Journal Publications & Informatics Engineering Research
Volume 4, Number 1, October 2019
DOI : https://doi.org/10.33395/sinkron.v4i1.10195

e-ISSN : 2541-2019
p-ISSN : 2541-044X

Application Of The Support Vector Machine And Neural Network Model Based On Particle Swarm Optimization For Breast Cancer Prediction

Esty Purwaningsih
Universitas Bina Sarana Informatika
Jakarta, Indonesia
esty.epw@bsi.ac.id

***Abstract*—** There are several studies in the medical field that classify data to diagnose and analyze decisions. To predict breast cancer, this study compares two methods, the Support Vector Machine method and the Neural Network method based on Particle Swarm Optimization (PSO) which is intended to determine the highest accuracy value in the Coimbra dataset data. To implement the Support Vector Machine and Neural Network method based on PSO, RapidMiner software is used. Then the application results are compared using Confusion Matrix and ROC Curve. Based on the accuracy of the two models, it is known that the PSO-based Neural Network model has a higher accuracy value of 84.55% than the results of the PSO-based Vector Support Machine with an accuracy value of 80.08%. The calculation results, the accuracy of the AUC performance obtained by the results of the study are, the two methods are PSO-based Neural Network with AUC value of 0.885 and PSO-based Support Vector Machine with a value of 0.819 included in the category of Good Classification.

Keywords— breast cancer, support vector machine, neural network, particle swarm optimization, rapid miner

I. Introduction

Factors causing the death of most women in the world due to exposure to a fairly dangerous disease that is cancer (cancer) and one of them is a cancer that is very worried about most women is breast cancer.

According to Globocan data released by the (Biro Komunikasi dan Pelayanan Masyarakat, 2019) that in 2018 there were 18.1 million new cases with a mortality rate of 9.6 million deaths, of which 1 in 5 men and 1 in 6 women in the world experience cancer. The data also states 1 in 8 men and 1 in 11 women, die of cancer.

The incidence of cancer in Indonesia (136.2 / 100,000 population) ranks 8th in Southeast Asia, whereas in Asia it ranks 23. The highest incidence rate for women is breast cancer, which is 42.1 per 100,000 population with an average death rate of 17 per 100,000 population (Biro Komunikasi dan Pelayanan Masyarakat, 2019).

According to (Zamani & Amaliah, 2012) The process and method of treatment is to do either by chemotherapy or radiation, but it does not have a significant impact if the cancer-triggering cells have reached the final stage.

To be able to predict a disease, especially breast cancer, the use of data mining techniques is very potential to be applied into health services (Asria, Hiba, Moatassime, C, & Noeld, 2016).

To predict the presence of breast cancer, it has been done (Patrício et al., 2018) predicting the presence of breast cancer with a support vector machines model using glucose, resistin, age and BMI resulting in a range of sensitivity range between 82% and 88% and specificity ranging between 85 and 90%, AUC 0.87, 0.91.

Research conducted (Asria, Hiba et al., 2016) that breast cancer is one of the diseases that makes high mortality every year. In this study, a comparison of performance between different machine learning algorithms: Support Suppot Vector Machine (SVM), Decision Tree (C4.5), Naive Bayes (NB) and K-NN in Wisconsin Breast Cancer (original) dataset was done. The experimental results show that SVM provides the highest accuracy (97.13%) with the lowest error rate.

Research has also been conducted (Zamani & Amaliah, 2012) in predicting breast cancer using a combination of Neural Network (NN) and Association Rules (AR) methods. Both methods were tested using 10-fold cross validation. The results obtained that the Neural Network method with genetic algorithms get a high accuracy value of 97.00% compared to the Naïve Bayes method which produces an accuracy of 96.24% and the Neural Network method based on the Rules of the rules with an accuracy value of 95.6%.

Research conducted (Novianti & Purnami, 2012) in the diagnosis of benign (benign) and malignant (malignant) breast cancer patients based on the results of mammography and analyzing the causal factors using logistic regression methods and support vector machines (SVM). Where the accuracy of classification gets a value of 88.72% of binary logistic regression, where intermediate findings and BIRADS factors affect malignant breast cancer. In the L1-Norm SVM variable, classification accuracy is found 94.34% of all predictor variables that can affect the presence of malignant breast cancer including intermediate findings, then BIRADS, suspicious for malignancy, abnormal location, and age.

