Qualitative and Quantitative Identification of Components in Mixture by Terahertz Spectroscopy

Yan Peng, Chenjun Shi, Mingqian Xu, Tianyi Kou, Xu Wu, Bin Song, Hongyun Ma, Shiwei Guo, Lizhuang Liu, and Yiming Zhu

***Abstract*—Typical methods for the analysis of mixture components include multiple linear regression, partial linear squares, and artificial neural network. However, these methods need large amount of samples and time to improve recognition accuracy. In this paper, based on the data obtained from terahertz spectroscopy, an identification method with less sample requirements and lower calculation time but higher accuracy is proposed. Based on the wavelet transform, baseline elimination, support vector regression, and loop iteration of samples, the specific substance in the mixture can be identified effectively. For example, seven substances that exist in brain glioma are chosen as the components of a mixture, where the key substances used for glioma diagnosis are set as the target substances and the spectra of mixtures with different mix proportions serve as training data. The average correlation coefficient of identification achieves 99.135% and the root-mean-square error is 0.40%. These results have profound implications for the eventual practical application of exact qualitative and quantitative identification of components in mixtures.**

***Index Terms*—Mixture identification, terahertz (THz) spectroscopy.**

I. Introduction

TERAHERTZ (THz) radiation is the electromagnetic radiation in the frequency interval from 0.1 to 10 THz. Due to its low photon energy (4 meV @ 1 THz), THz wave offers the advantages of being noninvasive and nonionizing, thus presenting little harm to biological tissue [1]–[4]. Combining good accuracy, sensitivity, and rapidity, THz spectroscopy has been widely used in identifying substances, especially cancer tissues. Recently, the qualitative and quantitative analyses of mixtures using THz spectroscopy became an important research topic [4]–[7]. However, in those cases, such as multiple linear regression and least square method, the spectrum of each component in the mixture has to be known in advance (single-spectrum-known pattern), which is infeasible for the identification of complicated mixtures consisting of over hundreds of substances. While for some mixture identification methods that do not need to measure the spectrum of each mixture component in advance, such as partial least squares (PLS) and principal component analysis [4], [8], [9], only the datasets containing strong linear relationships can be identified with high accuracy. However, there always exists unknown or uncertain relationship mapping between the spectral data and the concentration of the component. This can only result in poor performance of these common algorithms. On the other hand, backpropagation (BP) neural network is considered to be a good nonlinear regression method [10]. However, the BP neural network requires a massive database to improve its accuracy, whereas the larger database also needs more computation time. Therefore, a new mixture identification method with unknown relationship identification capability and few training sample requirement is urgently needed.

Here, the human brain, one of the most complicated mixtures, is chosen as the example. Brain tissue cells contain lots of substances, such as γ-aminobutyric acid (GABA), L-glutamic acid (L-Glu), D-myo-inositol (D-MI), creatine monohydrate (CMH), cholesterol (CHO), noradrenaline (NE), and N-acetylaspartate (NAA). But only a small part of them will have obvious concentration change when brain glioma occurs. NE is the major transmitter of various inhibitory neurons and interneurons in the human brain, the increase of its concentration means the invasion of glioblastoma cells. NAA is a cellular structure in neurons, while a decreased concentration of NAA reflects neuronal death and an increased concentration of NAA is related to inflammation, demyelination, and membrane synthesis or repair [11]–[13]. Thus, NE and NAA are chosen as the target substances for the study.

