Comparison of Kernel Support Vector Machine In Stroke Risk Classification (Case Study: IFLS data)

L R Safitri[1], N Chamidah[1,*], T Saifudin[1], and G T Alpandi [1]

[1] Faculty of Science and Technology, Universitas Airlangga, Indonesia

*Corresponding author's email: nur-c@fst.unair.ac.id

Abstract. Stroke s a disability main source and main disability source to lost years of disability-adjusted life. Currently the information technology development, especially the field of machine learning has an important role in early warning of various diseases, such as strokes. One of the methods used for stroke classifying is Support Vector Machine (SVM). In this study, we aim to compare several kernel functions in SVM such as linear, radial basis function(RBF), polynomial, and sigmoid for classifying stroke risk. We determine the best kernel based on accuracy, sensitivity, and specificity values. The result of this study shows that linear kernel function gives the best performance in classifying with values of classification accuracy 99.0%, specificity 100.0%, ,and sensitivity 97.0%. Those scores are the highest scores among the other kernel , that means the linear kernel function is the best method for classifying strokes risk.

1. Introduction

Strokes as one of non-communicable diseases has caused 74% of all deaths worldwide. The strokes is not only a primary resource of disability, but also a primary contributor to lost years of disability-adjusted life, particularly in both low-income country and middle-income country[1]. Based on the World Strokes Organization (WSO) report, there are more than 12.2 million new strokes each year. In global, every one in four people who have age over 25 will be attacked by strokes in their lifetime [2]. Currently, the development of information technology, especially the field of Machine Learning (ML), have an important role in the early prediction of various diseases, one of which is strokes. There are several classification methods in ML that can be used, one of them is the SVM (Support Vector Machine). As a supervised learning algorithm, SVM charts examples of training to spots in space such that the gap width between one category and another is maximum [3]. In SVM, the new examples are then charted into that same space and thought to belong one of the categories based on which side of the crack they fell. Assumption of this algorithm is that target is a nominal or ordinal variable where the feature variables can consist of continuous, nominal, or ordinal variables.

Researches on comparing SVM kernels has been done by several previous researchers, for examples a comparison study on the performance of different SVM's kernels for classifying multi-temporal full-polarimetric L-band SAR data in an agriculture region was discussed by [4]; the classification of Human Development Index (HDI) using Kernel SVM has been discussed by [5]; the comparison kernel functions in SVM has been studied by [6] and the results showed that among kernel functions investigated, the polynomial kernel function has the highest accuracy and specificity values that are 0.91 and 0.99, respectively; an optimal picture of how this kernel functions can be implemented in case of a polynomial or RBF method argument has been presented by[7] ; comparison the performances of linear kernel and polynomial kernel using SVM method has been investigated by [8] with the results that the

best performance of kernel in SVM method is polynomial kernel with an accuracy of 51.2%; the performance of scintillation detection based on SVM using different kernel functions was studied by [9]; and [10] presented a better performance of RBF kernel than linear kernel in SVM.

However, all these previous researches gave different conclusions related to which kernel function is the best among other kernel functions. Therefore, in this study, we proposed a more accurate stroke risk classification method based on covariate variables such as age, sex, hypertension, diabetes mellitus (DM), obesity, and smoking status by comparing several kernel functions in Support Vector Machine, namely linear, RBF, polynomial, and sigmoid by determining accuracy, sensitivity, and specificity as measures of the goodness of classification.

2. Research Methods

In the following we present explanations of the research methods which include a brief explanation of data and research variables, Support Vector Machine, and performance of classifications.

2.1. Data and Research Variables

In this study, to classify stroke risk based on age, sex, hypertension, diabetes mellitus (DM), obesity and smoking status we use a machine learning in SVM algorithm. Next, we use a dataset of IFLS (Indonesia Family Life Survey) provided on the web https://www.rand.org/well-being/social-and-behavioral-policy/data/FLS/IFLS/ifls5.html. The descriptions of data and research variables are presented in Table 1. Then, we select randomly select 400 observations from the IFLS that consist of 200 strokes and 200 non-strokes. For validation accuracy, we divide the dataset into two parts for each validation, namely 80% and 20% for training and testing, respectively and used 10 fold cross validation. Metrics of performances are calculated by testing the dataset, which contains data not processed by the model during training.

