DOI 10.1007/s10812-024-01713-z
Journal of Applied Spectroscopy, Vol. 91, No. 1, March, 2024 (Russian Original Vol. 91, No. 1, January–February, 2024)

COMPARATIVE STUDY ON CALIBRATION MODELS USING NIR SPECTROSCOPY DATA

Ning Pan,* Zhixin Yu, Wei Ling, Jie Xu, and Yumei Liao

UDC 543.42

The quality of pork is largely influenced by moisture, fat, and protein. In the meat industry, the establishment of a fast and accurate prediction system is always welcomed. Near infrared spectroscopy (NIRS) can satisfy the requirements of the evaluation. An automatic routine based on support vector regression (SVR), a backpropagation neural network (BPNN), and principal component analysis–backpropagation neural network (PCA–BPNN) was developed to predict three components of pork using 16 combinations of pretreatment (convolution function-based moving average, detrending based on the standard normal variate, and multiplicative scatter correction). Model comparisons were implemented to evaluate the influence of pretreatment and calibration models on the prediction ability of models. The correction method and smoothing methods can significantly reduce the model prediction error. Most of the SVR models have high prediction accuracy and are suitable for predicting moisture and protein. The BPNN and PCA–BPNN are more suitable for dealing with nonlinearity between fat and NIR observations.

Keywords: *fatty acids, near-infrared spectroscopy, support vector regression, back-propagation neural network, principal component analysis.*

Introduction. Meat factories can develop a feeding program based on meat samples after slaughter to produce various meat products [1]. A meat's fat, oil, and protein are impacted by the feeding procedure. This impact, along with other factors, contribute to the production of several types of pigs [2]. A low-cost, real-time monitoring, and control system, near-infrared spectroscopy (NIR), has been utilized for the rapid evaluation of a large amount of pig production data and parameters. Meat producers find accurate prediction models attractive because they can control the quality of products by changing the feeding program according to the output of the NIR model.

Nevertheless, the performance of NIR analysis is susceptible to uncontrolled factors, such as changes in instruments, the environment, and sample preparation protocols. These variations are essential in model development. Therefore, chemometrics procedures are widely implemented, and pretreatment techniques and related algorithms are carefully selected to reduce the errors introduced by uncontrolled variations.

In the development of a calibration model, the use of various types of data pretreatment is a regular practice to assist in developing the best models with minimal residuals. Some transformation operations can significantly reduce the variations caused by unknown factors. Therefore, it is crucial to determine whether pretreatment can correct model errors and to obtain a clear idea of the degree and extent to which pretreatment plays a role in prediction variation.

The calibration model was shown to be essential in addition to pretreatment to help to create the best models with minimal residuals. Different models were implemented based on research purposes. Certain machine-learning algorithms, including the backpropagation neural network (BPNN) and support vector regression (SVR), can quickly and accurately accomplish the goal of research on pattern recognition [3]. Recently, these machine-learning tools were also proven to have excellent performance in regression analysis. Neural networks can be used for prediction because of their nonlinear mapping capability. However, their learning algorithms lack theoretical support. Therefore, in academic research, there is no universally accepted calculation method that can determine the number of neurons in a neural layer. Support vector machines, proposed by Chapelle et al. [4], can solve problems that neural networks cannot overcome, such as small sample sizes and high-dimensionality issues. Nevertheless, spectral data usually include hundreds of variables in all wavelength ranges. Therefore, many scholars have suggested that some dimensionality reduction methods, such as principal component analysis (PCA) and partial least squares (PLS), can be leveraged in advance to transform mutually dependent NIR spectra

School of Mathematics and Big Data, Guizhou Education University, China; email: ning.cecil.pan@outlook.com..
Abstract of article is published in Zhurnal Prikladnoi Spektroskopii, Vol. 91, No. 1, p. 172, January–February, 2024.

into several factors [5]. Therefore, evaluating the comparison results of various models with different pretreatments is an important issue to be studied.

