Regular Article

ANALYTICAL
AND BIOANALYTICAL
CHEMISTRY
RESEARCH

Published by the
Iranian Chemical Society

***Anal. Bioanal. Chem. Res.**, Vol. 10, No. 3, 301-317, July 2023.*

Multivariate Authentication of Herbs and Spices through UV-Vis and FT-IR Fingerprint

Maryam Abbasi Tarighat[a,*], Gholamreza Abdi[b], Farideh Heidari Ghorghosheh[a] and Kowsar Shahmohammadi Bayatiyani[a]

[a]*Faculty of Nano and Bio Science and Technology, Persian Gulf University, Bushehr, 75169, Iran*

[b]*Department of Biotechnology, Persian Gulf Research Institute, Persian Gulf University, Bushehr, Iran*

(Received 10 October 2022, Accepted 23 November 2022)

The aim of this study was to investigate the applicability of UV-Vis and FT-IR fingerprints combined with multivariate statistical tools to classify and authenticate Iranian standard herbs and spices, and their mislabeled and adulterated samples in single and fusion models. The proposed strategy is an alternative, rapid, easy, and economical approach for herbs and spices authentication. Sixty-three samples of different herbs and spices were collected across several cities in Iran. The potency of Savitzky-Golay(SG) smoothing in combination with autoscaling for improving the accuracy of clustering well studied and principal component analysis (PCA), PCA-linear discriminant analysis (PCA-LDA) and partial least squares-discriminant analysis (PLS-DA) were applied for classification. Additionally, data mining of spectral sets was performed using Kohonen self-organization maps (SOMs) of smoothed and unsmoothed individual data sets and classification results were compared. Also, the discriminant models using fusion matrix were built by concatenation of SG smoothed-autoscaled SOMs clusters of FTIR and UV-Vis (SG-autoscaled-SOMs) spectra. The results of different models showed that the accuracy of single SG-autoscaled-SOMs-FTIR data was better than SG-autoscaled-UV-Vis data and the accuracy of SG-autoscaled-SOMs-fusion technique was better than the other models. This method predicted the class of samples more accurately (more than 95%). The authentication and quality of fraud samples were identified more correctly with respect to raw data.

Keywords: Spice and Herbs, FTIR, UV-Vis, Savitzky-Golay, SOM, Data fusion

INTRODUCTION

Herbs and spices have been widely applied in the food industry for improving the taste and, or aroma and preservation of food. Also, they contain pharmacologic properties and have a particular impact on the prevention and treatment of diseases. They fall under the class Angiospermae or the flowering plants and use plants for adding flavor, colure, and aroma to food. Herbs are taken from the leaves of a plant or fresh parts of the plants (non-woody). But spices are usually obtained from parts of plants other than the leaves. Spices are the dried root, stalk, seed, or dried fruit of the plant and are always dried, not fresh.

Herbs and spices have a long culinary usage and health-beneficial history in Iran. The history of usage in Iran dates back to the Aryan civilization from about 6500 to 7000 BC [1-2]. Iranian physicians, such as Avicenna and Razi made high development in Persian medicine. In the 13th century, Ibn al-Baitar explained the properties of over 1400 plants [3].

Also, these samples are rich sources of large groups of bioactive compounds like flavonoids, phenolic compounds, carotenoids, plant sterols, glucosinolates, sulfur-containing compounds, tannins, alkaloids, and phenolic diterpenes. These compounds display different antioxidant activities [4-7]. Some recent studies have suggested that spices and herbs like rosemary, sage, and oregano with a high content of

*Corresponding author. E-mail: matarighat@pgu.ac.ir

phenolic compounds act as strong antioxidants [8].

It is not at all true to claim that spices and herbs have natural sources and are harmless. As a result, herbal products' improper use has dangerous side effects and can cause allergic reactions, liver or kidney damage, colon perforation, carcinoma, coma, and death. Powdered and milled spices can increase the risk of fraud. The flour, chalk powder, or bread are mixed with turmeric and sold as turmeric powder. Chilli, buckwheat or millet, and papaya have been used for adulteration of black pepper or fibers, dyed corn stigmas, red-dyed silk fibers, safflower, marigold to red stigma, stigmas of other saffron types, flowers, starch, glucose for saffron adulteration, and almond, peanut, tree nuts, peach, and cherry, fennel seeds, and Peanut shell for cumin have been applied [8-9].

Sensory, physical, and chemical examinations have been used to determine the quality of herbs and spices. Recently, the electronic nose, electronic tongue, and electronic eye have been replaced by human senses. Typically, physical characteristics like color, density, texture, solubility, or calorimetry are used to analyze the physical aspects [10]. To determine the safety and quality of herbs and spices, different techniques have been used. Some of the instrumental procedures for examination of the quality and adulteration of spices and herbs are thin-layer chromatography (TLC) [11], high-performance liquid chromatography (HPLC) [11], ultra-high performance liquid chromatography (UHPLC) [12], gas Chromatography [13], plasma atomic emission spectrometry (ICP-AES) [14-15], and inductively coupled plasma-optical Emission Spectroscopy (ICP-OES) [15]. Chromatographic methods are useful for the identification of spices and herbs but need a time-consuming preparation process and consume large volumes of environmentally unfriendly solvents. ICP-AES, ICP-OES, and ICP-MS require expensive equipment, and equipment maintenance and repair cost is high.

The use of spectroscopy technology to carry out quality assessment has been approved by the World Health Organization. The technology used in spectroscopy is based on non-target measurement, which perfectly matches the chemical properties of herbs and spices [10,16-18]. However, rapid and robust assessment methods with low cost and technical skills, and little or no sample preparation are desired. Among the different solutions, special attention has been paid to FT-IR as a fast and cheap analytical technique [19-26].

Through the application of multivariate chemometric tools, spectroscopic techniques generate a high amount of information about the profiles of species. Several non-destructive procedures based on spectroscopic techniques (mostly Near-Infrared NIR paired with chemometric tools) have recently been proposed for the quality of food, spices, and herbs due to their several advantages. UV-Vis and FT-IR fingerprints contain complex information that describes the overall signal from many chromophores, and vibrational modes in the sample, respectively. Also, discrimination based on visual inspection of UV-Vis or IR spectra is difficult because the spectra are very similar to each other and only changing intensities of bands occur from one sample to others. The use of FT-IR and UV-Vis with chemometrics has been applied for adulteration and detection of fraud in herbs and spices. Discrimination of Iranian and Spanish saffron samples has been performed using principal component analysis [27-28]. Checking methods for the food industry must be easy to use, rapid, and low costs [21].

Preprocessing the original spectrum, wavelength selection, and feature extraction are critical steps in multivariate analysis. These data include variables that are uninformative due to excessive noise or background signal. Variable and feature selection are becoming increasingly important for data reduction and improving interpretability as measurement data complexity increases. While feature extraction chooses variables that vectors have integrated, the variable selection focuses on choosing the original data variables that contribute the most to the desired attribute. These procedures improve the predictive performance of models and produce predicted values more quickly and affordably [19].

Feature selection facilitates classification by removing irrelevant features. Classification algorithms frequently overfit training data when there are numerous irrelevant features present. Because a learning model would allow the model to adjust to random factors that are unique to the training data but not to the full data. Hence, overfitting, with high dimensional data, causes a widening of the estimated and true accuracy gap as well as a crucial decrease in classifier performance.

Self-Organizing Maps (SOMs) are one of the most effective tools which combine visualization and clustering to provide amazing insights from large data sets. SOMs widely use unsupervised neural network architecture to discover a group of structures in a dataset. SOMs have been used for feature selection, finding the spacing and position of high dimensional clusters, and patterns of data. SOMs have been used for the extraction of essential patterns from noisy data [19-22].

We think that applying SOMs facilitates the classification of a high-dimensional dataset by providing useful and relevant information. So, we used clusters of SOMs for the classification and authentication of samples.

The variety and geographical origin of spices and herbs are critical for food processing companies, and consumers. By increasing the consumption of spices and herbs, food fraud has also increased. One of the main problems of food adulteration is the replacement of expensive ingredients with cheaper allergenic or toxic ones. So, it is necessary to introduce modern diagnostic non-destructive techniques for the differentiation and authentication of these samples. It is insufficient to perform a comprehensive comparison of quality or properly characterize the properties of complicated products using a single analytical technique. Although, a single technique can present a lot of sample information. Consequently, data fusion from several sources may produce complementary or collaborative information for highly accurate, precise, and complete detection using synergistic effects.

