Prediction of the xanthine oxidase inhibitory activity of celery seed extract from ultraviolet–visible spectrum using machine learning algorithms

Nguyen Thu Hang¹ · Nguyen Van Phuong¹ · Le Viet Hoang¹ · Ngo Minh Khoa¹ · Le Thi Trang¹

Received: 10 June 2020 / Accepted: 17 September 2020 / Published online: 29 September 2020
© Springer Nature Switzerland AG 2020

Abstract

Xanthine oxidase inhibitory activity of celery seed extract was predicted from UV–Vis spectrum using five different algorithms: multiple linear regression, artificial neural network, support vector regression, random forest and partial linear regression. The optimal parameters of each model established by these methods were examined carefully and the best model was selected. The results showed that the model obtained by partial linear regression had the best quality with coefficient of determination $R^2=0.9618$, correlation coefficient of leave-one-out cross validation $Q^2=0.8746$, average absolute deviation $MAD=0.0671$, root mean square error $RMSE=0.0761$ and the accuracy of the model with train and test were 99.29% and 96.69%, respectively. Our findings suggest that this model could be an effective method for rapidly determining xanthine oxidase inhibitory activity of herbal medicine, particularly celery seed extract from UV–Vis spectra. However, more investigations are required in order to improve the performance of the partial linear regression model.

Electronic supplementary material The online version of this article (<https://doi.org/10.1007/s42452-020-03542-z>) contains supplementary material, which is available to authorized users.

✉ Nguyen VanPhuong, phuongnv@hup.edu.vn | ¹Department of Pharmacognosy, Hanoi University of Pharmacy, Hanoi, Vietnam.

SN Applied Sciences (2020) 2:1746 | <https://doi.org/10.1007/s42452-020-03542-z>

Graphic abstract

Keywords Celery seed · machine learning · Partial linear regression · UV-vis spectrum · Xanthine oxidase inhibitor

1 Introduction

Xanthine oxidase (XO) is an important enzyme that catalyzes the oxidation of xanthine to uric acid, the last reaction of metabolic pathway to breakdown purine in the body. Inhibiting the activity of the enzyme xanthine oxidase (XO), thereby reducing the formation of uric acid has been shown to be one of the most effective treatments for gout. However, the XO inhibitors such as allopurinol and febuxostat still have some undesirable effects including skin rash, nausea, vomiting, kidney failure, Steven—Johnson syndrome... [16]. Therefore, searching for compounds of natural origin with xanthine oxidase inhibitory effect and low toxicity has been recently considered as an alternative solution in the discovery and development of new drug for gout treatment.

Celery (*Apium graveolens* L.) has been widely used in folk medicine for treatment of rheumatoid arthritis [19]. The hydroalcoholic extract of celery was also demonstrated to have effects on lowering uric acid level in vivo by inhibiting

the XO activity [6]. Flavonoids, which are the main active constituents of celery extract, have been confirmed to have XO inhibitory effect in many research [4, 14]. In this case, the activity of celery extract may result from a combination of several flavonoids instead of a single compound [27].

In the traditional approach to quality control of herbal medicine, one or several chemical compounds, as known as markers, are chosen and analyzed. However, in most case, these compounds were not tested whether they have pharmacological activities the same as the herbal medicine [3]. On the other hand, because of the complexity of the chemical components and mechanism of action, the quality of herbal product can not be accurately reflected using this approach [24]. To overcome this shortcoming, *spectrum-effect relationship* approach was proposed and has been considered as an interdisciplinary and innovative science which allows to study plant materials more accurately [26]. The number of publishes on this field has risen dramatically in the past two decades. Some remarkable results were listed in Table 1:

Table 1 Some recent spectrum-effect relationship researches

Extract	Spectrum data	Activity	Algorithm	Reference
<i>Turnera diffusa</i>	HPLC	Antioxidant	PLS	[12]
<i>Curcuma longa</i>	HPLC	Anti cancer	SVR	[9]
<i>Tinospora sagittata</i>	HPLC	Analgesic	GA-SVR	[28]
<i>Chrysanthemum morifolium</i>	HPLC	XO inhibition	PLS	[18]
<i>Cureuma kwangsiensis</i>	GC	Antitumor	GRDA	[11]
<i>Juglans mandshurica</i>	TLC	Antitumor	CA	[29]
<i>Zanthoxylum nitidum</i>	IR	Antitumor	MLR	[25]