Table 1. Descriptions of Data and Research Variables

Variable			Scale	Category
Dependent	y	***Stroke***	Nominal	0: Non *Stroke*; 1: *Stroke*
Independent	x_1	**Gender**	Nominal	0: Male; 1: Female
	x_2	**Age**	Nominal	0: <45 years; 1: >45 years
	x_3	**Obesity**	Nominal	0: No; 1: Yes
	x_4	**Hypertension**	Nominal	0: No; 1: Yes
	x_5	**DM**	Nominal	0: No; 1: Yes
	x_6	**Smoking Status**	Nominal	0: No; 1: Yes

Further, for reliable results, we use a 10-fold validation method by dividing the dataset into 10 subsets. Here, we use four subsets to train the model to be representative and has the power of generalization for each validation fold, and the model is validated with the remaining subsets. Next, we did five validations, and the average of all results is represented by the performance matrix. Figure 1 illustrates procedure of 10-fold validation.

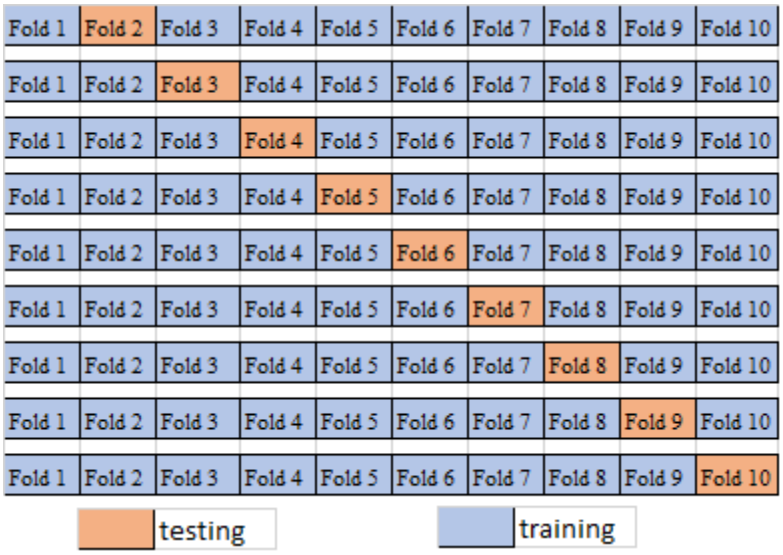

Figure 1. Procedure of five-fold validation.

2.2. Support Vector Machine

A classification method commonly used to form a binary non-probabilistic classification is Support Vector Machine. Principally, in this Support Vector Machine method, results of the training model are used to determine the best hyperplane for classifying data [12]. In classifying using SVM, it is necessary to have both training and testing stages. The main purpose of this SVM method is to determine the optimal classifier function that can be used to separate two different datasets [3]. If a hyperplane is caught in the middle of two objects from both classes, then that hyperplane is the best hyperplane or separator function. Formula for linear kernel function in Support Vector Machine is presented as follows:

$$\mathbf{w}\mathbf{x}^T + \boldsymbol{\gamma} = \mathbf{0}\,. \tag{1}$$

This SVM manipulates the model to allow linear domain division. SVM can be divided into linear and nonlinear models [13]. There are many techniques of ML or data mining developed under assumptions of linearity. This results in the resulting algorithm also only limited to linear cases. Generally, cases that occur in the real world are not non-linear cases. To overcome this non-linearity, kernel methods can be used [14].

A kernel function is a function given the original feature vector, returning a value equal to the dot product of the corresponding feature vectors are mapped. The feature vectors cannot be hided into a higher dimensional space explicitly by kernel function. Also, the dot product of the mapped vectors cannot be calculated by kernel function. The kernel returns the equal value using an unequal operations set which can frequently be calculated more efficiently. The essential reason we use kernel functions is to remove the need for processing to obtain a vector space with higher dimensions than a defined underlying space of vectors, which allows data to be detached linearly in higher dimensions. The following are kernel functions commonly used in SVM [3], [15]:

2.2.1 Linear. The linear kernel function is expressed as follows:

$$K\left(x, y\right) = x^T y\,. \tag{2}$$

This function obviously does not change native representation and does not get over the linearity constraints of linear classification and linear regression models. However, this allows linearity of dot product based algorithms (such as linear support vector machines and linear support vector regression algorithms) to be considered as special cases of suitable kernel based algorithms.

