BAREKENG: Jurnal Ilmu Matematika dan Terapan
March 2022 Volume 16 Issue 1 Page 129–136
P-ISSN: 1978-7227 E-ISSN: 2615-3017
https://doi.org/10.30598/barekengvol16iss1pp129-136

CLASSIFICATION SUPPORT VECTOR MACHINE IN BREAST CANCER PATIENTS

Siti Hadijah Hasanah[*]

Study Program of Statistics, Faculty of Science and Technology, Universitas Terbuka
Jl. Pd. Cabe Raya, Pd Cabe Udik Pamulang, Tangerang Selatan, 15418, Indonesia

Corresponding author e-mail: [1*] *sitihadijah@ecampus.ut.ac.id*

Abstract. *Support vector machine is one of the supervised learning methods in machine learning that is used in classification. The purpose of this study is to measure the accuracy of classification by using 4 hyperplane functions in SVM, namely linear, sigmoid, polynomial, and radial basis function (RBF). Based on the simulation results of training data and testing data on female breast cancer patients, SVM with hyperplane RBF has better accuracy than hyperplane polynomial, linear and sigmoid. The RBF results for the training and testing data were 89,1% and 73,2%, respectively. Based on the results of the classification of training data for female breast cancer patients, 88.07% had no recurrence and 93,33% had recurrence events. Meanwhile, based on the results of the classification of testing data, female patients did not recurrence events and recurrence events was 72,55% and 80,00%, respectively. So from this article, it can be concluded that SVM with hyperplane RBF is one of the best methods in the application of the method of classifying female breast cancer patients.*

Keywords: *support vector machine, radial basis function, classification*

Article info:

Submitted: 23rd November 2021. *Accepted: 5th February 2022*

How to cite this article:

S. H. Hasanah, "CLASSIFICATION SUPPORT VECTOR MACHINE IN BREAST CANCER PATIENTS", *BAREKENG: J. Il. Mat. & Ter.*, vol. 16, iss. 1, pp. 129-136, Mar 2022.

1. INTRODUCTION

Breast cancer is an abnormal growth that occurs in breast cells that can be felt as a lump or tumor [1]. Tumors occur when breast cells divide uncontrollably and produce additional tissue. A breast tumor can be benign or malignant (cancerous), a cancerous breast tumor can spread from within the breast tissue into the lymph nodes in the armpit and to other parts of the body. The cause of breast cancer is not known for certain, breast cancer can occur in women who have offspring with breast cancer or it can even occur in young women who are still menstruating.

Symptoms of breast cancer include a lump in the breast that does not cause pain, bleeding or unusual discharge from the nipple, pulling inward of the breast skin and areola, persistent itching and rash around the areola, and breast skin. swollen or thickened. Early examination of the breast can prevent women from getting breast cancer as early as possible or if it is found out as soon as possible, then these women immediately know the fast and appropriate action in dealing with the growth of abnormal cells in the breast. If you find a lump in the breast and suspect that there is a possibility of having breast cancer, it is advisable to consult a doctor and run tests such as mammograms and breast MRIs [2], [3].

Breast cancer is the second leading cause of death in women [4]. In the United States, there are approximately 250,000 women diagnosed with breast cancer [5]. Although the overall mortality of breast cancer patients has decreased in the country. Based on data obtained from the University Medical Center in the form of a dataset of breast cancer patients with a total of 286 records [6].

This article discusses scientifically using the Support Vector Machine (SVM) classification method which is useful for knowing the value of the classification comparison between women who did not recurrence events and recurrence events. suffering from breast cancer. The results of this classification itself can later be used for handling other patients which is useful for reducing the number of patients who experience recurrence of breast cancer. Several studies use the SVM method as a classification method, namely the application of SVM in economics [7], transportation [8], and social media data analysis [9].