In the present study, multivariate analyses of UV-Vis and FTIR fingerprints have been applied to classify the Iranian standard herbs and spices and their mislabeled and adulterated samples. The data of two spectroscopic techniques individually were analyzed using PCA as a usual clustering method but not obtained satisfactory results. Hence, classification results were investigated using PCA-LDA and PLS-DA techniques.

The main objectives of this study include 1-Classifying and authenticating Iranian standard herbs and spices, and their mislabeled using UV-Vis and FTIR fingerprints using multivariate chemometrics approaches; 2-Examining various preprocessing and smoothing criteria for quality assessment of the samples; 3-Application of SOMs clusters as selected variables; 4-Data fusion, combining data, from UV-Vis and FTIR SOMs clusters and 4-Better classification of samples by Savitzky-Golay-autoscaled-SOMs-principal component analysis-linear discriminant analysis (PCA-LDA).

Based on our knowledge, the current study is the first work that examined the applicability of SOMs variables-data fusion for the classification and authentication of herbs and spices. The results show preprocessing by Savitzky-Golay(SG) filter and then autoscaling provided better classification. Also, SG-autoscaled-SOMs-PCA-LDA and SG-autoscaled-SOMs-PLS-LDA identified original and mislabeled samples (*e.g.* saffron samples from fraud samples) from each other. Also, the qualities of mint samples from the same origin at different harvests were distinguished. A mislabeled Caraway sample was differentiated from Cumin samples. The results of classification parameters indicated that the accuracy of prediction samples by SG-autoscaled-SOMs-PCA-LDA and SOMs-PLS-LDA was greater than other models.

MATERIALS AND METHODS

Chemicals and Plant Extraction

Used solvents were of analytical grade and were purchased from Merck (Darmstadt, Germany). A total set of 63 different samples were collected across several cities of Iran, including Bushehr, Borazjan, Hamedan, Sanandaj, Mashhad, Shahrekord, Dena, Ghazvin, Urumayeh, Qom and Chalos, Yasooj, Kerman, and Shiraz. The Saffron sample was mixed with turmeric as an ingredient and one food colorant was mislabeled as saffron. Three *Caraway* samples were incorrectly labeled as *Cumin* class. Mint samples of Hamedan belong to different harvests. The descriptions of the samples are given in Table 1.

After obtaining, fresh samples of Mint, Thyme, Savory, Oregano, and Ajwain were dried. The other samples were purchased from local markets of different cities. The dried samples were milled to a homogenous powder in the ball mill. 10.0 g of each dried sample was transferred to the grinding jars and again they were milled at 500 rpm for 5 min. The 5.0 g of powdered samples were extracted with 100 ml of 65% (v/v) ethanol/water with ultrasonication for 40 min. The extracted samples were filtered and transferred to a 10 ml volumetric flask. These samples were used for UV-Vis analysis.

The sample powders were mixed with KBr at a suitable ratio of 1/180 (w/w), homogenized, and compressed by applying 200 MPa for 1 min, and FT-IR spectra were recorded. The clean KBr disc was made and applied for background correction. The spectra were recorded in transmittance mode from 400 to 4000 cm^{-1} using Bruker Vector 22 FT-IR equipment. Each measurement was carried out three times (Fig. 2).

Measurements

For FT-IR transmittance measurements, all samples were mixed with KBr at a suitable ratio of 1/180 (w/w) and homogenized. This mixture by applying 200 MPa for 1 min was compressed and a thin KBr disc was obtained. For each sample, the disc preparation procedure was carried out in triplicate. The spectrum of a clean KBr disc (without herbs and spices) was used for background correction. A Bruker Vector 22 FT-IR with a spectral range from 4000 to 400 cm^{-1} in transmittance mode was used for spectral measurement. The spectra were stored using the OPUS software supplied by the same manufacturer.

UV-Vis spectra of all extract solutions were determined by using a double-beam spectrophotometer Analytical Jena SPECORD250-22P16 UV-Vis. The measurement UV-Vis spectra were made in the range of wavelength from 250-600 nm with 1 nm interval in 1 cm quartz cuvette. Ethanol was used as a blank in the measurements.

Computational Software

Calculations were done by MATLAB software (version 7.8, MathWorks, Natick, MA, USA). Classification toolbox and PCA toolbox for MATLAB were used for the identification and discrimination of samples [29-30].

Multivariate Analysis

Multivariate analysis is used to study high-dimensional data. More variables must provide more information for measurements. However, the model with more variables requires more parameters to explain. Hence, the efficiency of analysis would be lost. Therefore, suitably reducing a large number of variables into a smaller number of orthogonal and uncorrelated factors is interesting [14,17,24]. The techniques of PCA, linear discriminant analysis (LDA), partial least square discriminant analysis (PLS-DA), soft independent modeling of class analogy (SIMCA), and Kohonen maps are widely used in multivariate analysis. In this study, we used self-organizing maps (SOM) for variable selection and PCA-DA and PLS-DA for the classification of herbs and spices.

Data matrix and Preprocessing

Before multivariate analysis, data preprocessing of data can be useful. Data treatment can isolate the interested signal from the interferences and noise signals and minimize problems arising from baseline shifts. By removing irrelevant information, they can improve the quality and predictive ability of the models, and hence enhance the accuracy and repeatability of results. Centering, smoothing, autoscaling, Pareto scaling, range scaling, vast scaling, log transformation, and power transformation, were proposed for preprocessing. The appropriate preprocessing method depends on the properties of the data set and the method of analysis [31-33].

Savitzky-Golay (SavGol) is a widely-used pretreatment method for smoothing spectral data. It can effectively eliminate the noises (baseline-drift and reverse) without loss of resolution. The number of smoothing points (SP) to the left and right of the reference point and the degree of the least squares polynomial of the Savitzky-Golay filter should be optimized. If the SP is small, a calculation error occurs and decreases model precision, while a too-big SP number causes over-smoothing, polishing the spectral data and decreasing accuracy. In the present study, we represent that the SavGol smoothing enhances the accuracy of classification results. At autoscaling, firstly the average of variables is subtracted from observed values then the variance of the variables is set to the unit.

Spectral FTIR and UV-Vis Spectra smoothed using Savitzky-Golay with 8 smoothing point 4^{th} order polynomial filter. The complete smoothed data sets or predicted clusters of SOMs were subjected to classification tools and then autoscaling was performed.

SOMs. The SOMs have been introduced by Teuvo Kohonen as a highly efficient visualization tool for visualizing and abstraction of high-dimensional and complex data. The SOMs lead to the automatic formation of clusters of similar data segments [23, 24]. The SOM network usually contains two the input layer and the Kohonen layer. The input

layer is completely connected to a two-dimensional Kohonen layer. At the training step, input data are fed to the network through the processing nodes in the input layer. The input pattern is defined as:

$$x_l = x_{l1}, x_{l2}, x_{l3,}, \dots, x_{lm}$$

where, x_{li} is the ith input of signal and *m* is the number of signals in each pattern. For n×n Kohonen layer, a weight vector of $w_i = w_{i1}, w_{i2}, w_{i3,}, \dots, w_{im}$ is associated with the *i*th node to the *j*th signal of the input vector. At the training step, the nodes modify their weight values according to the topological relations in the input data. The node or neuron with the minimum distance is the winner and adjusts its weights to be closer to the value of the input pattern.

RESULTS AND DISCUSSION

In this study, about 63 samples were measured using FTIR and UV-Vis techniques. The FTIR spectra were measured at 500.102 to 4000.60 cm^{-1} (3736 points per sample). UV-Vis spectra were recorded in the wavelength range between 250-500 nm with an interval of 1 nm three times. From these UV-Vis spectra, we will have 201 absorbance data for each sample. The average of three separate UV-Vis spectra of spices and herbs samples was calculated and used each sample spectrum for classification (Fig. 1). The Kennard-Stone algorithm was used for the selection of calibration and prediction samples. The calibration and prediction sets were built using 43 and 20 samples, respectively.

Classification of UV-Vis and FTIR Data Using PCA-DA

The dimension of data can be reduced by performing PCA as a common groupings method and extraction of PCs [14,17]. This method is appropriate for the initial evaluation of similarity or dissimilarity between clusters or classes. So, we tried to extract informative and important features from the pre-processed UV-Vis and FTIR data. Using the two first PC scores, the classification of samples was investigated. The used PCs accounted for 91% and 99% of variances of UV-Vis and FTIR data, respectively (Fig. 2). Due to the high similarity and complexity of FTIR and UV-Vis spectra, the samples were not well distinguished into their groups and classification model was unable to produce a satisfactory result. So, to estimate the classification accuracy, PCA-DA, and PLS-DA algorithms were further used.

