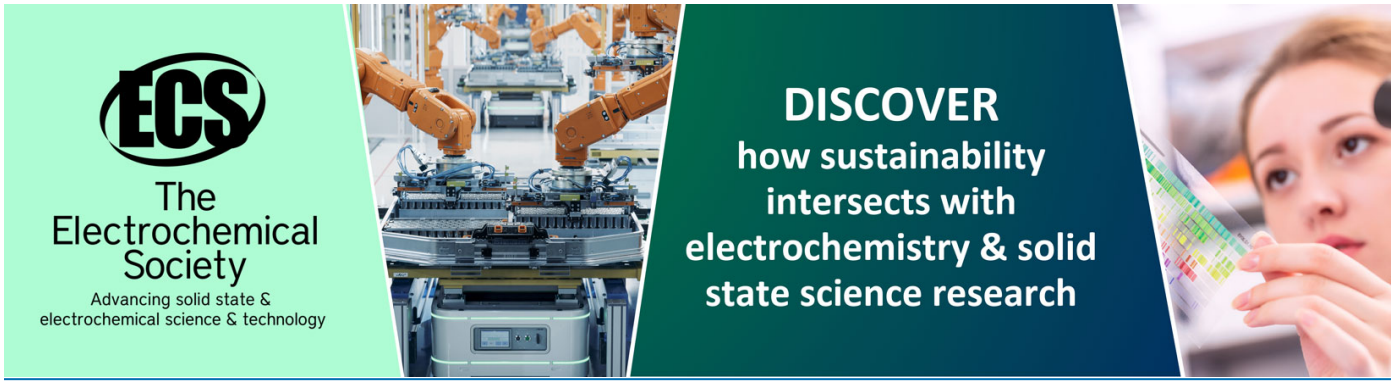
Comparison of Cubic SVM with Gaussian SVM: Classification of Infarction for detecting Ischemic Stroke



IOP Conf. Series: Materials Science and Engineering **546** (2019) 052016 doi:10.1088/1757-899X/546/5/052016

Comparison of Cubic SVM with Gaussian SVM: Classification of Infarction for detecting Ischemic Stroke

Amanda Rizki Bagasta[1], Zuherman Rustam[1*], Jacub Pandelaki[2], Widyo Ari Nugroho[2]

[1]Department of Mathematics, FMIPA Universitas Indonesia, Kampus UI Depok, Depok 16424, Indonesia

[2]Medical Department of Radiology, RSUPN Dr. Cipto Mangunkusumo, Kota Jakarta Pusat, DKI Jakarta 10430, Indonesia

Corresponding author's email: rustam@ui.ac.id

Abstract. Ischemic Stroke is a condition whereby the blood supply to the brain is disrupted or reduced due to a blockage and if it is not treated immediately will cause the death of the brain. A decrease in blood flow resulting in dead brain tissue can be called an infarction. The classifications of infarction help the health sector in detecting ischemic stroke in patients. In medicine, CT scans can be used to identify Infarctions and for detecting Ischemic Stroke in patients. Therefore, studying the CT scans is crucial in helping doctors obtain functional information about the surrounding brain tissues which will be used for detecting infarction in the brain. Since it is important to pay more attention at the time of choosing the best method that gives the best results, therefore this study proposes to compare between two types of methods, Gaussian Support Vector Machine (Gaussian SVM) and Cubic Support Vector Machine (Cubic SVM). The Cubic SVM could be an efficient method for infarction classification with accurate performances as high as 80%.

Keywords:infarction; Ischemic stroke; CT Scan; Cubic SVM; Gaussian SVM

1. Introduction

Ischemic stroke (Cerebrovascular Infarction), commonly known as brain attack, has become one of the main concerns in global health problems. High rates of stroke require immediate and proper strategies of treatment. Globally, Ischemic Stroke is the most commonly occurring stroke, if an artery carrying adequate amounts of blood to the brain is blocked, that might lead to particular brain tissues dying and long-term disabilities [1]. According to the World Health Organization data published in 2016, stroke has been one of the top 10 leading causes of death globally with approximately 5 million deaths [2]. The lack of blood flow leads to severe damage (infarct) to some of the brain tissue and it can be detected and analyzed through the use of the CT Scan or MRI (Magnetic Resonance Imaging) [3]. This study would be focused on analyzing images from the CT Scan of Cerebrovascular Infarction in patients. As a life-saving medical diagnostic tool, CT scan offers many advantages that will help the health sector, most especially it has been frequently used for clinically evaluating stroke.

