Food Anal. Methods (2015) 8:1042–1052
DOI 10.1007/s12161-014-9985-5

Apple Variety Identification Based on Dielectric Spectra and Chemometric Methods

Liang Shang · Wenchuan Guo · Stuart O. Nelson

Received: 9 July 2014 / Accepted: 27 August 2014 / Published online: 7 September 2014

Abstract The dielectric properties of 160 apples of three varieties were obtained from 10 to 1,800 MHz. Based on the Kennard-Stone algorithm, 106 apples were selected for calibration set and the remaining 54 apples were used for validation set. Principal component analysis (PCA) and successive projections algorithm (SPA) were used to extract characteristic variables from original full dielectric spectra (FS). The learning vector quantization (LVQ) network, support vector machine (SVM), and extreme learning machine (ELM) modeling algorithms were applied to build models to identify the varieties of apples. Results showed that the first three principal components, and two dielectric constants and ten loss factors were selected as characteristic variables by PCA and SPA, respectively. SPA-ELM and PCA-ELM, whose total average accuracy reached 99.5 and 99.0 %, respectively, had good potential in identifying apple varieties. The study indicates that the dielectric spectra with chemometrics are promising for identifying apple varieties nondestructively and accurately.

Keywords Apple · Dielectric properties · Learning vector quantization network · Support vector machine · Extreme learning machine

Data were obtained while Wenchuan Guo was serving as a visiting scholar at the Russell Research Center, USDA, ARS, Athens, GA, USA.

L. Shang · W. Guo (✉)
College of Mechanical and Electronic Engineering,
Northwest A&F University, Yangling, Shaanxi 712100, China
e-mail: guowenchuan69@126.com

W. Guo
e-mail: wencg915@sina.com

S. O. Nelson
Agricultural Research Service, Russell Research Center,
US Department of Agriculture, Athens, GA 30605, USA

Introduction

Apples, as a widely grown crop, have been appreciated by consumers because of their nutritional and delicious characteristics (Giovanelli et al. 2014). Apples are an important agricultural commodity in the global market of fresh products. The quality for an apple depends on its external characteristics, such as color, size, and surface texture, and internal parameters, such as sweetness, acidity, firmness, tissue texture, ascorbic acid, and polyphenolic compounds (Wojdyło et al. 2008). These characteristics, especially internal parameters, are similar within a variety. However, each variety has its special characteristics and flavor, which results in different prices and preferences by different people.

Generally, more than one apple variety is planted in an apple orchard, and several varieties are sold by sellers at one time. Therefore, different apple varieties can be easily mixed during harvesting and marketing. A means for distinguishing apple varieties is needed by apple sellers. Therefore, some reliable technique is needed to discriminate varieties of apples rapidly and nondestructively.

Dielectric properties as a useful parameter for materials have been noted increasingly by researchers (Feng et al. 2002; Guo et al. 2011b; Ndife et al. 1998; Nelson et al. 1953). The dielectric properties of usual interest in most applications are the dielectric constant ε' and loss factor ε'', the real part and imaginary part, respectively, of the relative complex permittivity, $\varepsilon* = \varepsilon' - j\varepsilon''\left(j = \sqrt{-1}\right)$. The dielectric constant indicates the ability of a material to store electric energy in the material, and the loss factor is associated with energy dissipation or conversion from electric energy to heat energy. It has been shown that the dielectric properties of agricultural products and food materials are linked to their internal features or composition, such as for grain (Nelson et al. 1953), pomegranate (Castro-Giráldez et al. 2013), honeydew melons (Guo et al. 2007b), and apples (Guo et al.

2011a). Previous studies showed that the dielectric properties had potential in predicting sweetness of apples and nectarines (Guo et al. 2013; Shang et al. 2013). If not only the main internal qualities but also the varieties can be nondestructively identified by using dielectric properties, a more efficient fruit quality and variety classification system might be developed in the future. However, to our knowledge, no research has been reported on determining varieties of fruits based on dielectric properties.

The artificial neural network (ANN), as a classification approach, has been extensively applied to establish variety identification models and has obtained good classification results by using visible/near-infrared (Vis/NIR) spectra (Bao et al. 2014; Liu et al. 2012) or hyperspectra (Chen et al. 2013). However, the precision of ANN models is usually influenced by data overlap and noise in original data or other unstable factors. To overcome these problems, chemometric methods, such as principal component analysis (PCA), the uninformative variables elimination method based on partial least squares (PLS-UVE), or the successive projection algorithm (SPA), are usually adopted to extract indispensable and useful information, usually called characteristic variables, from original data. Several studies have shown that models established by combining ANN approaches and chemometric methods together achieved better results than classical linear discriminant analysis (Cheng et al. 2014; Grunert et al. 2013).

To assess the potential of dielectric spectra in distinguishing varieties in fruits, dielectric properties of three varieties of apples ('Fuji', 'Red Rome', and 'Pink Lady') were obtained over the frequency range from 10 to 1,800 MHz. ANN analysis models, including the learning vector quantization (LVQ) network, support vector machine (SVM), and extreme learning machine (ELM), combined with chemometric approaches, such as PCA and SPA, were used to establish apple variety identification models. The study was expected to offer a new approach and useful information in applying dielectric spectra and in fruit variety classification.

Materials and Methods

Apples

Fresh apples, *Malus domestica* Borkh., of three varieties, 'Fuji', 'Pink Lady', and 'Red Rome', were obtained from refrigerated apple storage rooms in north Georgia within 2 weeks of harvest for the study. Measurements were taken initially and at 2-week intervals during 10 weeks of storage. At each measurement time, ten apples of each variety were measured. Difficulty in expressing juice from the 'Red Rome' apples resulted in suspension of the measurements on that cultivar after the sixth week of storage. Therefore, 160 apples were used in the work, including 60 'Fuji', 60 'Pink lady', and 40 'Red Rome' apples. The detailed information on apple samples was described previously (Guo et al. 2007a). The mean values and standard deviations of moisture content and firmness of pulp and the soluble solids content and pH of the juice of the three varieties of apples used in the study are listed in Table 1.

Dielectric Properties and Internal Qualities Measurement

A Hewlett-Packard (Palo Alto, CA) 4291A impedance/material analyzer, a Hewlett-Packard 85070B open-ended coaxial-line probe, a computer, and a laboratory jack constituted the measurement system for determining the dielectric properties in this study. The setup of the dielectric properties measurement system is shown in Fig. 1. Agilent Technologies 85070D dielectric probe kit software was applied to calculate the permittivities (ε' and ε'') from the reflection coefficient of the material in contact with the active tip of the coaxial-line probe. The frequency range of the 4291A impedance/material analyzer was from 1 to 1,800 MHz. Since the dielectric spectra contained much noise below 10 MHz, 10 MHz was set as the lower frequency limit in the study. The permittivity measurements were set at 51 discrete frequencies on a logarithmic scale from 10 to 1,800 MHz and were done with the probe in contact with the surface of the intact apples firmly in the equatorial region at four points about 90° apart around the perimeter of the fruit.