2.2.2. Polynomial. Polynomial kernel functions, especially those with a degree of two, are widely used for classification purposes. For example, Vladimir N Vapnik, an SVM creator, built a degree of two kernel function to classify handwritten numbers. The following is the formula for the polynomial kernel function:

$$K(x, y) = (\alpha x^T + y), \alpha > 0 \tag{3}$$

2.2.3. Radial Basis Function (RBF) (also called Gaussian). The RBF or Gaussian kernel is the best choice for problems requiring non-linear models. A decision limit in the feature space that are mapped, namely a hyperplane, is similar to a decision limit in the genuine space, namely a hypersphere. The space of feature generated by the RBF or Gaussian kernel is able to have dimensions with infinite number, a feat that would have been unlikely otherwise. The RBF or Gaussian kernel function follows formula as follows:

$$K(x, y) = \exp(\alpha \|x - y\|), \alpha > 0 \tag{4}$$

2.2.4. Sigmoid. The sigmoid kernel function can be written as follows:

$$K(x, y) = \tanh(\alpha x^T \, y + h) \tag{5}$$

The sigmoid function has gained popularity for the kernel approach because it is often used as the activation function for neural networks (multilayer perceptrons). If we use it correctly, it will similar to the family of RBF kernel. It can describe complex nonlinear interactions with multiple parameters. In some parameter configurations, it resembles a RBF kernel. This sigmoid kernel, however, probably not really represent a suitable kernel for some parameters because it is not completely positive.

2.3. Performance of Classifications

Confusion matrix is an algorithm used to measure the confusion matrix to describe the performance of the method (system) with the matrix. In the case of binary classification, the output matrix form is presented in Table 2.

Table 2. Confussion Matrix 2X2 of Binary Classification

Actual	**Prediction**	
	Positive	**Negative**
Positive	TP (True Positive)	FN (False Negative)
Negative	FP (False Positive)	TN (True Negative)

Based on the confusion matrix presented in Table 2, we can calculate values of accuracy, sensitivity and specificity, respectively by using the following formulas [16]:

$$\text{Accuracy} = \frac{\text{TP+TN}}{\text{TP+TN+FP+FN}} \times 100\% \tag{6}$$

$$\text{Sensitivity} = \frac{\text{TP}}{\text{TP+FN}} \times 100\% \tag{7}$$

$$\text{Spesificity} = \frac{\text{TN}}{\text{TN+FP}} \times 100\% \tag{8}$$

where TP represents true positive (or recall), which is the amount of data that is correctly classified from the positive class; TN represents true negative, namely the amount of data that is correctly classified from the negative class; FP represents a false positive, namely the amount of data that is predicted to be in the positive class but actually belongs to the negative class; and FN represents a false negative, namely the amount of data that is predicted to be in the negative class but actually belongs to the positive class.

3. Results And Discussion

In the following we present the results and discussion which are provided in two sub-sections, those are bivariate analysis and Comparison of kernel SVM for classifying strokes risk.

3.1. Bivariate Analysis

Before discussing the SVM algorithm, it is necessary to determine which factors have a statistically significant correlation with stroke risk. Next, we used the open source software R to perform bivariate analysis using the Chi-Square test and the results of analysis are given in the following Table 3.

Table 3. Bivariate Analysis Toward Strokes Risk

		Strokes Risk				P Value
		No (n=200)		Yes (n=200)		
		n	%	n	%	
Gender	**Male (n =304)**	196	64.5%	108	35.5%	<0.001***
	Female (n= 96)	4	4.2%	92	95.8%	
Age	**<45 (n= 228)**	196	86.0%	32	14.0%	<0.001***
	>45 (n= 172)	4	2.3%	168	97.7%	
Obesity	**No (n= 312)**	194	62.2%	118	37.8%	<0.001***
	Yes (n= 88)	6	6.8%	82	93.2%	
Hypertension	**No (n= 254)**	190	74.8%	64	25.2%	<0.001***
	Yes (n= 146)	10	6.8%	136	93.2%	
DM	**No (n= 367)**	196	53.4%	171	46.6%	<0.001***
	Yes (n= 33)	4	12.1%	29	87.9%	
Smoking Status	**No (n= 307)**	196	63.8%	111	36.2%	<0.001***
	Yes (n= 93)	4	4.3%	89	95.7%	

*** significant at level 0.1%

Based on the results of the Chi-square test in Table 3, we conclude that all covariate variables in this study (gender, age, hypertension, diabetes mellitus, obesity, and smoking status) have a statistically significant relationship with strokes risk.

3.2. Comparison of kernel SVM for Classifying Stroke Risk

We use Phyton to run the analysis for find the best kernel in classifying stroke risk. The comparison performances of four kernels in 10-fold validation are given in Table 4.