In order to obtain spectrum data, a various number of methods could be used, including high-performance liquid chromatography (HPLC), gas chromatography (GC), infrared spectroscopy (IR)... [26]. However, the main disadvantages of these methods are the requirement of sophisticated and expensive equipment which are not always affordable in a common laboratory. Another option which is much simpler and more cost-effective is ultraviolet-visible (UV-Vis) spectroscopy. Therefore, our study was conducted with the aim of developing a quantitative model that demonstrates the relationship between the UV-Vis spectrum and xanthine oxidase inhibitory effect of celery seed extract.

2 Materials and methods

2.1 Materials

Celery seeds were collected in June 2018 in Hai Hau, Nam Dinh, Vietnam. The plant material was identified and authenticated by Prof. Hang Nguyen Thu, Department of Pharmacognosy, Hanoi University of Pharmacy. The voucher specimen (ID: HNIP/18,542/19) was deposited at the Herbarium of Department of Botany—Hanoi University of Pharmacy. Post-harvest seeds were dried and stored in a cool, dry place.

2.2 Extraction

Celery seeds were ground, accurately weighed (about 15.00 g) then transferred into the round bottom flask. After that, the solvent (ethanol/water) was added and the mixture was heated for 1 h and filtered. The process was repeated one more time. Next, combine all the extract then 1.50 g solid paraffin was added followed by stirring the mixture for 10 min at 70 °C. Allow the mixture to cool then remove the paraffin and evaporate the solvent. By modifying three extraction conditions: ethanol concentration, temperature and solvent:solid ratio, 17 different celery seed extracts were prepared.

2.3 Ultraviolet-visible (UV-Vis) spectroscopy

The mixture of 0.1500 g of celery seed extract and 20 ml of methanol was homogenized by ultrasound for 10 min followed by centrifuging at the speed of 3000 rpm for 10 min. After the supernatant was collected, the residue was ultrasonically extracted and then centrifuged with the aforementioned parameters three more times. The stock solution was prepared by combining all the after-centrifugation supernatant then transferring into a 100 ml volumetric flask, making up to the mark with methanol and mixing. Taking exactly 0.5 ml of the stock solution into a 10 ml volumetric flask, followed by adding 5 ml of 1% triethylamine in methanol and making up to the mark with methanol to obtain the test solution. The absorbance of the test solution (A) was scanned over a range of wavelength, from 190 to 600 nm with a step of 5 nm using UV-Vis Hitachi U-1900 spectrophotometer.

2.4 Xanthine oxidase inhibitory activity assay

The XO inhibitory effect of celery seed extract was assessed on Corning 96-well UV plates using the method described by Nguyen et al. [15] with minor modifications. Each well contained 50 µl of test solution; 35 µl of 70 mM phosphate buffer (pH = 7.5); 30 µl of 0.01 U/ml enzyme solution (in 70 mM phosphate buffer) which was prepared immediately before use. After incubation at 25 °C for 15 min, 60 µl of 150 µM xanthine solution was added. The mixture was continued to incubate at 25 °C for 30 min then 25 µl of 1 N HCl was added. The absorbance was measured at the wavelength of 290 nm. Each experiment was repeated 3 times. The half maximal inhibitory concentration (IC₅₀) of the extracts was determined using GraphPad Prism 8.0 software. Quercetin was used as positive control.

2.5 Model building

2.5.1 Data preprocessing and feature selection

The response variable (logIC₅₀) was obtained by taking the common logarithm of the half maximal inhibitory

concentration (IC_{50}) while the absorbances were considered as independent variables (features). Then the initial dataset was randomly divided into a training set (12 samples-70%) to select important features and establish models and a test set (5 samples-30%) for evaluation of the models using SPSS 22.0 software. The independent variables (features) that had the most influence on $logIC_{50}$ were selected by correlation-based feature selection algorithm using Weka 3.8 software.

