Journal of Scientific and Innovative Research 2014; 3 (2): 189-202
Available online at: www.jsirjournal.com

Research Article

ISSN 2320-4818
JSIR 2014; 3(2): 189-202

Received: 17-01-2014
Accepted: 11-04-2014

Neda Khorshidi
Department of Chemistry, Faculty of Science, Islamic Azad University, Arak Branch, Arak, Iran

Maryam Sarkhosh
Department of Chemistry, Faculty of Science, Islamic Azad University, Arak Branch, Arak, Iran

Ali Niazi
Department of Chemistry, Faculty of Science, Islamic Azad University, Arak Branch, Arak, Iran

Correspondence:
Ali Niazi
Department of Chemistry, Faculty of Science, Islamic Azad University, Arak Branch, Arak, Iran
Tel: +98 861 3670017
Fax: +98 861 3670017
E-mail: a-niazi@iau-arak.ac.ir, ali.niazi@gmail.com

QSPR study of maximum absorption wavelength of various flavones by multivariate image analysis and principal components-least squares support vector machine

*Neda Khorshidi, Maryam Sarkhosh, Ali Niazi**

Abstract

A novel quantitative structure-property relationships (QSPR) model has been developed for the maximum absorption wavelength (λ_{max}) of 69 flavones. Modeling of λ_{max} of these compounds as a function of the bidimensional images as descriptors was established by chemometrics methods. The resulted descriptors were subjected to principal component analysis (PCA) and the most significant principal components (PCs) were extracted. Multivariate image analysis applied to QSPR modeling was done by means of principal component-least squares support vector machine (PC-LSSVM) method. This model was applied for the prediction of the λ_{max} of flavones, which were not in the modeling procedure with low standard errors and high correlation coefficient. The resulted model showed high prediction ability with root mean square error of prediction of 0.3815 for PC-LSSVM.

Keywords: QSPR, Flavones, Maximum absorption wavelength, PC-LSSVM, Multivariate image analysis.

Introduction

Flavones are a class of flavonoids based on the backbone of 2-phenylchromen-4-one (2-phenyl-1-benzopyran-4-one) shown on Figure 1. In recent years, scientific and public interest in flavones has grown enormously due to their putative beneficial effects against atherosclerosis, osteoporosis, diabetes mellitus and certain cancers. Flavones intake in the form of dietary supplements and plant extracts has been steadily increasing.[1, 2]

Studies on quantitative structure-activity/property relationships (QSAR/QSPR) represent an important tool in agrochemistry, pharmaceutical chemistry, toxicology, and eventually most facts of chemistry and for this reason several investigations have been carried out in order to improve the results. QSAR/QSPR is mathematical model of activity in terms of structural descriptors. The QSAR/QSPR model is useful for understanding the factors controlling activity, prediction of activity and for designing new potent compounds.[3] The main aim of QSAR/QSPR studies is to establish an empirical rule or function relating the descriptors of compounds under investigation to activities or properties. This rule of function is then utilized to predict the same activities/properties of the compounds not involved in the training set from their descriptors. Whether the activity/property can be predicted with satisfactory accuracy

depends to a great extent on the performance of the applied multivariate data analysis method provided the property being predicted is related to the descriptors. Model development in QSAR/QSPR studies comprises different critical steps as (1) descriptor generation, (2) data splitting to calibration (or training) and prediction (or validation) sets, (3) variable selection, (4) finding appropriate model between selected variables and activity/property and (5) model validation.[4]

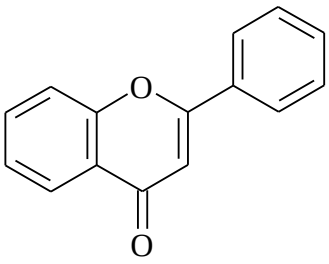


Figure 1: Chemical structure of 2-phenylchromen-4-one (2-phenyl-1-benzopyran-4-one)

Among the investigation of QSAR/QSPR, one of the most important factors affecting the quality of the model is the method to build the model. Many multivariate data analysis methods such as multiple linear regressions (MLR) and artificial neural network (ANN) have been used in QSAR/QSPR studies. However, the practical usefulness of MLR in QSAR/QSPR studies is rather limited, as it provides relatively poor accuracy. ANN offers satisfactory accuracy in most cases but tends to over fit the training data. The support vector machine (SVM) is a popular algorithm developed from the machine learning community. Due to its advantages and remarkable generalization performance over other methods, SVM has attracted attention and gained extensive applications.[5, 6] As a simplification of traditional of SVM, Suykens and Vandewalle have proposed the use of least-squares SVM (LS-SVM).[7, 8] LS-SVM encompasses similar advantages as SVM, but its additional advantage is that it requires solving a set of only linear equations (linear programming), which is much easier and computationally more simple. Theory of LS-SVM has been described clearly by Suykens *et al*.[7] and application of LSSVM in quantification and QSAR reported by some of the workers.[9-14]

A major step in constructing the QSAR/QSPR models is finding one or more molecular descriptors that represent variation in the structural property of the molecules by a number. Different descriptors have been studied to be used in QSAR analysis.[15] Nowadays, image analysis is becoming more important because of its ability to perform fast and non-invasive low-cost analysis on different processes in chemistry. Image analysis is a wide denomination that encloses classical studies on gray scale or (red-green-blue) RGB images.[16] Geladi and Esbensen[17] have demonstrated that image analysis may provide useful information in chemistry; through the descriptors do not have a direct physicochemical meaning, since they are binaries. In QSAR/QSPR, images (2D chemical structure) have shown to contain chemical information[18-20], allowing the correlation between chemical structures and properties.

The present study is focused on the application of 2D images, which are the proper structures of the compounds that can be drawn with aid of any appropriate program, as descriptors in QSAR/QSPR. Then, multivariate image analysis-quantitative structure property relationship study (MIA-QSPR) is proposed to model and predict the λ_{max} of a series of flavones by principal component–least squares support vector analysis (PC-LSSVM) modeling method.

