Journal of Innovative Optical Health Sciences
Vol. 6, No. 4 (2013) 1350036 (7 pages)
© The Authors
DOI: 10.1142/S1793545813500363

World Scientific
www.worldscientific.com

CLASSIFICATION OF SKIN AUTOFLUORESCENCE SPECTRUM USING SUPPORT VECTOR MACHINE IN TYPE 2 DIABETES SCREENING

YUANZHI ZHANG[*], LING ZHU[*,§], YIKUN WANG[*],
LONG ZHANG[*], SHANDONG YE[†], YONG LIU[*] and GONG ZHANG[‡]

[*]*Anhui Institute of Optics and Fine Mechanics*
Chinese Academy of Sciences
Hefei, 230031, P. R. China

[†]*Anhui Medical University*
Affiliated Anhui Provincial Hospital
Hefei, 230001, P. R. China

[‡]*University of Manitoba, Winnipeg, R3T6A5, Canada*
[§]*zhul@aiofm.ac.cn*

Received 25 April 2013
Accepted 3 August 2013
Published 4 September 2013

Advanced glycation end products (AGEs) are a complex and heterogeneous group of compounds that have been implicated in diabetes related complifications. Skin autofluorescence was recently introduced as an alternative tool for skin AGEs accumulation assessment in diabetes. Successful optical diagnosis of diabetes requires a rapid and accurate classification algorithm. In order to improve the performance of noninvasive and optical diagnosis of type 2 diabetes, support vector machines (SVM) algorithm was implemented for the classification of skin autofluorescence from diabetics and control subjects. Cross-validation and grid-optimization methods were employed to calculate the optimal parameters that maximize classification accuracy. Classification model was set up according to the training set and then verified by the testing set. The results show that radical basis function is the best choice in the four common kernels in SVM. Moreover, a diagnostic accuracy of 82.61%, a sensitivity of 69.57%, and a specificity of 95.65% for discriminating diabetics from control subjects were achieved using a mixed kernel function, which is based on liner kernel function and radical basis function. In comparison with fasting plasma glucose and HbA_{1c} test, the classification method of skin autofluorescence spectrum based on SVM shows great potential in screening of diabetes.

Keywords: Skin autofluorescence; support vector machines algorithm; type 2 diabetes; noninvasive screening.

J. Innov. Opt. Health Sci. 2013.06. Downloaded from www.worldscientific.com
by 103.240.126.9 on 10/21/18. Re-use and distribution is strictly not permitted, except for Open Access articles.

This is an Open Access article published by World Scientific Publishing Company. It is distributed under the terms of the Creative Commons Attribution 3.0 (CC-BY) License. Further distribution of this work is permitted, provided the original work is properly cited.

J. Innov. Opt. Health Sci. 2013.06. Downloaded from www.worldscientific.com
by 103.240.126.9 on 10/21/18. Re-use and distribution is strictly not permitted, except for Open Access articles.

1. Introduction

Advanced glycation end products (AGEs) are biochemical end-products of non-enzymatic glycation.[1] The accumulation of AGEs in human skin has been implicated in the progression of diabetes mellitus and the related complications. Several studies, including the diabetes control and complications trial (DCCT) and epidemiology of diabetes interventions and complications study (EDIC), have demonstrated that elevated skin AGEs are biomarkers of diabetes, are highly correlated with the complications of diabetes, and are predictive of future diabetic retinopathy and nephropathy.[2]

AGEs have fluorescent properties, light with wavelength between 300 and 420 nm can be employed as excitation source, and the emission spectrum distributes in the $420 \sim 600$ nm. Thus, the level of AGEs in human skin can be assessed through measuring skin autofluorescence, and then the risk of diabetes mellitus and related complications evaluated.

