Indonesian Journal of Modern Science and Technology Modern (IJMST)
Vol. 1, No. 1, January 2025, pp. 9-15

Classification of Heart (Cardiovascular) Disease using the SVM Method

Minhajul Abidin [1,*], Misbahul Munzir [2], Adi Imantoyo [3], Nuraqilla Waidha Bintang Grendis [4], Ahmad Syahrul Hadi San [5], Ahmed A. Mostfa [6], Furizal [7], Abdel-Nasser Sharkawy [8]

[1,2] Department of Computer Science, Universitas Qamarul Huda Badaruddin, 83562, Central Lombok, Indonesia
[3] Department of International Business, Hankuk University of Foreign Studies, 02450, Seoul, Korea
[4] Department of Information Technology, Universitas Qamarul Huda Badaruddin, 83562, Central Lombok, Indonesia
[5] Bachelor of Information Technology, Universiti Tun Hussein Onn Malaysia, 86400, Johor, Malaysia
[6] Department of Computer Science, College of Education, University of Al Hamdaniya, Mosul, Iraq
[7] Department of Research and Development, Peneliti Teknologi Teknik Indonesia, Sleman 55281, Indonesia
[8] Mechatronics Engineering, Mechanical Engineering Department, South Valley University, Qena, 83523, Egypt; Mechanical Engineering Department, Fahad Bin Sultan University, Tabuk 47721, Saudi Arabia

ARTICLE INFO

Article history:

Received November 23, 2024
Revised November 28, 2024
Published January 12, 2025

Keywords:

cardiovascular; classification; gridsearchcv; heart disease prediction; support vector machine;

ABSTRACT

Cardiovascular disease is one of the leading causes of death worldwide, with a high mortality rate, especially in developing countries like Indonesia. This highlights the importance of developing systems to identify and detect heart disease at an early stage. In this study, the Support Vector Machine (SVM) algorithm was used to classify cardiovascular diseases by utilizing a dataset consisting of 303 patient records obtained from Kaggle. The dataset was divided into 70% for training and 30% for testing. Before optimization using GridSearchCV, the SVM model achieved an accuracy of 79%, precision of 79%, recall of 73%, and F1-score of 76%. However, after adjusting the hyperparameters with GridSearchCV, the model's accuracy slightly decreased to 77%, with precision remaining at 79%, recall dropping to 66%, and F1-score at 72%. Despite this decline in performance after optimization, the results indicate that although SVM has potential for classifying heart disease, its performance is highly influenced by data quality and the selection of appropriate hyperparameters. Even with the performance decrease post-optimization, the model still provides useful predictions, showing consistent results and a proportional class distribution.

Corresponding Author:

Minhajul Abidin, Department of Computer Science, Universitas Qamarul Huda Badaruddin, 83562, Central Lombok, Indonesia
Email: minhajulabidin333@gmail.com

1. INTRODUCTION

The heart is a vital organ of the human body, responsible for pumping blood to every part of the anatomy. If it fails to function properly, the brain and other organs will cease to operate, leading to death within a matter of minutes. Lifestyle changes, including high-stress work environments and unhealthy eating habits, contribute to the rising prevalence of heart-related diseases [1]. Cardiovascular disease is a non-communicable disease that stands as the leading cause of death worldwide. Around 80% of fatalities caused by this condition take place in low- and middle-income nations, including Indonesia. This category of diseases encompasses disorders of the heart and blood vessels, such as coronary heart disease, heart failure, hypertension, and stroke. In addition to causing fatalities, cardiovascular diseases can lead to disabilities and significantly reduce a person's quality

