1 **Using FTIR spectra and Pattern Recognition for Discrimination of Tea Varieties**

2 **,Jian-xiong Cai^a, Xiong-gang Xi^a, Yuan-feng Wang^a, Xin-lin Wei^{a*}**

3 a:Institute of Food Engineering, College of Life& Environment Science, Shanghai Normal University, 100 Guilin

4 Road, Shanghai 200234, PR China

5 *Corresponding author.

6 Tel:+862127463960.

7 E-mail address: foodlab2010@163.com (X.Wei)

8 Abstract

9 In order to **classify** typical Chinese tea varieties, **Fourier Transform infrared spectroscopy (FTIR)**
10 **of tea polysaccharides** (TPS) was used as an accurate and economical method. Partial Least Squares
11 (PLS) modeling method along with a Self-Organizing Map (SOM) neural network method **was**
12 utilized due to the diversity and heterozygosis between teas. FTIR spectra results of tea extracts after
13 spectra pre-processing were used as input data for PLS and SOM multivariate statistical analyses
14 respectively. The predicted correlation coefficient of optimization PLS model was 0.9994, root mean
15 square error of calibration (RMSEC) was 0.08914, and root mean square error of calibration and
16 cross-validation (RMSECV) was 0.03285. The features of PLS **can** be visualized in principal
17 component (PC) space, **contributing to** discover correlation **between** different class of spectra samples.
18 After that, a data matrix consisted of the scores on the selected 3PCs computed by principle
19 component analysis (PCA) and the characteristic spectrum data was used as inputs for training of
20 SOM neural network. **Compared** with the PLS linear technique's recognition rate of 67% only, the
21 correct recognition rate of the PLS-SOM as a non-linear classification algorithm to differentiate types
22 of tea **reaches** up to 100%. And the models **come** to be reliable and **provide** a reasonable clustering of
23 tea varieties.

24

25 **Keywords:** Tea Polysaccharides; Fourier Transform infrared spectroscopy; Spectra Preprocessing;
26 Artificial Neural Network; Partial Least Squares.

1. Introduction

Green tea, made from the leaves of *Camellia Sinensis L*, is one of the most popular beverages consumed worldwide; furthermore, it is of great importance in economy and medicine. [1-3] The chemical compounds of tea mainly contain polyphenols, polysaccharides, amino acids, vitamins. [4] It was reported that tea polysaccharides (TPS) played a key role in preventing the risk of diabetes, cardiovascular disease and cancer. [5-6] Improved understanding of chemical and bioactive studies of TPS, the polysaccharides is reported to possess antioxidant [6-7], antibacterial [8], anticoagulant [9], antitumor [10], and immunological activities [11]. It has been proved that these polysaccharides are safe and non-toxic [12]. TPS have become the greatly important bioactive substances after other active components found in tea.

With tea market expanding rapidly, the required quality of tea is higher than before. And the diversity of tea also plays an critical role in the tea polysaccharides industry.[13] Therefore, to establish an accurate and rapid identification of tea varieties becomes an important assignment for analyzers, agricultural institutions and most business, and it is also necessary for practical breeding purposes.

Recently, several prior studies have been developed to classify tea varieties based on pattern recognition techniques applied to chemical composition data, such as high-performance liquid chromatography (HPLC) [14,15], gas chromatographic (GC)[16], capillary electrophoresis[17], plasma atomic emission spectrometry[18], etc. However, analytical methods are sophisticated and expensive to be applied in routine works in some cases. [19] Generally, employing professional tea tasters has been worked, but it takes a considerable amount of time and money to cultivate special talents. [20] Moreover, tea plants are not only highly diverse and heterozygous, but also many overlapping morphological, biochemical and physiological features. [21]

In the view of which, a new, accurate and reliable pattern recognition method for identification of tea varieties is important for practical purposes. FTIR analytical technique has been proved to be a powerful analytical tool because of its low cost and being non-destructive. FTIR spectra have been used to discriminate biodiesel [22], Ghana cocoa beans [23], oat [24], etc. The modern chemometrics combined with FTIR spectra enjoys higher level of efficiency, and has also been applied to the field of food as well as pharmacy industry.

Great advances have been made in the quantification models of TPS developed by PLS model recently.[25-29] PLS method is a quantitative spectral decomposition technique related to principle component regression (PCR), and it is also one of the most common regression algorithms in the field of chemometrics in general, particularly for spectroscopy. Through PLS, we aimed to describe the linear relationship between a response variable y , and a set of explanatory variables. In this process, PLS actively used the variables included in the concentration matrix as the data matrix, called “scores”. Then PLS extracted a set of orthogonal factors called latent variables (LVs) and evaluated by the optimal function of RMSECV, RMSEC, root mean square error (RMSE), and Correlation Coefficient. PLS is particularly suitable when the number of variables is more than that of samples and when there is multicollinearity among independent variables (spectra). [30]

Artificial neural network is an intelligent non-linear science algorithm, which could be effectively applicable to classification. Back Propagation-Artificial neural network (BP-ANN) model is a powerful study system, and can highly actualize difference between varieties non-linearity mapping, [19,31,32] and Self-Organizing Map (SOM) neural network is an unsupervised self-organizing competition network[33], which is used in the field of economy and engineering currently, such as tea aroma [34], tea granules[35]. SOM network can extract the internal features of parameter by self-organizing

and reflect them on the classified map. And it can be used in solving problems of pattern recognition. However, there have been few reports on employing SOM for classification of Chinese tea varieties depended on TPS using FTIR.