In addition to dielectric properties, other internal qualities, such as firmness and moisture content of pulp, soluble solids content, and pH value of juice, were obtained. Pulp firmness was measured with a Wagner (Wagner Instruments, Greenwich, CT) Fruit Test FT Series Fruit Tester equipped with an 11-mm-diameter penetrometer tip and with the Fruit Test instrument in a motor-driven penetrometer mount that advanced the tip at a constant speed (1.33 mm/s) into the fruit. Moisture content, wet basis, of pulp was determined by drying triplicate samples of about 10–14 g in disposable 57-mm aluminum weighing dishes that were placed in a forced-air drying oven for 24 h at 70 °C. Soluble solids content of juice was determined with an Atago Palette Series Model PR101α digital refractometer (Atago Co. Ltd., Tokyo, Japan). The pH values of juice were determined with a Sentron pH meter (Model 2001, Integrated Sensor Technology, Inc., Gig Harbor, WA). Averages of 3 or 4 repeated measurements for permittivities, firmness, soluble solids content, moisture content, and pH values of each sample were used.

The detailed calibration procedures for the open-ended coaxial-line probe and the impedance/material analyzer, dielectric properties, and internal qualities measurements were described previously (Guo et al. 2007a). All experiments were done at room temperature (24±1 °C).

Table 1 The mean values and standard deviations of internal qualities of the three varieties of apples used in the study

Varieties	Moisture content (% wet basis)	Firmness (kg/cm^2)	Soluble solids content (%)	pH
Fuji	80.48±1.31 a *	7.52±1.39 a	18.32±1.32 a	3.64±0.22 a
Pink Lady	83.21±0.69 b	6.69±1.13 b	15.18±0.57 b	3.41±0.12 b
Red Rome	83.83±0.88 b	4.26±0.76 c	14.57±0.84 c	3.45±0.18 b

* means within a column followed by different letters are significantly different at the 5 % probability level

Dielectric Spectra

For each sample, there were 51 values of dielectric constants and another 51 values of dielectric loss factors. From 10 to 1,800 MHz, all the values were numbered. The obtained ε' was numbered from 1 to 51 and ε'' was numbered from 52 to 102. Therefore, each sample contained 102 variables.

Sample Division Method

Rational division of the sample sets is important to improve the validation accuracy. The Kennard-Stone (KS) algorithm, for selecting representative calibration samples from all samples, is a classic method used in qualitative analysis (Galvao et al. 2005). The main process of the KS algorithm is as follows:

Step 1: Calculate the Euclidean distance between every two samples of all samples. Euclidean distance $d_x(p,q)$ is calculated by Eq. 1.

$$d_x(p,q) = \sqrt{\sum_{j=1}^{J} \left[x_p(j) - x_q(j)\right]^2}; p,q \in [1,N] \quad (1)$$

where $x_p(j)$ and $x_q(j)$ are the instrumental response at the jth variable of samples p and q, respectively. J refers to the number of all variables, and N refers to the number of all samples. In this study, J and N are 102 and 160, respectively. The samples with the largest Euclidean distance are chosen as the first and second samples in the calibration set.

Step 2: Calculate every remaining sample's Euclidean distance to the selected samples, and the minimum distance was selected. Until every remaining sample's distance is calculated, the sample with the largest minimum Euclidean distance is chosen as the next sample in the calibration set.

Step 3: Repeat step 2 until the set sample number of the calibration set is reached.

Through this algorithm, a calibration set including utmost main information from original samples is decided. Remaining samples are used as the validation set to assess established models.

In this study, 160 apples were divided into calibration and validation sets with the KS algorithm. The ratio of samples in the calibration and validation sets was 2:1. Therefore, the calibration set consisted of 40 'Fuji', 40 'Pink Lady', and 26 'Red Rome' apples, and the validation set had 20 'Fuji', 20 'Pink Lady', and 14 'Red Rome' apples.

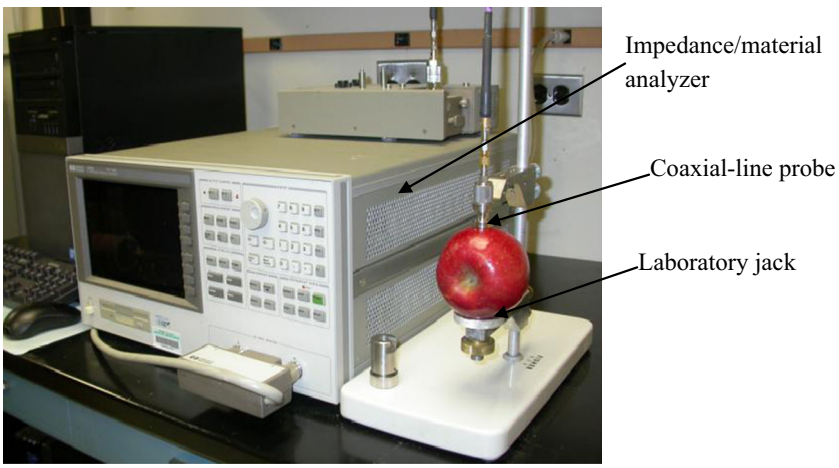

Fig. 1 The setup of the dielectric property measurement system

Characteristic Variable Selection Methods

Principal Component Analysis

PCA is a useful statistical method for data reduction (Huang et al. 2014). Based on the contribution rate applied by PCA, the number of principal components can be decided. Therefore, a reduced data set which contains the principal message of the original data can be obtained to replace the old one. It has been reported that the models established with data processed by PCA were superior to those established by original data (Li et al. 2009).

Successive Projections Algorithm

SPA is a forward-loop variable selection algorithm which can effectively solve collinearity problems by selecting variables whose information is minimally redundant (Pontes et al. 2005; Ye et al. 2008). With SPA, the first variable of the spectral variables is used as an initial variable x_0 and another new variable is incorporated in each orthogonalization iteration, until the preset variable number is reached. At the next iteration, the second variable of the spectral variables is chosen as the initial variable x_1, until all variables of the spectra are chosen as the initial one. More detailed processing of SPA can be found in other publications (Araujo et al. 2001; Pontes et al. 2005).