Support vector machine (SVM) is a machine-learning method, based on the principle of structural risk minimization, which performs well when applied to data outside the training set, and developed by Vapnik[3] and Burges.[4] During the past two decades, SVM has attracted great attention due to its capability of representing nonlinear relationships and producing models that generalize well in classifying the unseen data. The SVM algorithm has now emerged as an efficient approach to the classification of spectral data for tissue diagnosis. For instance, Lin *et al.*[5] used linear and nonlinear SVM to classify autofluorescence spectrum of nasopharyngeal carcinoma (NPC) from normal tissue and demonstrated that SVM has higher diagnostic accuracy than using PCA-LDA. Widjaja *et al.*[6] combined near-infrared (NIR) Raman spectroscopy with SVM for improving multi-class classification between different histopathological groups in tissues, and compared the performances when using different kernel functions and different types SVM.

In this paper, skin autofluorescence of 63 patients with type 2 diabetes and 140 control subjects were collected using a self-designed optical system.[7] SVM algorithm was implemented for classification of the skin autofluorescence. Based on the training set, cross-validation and grid-optimization methods were employed to calculate the optimal parameters in the four common kernel functions respectively, and corresponding model was set up and then verified with the testing set. In addition, to further improve the classification performance, a new mixed kernel function based on liner kernel function and radical basis function was established.

2. Materials and Methods

2.1. *Support vector machine*

Support vector machine is a relatively new type of learning algorithm. It has many unique advantages in solving small sample, nonlinear and high dimensional pattern recognition.[8] The main mechanism of SVM is to hunt an optimal separating hyperplane that meets the classification requirements. The plane should ensures the required classification accuracy, as well as makes the classification interval maximum. In theory, SVM can achieve the optimal classification for linearly separable problems. For nonlinear separable problems, they were first mapped into a high-dimensional linearly separable space through a nonlinear mapping, and then be traded as linearly separable problems. The nonlinear mapping is defined by an inner product function called kernel function. The most used kernel functions are as follow:

(a) Linear kernel:

$$K(x, x_i) = (x \bullet x_i) \tag{1}$$

(b) Polynomial kernel:

$$K(x, x_i) = (\gamma * (x \bullet x_i) + m)^d \tag{2}$$

(c) Gaussian radical basis function (RBF):

$$K(x, x_i) = \exp(-\gamma * \|x - x_i\|^2) \tag{3}$$

(d) Sigmoid tanh:

$$K(x, x_i) = \tanh(\gamma * (x \bullet x_i) + m) \tag{4}$$

When using SVM, two problems should be considered: how to choose the optimal input feature subset for SVM, and how to set the best kernel parameters. These two problems are crucial, because the feature subset choice influences the appropriate kernel parameters and vice versa. Therefore, obtaining the optimal feature subset and SVM parameters must occur simultaneously.[9] In this paper, skin autofluorescence was chosen as input features. The parameters that should be

J. Innov. Opt. Health Sci. 2013.06. Downloaded from www.worldscientific.com
by 103.240.126.9 on 10/21/18. Re-use and distribution is strictly not permitted, except for Open Access articles.

optimized include penalty parameter C and the kernel function parameters such as parameter γ for the RBF kernel. To design an SVM, one must choose a kernel function, set the kernel parameters and determine the penalty parameter C. Penalty parameter represents the compromise on training error and generalization ability. Cross-validation and grid-optimization methods are alternative to determine the kernel function and the optimal parameters.

2.2. *Data preprocessing*

Skin autofluorescence was assessed using the self-designed AGEs fluorescence spectroscopy detection device. The method has been described in detail elsewhere.[7] In short, the device illuminates approximately $0.1\,\text{cm}^2$ of skin, guarding against surrounding light, using an excitation light source with peak intensity at 370 nm. Emission light from the skin is measured with a spectrometer in 420–600 nm range through a fiber probe. Typical skin autofluorescence spectra of the diabetes and control subjects were show in Fig. 1. For emission light, the average sampling interval is 0.25 nm and 718 feature points are collected. It means that each spectrum is treated as a vector with 718 dimensions and labeled according to the clinical examination.