2.5.2 Model building

Five models were constructed using five methods: multiple linear regression (MLR), artificial neural network (ANN), support vector regression (SVR), random forest (RF) and partial linear regression (PLS) respectively, using Weka 3.8 software.

2.5.2.1 Multiple linear regression (MLR) MLR model is expressed as following equation:

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \dots + \beta_n X_n$$

where the dependent variable Y is $logIC_{50}$, the independent variable X_i is the absorbance at selected wavelength, β_i is the coefficient and β_0 is the intercept [10].

2.5.2.2 Artificial neural network (ANN) Artificial neural network is the method that simulates the learning and information processing of human brain [17]. The structure of a typical artificial neuron network consists of artificial neurons that are group in three type of layers: one input layer, one output layer and one or more hidden layer(s). In particular, the input layer represents the independent variables (absorbance), the output layer is specific to the dependent variable ($logIC_{50}$), the

hidden layer is specific to the information processing process of the network. During the learning process, based on the difference between the predicted value and the observed value of the output variable in the previous loop, the weight of the input variables will be adjusted to minimize the above error [2]. In this study, three parameters that affect to the quality of ANN model were investigated, including: number of neurons in the hidden layer (1–7), learning rate (0.1–0.9 with increments of 0.2) and momentum (0.1–0.9 with increments of 0.2).

2.5.2.3 Support vector regression (SVR) The support vector machine algorithm was first proposed by V. Vapnik [5] and could be applied to solve both classification and regression problems. In support vector regression, a type of support vector machine, which deals with regression tasks, the object is to find a function $y=f(x)$ that is as flat as possible so that the error between y_i predicted by the model and the observed value is not greater than the given value ϵ . The SVR model could be either linear or nonlinear functions [23]. In this study, nonlinear SVR with RBF kernel function was used. Two parameters that affect the quality of the SVR model were investigated, including: C (1–500 with increments of 50) and gamma γ (0.1–0.9 with increments of 0.2).

2.5.2.4 Random forest (RF) Random forest is an ensemble learning method which was first proposed by L. Breiman and has become one of the best-performing learning algorithms [20, 22]. In this study, different values of *number of trees* (from 10 to 100) were investigated to obtain the best quality model.

2.5.2.5 Partial linear regression (PLS) In case of high dimension data, to use MLR or other algorithms such as

Fig. 1 The UV–Vis spectra of 17 celery extracts

ANN, SVR, RF, feature selection needs to be performed. Another solution is to reduce the dimension of data by creating new variables using PLS regression. This algorithm is particularly suited when there are more variables than observations or multicollinearity exists among independent variables [8]. In this study, the number of latent variables which were created from 83 independent variables was investigated to obtain the best model.

2.5.3 Validation

The established models were chosen based on determination coefficient R^2 , *leave one out* (LOO) correlation coefficient Q^2 , LOO average absolute deviation (MAD), LOO root mean square error (RMSE) and the accuracy of predicted values of dependent variable of training set and test set.

2.5.4 Y randomization

Y randomization is performed by randomly shuffling the dependent variable while keeping the remain variables unchanged to confirm that the developed model is not obtain by chance. The established model is considered reliable if the values of Q^2_{LOO} of new models are significantly lower than Q^2 of the initial model [1].

2.6 Application domain

The application domain of the best model was determined using leverage approach. An observe is considered as a

outlier if the h_i value is larger than the warning leverage h^* and the standardized residual is greater than 2.0. The warning leverage h^* is calculated using the following equation:

$$h^* = \frac{3(k+1)}{n}$$

where k is the number of features and n is the number of observes [21].

2.7 Sensitivity analysis

Sensitivity analysis studies how the output (Y) is affected when the features (X) changes within a certain range. The influence of the independent variables on XO inhibitory activity of celery seed extract were determined based on the sensitivity coefficient (SC) which was calculated as follow:

$$SC = \frac{\Delta Y/Y}{\Delta X/X}$$

where $\Delta X/X$ and $\Delta Y/Y$ were the rates of change in the dependent and independent variables, respectively [7].