From previous studies related to the classification of breast cancer using a variety of methods used, this study compares 2 (two) methods, namely the Support Vector Machine method and the Neural Network method based on Particle Swarm Optimization (PSO) which is intended to determine the accuracy value the highest.

In this study, data sourced from the UCI machine learning repository, namely the breast cancer coimbra in 2018 by raising the problem that is limited to the application of the Support Vector Machine method and Neural Network method based on Particle Swarm Optimization (PSO), which will be determined in the Breast Cancer Coimbra dataset.

This study aims to determine the highest accuracy value by comparing the Support Vector Machine method and the Neural Network method based on Particle Swarm Optimization (PSO).

In applying the Support Vector Machine method and the Particle Swarm Optimization (PSO) based Neural Network method, RapidMiner software is used, where the software has a comprehensive system for analyzing data and has the ability and flexibility and ease of use. After being applied with RapidMiner then the level of accuracy is compared using the Confusion Matrix and ROC (Receiver Operating Characteristic) curve to find out the method that has the highest level of accuracy so that the purpose of applying the method for the Breast Cancer Coimbra dataset can be achieved.

II. LITERATURE REVIEW

Data mining has its own interest in the community and the world in the field of information systems, due to the need and willingness of large amounts of data to make useful information (Witten, I. H., Frank, E., & Hall, 2011).

According to (Maimon, Oded&Rokach, 2010), there are 3 (three) steps in the data mining process, including:

1. Preparation
   Data is selected, then cleaned and then preprocessed in accordance with the guidelines and knowledge of domain experts who take and integrate data both internal and external into the organization as a whole.
2. Data Mining Algorithm
   Makes it easy to identify data and integrate all data.
3. Analysis
   The output of the data mining process is evaluated.

According to (Vercellis, 2009), classification in data mining is a method of learning data to predict a value from a set of attributes. The classification algorithm will produce a set of rules called rules which will be used as indicators to be able to predict the class of data you want to predict.

According to (Vercellis, 2009), the process of making classification models can be divided into three major stages

a. Learning Stage
   The stage where the classification algorithm is applied to the sample data to get the data relations in each class. This stage will form a model that contains the attribute rules in determining data classes.
b. Testing Phase
   The testing phase is the stage of applying the rules that have been formed at the learning stage into sample data that is not included in the learning data. In this stage, the rules of the model will be applied to each attribute in the test data and see the match between the predicted class and the actual data class.

c. Prediction Stage
This stage is the stage where the resulting model is actually applied to data that is not yet known for its class. Rating algorithm classification is usually seen from the accuracy of the model. Model accuracy is the accuracy of the model in predicting data classes. In addition to the accuracy of the speed of the model formation, the ability of the algorithm to deal with irrelevant or even incomplete data, and the ability of the algorithm when applied to large or small amounts of data.

The selection of variables, also called attribute selection, is used in datasets for pattern formation in data mining (Witten, I. H., Frank, E., & Hall, 2011).

According to Gorunescu in (Harafani, 2015) SVM is a linear machine that is equipped with special features. Support Vector Machine (SVM) has advantages that are quite popular and good in the use of classification because it does not depend on the number of features and can overcome the problem of dimensions.

Neural Networks tend to provide accurate predictions. One of the advantages of Neural Networks is that they can work effectively with data that are not normally distributed (Satapathy, Chittineni, Mohan Krishna, Murthy, & Prasad Reddy, 2012).

Particle Swarm Optimization (PSO) has a simple concept, easy implementation, converges quickly, and can be applied to various applications and fields to solve optimization problems (Vieira, Mendonça, Farinha, & Sousa, 2013).

In testing the Breast Cancer Coimbra dataset using 10-Fold Cross Validation. K-Fold Cross Validation carried out experiments as much as k. One testing data is used in each trial and training data is obtained from part k-1, then one training data is an exchange of testing data.

III. Proposed Method

Data collection techniques used in this study using secondary data, where the dataset used was taken from the UCI Repository Machine Learning. This research was designed by conducting a Support Vector Machine and Particle Swarm Optimization (PSO) based Neural Network method to determine the highest level of accuracy in the classification of breast cancer.