We recruited 245 subjects in total, some of which are newly admitted patients and the rest is the patient's family or hospital staff. After excluding the cases which received anti-diabetic treatment, we noninvasively measured skin autofluorescence from 203 subjects (age, 61 ± 9 years; body mass index, $26.6 \pm 5.5\,\text{kg/m}^2$; male/female ratio 95/108; smokers/nonsmoker ratio, 0/203; 63 patients with type 2 diabetes and 140 control subjects (the examination results show they are in good health). Here, diabetes was defined as a fasting glucose level of ≥ 7.0 mmol/L or a glucose level of > 11.0 mmol/L at 2 h in the OGGT. Control was defined as a fasting glucose level of < 7.0 and a glucose level of ≤ 11.0 mmol/L at 2 h in the OGGT). Both the diabetic patients and the control subjects were randomly recruited from Anhui Provincial Hospital. Measurements were performed at the volar side of the arm and avoid the location of the blood vessels, scars, lichenification, sclerosis plaques as well as deformity skin.

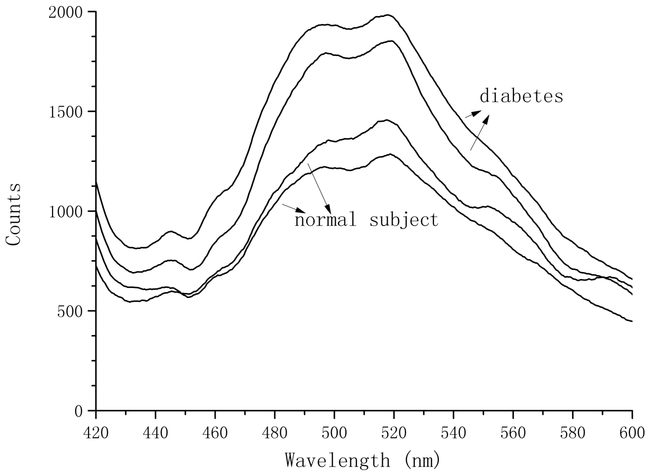

Fig. 1. Typical skin autofluorescence spectra of the diabetes and control subjects.

Analysis of all the data are performed in MATLAB environment using the Libsvm tool package[10] which was developed by Prof Lin.

The classification process of SVM is shown in Fig. 2. The selection of training set and testing set includes the following step: First, list all subjects in consecutive numerical order. Second, generate a group of random numbers between 1 and the number of subjects as the order of training set. Finally, select the training set according to the serial number and the rest is testing set. In this paper, 40 diabetics (age, 64 ± 7 years; body mass index, $27.6 \pm 4.5\,\text{kg/m}^2$; male/female ratio 19/21) and 117 control subjects (age, 60 ± 6 years; body mass index, $25.9 \pm 4.4\,\text{kg/m}^2$; male/female ratio 55/62) were selected as training set, the remaining 23 diabetics (age, 62 ± 8 years; body mass index, $27.9 \pm 4.8\,\text{kg/m}^2$; male/female ratio 11/12) and 23 control subjects (age, 60 ± 7 years; body mass index, $26.2 \pm 4.6\,\text{kg/m}^2$; male/female ratio 10/13) as testing set. To avoid the singular sample data,

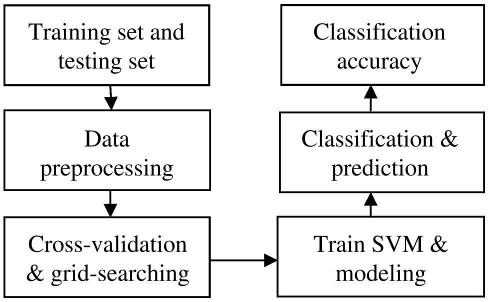
Training set and testing set
Data preprocessing
Cross-validation & grid-searching
Train SVM & modeling
Classification & prediction
Classification accuracy

Fig. 2. The flow chart of SVM. Classification process includes training and testing set selection, data preprocessing, parameter optimization, SVM modeling, classification judgment and solving of classification accuracy.