3 Results and discussion

3.1 Database preparation

17 different celery seed extracts were obtained by modifying three extraction conditions: ethanol concentration,

Table 2 Absorbance at 6 wavelengths and IC_{50} values of 17 samples

Sample	Absorbance (A1000)						IC_{50} (ng/ml)	Log IC_{50}
	500 nm	495 nm	410 nm	405 nm	230 nm	210 nm		
S1	12.75	13.73	312.75	330.39	360.78	128.43	78,343	4.894
S2*	28.09	30.64	351.06	371.49	471.06	153.19	100,000	5.000
S3*	17.12	18.75	339.13	366.85	1027.99	109.24	264,241	5.422
S4	14.38	16.44	447.95	474.66	376.03	152.05	106,905	5.029
S5	14.79	16.9	460.56	488.03	386.62	156.34	145,211	5.162
S6	28.97	32.07	486.21	516.21	493.45	212.07	91,201	4.960
S7	23.84	25.83	328.81	349.67	427.15	128.15	418,794	5.622
S8*	21.58	23.74	611.87	645.32	568.71	184.53	132,130	5.121
S9	30.79	32.78	480.79	507.62	493.71	128.15	232,809	5.367
S10	24.00	26.00	663.00	701.00	494.00	149.00	125,603	5.099
S11	20.27	22.3	400.34	423.65	486.49	168.24	201,372	5.304
S12*	21.43	23.21	408.93	432.14	458.93	184.82	138,676	5.142
S13	18.35	20.5	424.1	448.92	560.07	209.35	253,513	5.404
S14	17.71	18.75	319.79	338.54	396.88	136.46	418,794	5.236
S15*	18.00	19.71	401.14	424.29	474.00	173.14	159,588	5.203
S16	43.01	44.12	432.35	455.51	515.07	207.35	156,315	5.194
S17	36.36	38.18	503.64	531.82	516.36	173.64	314,775	5.498

*Test set

Table 3 Evaluation of five models: MLR, ANN, SVR, RF and PLS

Model	Training set					Test set			
	R ²	Q ²	MAD	RMSE	Accuracy (%)	Q ²	MAD	RMSE	Accuracy (%)
MLR	0.5570	0.0406	0.2038	0.2490	97.85	0.7198	0.5844	1.0075	89.03
ANN	0.7820	0.5038	0.1875	0.2033	97.04	0.0372	0.2215	0.2639	95.66
SVR	0.9347	0.9366	0.1875	0.2284	98.43	0.0000	0.1317	0.1499	97.45
RF	0.9153	0.1055	0.2109	0.2600	98.53	0.1962	0.1614	0.1976	96.82
PLS	0.9618	0.8746	0.0671	0.0761	99.29	0.9279	0.1716	0.1826	96.69

MLR multiple linear regression, ANN artificial neural network, SVR support vector regression, RF random forest, PLS partial linear regression, R² determination coefficient, Q²: leave one out correlation coefficient, MAD average absolute deviation, RMSE root mean square error

temperature and solvent:solid ratio. The half maximal inhibitory concentration (IC₅₀) value and the absorbance of 17 celery seed extracts over a range of wavelength, from 190 to 600 nm were determined using the method described in 2.3 and 2.4. The statistical properties of the input and output variables were presented in the Supplementary Material. The UV–Vis spectra of 17 celery extracts were presented in Fig. 1.

Using the SPSS 22.0 software, 12 samples were randomly selected for selecting variables and establishing model between XO inhibitory activity and absorbance of celery seed extract while the remaining 5 samples (**S2, S3, S8, S12, S15**) were selected for the test set.

Feature selection is a critical step in application of machine learning to establish model. Too much features will take a long time to train model and increase the risk of overfitting, sometime, reduce the accuracy of established model. Feature selection could overcome these problems by keeping the important variables, reducing the number of features without losing necessary information. In this study, to identify which variables (absorbances at which wavelengths) had the greatest impact on logIC₅₀, the correlation base feature selection (Cfs) algorithm was used. The results showed that, the absorbances of celery seed extracts at 6 wavelengths: 500, 495, 410, 405, 230 and 210 nm were the most important features. These six absorbances were presented in Table 2.