Discriminant Analysis (DA)

The LDA by creating a discriminant function of each class can classify a group of observations. Since the UV-Vis or FTIR spectra are correlated with each other and the numbers of variables (wavelength or wavenumber) are more than the number of samples, DA could not be used directly. In other words, DA cannot work effectively when the number of variables is more than the number of samples. Therefore, the number of applied variables was reduced. The common methods for data reduction are PCs of PCA and latent vectors of PLS analysis.

PCA- Linear Discriminant Analysis (LDA) Analysis

Linear Discriminant Analysis (LDA) as a linear classifier tries to find the best projection space and then performs classification on the projected subspaces. The projections are done in such a way as to keep maximum projected distances between classes while the projection distance between subjects in the same classes is minimum [34]. Determination of the optimal number of PCs at the construction of the PCA-LDA model is a very important step. The applied PCs should contribute all of the information. If the number of PCs is too large, overfitting will happen and if low PCs are selected all the information will not be transferred.

Raw data was analyzed using PCA-LDA analysis. The results of the analysis are presented in Table 1. The accuracy of the results is not good. So, pre-processing of analysis was considered.

Spectral FTIR and UV-Vis Spectra smoothed using Savitzky-Golay with 8 smoothing point 4th order polynomial filter. The complete smoothed data sets or predicted clusters of SOMs of spices and herbs were subjected to classification tools and then autoscaling was performed. It means that the smoothed and autoscaled data was used during classification aims.

The optimal number of PCs was employed by internal cross-validation of the training set. So, the optimal number of PCs 12 and 8 was selected for FTIR and UV-Vis analysis, respectively. According to Table 1 results, 91 and 85% of

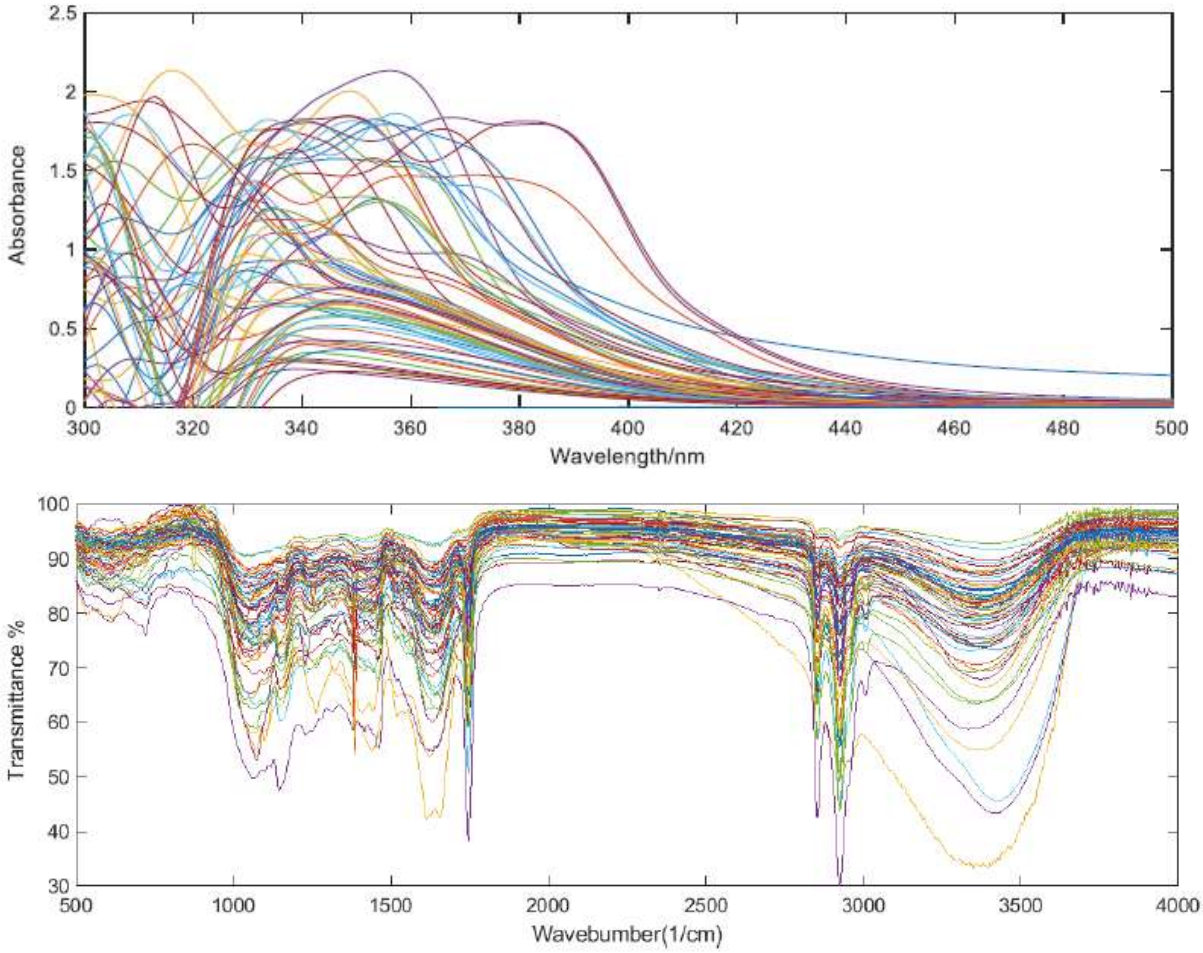

Fig. 1. (a) UV-Vis Spectra and (b) FTIR fingerprint of Iranian herbs and spices.

herbs and spices samples were correctly classified according to their classes using FTIR and UV-Vis data, respectively. Also, the classification of samples using FTIR fingerprints creates better accuracy, selectivity, and sensitivity for samples with respect to UV-Vis data classification.

Partial Least Squares Discriminate (PLS-DA) Analysis

PLS-DA can be assumed as the development of the LDA algorithm. This algorithm uses the latent variables of predicting one (or several) binary responses(s) (Y) from a set of variables in X. The extracted class variables not only have maximal variances of the original variables but also should be correlated with the Y-variables (class variable). PLS-DA is fundamentally based on the PLS2 algorithm that examined for finding latent variables with a maximum covariance (discriminant power) with the Y-variables (class of samples). The classification of raw data was performed using PLS-DA. The calibration parameters were derived and shown in Table 2. According to the results of this table, the results were not satisfactory. Hence, further analysis was done using SG-smoothed data.

PLS-DA makes a dimension reduction and uses the `extracted samples scores for discrimination at calibration and prediction sets. The calibration models were validated by internal cross-validation and root mean square error of calibration (RMSEC) was used for the selection of latent vectors (LVs) of models. The number of 15 and 12 LVs was applied for the classification of FTIR and UV-Vis data sets. The applied LVs explained more than 80% of the variance in the Y block with 95% of the information X matrix (FTIR and UV-Vis spectra).

