International Journal of Advances in Data and Information Systems
Vol. 3, No. 2, October 2022, pp. 74~86
ISSN: 2721-3056, DOI: 10.25008/ijadis.v3i2.1243

Performance Comparison of the SVM and SVM-PSO Algorithms for Heart Disease Prediction

Dedi Saputra[1], Weishky Steven Dharmawan[2], Windi Irmayani[3]
[1,2]Information System, Universitas Bina Sarana Informatika
[3]Accounting Information System, Universitas Bina Sarana Informatika

Article Info

Article history:

Received Sep 11, 2022
Revised Oct 15, 2022
Accepted Oct 30, 2022

Keywords:

Support Vector Machine (SVM)
Particle Swarm Optimization (PSO)
Classification
Heart Disease

ABSTRACT

Data analysis for datasets with very large dimensions, classification is needed to predict from large datasets, in this study compare a method for classifying large data where the data will be processed to obtain the desired data prediction information. In this study, the Support Vector Machine (SVM) is used to provide the classification results of an algorithm that will be compared with the incorporation of the Support Vector Machine (SVM) and Particle Swarm Optimization (PSO) where the test results will be compared with the SVM classification algorithm only as a comparison algorithm. better at predicting than data sets. SVM is used as a single algorithm to see different experimental results when SVM is combined with PSO. From the experiments carried out, SVM got an Accuracy value of 81.85% and an AUC value of 0.823 while SVM-PSO got an Accuracy value of 84.81% and an AUC value of 0.898.

Corresponding Author:

Weishky Steven Dharmawan,
Information System,
Universitas Bina Sarana Informatika, Indonesia
Email: weishky.wvn@bsi.ac.id

1. INTRODUCTION

In the world of health and medicine, the accuracy of predicting a disease is very important and requires appropriate and effective decisions in taking an analysis and predicting the accuracy of a patient's disease, such as heart disease which really requires accuracy in predicting an existing disease symptom. One of the non-communicable diseases (NCDs) that is prone to occur especially when an individual is at a productive age, namely heart disease. The high mortality factor from heart disease is due to the lack of public knowledge of the symptoms or signs when a person has this disease. Heart disease is one of the diseases that is quite dangerous when attacking a person, where the main cause of heart disease comes from the lifestyle of an unhealthy individual, consuming high-cholesterol foods, using alcohol, tobacco, extreme diets and other causes.

Heart disease is a disturbance in the balance between blood supply and demand caused by blockage of blood vessels. Deaths due to heart disease reached 959,227 patients, namely 41.4% of all deaths or every day 2,600 people died from heart disease [1] [2].

Many methods of prediction of heart disease have been proposed using Genetic Algorithms, native bayes and decision trees supporting naive bayes, Multilayer Perceptron [3]. Previous studies [4] used The Statlog (heart disease) dataset, where the highest accuracy (92.59%) was obtained using the C4.5 decision tree ensemble classification compared to other algorithms. Furthermore, in another

study [5] also using the Statlog (heart disease) dataset, the highest accuracy result was obtained by the Naive Bayes method of 84%.

Symptom factors diagnosed as heart disease include the type of chest pain (cheasr pain), high blood pressure (tresbps), cholesterol (chol), ECG test score (resting electrodiagraphic "restacg"), heart rate (thalach) and blood sugar level (Fasting blood sugar “FBS”), and several other factors that identify that a person has heart disease.

Heart disease includes aortic regurgitation, cardiogenic shock, congenital heart disease, cardiomyopathy, peripartum cardiomyopathy, tricuspid regurgitation which often affects children, adults and is still a major problem in developing countries [6].

In this study, we will compare two Support Vector Machine (SVM) classification algorithms as a single algorithm and compare them with Support Vector Machine (SVM) in combination with Particle Swarm Optimization (PSO) to determine which result is more accurate in predicting more common heart disease good [7].

2. RESEARCH METHOD

Self-efficacy has three dimensions that are magnitude, the level of task difficulty a person believes she can attain; strength, the conviction regarding magnitude as strong or weak; and generality, the degree to which the expectation is generalized across situations. Self-efficacy is judgement of a person to his capabilities to plan and implement the action to reach certain goals [8]. In an academic context, self-efficacy reflects how confident students are in performing specific tasks. Self-efficacy plays a role in academic motivation and learning motivation (especially students' ability to manage their learning activities), and resistance to learning [9].

Self-efficacy in mathematics is described as an individual’s mathematics self-efficacy is his or her confidence about completing a variety of tasks, from understanding concepts to solving problems, in mathematics [10]. High mathematics self-efficacy will encourage the achievement of good learning outcomes, and when students have good learning outcomes, they will be more motivated in the learning process. Higher self-efficacy expectations can lead to better results and therefore increase the motivation for learning mathematics. Based on the description above, it can be concluded that mathematics self-efficacy is a belief or self-assessment of the student's ability in overcoming certain mathematical problems and tasks related to mathematics in the three dimensions that are magnitude, strength and generality.

