International Journal of Technology (2016) 5: 849-858
ISSN 2086-9614

DIAGNOSIS OF DIABETES USING SUPPORT VECTOR MACHINES WITH RADIAL BASIS FUNCTION KERNELS

Abdul Azis Abdillah[1,2*], Suwarno[2]

[1] *Department of Mechanical Engineering, Jakarta State Polytechnic, Kampus Baru UI, Depok, 16424, Indonesia*
[2] *Department of Mathematics Education, STKIP Surya, Tangerang, 15810, Indonesia*

(Received: June 2015 / Revised: March 2016 / Accepted: June 2016)

ABSTRACT

Diabetes is one of the most serious health challenges in both developed and developing countries. Early detection and accurate diagnosis of diabetes can reduce the risk of complications. In recent years, the use of machine learning in predicting disease has gradually increased. A promising classification technique in machine learning is the use of support vector machines in combination with radial basis function kernels (SVM-RBF). In this study, we used SVM-RBF to predict diabetes. The study used a Pima Indian diabetes dataset from the University of California, Irvine (UCI) Machine Learning Repository. The subjects were female and ≥ 21 years of age at the time of the index examination. Our experiment design used 10-fold cross-validation. Confusion matrix and ROC were used to calculate performance evaluation. Based on the experimental results, the study demonstrated that SVM-RBF shows promise in aiding diagnosis of Pima Indian diabetes disease in the early stage.

Keywords: Diabetes; Pima dataset; SVM-RBF

1. INTRODUCTION

Diabetes is a major health problem in both developed and developing countries. It is a chronic disease that occurs either when the pancreas does not produce enough insulin or when the body cannot effectively use the insulin it produces (WHO, 2013). Untreated diabetes can result in such potentially fatal complications as increased risk of heart disease, stroke, kidney failure, blindness, and limb amputation. There are two main types of the disease (WHO, 2014): type 1 (T1B) and type 2 diabetes (T2B). T1B usually develops in childhood or adolescence, while T2B develops in adulthood.

As of 2014, an estimated 387 million people had diabetes worldwide (IDF, 2014), with T2B accounting for the majority of cases (WHO, 2014). At present there is no known cure for diabetes. However, early detection and accurate diagnosis can lower the risk of complications arising (IDF, 2014). To prevent complications, medical professionals usually suggest certain lifestyle choices such as a healthy diet, physical activity, maintaining an appropriate body weight, and not smoking (WHO, 2013). In recent years there has been a gradual increase in the use of machine learning in medical fields, particularly in disease prediction. Pattern recognition studies are already widely used in disease detection and have yielded promising results in detection of Alzheimer's disease (Liu et al., 2015), breast cancer (Cheng et al., 2010), and heart attack (Polat et al., 2006). Preliminary research related to diabetes conducted by Smith et al.

*Corresponding author's email: abdul.abdillah@mesin.pnj.ac.id, Tel. +62-21-7270044, Fax. +62-21-7270034
Permalink/DOI: http://dx.doi.org/10.14716/ijtech.v7i5.1370

(1988) using the Pima dataset generated sensitivity and specificity values of 76%.

One of the promising classification techniques in machine learning is the support vector machine (SVM). SVMs operate on the basic principle of binary classification with maximum margin and can be further developed to solve nonlinear cases. SVMs have been used successfully in the areas of face recognition (Khan et al., 2012), handwriting recognition (Adankon & Cheriet, 2009), iris recognition (Rai & Yadav, 2014), information retrieval (Abdillah et al., 2015), and many others. In this study, we used SVMs with radial basis function kernels (SVM-RBF) with an experiment design based on 10-fold cross-validation to predict diabetes.

2. SVMs

In this section, we summarize the basic concepts involved in using SVMs to perform two-class classification in linear and nonlinear cases.