In a classification situation, the number of characteristic variables selected by SPA was determined by the minimum of the average risk G under different variable numbers (Pontes et al. 2011). G is calculated in the validation by Eq. 2.

$$G = \frac{1}{K_v}\sum_{k=1}^{K_v} g_k \tag{2}$$

where K_v is the number of validation samples. g_k is the risk of misclassification of the kth validation object, and it is defined as Eq. 3.

$$g_k = \frac{r^2(\boldsymbol{x}_k, \boldsymbol{\mu}_{IL})}{\min_{Ij\neq IL} r^2\left(\boldsymbol{x}_k, \boldsymbol{\mu}_{Ij}\right)} \tag{3}$$

where the numerator $r^2(\boldsymbol{x}_k, \boldsymbol{\mu}_{IL})$ is the squared Mahalanobis distance between $\boldsymbol{x}_k$ (the kth sample in class index IL) and $\boldsymbol{\mu}_{IL}$ (the mean of its true class). Both $\boldsymbol{x}_k$ and $\boldsymbol{\mu}_{IL}$ are row vectors. $\boldsymbol{\mu}_{Ij}$ is the mean of the Ij variety (j=1, 2, 3) in the calibration set. More detailed information can be found elsewhere (Pontes et al. 2011).

Modeling Methods

Learning Vector Quantization

The LVQ network is a mostly used supervised learning algorithm for classification based on a self-organizing map (Kohonen 1982; Sun et al. 2011). The main work of LVQ is to confirm the decision boundaries between neighboring categories in order to minimize misclassifications (Paola and Schowengerdt 1995). An LVQ network has three layers: an input layer, a competitive layer which learns and performs the classification, and a linear output layer (Liu et al. 2010). There exist five versions of the LVQ training algorithm, i.e., LVQ1, LVQ2.1, LVQ3, OLVQ1, and combined LVQ (CLVQ) (Vakil-Baghmisheh and Pavešić 2003). In this study, LVQ1 was applied to establish the classification models. More detailed information can be found in former literature (Bashar et al. 2005).

Support Vector Machine

As a powerful tool, the SVM algorithm has been widely applied in many analytical problems in many fields. Its theory for classification and regression has been described in detail in several papers (Brereton and Lloyd 2010; Luts et al. 2010). The main thought of SVM is to represent the original category of samples in a higher dimensional space, which is usually called feature space. A complex nonlinear mapping of the variables to a feature space can be realized by SVM (Gómez-Carracedo et al. 2012; Laurentino Alves and Poppi 2013). Based on the kernel functions, the variables are expressed in the feature space and are distinguished more easily than those in the original space. Classification problems can be described as Eq. 4:

$$f(\boldsymbol{x}) = \mathrm{sgn}\left(\sum_{s=1}^{N_s}\alpha_i y_i K(\boldsymbol{x}, \boldsymbol{x}_i) + b\right) \tag{4}$$

where N_s is the number of support vectors, $\alpha_i(0\leq\alpha_i\leq c)$ is the ith Lagrange multiplier, the constant c is defined as a penalizing factor, which determines the trade-off between error minimization and margin maximization, $\boldsymbol{x}_i$ is the ith training vector, y_i is the classification label of $\boldsymbol{x}_i$, $K(\boldsymbol{x},\boldsymbol{x}_i)$ is the kernel function, and b is the bias.

It has been proven that radial basis function (RBF) is more effective than other kernel functions (Kuo et al. 2014; Prashanth et al. 2014). RBF is defined as follows:

$$K(x, x_i) = \exp\left(-\|x-x_i\|^2/\left(2g^2\right)\right) \tag{5}$$

where $\|x-x_i\|$ is the distance from the ith input vector to threshold vector and g is the width vector (a kernel parameter of RBF).

Therefore, selecting c and g is the key step in SVM modeling. In this study, the Libsvm (version 2.81) (Chang and Lin 2011) package was adopted to establish SVM models.

Extreme Learning Machine

ELM is a single-hidden-layer feed-forward neural network (SLFN) with a perfect generalization performance. It has been widely applied in classification by using hyperspectra (Bazi et al. 2014) or in determination of internal qualities of fruits by using visible/near-infrared spectra (Jiang and Zhu 2013; Ouyang et al. 2013).

Some studies (Ouyang et al. 2013; Zhu et al. 2005) have shown that the generalization performance of ELM is better than that of the traditional learning models (such as back propagation network). Another advantage is that ELM can overcome some difficulties (such as learning rate and learning epochs) which are usually faced by classical learning models. However, selecting the number of hidden layer nodes is an essential problem to be solved in establishing ELM models. After randomly choosing and fixing the weights between input neurons and hidden neurons, an ELM model which identifies the varieties of samples can be established. More detailed theories of ELM can be found in other references (Huang et al. 2006, 2012).

Evaluation of Identification Performance

The identification performance of a classification framework is always evaluated by three different metrics: recall, precision, or accuracy (Yousef and Moghadam Charkari 2013). Accuracy has more advantages than recall or precision because it considers true positive (TP), true negative (TN), false positive (FP), and false negative (FN) simultaneously. It can be calculated as follows:

$$\text{Accuracy} = \frac{\text{number of } (\text{TPs} + \text{TNs})}{\text{number of } (\text{TPs} + \text{TNs} + \text{FPs} + \text{FNs})} \times 100\% \quad (6)$$

where TP is the event that a positive sample is classified as a positive example, TN is the event that a negative sample is classified as a negative example, FP is the event that a negative sample is classified as a positive example, and FN is the event that a positive sample is classified as a negative example. The higher the accuracy, the better the model.

Software

Besides the above-described software 85070D for dielectric spectra acquisition, SPSS 17.0 (SPSS Inc., Chicago, IL) was used to do analysis of variance (ANOVA) on the internal qualities of the three varieties of apples used, and MATLAB R2012a (MathWorks, Natick, MA) was applied to establish classification models in this study.

Results and Discussion

The Internal Properties of the Three Varieties of Apples

Table 1 shows that 'Fuji' apples had the highest firmness, soluble solids content, and pH values and lowest moisture content. 'Red Rome' had the lowest values for firmness and soluble solids content. Several studies had reported that the moisture content of fruits had a negative linear relationship with the soluble solids content (Guo et al. 2007a, 2011a; Nelson et al. 2007). Similar results were also noted here.

The results of ANOVA on the internal quality parameters of the three varieties showed that the moisture content and pH value of 'Fuji' apples had a significant difference with those of 'Pink Lady' and 'Red Rome' apples at a significance level of 5 %, but they had no significant difference for 'Pink Lady' and 'Red Rome' apples. The firmness and soluble solids content of the three varieties all had a significant difference with each other at 5 % probability level.

