Institute of Research and Publication Indonesia (IRPI)
Public Research Journal of Engineering, Data Technology and Computer Science
Journal Homepage: https://journal.irpi.or.id/index.php/predatecs
Vol. 1 Iss. 1 July 2023, pp: 20-25
ISSN(P): 3024-921X | ISSN(E): 3024-8043

Implementation of C4.5 and Support Vector Machine (SVM) Algorithm for Classification of Coronary Heart Disease

Muhammad Ridho Anugrah[1], Nanda Nazira[2], Nola Ardelia Al-Qadr[3], Nurul Ihza[4]

[1,2,3]Department of Information Systems Faculty of Science and Technology,
Universitas Islam Negeri Sultan Syarif Kasim Riau, Indonesia
[4]Department of Islamic Law, Faculty of Islamic Studies, Al-Azhar University, Egypt

E-Mail: [1]mridhoanugrah145@gmail.com, [2]nandanazira0211@gmail.com,
[3]nolaardeliaa@gmail.com, [4]nurulihza2002@gmail.com

Received March 13th 2023; Revised May 27th 2023; Accepted Jun 25th 2023
Corresponding Author: Muhammad Ridho Anugrah

Abstract

Coronary Heart Disease (CHD) is a chronic disease that is not contagious and can cause heart attacks. This makes CHD one of the diseases that cause the highest mortality globally. CHD can be caused by the main factor, namely an unhealthy lifestyle, so that in an effort to identify and deal with CHD, many studies have been conducted, one of which is the use of information technology. With so many CHD patient data, data mining can be used using. classification methods include C4.5 algorithm and Support Vector Machine (NBC). The C4.5 algorithm is a decision tree-like algorithm that groups attribute values into classes so that it resembles a tree, while SVM is an algorithm that separates data with a hyperplane. This study aims to classify the CHD dataset by comparing the C4.5 and SVM algorithms. So that the best accuracy value for this data is produced, namely the SVM algorithm of 64.51% and followed by the C4.5 algorithm of 64.30%.

Keywords**:** C4.5, Chronic Heart Disease, Classification, Data Mining, Support Vector Machine

1. INTRODUCTION

Chronic diseases are long-lasting and develop gradually. Chronic diseases are also referred to as Non-Communicable Diseases (NCDs) and are typically the result of genetics, environment, or lifestyle factors.[1]. Unhealthy lifestyles can lead to the onset of NCDs, which can reduce the quality of life[2][3][4]. One of the NCDs with a risk factor of an unhealthy lifestyle is Cardiovascular Disease, which includes Coronary Heart Disease (CHD)[5][6]. Coronary Heart Disease (CHD) is an NCD caused by atherosclerosis, which is the buildup of plaque in the arteries over time due to increased levels of low-density lipoprotein (LDL) cholesterol in the bloodstream. The plaque accumulates over time, narrowing the arteries and reducing blood flow to the heart muscle, resulting in a decreased heart function. If the plaque buildup completely blocks an artery, it can increase the risk of a heart attack[7]. Cardiovascular disease is the leading cause of death worldwide. According to the World Health Organization (WHO), in 2019, cardiovascular disease claimed 17.9 million lives, with 85% of these deaths attributed to heart attacks[8]. The significant numbers mentioned above drive the need to classify and predict CHD by identifying the factors that can cause individuals to develop CHD through data mining. Data mining is a learning technique that utilizes machine learning algorithms to analyze diverse datasets from the past, aiming to discover valuable information, patterns, and relationships that can be useful for decision-making in the future[9][10]. Classification in data mining is a technique used to process a collection of data into specific pre-defined groups or classes, using mathematical techniques such as decision trees, linear programming, neural networks, and statistics[11].

Among the classification algorithms in data mining are the C4.5 algorithm and Support Vector Machine (SVM). The C4.5 algorithm is a decision tree algorithm that works by selecting attributes with the highest normalized information gain value. The C4.5 algorithm iteratively constructs branches, forming a tree-like structure when visualized[12]. The advantages of the C4.5 algorithm are its ability to classify each value into separate branches for categorical attributes, thus facilitating classification. C4.5, being a decision tree algorithm, also exhibits high accuracy levels[13]. However, C4.5 is not always optimal and tends to favor attributes with higher values.[14]. SVM is one of the data mining techniques that works by categorizing data through finding a hyperplane. This hyperplane can separate data with the highest margin[15]. SVM excels in

accuracy, efficient time usage, both in generalization, and has low computational burden[16]. On the other hand, SVM has a limitation that makes it challenging to handle data with a large scale[17].