2.1. Maximal Margin Classifier

Given a linearly separable training sample $S = \{(\boldsymbol{x}_i, t_i)\}_{i=1,\dots,m}$, where $\boldsymbol{x}_i \in R^n$ are n-dimensional vectors and $t_i = \{1, -1\}$ are corresponding labels, S is called linearly separable when there exists a hyperplane that correctly classifies every point in S. The SVM's goal is to find the hyperplane that can separate the two classes with maximum margin; this is known as a hard-margin SVM problem. The hyperplane and decision function of the SVM were defined as $\boldsymbol{w}^T\boldsymbol{x} + b = 0$ and $f(x) = sign(\boldsymbol{w}^T\boldsymbol{x} + b)$ respectively, where $\boldsymbol{w}$ is the weight vector and b is bias. The margin is the additional distance from the hyperplane to the closest positive data point and from the hyperplane to the closest negative data point. In order to find the margin, we first defined a canonical hyperplane such that $(\boldsymbol{w}^T\boldsymbol{x}_i^+ + b) = 1$ for the closest positive data and $(\boldsymbol{w}^T\boldsymbol{x}_i^- + b) = -1$ for the closest negative data point. Hence, all data satisfied:

$$\boldsymbol{w}^T\boldsymbol{x}_i + b \geq 1, \text{ if } t_i = 1, \text{ for } i = 1,2, \dots, m \tag{1}$$

$$\boldsymbol{w}^T\boldsymbol{x}_i + b \leq -1, \text{ if } t_i = -1, \text{ for } i = 1,2, \dots, m \tag{2}$$

Equations 1 and 2 is equivalent to

$$t_i(\boldsymbol{w}^T\boldsymbol{x}_i + b) \geq 1, i = 1,2, \dots, m \tag{3}$$

Let d_+ be the shortest distance from the separating hyperplane to the closest positive data and d_- be the shortest distance from the separating hyperplane to the closest negative data. The distance from the hyperplane to the closest data $\boldsymbol{x}_i$ is formulated as $\frac{|F(\boldsymbol{x}_i)|}{\|\boldsymbol{w}\|}$. Hence, we get the margin as:

$$\begin{aligned} d &= d_+ + d_- \\ &= \frac{1}{\|\boldsymbol{w}\|}(|\boldsymbol{w}^T\boldsymbol{x}_i^+ + b| + |\boldsymbol{w}^T\boldsymbol{x}_i^- + b| \\ &= \frac{2}{\|\boldsymbol{w}\|}. \end{aligned} \tag{4}$$

Therefore, maximizing the margin $\frac{2}{\|\boldsymbol{w}\|}$ is equivalent to minimizing $\frac{1}{2}\|\boldsymbol{w}\|^2$. Now, we have our problem in its primal form (Bishop, 2006; Cristianini & Shawe-Taylor, 2000; Scholkopf & Smola, 2001):

$$\min_{w,b} \frac{1}{2}\|\boldsymbol{w}\|^2 \tag{5}$$

Subject to:

$$t_i\left(\boldsymbol{w}^T\boldsymbol{x}_i + b\right) \geq 1, i = 1,2, \ldots, m \tag{6}$$

Now we have a quadratic optimization problem and we need to solve for $\boldsymbol{w}$ and b. In order to solve this optimization problem, we first defined the Lagrangian,

$$L(\boldsymbol{w}, b, a) = \frac{1}{2}\|w\|^2 - \sum_{i=1}^{m} a_i\left[t_i(\boldsymbol{w}^T\boldsymbol{x}_i + b) - 1\right] \tag{7}$$

where $a_i \geq 0$ are Lagrange multipliers. We want to find $\boldsymbol{w}$ and b, which minimizes, and a, which maximizes, by differentiating L with respect to $\boldsymbol{w}$ and b and setting the derivatives to zero. So we have:

$$\frac{\partial L}{\partial \boldsymbol{w}} = \mathbf{0} \rightarrow \boldsymbol{w} = \sum_{i=1}^{m} a_i t_i \boldsymbol{x}_i \tag{8}$$

$$\frac{\partial L}{\partial b} = 0 \rightarrow 0 = \sum_{i=1}^{m} a_i t_i \tag{9}$$

Then we substituted Equations 8 and 9 back into the primal Equation 7, so we could give the dual form (Bishop, 2006; Cristianini & Shawe-Taylor, 2000; Scholkopf & Smola, 2001):

$$\max_{a} \hat{L}(a) = \sum_{i=1}^{m} a_i - \frac{1}{2}\sum_{i=1}^{m}\sum_{j=1}^{m} a_i\, a_j\, t_i t_j\, \boldsymbol{x}_i^T\boldsymbol{x}_j \tag{10}$$