The Dielectric Properties of the Three Varieties of Apples

The averages of the dielectric constants and loss factors, with standard deviations, of the three varieties of apples over the frequency range from 10 to 1,800 MHz are shown in Fig. 2. The dependence of ε' and ε'' on frequency is similar for the three varieties. The ε' value decreased with increasing frequency over the whole frequency range. However, an overriding dielectric relaxation behavior was observed for ε''. The behavior may involve bound water and Maxwell-Wagner relaxations (Guo et al. 2007a).

Selection of Characteristic Variables

Selection of Characteristic Variables by PCA

Table 2 lists the contribution rate and accumulative contribution rate of the first seven PCs. It shows that the first principal component (PC1) offers the main contribution (64.84 %), and the second principal component (PC2) contributed 26.24 %. Generally, when the PCs have more than 85 % accumulative contribution of the original dataset, these PCs can be used to replace the original one (He et al. 2006). In this study, the first two PCs contributed 91.07 %, more than 85 %, of the whole contribution, but the validation capability was not good when the first two PCs were used to identify apple varieties. The

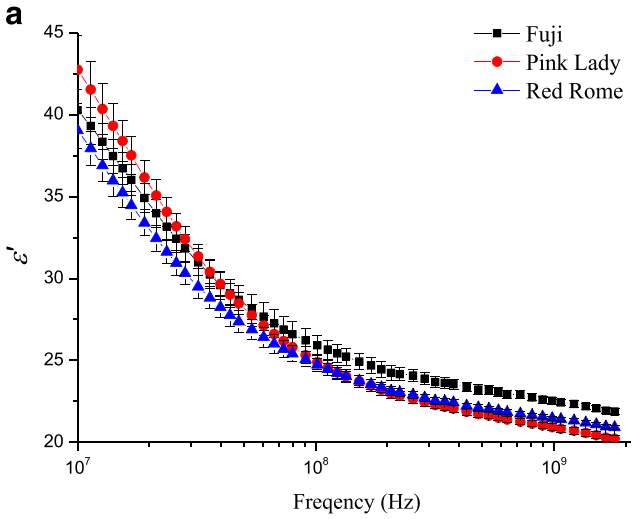

Fig. 2 Frequency dependence of ε' (**a**) and ε'' (**b**) of the three different varieties of apples

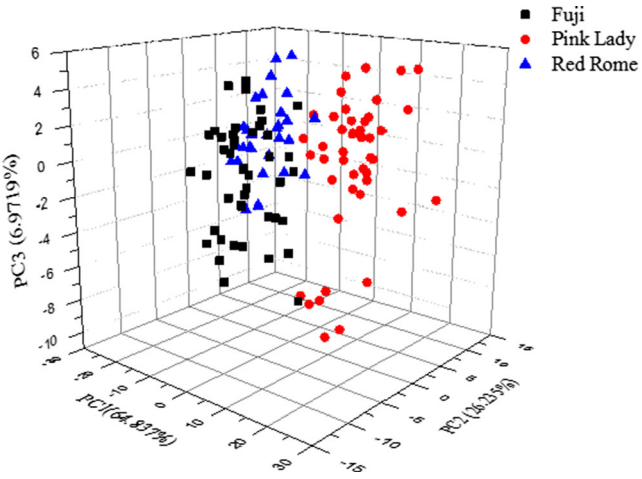

Fig. 3 Scatter plots by PC1 × PC2 × PC3 of the three varieties of apples

scores plot of PC1, PC2, and PC3 in three-dimensional space was examined for clustering results relating to the three varieties, 'Fuji', 'Pink Lady', and 'Red Rome' (Fig. 3). As shown in Fig. 3, 'Fuji' and 'Pink Lady' can be separated easily, but it is difficult to separate 'Red Rome' from 'Fuji'. This means that it is hard to separate these three varieties of apples by linear discriminant analysis. Nonlinear discriminant methods need to be developed. In the study, the first three PCs, whose accumulative rate higher than 98.04 %, were adopted and were considered as the inputs for the LVQ, SVM, and ELM models.

Selection of Characteristic Variables by SPA

The change in *G* value with the number of variables selected by SPA is shown in Fig. 4, where the *G* value decreases with the increasing number of variables. Since more variables will slow down the computation speed, usually the number of variables is determined when the *G* value has a small decrease with increasing numbers of variables. In this study, when the decrease in *G* was less than 0.4, the smallest number, 12, was chosen. The selected 12 characteristic variables, including two variables of ε' and ten variables of ε'' at different frequencies, are listed in Table 3. The amount of dielectric variables selected by SPA was 11.8 % of the 102 variables in the full dielectric spectra.

Variety Identification Models

Models Developed by LVQ

To achieve the best results for the LVQ models, all training epochs were set as 500. Furthermore, the learning rate and goal of the LVQ model were set as 0.1 and 0.05, respectively. An essential parameter of the LVQ is the number of hidden layer nodes, and it is usually selected by trial and error. The range of the number of hidden layer nodes was set from 1 to 30 at first and then increased 1 by 1 at a time, and the optimal number of nodes was confirmed based on the highest variety identification accuracy of the calibration set. The identification accuracy of the calibration set changed with the number

Table 2 The contribution rates and accumulative contribution rates of the first seven PCs

The number of principal components	PC1	PC2	PC3	PC4	PC5	PC6	PC7
Contribution rate (%)	64.84	26.24	6.97	1.31	0.25	0.21	0.05
Accumulative contribution rate (%)	64.84	91.07	98.04	99.35	99.60	99.82	99.87

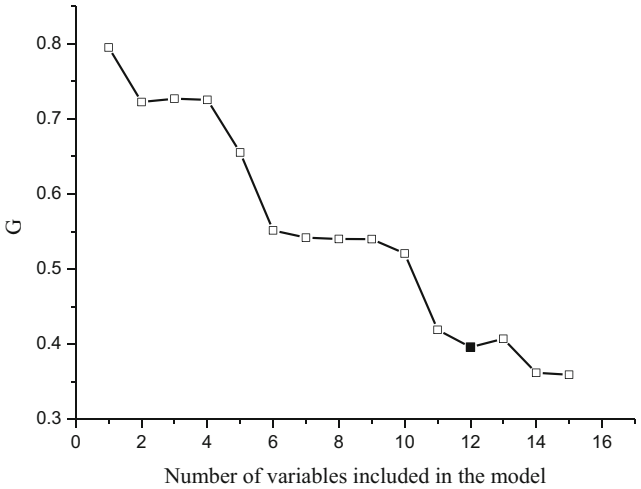

Fig. 4 *G* value dependence on the number of variables selected by SPA in the calibration set. The *solid square* represents the point at which the number of variables was finally selected by SPA