3.2 Model building

3.2.1 Multiple linear regression (MLR)

The model (**M1**) established from absorbances at six wavelengths and logIC₅₀ was obtained using MLR algorithm and presented as follows:

$$\text{LogIC}_{50} = 4.7945 - 0.9872 \times A_{405} + 3.6022 \times A_{230} - 4.6717 \times A_{210}$$

The results showed that **M1** model has a low coefficient of determination (R² = 0.5570) and a low leave one out

correlation coefficient (Q² = 0.0406) which indicated that this model was not stable (Table 3).

3.2.2 Artificial neural network (ANN)

In this study, backpropagation neural network with sigmoid as activation function and number of iterations = 500 was used to establish the relationship between input and output variables. The effects of three parameters (number of nodes in hidden layer, learning rate and momentum) on the quality of the model built by ANN algorithm were investigated using one factor at a time method. Among 17 obtained models, model **M2** (number of hidden neurons = 1, learning rate = 0.5, momentum = 0.3) showed the best quality with the lowest RMSE. However, the low value of predicted Q² leave one out indicated that this model might not be suitable to predict XO inhibitory activity of celery seed extract from UV–Vis spectra. The plot between Q², MAD, RMSE and model parameters was shown in Fig. 2.

3.2.3 Support vector regression (SVR)

Similar to ANN model, two parameters C (from 1 to 500) and gamma, γ (0.1–0.9) were investigated in order to construct the best quality SVR model using one-factor-at-a-time method. However, the results showed that these two parameters did not have any affected on the performance of SVR model. Therefore, model **M3** with C = 1, γ = 0.1 was selected. The high values of determination coefficient (R² = 0.9347) and leave one out correlation coefficient (Q² = 0.9366) indicated that a strong relationship existed between input and output variables. The model also showed the ability to accurately predict XO inhibitory effect of celery seed extract with small value of MAD,

RMSE and high accuracy (MAD = 0.1875, RMSE = 0.2284 and Accuracy = 98.43% on training set and MAD = 0.1317,

Fig. 2 The plot between Q^2 , MAD, RMSE and **a** number of hidden neurons, **b** learning rate, **c** momentum

RMSE=0.1499 and Accuracy=97.45% on test set, respectively). The observed and predicted values of $\log IC_{50}$ were plotted in Fig. 5.

3.2.4 Random forest (RF)

In order to identify the optimal number of the number of trees, 10 models were constructed and the change of Q^2 , MAD and RMSE value according to number of trees (from

Fig. 3 Q^2 , MAD and RMSE versus number of trees of RF **Fig. 4** Q^2 , MAD and RMSE versus number of latent variables of PLS

1 to 10) was investigated. The results were presented in Fig. 3.

As can be seen from the graph, among 10 models, the **M4** model corresponding to the number of trees of 100 gives the best quality with the RMSE reaching the lowest point of 0.2600.

3.2.5 Partial linear regression

Unlike other algorithms such as MLR, ANN, SVR or RF, high dimension data is not a problem for PLS. Therefore, from 83 initial independent variables, a number of latent variables were created without any other feature selection process. In this study, 9 models were established with 9 different numbers of new parameters. The results were presented in Fig. 4.

As can be seen from the graph, using 3 or more variables gave an acceptable results with the high *leave one out* Q^2 and low MAD, RMSE. However, when the number of parameters was equal or greater than 5, overfitting turn out to take place with the very low predicted Q^2_{LOO} of test

set. Therefore, the PLS model with 4 latent variables was chosen for prediction of xanthine oxidase inhibitory of celery seed extract from UV–Vis spectra (**M5** model).