CT scans are a form of X-Ray imaging created by a computer which processes the data and converts it into X-Ray. The description of imaging CT Scans of various layers in multiple ways is conducted by analyzing the measure and density of the substance through X-Ray. For certain body tissues, the amount of attenuation is relatively constant and well known as the network damping coefficient, which is related with the Hounsfield Unit (HU) that plays an important role for predicting Cerebrovascular Infraction in the CT Scan of the brain we are talking about [4].

Imaging of certain body tissues are radiodensity of air at STP = -1000 Hounsfield Unit with imaging color of black, radiodensity of cerebrospinal fluid at STP = 0 Hounsfield Unit with imaging color of black, radiodensity of brain at STP = 30 Hounsfield Unit with imaging color of grey, radiodensity of blood at STP = 100 Hounsfield Unit with imaging color of white, radiodensity of bones at STP = 1000 Hounsfield Unit with imaging color of white at STP: standard temperature is 0° and pressure is 10^5 Pascals [5].

By analyzing the image of the CT Brain, the presence or absence of infarction in the patient can be observed. For this study, the Hounsfield Unit would be used in the tissue comparison, whereby the Hounsfield Unit is used by taking one point from the abnormal tissue side of the brain and another from the normal tissue side of the brain which is collateral from one point.

Several researchers have shown that, machine learning technologies has played an important role in the application of classification, prediction and pattern recognition on large datasets that can be applied to daily life such as character/facial recognition, gaming, robotics, and biomedical data [6]. Machine learning is a technique that enables computers to learn from an experience in the form of training data that is then converted into an optimization algorithm that can be used to solve problems and produce accurate predictions. [7] Therefore, this study will be focused on machine learning to be the solution to classifying Ischemic Stroke in the patient by analyzing points from the abnormal and normal tissues from the brain. This study works with CT brain scans of 206 patients from Dr. Cipto Mangunkusumo National Central General Hospital, Department of Radiology. In this study, the classification algorithms will be applied to the entire datasets with the intention to diagnosis the probabilities of the existence of infarction in the brain. This study compares the infarction from SVM with kernel Gaussian and kernel Cubic to analyze Cerebrovascular Infarction of CT brain scan datasets with dozens 7 features

2. Experimental procedures

2.1. Proposed Method

In this study, comparisons would be made between the Support Vector Machine (SVM) with kernel RBF and Support Vector Machine (SVM) with kernel Cubic (Polynomial) to run classification algorithms in large sets so as to obtain more efficient results.

2.2. Datasets

In this study, datasets of CT scans of the brain are used. The datasets are taken from the Department of Radiology, Dr. Cipto Mangunkusumo National Central General Hospital. This study entails the use of 7 Features and also analyzing from an input of computerized images of a CT scan of the brain. The data defines characteristics of the particular brain displayed from the computerized Ct scan image and also the diagnosis of patient which is abnormal tissues (has infarction) or normal tissues (has no infarction). Using the description CT brain scan of seven features:

- Area (cm^2) : The size of the area from the infarction point
- Minimum : Minimum value of infarction
- Maximum : Maximum value of infarction
- Average : Average value of infarction
- SD : Standard error value of infarction
- Sum : Total amount value of infarction
- Length (cm) : Length of infarction point

This dataset is provided from Dr. Cipto Mangunkusumo National Central General Hospital. The following images are given one point of example from the abnormal (has cerebrovascular infarction) and normal tissues (has no cerebrovascular infarction) with collateral position. See Figure1 and 2.

Previous researches have shown that some other ischemic stroke dataset from different sources have been diagnosed with machine learning techniques such as implementing Support Vector Machine (SVM) with multilayer perception kernel, automated SVM, and manual SVM. These methods are consistent with an accuracy estimated to be about 60% to 75% [9]