A recent study in 2021 by Sajja et al compared SVM, KNN, RF, C4.5, and ID3 algorithms and arrived at the conclusion that SVM is the most effective method for classifying heart disease[18]. In a study written in 2022 by Phasinam et al, the comparison of performance among machine learning techniques in predicting diseases was discussed. The study concluded that among ID3, C4.5, SVM, and NB, SVM exhibited better results and lower error rates[19]. In a research study conducted in 2020 by Yahaya, Oye, and Adamu on the analysis of machine learning algorithms in predicting heart disease, it was concluded that among NB, J-48, SVM, RF, KNN, LR, MLP, and FCM algorithms, RF and SVM were found to be more accurate and recommended as baseline methods for predicting heart disease[20]. A previous study by Mailana, Agus et al. in 2021 compared the C4.5 algorithm with SVM for predicting student graduation. The comparison of accuracy and precision on a 20% testing data showed that SVM achieved a higher accuracy of 85%, while C4.5 had an accuracy of 80%. However, it was noted that C4.5 is better suited for visualization purposes[21].

Based on the discussion provided, where no recent research has been found classifying coronary heart disease data using C4.5 and SVM, a classification of CHD was conducted by implementing the C4.5 and SVM algorithms. Therefore, the research conducted is titled Implementation of C4.5 and Support Vector Machine (SVM) Algorithm for Classification of Coronary Heart Disease.

2. MATERIALS AND METHODS

The material under study in this research is the dataset "Replication Data for: South African Heart Disease" obtained from the Harvard Dataverse. The dataset can be accessed through the following link: https://dataverse.harvard.edu/dataset.xhtml?persistentId=doi:10.7910/DVN/76SIQD. The dataset consists of 10 attributes, including chd, sbp, tobacco, ldl, adiposity, famhist, typea, obesity, alcohol, and age[22]. This study is experimental in nature and follows a specific methodology outlined in Figure 1:

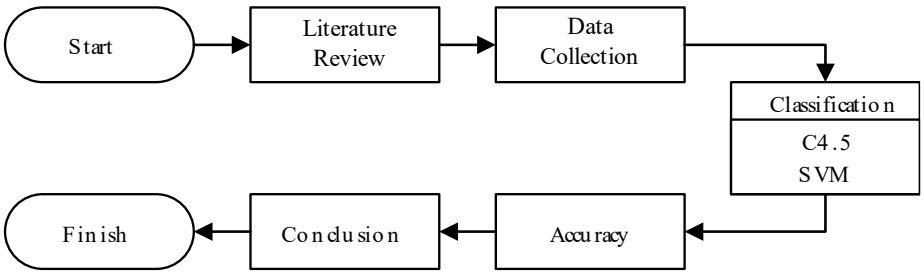

Figure 1. Research Methodology

The stages of this research begin with (1) reviewing literature, scientific articles, modules, and other relevant sources of information related to the research topic; (2) Data collection obtained from https://dataverse.harvard.edu/dataset.xhtml?persistentId=doi:10.7910/DVN/76SIQD; (3) Classification using the C4.5 and SVM algorithms; (4) Testing accuracy; and (5) Drawing conclusions from the research findings.

2.1 Coronary Heart Disease

Coronary Heart Disease (CHD) is a term used to describe a chronic condition caused by the narrowing of the arteries due to the accumulation of a waxy substance known as plaque. The narrowing of the arteries reduces blood flow to the heart muscles and tissues, leading to weakened heart function and the potential for a heart attack. The main risk factors for CHD include unhealthy cholesterol levels, diabetes, high blood pressure, obesity, smoking, unhealthy lifestyle, age, gender, genetics, and stress[23].

2.2 Decision Tree

A decision tree is a data mining technique used to classify objects by dividing the data into multiple sets based on input variables, thereby forming a hierarchical tree[22].