3.2.6 Selection of the optimal model

In order to predict XO inhibitory activity of celery seed extract, five models: **M1**, **M2**, **M3**, **M4** and **M5** were built using MLR, ANN, SVR, RF and PLS algorithm, respectively. The statistical parameters of these 4 models are shown in Table 3 and the scatter plot between observed and predicted values of $\log IC_{50}$ were presented in Fig. 5. The results showed that the **M5** model constructed by PLS method had the best quality and predictive ability with high correlation coefficient ($R^2=0.9618$, $Q^2=0.8746$), small error (MAD=0.0671; RMSE=0.0761, Accuracy=99.29%). SVR model (**M3**) was also performed well with the high value of *leave one out* correlation coefficient ($Q^2=0.9366$) suggested that this model was also able to predict biological activity of celery seed extract with high accuracy. It is appeared that ANN, MLR and RF was not good choices

Fig. 5 The scatter plots of observed and predicted values

in this case. Therefore, PLS model was chosen for further analysis. Our results were also consistent with other research in which the relationship between spectrum data and biological activity could be established well using partial linear regression [12, 13, 18]. These results also allow to affirm that there was a quantitative relationship between the UV–Vis spectra and the XO inhibitory effect of celery seed extract. However, because of the small sample size,

more investigations are required in order to apply this model to predict the XO inhibitory effect of celery extract from UV–Vis absorbances at 6 different wavelengths.

3.2.7 Y randomization

To confirm that the established PLS model was not obtained by chance, the Y-randomization test was

Fig. 6 Application domain of PLS model **Fig. 7** Sensitivity coefficient of 16 input variables

applied. The dependent variable was random shuffles ten times and 10 new PLS models were obtained. The results showed that the average of Q^2 values of 10 models was 0.1976 which was lower than the Q^2 value of initial model (0.8746), therefore, proved the robustness of PLS model.

3.3 Application domain

The Williams plot of the PLS model was illustrated in Fig. 6. The results showed that all 17 extracts (12 in the training set and 5 in the test set) had the h_i values that were lower than the warning leverage $h^* = 1.25$ and the standardized residual was less than 2.0. On the other words, there was

no outlier in dataset and the biological activities predicted by the establish model were reliable.

3.4 Sensitivity analysis

Using PLS transformation, it was identified that absorbances at 16 wavelengths: 210, 230, 235, 240, 245, 250, 255, 320, 325, 330, 335, 340, 400, 405, 410, 415 nm might play an important role in the biological activity of celery seed extract. Sensitivity analysis was performed on all these 16 variables in order to explain the influence of input parameters on the output. Each variable was increased by 10%, 20%, 30%, 40% and 50% then the average sensitivity coefficients (SC) of all independent variables were determined.

The results were showed in Fig. 7. Among 16 parameters, the SC values of the absorbance at 230 nm was the highest. Therefore, the influence of this variable on the XO inhibitory activity of celery seed extract should be taken account.

Our study has been the first research so far that establish a quantitative relationship between UV–Vis spectra and the XO inhibitory effect of celery seed extract. Compared to other chromatographic methods (HPLC, GC...), UV–Vis spectroscopy had some advantages such as simple, easy to implement and low-cost. However, one of the limitations of this method is that UV–Vis spectrum only indirectly reflected chemical components, so it is hard to identify exactly which compound(s) was responsible for the biological effect of the extract. Despite that, UV–Vis spectroscopy was still an efficient method used in quality control of herbal medicine.

4 Conclusion

In this study, the xanthine oxidase inhibitory effect of celery extract was predicted from UV–Vis absorbances at 6 different wavelengths using 5 methods: multiple linear regression (MLR), artificial neural network (ANN), support vector regression (SVR), random forest (RF) and partial linear regression (PLS). The best model was obtained by PLS algorithm with the determination coefficient $R^2 = 0.9618$, leave one out correlation coefficient $Q^2 = 0.8746$, $MAD = 0.0671$, $RMSE = 0.0716$ and the accuracy on training set and test set were 99.29% and 96.69%, respectively. Our results suggested that the PLS model from UV–Vis spectra could be an effective method for rapid determination of xanthine oxidase inhibitory activity of herbal medicine, particularly celery seed extract.

Acknowledgment We are grateful to Ms. Nong Thi Bich Van and Ms. Nguyen Kieu Anh for their diligent proofreading of this manuscript.