Subject to:

$$a_i \geq 0, i = 1, \ldots, m \tag{11}$$

$$\sum_{i=1}^{m} a_i t_i = 0 \tag{12}$$

We found all a_i by solving Equation 10 with constraint in Equations 11 and 12 using a quadratic programming solver. Only $\boldsymbol{x}_i$ with $a_i > 0$, which carry all relevant information about the classification problem and these a_i are called support vectors (SVs). Furthermore, the solution for $\boldsymbol{w}$ could be obtained from Equation 5

$$\boldsymbol{w}^* = \sum_{i \in SV} a_i t_i \boldsymbol{x}_i \tag{13}$$

and we were able to calculate b using Karush-Kuhn-Tucker (KKT) conditions from the primal problem:

$$a_i \geq 0 \tag{14}$$

$$t_i\left(\boldsymbol{w}^T\boldsymbol{x}_i + b\right) - 1 \geq 0 \tag{15}$$

$$a_i\left(t_i\left(\boldsymbol{w}^T\boldsymbol{x}_i + b\right) - 1\right) = 1 \tag{16}$$

Substituting Equation 13 into Equation 16, we got b as (Bishop, 2006; Cristianini & Shawe-Taylor, 2000; Scholkopf & Smola, 2001):

$$b^* = \frac{1}{N_{SV}} \sum_{j \in SV} (t_j - \sum_{i \in SV} a_i t_i \boldsymbol{x}_i^T \boldsymbol{x}_j). \tag{17}$$

Therefore, the decision function that performs the classification could be written as:

$$f(\boldsymbol{x}) = Sign\left(\sum_{i \in SV} a_i t_i \boldsymbol{x}_i^T \boldsymbol{x} + b^*\right), \tag{18}$$

where SV is the number of support vectors.

2.2. Soft-margin SVMs

Many real datasets cannot be separated without misclassification or errors. To handle these cases, the SVM algorithm was modified by adding slack variables $\xi_i \geq 0,\ i = 1,2,\ldots,m$. This is known as a soft-margin SVM. The modification of the primal form in Equations 5 and 6 to get the soft-margin optimization problem can be written as follows (Bishop, 2006; Cristianini & Shawe-Taylor, 2000; Scholkopf & Smola, 2001):

$$\min_{w,b} \frac{1}{2}\|\boldsymbol{w}\|^2 + C\sum_{i=1}^{m} \xi_i \tag{19}$$

subject to:

$$t_i\left(\boldsymbol{w}^T\boldsymbol{x}_i + b\right) \geq 1 - \xi_i, i = 1,2,\ldots,m \tag{20}$$

$$\xi_i \geq 0, i = 1,2,\ldots,m \tag{21}$$

where the parameter $C > 0$ controls the trade-off between the slack variable penalty and the margin. Meanwhile, the dual form of the soft-margin optimization problem in Equation 10 with constraint in Equations 11 and 12 can be written as follows (Bishop, 2006; Cristianini & Shawe-Taylor, 2000; Scholkopf & Smola, 2001):

$$\max_a \hat{L}(a) = \sum_{i=1}^{m} a_i - \frac{1}{2}\sum_{i=1}^{m}\sum_{j=1}^{m} a_i\, a_j\, t_i t_j\, \boldsymbol{x}_i^T \boldsymbol{x}_j \tag{22}$$

subject to:

$$0 \leq a_i \leq C, i = 1,\ldots,m \tag{23}$$

$$\sum_{i=1}^{m} a_i t_i = 0 \tag{24}$$

The solution of this optimization problem is identical to the dual hard-margin SVM formulation; the only difference is the constraint $0 \leq a_i \leq C$.