Materials and computational methods

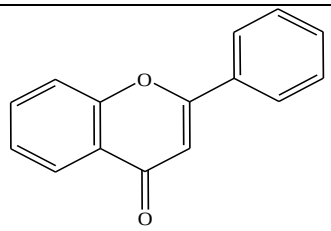
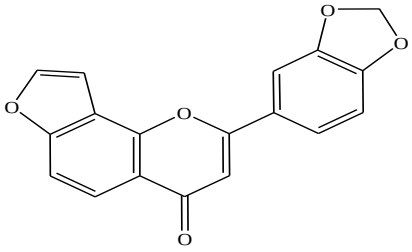
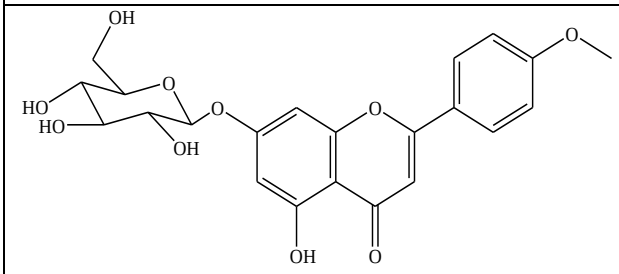
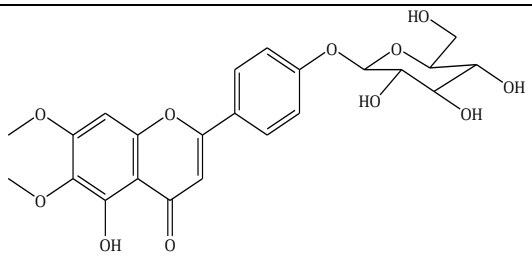
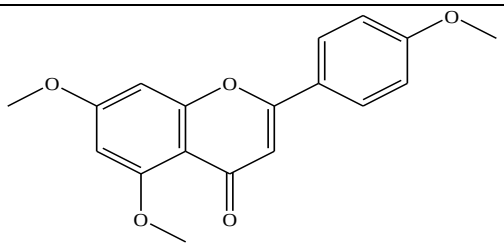
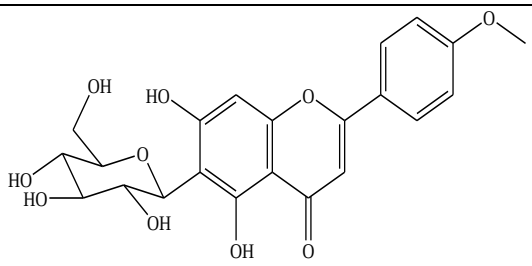
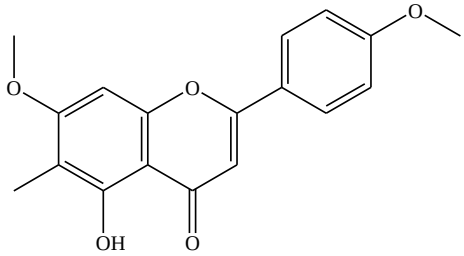
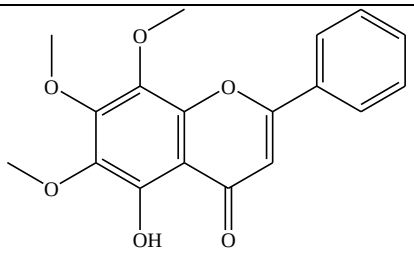
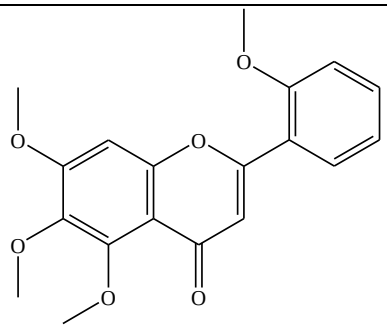
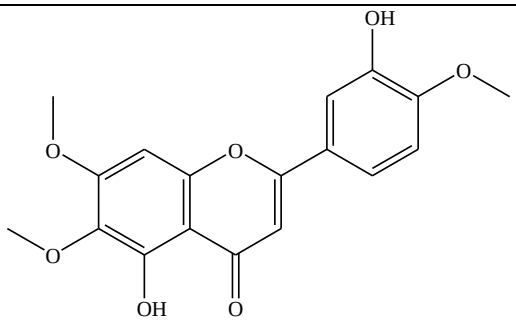
Hardware and software

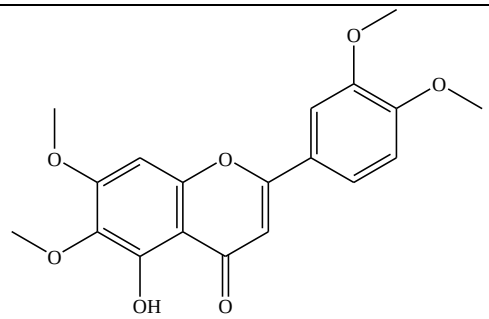
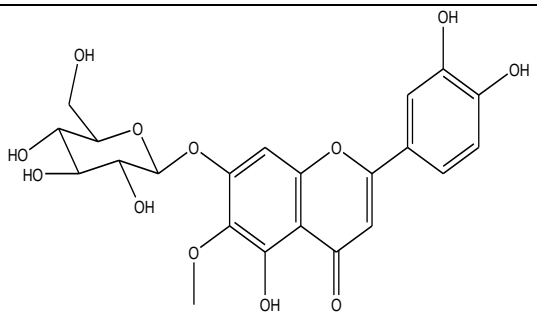
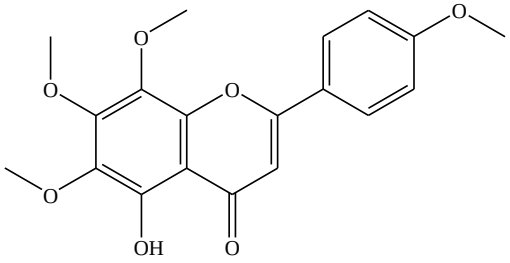
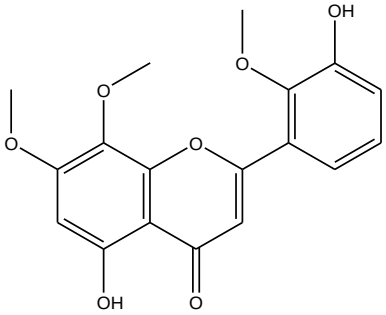
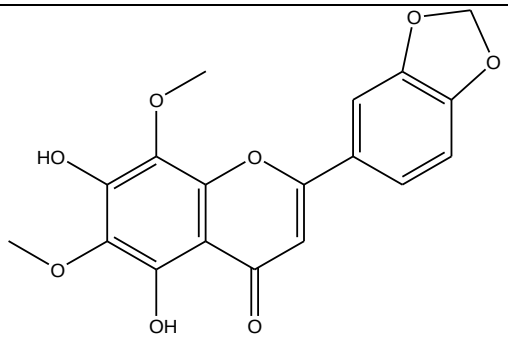
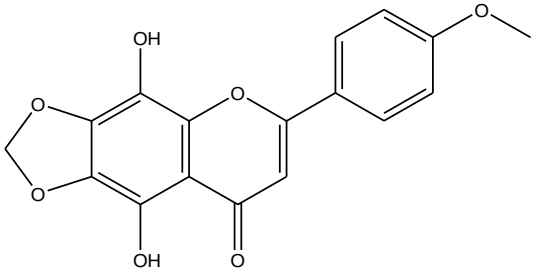
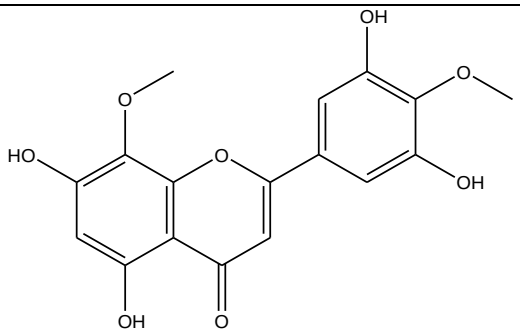
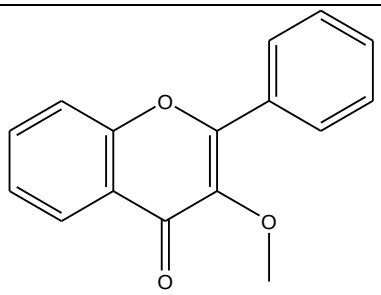
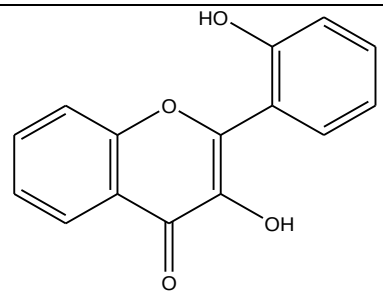
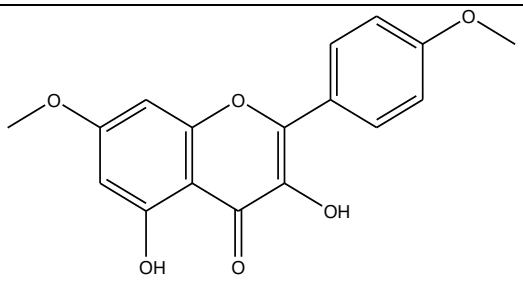
The HP Personal Computer (1 GB RAM) equipped with the Windows Vista operating system and MATLAB (Version 7.13, Mathwork Inc.) was used. The LS-SVM optimization and model results were obtained using the LS-SVM lab toolbox (Matlab/C Toolbox for Least-Squares Support Vector Machines)[21] and ChemOffice 2010 package was used to draw the molecular structure. Principal component analysis[22] and Kennard-Stones[23, 24] programs were written in MATLAB according to the algorithm.