2. RESEARCH METHODS

The data for this article was obtained from the University Medical Center in the form of a dataset of women with breast cancer with a total of 277 data records. This data is divided into training data (80%) and testing data (20%), using the Support Vector Machine (SVM) classification method which aims to classify women who recurrence events and do not recurrence events against breast cancer. The analytical method used in this article is the SVM method with the help of Python applications and the kernel functions used in SVM are linear, sigmoid, polynomial and RBF.

Table 1. Characteristics of Female Breast Cancer Patients

Variable	Description	Scale	Category
X1	Age	Interval	1 = 20-29 2 = 30-39 3 = 40-49 4 = 50-59 5 = 60-69 6 = 70-79
X2	Menopause	Nominal	1 = ge40 2 = lt40 3 = premeno
X3	Tumor size	Interval	1 = 0-4 2 = 5-9 3 = 10-14 4 = 15-19 5 = 20-24 6 = 25-29 7 = 30-34 8 = 35-39 9 = 40-44 10 = 45-49

Variable	Description	Scale	Category
			11 = 50-54
X4	Inv nodes	Interval	1 = 0-2 2 = 3-5 3 = 6-8 4 = 9-11 5 = 12-14 7 = 15-17 8 = 24-26
X5	Node caps	Nominal	1 = no 2 = yes
X6	Deg malign	Nominal	1 2 3
X7	Breast	Nominal	1 = left 2 = right
X8	Breast quad	Nominal	1 = central 2 = left low 3 = left up 4 = right low 5 = right up
X9	Irradiate	Nominal	1 = no 2 = yes
Y	Class	Nominal	0 = no recurrence events 1 = recurrence events

3. RESULTS AND DISCUSSION

3.1 Multicollinearity

The multicollinearity problem occurs when there is a linear relationship between the predictor variables. One way to indicate a linear relationship between independent variables is to calculate the correlation coefficient between the predictor variables [10]. If the value of the correlation coefficient exceeds 0.8, it indicates the existence of multicollinearity [11]. One way to overcome this multicollinearity is to eliminate predictor variables that have a high correlation [12].

3.2 Support Vector Machine (SVM)

Support Vector Machine (SVM) is one of the supervised learning methods used in classification (Support Vector Classification) and regression (Support Vector Regression). SVM in classification modeling has advantages compared to other classification techniques, namely, it can solve linear and non-linear classification and regression problems. SVM on a non-linear problem that uses the concept of a kernel in a high-dimensional space. In this dimensional space, the best hyperplane will be sought, namely by maximizing the distance between classes. The process of finding the best hyperplane is what is at the core of the SVM process. The most important hyperplane in SVM is the kernel, the kernels commonly used in SVM include the following [13]:

1. Linear
   Linear is the simplest kernel function that is used when the analyzed data is linearly separated. Here is the linear kernel equation.

   $$\mathrm{K}(\mathrm{x},\mathrm{z}) = \mathrm{x}^{\mathrm{T}}\mathrm{z}$$

2. Radial Basis Function (RBF)
   RBF is a kernel function that is used when the data is not linearly separated. This kernel has two parameters, namely Gamma and Cost. The Gamma parameter determines how far the influence of one sample training dataset is. While the Cost (C) parameter works as an SVM optimization to avoid misclassification in each sample in the training dataset. The following is the RBF kernel equation.

$$\mathrm{K(x,z)} = \exp[-\gamma\|x - z\|^2]$$

3. Sigmoid
   Sigmoid is a kernel function similar to the activation function in the neural network model. The following is the sigmoid kernel equation.

   $$\mathrm{K(x,z)} = \tanh(\gamma\, \mathrm{x^T z} + \mathrm{r})^d, \quad \gamma > 0$$

4. Polynomial
   Polynomial is a kernel function that is used when data is not linearly separated, generally, this kernel is applied to normalized datasets.

   $$\mathrm{K(x,z)} = (\mathrm{x^T z})^d \text{ atau } (1 + \mathrm{x^T z})^d$$

3.3 Normalization and Accuracy

Normalization is used to make the classification accuracy results are quite high [14]. Normalization in the SVM method is in the range of values from 0 to 1 in breast cancer data. Accuracy is the number of true or false predictions, which is obtained from the number of positive data that is predicted to be positive and negative data that is predicted to be negative divided by the total number of data in the dataset.