$$\boldsymbol{x} \rightarrow \Phi(\boldsymbol{x}). \tag{25}$$

However, finding the best transformation function is very hard. To avoid this problem, we applied the kernel trick. We do not need to know the function Φ; the kernel trick replaces the inner product $\Phi(\boldsymbol{x}_i)^T\Phi(\boldsymbol{x}_j)$ with the kernel function $K(\boldsymbol{x}_i,\boldsymbol{x}_j)$ in the original input space. Kernel function is defined as $K(\boldsymbol{x}_i,\boldsymbol{x}_j) = \Phi(\boldsymbol{x}_i)^T\Phi(\boldsymbol{x}_j)$. Using the kernel function, the dual problem can be rewritten as (Bishop, 2006; Cristianini & Shawe-Taylor, 2000; Scholkopf & Smola, 2001):

$$\max_a \hat{L}(a) = \sum_{i=1}^{m} a_i - \frac{1}{2}\sum_{i=1}^{m}\sum_{j=1}^{m} a_i\, a_j\, t_i t_j\, k(\boldsymbol{x}_i,\boldsymbol{x}) \tag{26}$$

subject to:

$$0 \leq a_i \leq C, i = 1, \ldots, m \tag{27}$$

$$\sum_{i=1}^{m} a_i t_i = 0 \tag{28}$$

Meanwhile, the classification function for nonlinear SVM becomes

$$f(\boldsymbol{x}) = Sign\left(\sum_{i \in SV} a_i t_i k(\boldsymbol{x}_i,\boldsymbol{x}) + b^*\right), \tag{29}$$

where SV is the number of support vectors.

There are many kernel functions that can be employed in SVM. The kernel function used in this study is the radial basis function (SVM-RBF). The formula of the RBF kernel can be written as (Bishop, 2006; Cristianini & Shawe-Taylor, 2000; Scholkopf & Smola, 2001):

$$K(\boldsymbol{x},\boldsymbol{y}) = exp\left(-\frac{\|\boldsymbol{x}-\boldsymbol{y}\|^2}{\sigma^2}\right), \sigma > 0 \tag{30}$$

3. EXPERIMENTAL METHODS

A general diagram of diabetes prediction system using SVM-RBF can be seen in Figure 1. Details of dataset used, experiment design and the evaluation experiments for the proposed model are described in the following subsections:

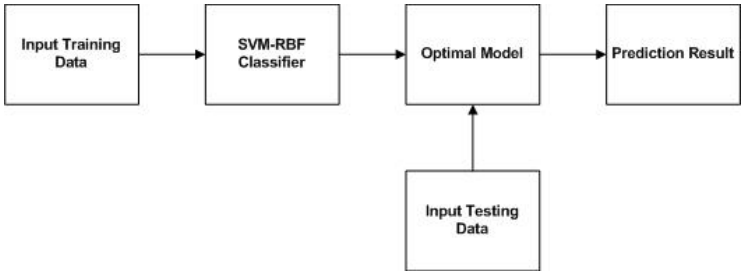


Figure 1 General diagram of diabetes prediction system using SVM-RBF

3.1. Datasets

We conducted experiments on the Pima Indian diabetes dataset (Lichman, 2013). This dataset contains 768 subjects. The subjects were female and ≥ 21 years of age at the time of the index examination. The dataset consists of two classes (positive and negative for diabetes) and eight variables;

- Number of pregnancies;
- Concentration of plasma glucose using a 2-h oral glucose tolerance test;
- Diastolic blood pressure (mm Hg);
- Thickness of triceps skin fold (mm);
- 2-h serum insulin (mu U/ml);
- Body mass index (kg/m^2)
- Diabetes pedigree function; and
- Age (years)

3.2. Experiment Design

In this study, 10-fold cross-validation (Bishop, 2006) was used to build the best models. The data set was partitioned into 10 subsets; one subset was used as the testing set and the rest were used as training sets. This process was repeated 10 times, with each of the subsets used exactly once as the testing data. Then, to evaluate the performance of the models, the average of the performances across all 10 trials was calculated.

3.3. Performance Evaluation Methods

To evaluate the performance of this method, we used a confusion matrix and a receiver operating characteristic (ROC) curve.