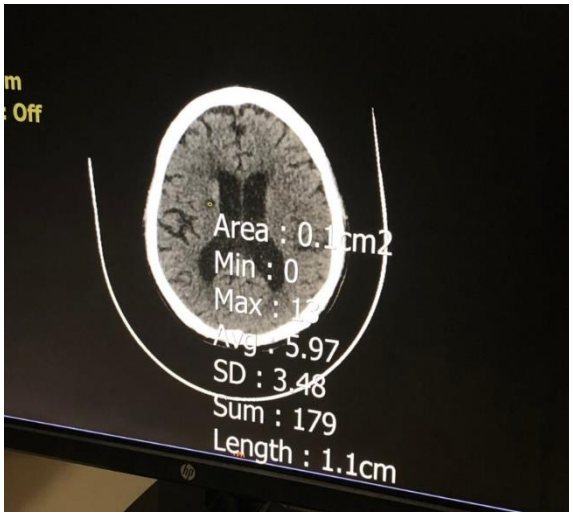


Figure 1. Normal tissues (has no infarction)

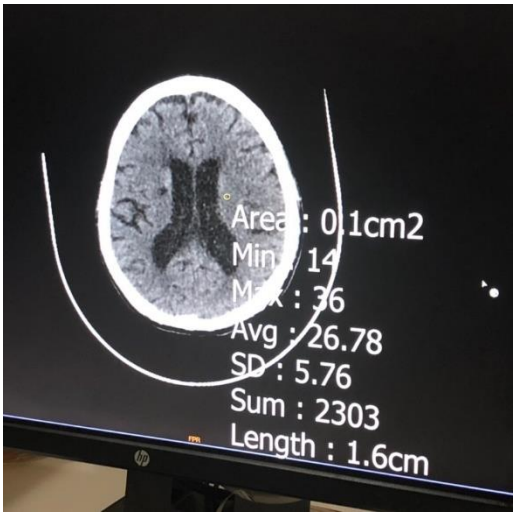


Figure 2. Abnormal tissues (has infarction)

Therefore, this study proposes to explore more possibilities of efficient machine learning algorithms for large datasets. The Support Vector Machine (SVM) with kernel RBF and kernel Cubic Polynomial are used. This study was ethically approved by the Department Radiology, Dr. Cipto Mangunkusumo National Central General Hospital but did not require patient consent for we made use of non-identifiable data.

2.3. Support Vector Machine (SVM)

Multiclass Classification is a process that aims to optimize an object into some class or category which are already decided beforehand, for example in a classification, the machine will help to approximate a function by mapping a vector into classes by considering that output of example from function [10]. As the definition of classification system which is a structured organizer used to determine groups based on similar characteristics of their attributes values, it is supposed to be the main concern of the steps from classifier design is data processing. In machine learning methods, whenever a conceptual error occurs, this might potentially create room for over fitting due to the data not being well generalized, making the evaluation of the accuracy of a classification crucial. [11]. The classification process is additionally divided into two phases: training and testing. The training set occurs as the algorithm uses information from the training set to build a classification model and the testing set occurs as the algorithm is allowed to see the actual class of just-classified examples [12]. One of the major purposes of this classification algorithm is to maximize the predictive accuracy of the classification model resulting from the training phases. It is crucial that the classification system can also be used for other important matter or data for example biomedical imaging data. Additionally, in this study, the following types of classification which is Support Vector Machines Method (SVM) and addition of various kernels were used.

SVM is a machine learning algorithm which is related to the classification or regression of data by learning and predicting from data training. But for some of cases, SVM is frequently used in

classification, corporate finances, motors and circuit diagnosis, enterprise market competition, and artificial intelligence. The SVM has been considered as one of the most effective and convenient method of supervised technology in machine learning [13]. The SVM method, chooses data points that lie closest to the classification data [14]. By finding the hyperplanes that differentiate the two classes very well, it will then proceed and produce the value of accuracy which depends on the value of kernel and parameters that were used. Therefore, this study uses SVM Non-Linear and adjusts with kernel of Gaussian and Cubic (polynomial). In following section, we will simply discuss the basic concepts involved in using SVM to produce two-class classification.

First, have m separable training sample $S = \{(x_i, t_i)\}_i = 1, \ldots, m$, where $x_i \in R^n$ and n - dimensional training vectors and $t_i = \{-1,1\}$ are corresponding labels [15]. The main purpose is to find the hyperplane that can separate the two classes with maximum margin, the model equation defining the decision surface separating the classes is a hyperplane of the form [15]

$$w^T.x + b = 0 \quad \ldots(1)$$

Allows us to write

$$w^T.x + b \geq 0 \qquad for\ t_i = +1 \quad \ldots(2)$$
$$w^T.x + b < 0 \qquad for\ t_i = -1 \quad \ldots(3)$$