J. Innov. Opt. Health Sci. 2013.06. Downloaded from www.worldscientific.com
by 103.240.126.9 on 10/21/18. Re-use and distribution is strictly not permitted, except for Open Access articles.

speed up the process convergence rate and reduce the dimensionality of the data so as to simplified SVM operation, the skin autofluorescence were filtered using nine points Savitzky–Golay smoothing, normalized by dividing the mean intensity, and through principal component analysis, the first two principal components (representing more than 99% of the total variability) were chose as SVM input features.

3. Results

In this paper, we investigated four commonly used kernel functions. Figure 3 shows how the classification accuracy of a linear kernel SVM algorithm depends on the parameter C in a wide range search for optimal C. The algorithm was optimized by choosing the value of C that maximized classification accuracy for the cross-validation and grid-optimization.

In the development of nonlinear SVM algorithm using an RBF function, an optimal C that maximizes classification accuracy can be determined for each selected value of parameter γ. In a wide range of values for C and γ, many sets of C and γ could be found to yield the same classification accuracy. Figure 4 shows the optimal sets of C and γ that yield the maximal classification accuracy in the training set.

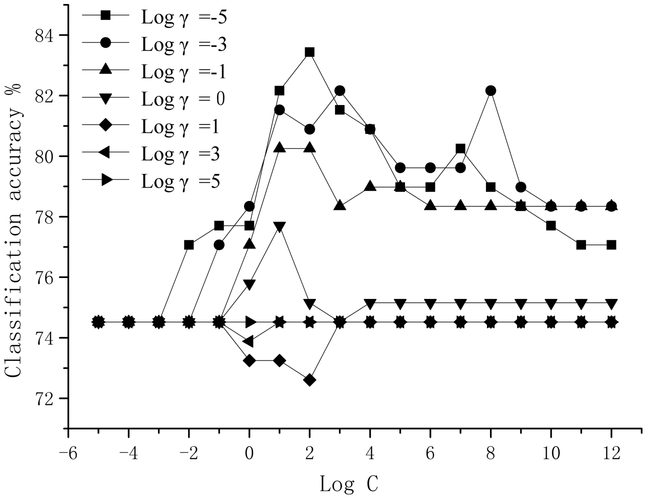

Fig. 4. Dependence of classification accuracy on parameters C and γ for a RBF SVM. The base of log C is 2 and the base of log γ is 2.

For the polynomial SVM algorithms, the maximum diagnostic accuracy was not very sensitive to the parameter d and m, and the calculation would be more time-consuming with the increase of d and m.[6] Therefore, in the development of polynomial SVM algorithms, the values of d and m were set as 3 and 1, respectively. Similarly, the value of m was set as 1 in the development of sigmoid SVM algorithms. Figures 5 and 6 show the optimal sets of C and γ that yield the maximal classification accuracy in polynomial SVM algorithms and sigmoid SVM algorithms.

When it does not matter in the classification accuracy, the penalty parameter C should be as small as possible, since higher value of C would lead

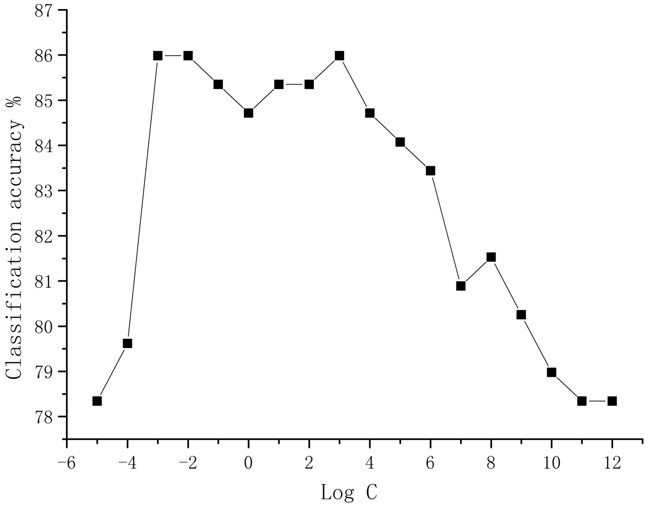

Fig. 3. Dependence of classification accuracy on parameter C using a linear SVM. The base of log C is 2.