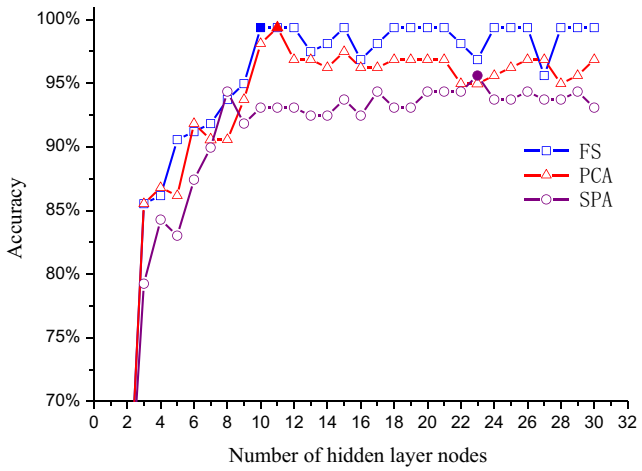

Fig. 5 The accuracies under different numbers of hidden layer nodes for LVQ networks in the calibration set. The *blue square*, *red triangle*, and *purple circle* represent the points at which the optimal numbers for FS, PCA, and SPA, respectively, were finally selected

of hidden layer nodes as shown in Fig. 5. The smallest numbers, where higher or highestaccuracies were obtained, were chosen. The numbers of hidden layer nodes for full dielectric spectra (FS)-LVQ, PCA-LVQ, and SPA-LVQ were 10, 11, and 22, respectively. The data are listed in Table 4.

The apple variety identification accuracies for the calibration and validation sets of LVQ models under different characteristic variable selection methods are given in Table 5. The results show that for the calibration set, the FS-LVQ and PCA-LVQ models had higher average accuracy (99.4 %) than SPA-LVQ (94.3 %). For the validation set, the accuracy of FS-LVQ reached 98.8 %, followed by SPA-LVQ and PCA-LVQ (93.8 %). As for total average accuracy, FS-LVQ had the highest accuracy (99.1 %), followed by PCA-LVQ (96.6 %) and SPA-LVQ (94.1 %). The accuracy was higher than 90.7 % for each apple variety. Especially, FS-LVQ had 100 % accuracy for 'Pink Lady' both in the calibration and validation sets.

Models Developed by SVM

The RBF was used as the kernel function in this study. Fivefold cross validation was applied to select *c* and *g*. When the ranges of *c* and *g* were set from 2^{-6} to 2^{6}, from 2^{-7} to 2^{7}, or from 2^{-8} to 2^{8}, the determined values of *c* and *g* were same as when the range was set from 2^{-5} to 2^{5}. Therefore, the range of 2^{-5} to 2^{5} with an increment of $2^{0.5}$ was used to select *c* and *g* for the SVM.

For each combination of *c* and *g*, the SVM model was established, and the accuracy was calculated. The optimal values of *c* and *g* were determined by the highest accuracy in all combinations of *c* and *g*. Detailed processing for parameter selection was discussed elsewhere (Chang and Lin 2011; Cheng et al. 2013). Figure 6 summarizes the process of selecting these parameters. The determined optimal *c* and *g* for FS-SVM, PCA-SVM, and SPA-SVM were 2 and 0.125, 16 and 2, and 4 and 0.345, with the apple variety identification accuracy of 100, 98.1, and 97.5 %, respectively, for the calibration set (Fig. 6).

Table 5 lists the apple variety identification accuracies for the calibration and validation sets of SVM models under different characteristic variable selection methods. The results show that for the calibration set, the average accuracy of FS-SVM was 100 %, followed by PCA-SVM (98.1 %) and SPA-SVM (97.5 %). For the validation set, both FS-SVM and SPA-SVM had 100 % average accuracy, higher than that of PCA-SVM (97.5 %). As for total average accuracy, FS-SVM (100 %) was better than SPA-SVM (98.8 %) and PCA-SVM (97.8 %). The lowest accuracy for each apple variety was 96.2 %. FS-SVM obtained 100 % identification rate for each apple variety, not only in the calibration set but also in the validation set.

Models Developed by ELM

The "sig." function was selected as the excitation function to develop the ELM models. The numbers of hidden layer nodes

Table 3 The 12 variables selected by SPA

No.	Frequency (MHz)	Dielectric properties	No.	Frequency (MHz)	Dielectric properties
1	10.000	ε'	7	188.971	ε''
2	32.099	ε'	8	379.045	ε''
3	10.000	ε''	9	460.659	ε''
4	23.675	ε''	10	585.501	ε''
5	66.887	ε''	11	723.870	ε''
6	134.164	ε''	12	1,800.000	ε''

Table 4 The optimal training parameters of different models

Pretreatment methods	LVQ			SVM		ELM		
	Input layer nodes	Hidden layer nodes	Output layer nodes	*c*	*g*	Input layer nodes	Hidden layer nodes	Output layer nodes
FS	102	10	1	2	0.125	102	38	1
PCA	3	11	1	16	2	3	35	1
SPA	12	22	1	4	0.354	12	30	1

of ELM models were also determined by the trial and error method (Chen et al. 2012). The range of number of hidden layer nodes was set from 1 to 50 at first; then, the number of hidden layer nodes was gradually increased by 1 at a time. The optimal numbers of nodes were determined based on the highest variety identification accuracy. Since the initial weight value of ELM was random, the performance of the model was unstable. To overcome the problem, a model repeated 1,000 times was employed here. According to the average accuracies, the number of hidden layer nodes was chosen. The average accuracies of the calibration set in 1,000 times repetition under different numbers of hidden layer nodes for ELM with different characteristic variable selection methods of FS, UVE-PLS, and SPA are shown in Fig. 7. Based on the highest accuracy, the numbers of hidden layer nodes for FS-SVM, PCA-SVM, and SPA-SVM were decided as 38, 35, and 30, respectively (Table 4).

The apple variety identification accuracies for the calibration set and the validation set of ELM models under different characteristic variable selection methods are also listed in Table 5. Table 5 shows that for the calibration set, the accuracy of FS-ELM reached 99.8 %, higher than that of SPA-ELM (99.3 %) and PCA-ELM (98.7 %). For the validation set, the accuracies of three ELM models were higher than 99.3 %. The highest total average accuracy was 99.8 % (FS-ELM), followed by 99.5 % (SPA-ELM) and 99.0 % (PCA-ELM). As for each apple variety, the accuracy was higher than 98.0 %.