Author contributions Concept: H.NT, Design: H.NT, P.NV, Supervision: H.NT, Data Collection and Processing: H.LV, K.NM, T.LT, Analysis and Interpretation: H.NT, P.NV, H.LV., Literature Search: K.NM, T.LT, Writing: P.NV., Critical Reviews: H.NT, P.NV, H.LV, K.NM, T.LT.

Compliance with ethical standards

Conflict of interest The authors declare that there is no conflicts of interest.

References

- Adeniji SE, Uba S, Uzairu A, Arthur DE (2019) A derived QSAR model for predicting some compounds as potent antagonist against *Mycobacterium tuberculosis*: a theoretical approach. *Adv Prev Med* 2019:1–18. <https://doi.org/10.1155/2019/5173786>
- Alizamir M, Kim S, Kisi O, Zounemat-Kermani M (2020) A comparative study of several machine learning based non-linear regression methods in estimating solar radiation: Case studies of the USA and Turkey regions. *Energy*, p 117239
- Calixto JB (2000) Efficacy, safety, quality control, marketing and regulatory guidelines for herbal medicines (phytotherapeutic agents). *Brazilian J Med Biol Res* 33:179–189. <https://doi.org/10.1590/S0100-879X2000000200004>
- Chang W-S, Lee Y-J, Lu F-J, Chiang H-C (1993) Inhibitory effects of flavonoids on xanthine oxidase. *Anticancer Res* 13:2165–2170
- Cortes C, Vapnik V (1995) Support-vector networks. *Mach Learn* 20:273–297. <https://doi.org/10.1007/BF00994018>
- Dolati K, Rakhshandeh H, Golestani M, Forouzanfar F, Sadeghnia R, Sadeghnia HR (2018) Inhibitory effects of apium graveolens on xanthine oxidase activity and serum uric acid levels in hyperuricemic mice. *Prev Nutr Food Sci* 23:127–133. <https://doi.org/10.3746/pnf.2018.23.2.127>
- Du M, Zhang D, Hou Y, Zhao X, Li Y (2019) Combined 2D-QSAR, Principal Component Analysis and Sensitivity Analysis Studies on Fluoroquinolones' Genotoxicity. *Int J Environ Res Public Health* 16:4156
- Huerta M, Leiva V, Liu S, Rodríguez M, Villegas D (2019) On a partial least squares regression model for asymmetric data with a chemical application in mining. *Chemom Intell Lab Syst* 190:55–68
- Jiang J-L, Zhang H, Zhou P-P, Han S-N, Han Y-D, Yuan Y-J (2013) Composition–activity relationship modeling to predict the anti-tumor activity for quality control of curcuminoids from *Curcuma longa* L. (turmeric). *Anal Methods* 5:641–647
- Khademi F, Akbari M, Jamal SM, Nikoo M (2017) Multiple linear regression, artificial neural network, and fuzzy logic prediction of 28 days compressive strength of concrete. *Front Struct Civ Eng* 11:90–99
- Liao W, Zhang JM, Fu S, Wang JS, Fu CM (2013) Chromatography-activity relation of stagnation of vital energy and blood stasis syndrome influencing hemorheology by stir-baked *Curcuma* Rhizoma with vinegar. *Chin Tradit Pat Med* 35:330–334
- Lucio-Gutierrez JR, Garza-Juárez A, Coello J, Maspoch S, Salazar-Cavazos ML, Salazar-Aranda R, de Torres NW (2012) Multi-wavelength high-performance liquid chromatographic fingerprints and chemometrics to predict the antioxidant activity of *Turnera diffusa* as part of its quality control. *J Chromatogr* 1235:68–76
- Luo P, Liu Y, Lv L-Y, Zhang Z-F (2013) Spectrum-effect correlation analysis of traditional Tibetan medicine "Morina nepalensis" on nitric oxide production inhibition. *Zhongguo Zhong yao za zhi= Zhongguo zhongyao zazhi= China. J Chin Mater medica* 38:2882–2885
- Nagao A, Seki M, Kobayashi H (1999) Inhibition of xanthine oxidase by flavonoids. *Biosci Biotechnol Biochem* 63:1787–1790. <https://doi.org/10.1271/bbb.63.1787>
- Nguyen MTT, Awale S, Tezuka Y, Tran QL, Watanabe H, Kadota S (2004) Xanthine oxidase inhibitory activity of vietnamese medicinal plants. *Biol Pharm Bull* 27:1414–1421. <https://doi.org/10.1248/bpb.27.1414>
- Pacher P, Nivorozhkin A, Szabó C (2006) Therapeutic effects of xanthine oxidase inhibitors: renaissance half a century after the discovery of allopurinol. *Pharmacol Rev* 58:87–114. <https://doi.org/10.1124/pr.58.1.6>
- Park DC, El-Sharkawi MA, Marks RJ, Atlas LE, Damborg MJ (1991) Electric load forecasting using an artificial neural network. *IEEE Trans Power Syst* 6:442–449. <https://doi.org/10.1109/59.76685>
- Peng A, Lin L, Zhao M (2020) Screening of key flavonoids and monoterpenoids for xanthine oxidase inhibitory activity-oriented quality control of *Chrysanthemum morifolium* Ramat.