In the current study, one of the main objectives was to investigate the relationship between the spectral reflectance characteristics of a tea plant and its varieties. Spectral pre-treatment methods were considered to eliminate noise and optimize the length scale of spectrum. [36] Afterward, the application of FTIR spectra and pattern recognition methods was proposed to differentiate the four main tea varieties. The characteristic absorption peaks was to build and train SOM model to rapidly discriminate 6 accessions of tea, selected from the widely diffused types in three provinces in China.

2. Materials and methods

2.1 Sample preparation and preprocessing

Six typical samples of tea were collected from three different major producers of green tea regions plucked in different season (Hubei, Fujian, Guizhou provinces of China) (Table 1), before the plant materials were stored in Institute of Food Engineering, Shanghai Normal University. Green tea leaves were dried and passed through a 30-mesh sieve to make uniform powder (0.600mm). Each powder (150 g) was immersed in 95% ethanol (v/v) three times every two hours to remove most of the polyphenols, pigments and then extracted with water (1:8, m/v) in bath at 90 °C for twice every four hours. After the water extract was filtered, the combined filtered solutions were centrifuged (10 min, 8000 rpm) to remove the contaminants. Then the supernatant was concentrated in a rotary evaporator and dialyzed. After that the filtrate was precipitated with 75% ethanol. The dialyzed solution was freeze-dried to yield polysaccharides powder. [37] All chemicals were used with analytical-grade from Sinopharm Chemical Reagent Co. (Shanghai, China) and doubly distilled water was used throughout

the work.

2.2 Analytical methods of components in polysaccharide samples

The total sugars were determined according to the phenol-sulphuric acid method with D-glucose as standard. [38] The soluble protein (SP) was determined by the Coomassie brilliant blue G-250 method with bovine serum albumin as a standard. [39]

2.3 Spectra acquisition

A Nexus FT-IR 5700 spectrometer (Nicolet, USA) with fibre optic accessory in reflectance mode, equipped with a DTGS detector, was used for spectral scanning in the range of 4000-400 cm^{-1} at a resolution of 4 cm^{-1} . Each sample (2 mg) was incorporated into 200 mg of KBr (spectroscopic grade) powder to be pressed into 1 mm pellets for FT-IR measurement. The interference of H_2O and CO_2 was minimized when scanning. The absorbance spectra were obtained using OMNICESPV6 software. [40]

2.4 Spectra preprocessing

Data preprocessing is essential for enhanced performance of the classifiers. To reduce the effect of noise, remove redundant information and enhance sample-to-sample differences, data pre-processing techniques as follows: (1) Raw data, (2) Multiple scatter correction (MSC), (3) Savitzky-Golay smoothing filter (Savitzky-Golay), (4) Standard normal variates (SNV), (5) Savitzky-Golay followed by MSC, (6) Savitzky-Golay followed by SNV. MSC technique was used to correct for additive and multiplicative effects in the spectra. Savitzky-Golay was employed to provide noise reduction. SNV was originally designed to reduce scattering effects in the spectra but was also proved to be effective in correcting the interference caused by variations in pellet thickness or optical path. [41, 42]

2.5 Partial least squares model

The partial least squares model (PLS) is assessed to describe the linear relationship between a

response variable, y , and a set of explanatory variables, X [43-45]. The PLS model taken as the basic classification learner **helpfully deals with** the multicollinearity problem. This is achieved by replacing the original variables with a few orthogonal latent factors, T , with maximal covariance with y . [25] The PLS model can be presented as:

$$\mathbf{X} = \mathbf{T}\mathbf{P}^T + \mathbf{E}$$

$$\mathbf{y} = \mathbf{T}\mathbf{q} + \mathbf{r} = \mathbf{X}\mathbf{b} + \mathbf{r}$$

Where $\mathbf{X}$ is the data matrix with m rows and n columns, $\mathbf{q}$ holds f regression coefficients associated with f PLS factors, $\mathbf{T}$, while $\mathbf{E}$ represents the differences between the observed and predicted $\mathbf{X}$ and $\mathbf{r}$ holds the differences between the observed and predicted $\mathbf{y}$. The regression coefficients, $\mathbf{b}$, are obtained as:

$$\mathbf{b} = \mathbf{W}(\mathbf{P}^T\mathbf{W})^{-1}\mathbf{q}$$

Where $\mathbf{W}$ is the matrix of loadings maximizing the covariance criterion, and $\mathbf{P}$ is equal to $\mathbf{X}^T\mathbf{X}\mathbf{W}$.