Many previous works have evaluated the prediction ability of models by using some special statistics. However, it is not appropriate to compare the prediction results of various preprocessing methods by simply using statistical values as the degree of improvement brought by calibration models and pretreatment is still unknown. A selected model does not necessarily retain its superiority when the data change. Recently, the bias between models has been compared according to the method proposed by Roggo et al. [6]. This model outputs the confidence interval of the standard deviation of the model with the slightest prediction error. All other models with prediction errors within this range are not significantly different from this model.

We attempt to develop a rapid and accurate system that can predict moisture, fat, and protein by analyzing the effect of pretreatment methods and calibration models on reducing unpredictable errors in spectral data.

Calculation. The Tecator data used in this study were obtained from open sources [7]. NIR spectra were recorded for 215 meat samples with a fat content of 0.9–49% using a Tecator Infratec spectrometer. The Soxhlet method is used as a laboratory reference for fat determination. The spectra ranged from 850 to 1050 nm with an interval of 2 nm (100 wavelengths). The fat, moisture, and protein of ground pork were provided (three response variables). These samples were previously divided into two parts by Borggaard and Thodberg [8], i.e., a training set with 172 data points (N1 = 172) and a test set with 43 data points (N2 = 43). Reflectance data were stored as the logarithm of the reciprocal of reflectance log (1/*R*).

All spectra were processed, and all calibration equations and validation results were obtained using *R* software 4.0.5. The statistical tests for prediction error comparisons were performed with Microsoft Excel 2019.

The pretreatment code in *R* software can be summarized by a three-digit notation *a b c*, where *a* refers to the derivative order, *b* is the filter order, and *c* indicates the filter length (*c* must be odd) [9]. Three calibration models were applied to datasets with a total of 16 combinations of pretreatments (Table 1). The BPNN model is used in Formulas 17–32, and the principal component analysis–backpropagation neural network (PCA–BPNN) model is used in Formulas 33–48.

The pretreatment methods initially utilized were standard normal variate (SNV), which was mainly used to eliminate the effects of the solid particle size, surface scattering, and optical path changes on NIR diffuse reflectance spectra; standard normal variate and detrending (SDE), which was used to eliminate baseline drift in diffuse reflectance spectra [10]; multiplicative scatter correction (MSC), which was used to achieve better fitting results and improve the prediction results [11]; and the Savitzky–Golay filter (SG filter), which was used to reduce noise [12]. Some studies have shown that order is critical when joint pretreatment is carried out. SG filtering after SNV can eliminate noise and interference and make the model fit the window better. For MSC, the smoothing function appears first [13].

Different calibration techniques, PCA, BPNN, and SVR, were used in this study. One of the most well-known tools for assessing spectral data is the BPNN, a feedforward neural network that propagates backward based on the model error. This method is a calibration technique that provides accurate prediction results [14]. The approximate function is obtained through a constant iteration and correction process. Then, the explanatory variables are closely associated with the response variables. Eventually, the input data can be leveraged to predict the output variables accurately.

A four-layer BPNN consisting of an input layer, two hidden layers, and an output layer was applied. These layers are mutually connected by nodes, which are associated with the activation functions that transform the linearity between the output and input in an artificial neural network (ANN) into a nonlinear relationship [15]. The sigmoid and linear functions were applied in the hidden and output layers respectively [16]. After calibration, the structure of the neural network is two hidden layers with six nodes each. The learning rate was set to 0.2, and the sum of squares of errors was used as the error function. The resilient backpropagation algorithm with and without weight backtracking (rprop) was used to calculate the neural network. All data were standardized before neural network analysis [17].

Two drawbacks of applying BPNN calibration to the spectral data were found: the long calibration time and overfitting [18, 19]. Applying PCA in advance of the BPNN would be an effective way of transforming a significant number of mutually correlated variables into several independent components as input into the network [19]. In this study, a comparison was carried out between the BPNN and PCA–BPNN. A straightforward method was leveraged to select the number of factors, where PCs were selected when the cumulative interpretation variance of the model was greater than 90%. In addition, the Kaiser–Meyer–Olkin measure of sampling adequacy and Bartlett's test of sphericity were completed before PCA to provide some reliability in performing PCA on the spectral data [20].