One type of solution that can be done to deal with the problem of large data dimensions is to perform feature selection. It is a process that reduces the dimension of features by selecting important attributes and eliminating irrelevant, redundant and noisy attributes to get a more accurate data classification. Thus, feature selection is an important step in text classification and directly affects the classification performance [11] [12].

Classification is how to place specific objects into a group based on their nature. This method aims to study the different functions that describe each of the data selected into one of the predefined groups of classes [13].

2.1. Supports Vector Machines (SVM)

Support Vector Machine is a machine learning method that works based on the principle of Structural Risk Minimization (SRM) with the aim of finding the best hyperplane that separates two classes in the input space. The best hyperplane is a hyperplane that is located in the middle between two sets of objects of two classes. The best dividing hyperplane between the two classes can be found by measuring the hyperplane margin and finding its maximum point. Margin is the distance between the hyperplane and the closest pattern of each class. This closest pattern is called a support vector [14].

Support Vector Machine (SVM) is a machine learning method that works based on the principle of Structural Risk Minimization (SRM) with the aim of finding the best hyperplane that separates two classes in the input space [15] [16]. Support Vector Machine is defined as a set of related learning methods that analyze data and recognize patterns, which are then used for classification and regression analysis [17]. SVM takes a set of input data and predicts for each given input, which comes from two classes which are then classified by finding the best hyperplane value.

Support Vector Machine is a classification method to find the best hyperplane value that is able to find the optimal global solution [18], [19]. So the accuracy value is not easy to change because not all training data will be seen to be involved in every training iteration [20]. The data that contributes is referred to as a support vector, so this method is called a Support Vector Machine.

The characteristics of the Support Vector Machine (SVM) are as follows:

1. In principle, SVM is a linear classifier.
2. Pattern recognition is done by transforming the data in the input space to a higher-dimensional space, and optimization is carried out on the new vector space. This distinguishes SVM from typical pattern recognition solutions, which perform parameter optimization on the transformed result space with lower dimensions than the input space.
3. Implementing a Structural Risk Minimization (SRM) strategy.
4. The working principle of SVM is basically only able to handle the classification of two classes.

In simple terms, the SVM concept is an attempt to find the best hyperlane that functions as a separator of two classes in the input space, which can be seen in the image below:

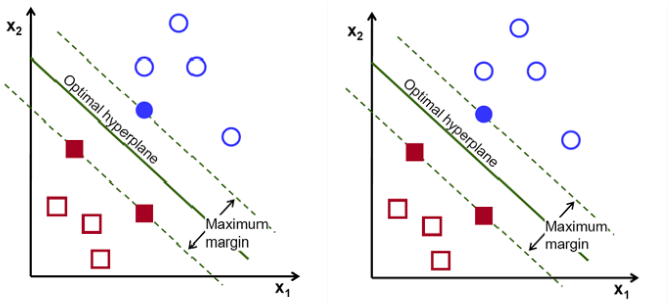


Figure. 1. SVM concept to find the best hyperplane

The figure above shows some patterns that are members of two data classes: +1 and -1. Data belonging to class -1 is symbolized by a circle, while data in class +1 is symbolized by a square.

2.2. Particle Swarm Optimization (PSO)

Particle Swarm Optimization (PSO) is a global heuristic optimization technique introduced by Doctors Kennedy and Eberhart in 1995 which was inspired by the social behavior of flocks of birds trying to achieve an unknown goal [21]. Particle Swarm Optimization (PSO) is a type of intelligence algorithm that is able to optimize the related variables most effectively.

According to Liu, Particle Swarm Optimization (PSO) is an evolution of computational engineering. Similar to genetic algorithms, PSO is an optimization tool. It is inspired by social behavior between individuals. Particles (individuals) that represent potential solutions to the problem move through the n-dimensional search space. Each particle i keeps a record of the best performing position in a vector called pbest [22].

According to Y. Yin et al., Particle Swarm Optimization (PSO) is a computational method that iteratively optimizes a problem to increase candidate solutions to a certain size. Quality The movement of each particle is affected by a local position that is guided towards the most recognized position in the search for space, which is updated as a better position than the other particles [23]. Particle Swarm Optimization (PSO) is also an evolutionary computational technique capable of generating global optimal solutions in the search space through the interaction of individual particles in a swarm. Each particle conveys information in the form of its best position to other particles and adjusts the position and speed of each based on the information received about the best position [24].

According to Zhao, Liu, Zhang, & Wang, Particle Swarm Optimization (PSO) is an evolutionary computational technique that is able to generate a global optimal solution in the search space through the interaction of individuals in a swarm of particles [25]. Each particle conveys information in the form of its best position to other particles and adjusts the position and speed of each based on the information received about the best position.

Particle Swarm Optimization (PSO) is a tool for dealing with optimization problems. Although relatively new, many have implemented the PSO algorithm, because it is quite simple and has a faster

computational speed than other optimization algorithms such as Genetic Algorithm (GA). Each particle in PSO is also associated with the velocity of the particle that flies through the search space at a speed that is dynamically adjusted to its historical behavior. Therefore, the particles have a tendency to fly towards a better search area during the search process [26].