2.6 Evaluation of model complexity, its fit and prediction abilities

Selection of the optimal number of factors is crucial in constructing a calibration model. Normally, model complexity is estimated using the cross-validation procedure. [46] In this way, n values of RMSECV, for models with different complexity, are obtained. [47] The RMSECV values are reported as the average value of the obtained RMSECVs for n models with f factors:

$$\text{RMSECV}(f) = \sqrt{\frac{\sum_{i=1}^m \left(y_i - \hat{y}_{-i} \right)^2}{m}}$$

Where y_i is the experimental value of the dependent variable of the i th object, and $\hat{y}_{-i}$ is the predicted value of the dependent variable of the i th object from the model built without the i th sample.

In order to characterize model fit and its prediction abilities, usually root mean square error

136 (RMSE) are calculated:

$$137 \quad \text{RMSE}(f) = \sqrt{\frac{\sum_{i=1}^m \left(y_i - \hat{y}_i \right)^2}{m}}$$

138 Where y_i and $\hat{y}_i$ correspond to the actual and predicted adulteration levels of sample i .

139 2.7 Self-Organizing Maps (SOM)

140 SOM algorithm, widely applied to clustering problems and data exploration in industry, medicine
141 finance, and natural sciences, is an unsupervised neural network method of ANN. [48]SOM has
142 properties of both vector quantization and vector projection algorithms. The prototype vectors are
143 positioned on a regular low-dimensional grid in an ordered fashion, making the SOM a powerful
144 visualization tool. SOM usually consists in two layers of neurons: an input layer is connected to each
145 vector of the data set (i.e., one vector is infrared absorption characteristic peaks of a real tea sample)
146 and an output layer forms a two-dimensional array of neurons arranged on a hexagonal lattice. Every
147 output neuron competes with others to become the winning neuron. When SOM is employed in judging
148 which cluster a new input point belongs to, the point is assigned to the cluster of the winning neuron.
149 [49]The learning process of the SOM goes as follows:

150 (1) .One sample vector X_k : $X_k = \{X_{1k}, X_{2k}, \dots, X_{nk}\}$ is randomly drawn from the input data set. And
151 network weights initially assigned random values between +1 and -1.

152 (2) .Its similarity (distance: d_{jk}) to the codebook vectors is computed by using the common Euclidean
153 distance measure:

$$154 \quad \|X_k - W_c\| = \min_j \{d_{jk}\}$$

155 (3) .The update rule for the reference vector of unit i is the following:

$$w_{ij}(t+1) = w_{ij}(t) + \eta(t)(x_i - w_{ij}(t))$$

The magnitude of the attraction is governed by $\eta(t)$ (the learning rate).

(4) .After the Best Matching Unit (BMU) has been found, the codebook vectors are updated.

The BMU itself as well as its topological neighbors are moved closer to the input vector in the input space i.e. the input vector attracts them. As the learning proceeds and new input vectors are given to the map, the learning rate gradually decreases to zero according to the specified learning rate function type. Along with the learning rate, the neighbourhood radius decreases as well. The steps 1 and 2 together constitute a single training step and they are repeated until the training ends.

In the present work, some important parameters were chosen as follows: The number of training steps was set at 1000 to train the SOM, and the rate of convergence in the neighbourhood function and the learning rate was 0.1, respectively. In the output layer, the units of the grid (virtual tea samples) give a representation of the distribution of the tea samples in an ordered way as 1~4.

2.8 Statistical Analysis

Matlab 2013a programs were used in the statistical analysis. The significance of differences among values was analyzed by one-way analysis of variance (ANOVA) followed by Bonferroni post hoc tests. P -values < 0.05 were considered statistically significant. Data are expressed as mean $\pm$ SD.

2.9 Software

All methods were performed in Matlab 2013a (MathworksCo., USA) for Windows 2007 and The Unscrambler, Ver 9.7.CAMO Software. For the spectral acquisition, Result Integrating Program (Thermo Electron Result Integration and Result Operation Software, Thermo Electron Corporation, USA) was used. All calculations were performed using a computer equipped with an Intel Pentium® III-2GHz processor.

3. Results and discussion

3.1 Components of polysaccharide samples

As shown in Table1, the total sugars and proteins content of TPS were 52.36~59.48% and 0.01~13.29%, respectively. In the phenol-sulphuric acid assay, The Contents of Protein of Small Leaf Tai cha (SLTCSP) were significant than others. ($p=0.03<0.05$)The Contents of total Sugar of Small leaf Tai cha (SLTCA) were significant than others. ($p=0.01<0.05$)Remarkable content level was detected in the total sugars content of samples compared with proteins content ($p=0.003<0.05$), that is, the extractions can be recognized as polysaccharides by this method.