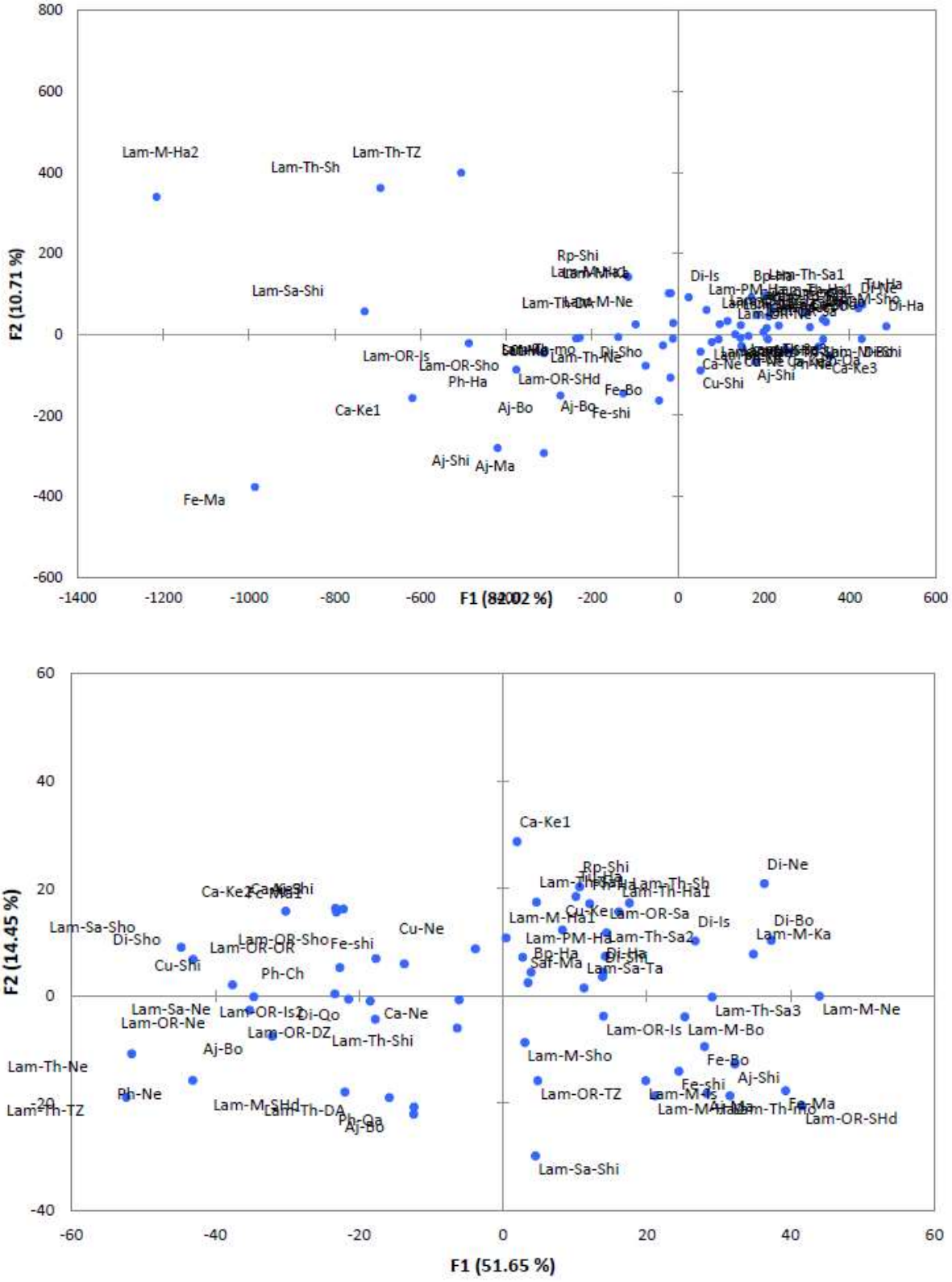

Fig. 2. (a) Score plot of FTIR fingerprint and (b) Score plot of UV-Vis spectra of spices and herbs samples.

Table 1. The Characteristic of Herb and Spices Samples

Number	Sample	Origin	Abbreviation	Number	Sample	Origin	Abbreviation
1	Peppermint	Shiraz	Lam-PM-Shi	33	Ajwain	Shiraz	Aj-Shi
2	Mint	Hamadan	Lam-M-Ha1	34	Ajwain	Borazjan	Aj-Bo
3	Mint	Shoush	Lam-M-Sho	35	Ajwain	Shiraz	Aj-Shi
4	Mint	Hamadan	Lam-M-Ha2	36	Ajwain	Mashhad	Aj-Ma
5	Mint	Isfahan	Lam-M-Is	37	Ajwain	Borazjan	Aj-Bo
6	Mint	Kashan	Lam-M-Ka	38	Black pepper	Hamadan	Bp-Ha
7	Mint	Neyshabur	Lam-M-Ne	39	Red pepper	Shiraz	Rp-Shi
8	Mint	Borazjan	Lam-M-Bo	40	Turmeric	Hamadan	Tu-Ha
9	Mint	Shahrekord	Lam-M-SHd	41	Saffron	Mashhad	Saf-Ma
10	Thyme	Hamadan	Lam-Th-Ha1	42	Dill	Hamadan	Di-Ha
11	Thyme	Sanandaj	Lam-Th-Sa1	43	Dill	Shiraz	Di-Shi
12	Thyme	Shush	Lam-Th-Sh	44	Dill	Isfahan	Di-Is
13	Thyme	Sanandaj	Lam-Th-Sa2	45	Dill	Borazjan	Di-Bo
14	Thyme	Tabriz	Lam-Th-TZ	46	Dill	Ghom	Di-Qo
15	Thyme	Dena	Lam-Th-DA	47	Dill	Neyshabur	Di-Ne
16	Thyme	Sanandaj	Lam-Th-Sa3	48	Dill	Shush	Di-Sho
17	Thyme	morazjan	Lam-Th-mo	49	Caraway	Kerman	Ca-Ke1
18	Thyme	Shiraz	Lam-Th-Shi	50	Caraway	Kerman	Ca-Ke2
19	Thyme	Neyshamur	Lam-Th-Ne	51	Caraway	Kerman	Ca-Ke3
20	Oregano	Isfahan	Lam-OR-Is	52	Caraway	Neyshabur	Ca-Ne
21	Oregano	Tabriz	Lam-OR-TZ	53	Cumin	Kerman	Cu-Ke
22	Oregano	Isfahan	Lam-OR-Is2	54	Cumin	Shiraz	Cu-Shi
23	Oregano	Sanandaj	Lam-OR-Sa	55	Cumin	Neyshabur	Cu-Ne
24	Oregano	Shush	Lam-OR-Sho	56	Persian hogweed	Hamadan	Ph-Ha
25	Oregano	Dezful	Lam-OR-DZ	57	Persian hogweed	Qazvin	Ph-Qa
26	Oregano	Neyshabur	Lam-OR-Ne	58	Persian hogweed	Neyshabur	Ph-Ne
27	Oregano	Orumiyeh	Lam-OR-OR	59	Persian hogweed	Chalus	Ph-Ch
28	Oregano	Shahrekord	Lam-OR-SHd	60	Fennel	Shiraz	Fe-shi
29	Savory	Tabriz	Lam-Sa-Ta	61	Fennel	Shiraz	Fe-shi
30	Savory	Shush	Lam-Sa-Sho	62	Fennel	Mashhad	Fe-Ma
31	Savory	Neyshabur	Lam-Sa-Ne	63	Food colorant	Borazjan	FC-Bo
32	Savory	Shiraz	Lam-Sa-Shi				

The initial estimation of the similarity of samples can be considered by the SG-autoscaled-PLS-DA. The estimated class of spices and herbs using FTIR data was illustrated in Fig. 3a. It can be said that, according to the pattern of samples, they were distributed based on their major content and divided into two groups, one group at the above borderline and the others at the bottom. It can be seen the mint family (Lamiaceae*)* was limited to 0.5 and 1.5 above the zero lines. Ajwain samples between 0 and 0.5 values, Caraway, Cumin, Dill, Persian hogweed, and Fennel and Saffron between -0.5 and -1.5 (real and fake samples) under the zero line are located. Hence, discrimination patterns are dependent on the chemical contents of samples. Lamiaceae samples were differentiated from other spices and

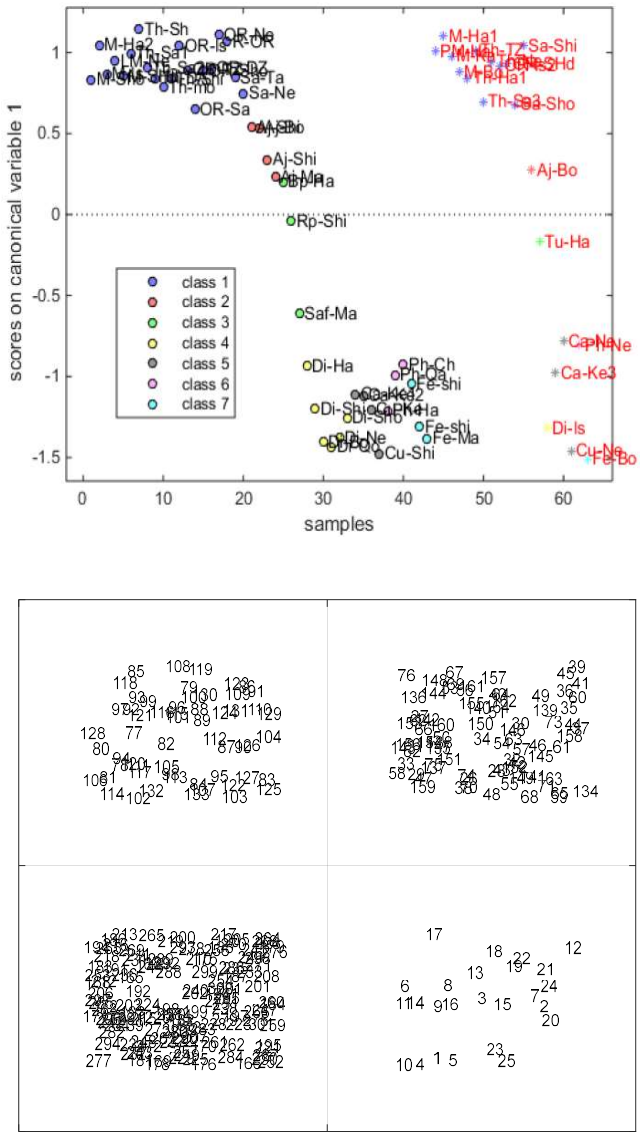

Fig. 3. (a)Estimated class of spice and herb samples by the PLS-DA model versus the calibration and prediction sets using FTIR data and (b) top map of SOM network 2×2 of UV-Vis spectra.