Based on the above understanding, it can be concluded that Particle Swarm Optimization (PSO) is an optimization method that is able to optimize the closest variable to achieve maximum accuracy. Swarm Intelligence (SI) is an innovative distributed intelligent paradigm to solve optimization problems which originally took inspiration from biological examples with the phenomena of swarming, flocking and herding in vertebrate animals. Particle Swarm Optimization (PSO) combines the swarming behavior of sampled animals in a flock of birds, a group of fish, or a swarm of bees, and social behavior in humans [27].

To find the optimal solution, each particle will move to the previous best position (pbest) and the global best position (gbest). For example, the i-th particle is expressed as: xi = (xi,1,xi,2... x-i,d) in d-dimensional space. The previous best position of the i-th particle is stored and expressed as pbesti = (pbesti,1, pbesti,2,... pbesti,d). Change the speed and position of each one The particles can be calculated using the current velocity and the distance pbesti,d to pbestd as shown by the following equation:

$$v_{i,m} = w.v_{i,m} + c1 * R * (pbest_{i,m} - x_{i,m}) + c2 * R * (gbest_m - x_{i,m})$$

$$x_{id} = x_{i,m} + v_{i,m}$$

Where:
n : number of particles in group d : dimension
vi,m : velocity of particle i in iteration i
w : weight factor of inertia
c1, c2 : acceleration constant (learning rate) R : random number (0-1)
xi,d : the current position of the i-th particle in the i-th iteration
pbesti : the previous best position of the i-th particle
gbest : the best particle among all the particles in a group or population

The above formula calculates the new velocity for each particle (potential solution) based on the previous velocity (Vi,m), the location of the particle that has achieved the best fitness value (pbest), and the location of the global population (gbest for the global version, lbest for the local version). or the local environment in the local version of the algorithm where the best fitness value has been achieved.

The following equation updates the position of each particle in the solution space. Two random numbers c1 and c2 are generated independently. The use of inertial weights w has provided improved performance in a number of applications. Broadly speaking, the basic structure of PSO can be depicted in the graph below:

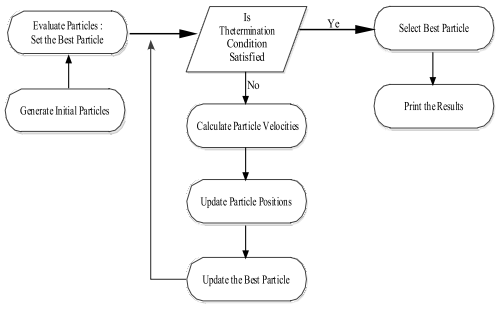


Figure 2. PSO Basic Structure

2.3. K-Fold Cross Validation Test

One alternative approach to "train and test" that is often adopted in some cases (and some regardless of size) is called k-fold cross-validation, by testing the magnitude of the error in the test

data.

Cross validation is a validation technique by dividing the data randomly into k parts and each part will be classified as a process [28]. By using cross validation, a k test will be carried out. The data used in this experiment is training data to find the overall error rate. In general, testing the value of k is done 10 times to estimate the accuracy of the estimate. In this study, the value of k used is 10 or 10 times the cross validation. each experiment will use one test data and the k-1 part will be the training data, then the test data will be exchanged with one training data so that for each trial different test data will be obtained. Training data is data that will be used in conducting learning while test data is data that has never been used for learning and will be used as data to test the truth or accuracy of learning outcomes [29].

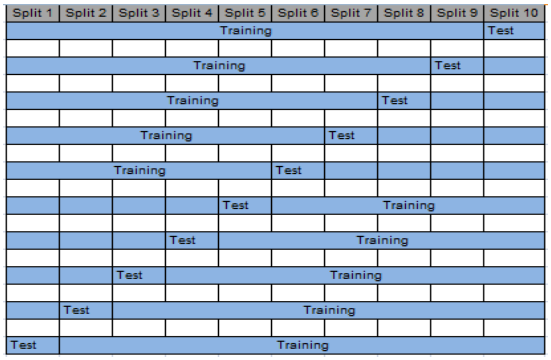


Figure. 3. Ilustrasi 10-Fold Cross validation

2.4. Confusion Matrix

The confusion matrix provides decisions obtained in training and testing, the confusion matrix provides an assessment of the classification performance based on objects correctly or incorrectly [30]. The confusion matrix contains actual (actual) and predicted (predicted) information on the classification system.