Generally, the main bioactive components of tea vary with species, season, and age of the leaf (plucking position), climate, and horticultural practices. And Chinese green tea can break down into two basic types: small leaf tea and the large one. There are abundant lutein, carotene, and TPS in the Small leaf tea which can be made to be high aroma of tea and high content of Tea polyphones (TPP) and TPS in the big leaf tea, which both can reduce the risk of developing cancer. “Fudingdahaocha”, which is originated in Wang Jiayang Village of Fujian Province, mainly consisted of TPP and TPS, and significantly suggested to be applied as nutraceuticals, and functional foods. In this work, TPS were prepared by the same process but its polysaccharides content cannot be considered as a main basis for evaluation. Therefore, we investigated distinctive infrared spectrum features and pattern recognition to discriminate tea varieties.

3.2 IR spectra fingerprint of polysaccharide samples

The vibrations of the functional groups of the compounds were discussed as to their homologous infrared absorption characteristic peaks. [50, 51] The examples we used are NIR. As shown in Fig. 1., the sharp absorption bands of raw data spectrum seen at 2929 cm^{-1} , $1700\sim1460\text{ cm}^{-1}$, $1400\sim1200\text{ cm}^{-1}$,

and 800–500 cm^{-1} , are indicative of compounds containing TPS. These bands can be partly assigned to polysaccharides presented in green, black, or oolong tea. The FT-IR spectra of the tea polysaccharides showed the characteristic strong broad absorption in the range of 1659~1600 cm^{-1} corresponding to stretching vibration of sugar rings. The band at 1431.87 cm^{-1} gave evidence for a weak C-H stretching vibration, absorption at around 1643 cm^{-1} indicated N-H vibration or C-O asymmetric vibration of carboxyl group, and the band between 1000 and 1100 cm^{-1} , i.e. was characteristic for the presence of tea polysaccharide due to O-substituted glucose residues. The band in the region of 1740.28 cm^{-1} was acetyl and the absorption peaks at around 1643.12 cm^{-1} and 1416.53 cm^{-1} were the characteristic IR absorptions of protein, which showed that the TPS were a glycoprotein containing polysaccharide and protein.

3.2 Spectral pre-processing of the PLS model justification

PLS models were constructed using the data from 6 typical samples of Green tea, based on the full FTIR spectra approach and employing the different. The pretreatments that provided a surprisingly vital improvement in model performance, a decrease in RMSECV values coupled with an increase in correlation values, following: Savitzky-Golay、MSC、SNV. The model based on Savitzky-Golay/SNV data preprocess was considered to be satisfactory, and because its Correlation Coefficient (0.9994) significantly higher than others, RMSE (0.0330) and lower RMSECV (0.03285) (Table 2). SNV only was not effective, although it improved model performance. Therefore, it was chosen as the most appropriate of the pre-treatment schemes tested.

3.3 Results of PLS model

PLS allowed an overall view of the differences among tea samples after using the optimizing spectrum data. The first three principal components explained 99.03% of total variability, which could nearly cover information of tea samples, and the first two PLS scores explained 97.40% of total

variability. The PLS analysis clearly identified different tea samples from FTIR spectra, although they followed some parallel trajectories (see Fig. 2(A)). But as Fig.2 (B) showed, PLS model fails to distinguish the types of tea by measuring the original spectrum data. Fig.2 (A) showed the first principal component (PC1), explained 93% of total variability, clearly distinguish “the Small leaf Tai cha”, “Local small leaf tea”, and “Local big leaf tea” from the others. Normally, the quality characteristic of tea varies with not only species, but also season. Thus, PC1 also recognized the differences in season of “the Small leaf Tai cha” as well as “Fudingdahaocha”. The second principal component (PC2) explained 4%, separates “Local small leaf” from “Local big leaf”, and different season of “the Small leaf Tai cha”. The results demonstrated that the preprocessing and transforming of the spectrum data is an effective and reliable method for our research once again. However, FDDHCS and LSLCSM did not accurately differentiate via examination of PCs loadings.

PLS assumes the spectra to be a linear function of scores. For avoidance of complexity of the model and over-fitting, 3 latent factors (3PCs) which were linear combinations of the original variables were selected. After PCA analysis, by removing the region a absorption of water in the Factor 1, the results showed that the $781\text{-}1905\text{ cm}^{-1}$ and $2280\text{-}2450\text{ cm}^{-1}$ spectral regions with 670 data point have high correlation coefficients and better result for real samples, as shown in Fig.3. Afterwards, a data matrix of scores contained 3PCs and the scores on the selected PCs were subjected to ANN as input.

3.4 Results of SOM model

Based on the data-set, we focused on extracting and encoding the most meaningful information from FTIR spectroscopy to use as the input matrix in ANN. As shown in Table 3, Cluster 1 contained “the Small leaf Tai cha” which mainly caused by different kinds of tea no matter plucking time different. In the cluster 2, “Fudingdahaocha” is positioned from the other accessions; Cluster 3

belonged to the so-called Chinese ‘small leaf’ type and Cluster 4 included the majority of the “Local big leaf tea”. In this model, computing is simple; learning is fast, the running time in only 15-20 seconds. The PLS-SOM recognition rate for tea types differentiation was 100%, and the result showed that the recognition rate of nonlinear SOM-ANN model was much better linear PLS models.