2.3 C4.5

The C4.5 algorithm is essentially a decision tree. In constructing the decision tree, the C4.5 algorithm follows the following steps:

1. Selecting the root attribute.
2. Creating branches for each attribute value.
3. Dividing cases into branches.
4. Repeating the process until each branch contains all cases with the same class.

To determine the root attribute of the decision tree, C4.5 uses entropy, gain, split info, and gain ratio. Entropy is a parameter that reflects the level of heterogeneity in the data. The calculation of entropy is determined by equation (1).

$$\text{Entropy}\ (S) = \ -\ \sum_{i=1}^{n} p_i \log_2(p_i) \tag{1}$$

Explanation:

S : Set of cases
n : Number of classification classes
p_i : Number of samples proportional to class i

Gain is the difference between the total entropy and the entropy of an attribute, used to measure the effectiveness of the attribute in classifying data. Gain is formulated by equation (2).

$$\text{Gain}\ (S, A) = \ -\ \text{Entropy}\ (S) - \sum_{i=1}^{n} \frac{|S_i|}{|S|} \times \text{Entropy}(S_i) \tag{2}$$

Explanation:

A : Attribute
|Si| : Number of samples in class i
|S| : Total number of samples in the entire dataset

The value of gain plays a role in determining the attributes that will become nodes or leaves in the decision tree[24].

2.4 Support Vector Machine

The Support Vector Machine (SVM) classification algorithm is an algorithm that creates decision boundaries between two classes, allowing the prediction of labels for one or more attribute vectors. These decision boundaries are known as hyperplanes because they are located farthest from the closest data points to each other. These closest data points are referred to as support vectors. Given a labeled training set, the SVM algorithm can be represented by equation (3).

$$(x_1, y_1), \ldots, (x_n, y_n), x_i \in R^d \text{and}\ y_u \in (-1, +1) \tag{3}$$

x_i is the attribute vector, and y_i is the class label. The optimal hyperplane is defined by equation (4).

$$wx^T + b = 0 \tag{4}$$

w is the weight vector, x is the input attribute vector, and b is the bias. w and b satisfy the inequalities of the training set as shown in equations (5) and (6).

$$wx^T + b\ \geq\ +1 = \text{if}\ y_i = 1 \tag{5}$$

$$wx^T + b\ \leq\ -1 = \text{if}\ y_i = -1 \tag{6}$$

The SVM model aims to determine the values of w and b in order to separate the data with a hyperplane and maximize the margin $1 / \| w \|^2$. The vectors x_i where$|y_i|(wx_i^T + b) = 1$ are referred to as support vectors[25].

3. RESULTS AND ANALYSIS

3.1 Data Collection

The data used in this study is the Coronary Heart Disease dataset sourced from the Replication Data for: South African Heart Disease available on the Harvard Dataverse website. It consists of 10 attributes, namely Class, Sbp, Tobacco, Ldl, Adiposity, Famhist, Typea, Obesity, Alcohol, and Age. The confirmed data in the Coronary Heart Disease dataset consists of 463 records. The details of the South African Heart Disease dataset are presented in Table 1:

Table 1. South African Heart Disease Dataset

Class	Sbp	Tobacco	Ldl	Adiposity	….	Age
Yes	160	12	5,73	23,11	….	52
Yes	144	0,01	4,41	28,61	….	63
No	118	0,08	3,48	32,28	….	46
Yes	170	7,5	6,41	38,03	….	58

Class	Sbp	Tobacco	Ldl	Adiposity	….	Age
Yes	134	13,6	3,5	27,78	….	49
No	132	6,2	6,47	36,21	….	45
No	142	4,05	3,38	16,2	….	38
Yes	114	4,08	4,59	14,6	….	58
no	114	0	3,83	19,4	….	29
Yes	132	0	5,8	30,96	….	53
….	….	….	….	….	….	….
Yes	132	0	4,82	33,41	….	46

3.2 Implementation of C4.5 and SVM Algorithms

The South African Heart Disease dataset was processed by implementing the C4.5 and SVM algorithms to classify individuals with coronary heart disease (CHD). The class attribute in this dataset is "Class," with the class "Yes" indicating the presence of CHD (positive) and the class "No" indicating the absence of CHD (negative). The implementation of the C4.5 algorithm was evaluated using a confusion matrix, and the results are shown in Table 2.