of life [2]. Cardiovascular disease (CVD) is a term that encompasses a range of disorders affecting the heart, coronary arteries, and blood vessels, including veins, arteries, and capillaries. According to data from the World Health Organization (WHO), cardiovascular diseases are responsible for approximately 17.9 million deaths annually, accounting for about 32% of total yearly fatalities [3]. With the increasing number of cases each year, a classification system is needed to provide information about the disease and enable early detection of heart attacks in individuals [4]. Advancements in science and information technology have led to the emergence of a new field in computer science known as data mining. Classification in data mining involves grouping objects into predefined categories. This study utilizes a classification algorithm, Support Vector Machine (SVM), to evaluate its performance when applied to a small-scale dataset[5]. The Support Vector Machine (SVM) algorithm offers several advantages, including its effectiveness for high-dimensional data, strong performance on both linear and non-linear data, the ability to measure distances efficiently, which accelerates computation, and flexibility through the use of kernels. However, it also has some limitations. SVM is less efficient for large-scale problems, especially when processing a large number of samples. It is sensitive to hyperparameter selection, performs poorly with noisy or overlapping data, and struggles with non-linear data without an appropriate kernel [6].

Several studies have utilized the SVM method, including a web-based heart disease prediction system using SVM and the Streamlet framework. The findings revealed that the system could classify heart disease with an accuracy of 85%. The hyperparameters used were gamma 0.01, Const 0.1, and a polynomial kernel, as the dataset was non-linear. The precision achieved was 78%, while the recall reached 89%. The training score was recorded at 85%, and the testing score at 87%. With a minimal difference between training and testing scores, the model was determined to neither overfit nor underfit [7]. The implementation of the Support Vector Machine (SVM) method enhanced by Particle Swarm Optimization (PSO) for heart disease prediction yielded notable results. Initial experiments showed that the SVM method achieved an accuracy of 79.55% with an AUC value of 0.850. After adjusting the parameters C and epsilon, the best accuracy improved to 81.85%, with an AUC value of 0.899. In a second experiment, the SVM method integrated with PSO demonstrated an initial accuracy of 82.50% and an AUC value of 0.825. Following adjustments to parameters such as C, epsilon, and population size, the best accuracy achieved by the SVM-PSO method was 88.61%, with an AUC value of 0.919 [8]. The study on heart failure classification using the Support Vector Machine (SVM) method concluded that SVM is quite effective for classifying heart failure diseases. In the trials, the highest accuracy achieved was 86.92%, using a linear kernel and cost values of 1, 10, and 100. In the best model with a cost of 100, the test results showed a precision of 88.68%, an F1-score of 88.26%, an accuracy of 86.41%, and a recall of 87.85% [9]. In a study comparing the SVM (Support Vector Machine) algorithm with KNN (K-Nearest Neighbors), the classification results demonstrated that the SVM method performed exceptionally well in classifying heart disease. SVM achieved an accuracy of 84.61% without normalization and 90.10% with normalization. In contrast, the KNN algorithm showed an accuracy of 64.83% without normalization and 81.31% with normalization [10]. Another study comparing SVM with Random Forest for breast cancer classification found that SVM achieved an accuracy of 95%, outperforming Random Forest, which had an accuracy of 90% [11]. By analyzing prior studies, it can be concluded that SVM is a suitable method for classification tasks based on its successful implementation in previous research.

2. METHODS

2.1. Data collection

This dataset was sourced from the Kaggle website and contains 303 cardiovascular patient records in a preprocessed CSV format. It includes 14 attributes used for classifying cardiovascular diseases, such as age, sex, cp, trestbps, chol, fbs, restecg, thalach, exang, oldpeak, slope, ca, thal, and target.
Dataset link: https://www.kaggle.com/datasets/rashikrahmanpritom/heart-attack-analysis-prediction-dataset.

2.2. Data Preparation (Dataset Division)

The dataset underwent preprocessing, including the Famhist attribute, which indicates a family history of heart disease, with a value of 0 representing no history (healthy) and 1 indicating a history (unhealthy). The Chd attribute reflects the disease response, where 0 represents inactive and 1 represents active cases. The data was then classified using an SVM Classifier model to predict cardiovascular disease. Subsequently, the dataset was split into two parts using the percentage split method, allocating 70% for training and 30% for testing the Support Vector Machine algorithm [12].