As an alternative to ANN methods, support vector machines (SVMs) are commonly leveraged to solve data classification and spectral regression tasks [18]. The SVM was first used as a robust classification technique to solve

TABLE 1. Combinations of Pretreatment Used in Calibration in the SVR Models

Formula	Calibration model	Scatter correction	Derivative	Formula	Calibration model	Scatter correction	Derivative
1	SVR	None	0 2 0	25	BPNN	SNV	0 2 0
2		None	1 2 3	26		SNV	1 2 3
3		None	2 2 3	27		SNV	2 2 3
4		None	2 2 5	28		SNV	2 2 5
5		MSC	0 2 0	29		SDE	0 2 0
6		MSC	1 2 3	30		SDE	1 2 3
7		MSC	2 2 3	31		SDE	2 2 3
8		MSC	2 2 5	32		SDE	2 2 5
9		SNV	0 2 0	33	PCA–BPNN	None	0 2 0
10		SNV	1 2 3	34		None	1 2 3
11		SNV	2 2 3	35		None	2 2 3
12		SNV	2 2 5	36		None	2 2 5
13		SDE	0 2 0	37		MSC	0 2 0
14		SDE	1 2 3	38		MSC	1 2 3
15		SDE	2 2 3	39		MSC	2 2 3
16		SDE	2 2 5	40		MSC	2 2 5
17	BPNN	None	0 2 0	41		SNV	0 2 0
18		None	1 2 3	42		SNV	1 2 3
19		None	2 2 3	43		SNV	2 2 3
20		None	2 2 5	44		SNV	2 2 5
21		MSC	0 2 0	45		SDE	0 2 0
22		MSC	1 2 3	46		SDE	1 2 3
23		MSC	2 2 3	47		SDE	2 2 3
24		MSC	2 2 5	48		SDE	2 2 5

real-world problems in many fields. Then, with the growing demand for regression forecasting in various industries, the application of SVM was expanded into support vector regression (SVR). SVR selects more effective support vectors from the training data and obtains target data predictions through regression analysis of the raw data. From a theoretical basis, SVR can perform better than the BPNN in solving the problem of high-dimensional data [21].

In SVR, a kernel function is a helpful feature mapping technique that converts the input data into the required data format. Typically, there are four types of kernels in SVM: linear, polynomial, radial basis function (RBF), and sigmoid. A linear kernel is a special form of RBF kernel, and there are similarities between RBF and sigmoid Gaussian kernels [22]. Therefore, the RBF kernel was used in this study as it can address most problems related to SVR. The leave-one-out cross-validation method was employed to calculate the fitting error [23]. In addition, a grid search was implemented as a parameter search method to find the best values of gamma and cost. Gamma is the parameter in the RBF kernel function that determines the distribution of data mapped to a new feature space and is related to the number of support vectors. Cost defines the contribution of the sample weight within the SVR edge to the overall error. The ranges of parameter values are first set as 1 to 10 and 0.1 to 1 and are divided into grids. Then, SVR models with different parameter values are used for verification until the best results are achieved.

The performance of the NIRS model is evaluated by comparing the results of the statistical values. The validation statistics are the root mean square error of prediction (RMSEP), coefficient of determination R^2, and residual prediction deviation (RPD). Other statistics used to show the calibration results are the standard error of cross validation (SECV), coefficient of determination of calibration R_c^2, and residual prediction deviation of calibration (RPD_{cv}). The equations for RMSEP and SECV are the same. The RPD is defined as the standard deviation of the observed values divided by the RMSEP. The *R*-squared value represents the degree of interpretation of the regression model to the data. The bias between models was compared according to the method proposed by Roggo et al. [6]. This method is based on the Fisher test and defines the confidence interval for errors that are not significantly different from the minimum error. The confidence limit of the standard deviation of the minimum prediction error ($ERROR_{min}$) was calculated:

$$ERROR_{min}, \quad ERROR_{min} \sqrt{F_{1-\alpha,n-1,n-1}} , \tag{1}$$

where α is the significance level (5%); n–1is the degree of freedom; $\sqrt{F_{1-\alpha,n-1,n-1}}$ is read in Fisher's table.

