvements over the baseline model. These results highlight the SVM model's effectiveness as a tool for early heart disease detection, offering potential for enhanced predictive healthcare, particularly in Indonesian clinical settings.

Corresponding Author:

Septian Simatupang,
Department of Software Engineering Technology
Polytechnic Wilmar Bisnis Indonesia, Medan, Indonesia
Email: septian.simatupang@wbi.ac.id

1. INTRODUCTION

Heart disease continues to be the leading cause of death globally, including in Indonesia, where it represents a significant portion of mortality rates. In 2021, healthcare expenditures for heart disease in Indonesia reached Rp7.7 trillion, underscoring the growing burden it places on the healthcare system both clinically and economically [1], [2]. The prevalence of heart disease continues to rise, posing substantial risks to individuals, families, and society at large. This escalating incidence also contributes to the increasing burden on healthcare resources and highlights the critical need for effective detection and intervention strategies.

Early detection of heart disease is crucial to reducing both mortality and morbidity rates. However, conventional diagnostic methods face several limitations, such as reliance on subjective medical interpretation, lengthy procedures, and high costs. These methods often result in late-stage diagnoses, which increase the risk of severe complications and death. Therefore, there is a pressing need for alternative, cost-effective, and efficient diagnostic methods to improve early detection [3], [4].

Recent advancements in artificial intelligence (AI) and machine learning (ML) offer promising solutions to enhance early diagnosis, including the prediction of heart disease risk. Among the various machine learning algorithms, Support Vector Machine (SVM) has emerged as a highly effective tool for disease classification due to its ability to handle high-dimensional data and perform well with non-linear separability. SVM has been successfully applied in healthcare, particularly for heart disease prediction, by using clinical data such as age, blood pressure, cholesterol levels, and medical history to classify individuals into risk categories [5], [6], [7].

Although numerous studies have developed prediction models for heart disease using machine learning algorithms (e.g., SVM, Random Forest, Neural Networks), significant gaps remain that impede their applicability in Indonesian clinical contexts. First, many investigations rely on datasets drawn from Western or East Asian populations; these datasets may not reflect Indonesia's unique demographic, lifestyle, and clinical-profile characteristics [8]. Second, a number of the existing studies focus primarily on performance metrics like accuracy, but neglect to explicitly address class-imbalance issues or apply rigorous external validation, which are critical for ensuring model generalizability in real-world settings. For instance, class imbalance has been identified as a recurring challenge in machine-learning for cardiovascular risk prediction [9]. Third, the literature shows limited examination of how these models perform in Indonesian hospital settings where data access, electronic health record integration, and clinical process differences pose additional barriers. Consequently, despite promising numeric results, the readiness of these predictive models for real-world deployment remains questionable.

Based on these gaps, this study aims to develop an SVM-based heart disease risk prediction model using Indonesian clinical data, explicitly address class imbalance (e.g., via oversampling techniques), and conduct external validation in a hospital setting, thereby enhancing both accuracy and practical applicability. Through data preprocessing, feature selection, and model optimization, this research contributes to the growing body of knowledge on machine learning applications in healthcare, with a particular focus on heart disease risk prediction for the Indonesian context [5], [8].

2. RESEARCH METHOD

This study utilizes the Support Vector Machine (SVM) algorithm for heart disease risk prediction. SVM is a supervised machine learning technique commonly used for classification tasks, particularly effective in handling high-dimensional and non-linear data. Additionally, we discuss the potential use of other models, such as Random Forest (RF) and Neural Networks (NN), for comparison.

The Support Vector Machine (SVM) seeks to find a hyperplane that maximizes the margin between two classes in the dataset. The mathematical formulation of the SVM is as follows:

For a dataset with Ndata points $\{(x_1, y_1), (x_2, y_2), \dots, (x_N, y_N)\}$, where $x_i \in \mathbb{R}^n$are the feature vectors and $y_i \in \{-1,1\}$are the labels (class labels), the goal is to find a hyperplane $w \cdot x + b = 0$that separates the data points of one class from the other with the maximum margin.