Great advances have been made in classification studies of tea varieties in recent decades. It was reported that the FTIR analysis and the creation of a BP neural network through infrared spectra were the capacity to characterize, univocally discriminate and associate most of the accessions in the correct clusters without a molecular or morphological analysis, as previously done by LUO et al.[52]However, the limitations of the BP-ANN method are extremely the quantity, validity, and accuracy of training data, namely, neural networks train best and learn to generalize best when presented data rich in variation. [53]In contrast to back-propagation neural network (BPNN) using supervised learning, the SOM is based on unsupervised learning, especially if we don't know sample varieties in details. [54]In other words, the SOM can also be available to apply to other tasks which is unclassified data.

4. Conclusion

On the basis of the results above, these two algorithms (PLS-SOM) tried to locate probable clusters within the data in order to find out the most adequate classification model. SOM-ANN was proven to be a powerful data-modeling tool to capture and represent complex correlation between inputs and outputs. PLS could also play an important role in different types of tea. The results also showed that TPS were suited to preprocess the FTIR spectra of tea samples, and data processing algorithms can play a vital role in correcting estimate of results about probable cluster positions within data. Then an appropriate approach (Original spectrum /Savitzky-Golay/SNV) discovered for data visualization of complex datasets. PLS results showed that the first principal component (PC1) was

able to distinguish “the Small leaf Tai cha” from other tea samples and the second principal component (PC2) separate **dearly** “Local small leaf” from “Local big leaf”. But the “Fudingdahaocha” are not properly classified into categories. The PLS recognition rate of 67% was achieved. Although PLS model is possible to correctly tell a variety of tea, this isn't always turn out so well. It really cannot achieve our research aims effectively. Also single detection method has some shortcomings, **so that it** not enough to meet the needs of the huge variety of teas test. And then the PCA showed that the spectral features were integrated into 3PCs and their additive contributing rate came up to 99.03%. A data matrix of scores contained 3PCs and the scores on the selected PCs were subjected to ANN as input. Taking the learning parameter as 0.1 and the training iteration number as 1000, the PLS-SOM recognition rate to differentiate types of tea was 100%.Through a combination of two methods, it was higher than PLS method and PCA alone. This research attempted to explore an effective way to classify tea samples. In Conclusion, in our case, the creation of an artificial neural network along with PLS based on FTIR spectra can lead to an effective tea varieties recognition, even though the choice of the tea leaf materials must be healthy and well-developed. Moreover, the study has been found that the spectral preprocessing and analyzing could provide the scientific foundation for the related spectroscopy applications. Further studies can be directed so as to examine the effect of the different structures on TPS of tea varieties, as the current study is essentially devoted to the characterization of TPS in the acidic polysaccharides or neutral polysaccharides. And also using different kinds of neural network in combination maybe a better way for cultivar discrimination.

Acknowledgements

The authors are grateful for financial sponsored by Innovation Program of Shanghai Municipal Education Creative Commission (12ZZ125,11ZZ121),National High Technology Research and

289 Development(863) Program of China(2008AA10Z322),Shanghai Rising-Star Tracking Program
290 (11QH1401800), Shanghai Yangtze River Delta Science Joint Efforts Program (12495810600), the
291 Program of Food safety and Nutrition Team of Shanghai Normal University(DXL123), Local colleges
292 and universities capacity building Program (13430502900) and Development Center of Plant Germ
293 plasm Resources. Thanks are due to Yang-yang Liu for assistance with the experiments and to Yu Chao
294 for correcting manuscript. And we also thank Professor Bao-rong Lu in Fudan University and Professor
295 Kun-lin Jin in Department of Pharmacology and Neuroscience University of North Texas Health
296 Science Center for helpful conversations.