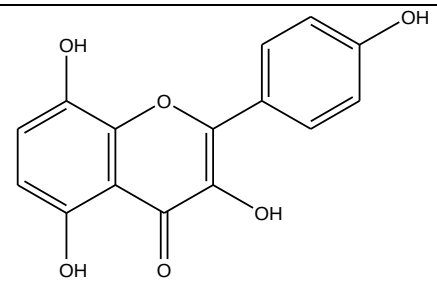
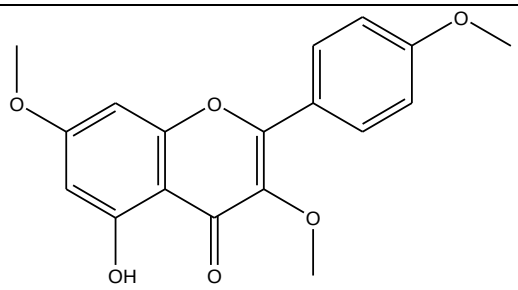
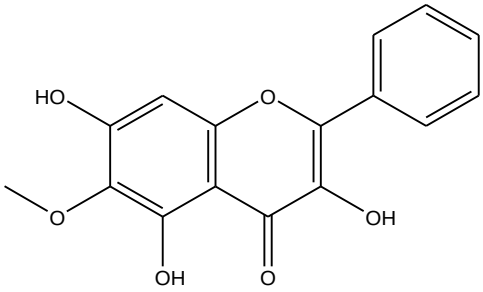
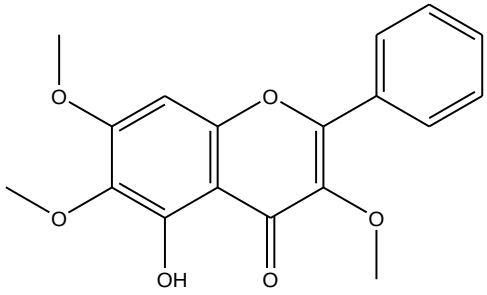
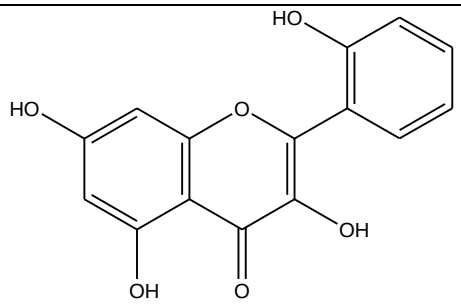
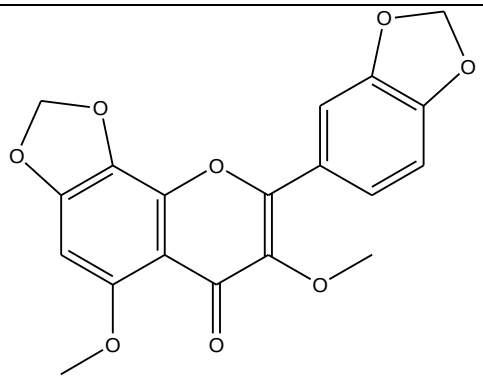
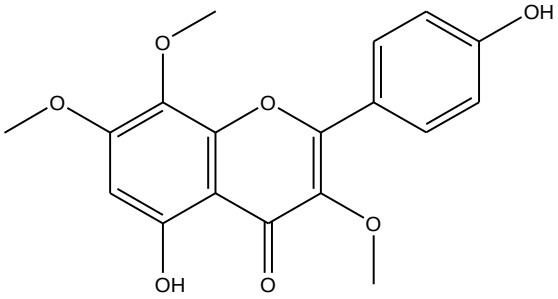
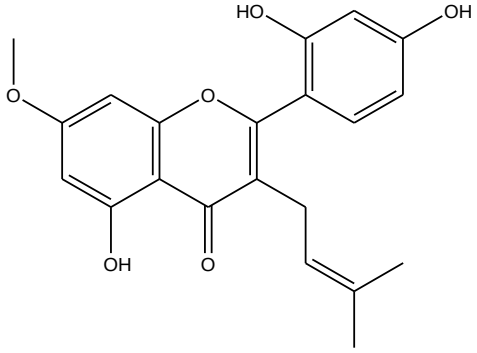
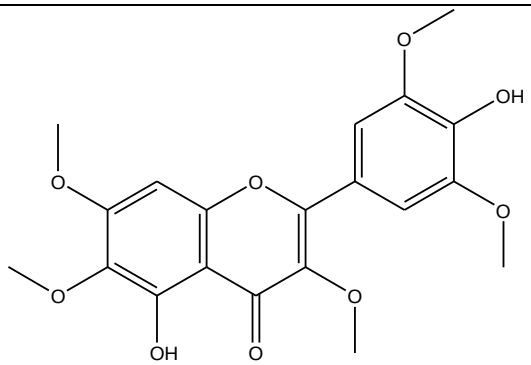
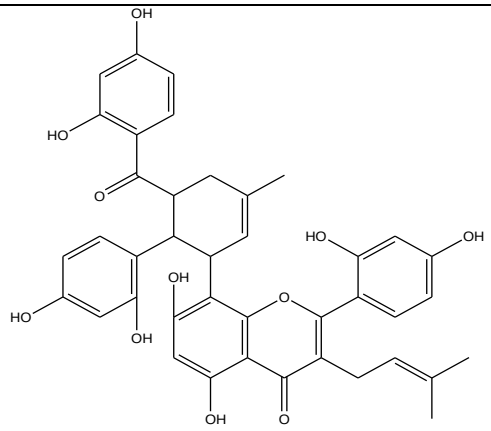
Data set