The optimization problem for SVM is shown in the following equation (1):

$$\min_{w,b} \frac{1}{2} \parallel w \parallel^2 \tag{1}$$

subject to the constraints:

$$y_i(w \cdot x_i + b) \geq 1, i = 1, \dots, N \tag{2}$$

where:

w is the weight vector, which determines the orientation of the hyperplane.

b is the bias term, which determines the offset of the hyperplane from the origin.

$\parallel w \parallel$ represents the norm of the weight vector, and the goal is to minimize this to maximize the margin between the classes.

To handle non-linear separability, we use the kernel trick, which maps the original data into a higher-dimensional space where it becomes linearly separable. Common kernels include:

Linear kernel: $K(x, x') = x \cdot x'$

Polynomial kernel: $K(x, x') = (x \cdot x' + 1)^d$

Radial Basis Function (RBF) kernel: $K(x, x') = e^{-\gamma \parallel x - x' \parallel^2}$, where $\gamma$is a parameter controlling the width of the kernel.

Random Forest (RF) is an ensemble learning method based on decision trees. It constructs multiple decision trees during training and outputs the class that is the mode of the classes predicted by individual trees. The basic idea is to train multiple decision trees on different subsets of the data and combine their predictions to improve overall performance.

Each decision tree is built by: Bootstrapping: Randomly sampling data points with replacement to create multiple training datasets. Feature Randomization: At each split, a random subset of features is chosen to determine the best split, reducing correlation between trees.

Mathematically, the output of the Random Forest classifier shown in the following equation (3):

$$\hat{y} = \text{mode}(f_1(x), f_2(x), \dots, f_T(x)) \tag{3}$$

where:
$f_t(x)$ is the prediction of the t-th tree for input x.
T is the number of trees in the forest.

This ensemble approach helps reduce overfitting, which is a common issue with individual decision trees, and improves generalization to unseen data. Neural Networks (NN) are a class of models inspired by biological neural networks. They consist of layers of interconnected nodes (neurons), where each neuron applies a non-linear activation function to the weighted sum of inputs from the previous layer. For a multi-layer perceptron (MLP) with one hidden layer, the forward pass is as follows:

Input layer to hidden layer:

$$h_j = \sigma(\sum_i w_{ij}x_i + b_j) \tag{4}$$

where:
h_j is the output of the j-th hidden neuron.
w_{ij} is the weight between the i-th input and the j-th hidden neuron.
x_i is the i-th input feature.
b_j is the bias term for the hidden neuron.
σ is the activation function (commonly sigmoid $\sigma(x) = 1/(1 + e^{-x})$or ReLU $\sigma(x) = \max(0, x)$).

Hidden layer to output layer:

$$y = \sigma(\sum_j w_{oj}h_j + b_o) \tag{5}$$

where:
y is the final output (predicted class label).
w_{oj} is the weight from the j-th hidden neuron to the output.
b_o is the bias term for the output neuron.
σ is the output activation function (typically a softmax for classification tasks).

Dataset Collection

The dataset used in this study consists of 303 patient records, each containing 14 clinical attributes such as age, blood pressure, cholesterol levels, chest pain type, fasting blood sugar, and maximum heart rate achieved during exercise. These records were sourced from local Indonesian medical data, ensuring that the model is tailored to the demographic and clinical characteristics specific to Indonesia. This is a significant strength, as many existing heart disease prediction models rely on datasets from other regions, which may not accurately reflect the health characteristics of the Indonesian population.

However, the dataset's sample size of 303 may be insufficient to fully capture the diversity of heart disease risk factors, and potential biases may exist due to missing data or sampling bias if the dataset is derived from a single institution. To improve the generalizability of the model, future research should aim to collect a larger, more diverse dataset from multiple hospitals across Indonesia or other regions.