297 **References**

- 298 [1]C.S. Yang, P Maliakal, X. Meng, (2002) .Inhibition of carcinogenesis by tea, Annual Review of
299 Pharmacology and Toxicology. 42, 25-54.
- 300 [2]V. Chopade, A. Phatak, A.B. Upaganlawar, A.A. Tankar, (2008). Green tea (*Camellia sinensis*):
301 Chemistry, traditional, medicinal uses and its pharmacological activities: A review, Pharmacog.
302 Rev, 2,157-162.
- 303 [3]V. Crespy, G. Williamson, (2004). A Review of the Health Effects of Green Tea Catechins in In Vivo
304 Animal Models. Journal of Nutrition. 134(12), 3431-3440.
- 305 [4]Y. K. Kim,Y. J.Oh, J. O. Chang,S. J. Lee, K.O. Kim, (2009). Chemical Composition of Green Teas
306 According to Processing Methods and Extraction Conditions, FOOD SCIENCE AND
307 BIOTECHNOLOGY. 18(5), 1212-1217.
- 308 [5]X. Zhou, D. Wang, P. Sun, P. Bucheli, L. Li, Y. Hou, J. Wang, (2007). Effects of soluble tea
309 polysaccharides on hyperglycemia in alloxan-diabetic mice, J Agric Food Chem., 55(14), 5523-5528.
- 310 [6]J.J Yang, B. Chen, G. Yan, (2012). Pharmacological evaluation of tea polysaccharides with
311 antioxidant activity in gastric cancer mice, Carbohydrate Polymers. 90(2), 943–947.
- 312 [7]Z.Y. Zhao, L.T. Huangfu, L.L. Dong, S.L. Liu, (2014). Functional groups and antioxidant activities
313 of polysaccharides from five categories of tea, Industrial Crops and Products. (58), 31–35.
- 314 [8]J.H. Lee,J.S. Shim,J.S. Lee, J.K. Kim, I.S. Yang,M.S. Chung, K.H. Kim, (2006). Inhibition of
315 Pathogenic Bacterial Adhesion by Acidic Polysaccharide from Green Tea (*Camellia sinensis*). Journal
316 ofAgricultural and Food Chemistry.54 (23), 8717-8723.
- 317 [9]W.R.Cai,L.L.Xie,Y. Chen,H. Zhang,(2013).Purification, characterization and anticoagulant activity
318 of the polysaccharides from green tea, Carbohydrate Polymers. 92(2), 1086-1090.

319 [10]R.J. Xu,H. Ye, Y. Sun, Y.YTu, X.X. Zeng,(2012).Preparation, preliminary characterization,
320 antioxidant, hepatoprotective and antitumor activities of polysaccharides from the flower of tea plant
321 (*Camellia sinensis*), Food and Chemical Toxicology.50 (7), 2473-2480.

322 [11]M.Monobe,K.Ema,F. Kato, M.Y. Mari, (2008).Immunostimulating Activity of a Crude
323 Polysaccharide Derived from Green Tea (*Camellia sinensis*) Extract, Journal of Agricultural and Food
324 Chemistry.56 (4), 1423-1427.

325 [12]H.X. Chen, M. Zhang, Z.S. Qu; B.J. Xie, (2007). Compositional analysis and preliminary
326 toxicological evaluation of a tea polysaccharide conjugate, J. Agric. Food Chem. 55, 2256-2260.

327 [13]X.L Li, H. Yong, (2008). Discriminating varieties of tea plant based on Vis/NIR spectral
328 characteristics and using artificial neural networks, Biosystems Engineering.99 (3):313-321.

329 [14]P. Valera, F. Pablo, A.G. Gonzalez, (1996).Talanta 43, 415–419.

330 [15]Y.G. Zuo, H. Chen, Y.W. Deng, (2002). Talanta 57, 307–316.

331 [16]N. Togari, A. Kobayashi, T. Aishima, (1995).Food Res. Int. 28, 495–502.

332 [17]H. Horie, T. Mukai, K. Kohata, (1997) J. Chromatogr. A 758, 332–335.

333 [18]M. Angeles Herrador, A. Gustavo Gonzalez, (2001). Talanta 53, 1249–1257.

334 [19]A. Palacios-Morillo,A.Alcázar,F. de Pablos, J.M. Jurado,(2013). Differentiation of tea varieties
335 using UV-Vis spectra and pattern recognition techniques, SpectrochimicaActa Part A: Molecular and
336 Biomolecular Spectroscopy.103, 79-83.

337 [20]L.G.Gwénaëlle,J.C. Ian, M. Defernez, (2004). Metabolite Profiling Using ¹H NMR Spectroscopy
338 for Quality Assessment of Green Tea, *Camellia sinensis*(L.), Journal of Agricultural and Food
339 Chemistry.52(4),692-700.

340 [21]K.C.Willson,M.N. Clifford, (1992). Tea: cultivation to consumption, Chapman & Hall, London.

341 [22]G.Veras, Ade. A. Gomes,A.C.da Silva, A.L. de Brito,P.B. de Almeida,E.P. de Medeiros,
 342 (2010).Classification of biodiesel using NIR spectrometry and multivariate techniques, Talanta.83 (2),
 343 565.

344 [23]E.Teye,X.Y. Huang,H. Dai,Q.S. Chen, (2013). Rapid differentiation of Ghana cocoa beans by
 345 FT-NIR spectroscopy coupled with multivariate classification, SpectrochimicaActa Part A: Molecular
 346 and Biomolecular Spectroscopy.114, 183-189.

347 [24]S.Serranti,D. Cesare, F.Marini, G. Bonifazi, (2013).Classification of oat and groat kernels using
 348 NIR hyperspectral imaging, Talanta.103, 276-284.

349 [25]M. Daszykowski, Y. Vander Heyden, B. Walczak, (2007).Robust partial least squares model for
 350 prediction of green tea antioxidant capacity from chromatograms,J Chromatogr A.1176(1-2), 12-18.

351 [26]J. Zhou, H. Cheng, J.M. Zeng, (2010).Study on Identification and Traceability of Tea Material
 352 Cultivar by Combined Analysis of Multi-Partial Least Squares Models Based on Near Infrared
 353 Spectroscopy, SPECTROSCOPY AND SPECTRAL ANALYSIS.30 (10), 2650-2653.