- 'Boju' based on spectrum-effect relationship coupled with UPLC-TOF-MS and HS-SPME-GC/MS. *Food Res Int* 137:109448
19. Powanda MC, Whitehouse MW, Rainsford KD (2015) Celery seed and related extracts with antiarthritic, antiulcer, and antimicrobial activities. In: *Novel natural products: therapeutic effects in pain, arthritis and gastro-intestinal diseases*. pp. 133–153. https://doi.org/10.1007/978-3-0348-0927-6_4
 20. Probst P, Wright MN, Boulesteix A (2019) Hyperparameters and tuning strategies for random forest. *Wiley interdiscip. Rev Data Min Knowl Discov* 9:e1301
 21. Qin L, Zhang X, Chen Y, Mo L, Zeng H, Liang Y (2017) Predictive QSAR models for the toxicity of disinfection byproducts. *Molecules* 22:1671
 22. Schonlau M, Zou RY (2020) The random forest algorithm for statistical learning. *Stata J* 20:3–29
 23. Sharifzadeh M, Sikinioti-Lock A, Shah N (2019) Machine-learning methods for integrated renewable power generation: A comparative study of artificial neural networks, support vector regression, and Gaussian Process Regression. *Renew Sustain Energy Rev* 108:513–538
 24. Song X-Y, Li Y-D, Shi Y-P, Jin L, Chen J (2013) Quality control of traditional Chinese medicines: a review. *Chin J Nat Med* 11:596–607. [https://doi.org/10.1016/S1875-5364\(13\)60069-2](https://doi.org/10.1016/S1875-5364(13)60069-2)
 25. Wang HH, Liu HG, Huang HX, Liu LM, Shen QR, Cao LJ (2011) Studies on chromatogram-effect relation of *Zanthoxylum nitidum* (Roxb.) DC. on tumor cell line HeLa. *Chin Acad J Elect Pub House* 27:84–89
 26. Xu G-L, Xie M, Yang X-Y, Song Y, Yan C, Yang Y, Zhang X, Liu Z-Z, Tian Y-X, Wang Y, Jiang R, Liu W-R, Wang X-H, She G-M (2014) Spectrum-effect relationships as a systematic approach to traditional chinese medicine research: current status and future perspectives. *Molecules* 19:17897–17925. <https://doi.org/10.3390/molecules191117897>
 27. Xue T (2003) Studying Traditional Chinese Medicine. *Science* 300:740–741. <https://doi.org/10.1126/science.300.5620.740>
 28. Yan S-K, Lin Z-Y, Dai W-X, Shi Q-R, Liu X-H, Jin H-Z, Zhang W-D (2010) Chemometrics-based approach to modeling quantitative composition-activity relationships for *Radix Tinosporae*. *Interdiscip Sci Comput Life Sci* 2:221–227
 29. Zhang L, Guan J, Liu L (2010) Preliminary exploration of antitumor profile-effect of fresh rejuvenated fruits of *Juglans mandshurica*. *Prog Mod Biomed* 10:751–752

Publisher's Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.