Data Preprocessing

Before training the model, it is crucial to ensure the quality and consistency of the dataset. The data preprocessing steps involved are as follows:

1. Handling Missing Values: Missing data is common in medical datasets. To maintain the integrity of the model, missing values are handled in two ways:
   a) For numerical attributes, missing values are imputed using the mean or median of the attribute, depending on the distribution of the data. The median is particularly useful when the data is skewed, as it is less affected by outliers.

b) For categorical attributes, missing values are imputed using the mode (the most frequent value) to prevent bias towards one class. This method ensures that the imputation is realistic and consistent with the existing data patterns.

2. Duplicate Removal: To avoid redundancy and ensure that each data point contributes unique information, duplicate records are identified and removed using standard data cleaning techniques. This step is essential to ensure the integrity of the model training process.
3. Normalization: SVM is sensitive to the scale of input features. To mitigate the risk of features with larger ranges dominating the learning process, Min-Max Scaling is applied to all numerical values. This scaling technique normalizes each feature to a [0,1] range. The formula used for Min-Max normalization is shown on (6):

$$X_{\text{normalized}} = \frac{X - X_{\min}}{X_{\max} - X_{\min}} \tag{6}$$

where:
X is the original feature value,
$X_{\min}$and $X_{\max}$are the minimum and maximum values of the feature.
This normalization ensures that all numerical features, such as cholesterol levels, blood pressure, and heart rate, are treated equally and have no disproportionate influence on the model.

4. Categorical Encoding: Some of the dataset's attributes, such as chest pain type and fasting blood sugar, are categorical in nature. These categorical variables are transformed into a numerical format using One-Hot Encoding, a common technique for converting categorical variables into a form that can be processed by machine learning algorithms. The One-Hot Encoding method creates a binary column for each category within the variable, where a "1" indicates the presence of that category, and "0" indicates its absence.

For example, if a variable chest pain type has three possible categoriesypical angina, atypical angina, and non-anginal pain One-Hot Encoding will create three new binary columns. Each column represents one of the categories, and each record will have a "1" in the column corresponding to its category and "0" in the others. This approach ensures that categorical data is appropriately incorporated into the SVM model, which requires numerical inputs.

Feature Selection

Feature selection is crucial for improving model performance by reducing dimensionality and avoiding overfitting. Two primary methods are used for feature selection:

1. Pearson Correlation: This method measures the linear relationship between each attribute and the target variable. Features with high correlation values are considered to be more relevant for the model [9].
2. Recursive Feature Elimination (RFE): RFE is used to recursively remove the least important features and build a model using the remaining features. This helps in identifying the most significant attributes influencing heart disease risk prediction [10].

Data Balancing

A common issue in healthcare datasets is class imbalance, where the number of instances in one class (e.g., patients at risk for heart disease) is significantly lower than the other class (e.g., healthy patients). To address this, the Synthetic Minority Over-sampling Technique (SMOTE) is applied to balance the class distribution. SMOTE works by generating synthetic samples for the minority class, ensuring that the model has a more balanced dataset to learn from, which improves the prediction accuracy for the minority class [11], [12].

SVM Model Training

Once the data is preprocessed and balanced, the SVM model is trained using the following steps:

1. Kernel Selection: Two types of kernels are tested in this study: the linear kernel, which is suitable for linearly separable data, and the Radial Basis Function (RBF) kernel, which is used for non-linear separation of data. The choice of kernel is crucial for improving the model's accuracy and generalization [13].
2. Hyperparameter Tuning: To optimize the SVM model's performance, hyperparameters such as the regularization parameter (C) and the kernel-specific parameters (e.g., gamma for RBF) are tuned using Grid Search and Particle Swarm Optimization (PSO). Grid Search involves exhaustively searching over a specified parameter grid, while PSO is a heuristic optimization technique that simulates the social behavior of birds or fish to find optimal solutions faster [14], [15].
3. Model Training: After selecting the kernel and tuning the hyperparameters, the model is trained on the preprocessed and balanced dataset. The SVM algorithm constructs a hyperplane that maximizes the margin between the "at risk" and "no risk" classes [16].