Table 1. Confusion Matrix

Classification	Predicted Class	
	Class = Yes	Class = No
Class = Yes	a(True Positive)	b(False Negative)
Class = No	c(False Positive)	d(True Negative)

Information:
True Positive (TP) = the proportion of positives in the data set that are classified as positive
True Negative (TN) = the proportion of negatives in the data set that are classified as negative
False Positive (FP) = the proportion of negatives in the data set that are classified as positive
False Negative (FN) = the proportion of negatives in the data set that are classified as negative
The following is the equation of the Confusion matrix model:
a. Accuracy value is the proportion of the number of correct predictions. Can be calculated using the equation:

$$Accuracy = \frac{TP + TN}{(TP + TN + FP + FN)}$$

b. Sensitivity is used to compare the proportion of TP to positive tuples, which is calculated using the equation:

$$Sensitivity = \frac{TP}{(TP + FN)}$$

c. Specificity is used to compare the proportion of TN to negative tuples, which is calculated using the equation:

$$Specificity = \frac{TN}{(TN + FP)}$$

d. PPV (positive predictive value) is the proportion of cases with a positive diagnosis, which is calculated using the equation:

$$PPV = \frac{TP}{(TP + FP)}$$

e. NPV (negative predictive value) is the proportion of cases with a negative diagnosis, which is calculated using the equation:

$$NPV = \frac{TN}{(TN + FN)}$$

2.5. ROC Curve

The ROC (Receiver Operating Characteristic) curve or commonly called the AUC value is a useful visual tool for comparing two classification models. ROC expresses the Confusion matrix. ROC is a two-dimensional graph with false positives as horizontal lines and true positives as vertical lines [9]. Using the ROC curve, we can see the trade off between the rate at which the model can accurately identify positive tuples and the rate at which the model incorrectly recognizes negative tuples as positive tuples. The ROC graph is a two-dimensional plot with the proportion of false positives (fp) on the X axis and the proportion of true positives (tp) on the Y axis. Points (0,1) are the perfect classification of all positive and negative cases. False positive values are absent (fp = 0) and true positive values are high (tp = 1). Points (0,0) are classifications that predict every case to be negative {-1}, and points (1,1) are classifications that predict every case to be positive {1}. The ROC graph depicts the trade-off between benefits (true positives) and costs (false positives). The following shows two types of ROC curves (discrete and continuous).

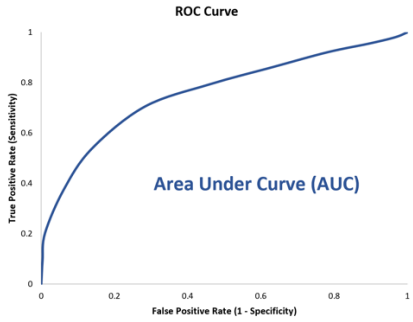


Figure. 4. ROC Curve

The point above the diagonal line is a good classification result, while the point below the diagonal line is a poor classification result. It can be concluded that, one point on the ROC curve is better than another if the direction of the line is from the bottom left to the top right on the graph. For the level of accuracy of the AUC value in data mining classification, it is divided into five groups [30] namely:

a. 0.90 - 1.00 = very good classification
b. 0.80 - 0.90 = good classification
c. 0.70 - 0.80 = fair classification
d. 0.60 - 0.70 = bad classification
e. 0.50- 0.60 = misclassification (failed)

3. RESULTS AND DISCUSSION

This research is a systematic problem solving activity, which is carried out carefully and attentively in the context of the situation at hand, research in the academic field is used to refer to a diligent and systematic investigation or investigation in an area, with the aim of finding or revising facts, theories , application and purpose for discovering and disseminating new knowledge This study uses experimental research methods involving the Heart Disease dataset. Heart disease is a heart database that has variables that must be predicted whether it has symptoms of heart disease or not.

Dataset Information:

This database contains 76 attributes, but all published experiments refer to the use of a subset of 14 attributes. In particular, the Cleveland database is the only one used by ML researchers to date. The "goal" field refers to the presence of heart disease in the patient. It is an integer rated from 0 (none) to 4. Experiments with the Cleveland database have concentrated on a simple attempt to distinguish existence (value 1,2,3,4) from absence (value 0). The patient's name and social security number were recently removed from the database, replaced with dummy values. Attribute Information:

Only 14 attributes used:
1. #3 (age)
2. #4 (sex)
3. #9 (cp)
4. #10 (trestbps)
5. #12 (chol)
6. #16 (fbs)
7. #19 (restecg)
8. #32 (thalach)
9. #38 (exang)
10. #40 (oldpeak)
11. #41 (slope)
12. #44 (ca)
13. #51 (thal)
14. #58 (num) (the predicted attribute)

Data Set Characteristics:	Multivariate	**Number of Instances:**	303	**Area:**	Life
Attribute Characteristics:	Categorical, Integer, Real	**Number of Attributes:**	75	**Date Donated**	1988-07-01
Associated Tasks:	Classification	**Missing Values?**	Yes	**Number of Web Hits:**	2029499

Figure. 5. Datasets heart disease

Source: UCI Repository
Link : https://archive.ics.uci.edu/ml/datasets/heart+disease