Table 4. Comparison Performance of Four Kernels in SVM

Kernel	Accuracy	Sensitivity	Specificity
Linear	0.99	0.97	1.00
RBF	0.85	0.72	1.00
Polynomial	0.89	0.79	1.00
Sigmoid	0.45	0.28	0.48

Table 4 shows the comparison performances of four kernels, i.e., linear, polynomial, RBF, and sigmoid based on values of accuracy, specificity, and sensitivity. From accuracy and sensitivity, kernel linear has the highest score among other, but from specificity, 3 kernels such as linear, RBF, and polynomial has the best score compared to sigmoid kernel. From this results, we conclude that linear kernel function in SVM gives the best performance in classifying strokes risk with 10-fold validation.

This results are supported with other researches that find linear kernel is best kernel among other such as [17], and [18]. The linear kernel obviously leaves the original representation unchanged and does not overcome the linearity limitations of linear classification and linear regression models in any way. However, it is possible to consider linear dot product-based algorithms (such as linear support vector machines and linear support vector regression algorithms) as special cases of the corresponding kernel-based algorithms [15].

Here are the the heatmap of confussion matrix for linear kernel SVM which are the best kernel in SVM for classifying stroke.

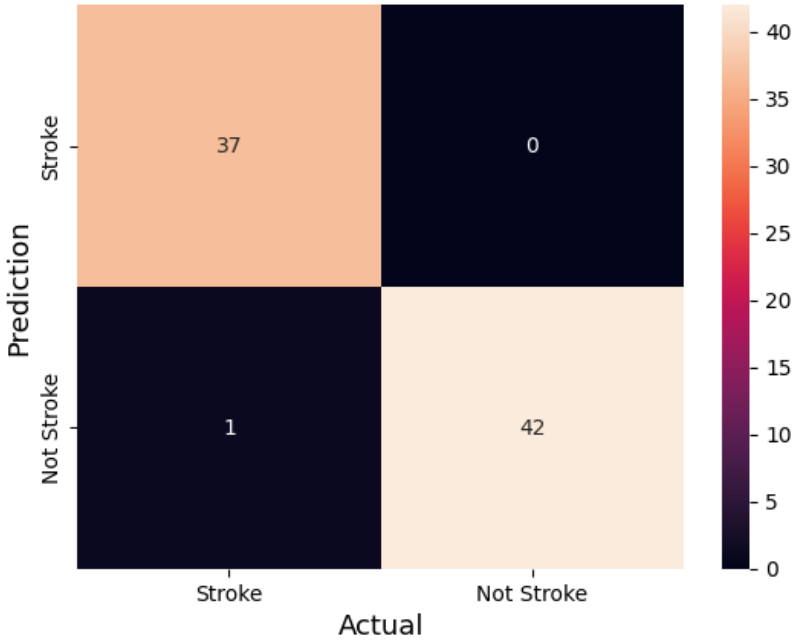

Figure 2. Heatmap of confussion matrix for linear kernel SVM
The Importance variable of SVM based on linear kernel can be seen in Figure 3

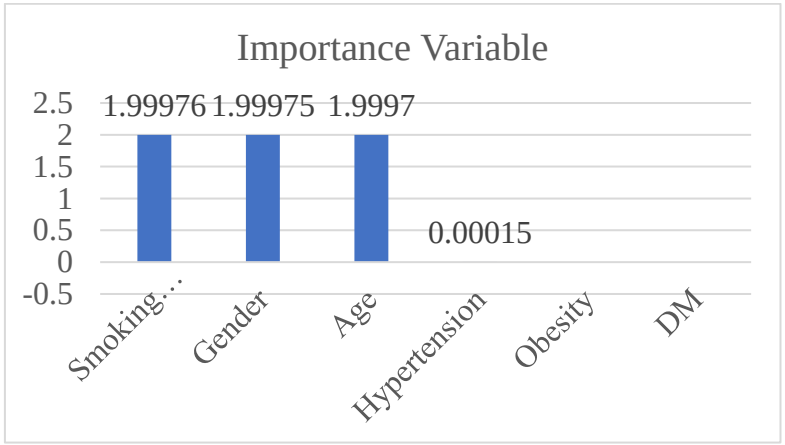

Figure 3. The Importance variable of linear kernel SVM

Figure 3 show the importance variable analyzing the variable weights provided by the SVM model to identify the most influential variabels in classification. The greater the weight of a feature, the more important the feature is in determining the decision boundary. Based on Figure 3**,** we can seen that the

order of variables that importance in model is smoking sttatus, gender, age, hypertension, obesity and the last is DM.