Model Evaluation

Model performance is evaluated using several metrics to ensure a comprehensive assessment of its predictive ability:

1. Accuracy: The proportion of correctly predicted instances (both positive and negative) over the total number of instances in the test set.
2. Precision: The proportion of true positive predictions relative to the total positive predictions made by the model. This metric is important when minimizing false positives is critical [17].
3. Recall (Sensitivity): The proportion of actual positive instances that are correctly identified by the model. Recall is crucial in medical applications to ensure that the model detects as many positive cases as possible [18].
4. F1-Score: The harmonic means of precision and recall, providing a balance between the two. A high F1-score indicates a good balance between precision and recall, which is particularly useful in imbalanced datasets [19].
5. Cross-Validation: To assess the model's robustness, k-fold cross-validation is used. This technique splits the data into k subsets and iteratively trains and evaluates the model on different subsets to reduce variance and avoid overfitting [20].

Model Optimization

In case the initial model performance is suboptimal, hyperparameter optimization is performed using Grid Search and PSO. This helps identify the best combination of parameters (C, gamma, and kernel type) that minimizes the classification error and maximizes the prediction accuracy [21]. Further optimization could involve exploring ensemble methods to combine multiple models and improve overall performance [22].

3. RESULT AND ANALYSIS

After performing the necessary data preprocessing steps, the dataset was cleaned and normalized. Missing values were imputed using the mean or mode for numerical and categorical features, respectively. Duplicate records were removed to maintain the integrity of the dataset. Feature normalization using Min-Max scaling was applied to all numerical attributes to ensure that no feature disproportionately affected the model's performance due to its scale. Categorical features were encoded using One-Hot Encoding to enable processing by the Support Vector Machine (SVM) model.

Feature selection was conducted using two methods: Pearson Correlation and Recursive Feature Elimination (RFE). Both methods identified key features critical to predicting heart disease risk, such as age, cholesterol level, and exercise-induced angina. These features were deemed the most influential for the model, which helped in reducing the dimensionality of the dataset and improving the model's efficiency.Given the imbalanced nature of the dataset, the Synthetic Minority Over-sampling Technique (SMOTE) was employed to balance the classes by generating synthetic data points for the minority class (patients at risk of heart disease). This ensured that the SVM algorithm was exposed to a balanced dataset, thus preventing the model from being biased towards the majority class.

Two different kernels were tested: a linear kernel and a Radial Basis Function (RBF) kernel. After training the initial models, hyperparameter tuning was performed using Grid Search and Particle Swarm Optimization (PSO) to find the optimal parameter combinations. This step ensured the model was fine-tuned for maximum predictive performance.The model's performance was evaluated using accuracy, precision, recall, and F1-score metrics, which were computed both for the training set and the test set.

The model achieved the following metrics on the training set as shown in table 1.

Table 1. Set Metrics

Metric	Value	Description
Accuracy	0.9050	Indicates that 90.50% of predictions on the training data are correct (both positive and negative).
Precision	0.8980	Of all predictions of patients at risk of heart disease, 89.80% are actually at risk.
Recall	0.9072	Of all patients who are actually at risk of heart disease, 90.72% were successfully detected by the model.
F1-Score	0.9026	The harmonic means between precision and recall, indicating the balance between the accuracy and completeness of the model.

On the test set, the model achieved the following results:

Table 2. Test Set Metrics

Metrics	Value	Explanation
Accuracy	0.9200	The model correctly predicted 92% of the test data.
Precision	0.9000	Of all predictions of patients at risk of heart disease, 90% were actually at risk.
Recall	0.9643	Of all patients who are actually at risk of heart disease, 96.43% were successfully detected.
F1-Score	0.9310	Shows an excellent balance between precision and recall on the test data.