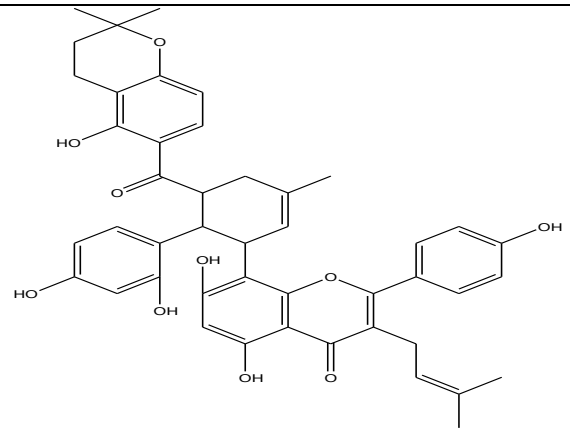
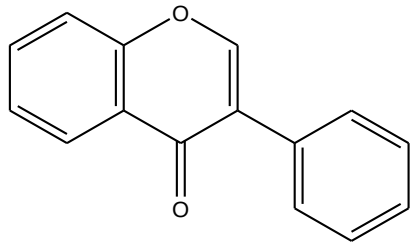
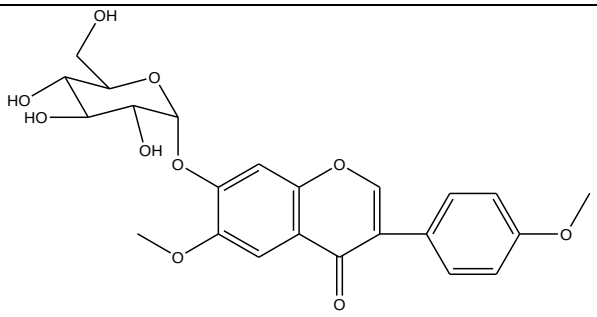
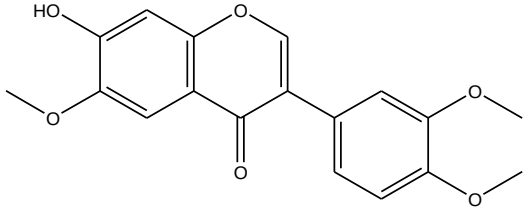
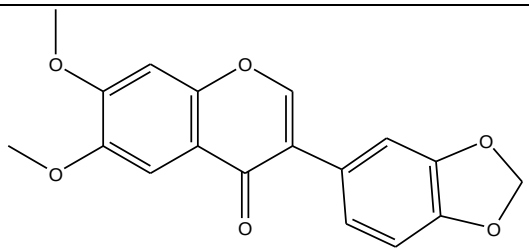
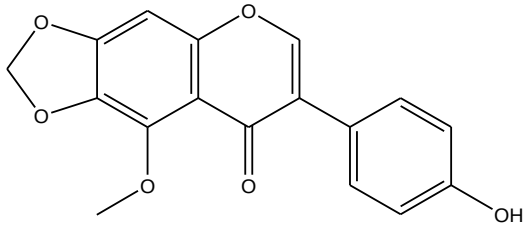
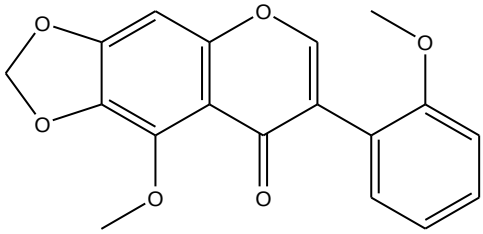
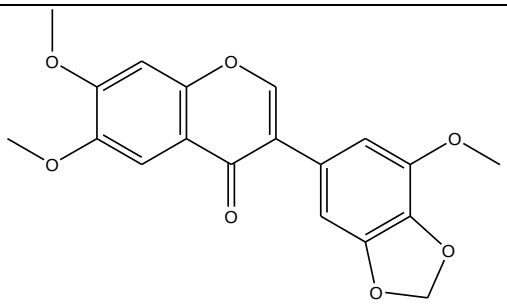
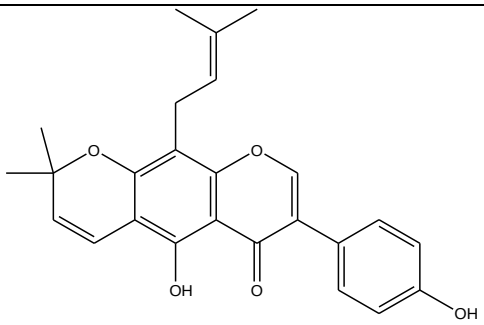
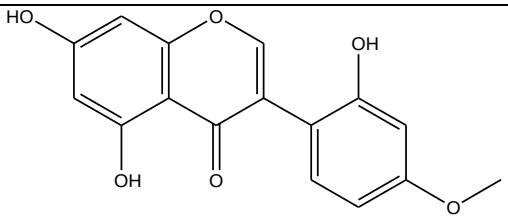
The λ_{max} of different flavones was taken from literature.[25] The parent structure of flavones is shown in Figure 1. The λ_{max} values were obtained in the same medium of ethanol. The chemical structures of these compounds and their corresponding λ_{max} are listed in Table 1. In order to guarantee that training and prediction sets cover the total space occupied by the original data set, the set was divided into the parts of training and prediction set according to the Kennard-Stones algorithm.[23, 24] The Kennard-Stones algorithm is known as one of the best ways of building training and prediction sets and it has been used in many QSAR/QSPR studies.

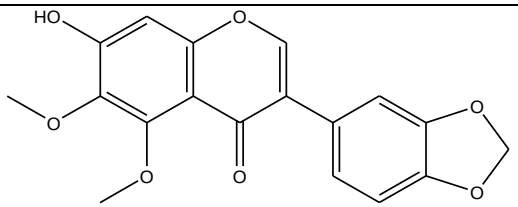
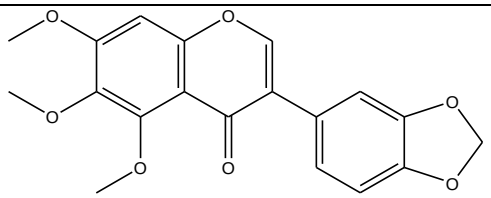
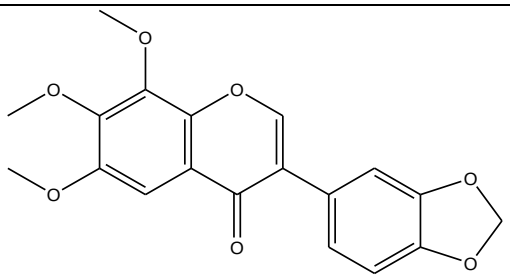
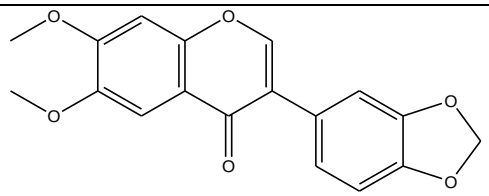
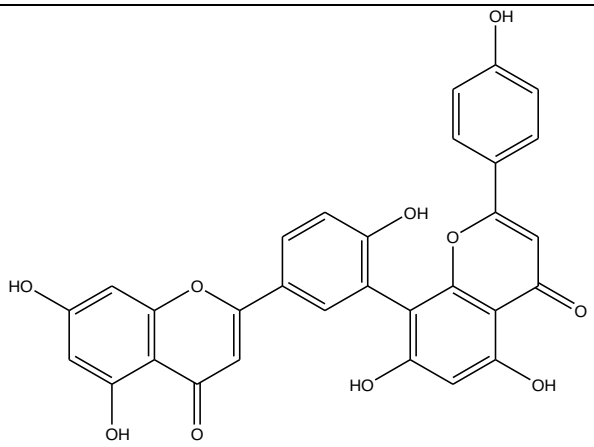
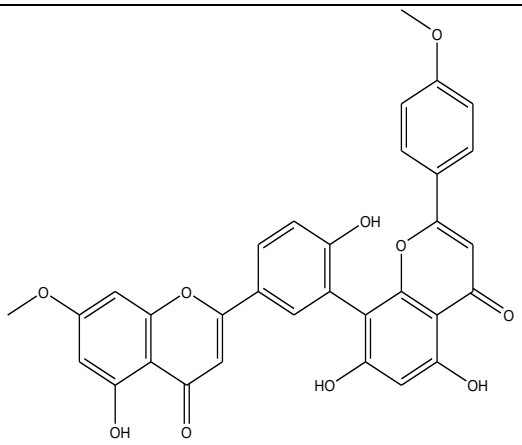
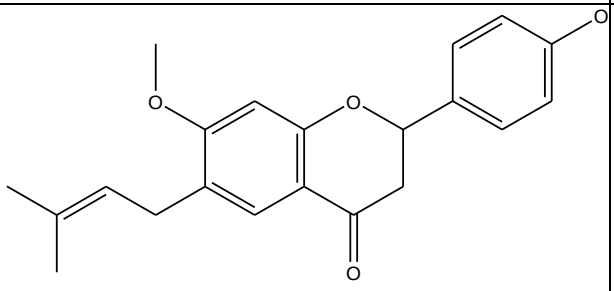
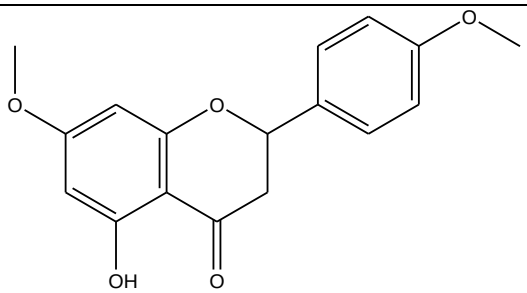
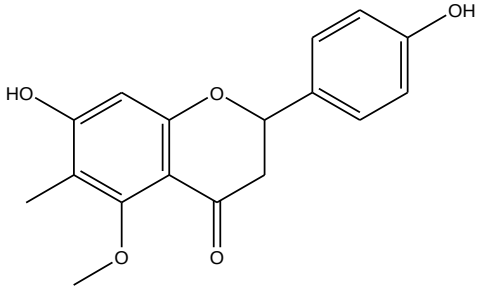
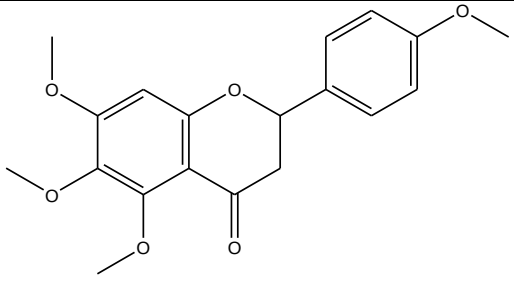
Table 1: Chemical structure of flavones and their corresponding λ_{max}