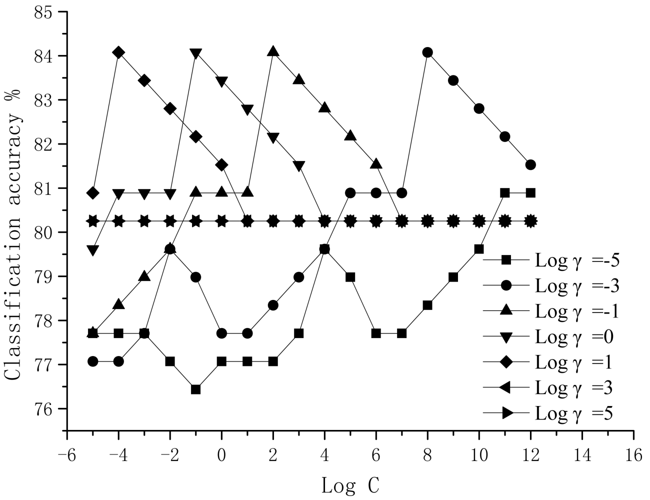

Fig. 5. Dependence of classification accuracy on parameters C and γ for a polynomial kernel SVM. The base of log C is 2 and the base of log γ is 2.

J. Innov. Opt. Health Sci. 2013.06. Downloaded from www.worldscientific.com by 103.240.126.9 on 10/21/18. Re-use and distribution is strictly not permitted, except for Open Access articles.

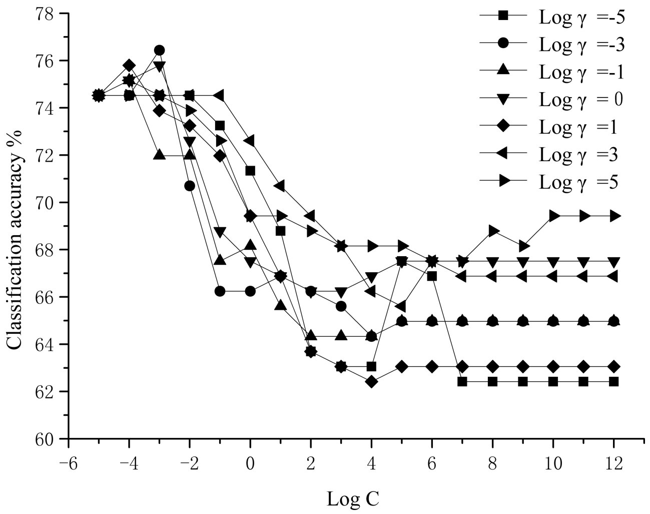

Fig. 6. Dependence of classification accuracy on parameters C and γ for a sigmoid SVM. The base of log C is 2 and the base of log γ is 2.

to over fitting and then reduce SVM's generalization ability. According to the cross-validation and grid-optimization, the optimal classification parameters and cross-validation average accuracy were achieved, and shown in Table 1.

Combining training set and the optimal parameters of each kernel function, we can build the classification model and evaluate the results of each model with testing set. Results of classification with four different kernels were shown in Table 2. It indicated that linear kernel and RBF kernel have better performances for the classification in this paper.

Models were set up with the training set and different kernels. The training set classification accuracy represent the classification result that model for the training set. The testing set classification accuracy represent the classification result that model for the training set.

Table 1. Optimal parameters of four different kernels.

	Linear kernel	Polynomial kernel	RBF	Sigmoid tanh
Parameter	$C = 0.125$	$C = 0.0625$ $\gamma = 2$	$C = 4$ $\gamma = 0.03125$	$C = 0.125$ $\gamma = 0.125$
Classification accuracy (%)	85.99	84.08	83.44	76.43

Note: Classification accuracy values in the training set data represent leave-one-out cross-validation values.

Table 2. Results of classification with four different kernels.