These results demonstrate that the optimized SVM model is highly effective in predicting heart disease risk, with strong performance across all evaluation metrics.

DISCUSSION

In the Support Vector Machine (SVM) model, the kernel transformation plays a crucial role in improving classification performance, especially in cases where the data is not linearly separable in the original input space. The kernel trick allows us to map the data from its original feature space into a higher-dimensional feature space, where the classes may become linearly separable. This transformation is key to improving the model's ability to classify complex, non-linearly separable data. By using the kernel trick, data transformed into a higher-dimensional space where the classes may become more separable.

The results of this study suggest that the Support Vector Machine (SVM) algorithm, when optimized with the Radial Basis Function (RBF) kernel and Particle Swarm Optimization (PSO), is highly effective for heart disease risk prediction based on patient health data. The model's accuracy of 92.00% on the test set indicates its potential for real-world applications in clinical settings. The use of SMOTE to address the class imbalance significantly improved the model's recall (96.43%), which is particularly critical for medical applications where detecting true positives is crucial.

The performance of this SVM model is consistent with previous studies in the field of heart disease prediction. For instance, Maulana et al. [3] reported an accuracy of 89.4% for SVM in heart disease prediction, while Natsir et al. [4] achieved an accuracy of 91.5%. However, unlike these studies, this research incorporated SMOTE, RFE, and PSO, which likely contributed to the higher performance observed in this study. The use of RBF as the kernel function also allowed for better handling of non-linear relationships in the data, which was not addressed in many previous studies.

The primary reason the RBF kernel outperformed the linear and polynomial kernels in this case is its ability to handle non-linearly separable data. In medical datasets, like the one used in this study, the relationship between clinical attributes (e.g., cholesterol, blood pressure, heart rate) and the target variable (heart disease risk) is often complex and not easily separable with a straight line or a simple polynomial boundary. In this study, the optimal gamma value was found through hyperparameter tuning (using methods like Grid Search or Particle Swarm Optimization). This allowed the RBF kernel to adapt to the local structure of the data by adjusting the influence of each data point. By fine-tuning gamma, the SVM was able to create a decision boundary that effectively separated the classes, leading to the high accuracy and recall rates (96.43%).

The preprocessing steps, including missing value imputation, duplicate removal, and normalization, were essential in ensuring that the data was clean and consistent. These steps prevented potential biases that could have impacted the model's performance. Feature selection, particularly through Pearson Correlation and RFE, helped focus the model on the most relevant attributes, improving both model efficiency and interpretability. This supports the findings of Hidayat et al. [5], who emphasized the importance of feature selection in enhancing model accuracy, particularly in healthcare applications.

The SMOTE technique played a pivotal role in addressing the class imbalance in the dataset. By generating synthetic samples for the minority class, SMOTE ensured that the model was not biased toward the majority class, which is a common issue in healthcare datasets. The resulting balance between classes helped improve both precision (90%) and recall (96.43%), making the model more reliable in identifying at-risk patients. This finding is in line with Chawla et al. [6] and Gupta et al. [7], who have shown that SMOTE is an effective technique for improving the performance of classifiers in imbalanced datasets.

The promising performance of the SVM model suggests its potential integration into Clinical Decision Support Systems (CDSS), where it could assist healthcare providers in early heart disease detection. By incorporating this model into clinical workflows, physicians could obtain real-time risk assessments for patients, allowing for timely interventions. Future research should focus on expanding the dataset to include a broader range of demographic and clinical data to enhance the model's generalizability. Additionally, exploring other optimization techniques, such as Genetic Algorithms or Bayesian Optimization, could further improve the model's performance. The integration of ensemble learning methods could also be explored to increase the robustness of the model.