The stages of research for the classification of breast cancer include:

1. Preprocessing the Breast Cancer Coimbra dataset with 116 data records, with 10 variables including 9 predictor attributes, namely age, BMI, Glucose, Insulin, HOMA, Leptin, Adiponectin, Resistin, MCP.1 and 1 objective attribute, namely Healthy Control (Healthy Control) 1) or Patients (2). Preprocessing data is done by deleting records that are missing value and duplicates. Discretization techniques are part of data transformation that is used to change data types
2. After the data is transformed, then the data is processed and tested by Rapid Miner using the Support Vector Machined and Neural Network model based on Particle Swarm Optimization.
3. Furthermore, this research produces an accuracy level, namely Confusion Matrix, to compare accuracy values in the Support Vector Machine method with the Neural Network method based on Particle Swarm Optimization.

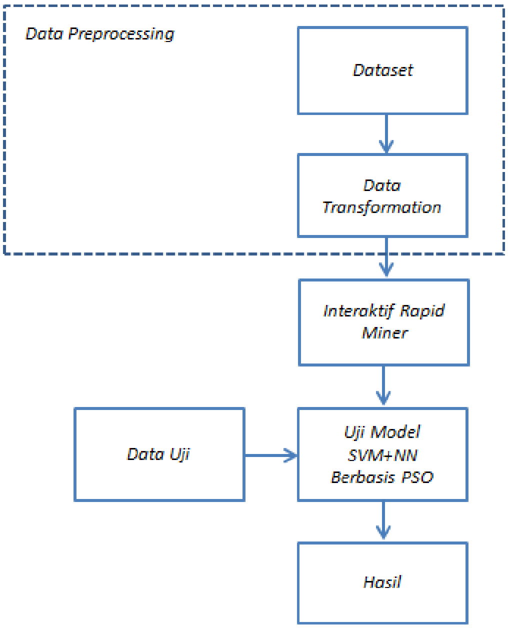


FIGURE I. STAGE OF RESEARCH

The method proposed in this study is to compare the Support Vector Machine method and the PSO-based Neural Network method. In processing the initial dataset, the data set is in the form of testing and learning data, the dataset is transformed into range 0 and range 1, then the dataset is divided into 10-fold cross validation.

The dataset was tested using the Support Vector Machine and Particle Swarm Optimization-based Neural Network method to validate the model accuracy.

The evaluation model applied in this study with the area under curve (AUC) as an indicator of accuracy to evaluate the performance of both methods. According to Gorunescu in (Ridwansyah & Purwaningsih, 2018) classification to test the accuracy of using AUC is formulated in the following range 0.90 - 1.00 (Excellent Classification), 0.80 - 0.90 (Good Classification), 0.70 - 0.80 (Fair Classification), 0.60 - 0.70 (Poor Classification), 0.50 - 0.60 (Failure).

IV. Result and Discussion

The data analyzed is a dataset obtained from the UCI machine learning repository using the Support Vector Machine and Neural Network based on Particle Swarm Optimization. The results of the analysis will then be compared to get the selected model according to the best model selection criteria, namely the method or model that has the highest accuracy.

a. Support Vector Machine (SVM)

The value of training cycles in the study is determined by conducting a trial run by entering C (1.0) and epsilon (0.0). Next is the kernel model in the Support Vector Machine training data.

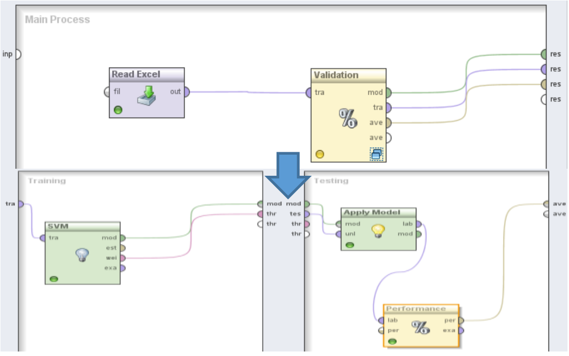
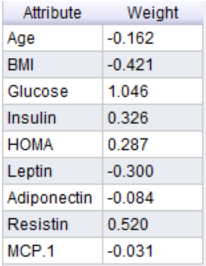