	Linear kernel	Polynomial kernel	RBF	Sigmoid tanh
The training set classification accuracy (%)	86.62	92.99	85.99	75.80
Sensitivity (%)	65	75	60	15
Specificity (%)	94.02	99.15	94.87	96.58
The testing set classification accuracy (%)	76.09	67.39	78.26	52.17
Sensitivity (%)	60.87	69.57	60.87	4.35
Specificity (%)	91.3	65.22	95.96	100

In addition to the above four commonly used kernel function, we can also customize the kernel function or make linear combination with two kernel functions to generate a mixed kernel function. In this paper, a new classification model employing a linear combination of linear kernel and RBF kernel was proposed. The mixed kernel function is as follows:

$$K_{\text{mix}} = \lambda \bullet K_{\text{line}} + (1 - \lambda) \bullet K_{\text{RBF}}. \qquad (5)$$

The parameter λ varies from $\lambda = 1$, when the mixed kernel function degenerate to linear kernel, to $\lambda = 0$ for the mixed kernel function degenerate to kernel RBF kernel. In practice, the value of λ depends on the characteristics of data. In this paper, we choose $\lambda = 0.8$, parameter $\gamma = 1$, penalty parameter $C = 2$. Classification model was established with training set. In comparison with the RBF kernel, the performance of mixed kernel show as following:

As shown in Table 3, the classification accuracy, sensitivity and specificity of mixture kernel are 82.61%, 69.57% and 95.65% (The PPV and NPV are 80% and 84.6%). For the fasting plasma glucose (FPG), at the impaired fasting glucose threshold (FPG = 100 mg/dl), the FPG testing sensitivity is

Table 3. Results of classification with RBF kernel and mixture kernel.

	Classification accuracy (%)	Sensitivity (%)	Specificity (%)
RBF function	78.26	60.87	95.96
Mixture kernel	82.61	69.57	95.65

Note: The classification accuracy, sensitivity and specificity were calculated under the optimal parameters.

J. Innov. Opt. Health Sci. 2013.06. Downloaded from www.worldscientific.com
by 103.240.126.9 on 10/21/18. Re-use and distribution is strictly not permitted, except for Open Access articles.

58% and the specificity is 77.4%. At that same specificity, the sensitivity for $\mathrm{HbA_{1C}}$ testing was 63.8% (The cut-off point is 6.8%).[2] Obviously, based on SVMs, skin autofluorescence classification method is more effective than traditional methods for diabetes screening.

4. Discussion

In this paper, there are several factors affecting the classification results:

(i) The autofluorescence that used in classification are distorted by the scattering and absorbing of human skin. It can be predicted that the development of the spectral correction technique[11] will promote the classification accuracy.

(ii) The data preprocessing of skin autofluorescence have the ability to reduce the interference of singular sample data and increase the speed of convergence, but at the same time it would inevitably lose some information.

(iii) In order to ensure a large enough optimization range during cross-validation and grid-optimization, we set the optimization stepping as 1, resulting in that the amount of the tested values of parameter C and γ is finite.

With the development of the attenuation correction techniques for tissue fluorescence, optimization of the normalized function and refinement of the optimization process, the above effects can be reduced or even eliminated, so the classification of skin autofluorescence based on SVM has a bright future.

Under the existing technical conditions, we can use RBF SVM algorithm to classify skin autofluorescence. If the classification result is not satisfactory, we can also consider linear combinations with linear and RBF kernels to construct a new kernel function. Cross-validation and grid-optimization methods are commonly used during the selection of the kernel function parameters and the penalty parameter. In order to ensure a large enough range of optimization, we are generally looking for the relationship between the logarithm of relevant parameters and classification rate. However, this will cause missing test of some parameters, and then the parameters we get may be not the best. At this time, we can first conduct wide range optimization, then pick up the parameters with higher classification accuracy from the results, and then reduce the step of optimization parameters to perform accurate optimization.