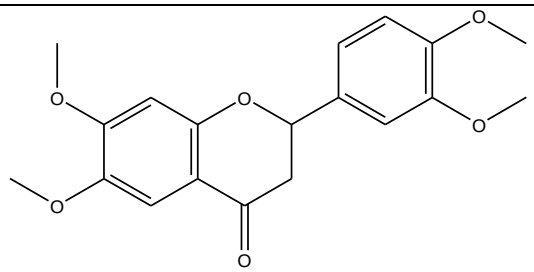
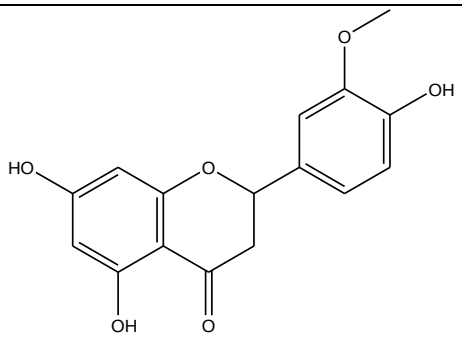
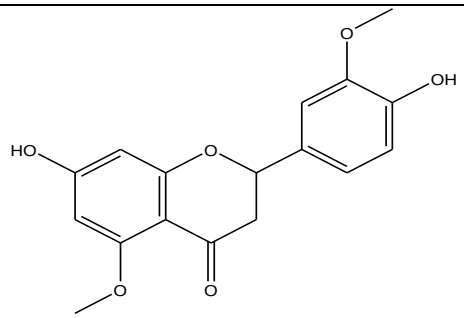
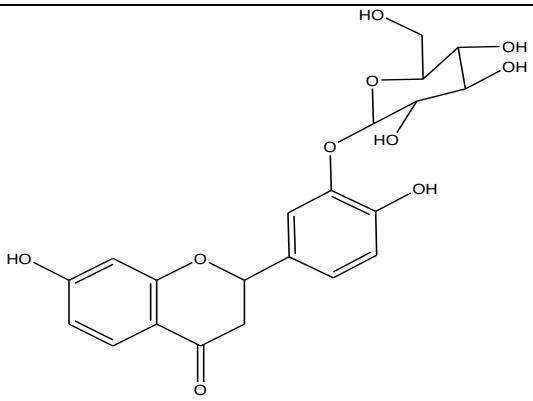
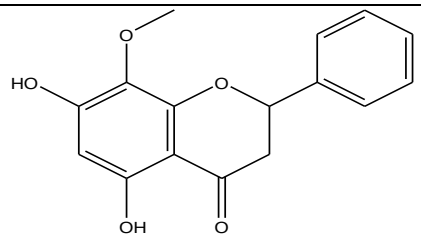
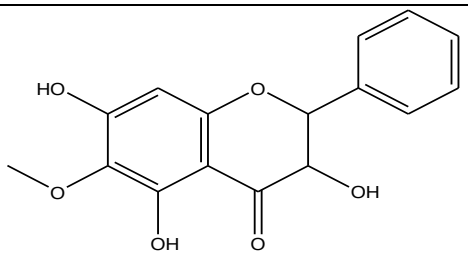
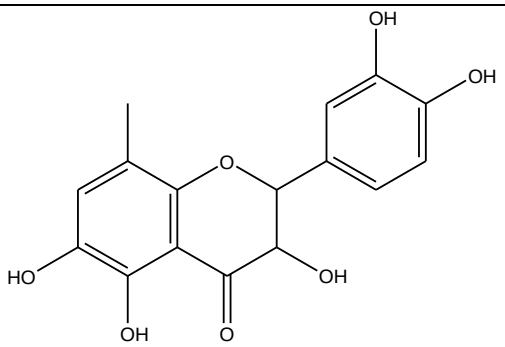
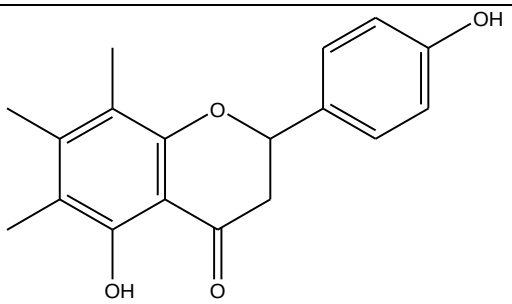
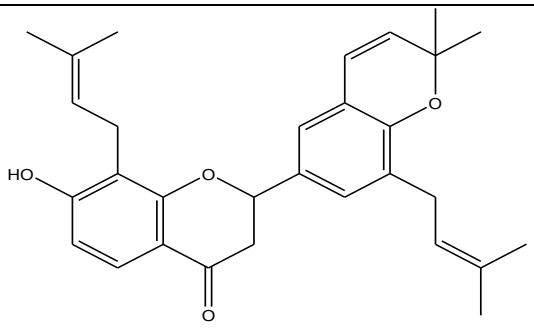
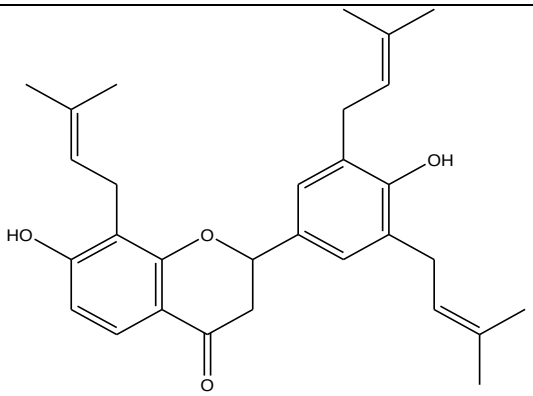
Compound	**λ_{max} (nm)**	**Compound**	**λ_{max} (nm)**
	250 test		249
	326		337
	325		332
	327 test		320
	322		342