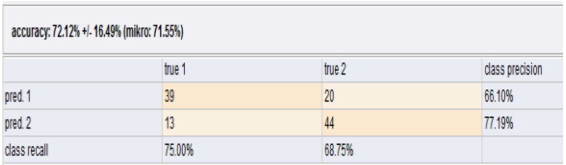
FIGURE II. Testing of Support Vector Machine Models

TABLE I. Kernel SVM



Attribute	Weight
Age	-0.162
BMI	-0.421
Glucose	1.046
Insulin	0.326
HOMA	0.287
Leptin	-0.300
Adiponectin	-0.084
Resistin	0.520
MCP.1	-0.031

TABLE II. Value of Accuracy Data Training Model Support Vector Machine



accuracy: 72.12% +/- 16.49% (mikro: 71.55%)

	true 1	true 2	class precision
pred. 1	39	20	66.10%
pred. 2	13	44	77.19%
class recall	75.00%	68.75%	

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

$$= \frac{(44 + 39)}{(44 + 20 + 39 + 13)}$$

$$= \frac{(83)}{(116)}$$

= 0,72 = **72%**

Testing training data with the Support Vector Machine model obtained an accuracy of 72,12% according to the confusion matrix table presented in Table.II out of 116 data as many as 39 data that were predicted correctly included in the Healthy Control classification, and as many as 20 data predicted Healthy Control but in fact entered into the Patients classification, 13 data were predicted by Patients but included into the Healthy Control classification, 44 data were predicted to be exact that is included in the Patients classification such as the confusion matrix table presented in Table II.

The results obtained from ROC processing for the Support Vector Machine model using training data of 0,797 can be seen in Figure III.

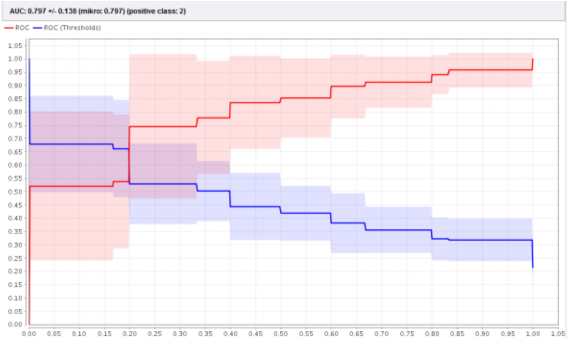
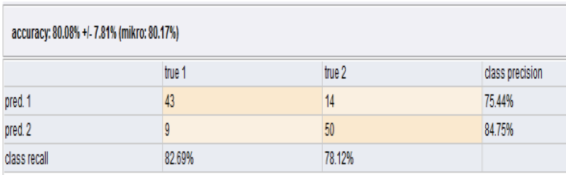


FIGURE III. AUC MODEL SUPPORT VECTOR MACHINE CURVE

b. SVM + Particle Swarm Optimization (PSO)

TABLE III. VALUE OF ACCURACY DATA TRAINING MODEL SVM + PSO



accuracy: 80.08% +/- 7.81% (mikro: 80.17%)

	true 1	true 2	class precision
pred. 1	43	14	75.44%
pred. 2	9	50	84.75%
class recall	82.69%	78.12%	

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

$$= \frac{(50 + 43)}{(50 + 14 + 43 + 9)}$$

$$= \frac{(93)}{(116)} \quad = 0{,}80 \quad = 80\%$$

Testing training data with Support Vector Machine + Particle Swarm Optimization model obtained 80,08% accuracy according to the confusion matrix table presented in Table.III out of 116 data as many as 43 data were correctly predicted namely included in the Healthy Control classification, and as many as 14 predicted data Healthy Control but apparently included in the Patients classification, 9 data predicted by Patients but included in the Healthy Control classification, 50 data were predicted to be exact that is included in the Patients classification such as the confusion matrix table presented in Table III.

The results obtained from ROC processing for the Support Vector Machine model using training data of 0,819 can be seen in Figure IV.