where w is a weight vector, x is input vector, and b is bias.
To find the maximum distance between hyperplane, can be written of the form

$$t_i(w.x_i + b) \geq 1, \quad 1 \leq i \leq m \quad \ldots(4)$$

From this study binary SVM was used. W is denoted as $W = w_1, w_2, \ldots, w_l$as our dataset, where $w_i = (x_i, y_i)$ and $i = 1, \ldots, m$ and$y\epsilon\{1, \ldots, n\}$. The response y_i is the class of predictor vectorx_i. To solve the SVM objective function is used. [16]

$$\min_{\vec{W}, b, \xi} \frac{1}{2}(W^T W) + C \sum_{i=1}^{l} \xi_1 \quad \ldots(5)$$

Subject to

$$t_i(W^T\, \phi(x_i) + b) \geq 1 - \xi_1, \qquad \xi_1 \geq 0, i = 1, \ldots, l \quad \ldots\ (6)$$

For this study non-linear Support Vector Machine and soft margin SVM are used because we also predict that datasets that cannot be separated without misclassification or errors, where K is the kernel function and C is penalty parameter to control the tradeoff between large margin and classification error. We minimize$\frac{1}{2}K(W, W)$, The margin between two groups of data is maximized. While $\min_{\vec{W}, b, \varepsilon} \frac{1}{2}\|w\|^2$, this imposes a preference over the hypothesis space and pushes for better generalization. From equation (2), b is scalar, and W is p-dimensional vector. Vector W shows as the perpendicular to the separating hyperplane. By adding parameter b allows us to increase the margin. The optimal w satisfies [16]

$$w = t_i \alpha_i \phi(x_i) \quad \ldots(7)$$

And the decision function is [16]

$$sgn\,(w^T \phi(x) + b) = sgn\,(\sum_{i=1}^{l} t_i \alpha_i K(x_i, x) + b) \quad \ldots(8)$$

Where α_i are called support vectors (SVs) and primal problem $\alpha_i \geq 0$.

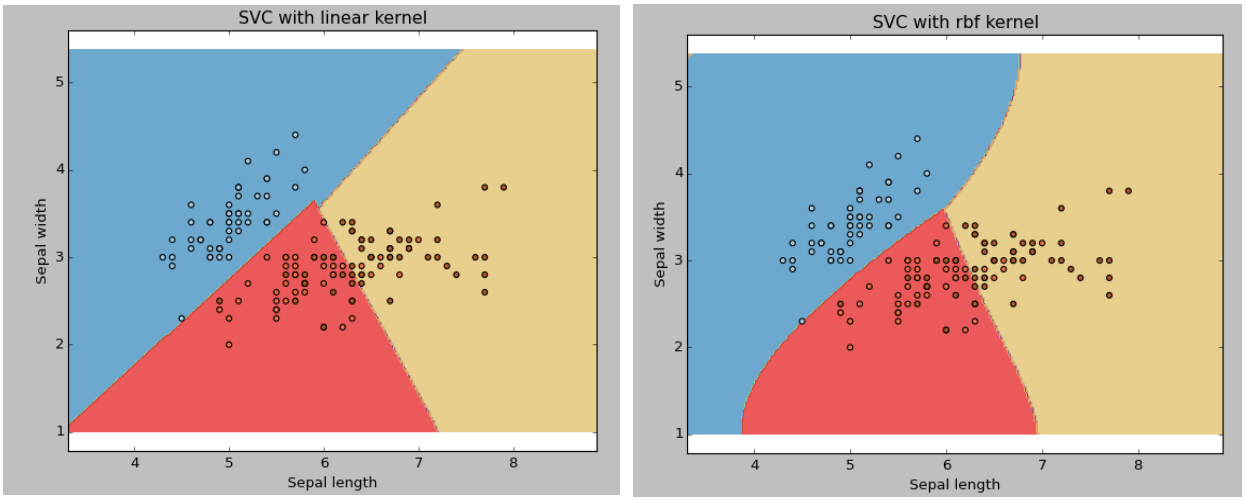


Figure 3. Comparison of linear SVM with non-linear SVM (RBF-SVM) [17].

This algorithm is modified by adding kernel function into non-linear SVM method (see figure 3). By using kernel function, the expectation is that this method will give better results of accuracy. In this study, the kernel used is the Radial Basis Function / Gaussian (RBF) kernel and Cubic (Polynomial degree 3) kernel, then we compare between these two kernels and find out which kernel has good performance on certain parameters with minimum training errors.