Comparison of Variety Identification Performance for Different Models

When the three modeling methods were compared, it was found that the ELM had the best identification accuracy, followed by SVM and LVQ, since the lowest accuracy rates of ELM, SVM, and LVQ for each apple variety were 98.0 %, 96.2 % and 90.7 %, respectively. The good prediction performance of ELM was also noted in other classification work (Heras et al. 2014; Termenon et al. 2013). At each modeling method, FS did excellent work in identifying apple varieties. For example, for SVM, FS-SVM had 100 % accuracy for each apple variety both in the calibration and validation sets. For LVQ and ELM, FS also had higher average accuracy not only in the calibration set but also in the validation set. The reason is that the FS includes all useful information. When SPA and PCA were compared, it was found that SPA did a little better work in SVM and ELM models. However, PCA played a better job than SPA in LVQ models. This indicates that different variable extraction methods have different effects in different models. Therefore, it is necessary to find the most

Table 5 Apple variety identification accuracies of LVQ, SVM, and ELM models under different variable selection methods

Modeling approach	Characteristic variables selection method	Accuracy for calibration set (%)				Accuracy for validation set (%)				Total average
		Fuji	Pink Lady	Red Rome	Average	Fuji	Pink Lady	Red Rome	Average	
LVQ	FS	99.1	100	99.1	99.4	98.2	100	98.2	98.8	99.1
	PCA	99.1	100	99.1	99.4	90.7	98.2	92.6	93.8	96.6
	SPA	92.5	98.1	92.5	94.3	90.7	100	90.7	93.8	94.1
SVM	FS	100	100	100	100	100	100	100	100	100
	PCA	98.1	99.1	97.2	98.1	96.3	100	96.3	97.5	97.8
	SPA	99.1	97.2	96.2	97.5	100	100	100	100	98.8
ELM	FS	99.9	99.8	99.7	99.8	99.8	99.3	99.8	99.8	99.8
	PCA	98.0	100	98.0	98.7	99.0	100	99.0	99.3	99.0
	SPA	99.6	99.3	98.9	99.3	99.4	100	99.3	99.6	99.5

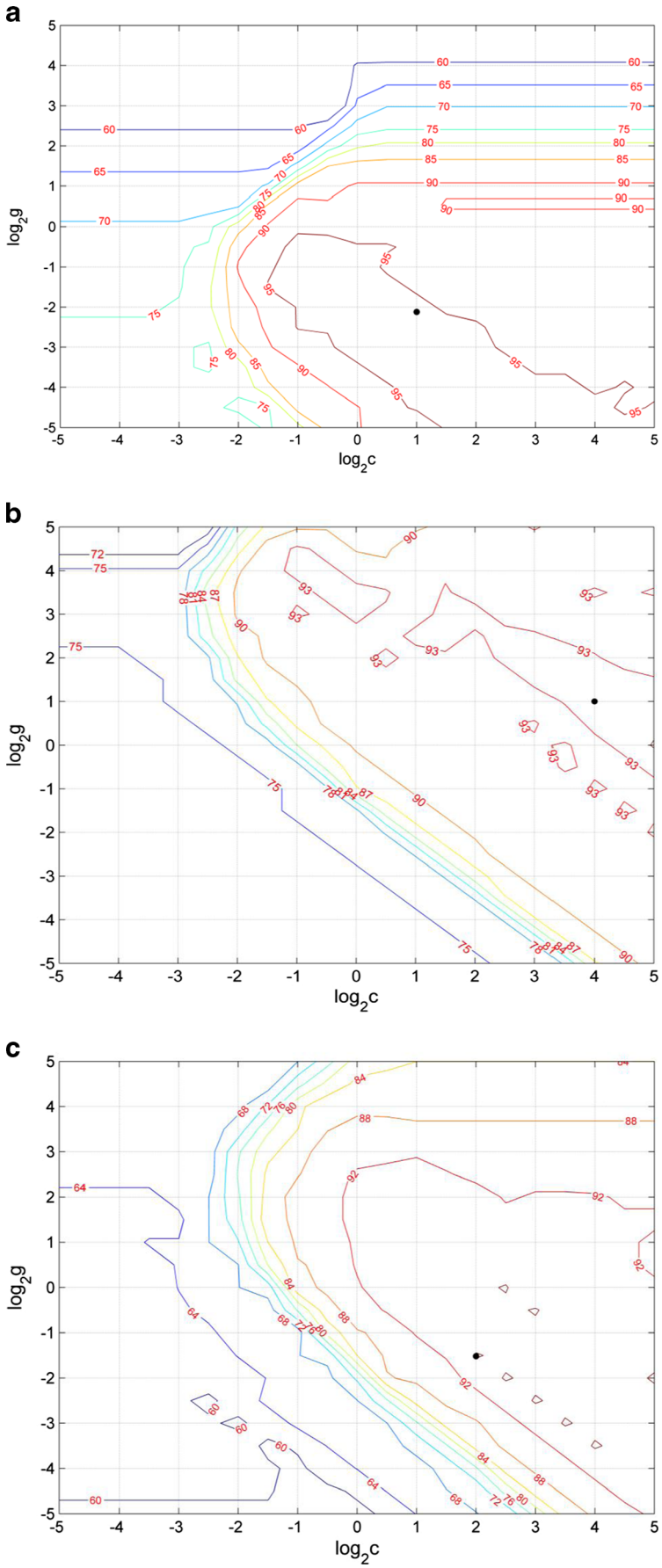

Fig. 6 The grid search process for penalty factor (c) and RBF kernel parameter (g) combination of three selection methods for SVM by fivefold cross validation. The *black circle* represents the point at which the optimal c and g were chosen

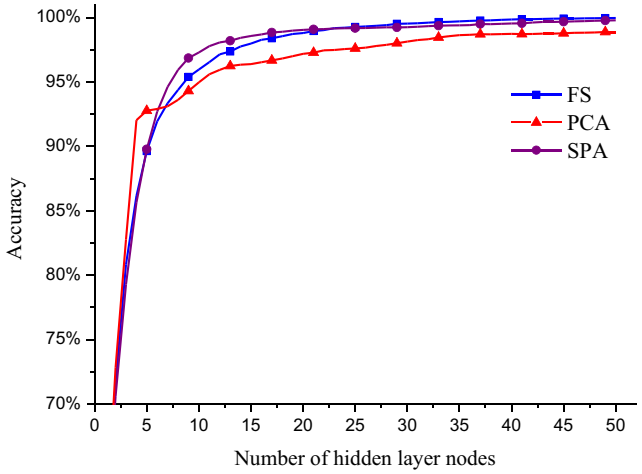

Fig. 7 The accuracies under different numbers of hidden layer nodes for ELM in the calibration set. The optimal numbers for FS, PCA, and SPA are 38, 35, and 30, respectively, which were finally chosen

suitable characteristic variable selection method for each ANN models.