Results and Discussion. An SG filter was used to smooth the data. After comparing several figures, we decided that the smoothing window length should be set to 3 when the derivative order is 1, and the smoothing window length should be set to 3 or 5 when the derivative order is 2. Eventually, SVR models with different pretreatment techniques were selected for calibration, as shown in Table 1.

For most of the model results, the MSC pretreatment is beneficial. The error can be greatly reduced by combining MSC and SVR. However, the prediction error worsens after SNV and SDE processing, with a decrease in the *R*-squared value. Derivative 1 2 3 is a better pretreatment than derivative 2 2 3 and has smaller prediction errors for moisture and fat.

All SVR models achieve an *R* squared value of 99.9%, indicating that these models are sufficient to explain the information in the data. The SVR model with pretreatment MSC+1 2 3 can lead to minimum prediction errors for moisture among the three components, with SECV = 0.087 and RPD_{cv} = 10.8.

From the fitting results of most models, the BP model outperforms the SVR model with the same preprocessing method. For moisture calibration, the preprocessing of MSC+1 2 3 leads to the minimum prediction error SECV = 0.080 for all models, with RPD_{cv} = 12.5. For validation, the preprocessing of MSC+1 2 3 leads to the minimum error RMSEP = 0.187, with RPD = 4.6. For fat, MSC+1 2 3 results in the minimum error SECV = 0.061 and RMSEP = 0.174. Validation datasets of protein can also be used to minimize the prediction errors using preprocessing MSC+1 2 3, with RMSEP = 0.230 and RPD = 4.3.

The overall fitting performance of the PCA–BPNN is not as good as that of the BPNN, and even the *R*-squared value decreases when adding pretreatment SDE or SNV. For the validation of fat, the MSC+2 2 3 pretreatment achieved the minimum value of prediction error, i.e., RMSEP = 0.099, with RPD = 10.0.

The model predicts fat better than moisture and protein, possibly because of significant variance within the fat data [24]. Figure 1 shows the comparison of prediction errors of fat calibration and validation. The calibration accuracy of the BPNN model is higher than that of the SVR and PCA–BPNN models. However, some of the validation results of the PCA–BPNN model are better than those of the other two models.

The range of SECV values with the calibration datasets usually differs slightly from that of the test set. In this paper, the ranges of the standard deviation of errors for the calibrations and validations of the PCA–BPNN and SVR model are basically the same. However, the range of SECV of the BPNN model is from 0.061 to 0.422 for the three components, but the range for the validation set is 0.174 to 3.760. Large RMSEP values were obtained when SDE and SNV were applied in the BPNN validation, which indicates that these two pretreatments are unsuitable for the BPNN model.

According to techniques proposed by Roggo et al. [6], the influence of pretreatment and fitting methods on the prediction system can be concluded by comparing the minimum prediction errors obtained from different equations. The conclusion is summarized in Table 2, including the maximum and minimum standard errors of validations for models and the upper confidence limit value of $ERROR_{min}$ (UL). UL is calculated as $ERROR_{min}\sqrt{F_{1-\alpha,n-1,n-1}}$ in Eq. (1). For the validation datasets, $F_{0.95,42,42}$ = 1.671. Models are regarded as having the same predictive power when the standard error of the models is between the $ERROR_{min}$ and the UL.