2.3. Data Pre-Processing

This stage involves checking for duplicate rows or missing values in the columns, followed by data normalization to reduce redundancy and ensure data consistency. Additionally, to ensure a valid evaluation process, the data is divided into 2 categories: training, and testing. Duplicate rows or missing values in specific columns are identified during this step. Subsequently, normalization is applied to minimize redundancy and maintain consistent relationships between the data. To further guarantee accurate model evaluation, the dataset is split into 2 parts: training data, testing data, according to predefined proportions.

Table 1. Data Split

No	Item	Percentage (%)
1	Train	70
2	Test	30

Data preprocessing is employed to structure unorganized datasets, making them suitable for the intended requirements and ready for subsequent processing stages [13].

2.4. Model Training (Modelling)

Support Vector Machine is an algorithm designed to identify the optimal hyperplane that separates two classes within the input space. The next step involves selecting an appropriate kernel, such as linear, polynomial, or RBF. For the RBF kernel, parameters like C and Gamma should be tuned to optimize performance. Cost (C) is a parameter in SVM that controls the extent to which the model penalizes misclassification errors in the training data [14]. A kernel is a component responsible for mapping low-dimensional input into a higher-dimensional space [15]. This study also utilizes the GridSearchCV module. GridSearchCV, a feature of the scikit-learn library, systematically and automatically validates various models with different combinations of hyperparameters to identify the optimal configuration [16].

Table 2. Prediction Parameters included in the Model

No	Parameter	Value
1	C	0.1; 1; 10; 100; 1000
2	Gamma	1; 0.1; 0.01; 0.001; 0.0001
3	Kernel	‘rbf’

2.5. Model Testing

In this study, 30% of the total data was allocated as test data to evaluate the performance of the developed model. The classification model was assessed using four commonly used metrics: accuracy, precision, recall, and F1-score. These metrics are calculated based on the Confusion Matrix, which consists of four key terms: True Positive (TP), True Negative (TN), False Positive (FP), and False Negative (FN). The evaluation through the confusion matrix provides various metrics, including accuracy, precision, recall, and F1-score.

- Accuracy

  Accuracy is the ratio of the number of correct predictions to the total number of data points [17]. This metric is used to represent the model's overall correctness in performing classification. The formula for calculating accuracy is shown below.

$$Accuracy = \frac{\text{True Positive} + \text{True Negative}}{Total\ Prediction}$$

- Precision

Precision is the ratio that indicates the proportion of correctly predicted positive instances[18]. This metric measures the accuracy of the model's positive predictions relative to the expected data. The formula for calculating precision is shown in the following equation.

$$Precision = \frac{\text{True Positive}}{True\ Positive + False\ Positive}$$

- Recall
  Recall is the ratio of actual positive data that is correctly classified as positive[19]. The following equation is used to explain the formula for calculating recall.

$$Recall = \frac{\text{True Positive}}{True\ Positive + False\ Negative}$$

- F1 Score
  F1 Score is a metric that combines both precision and recall by calculating their harmonic mean[20]. The formula for calculating the F1 Score is shown in the following equation.

$$F1\ Score = \frac{\text{2 x Precision x Recall}}{Precision + Recall}$$

2.6. Evaluation of Results

Use the testing set to evaluate the classification results using the confusion matrix method. A confusion matrix is a table that shows the number of test data correctly classified and the number of test data misclassified. From the SVM algorithm, the results include a comparison of hyperparameter tuning with the matrix values.
The following is the processing structure in diagram form:

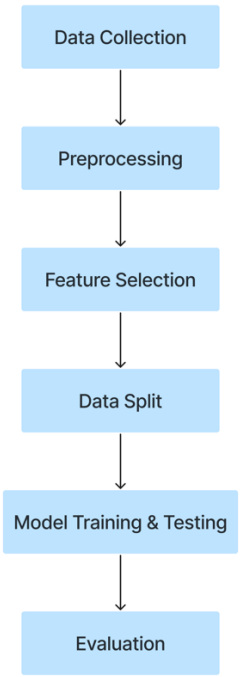

Fig. 2. Processing structure

3. RESULTS AND DISCUSSION

3.1. Model Training

Based on the testing results, it can be concluded that the model's performance before optimization through GridSearchCV as a hyperparameter tuning method shows an accuracy of 79%. This indicates that there are still some errors in the predictions/testing based on the available data. Additionally, the results from the confusion matrix before the testing process are shown in the table below.

Table 3. Results Model Training

No	Matrix	Value
1	Accuracy	0.79%
2	Precision	0.79%
3	Recall	0.73%
4	F1 Score	0.76%

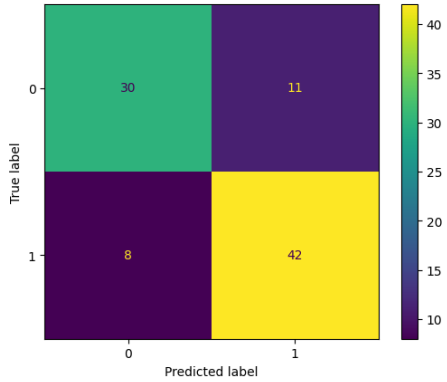

Fig 3. Confusion matrix Training Model

The model training results show an evaluation metric with Accuracy of 0.79%, which indicates the model has a low correct prediction rate over the entire data. The Precision of 0.79% indicates the model's ability to generate correct positive predictions, while the Recall of 0.73% indicates the model only captured 73% of all true positives. The F1 Score value of 0.76% reflects the balance between Precision and Recall. Overall, the performance of the model still needs to be improved, especially in reducing false negatives.

3.2. Testing

To enhance the model's performance during testing and improve the overall system quality, optimization was carried out using GridSearchCV as the hyperparameter tuning method. Through this hyperparameter tuning process, the tested parameter combinations yielded the best performance during the classification model development phase. After successfully identifying the optimal hyperparameters, the results are shown in the following table.

Table 4. Predicted Parameters to be Included in the Model

No	Parameter	Value
1	C	10
2	Gamma	0.01
3	Kernel	‘rbf’

Once the model is built using the specified parameters, the next step is to evaluate it according to the prepared testing scenario. The model evaluation results are then documented through calculations and visualizations of the Confusion Matrix diagram.

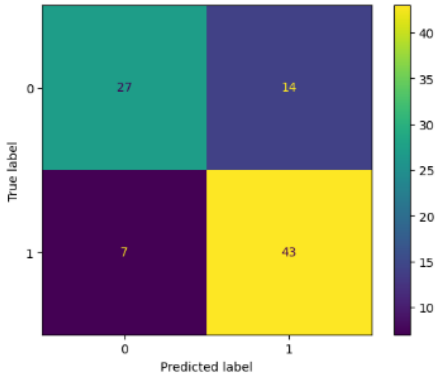

Fig 3. Confusion matrix Testing

Table 3. Results after testing

No	Matrix	Value
1	Accuracy	0.77%
2	Precision	0.79%
3	Recall	0.66%
4	F1 Score	0.72%

Based on the table above, it shows that the classification model using SVM for heart disease prediction experienced a relatively stable decrease in value (range), dropping from 79% to 77%. The performance metrics range from 0 to 1. However, it is worth noting that the model consistently produced reliable values for each performance metric. This indicates that the data used in the learning process was well-normalized and that the class distribution for each data type was proportionally organized. As seen in the image above, it can be concluded that the most accurate predictions are found in label 1, while the least accurate predictions are for label 0. This is due to the lower number of correct predictions for label 0 compared to the incorrect predictions. The analysis of the Confusion Matrix is supported by the color difference, with yellow indicating the best prediction performance.