RBF Kernel $$K(\vec{x_i}, \vec{y_i}) = exp\left(-\frac{||\vec{x_i} - \vec{y_i}||^2}{\sigma^2}\right) \quad \ldots(9)$$

Cubic Kernel $$K(\vec{x_i}, \vec{y_i}) = \left(\vec{x_i}^T \vec{y_i} + 1\right)^3 \quad \ldots(10)$$

Some of the advantages of SVM is its ability to process large amounts of data in high dimension also this method can be easily implemented because of the process of determining vector support can be formulated in a problem.

2.4. Performance Evaluation Methods

To evaluate the performance of Cubic SVM and Gaussian SVM, a confusion matrix is used where three metrics are used: accuracy, sensitivity, and specificity. Accuracy was defined by the following formula:

$$Accuracy = \frac{Sum\ of\ the\ correct\ prediction\ on\ testing\ set}{Sum\ of\ dataset} \times 100\% \quad \ldots(11)$$

The parameters with $\sigma = 0.0001$ for RBF and Cubic Kernels.

Accuracy can be defined as (Gorunescu 2011):

ACTUAL CLASS	**PREDICTION CLASS**	
	Positive Infarction	Negative Infarction
Positive Infarction	True Positive (TP)	False Negative (FN)
Negative Infarction	False Positive (FP)	True Negative (TN)

In the medical field, the greater the value of Recall (Sensitivity) and Specificity, the better the method, which can be defined by the following equations [15]:

$$Recall = \frac{TP}{(TP + FN)} \quad \ldots.(12)$$

$$Specificity = \frac{TN}{(TN + FP)} \quad \ldots(13)$$

While the ratio of the test positive sample class that is predicted correctly with the overall positive class sample is the definition of Precision, which described as the following equation:

$$Precision = \frac{TP}{(FP + TP)} \quad \ldots \ (14)$$

And F1-score is the average harmonic between Precision and Recall. The best classifiers have a value close to 1 and the worst classifiers have a value close to 0. F1-score described by the following equation:

$$F1 = \frac{Precision\ x\ Recall}{Precision + Recall} \quad \ldots \ (15)$$

3. Result and discussions

In this study comparison is made between the accuracies of two kind of kernels. The first one is Support Vector Machine (SVM) method with RBF Kernel (Gaussian) and the second one is Support Vector Machine (SVM) method with Cubic Kernel (Polynomial). The parameters with $\sigma = 0.0001$ for RBF & Cubic Kernels.

The datasets that are used are classified as large-scale medical dataset provided by the Department of Radiology, Dr. Cipto Mangunkusumo National Central General Hospital which consists of 206 samples and 7 Attributes. The creators of this datasets are Dr. Jacub and Dr. Widyo from the National Central General Hospital, while the donor is from various patient of the Department of Radiology, Dr. Cipto Mangunkusumo National Central General Hospital. Data is taken from the year 2019 while the CT scans are taken from early 2018 to late 2018.

This dataset has 207, with 206 samples and 1 row to explain the attributes, also it has 7 columns for which every 1 column describe 1 attribute. All dataset is in excel form. From the random data set at hand, we can examine the accuracy of the prediction. After that the calculation of the specific number of the accuracy can be attained and analysis the confusion matrices can be done.

From the experiment on the Infarction Cerebrovascular dataset, the results as compared are described below. For the summary of the experiment can be seen in table 1.

Table 1. Accuracy Score of SVM (with Kernel RBF), sigma=0.0001

% Data Training	Number of Data Training	%Accuracy
10	20	83.87
20	41	91.93
30	61	83.87
40	82	80.64
50	103	85.48
60	123	77.41
70	144	91.93
80	164	77.41
90	185	79.03

Table 2. Accuracy Score of SVM (with Kernel Cubic), sigma=0.0001

%Data Training	Number of Data Training	%Accuracy
10	20	80.64
20	41	82.25
30	61	79.03
40	82	79.03
50	103	79.03
60	123	80.64
70	144	77.41
80	164	79.03
90	185	75.80

Table 3. Values of Precision, Recall, and F1-score for experiments with best accuracy score

Kernel	% Data Training	Classification	Precision	Recall	f1-score
Cubic	20	0	0.90	0.76	0.83
		1	0.76	0.89	0.82
RBF	20	0	0.97	0.89	0.93
		1	0.86	0.96	0.91
	70	0	0.96	0.86	0.91
		0	0.92	0.92	0.92