From the prediction errors in Table 2, the models that achieved the minimum error are mostly BPNN models. However, according to Roggo et al. theorem [6], the predictive power could be ranked as BPNN ≈ SVR > PCA–BPNN. The prediction results of both SVR and PCA–BPNN for the three components differed significantly, with SVR generally having better prediction results.

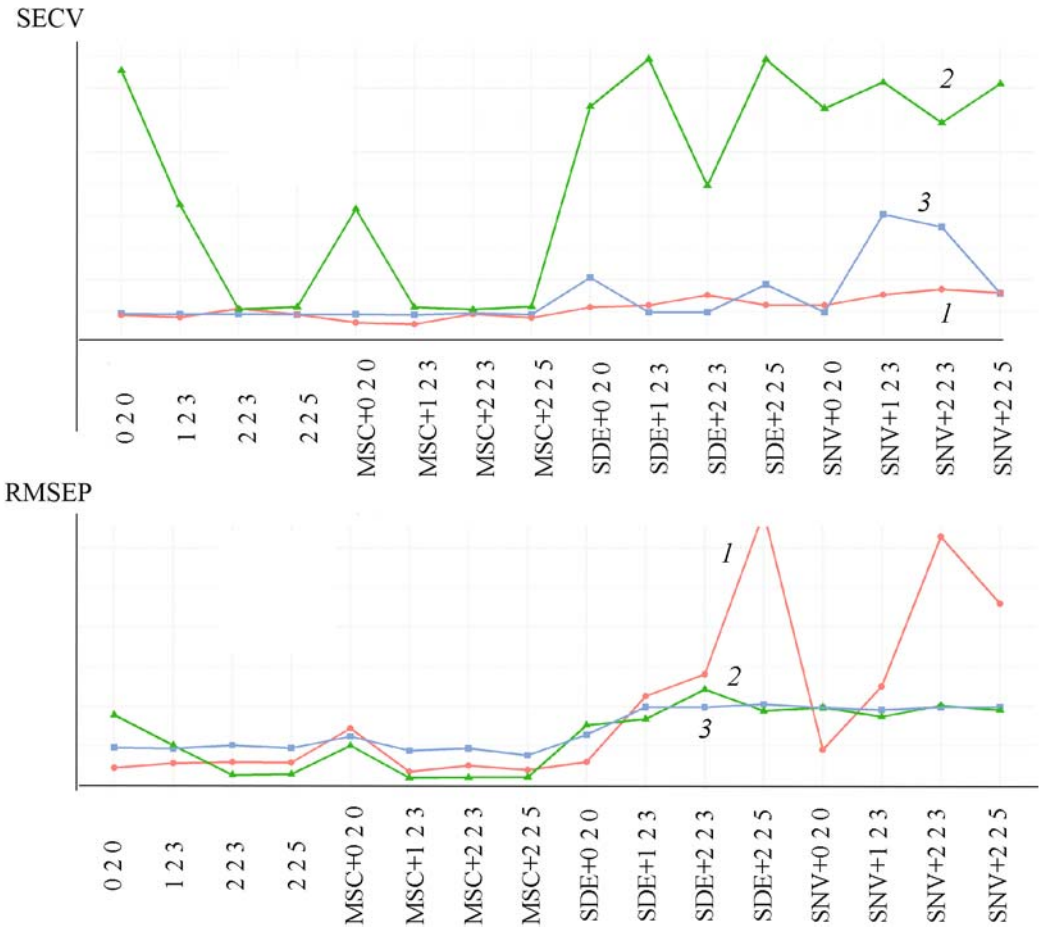

Fig. 1. Comparison of prediction errors of fat calibration and validation: 1, BPNN; 2, PCA-BPNN; 3, SVR model.

TABLE 2. Maximum and Minimum RMSEP Values for Moisture, Fat, and Protein Validations for all Models and the Upper Confidence Limit Value of $ERROR_{min}$ (UL)

Value	Moisture			Fat			Protein		
	Formula	Model	RMSEP	Formula	Model	RMSEP	Formula	Model	RMSEP
Minimum	22	BPNN	0.187	38	PCA–BPNN	0.099	22	BPNN	0.230
Maximum	32	BPNN	3.335	32	BPNN	3.447	27	BPNN	3.760
UL			0.242			0.128			0.297

Note. RMSEP: root mean square error of prediction; UL: the upper confidence limit value of $ERROR_{min}$.