4. CONCLUSION

Research on heart disease classification using the Support Vector Machine (SVM) method indicates that SVM can be used to predict heart disease with a certain level of accuracy, but the results obtained are not yet stable or optimal. Before optimization with GridSearchCV, the model showed an accuracy of 79%, with precision of 79%, recall of 73%, and an F1-score of 76%. After hyperparameter optimization, the model's accuracy slightly decreased to 77%, with precision of 79%, recall of 66%, and an F1-score of 72%. These results suggest that while the SVM method has potential for heart disease classification, its performance is highly influenced by the quality and quantity of data, as well as the selection of appropriate hyperparameters.

REFERENCES

[1] H. Hasanah and Nurmalitasari, "Perbandingan Tingkat Akurasi Algoritma Support Vector Machines (SVM) dan C45 dalam Prediksi Penyakit Jantung," *Pros. Semin. Nas. Teknol. dan Sains*, vol. 2, pp. 13–18, 2023.

[2] J. Teknik Elektro dan Komputasi *et al.*, "Implementasi Algoritma Naïve Bayes dan Support Vector Machine (SVM) Pada Klasifikasi Penyakit Kardiovaskular," *J. Tek. Elektro dan Komputasi*, vol. 4, no. 2, pp. 207–214, 2022, [Online]. Available: http://jurnal.unmuhjember.ac.id/index.php/ELKOM/article/view/7691

[3] S. N. N. Arif, A. M. Siregar, S. Faisal, and A. R. Juwita, "Klasifikasi Penyakit Serangan Jantung Menggunakan Metode Machine Learning K-Nearest Neighbors (KNN) dan Support Vector Machine (SVM)," *J. Media Inform. Budidarma*, vol. 8, no. 3, p. 1617, 2024, doi: 10.30865/mib.v8i3.7844.

[4] M. A. Bianto, K. Kusrini, and S. Sudarmawan, "Perancangan Sistem Klasifikasi Penyakit Jantung Mengunakan Naïve Bayes," *Creat. Inf. Technol. J.*, vol. 6, no. 1, p. 75, 2020, doi: 10.24076/citec.2019v6i1.231.

[5] S. Syarli and A. A. Muin, "Metode Naive Bayes Untuk Prediksi Kelulusan (Studi Kasus: Data Mahasiswa Baru

Perguruan Tinggi)," *J. Ilmu Komput.*, vol. 2, no. 1, pp. 22–26, 2018, [Online]. Available: https://fikom-unasman.ac.id/ejournal/index.php/jikom/article/view/26/17

[6] P. A. Octaviani, Yuciana Wilandari, and D. Ispriyanti, "Penerapan Metode Klasifikasi Support Vector Machine (SVM) pada Data Akreditasi Sekolah Dasar (SD) di Kabupaten Magelang," *J. Gaussian*, vol. 3, no. 8, pp. 811–820, 2014, [Online]. Available: http://download.portalgaruda.org/article.php?article=286497&val=4706&title=PENERAPAN METODE KLASIFIKASI SUPPORT VECTOR MACHINE (SVM) PADA DATA AKREDITASI SEKOLAH DASAR (SD) DI KABUPATEN MAGELANG

[7] A. Putranto, N. L. Azizah, and A. I. Ratna Ika, "Sistem Prediksi Penyakit Jantung Berbasis Web Menggunakan Metode SVM dan Framework Streamlit," *J. Penerapan Sist. Inf. (Komputer Manajemen)*, vol. 4, no. 2, pp. 442–452, 2023, [Online]. Available: https://archive.ics.uci.edu/ml/datasets/heart+disease

[8] S. Nurajizah, "Penerapan Metode Support Vector Machine Berbasis Particle Swarm Optimization Untuk Prediksi Penyakit Jantung," *J. Techno Nusa Mandiri*, vol. 10, no. 1, pp. 216–266, 2013, [Online]. Available: http://www.bsi.ac.id

[9] L. N. Farida and S. Bahri, "Komputika : Jurnal Sistem Komputer Klasifikasi Gagal Jantung Menggunakan Metode SVM (Support Vector Machine) Classification of Heart Failure using the SVM (Support Vector Machine) Method," vol. 13, pp. 0–7, 2025, doi: 10.34010/komputika.v13i2.11330.