3.3.1. Confusion matrix

The confusion matrix is shown in Table 1. To calculate performance evaluation using a confusion matrix, we used three metrics: accuracy, sensitivity, and specificity. Accuracy was defined as the proportion of correct predictions to total number of predictions. Accuracy can be expressed as (Gorunescu, 2011):

Table 1 Confusion matrix

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

$$Accuracy = \frac{TP + TN}{TP + FP + FN + TN} \tag{31}$$

In the case of the evaluation of a system, the sensitivity of the test is the ability of the test to correctly identify those patients with the disease. Specificity is the ability of the test to correctly identify those patients without the disease. In the medical field, the greater the value of sensitivity and specificity, the better the method. Sensitivity and specificity can be expressed as (Gorunescu, 2011):

$$Sensitivity = \frac{TP}{TP + FN} \tag{32}$$

$$Specificity = \frac{TN}{TN + FP} \tag{33}$$

3.3.2. Receiver operating characteristic

An ROC curve was used to evaluate the performance of diagnostic tests visually. ROC was defined as a plot of sensitivity as the y-axis and 1-specificity as the x-axis. The perfect ROC will have a line going from the bottom left corner to the top left corner to the top right corner, with a perfect AUROC score of 1. The AUROC can be calculated as the sum of the areas of trapeziums below the ROC line. The formula of AUROC can be written as (Altman, 2006; Fawcett, 2006)

$$AUROC = \frac{1}{2}\sum_{i=1}^{n}(x_{i+1} - x_i)(y_{i+1} - y_i) \tag{34}$$

4. RESULTS AND DISCUSSION

All experiments were performed on a personal computer utilizing a 3.2 GHz Core i5 4460 CPU processor and 4 GB of RAM. This computer ran the Windows 7 operating system with MATLAB R2013a installed.

Table 2 Values of accuracy, sensitivity, and specificity for all experiments

	100 Data		200 Data		300 Data		400 Data		500 Data	
	Train	Test	Train	Test	Train	Test	Train	Test	Train	Test
Best Sigma	1.5	1.5	1.5	1.5	1.6	1.6	1.5	1.5	1.5	1.5
Accuracy (%)	83.00	72.76	84.00	76.49	82.00	75.37	83.75	80.22	82.40	80.22
Sensitivity (%)	89.19	76.74	85.33	80.23	86.84	76.74	86.18	81.40	84.07	82.56
Specificity (%)	79.37	70.88	83.20	74.73	79.03	74.73	82.26	79.67	81.45	79.12

In the early stages of implementation, the Pima dataset was divided into training and testing data. The training data used in this study consisted of 100, 200, 300, 400, and 500 data, while the testing consisted of 268 data. The value of parameter C was default and σ was tried between -10 and -3 in increments of 0.1, and between 1.5 and 10 in increments of 0.1. The results of the performance evaluation for optimal parameter values for both training and testing data can be seen in Table 2.

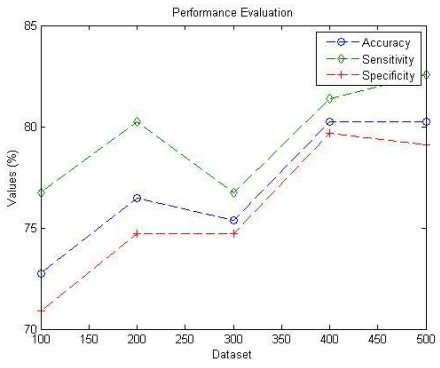
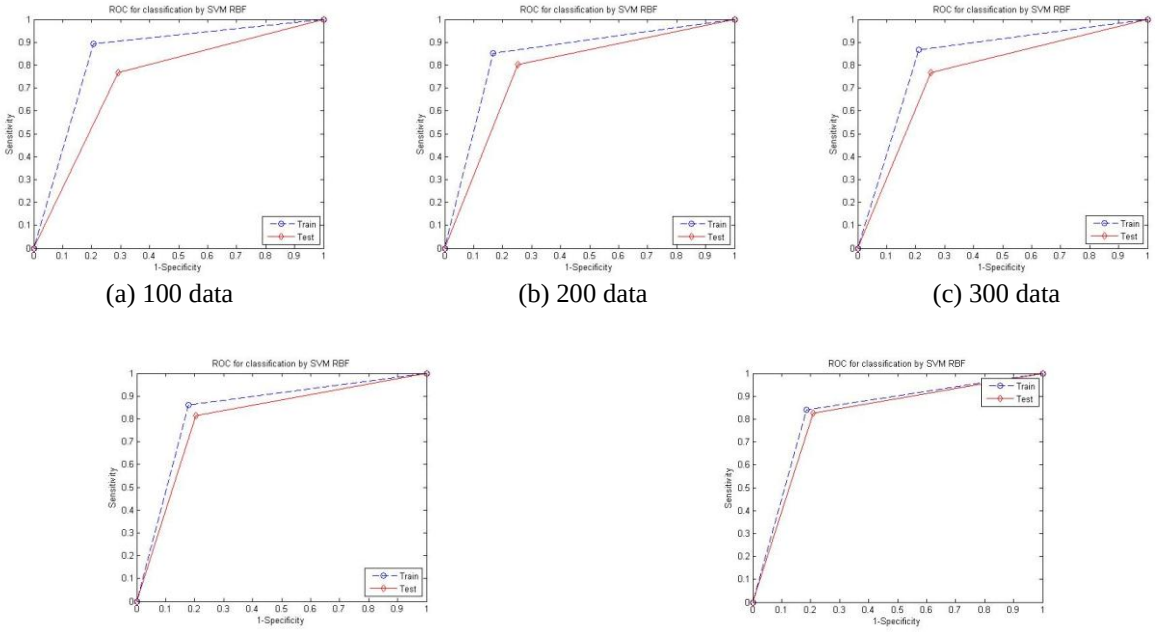