herbs. The flavonoids and phenolic acid content of the mint family (mint, savory, oregano, and thyme) are similar, they are distributed closer together. Dill samples contained rosmarinic acid and carvone and the major phenolic compounds offennel, Persian hogweed, cumin, and caraway are carvone and carvacrol. Therefore, samples with similar content are distributed close to each other. Also, the original saffron sample (Saf-Ma), and adulterated samples of saffron are separated from the original sample on discriminate factors space [32-35]. The classification performances of the models were investigated by comparing the classification parameters of training and prediction samples using optimized LVs (Table 2). The results of this table illustrate that the model based on smoothed FTIR data using 15 LVs could correctly classify different spice and herb samples just the Cu-Ne misclassified as Cumin class and La-Sa-Sho, Aj-Bo, and Cu-Ne were not assigned. Also, the samples La-M-Ha, La-Sa-Shi, Ca-Ke3, Ca-Ne, Cu-Ne2, and Aj-Bo of samples were not assigned by using a model based on the UV-Vis data. The predictability of the model was tested by the accuracy of the prediction set. The accuracy (AC) is evaluated using the ratio of correctly assigned samples:

$$AC = \frac{\sum_{g=1}^{G} n_{gg}}{n} \quad (1)$$

where n is the total number of samples. Not assigned samples are not considered for the accuracy calculation. The accuracy of PLS-DA models for prediction samples was obtained at 80% and 85% for FTIR and UV-Vis, respectively.

This pattern is similar to the PCA score plot; however, the separation of samples appears more obvious. It can be explained by the fact that the DA (at PCA-DA and PLS-DA algorithm) maximizes the variance between groups where the within-groups variances are minimized.

Description of Models Based on SOMs

One of the main objectives of this analysis is finding the wavenumbers or wavelengths containing important information that can predict classes better than full-spectrum ranges. SOMs are the unsupervised –clustering neural network and might be considered as a nonlinear generalization of PCA for this purpose. SOMs perform a characteristic of a nonlinear projection from the high dimensional space of input data onto low-dimensional patterns and produce a map of similarities (Fig. 3b).

FTIR spectrum and UV-Vis include 3736 and 201 points per sample, respectively. All of the wavenumbers do not contain useful classification information. So, SOMs were

Table 2. Comparing the Classification Results of Smoothed by SG and Non-smoothed Data

Data	Parameter	SG-Smoothed data				Non smoothed data			
		FTIR		UV-Vis		FTIR		UV-Vis	
		PCA-DA	PLS-DA	PCA-DA	PLS-DA	PCA-DA	PLS-DA	PCA-DA	PLS-DA
Calibration	ER[a]	0.0	0.05	0.08	0.38	0.05	0.08	0.05	0.11
	NER[b]	1.0	0.95	0.92	0.63	0.95	0.92	0.95	0.89
	N.A[c]	0.0	0.0	0.0	0.39	0.0	0.1	0.0	0.15
	Accuracy (%)	100	96	93	93	96	93	96	93
Prediction	ER[a]	0.19	0.20	0.16	0.2	0.20	0.46	0.62	0.47
	NER[b]	0.81	0.80	0.84	0.80	0.80	0.54	0.38	0.53
	N.A.[c]	0	0.22	0	0.22	0.0	0.0	0.0	0.0
	Accuracy (%)	91.0	80	90	85	83	87	44	61

[a]Error rates(ER). [b]Non-error rate (NER). [c]Not assigned.

Table 3. Sensitivity and Specificity of Prediction Set Classes Set Using PCA-DA

Data	Parameter	Map size			
		2×2	3×3	4×4	5×5
FTIR	Sensitivity (%)	0.78	0.95	0.88	0.83
	Specificity (%)	0.80	0.97	0.89	0.80
UV-Vis	Sensitivity (%)	0.85	0.80	0.72	0.79
	Specificity (%)	0.86	0.79	0.78	0.86

used for the selection of the best variables by clustering them. The number of Kohonen nodes is the first parameter that should be determined and optimized. Every n-node leads to n × n cluster of variables or n^2 clusters produces. The size of the map especially depends on the input and the purposes of the classification. If the aim of clustering is to find all of the patterns in the data sets, particularly containing those with a low probability of existence, the large size of maps must be used. But if the predominant patterns are concerned, smaller-sized SOMs can be employed. However, large sizes for the SOMs are used if the efficiency of the classification technique is not remarkably affected.

When FTIR and UV-Vis spectra were inserted into the Kohonen SOMs, similar variables (wavenumbers or wavelengths) were placed in the same cluster. The number of variables in each cluster depends on the similarityof valuable information. For the selection of better SOMs node number, different Kohonen SOMs networks from the node sizes of 2×2 to 5×5 have been checked and sensitivity (model ability to correctly recognize samples belonging to each class), specificity (model ability to reject samples of all the other classes from class g^{th} of classes), accuracy and the number of not-assigned samples have been considered.

Analysis of FTIR Data by SOMs Clustering

Different Kohonen SOM network (2×2 to 4×4) was created and distributions of variables were considered. Through the clusters, different models were built and classification results were investigated. Tables 3 and 4 show the mean values of sensitivities and specificities of different

SOMs for FTIR data. The results illustrate that model 3×3 is appropriate for FTIR data analysis. It is clear that the number of variables is different in each cluster.

The cluster characteristics, the interval of wavenumbers of each cluster, sensitivity, and specificity of classification of different models are given in Tables 4 and 5. As can be seen, the variables are not equally distributed in clusters and each cluster includes different numbers of wavenumbers. Using suggested clusters, SG-autoscaled-SOMs-PCA-LDA and SG-autoscaled-SOMs-PLS-LDA models were created. After the optimization of models, classification was done. The calibration and prediction results were used for the optimization of best model selections. But for simplicity, just the results of the prediction set were reported. Table 3 shows the statistical parameters of PCA-DA and PLS-DA models by some of the SOMs clusters. It can be seen that clusters, S_1 and S_2 of the 3×3 Kohonen SOM network classified samples with appropriate ER, NER, and accuracy values. At PCA-DA models, S_1 of 3×3 provided better accuracy than a cluster of S_2 (3×3). According to the results presented in Tables 2 and 3, SG-autoscaled-SOMs-PCA-LDA of network size 3 has higher values of sensitivity, sensitivity, accuracy, and lower ER values. It illustrates the correct recognition ability of the model for the prediction of samples with respect to other networks. Hence, wavenumbers of S_1 of 3×3 belonging to this cluster were used for further studies. The wavenumbers

Table 4. The Cluster Characteristics, Wavenumber Intervals of some Clusters of 2×2, 3×3 and 4×4 of SOMs Using SG-FTIR Data with PCA-DA and PLS-DA

Map	Range (Wavenumber cm^{-1})	ER	NER	Accuracy	ER	NER	N.A.	Accuracy
S_1 (3×3)	977.7-1004.7; 1128.1-1753.9; 2480.6-3193.5; 3525.2-3602.4	0.17	0.83	93	0.29	0.71	0.12	1.0
S_2(3×3)	1005.7-1127.2; 1587.1-1677.8; 2915.8-2396.1; 3194.5-3524.2	0.06	0.94	90	0.30	0.70	0.17	1.0
S_3(3×3)	729.9-944.9; 1771.9-2564.5; 2566.8-2640.3; 3669.8-4000.5	0.25	0.75	80	0.14	0.86	0. 2	1.0
S_1(4×4)	1806.9-2426.1; 3707.1-4000.5	0.33	0.67	85	0.30	0.70	0.15	1
S_2(4×4)	1521.6-1669.8; 2740.1-3488.7;	0.15	0.85	90	0.14	0.86	0.2	1
S_3(4×4)	1596.1-1669.1; 3246.5-3489.6	0.2	0.83	85	0.14	0.86	0. 2	1
S_4(4×4)	2808.8-3635.1	0.35	0.65	78	0.43	0.71	0.55	1