354 [27] S.M. Tan, R.M. Luo, Y.P. Zhou, (2012). Boosting partial least-squares discriminant analysis with
 355 application to near infrared spectroscopic tea variety discrimination, JOURNAL OF
 356 CHEMOMETRICS. 26(1), 34-39.

357 [28] W. He, J. Zhou, H. Cheng, L. Wang, K. Wei, W. Wang, X. Li, (2012).Validation of origins of tea
 358 samples using partial least squares analysis and Euclidean distance method with near-infrared
 359 spectroscopy data. SPECTROCHIMICA ACTA PARTA-MOLECULAR AND BIOMOLECULAR
 360 SPECTROSCOPY. 86,399-404.

361 [29]S. Wang, Z. Zhang,J.Ning, (2013). Back Propagation-Artificial Neural Network Model for
 362 Prediction of the Quality of Tea Shoots through Selection of Relevant Near Infrared Spectral Data via

363 Synergy Interval Partial Least Squares, ANALYTICAL LETTERS. 46(1), 184-195.

364 [30]D. Wu, J.Y. Chen,B.Y. Lu, L.N. Xiong, Y. He, Y. Zhang, (2012).Application of near infrared
 365 spectroscopy for the rapid determination of antioxidant activity of bamboo leaf extract, Food
 366 Chemistry.135 (4), 2147-2156.

367 [31]Q. Chen,J. Zhao,M. Liu, J. Cai, (2008). Nondestructive Identification of Tea (*Camellia Sinensis*L.)
 368 Varieties Using FT-NIR Spectroscopy and Pattern Recognition, Czech Journal of Food Sciences. 26(5),
 369 360-370.

370 [32]J.Zhao,Q.Chen,X. Huang, C.H. Fang,(2006). Qualitative identification of tea categories by near
 371 infrared spectroscopy and support vector machine, Journal of Pharmaceutical and Biomedical
 372 Analysis.41 (4), 1198-1204.

373 [33]A.M. Kalteh,P. Hjortha, R. Berndtssona, (2008). Review of the self-organizing map (SOM)
 374 approach in water resources: Analysis, modelling and application, Environmental Modelling and
 375 Software.23 (7), 835-845.

376 [34]S. Borah, E. L. Hines, M. S. Leeson, D. D. Iliescu, M. Bhuyan, J. W. Gardner (2008). Neural
 377 network based electronic nose for classification of tea aroma, Sensing and Instrumentation for Food
 378 Quality and Safety.2 (1), 7-14.

379 [35]S. Boraha,E.L. Hinesa, M. Bhuyanb, (2007).Wavelet transform based image texture analysis for
 380 size estimation applied to the sorting of tea granules, Journal of Food Engineering.79(2), 629-639.

381 [36]Y. D. Bao, Y. P. Wu, Y. He, (2004). Optimal mix forecasting method based on BP neutral network
 382 and its application, Journal of Agricultural Mechanization Research.3, 162-164.

383 [37]Y.F. Wang, F.F. Mao, X.L. Wei, (2012). Characterization and antioxidant activities of
 384 polysaccharides from leaves, flowers and seeds of green tea, Carbohydrate Polymers.88 (1), 146-153.

385 [38]M. Monobe, K. Ema, F. Kato, M. Maeda-Yamamoto, (2008). Immunostimulating activity of a
386 crude polysaccharide derived from green tea (*Camellia sinensis*) extract, *Journal of Agricultural and*
387 *Food Chemistry*.56 (4), 1423-1427.

388 [39]M. Tatsuga, M. Akio, I. Norimasa, M. Tokifumi, N. Shin-Ichiro, Y. C. Lee, (2005).Carbohydrate
389 analysis by a phenol-sulfuric acid method in microplate format, *Analytical Biochemistry*.339(1), 69-72.

390 [40]Y. Wang, H. Jiang, X. Xi, X. Wei, (2013).Multi-fingerprint and quality control analysis of Tea
391 Polysaccharides, *Carbohydrate Polymers*.92(1),583-590.

392 [41]L. Xu, D.H. Deng, C.B. Cai, (2011).Predicting the Age and Type of Tuocha Tea by Fourier
393 Transform Infrared Spectroscopy and Chemometric Data Analysis, *Journal of Agricultural and Food*
394 *Chemistry*.59 (19),10461-10469.

395 [42]Z. Boris, K.Achim, (2013). Optimizing Savitzky-Golay Parameters for Improving Spectral
396 Resolution and Quantification in Infrared Spectroscopy, *Applied Spectroscopy*.67 (8), 892-902.

397 [43]H. Martens, T. Næs, (1989). *Mutivariate Calibration*, Wiley, Chichester.

398 [44]T. Næs, T. Isaksson, T. Fearn, T. Davies, (2002).*Multivariate Calibration and Classification*, NIR
399 Publications, Chichester,

400 [45]S. Wold, M. Sjostrom, L. Eriksson, (2001).*Chemom. Intell. Lab. Syst.* 58,109.