Table 4. Accuracy Score of three method with and without Feature Selection

Classification Method	Accuracy Score
SVM and RBF Kernel	91.93%
SVM and Cubic Kernel	82.85%

In the probability of abnormal and normal brain tissues, the algorithm of Support Vector Machine (SVM) can be implemented with two kind of kernels which give various accuracies, as described from Table 1 and Table 2 (see Table 1 and Table 2). From Table 4 (see Table 4), it can be seen that the first column is the classification method which is Support Vector Machine method (SVM) with Radial Basis Function (RBF) and Cubic Kernel, and that the second column is the results of accuracy score from every method. The results of these experiments are 91.93% (or 92%) accuracy for the Gaussian SVM (RBF Kernel) method, and 82.85% (or 82%) accuracy for the Cubic SVM (Polynomial Kernel degree 3).

As observed from Table 1 and Table 2, choosing the right kernel for SVM method itself will help to give a good accuracy score of above 80%, but from the comparison between the final results of the score accuracy (see Table 4) SVM method with RBF Kernel gives a better accuracy score which is above 90% accuracy. Therefore, it can be concluded that the Classification Method will give the best result with the SVM method with RBF Kernel as it has a better performance. For other considerations, choosing the best amount of data training will help to give best accuracy score, as we can see from Table 1 and Table 2, 20% and 70% data training give the best result of accuracy score.

From Table 3, it describes precision, recall, and f1-score in each column. Based on the precision value, the Gaussian SVM method is better than the Cubic SVM because it has a precision value of 0.97 (97%), meanwhile Cubic SVM has a precision value of 0.90 (90%). Based on the recall value, the Gaussian SVM methods is better than the Cubic SVM because it has a recall value of 0.89 (89%), meanwhile the Cubic SVM has a recall value of 0.76 (76%). Based on the F1-score value, the Gaussian SVM method is better than the Cubic SVM because it has a F1-score value of 0.93 (93%), meanwhile

Cubic SVM has a precision value of 0.83 (83%). Therefore, because both methods have F1-scores approaches 1 this shows that these methods are good methods for classification of this data.

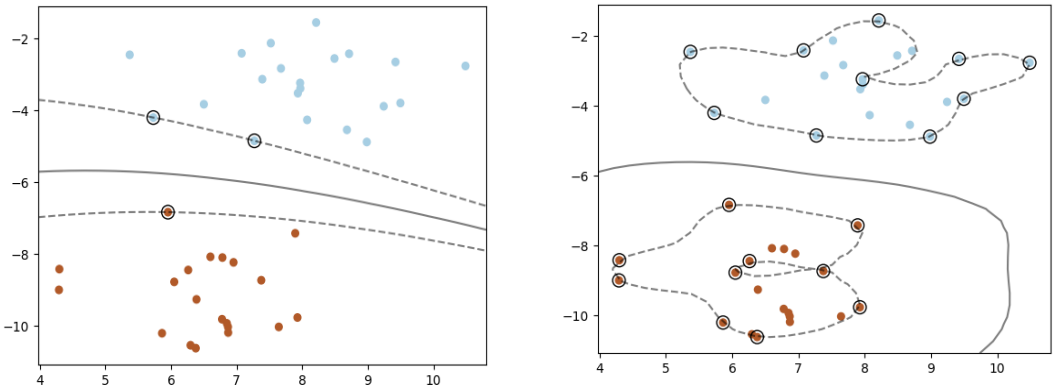


Figure 4. Comparison of linear Cubic SVM (left) with RBF SVM (right).

Using nonlinearly SVM by applying two kinds of kernels Gaussian (RBF) and Cubic (Polynomial) for detecting Ischemic Stroke, this method is applied to real world problems to identify the presence of cerebral infarction (blockage of the brain) that can cause ischemic stroke. The advantage of this method is it can produce an increased performance in SVM. This method achieves the highest accuracy 91%.

4. Conclusion

Recently, machine learning has been given more attention in the health sector. Medical data mining plays an important role in helping the health sector to search for possible indications of many kinds of diseases so that from possible factors, the experts could propose for better and effective treatment. Classification for Infarction Cerebrovascular which is found in patient's CT brain scan, has been very helpful for research be more focused on Classification for its good performance and accurate score. In this study, we recommend Support Vector Machine method (SVM) with RBF Kernel and Cubic Kernel.