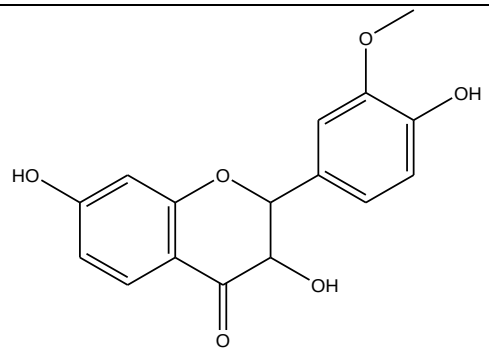
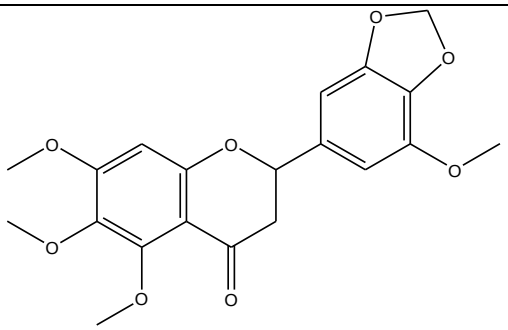
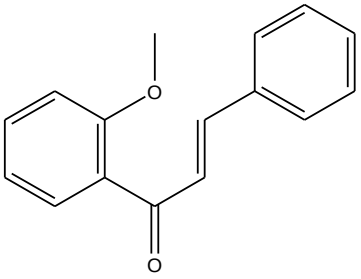
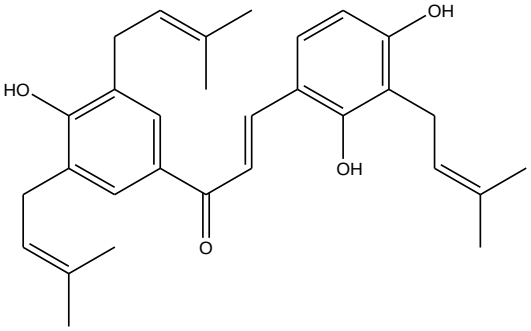
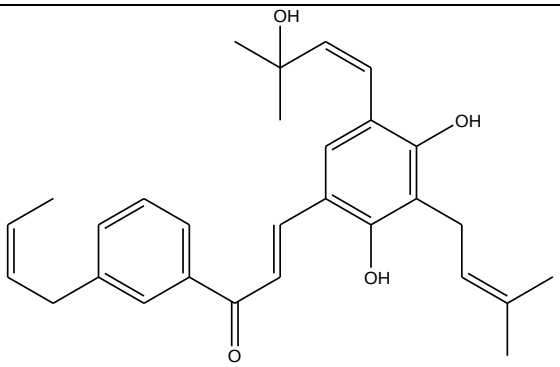
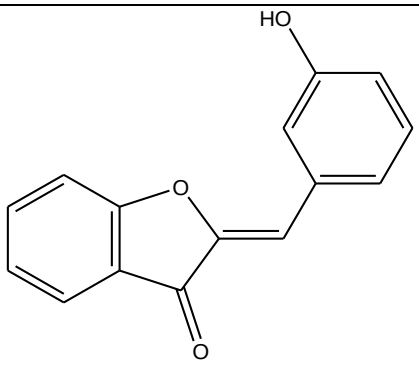
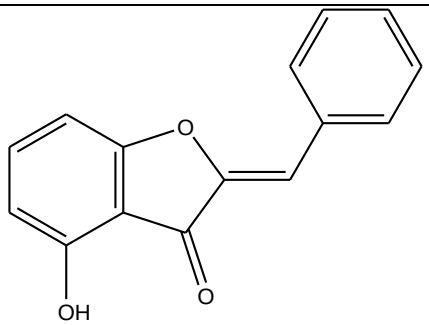
	342		340
	330		330
	285		280
	327		246
	244 test		322 test

	299		320
	325 test		317
	263		269
	331		315
	344		264

	264		245
	320		320
	266 test		270
	246		272
	276		290

	292		292
	296		246
	270		272
	272		288
	281		291

	278		276 test
	288		278
	290		295
	290 test		283
	286		286

	287 test		276
	366		380
	380		252
	307		

Multivariate image analysis descriptors

In the MIA-QSPR study, the pixel descriptors of images can be two or three dimensional. These pixels are correlated with dependent variables for making QSPR models. The 2D structures of each compound of Table 1 were systematically drawn in the ChemOffice program, and then, converted to bitmaps in 160×130 pixels workspace. All process in image is done in MATLAB.

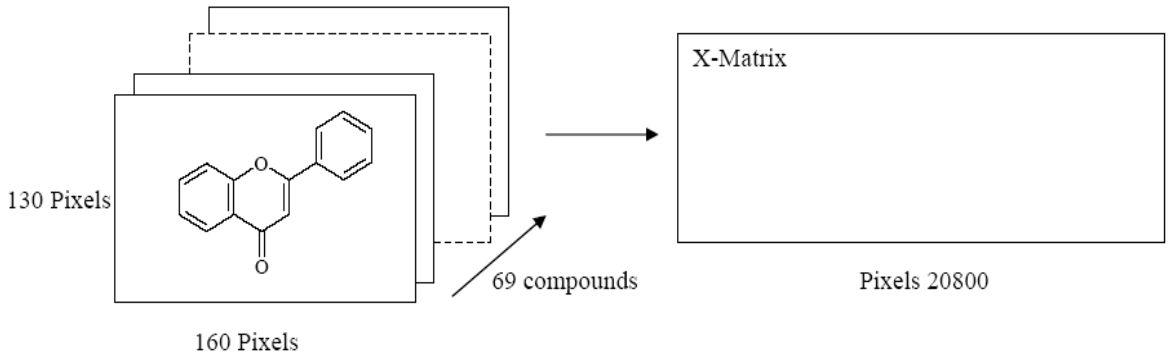


Figure 2: 2D images and unfolding step of the 69 chemical structures to give the X-matrix. The arrow in structure indicates the coordinate of a pixel in common among the whole series of compounds, used in the 2D alignment step

Results and Discussion

Multivariate image analysis descriptors

In the MIA-QSPR study is made according to the correlation of these pixels with dependent variables. The 2D structure of all flavones shown in Table 1, are drawn by ChemDraw program and then converted to bitmaps in 160×130 pixels workspace. All the drawn molecular structures were systematically fixed in a given coordinate. In this study, the pixel located at the 80×30 coordinate (carbonyl group), was used as reference in the alignment step, as illustrated in Figure 1. Each 2D image was read and converted into binaries (double array in MATLAB). Each image of dimension 160×130 pixels was unfolded at 20800 row and then the 69 images were grouped to form 69×20800 matrix. In order to minimize memory, the columns with zero variance were reduced. As the result the matrix will be reduced in 69×8014 dimensional and then all pixels data will be mean-centered.

Principal component analysis of the data set

To reduce the dimensionality of the independent variable space, a limited number of principal components (PCs) are used. Also, PCA was performed on the bidimensional images descriptors to the whole data set (Table 1), for investigation the distribution in the chemical space, which shows the spatial location of samples to assist the separation of data into training and prediction sets. The PCA results show that two PCs (PC1, PC2 and PC3) describe 90.70% of the overall variances: PC1=46.21%, PC2=28.63% and PC3=15.86% (Figure 3). Since almost all variables can be accounted for the first three PCs, their score plot is a reliable presentation of the spatial distribution of the points for the data set. As can be seen in Figure 3, there is not a clear clustering between compounds. The data separation is very important in the development of reliable and robust QSPR models. The quality of the prediction depends on the data set used to develop the mode. For regression analysis, data set was separated into two groups, a training set (60 data) and a prediction set (9 data) according to Kennard-Stones algorithm. As shown in Figure 3, the distribution of the compounds in each subset seems to be relatively well-balanced over the space of the principal components.

LS-SVM analysis

The LS-SVM multivariate calibration method is a powerful tool for modeling, because it extracts more information from the data and allows building more robust models. According to the basis of Kennard-Stones algorithm, 60 compounds of 69 were selected as the training set and the remaining 9 were selected as the test set. In first run (LSSVM), all pixels descriptors were considered for modeling; while in the second run (PC-LSSVM), after achieving PCs, PCs were used as the input to develop nonlinear model by LS-SVM. The quality of LS-SVM for regression depends on γ and σ^2 parameters. In this work, LS-SVM was performed with radial basis function (RBF) as a kernel function. To determine the optimal parameters, a grid search was performed based on leave-one-out cross-validation on the original training set for all parameter combinations of γ and σ^2 from 1 to 10 and 1 to 500, respectively, with increment steps of 1. Table 2 shows the optimum γ and σ^2 parameters for the LS-SVM and RBF kernel, using the training sets for 60 flavones compounds.