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

where,
TN = True Negative
TP = True Positive
FN = False Negative
FP = False Positive

3.4 Confusion matrix

Confusion matrix is a diagonal matrix consisting of correctly classified examples, while the other matrix elements show the number of examples that are misclassified as some other class and this confusion matrix is a fairly good way of analyzing various types of errors [15].

Table 2. Confusion matrix

Observation	**Prediction**		**Total**
	0	**1**	
0	TN	FP	n_0
1	FN	TP	n_1
Total			$\boldsymbol{n}$

The classification accuracy measure is formulated [16],
Accuracy = $(TN + TP)/n$,
Sensitivity = TP/n_1,
Specificity = TN/n_0.

where,
n_0 = TN + FP
n_1 = FN + TP
n = $n_0 + n_1$

Below are the results of data exploration based on descriptive statistics as follows:

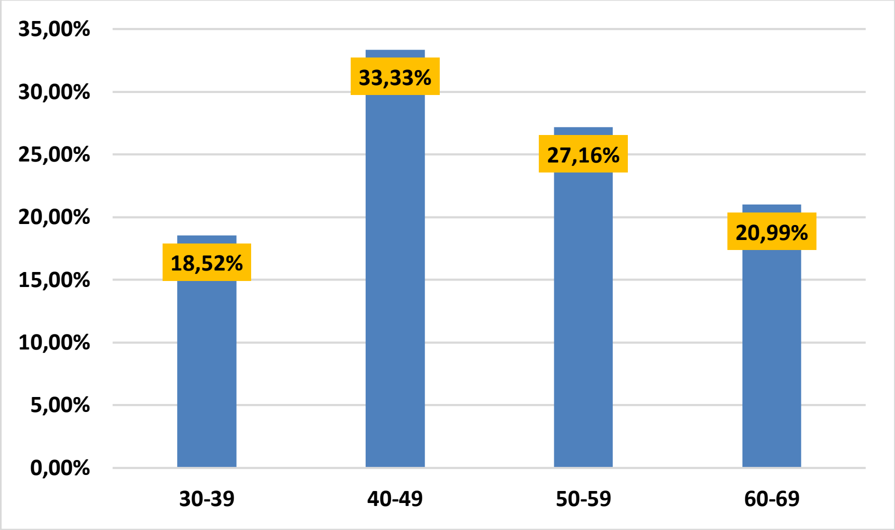

Figure 1. Breast cancer by age

Based on Figure 1, it is known that the age of a woman's risk of breast cancer begins when they are 30 years old and a woman's risk of breast cancer increases in the age range of 40-49 years. So based on these results, it is expected that women will be more aware of their breast health with breast self-examination and clinical breast examination.