Table 2. Confusion matrix dan akurasi C4.5

Accuracy: 64.51%

	True Yes	True No	Class Precision
Pred. Yes	6	10	37.50%
Pred. No	154	292	65.47%
Class Recall	3.75%	96.69%	

The confusion matrix for the implementation of the SVM algorithm shows that there are 6 data points correctly predicted as 'Yes' and 292 correctly predicted as 'No'. However, there are 10 incorrect predictions of 'Yes' and 154 incorrect predictions of 'No'. The performance of the SVM algorithm is evaluated using a confusion matrix, which provides detailed information on the classification results. The results are presented in Table 3.

Table 3. Confusion matrix and accuracy of SVM

Accuracy: 64.30%

	True Yes	True No	Class Precision
Pred. Yes	41	46	47.13%
Pred. No	119	256	68.27%
Class Recall	25.62%	84.77%	

The confusion matrix for the implementation with the SVM algorithm, as shown in Table 3, indicates that 41 data were correctly predicted as 'Yes' and 256 were correctly predicted as 'No'. On the other hand, 46 data were incorrectly predicted as 'Yes' and 149 were incorrectly predicted as 'No'. Based on the implementation of each algorithm, the performance of C4.5 and SVM is shown in Table 4 and Figure 2, with an accuracy of 64.51% for C4.5 and 64.30% for SVM.

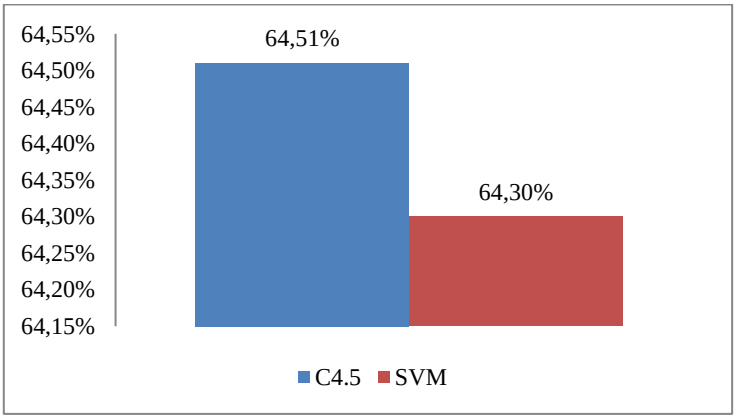

Figure 2. Accuracy Comparison of C4.5 and SVM Algorithms

The Receiver Operating Characteristic (ROC) graph illustrates the performance comparison between the C4.5 and SVM algorithms for the dataset used in this study. It is depicted in Figure 3.

Table 4. Performance Comparison of C4.5 and SVM Algorithms

No	Algorithm	Accuracy	Recall		Precision	
			True Yes	True No	True Yes	True No
1	C4.5	64.51%	3.75%	96.69%	37.50%	65.47%
2	SVM	64.30%	25.62%	84.77%	47.13%	68.27%

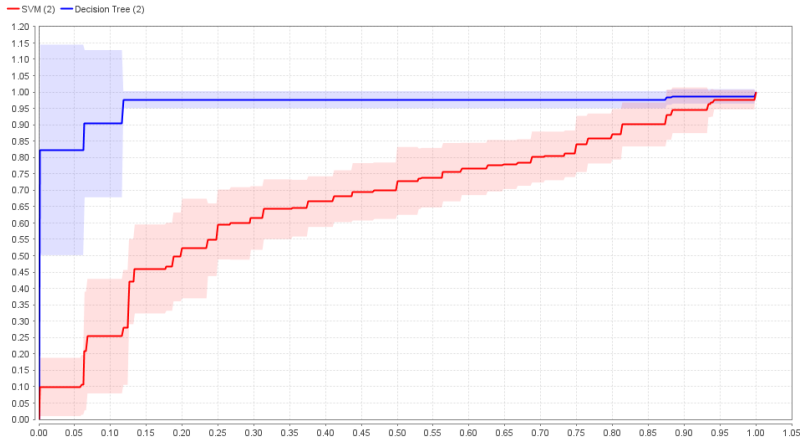

Figure 3. ROC Curve of C4.5 and SVM Algorithms

In the above Figure 3, it can be observed that the blue line represents C4.5 and the red line represents SVM, both of which have values not exceeding 1. This indicates that both C4.5 and SVM algorithms have good classification quality.