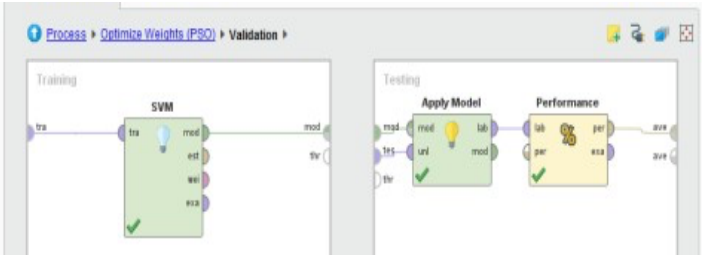
Complete attribute documentation:
1 id: patient identification number
2 ccf: social security number (I replaced this with a dummy value of 0)
3 age: age in years
4 sex: sex (1 = male; 0 = female)
5 painloc: chest pain location (1 = substernal; 0 = otherwise)
6 painexer (1 = provoked by exertion; 0 = otherwise)
7 relrest (1 = relieved after rest; 0 = otherwise)
8 pncaden (sum of 5, 6, and 7)
9 cp: chest pain type
-- Value 1: typical angina
-- Value 2: atypical angina
-- Value 3: non-anginal pain
-- Value 4: asymptomatic
10 trestbps: resting blood pressure (in mm Hg on admission to the hospital)
11 htn
12 chol: serum cholestoral in mg/dl
13 smoke: I believe this is 1 = yes; 0 = no (is or is not a smoker)
14 cigs (cigarettes per day)

15 years (number of years as a smoker)
16 fbs: (fasting blood sugar > 120 mg/dl) (1 = true; 0 = false)
17 dm (1 = history of diabetes; 0 = no such history)
18 famhist: family history of coronary artery disease (1 = yes; 0 = no)
19 restecg: resting electrocardiographic results
-- Value 0: normal
-- Value 1: having ST-T wave abnormality (T wave inversions and/or ST elevation or depression of > 0.05 mV)
-- Value 2: showing probable or definite left ventricular hypertrophy by Estes' criteria
20 ekgmo (month of exercise ECG reading)
21 ekgday(day of exercise ECG reading)
22 ekgyr (year of exercise ECG reading)
23 dig (digitalis used furing exercise ECG: 1 = yes; 0 = no)
24 prop (Beta blocker used during exercise ECG: 1 = yes; 0 = no)
25 nitr (nitrates used during exercise ECG: 1 = yes; 0 = no)
26 pro (calcium channel blocker used during exercise ECG: 1 = yes; 0 = no)
27 diuretic (diuretic used used during exercise ECG: 1 = yes; 0 = no)
28 proto: exercise protocol
1 = Bruce
2 = Kottus
3 = McHenry
4 = fast Balke
5 = Balke
6 = Noughton
7 = bike 150 kpa min/min (Not sure if "kpa min/min" is what was written!)
8 = bike 125 kpa min/min
9 = bike 100 kpa min/min
10 = bike 75 kpa min/min
11 = bike 50 kpa min/min
12 = arm ergometer
29 thaldur: duration of exercise test in minutes
30 thaltime: time when ST measure depression was noted
31 met: mets achieved
32 thalach: maximum heart rate achieved
33 thalrest: resting heart rate
34 tpeakbps: peak exercise blood pressure (first of 2 parts)
35 tpeakbpd: peak exercise blood pressure (second of 2 parts)
36 dummy
37 trestbpd: resting blood pressure
38 exang: exercise induced angina (1 = yes; 0 = no)
39 xhypo: (1 = yes; 0 = no)
40 oldpeak = ST depression induced by exercise relative to rest
41 slope: the slope of the peak exercise ST segment
-- Value 1: upsloping
-- Value 2: flat
-- Value 3: downsloping
42 rldv5: height at rest
43 rldv5e: height at peak exercise
44 ca: number of major vessels (0-3) colored by flourosopy
45 restckm: irrelevant
46 exerckm: irrelevant
47 restef: rest raidonuclid (sp?) ejection fraction
48 restwm: rest wall (sp?) motion abnormality
0 = none

1 = mild or moderate
2 = moderate or severe
3 = akinesis or dyskmem (sp?)
49 exeref: exercise radinalid (sp?) ejection fraction
50 exerwm: exercise wall (sp?) motion
51 thal: 3 = normal; 6 = fixed defect; 7 = reversable defect
52 thalsev: not used
53 thalpul: not used
54 earlobe: not used
55 cmo: month of cardiac cath (sp?) (perhaps "call")
56 cday: day of cardiac cath (sp?)
57 cyr: year of cardiac cath (sp?)
58 num: diagnosis of heart disease (angiographic disease status)
-- Value 0: < 50% diameter narrowing
-- Value 1: > 50% diameter narrowing
(in any major vessel: attributes 59 through 68 are vessels)
59 lmt
60 ladprox
61 laddist
62 diag
63 cxmain
64 ramus
65 om1
66 om2
67 rcaprox
68 rcadist
69 lvx1: not used
70 lvx2: not used
71 lvx3: not used
72 lvx4: not used
73 lvf: not used
74 cathef: not used
75 junk: not used
76 name: last name of patient (I replaced this with the dummy string "name")

3.1 Testing the Support Vector Machine (SVM) Method

The following are the results of the Support Vector Machine test using the Cross Validation method using RapidMiner. The results and discussion contain discussion and final results or program outputs or analysis of research methods.