Table 5. The Cluster Characteristics, Wavenumber Intervals of some Clusters of 2×2, 3×3 and 4×4 of SOMs Using SG-UV-Vis Data with PCA-DA and PLS-DA

		PCA-DA			PLS-DA			
Map	Range (nm)	ER	NER	Accuracy (%)	ER	NER	N.A.	Accuracy (%)
S1(2×2)	400-530	0.25	0.75	85	0	1.0	0.2	0.98
S2(2×2)	308-399	0.2	0.8	85	0.14	0.86	0.15	1
S1(3×3)	329-392	0.17	0.83	85	0.25	0.85	0.1	88
S2(3×3)	396-411	0.25	0.75	83	0.3	0.7	0.2	0.80
S3(3×3)	412-438	0.17	0.83	84	0.29	0.71	0.2	0.79

Table 6. The Results of PCA-LDA and PLS-DA Analysis Using Best-proposed Clusters by SOMs at Different Data Sets

Data	PCA-DA			PLS-DA			
	ER	NER	Accuracy (%)	ER	NER	N.A.	Accuracy (%)
FTIR	0.17	0.83	93	0.29	0.71	0.12	1
UV-Vis	0.2	0.8	85	0.14	0.86	0.15	1
Fusion matrix	0.12	0.88	96	0.14	0.86	0.05	1

977.7-1004.7 cm^{-1} corresponding to C–O stretching and1128.1-1753.9 cm^{-1} corresponding to C–H rocking, =C-O-C, phenyl, Isopropyl, C=C, and C=O stretching. The wavenumber regions 2480.6-3193.5 cm^{-1} and 3525.2-3602.4 cm^{-1} belong to saturated C-H and–OH, –NH stretching [30-33].

The discriminant function plot of S_1 Kohonen network size q = 3 of SG-autoscaled-SOMs-PCA-LDA and SG-autoscaled-SOMs-PLS-LDA models are shown in Fig. 4. It can be realized that a better separation was obtained with respect to autoscaled-PCA-DA and autoscaled PLS-DA models. According to this Fig. 4a, mint samples of Hamedan (M-Ha2) were located upper the zero line while other samples are lower than this line. This sample was old ample. So, the quality of samples is different and they are distinguished from others. The M-Ha1 and PM-Ha2 that were collected at different harvests at one year are similar to each other than the old sample (La-M-Ha2). Also, saffron original samples were differentiated from fraud samples.

Analysis of UV-Vis Data by SOMs Clustering

Also, the classification of samples was checked using UV-Vis data. In order to find the best model, different models using different sizes of clusters were built. Tables 2 and 4 include the different Kohonen network (q = 2 and 3) information.

For PCA-DA analysis, the five first PCs were used. The number of selected PCs in the LDA analysis remains relatively constant and is independent of the number of clusters. Like FTIR analysis, the calibration and prediction results were used for the selection of the best model. The results of prediction samples were reported for comparing the effect of clusters on classification results. Table 5 shows the summary of the results. As the SG- SOMs-PCA-LDA model reveals, the Kohonen nodes q = 2, (4 clusters) is better than q = 3, (9 clusters). So, the proposed clusters of network size q = 2 were used for classification. The results of classification parameters are depicted in Table 5. It is clear that the model with the best performance was obtained by excluding the

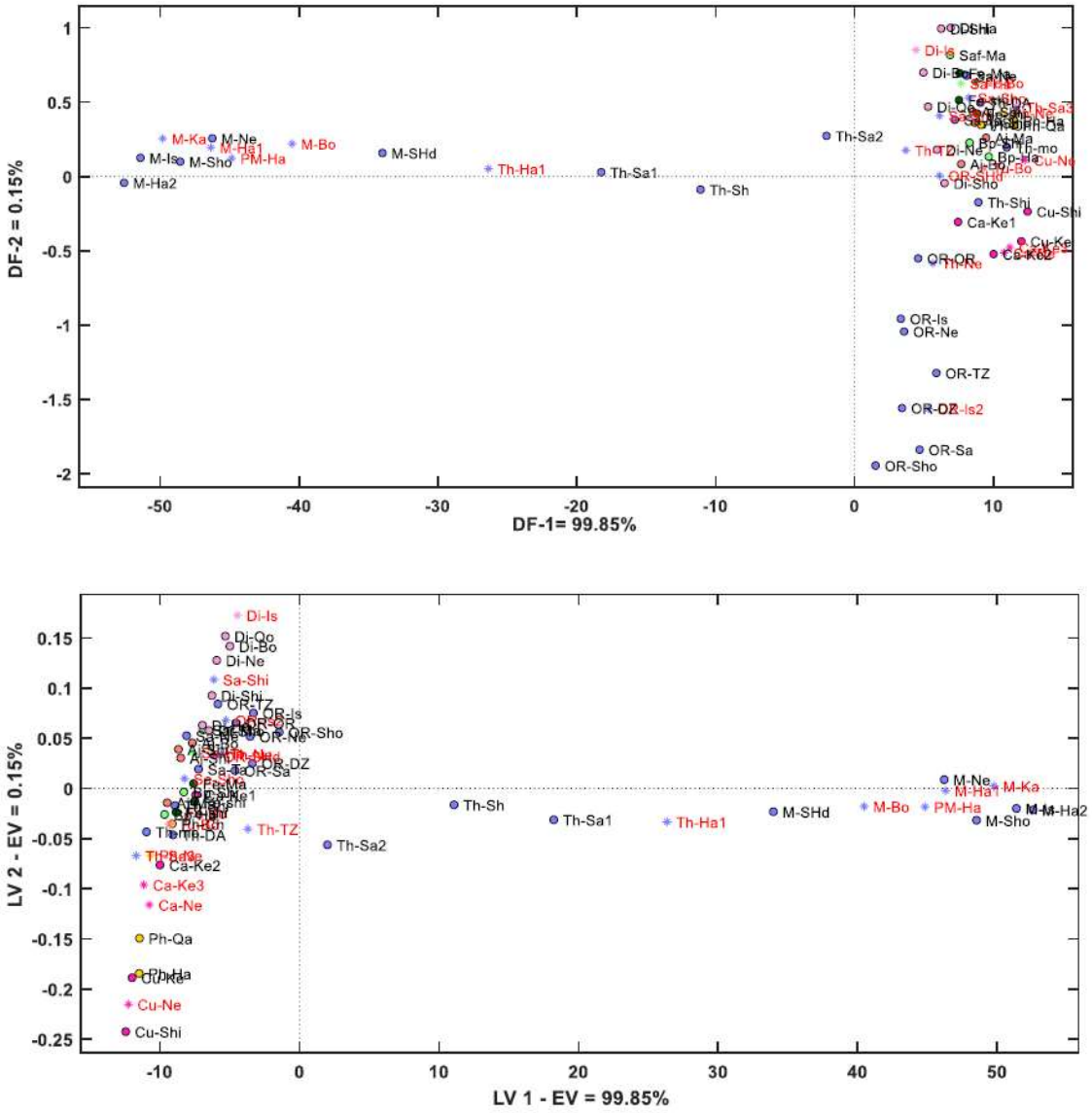

Fig. 4. Discriminant plot of S1 (3×3) of SG-autoscaled-FTIR using (a) PCA-LDA and (b) PLS-DA.

region 308-399 nm. So, the cluster predicted samples with high accuracy with respect to autoscaled-PCA-DA and autoscaled PLS-DA models. The discriminant function plot of S_2 of Kohonen network size q = 2 is shown in Figure 5a. The UV-Vis absorption at this region is dependent on the presence of flavonoid compounds like quercetin (333-373 nm), crocin and safranal (320-400 nm), and kaempferol (310-380 nm). Probably, this region contains mostly common spectral features and can be contributed to the detection and identification of original and adulterated samples.