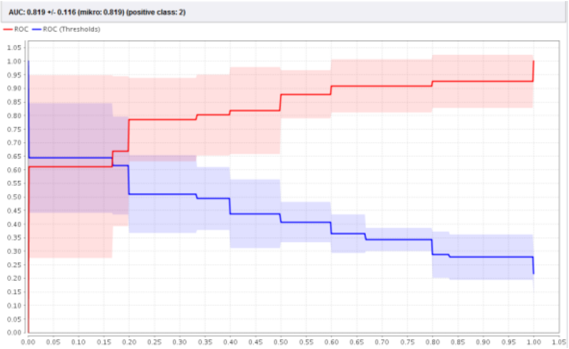


FIGURE IV. AUC MODEL SVM+ PSO CURVE

c. *Neural Network* (NN)

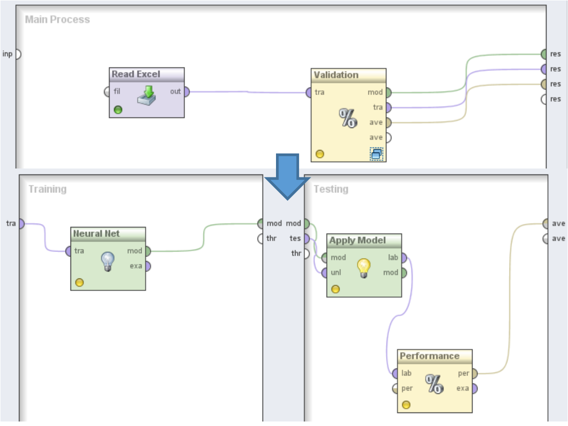
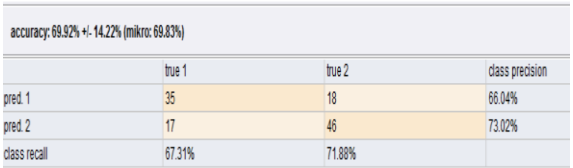


FIGURE V. TESTING OF NEURAL NETWORK

TABLE IV. VALUE OF ACCURACY DATA TRAINING MODEL NEURAL NETWORK



accuracy: 69.92% +/- 14.22% (mikro: 69.83%)

	true 1	true 2	class precision
pred. 1	35	18	66.04%
pred. 2	17	46	73.02%
class recall	67.31%	71.88%	

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

$$= \frac{(46 + 35)}{(46 + 18 + 35 + 17)}$$

$$= \frac{(81)}{(116)} \quad = 0{,}69 \quad = 69\%$$

Testing training data with Neural Network models obtained 69,92% accuracy according to the confusion matrix table presented in Table.IV out of 116 data as many as 35 data were predicted correctly that is included in the Healthy Control classification, and as many as 18 data predicted Healthy Control but it turned out to be included into the Patients classification, 17 data were predicted by Patients but included in the Healthy Control classification, 46 data were predicted to be exact that is included in the Patients classification such as the confusion matrix table presented in Table IV.

Next test with the Neural Network method produced in Figure VI by using three layers consisting of an input layer consisting of 9 nodes including age, BMI, Glucose, Insulin, HOMA, Leptin, Adiponectin, Resistin, MCP.1 plus one bias node. The second layer is a hidden layer consisting of seven vertices and one bias node. The third layer is the output layer, there are two nodes which represent the attributes of the healthy control and patient classes. Figure 6 is the result of a neural network experiment with one hidden layer consisting of nine vertices.

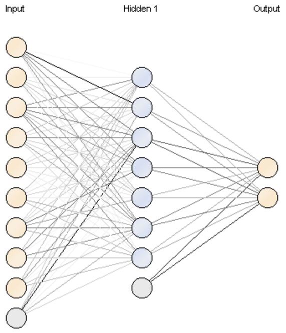


FIGURE VI.

NEURAL NET WITH SEVEN NODES IN HIDDEN LAYER

The results obtained from ROC processing for the Neural Network model using training data of 0,757 can be seen in Figure VII.