[10] D. A. Anggoro and D. Permatasari, "Performance Comparison of the Kernels of Support Vector Machine Algorithm for Diabetes Mellitus Classification," *Int. J. Adv. Comput. Sci. Appl.*, vol. 14, no. 1, pp. 580–585, 2023, doi: 10.14569/IJACSA.2023.0140163.

[11] C. Aroef, Y. Rivan, and Z. Rustam, "Comparing random forest and support vector machines for breast cancer classification," *Telkomnika (Telecommunication Comput. Electron. Control.*, vol. 18, no. 2, pp. 815–821, 2020, doi: 10.12928/TELKOMNIKA.V18I2.14785.

[12] S. Adi and A. Wintarti, "Komparasi Metode Support Vector Machine (Svm), K-Nearest Neighbors (Knn), Dan Random Forest (Rf) Untuk Prediksi Penyakit Gagal Jantung," *MATHunesa J. Ilm. Mat.*, vol. 10, no. 2, pp. 258–268, 2022, doi: 10.26740/mathunesa.v10n2.p258-268.

[13] E. Suryati, Styawati, and A. A. Aldino, "Analisis Sentimen Transportasi Online Menggunakan Ekstraksi Fitur Model Word2vec Text Embedding Dan Algoritma Support Vector Machine (SVM)," *J. Teknol. Dan Sist. Inf.*, vol. 4, no. 1, pp. 96–106, 2023, [Online]. Available: https://doi.org/10.33365/jtsi.v4i1.2445

[14] V. Blanco, A. Japón, and J. Puerto, "A mathematical programming approach to SVM-based classification with label noise," *Comput. Ind. Eng.*, vol. 172, no. PA, p. 108611, 2022, doi: 10.1016/j.cie.2022.108611.

[15] A. Gilardi, R. Borgoni, and J. Mateu, "a Nonseparable First-Order Spatiotemporal Intensity for Events on Linear Networks: an Application To Ambulance Interventions," *Ann. Appl. Stat.*, vol. 18, no. 1, pp. 529–554, 2024, doi: 10.1214/23-AOAS1800.

[16] Z. Maisat, E. Darmawan, and A. Fauzan, "Implementasi Optimasi Hyperparameter GridSearchCV Pada Sistem Prediksi Serangan Jantung Menggunakan SVM Implementation of GridSearchCV Hyperparameter Optimization in Heart Attack Prediction System Using SVM," *Unipdu*, vol. 13, no. 1, pp. 8–15, 2023.

[17] D. Chicco and G. Jurman, "The advantages of the Matthews correlation coefficient (MCC) over F1 score and accuracy in binary classification evaluation," *BMC Genomics*, vol. 21, no. 1, pp. 1–13, 2020, doi: 10.1186/s12864-019-6413-7.

[18] L. Qadrini, "Undersampling dan K-Fold Random Forest Untuk Klasifikasi Kelas Tidak Seimbang," *Build. Informatics, Technol. Sci.*, vol. 4, no. 4, pp. 1967–1974, 2023, doi: 10.47065/bits.v4i4.3141.

[19] C. A. Choquette-Choo, F. Tramèr, N. Carlini, and N. Papernot, "Label-Only Membership Inference Attacks," *Proc. Mach. Learn. Res.*, vol. 139, pp. 1964–1974, 2021.

[20] M. Shorfuzzaman and M. S. Hossain, "MetaCOVID: A Siamese neural network framework with contrastive loss for n-shot diagnosis of COVID-19 patients," *Pattern Recognit.*, vol. 113, p. 107700, 2021, doi: 10.1016/j.patcog.2020.107700.