4. Conclusions

The linear kernel function has the best performance for classifying stroke risk compared to the RBF kernel, the polynomial kernel. and the sigmoid kernel function in the SVM method with an accuracy value of 99.0%, specificity value 97.0%; and a sensitivity value of 100.0%. However, Kernel performance depends on the specific characteristics of the dataset you are dealing with. In some cases, non-linear kernels such as RBF or polynomial kernels may be more suitable for solving complex problems with non-linear constraints. The results of this research can be implemented as an additional reference regarding the best methods in classification particularly on stroke data. For further research it is suggested to use other method such as naive bayes, random forest, decision tree, etc, and also add other covariates such as physical activity, stress level, and eating habits.

References

[1] WHO. World Stroke Day [Internet]. 2021. Available from: https://www.who.int/southeastasia/news/detail/28-10-2021-world-stroke-day

[2] Feigin VL, Brainin M, Norrving B, Martins S, Sacco RL, Hacke W, et al. World Stroke Organization (WSO): global stroke fact sheet 2022. International Journal of Stroke. 2022;17(1):18–29.

[3] Dangeti P. Statistics for machine learning. Packt Publishing Ltd; 2017.

[4] Yekkehkhany B, Safari A, Homayouni S, Hasanlou M. A Comparison Study Of Different Kernel Functions For Svm-Based Classification Of Multi-Temporal Polarimetry Sar Data. Int Arch Photogramm Remote Sens Spatial Inf Sci. 2014 Oct 22;XL-2/W3:281–5.

[5] Al Azies H, Trishnanti D, Mustikawati P.H E. Comparison of Kernel Support Vector Machine (SVM) in Classification of Human Development Index (HDI). IJPS. 2019 Dec 30;0(6):53.

[6] Rochim AF, Widyaningrum K, Eridani D. Comparison of Kernels Function between of Linear, Radial Base and Polynomial of Support Vector Machine Method Towards COVID-19 Sentiment Analysis.

[7] Panja S, Chatterjee A, Yasmin G. Kernel Functions of SVM: A Comparison and Optimal Solution. In: Luhach AK, Singh D, Hsiung PA, Hawari KBG, Lingras P, Singh PK, editors. Advanced Informatics for Computing Research [Internet]. Singapore: Springer Singapore; 2019 [cited 2023 Feb 7]. p. 88–97. (Communications in Computer and Information Science; vol. 955). Available from: http://link.springer.com/10.1007/978-981-13-3140-4_9

[8]. Mukarramah R, Atmajaya D, Ilmawan LB. Performance comparison of support vector machine (SVM) with linear kernel and polynomial kernel for multiclass sentiment analysis on twitter. Ilk J Ilm. 2021 Aug 8;13(2):168–74.

[9] Savas C, Dovis F. The Impact of Different Kernel Functions on the Performance of Scintillation Detection Based on Support Vector Machines. Sensors. 2019 Nov 28;19(23):5219.

[10] Ardhani BA, Chamidah N, Saifudin T. Sentiment Analysis Towards Kartu Prakerja Using Text Mining with Support Vector Machine and Radial Basis Function Kernel. JISEBI. 2021 Oct 28;7(2):119.

[11] Kim H, Jeon J, Han YJ, Joo Y, Lee J, Lee S, et al. Convolutional Neural Network Classifies Pathological Voice Change in Laryngeal Cancer with High Accuracy. JCM. 2020 Oct 25;9(11):3415.

[12] Tripathy A, Agrawal A, Rath SK. Classification of Sentimental Reviews Using Machine Learning Techniques. Procedia Computer Science. 2015;57:821–9.

[13] Suthaharan S. Support Vector Machine-Machine Learning Models and Algorithms for Big Data Classification. Integrated Series in Information Systems. 2016;36.

[14] Schölkopf B, Smola A. Support vector machines and kernel algorithms. In: Encyclopedia of Biostatistics. Wiley; 2005. p. 5328–35.

[15] Cichosz, Pawel,(2015) “Kernel methods,” in Data Mining Algorithms, Chichester, UK: John Wiley & Sons, Ltd, pp. 454–497. doi: 10.1002/9781118950951.ch16.
[16] Goel A, Srivastava SK. Role of kernel parameters in performance evaluation of SVM. In: 2016 Second international conference on computational intelligence & communication technology (CICT). IEEE; 2016. p. 166–9.
[17] Baitharu TR, Pani SK. Comparison of Kernel Selection for Support Vector Machines Using Diabetes Dataset. Journal of Computer Sciences and Applications.
[18] Intan PK. Comparison of Kernel Function on Support Vector Machine in Classification of Childbirth. JMM. 2019 Oct 27;5(2):90–9.