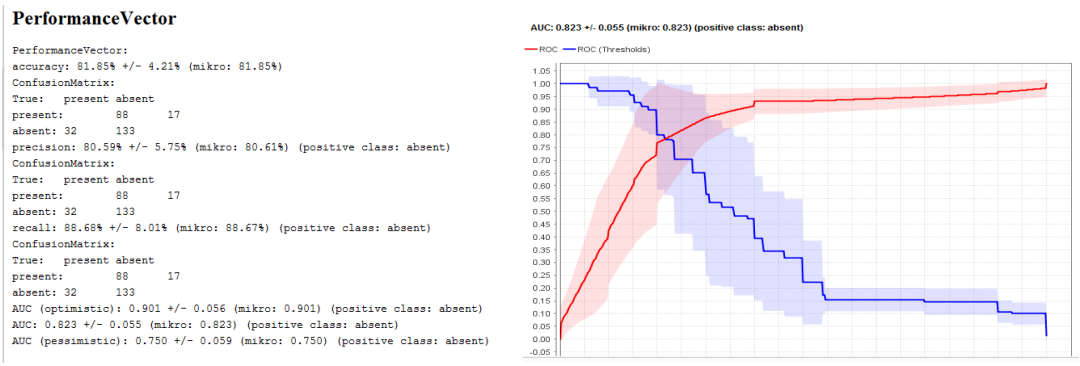


Figure. 6. Testing the Support Vector Machine Using RapidMiner

TP = 88; FP = 17; TN = 32; FN= 133

Accuracy = ((TP+TN) / (TP+TN+FN+FP)) *100%

Accuracy = ((88+133) / (88+133+17+32)) * 100%

Accuracy = 81.59%

The results of the SVM test from the haert disease dataset are shown in Figure 5. The accuracy value is 81.59% and the AUC is 0.823, so that value will be used in this study. Based on the classification of data mining according to Gorunescu Florin, the results of the completion of the classification carried out by SVM with the haert disease dataset have an AUC value between 0.80-0.90 with a good classification meaning.

Results show that the performance evaluation of the proposed method is carried out by calculating the test parameters in the form of precision with a value of 80.59%, recall with a value of 88.68% is calculated using the formula:

$$Precision = \frac{TP}{(TP + FP)}$$

$$Recall = \frac{TP}{(TP + FN)}$$

3.2 Testing the Support Vector Machine (SVM) and Particle Swarm Optimization (PSO) Method

The following are the results of the Support Vector Machine and Particle Swarm Optimization test using the Cross Validation method using RapidMiner.The results and discussion contain discussion and final results or program outputs or analysis of research methods.

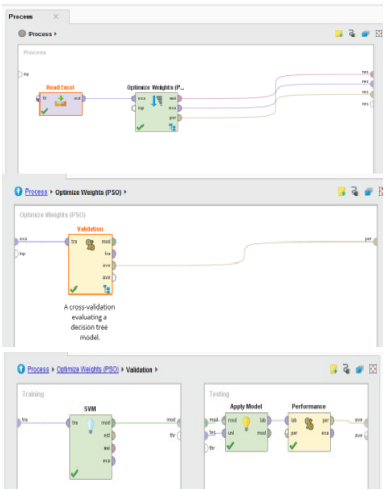


Figure. 7. Testing the Support Vector Machine and Particle Swarm Optimization Using RapidMiner

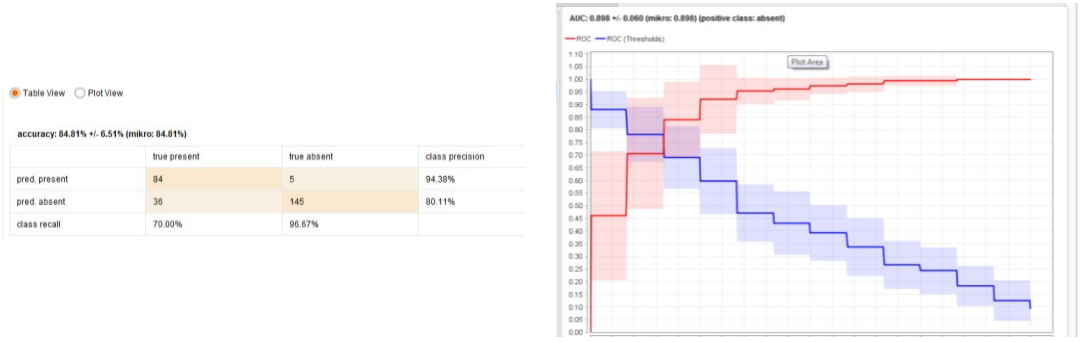


Figure 8. SVM-PSO Accuracy Results and AUC Value

TP = 84; FP = 5; TN = 36; FN= 145 Akurasi = ((TP+TN) / (TP+TN+FN+FP)) *100%

Accuracy = ((84+145) / (84+145+5+36)) *100%

Accuracy = 84,81%

From the results of the SVM-PSO test from the haert disease dataset in Figure 7. the accuracy value is 84.81% and the AUC is 0.898, so that value will be used in this study. Based on the classification of data mining according to Gorunescu Florin, the results of the completion of the classification carried out by SVM-PSO with the haert statlog dataset have an AUC value between 0.80-0.90 with a good classification meaning.