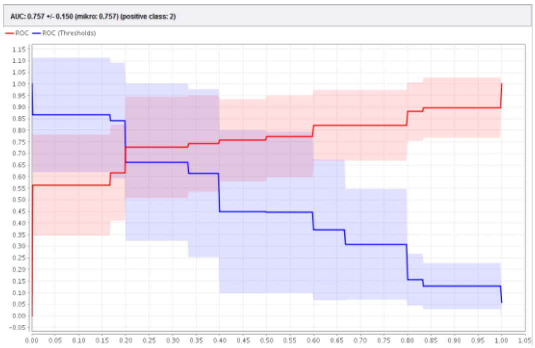


FIGURE VII. AUC MODEL NN

d. *Neural Network + PSO*

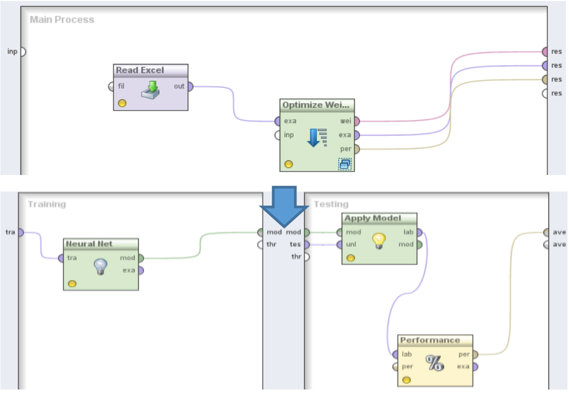
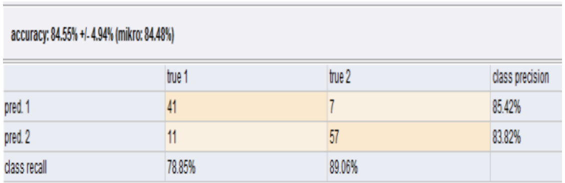


FIGURE VIII. TESTING OF NEURAL NETWORK+PSO

TABLE V. VALUE OF ACCURACY DATA TRAINING MODEL NN + PSO



accuracy: 84.55% +/- 4.94% (mikro: 84.48%)

	true 1	true 2	class precision
pred. 1	41	7	85.42%
pred. 2	11	57	83.82%
class recall	78.85%	89.06%	

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

$$= \frac{(57 + 41)}{(41 + 7 + 11 + 57)}$$

$$= \frac{(98)}{(116)} \qquad = 0{,}84 \quad = 84\%$$

The network obtained an accuracy of 84,55% according to the confusion matrix table presented in Table.V out of 116 data, 41 of them were predicted correctly, namely included in the Healthy Control classification, and as many as 7 data predicted by Healthy Control but turned out to be included in the Patients classification, 11 data Patients predicted but entered into the Healthy Control classification, 57 data predicted exactly that is included in the Patients classification such as the confusion matrix table presented in Table V.

The results obtained from ROC processing for Neural Network models using training data of 0,885 can be seen in Figure IX.

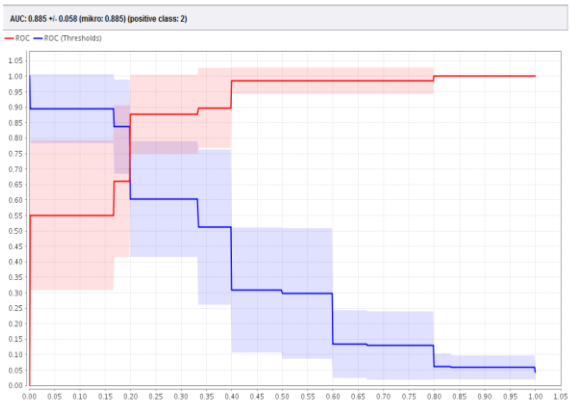


FIGURE IX. AUC MODEL NEURAL NETWORK+PSO

From data processing using rapidminer tools, a set of results is obtained as follows: Nilai Akurasi dan ROC Pada Metode SVM dan NN berbasis PSO

TABLE V. VALUE OF ACCURACY AND ROC ON SVM AND NN METHOD BASED ON PSO

Model	Accuracy		AUC	
	default	PSO	default	PSO
Support Vector Machine (SVM)	72,12%	80,08%	0,797	0,819
Neural Network (NN)	69,92%	84,55%	0,757	0,885

V. CONCLUSION AND SUGGESTION

a. Conclusion

This research compares two methods, namely Support Vector Machine and Neural Network based on Particle Swarm Optimization (PSO) which aims to measure the accuracy of the model. This study uses the Counfusion Matrix and ROC Curve testing methods.

Based on the measurement of the accuracy of the two models with PSO-based, it is known that the PSO-based Neural Network model has a higher accuracy value of 84.55% from the results of PSO-based Support Vector Machine with an accuracy value of 80.08%.