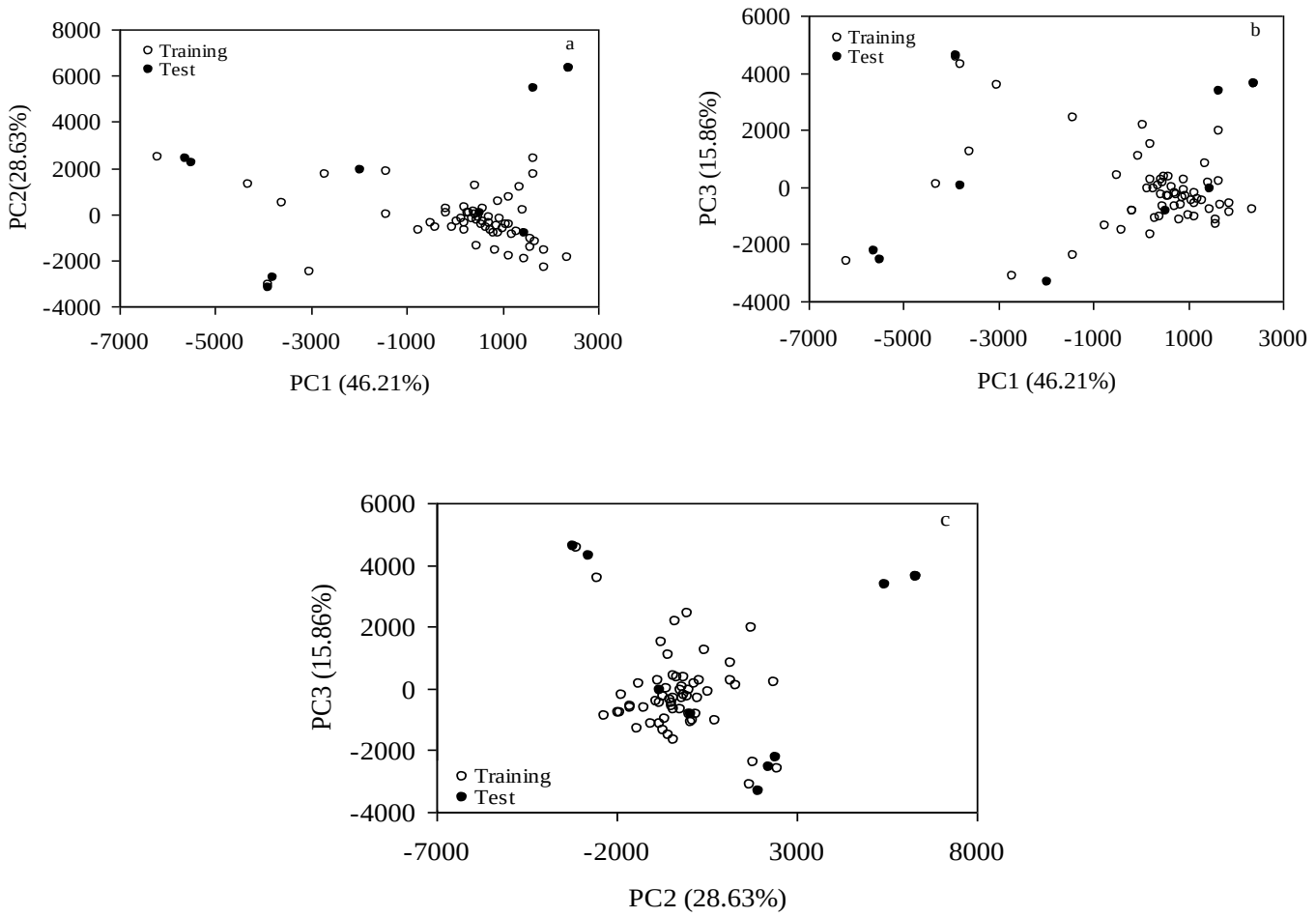


Figure 3: Principal components analysis of the 2D image descriptors for the data set, (a) PC1 versus PC2, (b) PC1 versus PC3 and (c) PC2 versus PC3

Model validation and prediction of λ_{max}

The predictive ability of these methods (LSSVM and PC-LSSVM) was determined using nine data (their structures are given in Table 1). Validation of predictive ability is another key step in QSPR studies. Several statistical parameters have been used for the evaluation of the suitability of the developed QSPR models for prediction of the property of the studied compounds this include cross validation coefficient (Q^2 or R^2), the root mean square error of prediction (RMSEP) and relative standard error of prediction (RSEP), validation through an external prediction set.

$$Q^2 = 1 - \frac{\sum_{i=1}^{n}(y_{i,pred} - y_{i,obs})^2}{\sum_{i=1}^{n}(y_{i,pred} - \bar{y})^2}$$

$$RMSEP = \sqrt{\frac{\sum_{i=1}^{n}(y_{i,pred} - y_{i,obs})^2}{n}}$$

$$RSEP(\%) = 100 \times \sqrt{\frac{\sum_{i=1}^{n}(y_{i,pred} - y_{i,obs})^2}{\sum(y_{i,obs})^2}}$$

where $y_{i,pred}$ is the predicted of the λ_{max} using different model, $y_{i,obs}$ is the observed value of the λ_{max}, and n is the number of molecules in the prediction set. The statistical parameters obtained by LSSVM and PC-LSSVM methods are listed in Table 2 and Table 3.

Table 2: Observation and calculation values of λ_{max} using LSSVM and PC-LSSVM models

Number of compounds (Table 1)	Observation λ_{max} (nm)	LSSVM Model		PC-LSSVM Model	
		Calculation λ_{max} (nm)	Error (%)	Calculation λ_{max} (nm)	Error (%)
1	250	245.3	-1.88	250.0	0.00
7	327	320.1	-2.11	326.9	-0.03
19	244	241.3	-1.11	244.0	0.00
20	322	318.6	-1.6	321.9	-0.03
23	325	314.3	-3.29	324.8	-0.06
35	266	263.1	-1.09	266.0	0.00
54	276	279.5	1.27	277.1	0.40
57	290	293.5	1.21	290.0	0.00
63	287	288.6	0.56	287.2	0.07
γ		0.2		0.1	
σ^2		100		50	
RMSEP		5.1479		0.3815	
RSEP (%)		1.7813		0.132	

Table 2 shows RMSEP, RSEP and the percentage error for prediction of λ_{max} of flavones. As can be seen, the percentage error was also quite acceptable only for PC-LSSVM. Good results were achieved in PC-LSSVM model with percentage error ranges from -0.06 to 0.40 for λ_{max} of Flavones. The plots of the predicted λ_{max} versus actual values are shown in Figure 4 for each model (line equations and R^2 values are also shown). The correlation coefficients (R^2) for PC-LSSVM model were better than the model and close to one. Also, it is possible to see that PC-LSSVM presents excellent prediction abilities when compared with LSSVM. Also, obtained results indicated that MIA descriptors are capable to recognize the physicochemical information and may be useful to predict λ_{max}. Also, it is possible to see that PC-LSSVM presents excellent prediction abilities when compared with LSSVM.