Figure 2 Performance evaluation for all testing data



(a) 100 data (b) 200 data (c) 300 data

(d) 400 data (e) 500 data

Figure 3 ROC curves for all experiments

Table 2 shows that the optimal parameter values were obtained when $\sigma = 1.5$, except when 300 training data were used ($\sigma = 1.6$). Figure 2 explains the performance results of all experiments. Comparing the performances using five different training data, it can be concluded that the more data used for training, the better the performance. While the performance decreased in the case of 300 training data, the difference in performances between the five training data amounts was not significant. The highest performance results were obtained when 500 training data were used; the results for accuracy, sensitivity, and specificity were 80.22%, 82.56%, and 79.12% respectively.

Table 3 Values of AUROC in All Experiments

	100 Data		200 Data		300 Data		400 Data		500 Data	
	Train	Test	Train	Test	Train	Test	Train	Test	Train	Test
AUROC	0.8428	0.7381	0.8427	0.7748	0.8294	0.7573	0.8422	0.8053	0.8276	0.8084

Figure 3 shows the ROC curve for all experiments. From ROC curves we can also get AUROC, as can be seen in Table 3. Comparing the AUROC in the cases of five different training data amounts in Figure 4, it can be seen that the highest AUROC value was obtained when 500 training data were used, with a value of 0.8084; the lowest value was in the case of 100 training data, with a value of 0.7381. In line with the values of accuracy, sensitivity, and specificity, the more training data used, the greater the value of AUROC generated. Based on the results with 500 training data, the training data and a value of $\sigma = 1.5$ could produce a performance which exceeded 80%.

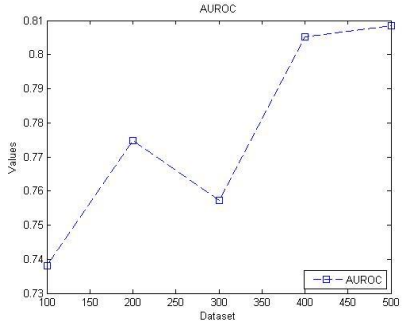


Figure 4 AUROC for all testing data

5. CONCLUSION

This study used support vector machines with kernel radial basis function (SVM-RBF) to predict diabetes. To evaluate the effectiveness of our method, the Pima Indian diabetes dataset was used. In order to achieve high classification performance, 10-fold cross-validation was used to build a model and search for the optimal parameters. In conclusion, the results can be summarized as follows: (1) The highest performance of SVM-RBF using 10-fold cross-validation was obtained from 500 training data with optimal parameter $\sigma = 1.5$, which yields accuracy, sensitivity, specificity, and AUROC of 80.22%, 82.56%, 79.12%, and 0.8084 respectively; (2) Based on the performance of the method, SVM-RBF obtained promising results as an aid in diagnosis of Pima Indian diabetes disease in the early stage.

In the future we plan to focus on using SVM combined with other kernel functions. Moreover, SVM can be combined with principal component analysis (PCA) for feature selection to enhance the accuracy of results.