The calculation results, the accuracy of the AUC performance obtained by the results of the study are, the two methods are PSO-based Neural Network with AUC value of 0.885 and PSO-based Support Vector Machine with a value of 0.819 included in the category of Good Classification.

So it can be concluded that the PSO-based Neural Network method has a higher level of accuracy than the PSO-based Vector Support Machine. Thus, the PSO-based Neural Network method can be used for classification in breast cancer prediction.

b. Suggestion

This research is suggested to be able to develop further methods from previous studies and to make an application in predicting breast cancer.

VI. REFERENCES

Asria, Hiba, H. M., Moatassime, H. Al, C, & Noeld, T. (2016). Using Machine Learning Algorithms for Breast Cancer Risk Prediction and Diagnosis. *Procedia Computer Science*, *83*, 1064–1069. Retrieved from https://ac.els-cdn.com/S1877050916302575/1-s2.0-S1877050916302575-main.pdf?_tid=e2d06037-c161-4e91-a842-8ce3db6bd633&acdnat=1552990694_8bc70a70ef15b5bad610be0e54cc9e58

Biro Komunikasi dan Pelayanan Masyarakat, K. K. R. (2019). Kementerian Kesehatan Republik Indonesia. Retrieved September 3, 2019, from Biro Komunikasi dan Pelayanan Masyarakat, Kementerian Kesehatan RI website: http://www.depkes.go.id/article/view/19020100003/hari-kanker-sedunia-2019.html

Cao, J., Cui, H., Shi, H., & Jiao, L. (2016). Big data: A parallel particle swarm optimization-back-propagation neural network algorithm based on MapReduce. *PLoS ONE*, *11*(6), 1–17. https://doi.org/10.1371/journal.pone.0157551

Harafani, H. (2015). Optimasi Parameter pada Support Vector Machine Berbasis Algoritma Genetika untuk Estimasi Kebakaran Hutan. *Journal of Intelligent Systems*, *1*(2), 82–90.

Maimon, Oded&Rokach, L. (2010). *Data Mining and Knowledge Discovey Handbook*. New York: Springer.

Novianti, F. A., & Purnami, S. W. (2012). Analisis Diagnosis Pasien Kanker Payudara Menggunakan Regresi Logistik dan Support

Vector Machine (SVM) Berdasarkan Hasil Mamografi. *Jurnal Sains Dan Seni ITS*, *1*(1), D147--D152.

Patrício, M., Pereira, J., Crisóstomo, J., Matafome, P., Gomes, M., Seiça, R., & Caramelo, F. (2018). Using Resistin , glucose , age and BMI to predict the presence of breast cancer. *BMC Cancer*, *DOI 10.118*, 1–8. https://doi.org/10.1186/s12885-017-3877-1

Ridwansyah, & Purwaningsih, E. (2018). PARTICLE SWARM OPTIMIZATION UNTUK MENINGKATKAN AKURASI PREDIKSI PEMASARAN BANK. *Jurnal PILAR Nusa Mandiri*, *14*(1), 83–88.

Satapathy, S. C., Chittineni, S., Mohan Krishna, S., Murthy, J. V. R., & Prasad Reddy, P. V. G. D. (2012). Kalman particle swarm optimized polynomials for data classification. *Applied Mathematical Modelling*, *36*(1), 115–126. https://doi.org/10.1016/J.APM.2011.05.033

Vercellis, C. (2009). *Business Intelligent: Data Mining and Optimizzation for Decision Making*. Southern Gate, Chichester, West Sussex, United Kingdom: John Wiley & Sons Ltd.

Vieira, S. M., Mendonça, L. F., Farinha, G. J., & Sousa, J. M. C. (2013). Modified binary PSO for feature selection using SVM applied to mortality prediction of septic patients. *Applied Soft Computing*, *13*(8), 3494–3504. https://doi.org/10.1016/J.ASOC.2013.03.021

Witten, I. H., Frank, E., & Hall, M. A. (2011). *Data Mining : Practical Machine Learning and Tools*. Burlington: Morgan Kaufmann Publisher.

Zamani, A. M., & Amaliah, B. (2012). Implementasi Algoritma Genetika pada Struktur Backpropagation Neural Network untuk Klasifikasi Kanker Payudara. *Jurnal Teknik POMITS*, *1*(1), 1–6. https://doi.org/10.12962/j23373539.v1i1.638