Results show that the performance evaluation of the proposed method is carried out by calculating the test parameters in the form of precision with a value of 80.84%, recall with a value of 96.97% is calculated using the formula:

$$Precision = \frac{TP}{(TP + FP)}$$

$$Recall = \frac{TP}{(TP + FN)}$$

3.1.1 Evaluation Analysis and Result Validation

The results of the tests carried out were to make direct predictions with the support vector machine (SVM) as a single method, and compared directly with the support vector machine (SVM) based on particle swarm optimization (PSO) to determine accuracy and AUC (Area Under the Curve) values. The classification model can be evaluated based on criteria such as accuracy, speed, reliability, scalability and interpretability (Vecellis, 2009). The results of the analysis of Accuracy and AUC calculations from the SVM vs SVM-PSO algorithm are summarized in the table below.

Table 2. The results of the analysis

Value	SVM	SVM-PSO
Accuracy	81.59	84.81%
AUC	0.823	0.898

The results shown in the graph in table 1, it can be stated that the results of the classification method focus on the Accuracy and AUC values in each method. The results of the performance evaluation of the proposed method are carried out by calculating the test parameters in the form of precision, recall and f-measure. In general, precision, recall and f-measure are calculated using the formula:

$$Precision = \frac{TP}{(TP + FP)}$$

$$Recall = \frac{TP}{(TP + FN)}$$

$$F - Measure = 2 \times \frac{Precision * Recall}{Precision + Recall}$$

Table 3. The results of the analysis

Value	SVM	SVM-PSO
Precision	80.59%	80.84
Recall	88.68%	96.97%
F-Measure	84.441	88.049

Based on the experimental results and data analysis in this study, it can be obtained that the comparison of the highest Accuracy and Area Under Curve (AUC) values in the SVM-PSO model testing is compared to using a single SVM algorithm. There are advantages to the SVM-PSO method, after optimization of the parameters with the testing phase. The test results show that SVM-PSO on average has better performance than the SVM comparison method only in terms of Accuracy and Area Under Curve (AUC) values and several Precision, Recall and F-Measure values.

4. CONCLUSION

In data analysis research for large-dimensional datasets, classification is very necessary in predicting from a dataset, in this study comparing a method for classifying large data where the data will be processed to obtain the desired data prediction information. From the experiments conducted by the SVM-PSO algorithm, the Accuracy value is 84.81% and the AUC value is 0.898, while the SVM Algorithm as a single algorithm only gets an Accuracy value of 81.85% and an AUC value of 0.823. From the results of experiments and tests that have been carried out, it can be concluded that the SVM-PSO algorithm method is better than using a single SVM algorithm in predicting and classifying data.

ACKNOWLEDGEMENTS

In this study, I as the author would first like to thank my teacher who has taught me how to process large-dimensional data and how to classify and predict data to be used as information. Second, thanks also to the authors of journals and books as references in this research. and lastly, I do not forget to thank the International Journal of Advances in Data and Information Systems which has given me the opportunity for my research to be able to contribute to this journal, I hope this research can be published, thank you again.

REFERENCES

[1] D. P. Utomo and M. Mesran, "Analisis Komparasi Metode Klasifikasi Data Mining dan Reduksi Atribut Pada Data Set Penyakit Jantung," *Jurnal Media Informatika Budidarma*, vol. 4, no. 2, pp. 437–444, 2020.

[2] J. E. Dalen, J. S. Alpert, R. J. Goldberg, and R. S. Weinstein, "The epidemic of the 20th century: coronary heart disease," *Am J Med*, vol. 127, no. 9, pp. 807–812, 2014.

[3] K. Thirumoorthy and K. Muneeswaran, "Feature selection using hybrid poor and rich optimization algorithm for text classification," *Pattern Recognit Lett*, vol. 147, pp. 63–70, 2021.

[4] X. Liu *et al.*, "A hybrid classification system for heart disease diagnosis based on the RFRS method," *Comput Math Methods Med*, vol. 2017, 2017.

[5] S. H. Wijaya, G. T. Pamungkas, and M. B. Sulthan, "Improving classifier performance using particle swarm optimization on heart disease detection," in *2018 International Seminar on Application for Technology of Information and Communication*, 2018, pp. 603–608.

[6] P. D. Putra and D. P. Rini, "Prediksi Penyakit Jantung dengan Algoritma Klasifikasi," in *Annual Research Seminar (ARS)*, 2020, vol. 5, no. 1, pp. 95–99.