The SVM and RBF Kernel gave 91% accuracy. Hence, this method can be applied to the problem of classification of Infarction Cerebrovascular datasets with the best results and provide results of accuracy of above 90%. The high accuracy score means SVM with RBF Kernel could predict the class on the data itself, which means that it is able to classify exactly according to the original class, and has better performance when compared to SVM method with Cubic Kernel. A way to get a higher accuracy of the SVM Method is to consider other alternatives of Kernels and other parameters for a better accuracy number.

Acknowledgments

This work supported by Department Radiology of Dr.Cipto Mangunkusumo's Hospital. We thank to all reviewers for the improvement of this article.

References

[1] Zhang Yin, Liu Yue, wang Guiqian, Sun Shuailing, Gao Yang, Xie Yanming. Risk assessment model for endpoints of Ischemic Stroke : A Study protocol for registry study (China, 2017)

[2] https://www.who.int/news-room/fact-sheets/detail/the-top-10-causes-of-death. Accessed on 26 Febuary 2019.

[3] J. M. Wardlaw, T. M. West, P. A. G. Sandercock, S. C Lewis, O. Mielke. Visible infarction on computed tomography is an independent predictor of poor functional outcome after stroke, and not of haemorrhagic transformation (Edinburgh, UK 2003)

[4] Siva P. Raman, MD, Mahadevappa Mahesh, MS, PhD, Robert V. Blasko, BS, RT, Elliot K. Fishman, MD, CT Scan Parameters and Radiation Dose: Practical Advice for Radiologist, (John Hopkins University, Baltimore, Maryland, 2013)
[5] https://radiopaedia.org/articles/hounsfield-unit?lang=us. Accessed on 26 Febuary 2019.
[6] Harleen Kaur, Vinita Kumari, Predictive modelling and analytics for diabetes using a machine learning approach, (New Delhi, India, 2018)
[7] Alexander Selvikvag Lundervold, Arvid Lundervold, An overwiew of deep learning in medical imaging focusing on MRI, (Norway, 2018)
[8] Yi Cai, Yifan Guo, Haotian Jiang, Ming-Chun Huang,Machine learning approaches for recognizing muscle activities involved in facial expressions captured by multi-channels surface electromyogram, (United States, 2017)
[9] Paul Bentley, Jeban Ganesalingam, Anoma Lalani Carlton Jones, Kate Mahady, Sarah Epton, Paul Rinne, Pankaj Sharma, Omid Halse, Amrish Mehta, Daniel Rueckert, Prediction of stroke thrombolysis outcome using CT brain machine learning (Imperial College London, 2014)
[10] T. M. Mitchell, Machine Learning, McGraw Hill, 1997.
[11] Aaron E. Maxwell, Timothy A. Warner, Fang Fang, Implementation of machine-learning classification in remote sensing: an applied review, (West Virginia University, Morgantown, USA, 2017)
[12] Sunita Beniwal, Jitender Arora, Classification and Feature Selection Techniques in Data Mining, (Maharishi Markandeshwar University, Mullana, India,2012)
[13] Nizar A., Dong Z., Wang Y., Power utility nontechnical loss analysis with extreme learning machine method, (2008)
[14] Aized Amin SOofi and Arshad Awan, Classification Techniques in Machine Learning: Applications and Issues, Allama Iqbal Open University, Islamabad, Pakistan, 2017
[15] Abdul Azis Abdillah, Suwarno, Diagnosis of diabetes using Support Vector Machine with Radial Basis Function Kernels, (Indonesia, 2016)
[16] Chih-Chung Chang and Chih-Jen Lin, LIBSVM: A Library for Support Vector Machines, (National Taiwan University, Taipei, Taiwan, 2013)
[17] https://www.analyticsvidhya.com/blog/2017/09/understaing-support-vector-machine-example-code/. Accessed on 26 Febuary 2019.
[18] Zuherman Rustam, Aini Suri Talita, Fuzzy Kernel K-Medoids Algorithm for Multiclass Multidimensional Data Classification, (Journal of Theoretical and Applied Information Technology Conference, Jakarta, 2015)
[19] Zuherman Rustam, Aini Suri Talita, Fuzzy Kernel C-Means Algorithm for Intrusion Detection System, (Journal of Theoretical and Applied Information Technology Conference, Jakarta, 2015)