For moisture, using BPNN and MSC is appropriate. The BPNN model with pretreatment MSC+1 2 3 (Formula 22) achieved a minimum prediction error of RMSEP = 0.187. No significant differences were observed in the calibration results obtained by the SVR models and the BPNN models combined with pretreatments MSC+1 2 3 and MSC+2 2 5. For validation, the smallest RMSEP was achieved for BPNN models with pretreatments MSC+1 2 3 and MSC+2 2 5 and PCA–BPNN models with derivatives 2 2 3, 2 2 5, MSC+1 2 3, MSC + 2 2 5, and MSC+2 2 5.

For the calibration of fat, the model with pretreatment MSC+1 2 3 achieved the lowest prediction error and was not different from the model with pretreatment MSC+0 2 0. For validation, no significant differences were observed

in the prediction errors of the PCA–BPNN models with combinations of pretreatments MSC+1 2 3, MSC+2 2 3, and MSC+2 2 5.

For protein, among the three components, approximately the same error results were obtained using the two fitting methods, BPNN and SVR, when the treatment method was MSC+1 2 3. The best calibration result was obtained when using the BPNN model to train the data with pretreatment MSC+2 2 5. Except for the results with the models with pretreatments SNV+1 2 3, SNV+2 2 3, SNV+2 2 5, SDE+0 2 0, and SDE+2 2 5, the results for the remaining SVR model were not significantly different from those of the best calibration result. For the validation, the BPNN models with pretreatment MSC+1 2 3 led to better prediction results than those obtained with the other models.

For the prediction of protein, SVR was more accurate than BPNN and PCA–BPNN. However, BPNN and PCA–BPNN were better at predicting fat according to the paired comparison proposed by Roggo et al. [6]. A significant advantage of the BPNN is that it can deal with nonlinear problems, whereas SVR is better at solving linear problems. Therefore, the BPNN is more suitable for dealing with the nonlinear relationship between fat and NIR data. In addition, the BPNN is more susceptible to outliers than SVR. SVR does not need to detect outliers, and penalty terms bound its performance concerning outliers. Therefore, it is not affected by extreme outliers. SVR and PCA–BPNN are more suitable for solving high-dimensional data problems. At the same time, the BPNN is prone to overfitting and may also converge to a local minimum rather than a global minimum. Thus, the prediction results are sometimes inaccurate.

The calibration and validation in this paper were slightly better than those of previous work [25]. Tecator data are well known, and several scholars have analyzed these data using partial least squares-related models. For moisture, a minimum SECV = 1.62 and an RMSEP = 1.36 were obtained. The minimum prediction errors for fat were SECV = 1.58 and RMSEP = 1.66. The minimum prediction errors for protein were SECV = 0.55 and RMSEP = 0.59. It shown that the prediction error of the models in this paper is minor. The different calibration models applied may explain the slight difference in the values. In this paper, the models presented excellent predictive power with high RPD values, which may result from a significant variance of fatty acids. When the data differed significantly, the models in this paper showed superiority in screening tests and prediction tasks with minor prediction errors.

Conclusions. As the characteristics of near-infrared spectroscopy spectra data are sharp peaks that are vulnerable to changes caused by unexpected variations, it is essential to utilize optimized pretreatment techniques before calibration is implemented. The three data preprocessing methods (SDE, MSC, and SG-filter) used in this paper were advantageous in minimizing the prediction errors. In particular, MSC can reduce model prediction errors very well. Generally, significant differences were detected in the prediction ability of the models. Owing to the relationships between the independent variable (X) and different dependent variables (Y), using SVR is suitable for solving the linear relationships and obtaining accurate predictions of protein and moisture. Using the BPNN and PCA–BPNN models is suitable for solving the nonlinear relationship between fat and near-infrared spectroscopy data.