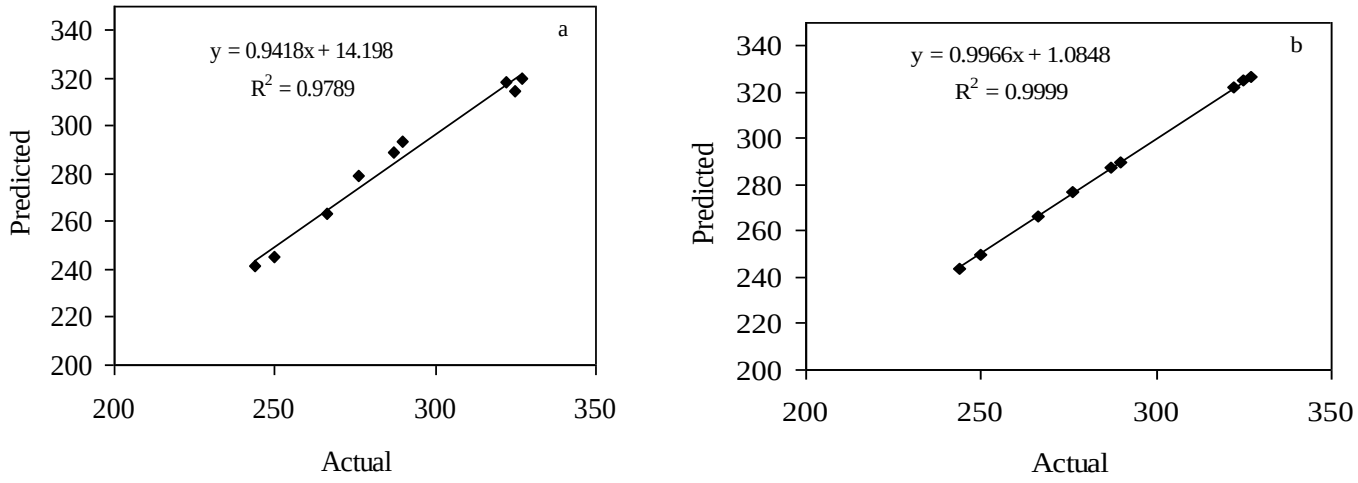


Figure 4: Plots of predicted versus actual λ_{max} (nm) for flavones with (a) LSSVM and (b) PC-LSSVM

Other statistical parameters have been used for the evaluation of the suitability of the developed models for prediction of the activity of the studied compounds this include cross validation coefficient (Q^2 and R^2). These parameters are listed in Table 3 and show the good statistical qualities.

Table 3: Comparison of the statistical parameters by different QSPR models for the prediction of the λ_{max}

Methods	Data set	R^2	Q^{2*}
LSSVM	Training	0.9862	0.9267
	Test	0.9706	0.9172
PC-LSSVM	Training	0.9999	0.9685
	Test	0.9994	0.9537

*Q^2 coefficient for the model validation by leave-one-out.

Conclusion

The QSPR model has been successfully developed with a good correlative and predictive ability for predicting λ_{max} property for 69 compounds based multivariate image analysis. This QSPR model exhibiting a high degree of accuracy was when validated by predicting the λ_{max} of experimental compounds in the external test. The results well illustrate the power of pixel descriptors in prediction of λ_{max} of flavones. The work is the first application of MIA descriptors and PC-LSSVM for QSPR study and shows that MIA descriptors are capable to recognize the physicochemical information and may be useful to predict the maximum absorption wavelengths.

References

1. Riter J.K., Chen F,. Sheen Y.Y., Tran H.M., Kimura S., Yeatman M.T., Owens I.S., A novel complex locus UGT1 encodes human bilirubin, phenol, and other UDP-glucuronosyltransferase isozymes with identical carboxyl termini. J. Bio. Chem. 1992; 267: 3257-3261.

2. Cermak R., Wolffram S., The potential of flavonoids to influence drug metabolism and pharmacokinetics by local gastrointestinal mechanisms. Curr. Drug. Metab. 2006; 7: 729-744.

3. Cui W., Yan X., Adaptive weighted least square support vector machine regression integrated with outlier detection and its application in QSAR. Chemometr. Intell. Lab. Syst. 2009; 98: 130-135.

4. Sarkhosh M., Ghasemi J., Ayati M., A quantitative structure-property relationship of gas chromatographic / mass spectrometric retention data of 85 volatile organic compounds as air pollutant materials by multivariate methods. Chem. Central. J. 2012; 6: 1-8.

5. Belousov A.I., Verzakov S.A., Von Frese A., Flexible classification approach with optimal generalization performance: support vector machines. Chemometr. Intell. Lab. Syst. 2002; 64; 15-25.

6. Burbidge R., Trotter M., Buxton B., Holden S., Drug design by machine learning: support vector machines for pharmaceutical data analysis. Comput. Chem. 2001; 26: 5-14.

7. Suykens J.A.K., Vandewalle J., Least squares support vector machine classifiers. Neural Process. Lett. 1999; 9: 293-300.

8. Suykens J.A.K., van Gestel T., de Brabanter J., de Moor B., Vandewalle J., Least-Squares Support Vector Machines, World Scientifics, Singapore, 2002.

9. Niazi A., Zolgharnein J., Afiuni-Zadeh S., Spectrophotometric determination of ternary mixtures of thiamin, riboflavin and pyridoxal in pharmaceutical and human plasma by least-squares support vector machines. Anal. Sci. 2007; 23: 1311-1316.

10. Niazi A., Ghasemi J., Yazdanipour A., Simultaneous spectrophotometric determination of nitroaniline isomers after cloud point extraction by using least-squares support vector machines. Spectrochim Acta Part A. 2007; 68: 523-530.

11. Niazi A., Ghasemi J., Zendehdel M., Simultaneous voltammetric determination of morphine and noscapine by adsorptive differential pulse stripping method and least-squares support vector machines. Talanta 2007; 74: 247-254.

12. Niazi A., Jameh-Bozorgi S., Nori-Shargh D., Prediction of toxicity of nitrobenzenes using ab initio and least squares support vector machines. J. Hazard. Mater. 2008; 151: 603-609.

13. Li J., Liu H., Yao Z., Liu M., Hu Z., Fan B., Structure–activity relationship study of oxindole-based inhibitors of cyclin-dependent kinases based on least-squares support vector machines. Anal. Chim. Acta 2007; 581: 333-342.

14. Xu Y., Zomer S., Brereton R., Support vector machines: a recent method for classification in Chemometrics. Crit. Rev. Anal. Chem. 2006; 36: 177-188.

15. Todeschini R., Consonni V., Handbook of Molecular Descriptors, Wiley-VCH, Weinheim, 2000.

16. Prats-Montalban J.M., de Juan A., Ferrer A., Multivariate image analysis: A review with applications. Chemometr. Intell. Lab. Syst. 2011; 107: 1-23.

17. Esbensen K., Geladi P., Strategy of multivariate image analysis (MIA). J. Chemometr. 1989; 7: 67-86.

18. Freitas M.P., MIA-QSAR modelling of anti-HIV-1 activities of some 2-amino-6-arylsulfonylbenzonitriles and their thio and sulfinyl congeners. Org. Biomol. Chem. 2006; 4: 1154-1159.

19. Freitas M.P., Brown S.D., Martins J.A., MIA- QSAR: A simple 2D image-based approach for quantitative structure-activity relationship analysis. J. Mol. Struc. 2005; 738: 149-154.

20. Garkani-Nejad Z., Poshteh-Shirani M., Application of multivariate image analysis in QSPR study of 13C chemical shifts of naphthalene derivatives: a comparative study. Talanta 2010; 83: 225-232.

21. Vapnik V., (Eds.) Suykens J.A.K., Vandewalle J., Nonlinear Modeling: Advanced Black-Box techniques, Kluwer Academic Publishers, Boston, 1998.

22. Martens H., Naes T., Multivariate Calibration, John Wiley, Chichester, 1989.

23. Kennard R.W., Stones L.A., Computer aided design of experiments. Technometrics 1969; 11: 137-148.

24. Daszykowski M., Walczak B., Massart D.L., A comparison of two algorithms for warping of analytical signals. Anal. Chim. Acta 2002; 468: 91-93.

25. Ke Y.K., Dong H.R., Handbook of Analytical Chemistry, Chemical Industry Press, Beijing, 1988.