In this paper, skin autofluorescence of 63 patients with type 2 diabetes and 140 control subjects was obtained using a self-designed optical system. According to the four commonly used kernel functions in SVM, cross-validation and grid-optimization methods were employed to calculate the optimal parameters that maximize classification accuracy. Based on training set, model was set up and then verified by testing set. The test result indicated that the best choice for classification is radical basis function. Otherwise based on liner kernel function and radical basis function, a kind of mixed kernel function was built. Its accuracy, sensitivity and specificity were 82.61%, 69.57% and 95.65%, respectively, which show a better classification performance than radical basis function. In a comparison with FPG and $\mathrm{HbA_{1c}}$ test, at the impaired fasting glucose threshold ($\mathrm{FPG} = 100\,\mathrm{mg/dl}$), the FPG testing sensitivity is 58% and the specificity is 77.4%. At that same specificity, the sensitivity for $\mathrm{HbA_{1c}}$ testing was 63.8% (The cut-off point is 6.8%). Obvious, the SVM algorithm produced better diagnostic accuracy in all instances.

SVM algorithm was successfully implemented for the classification of skin autofluorescence from patients with type 2 diabetes and control subjects. The results demonstrate that skin autofluorescence spectroscopy classified by an SVM algorithm can achieve high diagnostic accuracy in differentiating diabetics from control subjects.

Acknowledgments

This study is supported by the Knowledge Innovation Program of the Chinese Academy of Sciences (083RC11124). We thank Huayi Hou and Fei Li, the graduate students in Anhui Institute of Optics & Fine Mechanics, for helping with data collection assistance.

References

1. H. Obayashi, K. Nakano, H. Shigeta, M. Yamaguchi, K. Yoshimori, M. Fukui, M. Fujii, Y. Kitagawa, N. Nakamura, K. Nakamura, "Formation of crossline

J. Innov. Opt. Health Sci. 2013.06. Downloaded from www.worldscientific.com
by 103.240.126.9 on 10/21/18. Re-use and distribution is strictly not permitted, except for Open Access articles.

as a fluorescent advanced glycation end product in vitro and in vivo," *Biochem. Bioph. Res. Co.* **226**(1), 37–41 (1996).

2. J. D. Maynard, M. Rohrscheib, J. F. Way, C. M. Nguyen, M. N. Ediger, "Noninvasive type 2 diabetes screening," *Diabetes Care* **30**(5), 1120–1124 (2007).
3. V. N. Vapnik, *The Nature of Statistical Learning Theory*, Springer-Verlag, New York (2000).
4. C. J. C. Burges, "A tutorial on support vector machines for pattern recognition," *Data Min. Knowl. Disc.* **2**(2), 121–167 (1998).
5. W. M. Lin, X. Yuan, P. Yuen, W. I. Wei, J. Sham, P. C. Shi, J. Qu, "Classification of in vivo autofluorescence spectra using support vector machines," *J. Biomed. Opt.* **9**(1), 180–186 (2004).
6. E. Widjaja, W. Zheng, Z. Huang, "Classification of colonic tissues using near-infrared Raman spectroscopy and support vector machines," *Int. J. Oncol.* **32**(3), 653–662 (2008).
7. Y. Wang, L. Zhu, L. Zhang, G. Zhang, Y. Liu, A. Wang, "A portable system for noninvasive assessment of advanced glycation end-products using skin fluorescence and reflectance spectrum," *J. Appl. Spectrosc.* **79**(3), 431–436 (2012).
8. A. Ji, J. Pang, H. Qiu, "Support vector machine for classification based on fuzzy training data," *Expert Syst. Appl.* **37**(4), 3495–3498 (2010).
9. C. L. Huang, C. J. Wang, "A GA-based feature selection and parameters optimization for support vector machines," *Expert Syst. Appl.* **31**(2), 231–240 (2006).
10. C. C. Chang, C. J. Lin, "LIBSVM: A library for support vector machines," *ACM Trans. Intell. Syst. Technol.* **2**(3), 27–56 (2011).
11. R. S. Bradley, M. S. Thorniley, "A review of attenuation correction techniques for tissue fluorescence," *J. R. Soc. Interface.* **3**(6), 1–13 (2006).