The same procedure was applied for adulterant identification and quality determinations using PLS-DA analysis using cross-validation (Venetian blinds 5 cross-validation groups) procedure and further validation using a prediction set. Different clusters were introduced to the PLS-DA model and classification was performed. Tables 4 and 5 show some clusters (Si), ER, NER, not assigned samples, and the accuracy of test samples. Since the accuracy of the models is close to each other so, ER, NER, and not assigned samples were considered. The SOMs cluster of S_2 (2×2) with

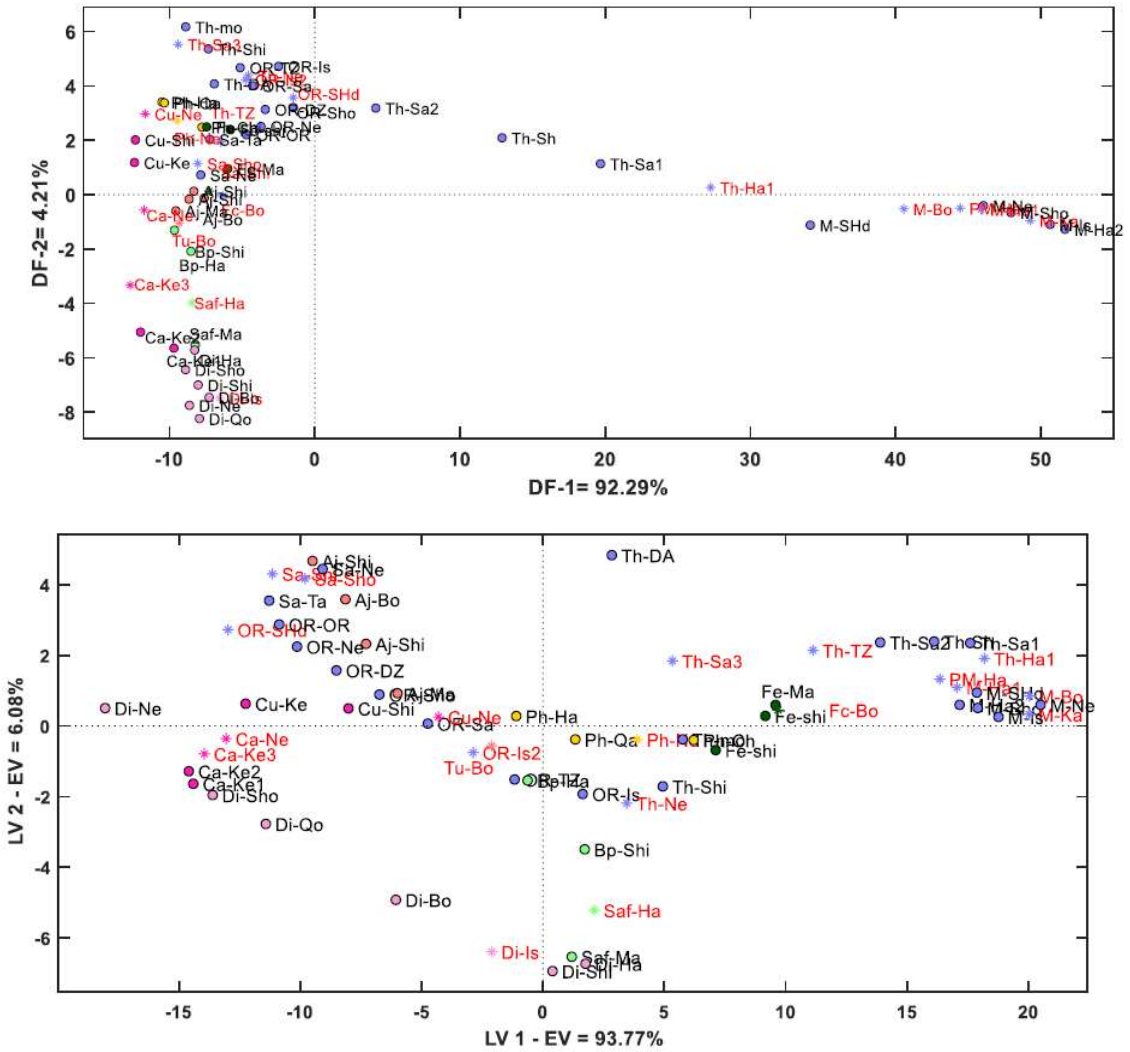

Fig. 5. The discriminant function plot of S2 of Kohonen network size q = 2 by (a) PCA-LDA and (b) PLS-DA.

lower ER value and not assigned samples and higher values of NER were selected for the construction of the PLS-DA model of UV-Vis data.

Figure 5b shows the PLS-DA latent variables scores plot of LV1 *vs*. LV2 of SG-autoscaled UV-Vis data. The first two LVs explained more than 99% and 95% of the variance of FTIR and UV-Vis data, respectively. This figure shows that the SG-autoscaled PLS-DA models can identify samples. Samples La-Sa-Shi, Ca-Ne, and Cu-Ne have not been assigned by the model.

Data Fusion

The present study may enable the introduction of a novel strategy for spice and herbs quality control. The classification results showed that FTIR data provide better discrimination of samples compared to UV-Vis spectroscopy. It is due to the high throughput of FTIR as an efficient technique for the characterization of sample profiles.

Also, we showed that the SOMs of single analysis (FTIR or UV-Vis) produced an accurate technique for data reduction and variable selection. Because, when the number of variables is much more than the number of samples data reduction and selection of informative variables are necessary. Here, the synergistic advantages of FTIR and UV-Vis were checked by data fusion strategy. The best Kohonen clusters of FTIR and UV-Vis data were used for

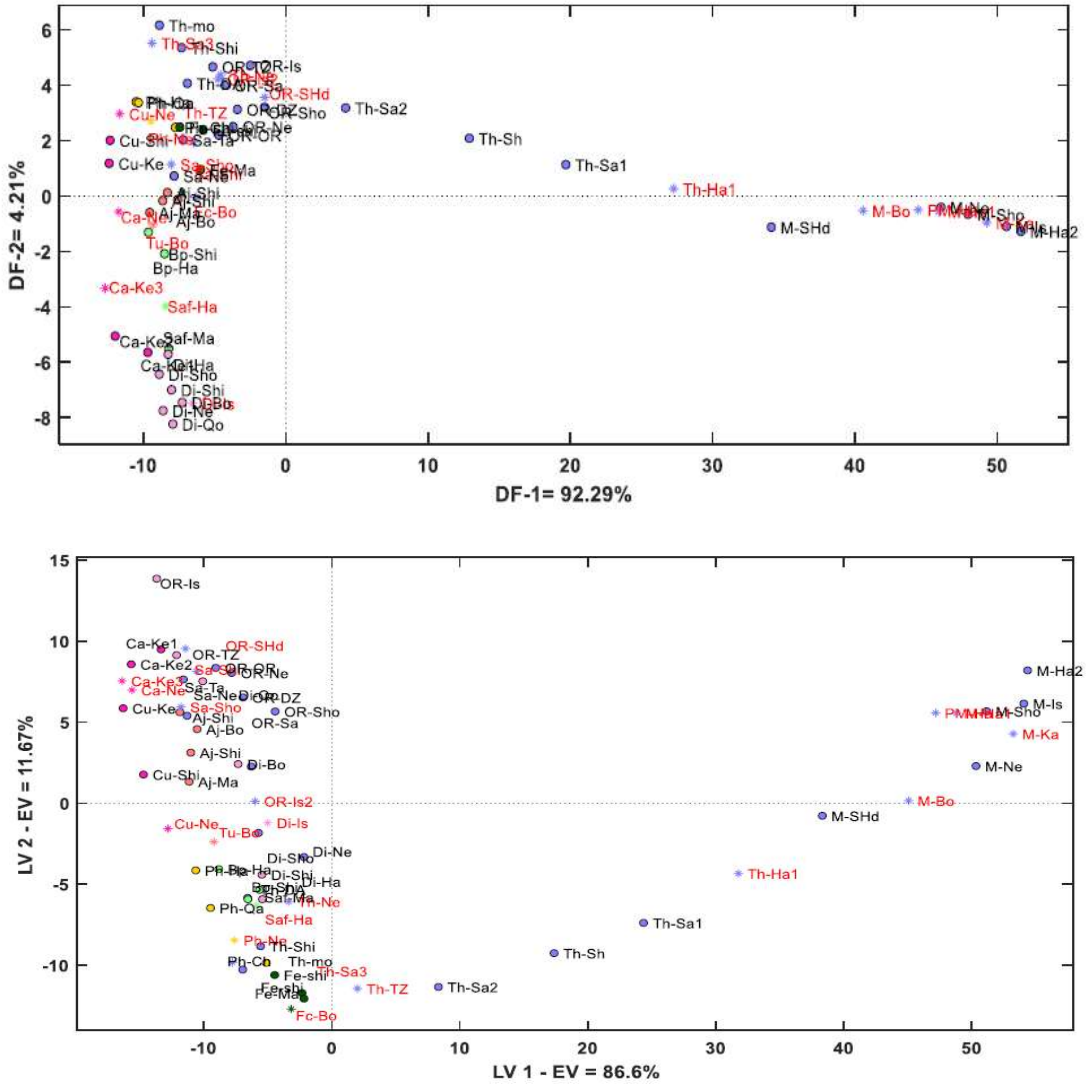

Fig. 6. The discriminant function plot of fusion data by (a) PCA-DA and (b) PLS-DA