4. CONCLUSIONS

The collected Coronary Heart Disease (CHD) data from the Replication Data for: South African Heart Disease consists of 463 instances. Classification using the RapidMiner tool implemented the C4.5 and SVM algorithms, with C4.5 achieving an accuracy of 64.51% and SVM achieving 64.30%. The results indicate that both algorithms yield similar accuracy. However, considering that the implementation of C4.5 and SVM for the CHD data from the South African Heart Disease dataset yielded relatively low results, it is suggested that future research explore alternative algorithm models for CHD data.

REFERENCES

[1] E. Anderson and J. L. Durstine, "Physical activity, exercise, and chronic diseases: A brief review," *Sport. Med. Heal. Sci.*, vol. 1, no. 1, pp. 3–10, 2019, doi: 10.1016/j.smhs.2019.08.006.

[2] J. Gronewold *et al.*, "Effects of life events and social isolation on stroke and coronary heart disease," *Stroke*, no. February, pp. 735–747, 2021, doi: 10.1161/STROKEAHA.120.032070.

[3] N. P. W. P. Sari and J. Artsanthia, "Lifestyle profile of elderly living with non-communicable disease in Bangkok and Surabaya," *Int. J. Public Heal. Sci.*, vol. 8, no. 4, p. 432, 2019, doi: 10.11591/ijphs.v8i4.20371.

[4] M. Z. Islam, M. M. Rahman, and M. A. H. Moly, "Knowledge about Non-Communicable Diseases among Selected Urban School Students," *J. Armed Forces Med. Coll. Bangladesh*, vol. 15, no. 1, pp. 90–93, 2020, doi: 10.3329/jafmc.v15i1.48654.

[5] D. De Bacquer *et al.*, "Poor adherence to lifestyle recommendations in patients with coronary heart disease: Results from the EUROASPIRE surveys," *Eur. J. Prev. Cardiol.*, vol. 29, no. 2, pp. 383–395, 2022, doi: 10.1093/eurjpc/zwab115.

[6] X. He, B. R. Matam, S. Bellary, G. Ghosh, and A. K. Chattopadhyay, "CHD Risk Minimization through Lifestyle Control: Machine Learning Gateway," *Sci. Rep.*, vol. 10, no. 1, pp. 1–10, 2020, doi: 10.1038/s41598-020-60786-w.

[7] K. H. Miao and J. H. Miao, "Coronary heart disease diagnosis using deep neural networks," *Int. J. Adv. Comput. Sci. Appl.*, vol. 9, no. 10, pp. 1–8, 2018, doi: 10.14569/IJACSA.2018.091001.

[8] WHO, "Cardiovascular diseases (CVDs)," Jun. 11, 2021. https://www.who.int/en/news-room/fact-sheets/detail/cardiovascular-diseases-(cvds) (accessed Dec. 11, 2022).

[9] A. Ishaq *et al.*, "Improving the Prediction of Heart Failure Patients' Survival Using SMOTE and

Effective Data Mining Techniques," *IEEE Access*, vol. 9, pp. 39707–39716, 2021, doi: 10.1109/ACCESS.2021.3064084.

[10] A. Lakshmanarao, Y. Swathi, and S. S. S. Puella, "Machine Learning Techniques for Heart Disease Prediction," *Int. J. Sci. Technol. Res.*, vol. 8, no. March, pp. 374–377, 2019, [Online]. Available: https://www.ijstr.org/final-print/nov2019/Machine-Learning-Techniques-For-Heart-Disease-Prediction.pdf

[11] K. Sumiran, "An Overview of Data Mining Techniques and Their Application in Industrial Engineering," *Asian J. Appl. Sci. Technol.*, vol. 2, no. 2, pp. 947–953, 2018, [Online]. Available: http://ajast.net/data/uploads/5029.pdf

[12] M. Thet, "Credit Card Classification using Integration of Hierarchical Agglomerative Clustering and C4. 5 Decision Tree," no. October, 2022, [Online]. Available: https://www.researchgate.net/profile/May_Thet2/publication/321995457_Credit_Card_Classification_using_Integration_of_Hierarchical_Agglomerative_clustering_and_C45_Decsion_Tree/links/5a3c8d71a6fdcc21d8780ab0/Credit-Card-Classification-using-Integration-of-

[13] A. Wibowo, D. Manongga, and H. D. Purnomo, "The Utilization of Naive Bayes and C.45 in Predicting The Timeliness of Students' Graduation," *Sci. J. Informatics*, vol. 7, no. 1, pp. 99–112, 2020, doi: 10.15294/sji.v7i1.24241.