[7] D. Saputra, F. Akbar, and A. Rahman, "Decision Support System for Providing Customer Reward Using Profile Matching Method: A Case Study at PT. Atlas Jakarta," *Bulletin of Computer Science and Electrical Engineering*, vol. 2, no. 1, pp. 28–37, 2021.

[8] H. Farmer, H. Xu, and M. E. Dupre, "Self-efficacy," in *Encyclopedia of Gerontology and Population Aging*, Springer, 2022, pp. 4410–4413.
[9] A. Zafra and S. Ventura, "Multi-instance genetic programming for predicting student performance in web based educational environments," *Appl Soft Comput*, vol. 12, no. 8, pp. 2693–2706, 2012.
[10] H. E. Zuya, S. K. Kwalat, and B. G. Attah, "Pre-Service Teachers' Mathematics Self-Efficacy and Mathematics Teaching Self-Efficacy.," *Journal of education and practice*, vol. 7, no. 14, pp. 93–98, 2016.
[11] M. Wahyudi, S. Sfenrianto, and W. S. Dharmawan, "Features Selection Based ABC-SVM and PSO-SVM in Classification Problem".
[12] J. Han, J. Pei, and H. Tong, *Data mining: concepts and techniques*. Morgan kaufmann, 2022.
[13] D. Saputra, W. S. Dharmawan, M. Wahyudi, W. Irmayani, J. Sidauruk, and Martias, "Performance Comparison and Optimized Algorithm Classification," *J Phys Conf Ser*, vol. 1641, pp. 12087–12093, 2020, doi: 10.1088/1742-6596/1641/1/012087.
[14] Y. Lee and J. Lee, "Binary tree optimization using genetic algorithm for multiclass support vector machine," *Expert Syst Appl*, vol. 42, no. 8, pp. 3843–3851, 2015.
[15] X.-D. Zhang, "Support vector machines," in *A Matrix Algebra Approach to Artificial Intelligence*, Springer, 2020, pp. 617–679.
[16] G. James, D. Witten, T. Hastie, and R. Tibshirani, "Support vector machines," in *An introduction to statistical learning*, Springer, 2021, pp. 367–402.
[17] P. Chen, C. Lin, and B. Schölkopf, "A tutorial on ν-support vector machines," *Appl Stoch Models Bus Ind*, vol. 21, no. 2, pp. 111–136, 2005.
[18] D. A. Pisner and D. M. Schnyer, "Support vector machine," in *Machine learning*, Elsevier, 2020, pp. 101–121.
[19] N. Guenther and M. Schonlau, "Support vector machines," *Stata J*, vol. 16, no. 4, pp. 917–937, 2016.
[20] G. Battineni, N. Chintalapudi, and F. Amenta, "Machine learning in medicine: Performance calculation of dementia prediction by support vector machines (SVM)," *Inform Med Unlocked*, vol. 16, p. 100200, 2019.
[21] J. Kim, K. Choi, G. Kim, and Y. Suh, "Classification cost: An empirical comparison among traditional classifier, Cost-Sensitive Classifier, and MetaCost," *Expert Syst Appl*, vol. 39, no. 4, pp. 4013–4019, 2012.
[22] Y. Liu, X. Yu, J. X. Huang, and A. An, "Combining integrated sampling with SVM ensembles for learning from imbalanced datasets," *Inf Process Manag*, vol. 47, no. 4, pp. 617–631, 2011.
[23] Y. Yin, D. Han, and Z. Cai, "Explore Data Classification Algorithm Based on SVM and PSO for Education Decision," *Journal of Convergence Information Technology*, vol. 6, no. 10, pp. 122–128, 2011, doi: 10.4156/jcit.vol6.issue10.16.
[24] R. Ma and X. Chen, "Intelligent education evaluation mechanism on ideology and politics with 5G: PSO-driven edge computing approach," *Wireless Networks*, pp. 1–12, 2022.
[25] Y. Lu, M. Liang, Z. Ye, and L. Cao, "Improved particle swarm optimization algorithm and its application in text feature selection," *Appl Soft Comput*, vol. 35, pp. 629–636, 2015.
[26] M. J. Zaki, W. Meira Jr, and W. Meira, *Data mining and analysis: fundamental concepts and algorithms*. Cambridge University Press, 2014.
[27] C. Grosan, A. Abraham, and M. Chis, "Swarm intelligence in data mining," in *Swarm Intelligence in Data Mining*, Springer, 2006, pp. 1–20.
[28] H. Jiawei, M. Kamber, J. Han, M. Kamber, and J. Pei, "Data Mining: Concepts and Techniques [Internet]. San Francisco, CA, itd." Morgan Kaufmann, 2012.
[29] I. H. Witten, E. Frank, and M. A. Hall, "Data Mining: Practical Machine Learning Tools and Techniques." Morgan Kaufmann Publishers, 2011.
[30] F. Gorunescu, "Data mining: concepts and techniques. Chemistry &," *Romania: Springer. https://doi. org/10.1007/978-3-642-19721-5*, 2011.