3.5 Multicollinearity Test

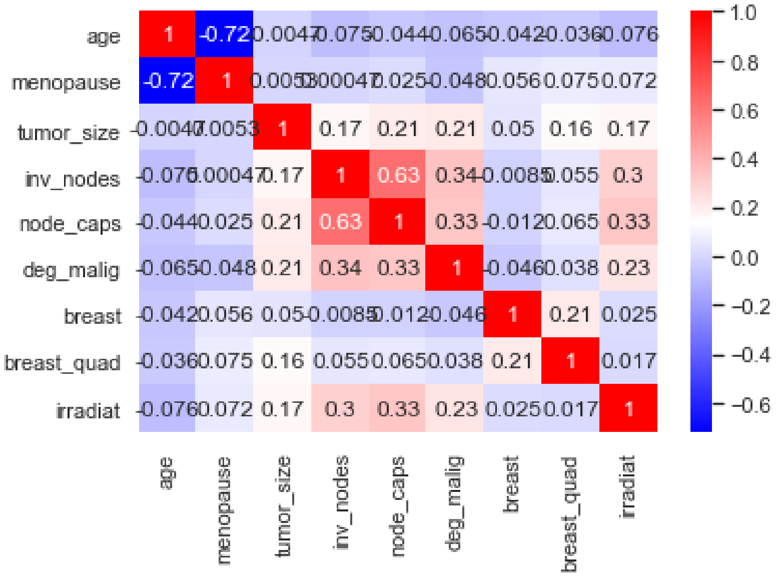

Figure 2. The relationship between independent variables

Based on Figure 2, it is known that the variable of the age with menopause has a very strong correlation value of 0,72, as well as a variable of the node_caps with inv_nodes, has a strong correlation value of 0,63. So to overcome the multicollinearity in this data, the authors decided to eliminate one of the variables that have a strong correlation value, namely eliminating the variable of menopause and inv_nodes. So in the SVM analysis, the indicator variables used in the SVM method are 7 variables.

3.6 SVM Simulation

Furthermore, data classification of recurrence events and no recurrence events breast cancer patients was carried out using the SVM method. Simulations were carried out using hyperplanes of linear, sigmoid, polynomial, and RBF. Then the application of SVM on training data and testing data is obtained as follows:

Table 3. Hyperplane Based on Accuracy

Data	Hyperplane			
	Linear	Sigmoid	Polynomial	RBF
Training	0,738	0,647	0,887	0,891
Testing	0,661	0,661	0,732	0,732

The accuracy results in table 3 show that the hyperplane with the RBF kernel function has a higher accuracy value than the hyperplanes of polynomial, linear and sigmoid in classifying breast cancer data. The RBF accuracy values for each training data and testing data are 89,1% and 73,2%, respectively. This means that 89,1% of the data were correctly predicted on the training data and 73,2% of the data were correctly predicted on the testing data.

Table 4. SVM Data Training Results

Class	Result of SVM (%)	
	No recurrence events	Recurrence events
No recurrence events	88,07	6,67
Recurrence events	11,93	93,33

Based on table 4, it was found that the percentage of SVM accuracy in the classification of training data in women who did not recurrence events with breast cancer was 88,07% and women who recurrence events with breast cancer were 93,33%. So it can be concluded that the SVM method is good for classifying training data.

Tabel 5. SVM Data Testing Results

Class	Result of SVM (%)	
	No recurrence events	Recurrence events
No recurrence events	72,55	20,00
Recurrence events	27,45	80,00

Based on table 5, it was found that the percentage of SVM accuracy in the classification of testing data for women who did not recurrence events with breast cancer was 72,55% and women who recurrence events with breast cancer was 80,00%, so it can be concluded that the SVM method good enough for testing data classification.

4. CONCLUSIONS

Based on the results of the analysis obtained the following conclusions:

1. There are 9 parameters used in the classification of recurrence and non-recurrence status of female breast cancer patients, namely age, menopause, tumor size, inv node, node caps, malignancy, breast, quad breast, and irradiation. In the multicollinearity test process there is a relationship between age and menopause as well as node_caps with inv_nodes, so to eliminate the effect of multicollinearity, 7 parameters are used without involving the menopause and inv_nodes parameters.
2. The best hyperplane applied to female breast cancer patient data is the radial function basis (RBF) with the classification accuracy results on training data 89,1% and testing data 73,2%. Based on the results of the classification of female patient training data, 88,07% did not experience a recurrence and 93,33%

experienced a recurrence. Meanwhile, based on the results of the classification of female patient examination data, 72,55% had no recurrence and 80,00% had a recurrence. So from this article, it can be concluded that SVM with hyperplane of RBF is one of the best methods in applying the classification method to female breast cancer patients.

REFERENCES

[1] National Cancer Institute, *Breast Changes*. National Institutes of Health (NIH), 2014.