The characteristic variables extracted by SPA and PCA were 12 and 3, which were only 11.8 and 2.9 % of the variables in FS. It has been reported that the training time increases linearly with the number of variables, i.e., the dimension of spectra (Chauchard et al. 2004). Speed is one of the most important factors for online detection. Although the full dielectric spectra obtained the best identification accuracy, too many variables included in FS reduce the operation speed. Therefore, it is suggested that SPA-ELM and PCA-ELM, whose total average accuracy was higher than 99.0 %, are the optimal models in identifying apple varieties based on an overall consideration of variables used as inputs for the models and variety identification accuracies for the calibration and validation sets.

Conclusions

The contribution of this work is to present a rapid and nondestructive approach for discriminating different varieties of apples. At present, there is only qualitative analysis in most of the discrimination of fruit varieties by using visible/near-infrared spectroscopy or hyperspectral spectroscopy, and no research had been devoted to fruit variety discrimination by using dielectric spectroscopy. In this research, qualitative analysis for three varieties of apples by means of combining dielectric spectra, ANN, and chemometric methods was made. The KS method was used for subset partitioning. Two chemometric methods (PCA and SPA) were adopted to extract characteristic variables from original dielectric spectra, and three ANN modeling approaches (LVQ, SVM, and ELM) were employed to establish variety determination models. By processing for PCA and SPA, 3 principal components and 12 characteristic variables were selected, respectively, as input data instead of full dielectric spectra with 102 variables. ELM had the best identification accuracy, followed by SVM and LVQ. FS performed well in determining apple varieties, but the many variables included in FS slow down the modeling speed. In ELM and SVM models, variables extracted by SPA performed better than PCA. Among the established nine models, the total average accuracy of SPA-ELM and PCA-ELM reached 99.5 and 99.0 %, respectively. SPA-ELM and PCA-ELM had good potential in identifying apple varieties quickly and efficiently. The research demonstrates that dielectric spectra associated with ANN and chemometric approaches can be adopted for fruit variety nondestructive determination.

Acknowledgments This research was supported by a grant from the National Natural Science Foundation of China (project no. 31171720).

Conflict of Interest Liang Shang declares that he has no conflict of interest. Wenchuan Guo declares that she has no conflict of interest. Stuart O. Nelson has no conflict of interest. This article does not contain any studies with human or animal subjects.

References

Araujo MCU, Saldanha TCB, Galvao RKH, Yoneyama T, Chame HC, Visani V (2001) The successive projections algorithm for variable selection in spectroscopic multicomponent analysis. Chemom Intell Lab Syst 57(2):65–73

Bao YD, Liu F, Kong WW, Sun DW, He Y, Qiu ZJ (2014) Measurement of soluble solid contents and pH of white vinegars using VIS/NIR spectroscopy and least squares support vector machine. Food Bioprocess Technol 7(1):54–61

Bashar MK, Ohnishi N, Matsumoto T, Takeuchi Y, Kudo H, Agusa K (2005) Image retrieval by pattern categorization using wavelet domain perceptual features with LVQ neural network. Pattern Recogn Lett 26(15):2315–2335

Bazi Y, Alajlan N, Melgani F, AlHichri H, Malek S, Yager RR (2014) Differential evolution extreme learning machine for the classification of hyperspectral images. IEEE Geosci Remote Sens Lett 11(6): 1066–1070

Brereton RG, Lloyd GR (2010) Support vector machines for classification and regression. Analyst 135(2):230–267

Castro-Giráldez M, Fito PJ, Ortolá MD, Balaguer N (2013) Study of pomegranate ripening by dielectric spectroscopy. Postharvest Biol Technol 86:346–353

Chang CC, Lin CJ (2011) LIBSVM: a library for support vector machines. ACM Trans Intell Syst Technol 2(3):27

Chauchard F, Cogdill R, Roussel S, Roger JM, Bellon-Maurel V (2004) Application of LS-SVM to non-linear phenomena in NIR spectroscopy: development of a robust and portable sensor for acidity prediction in grapes. Chemom Intell Lab Syst 71(2):141–150

Chen QS, Ding J, Cai JR, Zhao JW (2012) Rapid measurement of total acid content (TAC) in vinegar using near infrared spectroscopy based on efficient variables selection algorithm and nonlinear regression tools. Food Chem 135(2):590–595

Chen QS, Zhang YH, Zhao JW, Hui Z (2013) Nondestructive measurement of total volatile basic nitrogen (TVB-N) content in salted pork in jelly using a hyperspectral imaging technique combined with efficient hypercube processing algorithms. Anal Methods 5(22):6382–6388

Cheng PY, Fan WL, Xu Y (2013) Quality grade discrimination of Chinese strong aroma type liquors using mass spectrometry and multivariate analysis. Food Res Int 54(2):1753–1760

Cheng PY, Fan WL, Xu Y (2014) Determination of Chinese liquors from different geographic origins by combination of mass spectrometry and chemometric technique. Food Control 35(1):153–158

Feng H, Tang J, Cavalieri RP (2002) Dielectric properties of dehydrated apples as affected by moisture and temperature. Trans ASAE 45(1): 129–135

Galvao RKH, Araujo MCU, Jose GE, Pontes MJC, Silva EC, Saldanha TCB (2005) A method for calibration and validation subset partitioning. Talanta 67(4):736–740

Giovanelli G, Sinelli N, Beghi R, Guidetti R, Casiraghi E (2014) NIR spectroscopy for the optimization of postharvest apple management. Postharvest Biol Technol 87:13–20

Gómez-Carracedo MP, Fernández-Varela R, Ballabio D, Andrade JM (2012) Screening oil spills by mid-IR spectroscopy and supervised pattern recognition techniques. Chemom Intell Lab Syst 114:132–142

Grunert T, Wenning M, Barbagelata MS, Fricker M, Sordelli DO, Buzzola FR, Ehling-Schulz M (2013) Rapid and reliable identification of staphylococcus aureus capsular serotypes by means of artificial neural network-assisted fourier transform infrared spectroscopy. J Clin Microbiol 51(7):2261–2266

Guo W, Nelson SO, Trabelsi S, Kays SJ (2007a) 10–1800-MHz dielectric properties of fresh apples during storage. J Food Eng 83(4):562–569