401 [46]T. Næs, T. Isaksson, T. Fearn, T. Davies, (2002). *Multivariate Calibration and Classification*, NIR
402 Publications, Chichester.

403 [47]V. Saritporn,J.W.Zhao,Q.S. Chen, C. Rachata,(2008).Application of FT-NIR Spectroscopy to the
404 Measurement of Fruit Firmness of “Fuji” Apples, *Maejo International Journal of Science and*
405 *Technology*.2 (1), 13-23.

406 [48]T. Kohonen, (2013) *Essentials of the self-organizing map*, *Neural Netw*.37, 52-65.

407 [49]Y.K. Melody,Y. H. Michael, M.F. Dorothy, (2007). The effect of sample size on the extended
 408 self-organizing map network-A market segmentation application, Computational Statistics and Data
 409 Analysis.51 (12), 5940-5948.

410 [50]Y.F.Wang, X.L.Wei, Z.Y. Jin. (2009)Structure analysis of an acidic polysaccharide isolated from
 411 green tea. NATURALPRODUCT RESEARCH.23 (7):678-687.

412 [51]L.M. Wang, W.S. Xia. (2006) Isolation and analysis of a novel acidic polysaccharide with
 413 glucokinase-stimulating activity from coarse green tea. JOURNALOF FOOD BIOCHEMISTRY.30
 414 (2):187-202.

415 [52]Y.F. LUO,Z.F. GUO, Z.Y. ZHU,C.P. WANG, H.Y. JIANG, B.Y. HAN, (2005).Studies on ANN
 416 Models of determination of tea Polyphenol and polysaccharide in tea by Near-Infrared
 417 Spectroscopy.Spectroscopy and spectral Analysis. 25(8), 1230-1233.

418 [53]P. Camilla, M. Sergio, A. Elisa, B. Silvia, M. Elisa, M. Stefano.(2009)Artificial neural networks as
 419 a tool for plant identification: a case study on Vietnamese tea accessions.Euphytica. 166(3), 411–421

420 [54]Ritaban Dutta, E.L. Hines, J.W. Gardner, K.R. Kashwan, M. Bhuyan.(2003).Tea quality prediction
 421 using a tin oxide-based electronic nose: an artificial intelligence approach, Sensors and Actuators B.94
 422 (2), 228-237.

423 Table 1 the 6 selected tea accessions collected from the Institute of Food Engineering (n=3)

Abbr.	Name	Plucking Time	Contents of Protein (%)	Contents of total Sugar (%)
SLTCA	Small leaf Tai cha	Autumn	5.00 ± 1.40	59.48 ± 0.13
SLTCSP	Small leaf Tai cha	Spring	13.29 ± 2.70	52.36 ± 1.08
FDDHCA	Fuding dahaocha	Autumn	7.78 ± 3.70	52.63 ± 0.24
FDDHCS	Fuding dahaocha	Spring	0.64 ± 0.53	52.38 ± 0.23
LSLCSM	Local small leaf	Summer	2.03 ± 0.07	54.89 ± 0.15
LBLCSP	Local big leaf	Spring	0.01 ± 2.52	54.76 ± 0.77

424 The abbreviation of tea is typically capitalized of name and plucking time.

425 Table 2 Performance results of full spectrum PLS models based on different data preprocessing techniques.

Mathematical Treatment	Correlation Coefficient	RMSECV
Original spectrum	0.9981	0.3709
Original spectrum + Savitzky-Golay	0.9007	1.0478
Original spectrum + MSC	0.9952	0.3067
Original spectrum + SNV	0.9960	0.07131
Original spectrum + Savitzky-Golay + MSC	0.9993	0.1267
Original spectrum + Savitzky-Golay + SNV	0.9994	0.03285

426 Mathematical Treatment: plus sign means try changing the order of pre-processing.

427

Table3. The recognition results of predictive samples by SOM-ANN

Abbr.	Name	Output result
SLTCA	Small leaf Tai cha	1
SLTCSP	Small leaf Tai cha	1
FDDHCA	Fudingdahaocha	2
FDDHCS	Fudingdahaocha	2
LSLCSM	Local small leaf	3
LBLCSP	Local big leaf	4

428

Figure captions

Fig.1. Raw FTIR spectra data of tea samples. The dark blue line、red line、 green line、 light blue line、 scarlet line、 gray line correspond to SLTCA, SLTCSP, FDDHCA, FDDHCS, LSLCSM, LBLCSP, respectively.

Fig.2. plots of different kinds of tea (*Camellia sinensis* L.) after PCA. Fig. (2A), (2B) correspond to original spectrum data (Savitzky-Golay+SNV) and raw data spectral preprocessing, respectively.

Fig. 3.PC1、 PC2、 PC3 across the spectral region.

436 Fig. 1.

Fig. (2A)

444 Fig. (2B)

445

446

447 Fig.3.

448

449