[2] D. Roganovic, D. Djilas, S. Vujnovic, D. Pavic, and D. Stojanov, “Breast MRI, digital mammography and breast tomosynthesis: Comparison of three methods for early detection of breast cancer,” *Bosn. J. Basic Med. Sci.*, vol. 15, no. 4, pp. 64–68, 2015, doi: 10.17305/bjbms.2015.616.

[3] T. A. S. O. B. Surgeons, “Consensus Guideline on Diagnostic and Screening Magnetic Resonance Imaging of the Breast,” *Am. Soc. Breast Surg.*, pp. 76–99, 2017.

[4] P. S. K. K. S. Sharma, Ganesh N. Rahul Dave, Jyotsana Sanadya, “Various Types and Management of Breast Cancer: An Overview,” *J. Adv. Pharm. Tech. Res.*, vol. 1, no. 2, pp. 109–126, 2010.

[5] N. F. Idris and M. A. Ismail, “Breast cancer disease classification using fuzzy-ID3 algorithm with FUZZYDBD method: Automatic fuzzy database definition,” *PeerJ Comput. Sci.*, vol. 7, pp. 1–22, 2021, doi: 10.7717/PEERJ-CS.427.

[6] UCI Machine Learning, “UC Irvine Machine Learning Repository.” archive.ics.uci.edu (accessed Sep. 08, 2021).

[7] A. Krysovatyy, H. Lipyanina-Goncharenko, S. Sachenko, and O. Desyatnyuk, “Economic crime detection using support vector machine classification,” *CEUR Workshop Proc.*, vol. 2917, pp. 830–840, 2021.

[8] A. Jahangiri, “Developing a Support Vector Machine (SVM) Classifier for Transportation Mode Identification by Using Mobile Phone Sensor Data Collaborative Optimization and Planning for Transportation Energy Reduction View project,” no. February, 2014, [Online]. Available: https://www.researchgate.net/publication/270959050.

[9] N. Naw and A. C. Mon, “Social media data analysis in sentiment level by using support vector machine,” *J. Pharmacogn. Phytochem.*, vol. 7, no. 1S, pp. 609–613, 2018.

[10] J. I. Daoud, “Multicollinearity and Regression Analysis,” *J. Phys. Conf. Ser.*, vol. 949, no. 1, 2018, doi: 10.1088/1742-6596/949/1/012009.

[11] J. H. Kim, “Multicollinearity and misleading statistical results,” *Korean J. Anesthesiol.*, vol. 72, no. 6, pp. 558–569, 2019, [Online]. Available: https://stat.duke.edu/~kfl5/Lock_RREE_Results_2010.pdf.

[12] N. Herawati, K. Nisa, and E. Setiawan, “Regularized Multiple Regression Methods to Deal with Severe multicolinearity,” vol. 8, no. May 2018, pp. 167–172, 2018, doi: 10.5923/j.statistics.20180804.02.

[13] M. Awad and R. Khanna, *Efficient learning machines: Theories, concepts, and applications for engineers and system designers*, no. May 2016. 2015.

[14] S. Hasanah and S. Permatasari, “Metode Klasifikasi Jaringan Syaraf Tiruan Backpropagation Pada Mahasiswa Statistika Universitas Terbuka,” *BAREKENG J. Ilmu Mat. dan Terap.*, vol. 14, no. 2, pp. 243–252, 2020, doi: 10.30598/barekengvol14iss2pp249-258.

[15] J. Novakovic, A. Veljovi, S. Iiic, Z. Papic, and M. Tomovic, “Evaluation of Classification Models in Machine Learning,” *Theory Appl. Math. Comput. Sci.*, vol. 7, no. 1, pp. 39–46, 2017, [Online]. Available: https://uav.ro/applications/se/journal/index.php/TAMCS/article/view/158.

[16] A. Baratloo, M. Hosseini, A. Negida, and G. El Ashal, “Part 1: Simple Definition and Calculation of Accuracy, Sensitivity and Specificity.,” *Emerg. (Tehran, Iran)*, vol. 3, no. 2, pp. 48–9, 2015, doi: 10.22037/emergency.v3i2.8154.