6. ACKNOWLEDGEMENT

The authors would like to thank the Directorate General of Higher Education of the Republic of Indonesia (DIKTI) for providing financial support for this research via grant DIPA-023.04.1.673453/2015.

7. REFERENCES

Abdillah, A.A., Murfi, H., Satria, Y., 2015. Uji Kinerja Learning to Rank dengan Metode Support Vector Regression. *IndoMS Journal on Industrial and Applied Mathematics*, Volume 2(1), pp. 15–25 [in Bahasa]

Adankon, M.M., Cheriet, M., 2009. Model Selection for the LS-SVM: Application to Handwriting Recognition. *Pattern Recognition*, Volume 42(12), pp. 3264–3270

Altman, D.G., 2006. *Practical Statistics for Medical Research*. Chapman & Hall/CRC

Bishop, C.M., 2006. *Pattern Recognition and Machine Learning (Information Science and Statistics)*. Springer-Verlag New York, Inc., Secaucus, New Jersey, USA

Cheng, H., Shan, J., Ju, W., Guo, Y., Zhang, L., 2010. Automated Breast Cancer Detection and Classification using Ultrasound Images: A Survey. *Pattern Recognition*, Volume 43(1), pp. 299–317

Cristianini, N., Shawe-Taylor, J., 2000. *An Introduction to Support Vector Machines: and other Kernel-based Learning Methods*. Cambridge University Press, New York, New York, USA

Fawcett, T., 2006. An Introduction to ROC Analysis. *Pattern Recogn. Lett.*, Volume 27(8), pp. 861–874

Franciosi, M., De Berardis, G., Rossi, M.C., Sacco, M., Belfiglio, M., Pellegrini, F., Tognoni, G., Valentini, M., Nicolucci, A., 2005. Use of the Diabetes Risk Score for Opportunistic

Screening of Undiagnosed Diabetes and Impaired Glucose Tolerance: The Igloo (Impaired Glucose Tolerance and Long-term Outcomes Observational) Study. *Diabetes Care*, Volume 28(5), pp. 1187–1194

Gorunescu, F., 2011. *Data Mining Concepts, Models and Techniques*. Springer

IDF, 2014. *Key Findings 2014*. Available online at: http://www.idf.org/diabetesatlas/update-2014, Accessed on 2nd February, 2015

Khan, N., Ksantini, R., Ahmad, I., Boufama, B., 2012. A Novel SVM+NDA Model for Classification with an Application to Face Recognition. *Pattern Recognition*, Volume 45(1), pp. 66–79

Lichman, M., 2013. UCI Machine Learning Repository. Available online at: http://archive.ics.uci.edu/ml, Accessed on 9th February, 2015

Liu, X., Zhou, L., Wang, L., Zhang, J., Yin, J., Shen, D., 2015. An Efficient Radius-incorporated {MKL} Algorithm for Alzheimers Disease Prediction. *Pattern Recognition*, Volume 48(7), pp. 2141–2150

Polat, K., Gne, S., Tosun, S., 2006. Diagnosis of Heart Disease using Artificial Immune Recognition System and Fuzzy Weighted Pre-processing. *Pattern Recognition*, Volume 39(11), pp. 2186 – 2193

Rai, H., Yadav, A., 2014. Iris Recognition using Combined Support Vector Machine and Hamming Distance Approach. *Expert Systems with Applications*, Volume 41(2), pp. 588–593

Scholkopf, B., Smola, A.J., 2001. *Learning with Kernels: Support Vector Machines, Regularization, Optimization, and Beyond*. MIT Press, Cambridge, Massachusetts, USA

Smith, J.W., Everhart, J.E., Dickson, W.C., Knowler, W.C., Johannes, R.S., 1988. Using the Adapted Learning Algorithm to Forecast the Onset of Diabetes Mellitus. *In*: Proceedings of the Annual Symposium on Computer Application in Medical Care, pp. 261–265, American Medical Informatics Association

WHO, 2013. *Diabetes Fact Sheet*. Available online at: http://www.who.int/mediacentre/factsheets/fs312/en/, Accessed on 2nd February, 2015

WHO, 2014. *About Diabetes*. Available online at: http://www.who.int/diabetes/action_online/basics/en/, Accessed on 2nd February, 2015