[14] G. S. Reddy and S. Chittineni, "Entropy based C4.5-SHO algorithm with information gain optimization in data mining," *PeerJ Comput. Sci.*, vol. 7, pp. 1–22, 2021, doi: 10.7717/PEERJ-CS.424.

[15] M. Jain, S. Narayan, P. Balaji, K. P. Bharath, A. Bhowmick, and R. Karthik, "Speech Emotion Recognition using Support Vector Machine," 2020.

[16] C. Wang, Y. Zhang, J. Song, Q. Liu, and H. Dong, "A novel optimized SVM algorithm based on PSO with saturation and mixed time-delays for classification of oil pipeline leak detection," *Syst. Sci. Control Eng.*, vol. 7, no. 1, pp. 75–88, 2019, doi: 10.1080/21642583.2019.1573386.

[17] W. Xie, G. Liang, and P. Yuan, "Research on the incremental learning SVM algorithm based on the improved generalized KKT condition," *J. Phys. Conf. Ser.*, vol. 1237, no. 2, 2019, doi: 10.1088/1742-6596/1237/2/022150.

[18] G. S. Sajja, M. Mustafa, K. Phasinam, K. Kaliyaperumal, R. J. M. Ventayen, and T. Kassanuk, "Towards Application of Machine Learning in Classification and Prediction of Heart Disease," *Proc. 2nd Int. Conf. Electron. Sustain. Commun. Syst. ICESC 2021*, no. October, pp. 1664–1669, 2021, doi: 10.1109/ICESC51422.2021.9532940.

[19] K. Phasinam, T. Mondal, D. Novaliendry, C. H. Yang, C. Dutta, and M. Shabaz, "Analyzing the Performance of Machine Learning Techniques in Disease Prediction," *J. Food Qual.*, vol. 2022, 2022, doi: 10.1155/2022/7529472.

[20] L. Yahaya, N. D. Oye, and A. Adamu, "Performance Analysis of Some Se-Lected Machine Learning Algo-Rithms on Heart Disease Predic-Tion Using the Noble Uci Datasets," *Int. J. Eng. Appl. Sci. Technol.*, vol. 5, no. 1, pp. 36–46, 2020, doi: 10.33564/ijeast.2020.v05i01.006.

[21] A. Mailana, A. A. Putra, S. Hidayat, and A. Wibowo, "Comparison of C4.5 Algorithm and Support Vector Machine in Predicting the Student Graduation Timeliness," *J. Online Inform.*, vol. 6, no. 1, p. 11, 2021, doi: 10.15575/join.v6i1.608.

[22] C. Bartley, "Replication Data for: South African Heart Disease - Partial Monotonicity Datasets Dataverse," 2016. https://dataverse.harvard.edu/dataset.xhtml?persistentId=doi:10.7910/DVN/76SIQD (accessed Dec. 12, 2022).

[23] M. De Hert, J. Detraux, and D. Vancampfort, "The intriguing relationship between coronary heart disease and mental disorders," *Dialogues Clin. Neurosci.*, vol. 20, no. 1, pp. 31–40, 2018, doi: 10.31887/dcns.2018.20.1/mdehert.

[24] H. Sulistiani and A. A. Aldino, "Decision Tree C4.5 Algorithm for Tuition Aid Grant Program Classification (Case Study: Department of Information System, Universitas Teknokrat Indonesia)," *Edutic - Sci. J. Informatics Educ.*, vol. 7, no. 1, pp. 40–50, 2020, doi: 10.21107/edutic.v7i1.8849.

[25] S. Huang, C. A. I. Nianguang, P. Penzuti Pacheco, S. Narandes, Y. Wang, and X. U. Wayne, "Applications of support vector machine (SVM) learning in cancer genomics," *Cancer Genomics and Proteomics*, vol. 15, no. 1, pp. 41–51, 2018, doi: 10.21873/cgp.20063.