Guo W, Nelson SO, Trabelsi S, Kays SJ (2007b) Dielectric properties of honeydew melons and correlation with quality. J Microw Power Electromagn Energy 41(2):44–54

Guo W, Zhu X, Nelson SO, Yue R, Liu H, Liu Y (2011a) Maturity effects on dielectric properties of apples from 10 to 4500 MHz. LWT-Food Sci Technol 44(1):224–230

Guo W, Zhu X, Yue R, Liu H, Liu Y (2011b) Dielectric properties of Fuji apples from 10 to 4500 MHz during storage. J Food Process Preserv 35(6):884–890

Guo W, Shang L, Wang M, Zhu X (2013) Soluble solids content detection of postharvest apples based on frequency spectrum of dielectric parameters. Trans Chin Soc Agric Mach 44(9):132–137 (in Chinese with English abstract)

He Y, Feng SJ, Li XL, Qiu ZJ (2006) Study on fast discrimination of varieties of acidophilous milk using near infrared spectra. Spectrosc Spect Anal, 26(11):2021–2023 (in Chinese with English abstract).

Heras DB, Arguello F, Quesada-Barriuso P (2014) Exploring ELM-based spatial-spectral classification of hyperspectral images. Int J Remote Sens 35(2):401–423

Huang GB, Zhu QY, Siew CK (2006) Extreme learning machine: theory and applications. Neurocomputing 70(1–3):489–501

Huang GB, Zhou HM, Ding XJ, Zhang R (2012) Extreme learning machine for regression and multiclass classification. IEEE Trans Syst Man Cybern B Cybern 42(2):513–529

Huang Y, Min SG, Duan J, Wu LJ, Li QQ (2014) Identification of additive components in powdered milk by NIR imaging methods. Food Chem 145:278–283

Jiang H, Zhu WX (2013) Determination of pear internal quality attributes by fourier transform near infrared (FT-NIR) spectroscopy and multivariate analysis. Food Anal Methods 6(2):569–577

Kohonen T (1982) Self-organized formation of topologically correct feature maps. Biol Cybern 43(1):59–69

Kuo BC, Ho HH, Li CH, Hung CC, Taur JS (2014) A kernel-based feature selection method for SVM with RBF kernel for hyperspectral image classification. IEEE J Sel Top Appl Earth Obs Remote Sens 7(1):317–326

Laurentino Alves JC, Poppi RJ (2013) Determining the presence of naphthenic and vegetable oils in paraffin-based lubricant oils using near infrared spectroscopy and support vector machines. Anal Methods 5(22):6457–6464

Li W, Bagnol L, Berman M, Chiarella RA, Gerber M (2009) Applications of NIR in early stage formulation development. Part II. Content uniformity evaluation of low dose tablets by principal component analysis. Int J Pharm 380(1–2):49–54

Liu J, Zuo B, Zeng X, Vroman P, Rabenasolo B (2010) Nonwoven uniformity identification using wavelet texture analysis and LVQ neural network. Expert Syst Appl 37(3):2241–2246

Liu YD, Gao RJ, Hao Y, Sun XD, Ouyang AG (2012) Improvement of near-infrared spectral calibration models for brix prediction in 'Gannan' navel oranges by a portable near-infrared device. Food Bioprocess Technol 5(3):1106–1112

Luts J, Ojeda F, Van de Plas R, De Moor B, Van Huffel S, Suykens JAK (2010) A tutorial on support vector machine-based methods for classification problems in chemometrics. Anal Chim Acta 665(2):129–145

Ndife MK, Sumnu G, Bayindirli L (1998) Dielectric properties of six different species of starch at 2450 MHz. Food Res Int 31(1):43–52

Nelson SO, Soderholm LH, Yung FD (1953) Determining the dielectric properties of grain. Agric Eng 34(9):608–610

Nelson SO, Guo W, Trabelsi S, Kays SJ (2007) Dielectric properties of watermelons for quality sensing. Meas Sci Technol 18:1887–1892

Ouyang Q, Chen QS, Zhao JW, Lin H (2013) Determination of amino acid nitrogen in soy sauce using near infrared spectroscopy combined with characteristic variables selection and extreme learning machine. Food Bioprocess Technol 6(9):2486–2493

Paola J, Schowengerdt R (1995) A review and analysis of backpropagation neural networks for classification of remotely-sensed multi-spectral imagery. Int J Remote Sens 16(16):3033–3058

Pontes MJC, Galvao RKH, Araujo MCU, Nogueira P, Moreira T, Neto ODP, Jose GE, Saldanha TCB (2005) The successive projections algorithm for spectral variable selection in classification problems. Chemom Intell Lab Syst 78(1–2):11–18

Pontes MJ, Pereira CF, Pimentel MF, Vasconcelos FV, Silva AG (2011) Screening analysis to detect adulteration in diesel/biodiesel blends using near infrared spectrometry and multivariate classification. Talanta 85(4):2159–2165

Prashanth R, Roy SD, Mandal PK, Ghosh S (2014) Automatic classification and prediction models for early Parkinson's disease diagnosis from SPECT imaging. Expert Syst Appl 41(7):3333–3342

Shang L, Gu J, Guo W (2013) Non-destructively detecting sugar content of nectarines based on dielectric properties and ANN. Trans Chin Soc Agric Eng 29(17):257–264 (in Chinese with English abstract)

Sun H, Li M, Li D (2011) The vegetation classification in coal mine overburden dump using canopy spectral reflectance. Comput Electron Agric 75(1):176–180

Termenon M, Grana M, Barros-Loscertales A, Avila C (2013) Extreme learning machines for feature selection and classification of cocaine dependent patients on structural MRI data. Neural Process Lett 38(3):375–387

Vakil-Baghmisheh M-T, Pavešić N (2003) Premature clustering phenomenon and new training algorithms for LVQ. Pattern Recogn 36(8): 1901–1912

Wojdyło A, Oszmiański J, Laskowski P (2008) Polyphenolic compounds and antioxidant activity of new and old apple varieties. J Agric Food Chem 56(15):6520–6530

Ye SF, Wang D, Min SG (2008) Successive projections algorithm combined with uninformative variable elimination for spectral variable selection. Chemom Intell Lab Syst 91(2):194–199

Yousef A, Moghadam Charkari N (2013) A novel method based on new adaptive LVQ neural network for predicting protein–protein interactions from protein sequences. J Theor Biol 336:231–239

Zhu QY, Qin AK, Suganthan PN, Huang GB (2005) Evolutionary extreme learning machine. Pattern Recogn 38(10):1759–1763