integrating information from the two methods and creating characteristic, comprehensive, and unified data at data fusion. The multivariate statistical analysis of fusion data was performed to recognize the quality and adulteration of samples. The fusion matrix was analyzed by PCA-LDA using the first eight LVs. The NER value increased while ER values decreased (Table 5). These results show that the classification has been improved. The accuracy of PCA-DA of the fusion matrix (96%) is higher than the PCA-DA analysis of other classification models. Fig. 6a shows the DA plot of the first two discriminant functions. The fusion data were analyzed using the PLS-DA model, too. The fusion data were analyzed using the PLS-DA model, too. The first two LVs of the PLS-DA model, explained 98.2% of the variance, while twelve LVs were used in total and accounted for 100% of the variance. The predicted results of the test set are given in Table 5. According to this table, the percentage of not assigned samples was obtained 5, 85, and 88 for fusion, single FTIR, and single UV-Vis data, respectively. So, the classes of more samples were correctly predicted using this technique. Figure 6 shows mislabeled caraway samples were differentiated from cumin and food colorant (FC-Bo) and fraud saffron sample (with turmeric) from an original saffron sample. Also, the quality of different mint

samples from the different harvests was detected. The boundaries between very similar samples (mint family) are also easily differentiable, so interpretation will be easy. Therefore, the SOMs clusters at data fusion could utilize the synergies advantageous of original data sets and can help interpretation of results. The classification results clarified the potential of SG-autoscaled-SOMs -PCA-DA with respect to other models for reducing the dimensionality of spectral datasets, noise correction, data mining, and sample differentiation, as well as producing the highest classification accuracies.

CONCLUSION

The great concern of food producers and consumers enforce to apply high-throughput analytical methods for the authenticity of valuable products such as saffron, caraway, and black pepper. Coupling multivariate data of FTIR and Vis spectra with chemometrics techniques provides fast and accurate new methodologies for the identification of similarity/dissimilarity between different groups of samples. It is better to perform preprocessing to remove unwanted variances of background and noise in the data for enhancing the predictive ability of classification. The PCA-LDA and PLS-DA were applied on raw and preprocessed spectra but these techniques could not distinguish them into their groups. To the complexity and presence of irrelevant information in this data, the risk of overfitting during the analysis will be increased. In this paper, autoscaling, and SG smoothing were determined as the optimal preprocessing techniques. Also, the potential of classification tools was checked using SOMs clusters and a data fusion approach. The results show that all of the developed methods are favorable for the differentiation of herbs and spices. Since the clusters are informative and without redundant information, can greatly classify the samples. In conclusion, these features confirm the influence of SOMs as an effective tool for data mining, and authentication purposes, such as for determining the geographical origin, mislabeling, and quality of samples.

REFERENCES

[1] C. Elgood, A Medical History of Persia and the Eastern Caliphate: from the earliest times until the year AD 1932 Cambridge University Press, 2010.

[2] F. Jamshidi-Kia, Z. Lorigooini, H. Amini-Khoei, J. Herbmed. Pharmacol. 7 (2018) 1.

[3] A. Gurib-Fakim, Mol. Aspects. Med. 27 (2006) 1.

[4] D.J. Charles, Antioxidant Properties of Spices, Herbs and Other Sources; Springer: New York, NY, USA, 2013.

[5] I.B. Jaganath, A. Crozier, Dietary flavonoids and phenolic compounds In C.G. Fraga (Ed.), Dietary Flavonoids and Phenolic compounds Hoboken, NJ: JohnWiley Sons, 2010, pp. 1-49.

[6] K.V. Peter (Ed.), Handbook of Herbs and Spices, Elsevier, 2012.

[7] K. Srinivasan, Crit. Rev. Food Sc.i Nutr. 54 (2014) 352.

[8] L. Pizzale, R. Bortolomeazzi, S. Vichi, E. Uberegger, L.S. Conte, J. Sci. Food Agric. 82 (2002) 1645.

[9] N. Haghnegahdar, M. Abbasi Tarighatt, D. Dastan, J. Mater. Sci.: Mater. Electron. 32 (2021) 5602.

[10] Y. Xu, J. Zhang, Y. Wang, Food Chem. 398 (2022) 133939.

[11] J. Faghihi, X. Jiang, R. Vierling, S. Goldman, S. Sharfstein, J. Sarver, P. Erhardt, J. Chromatogr. A 915 (2001) 61.

[12] L. Nováková, L. Matysová, P. Solich, Talanta 68 (2006) 908.

[13] C. Tistaert, B. Dejaegher, Y. Vander Heyden, Anal. Chim. Acta 690 (2011) 148.

[14] S. Tokalıoğlu, F.K. Dokan, S. Köprü, LWT, 103 (2019) 301.

[15] I.M. Hwang, E.W. Moon, H.W. Lee, N. Jamila, K.S. Kim, J.H. Ha, S.H. Kim, Anal. Lett. 52 (2019) 932.

[16] R. Valand, S. Tanna, G. Lawson**,** L. Bengtström, Food Addit. Contam. Part A 37 (2020) 19.

[17] N. Rodrigues, F. Peres, S. Casal, A. Santamaria-Echart, F. Barreiro, A.M. Peres, J.A. Pereira, Food Chem. 398 (2023) 133945.

[18] M. Fathinezhad, M. Abbasi Tarighat, D. Dastan, Environ. Nanotechnol. Monit. Manage. 14 (2020) 100307.

[19] R. Sridevi, Glob. J. Comput. Sci. Technol. 10 (2019) 148.

[20] Y. Liu, R.H. Weisberg, Ch.N.K. Mooers, J. Geophys.

Res. 111 (2006) C05018.

[21] D. Dastan, N.B. Chaure, IJMMM (2014) 21.

[22] G.R. Lloyd, K. Wongravee, C.J.L. Silwood, M. Grootveld, R.G. Brereton, Chemom. Intell. Lab. Sys. 98 (2009) 149.

[23] A. Jafari, Kh. Tahani, D. Dastan, S. Asgary, Zh. Shi, X.T. Yin, W.D. Zhou, H. Garmestani, Ş. Ţălu Surf. Interfac. 18 (2020) 100463.

[24] A. Rezaei, M. Abbasi Tarighat, Kh. Mohammadi, J. Mater. Sci. Mater. Electron. 30 (2019) 13347.

[25] S.L. Panahi, D. Dastan, N.B. Chaure, Adv. Sci. Lett. 22 (2016) 941.

[26] W. Hu, T. Li, X. Liu, D. Dastan, K. Ji, P. Zhao, J. Alloys Comp. 818 (2020) 152933.

[27] A. Angelo Antonio, M.A. Maggi, Food Chem. 219 (2017) 408.

[28] S. Hassan, S. Ataolahi, Gh. Aliakbarzadeh, Sh. Zarre, Z. Poursorkh, J. Food Sci. Technol. 55 (2018) 1350.

[29] D. Ballabio, V.Consonni, R. Todeschini, Chemom. Intell. Lab. Syst. 98 (2009) 115.

[30] D. Ballabio, M.A. Vasighi, Chemom. Intell. Lab. Syst. 15 (2012) 24.

[31] N. Chavoshi, B. Hemmateenejad, Anal. Bioanal. Chem. Res. 2 (2022) 183.

[32] S.K. Oh, S.H. Yoo, W. Pedrycz, Expert Syst. Appl. 40 (2013) 1451.

[33] A. Fakhraei, M. Abbasi-Tarighat, Kh. Mohammadi, Gh. Abdi, A. Rezaei, Anal. Bioanal. Chem. Res. (2018) 317.

[34] R. Jannah, M. Rafi, R. Heryanto, A. Kautsar, D.A. Septaningsih, Int. Food Res. J. 25 (2018) 643.

[35] J. Mazina, M. Vaher, M. Kuhtinskaja, L. Poryvkina, M. Kaljurand, Talanta 139 (2015) 233.