Acknowledgments. This work was supported by the Fund of Guizhou Education Department Youth Science and Technology Talent Growth Project (QianJiaoHe-KY-Zi[2022]315, QianJiaoJi-KY-Zi[2022]263), Science Research Foundation of Guizhou Education University (2022YB008), and the National Natural Science Foundation of China (12001131).

REFERENCES

1. E. Zamora-Rojas, A. Garrido-Varo, E. De Pedro-Sanz, J. E. Guerrero-Ginel, and D. Pérez-Marín, *Food Chem.*, **129**, 1889–1897 (2011).
2. M. Bonneau and B. Lebret, *Meat Sci.*, **84**, 293–300 (2010).
3. I. Ramírez-Morales, D. Rivero, E. Fernández-Blanco, and A. Pazos, *Chemometr. Intell. Lab. Syst.*, **159**, 45–57 (2016).
4. O. Chapelle, P. Haffner, and V. N. Vapnik, *IEEE Trans. Neural Networks*, **10**, 1055–1064 (1999).
5. C. De Bleye, P. F. Chavez, J. Mantanus, R. Marini, P. Hubert, E. Rozet, and E. Ziemons, *J. Pharm. Biomed. Anal.*, **69**, 125–132 (2012).
6. Y. Roggo, L. Duponchel, B. Noe, and J. P. Huvenne, *J. Near Infrared Spectrosc.*, **10**, 137–150 (2002).
7. http://lib.stat.cmu.edu/datasets/tecator.
8. C. Borggaard and H. H. Thodberg, *Anal. Chem.*, **64**, 545–551 (1992).
9. M. B. Whitfield and M. S. Chinn, *J. Near Infrared Spectrosc.*, **25**, 363–380 (2017).
10. X. Chu, H. Yuan, and W. Lu, *Progress Chem.*, **16**, 528–542 (2004).
11. T. Naes, T. Isaksson, and B. Kowalski, *Anal. Chem.*, **62**, 664–673 (1990).

12. A. Savitzky and M. J. Golay, *Anal. Chem.*, **36**, 1627–1639 (1964).
13. S. R. Delwiche and R. A. Graybosch, *Appl. Spectrosc.*, **57**, 1517–1527 (2003).
14. Q. Liu and J. Wang, *IEEE Trans. Neural Networks*, **19**, 558–570 (2008).
15. S. Heo and J. H. Lee, *Comput. Chem. Eng.*, **127**, 1–10 (2019).
16. Y. Wang, Y. Li, Y. Song, and X. Rong, *Appl. Sci.*, **10**, 1897 (2020).
17. S. W. Lin, S. C. Chen, W. J. Wu, et al., *Knowledge and Inform. Systems*, **21**, 249–266 (2009).
18. C. H. Li and S. C. Park, *Inform. Proc. Manag.*, **45**, 329–340 (2009).
19. L. Zhang, Y. Li, Y. Gu, et al., *China Comm.*, **14**, 141–150 (2017).
20. C. Alcaraz, A. Vila-Gispert, and E. García-Berthou, *Diversity and Distributions*, **11**, 289–298 (2005).
21. R. M. Balabin and E. I. Lomakina, *Analyst*, **136**, 1703–1712 (2011).
22. H. T. Lin and C. J. Lin, *Neural Comput.*, **3**, 1–32 (2003).
23. S. Arlot and A. Celisse, *Statistics Surv.*, **4**, 40–79 (2010).
24. J. Li, S. Zhu, S. Jiang, and J. Wang, *LWT – Food Sci. Technol.*, **82**, 369–376 (2017).
25. P. Shan, S. Peng, Y. Bi, L. Tang, C. Yang, Q. Xie, and C. Li, *Chemometr. Intell. Lab. Syst.*, **138**, 72–83 (2014).