Despite its high performance, this study has several limitations. First, the dataset used is relatively small and may not capture all the possible variations in heart disease risk factors. Future studies should validate the model using data from multiple institutions and regions to assess its generalizability. Second, although the model performs well, its interpretability could be enhanced by using explainable AI techniques, such as SHAP values, to understand which features most influence the model's predictions.

4. CONCLUSION

The study's results demonstrate that the SVM-based heart disease risk prediction model is not only a statistical success but also has clear clinical utility. The strong performance metrics (92% accuracy, 96% recall) suggest that the model can be trusted in real-world clinical settings, provided it is integrated thoughtfully into hospital systems and workflows. Scalability: The model has the potential to scale across multiple healthcare settings, from large hospitals to remote clinics, without requiring significant resources beyond the necessary computational infrastructure for data processing and model inference. Clinical Impact: The improvements in recall (96%) indicate that the model is highly sensitive in identifying patients who are truly at risk, which is particularly valuable in preventive healthcare. This can significantly reduce false negatives, ensuring that high-risk patients do not slip through the cracks. Implementation Challenges: While the model shows promise, successful integration into hospital systems will require overcoming several challenges: Data Compatibility: Ensuring the model can work seamlessly with existing EHR systems and handle data from various hospital departments. Clinical Validation: The model will need to undergo continuous validation with real-world clinical data to ensure its ongoing reliability and relevance in different healthcare contexts.

However, there are some limitations to the study. The dataset used was relatively small and may not fully represent the diverse range of heart disease risk factors across the broader population. Additionally, the model was not externally validated using data from other medical institutions, which is a critical step to assess its generalizability. Future research should aim to expand the dataset by incorporating real patient data from multiple hospitals and regions, ensuring a more diverse sample. Further improvements could involve exploring other optimization techniques, such as Genetic Algorithms or Bayesian Optimization, and integrating ensemble learning methods to increase the model's robustness.

In conclusion, the SVM-based heart disease risk prediction model developed in this study holds significant promise for early detection and intervention in heart disease. If integrated into clinical practice, it could assist healthcare providers in making faster, data-driven decisions, improving patient outcomes, and reducing the burden of heart disease in Indonesia.

.

5. REFERENCES

[1] Arifuddin, A., Buana, G. S., Vinarti, R. A., & Djunaidy, A. (2024). Performance Comparison Comparison of of Decision Decision Tree Tree and and Support Support Vector Vector Performance Machine Algorithms Algorithms for for Heart Heart Failure Failure Prediction Prediction Machi. *Procedia Computer Science*, *234*, 628–636. https://doi.org/10.1016/j.procs.2024.03.048

[2] A. S. Prabowo and F. I. Kurniadi, "Improved parameter selection in support vector machines with a grid search method," J. Syst. Comput. Sci., vol. 24, no. 2, pp. 132-141, 2015.

[3] Ben-david, S. (2014). *Understanding Machine Learning: From Theory to Algorithms*.

[4] Chawla, I., Karthikeyan, L., & Mishra, A. K. (2020). A review of remote sensing applications for water security: Quantity, quality, and extremes. *Journal of Hydrology*, *585*(March), 124826. https://doi.org/10.1016/j.jhydrol.2020.124826

[5] Davis, J., & Goadrich, M. (2006). *The Relationship Between Precision-Recall and ROC Curves*. 233–240.

[6] Du, G., et al., "Integrated support vector machine with improved PSO: a two stage adaptive PSO algorithm," Computers & Electrical Engineering, 2025. https://doi.org/10.1016/j.cie.2025.111300

[7] F. M. Natsir, R. Y. Bakti, and T. Wahyuni. (2024). *Arus Jurnal Sains dan Teknologi (AJST) Analisis Deteksi Dini Penyakit Jantung dengan Pendekatan Support Vector Machine pada Data Pasien*. *2*(2).

[8] Kumar, R. et.al (2025). *A comprehensive review of machine learning for heart disease prediction: challenges, trends, ethical considerations, and future directions*. Front. Artif. Intell. Sec. Medicine and Public Health. Volume 8 - 2025 | https://doi.org/10.3389/frai.2025.1583459

[9] M. Altalhan, A. Algarni and M. Turki-Hadj Alouane, "Imbalanced Data Problem in Machine Learning: A Review," in *IEEE Access*, vol. 13, pp. 13686-13699, 2025, doi: 10.1109/ACCESS.2025.3531662.

[10] Hidayat, R., Sy, Y. S., Sujana, T., Husnah, M., & Saputra, H. T. (2024). *Implementasi Machine Learning Untuk Prediksi Penyakit Jantung Menggunakan Algoritma Support Vector Machine*. *5*(2), 161–168.

[11] Kohavi, R., & Edu, S. (1993). *A Study of Cross-Validation and Bootstrap for Accuracy Estimation and Model Selection*. 1137–1143.

[12] Lee, D. D., Laboratories, B., Hill, M., & Ý, H. S. S. (n.d.). *Algorithms for Non-negative Matrix Factorization*. *1*.

[13] M. R. Maulana, A. Sucipto, and H. M. Mulyo. (2024). Optimisasi Parameter Support Vector Machine Dengan Particle Swarm Optimization Untuk Peningkatan Klasifikasi Diabetes. Jurnal Informatika Teknologi dan Sains (JINTEKS) E-ISSN. 802–812.

[14] No, V., Hal, A., Agus, I. M., Gunawan, O., Ayu, I. D., Saraswati, I., Gede, I. D., Agung, R., & Eka, I. P. (2023). *Klasifikasi Penyakit Jantung Menggunakan Algoritma Decision Tree Series C4 . 5 Dengan Rapidminer*. *5*(2), 73–83.

[15] Prabowo, A. S., & Kurniadi, F. I. (2023). *Analisis Perbandingan Kinerja Algoritma Klasifikasi dalam Mendeteksi Penyakit Jantung*.

[16] R. K. Gupta, V. Gupta, and A. K. Sharma, "A study of SMOTE and its application to imbalanced data sets," International Journal of Computer Applications, vol. 8, no. 8, pp. 39-45, 2010.

[17] S. Simatupang, R. Ramadhansyah, R. Tumanggor, E. P. Tan, and S. A. Fajar, "Prediction of heart disease risk based on patient health history using the Support Vector Machine (SVM) algorithm," Zero: Jurnal Sains, Matematika dan Terapan, vol. 10, no. 1, pp. 1-9, 2024.

[18] S. Sitanggang, N. Nicholas, V. Wilson, A. R. A. Sinaga, and A. D. Simanjuntak, "Implementasi data mining untuk memprediksi penyakit jantung menggunakan metode k-nearest neighbor dan logistic regression," Jurnal Tekinkom (Teknik Informasi dan Komputer), vol. 5, no. 2, pp. 493-499, Dec. 2022. https://doi.org/10.37600/tekinkom.v5i2.698.

[19] Syaidah, I. B., Surono, S., & Goh, K. W. (2024). *Dynamic Weighted Particle Swarm Optimization - Support Vector Machine Optimization in Recursive Feature Elimination Feature Selection*. *23*(3), 627–640. https://doi.org/10.30812/matrik.v23i3.3963

[20] Tohka, J., & Gils, M. Van. (2021). Evaluation of machine learning algorithms for health and wellness applications : A tutorial. *Computers in Biology and Medicine*, *132*(February), 104324. https://doi.org/10.1016/j.compbiomed.2021.104324

[21] Vladimir N. Vapnik, " The Nature of Statistical Learning Theory". New York: Springer New York, 2021. https://doi.org/10.1007/978-1-4757-3264-1

[22] V. Vapnik, The Nature of Statistical Learning Theory, 2nd ed. New York, NY, USA: Springer-Verlag, 2000