To be specific, a method is proposed for the qualitative and quantitative identification of components in mixtures by THz spectroscopy, without building the databases of each component. First, under different mixture ratios, the absorption spectra of mixtures (CHO, L-Glu, NAA, GABA, NE, D-MI, and CMH) are obtained by THz time-domain spectroscopy (THz-TDS). Next, the baseline and noise are removed, and the absorption

Manuscript received May 9, 2018; revised July 25, 2018 and August 12, 2018; accepted August 12, 2018. Date of publication August 29, 2018; date of current version December 11, 2018. This work was supported in part by the Major National Development Project of Scientific Instrument and Equipment under Grant 2017YFF0106300 and Grant 2016YFF0100503, in part by the National Natural Science Foundation of China under Grant 61771314 and Grant 61722111, in part by the Shanghai Pujiang Program under Grant 16PJD033, in part by the Shanghai Rising-Star Program under Grant 17QA1402500, in part by the Shuguang Program under Grant 17SG45, and in part by the Shanghai Youth Talent Support Program and Young Yangtse River Scholar under Grant Q2016212. *(Corresponding authors: Yan Peng; Yiming Zhu.)*

Y. Peng, C. Shi, M. Xu, T. Kou, X. Wu and Y. Zhu are with the Shanghai Key Lab of Modern Optical System, University of Shanghai for Science and Technology, Shanghai 200093, China (e-mail: py@usst.edu.cn; 779113661@qq.com; 215705216@qq.com; 1050425105@qq.com; wuxumaomao@hotmail.com; ymzhu@usst.edu.cn).

B. Song, H. Ma, and S. Guo are with the Department of Pancreatic Surgery, Changhai Hospital, Second Military Medical University, Shanghai 200433, China.

L. Liu is with the Shanghai Advanced Research Institute, Chinese Academy of Sciences, Shanghai 201210, China.

Color versions of one or more of the figures in this paper are available online at http://ieeexplore.ieee.org.

Digital Object Identifier 10.1109/TTHZ.2018.2867816

spectra are used for building and training the support vector regression (SVR) model. After optimizing the SVR parameters, the testing spectral data are imported into the SVR model for the identification of two key diagnostic substances (NE and NAA). Finally, root-mean-square error (RMSE) and correlation coefficient are used to evaluate the accuracy rate of the mixture identification method.

II. Experiments

A. Sample Preparation

To ensure the wide adaptability of the method, mixtures of seven substances in glioma were selected as examples, including GABA, L-Glu, D-MI, CMH, CHO, NE, and NAA (all purchased from Sigma–Aldrich Corporation and stored according to the instructions of the supplier). Mixture samples were pressed as 13 mm tablets with 2 mm thickness. The masses of all the tablets are controlled to 130 ± 5 mg (the PE is 30 ± 1 mg) and the concentration of target components (NE and NAA) are within 0%–12% with random distribution. This range includes all their possible concentrations in the real human brain. To avoid damaging the compounds, the grinding force was controlled, and the pressing time was set at 1 min. Then, the absorption spectra of samples were measured by THz-TDS, whose effective spectral range is 1.0–3.0 THz, resolution is 0.001 THz, and SNR is 40 000:1 [4], [13]. The environmental temperature was controlled at 22 °C and humidity <3%. To ensure the reliability of the data, each sample was measured ten times and then averaged. The absorbance of the samples $\alpha(w)$ can be calculated by using the following equation [4], [13]:

$$\alpha(w) = \frac{1}{d}\ln\frac{l_{\mathrm{ref}}(f)}{l_{\mathrm{sam}}(f)}. \tag{1}$$

B. Data Preprocessing

The sources of major noise in THz spectra have the following three aspects.

1) When the particle size of the substance is comparable to the THz wavelength, scattering causes a significant loss in THz amplitude and yields a slanted baseline. This baseline will break the linearity between mixtures and their components and, therefore, increase the difficulty of identification and quantitative analysis.
2) Impurities existing in the substances (which are unavoidable) will also give out their THz spectra in the final spectral shape. This will complicate the recognition algorithms and then affect the efficiency and accuracy of identification.
3) Noises from the test system and the environment have a random distribution in the final spectrum, whose amplitudes are small but its irregular frequency makes spectral analysis difficult.

Therefore, the elimination of noise and baseline is necessary for the accurate quantitative measurement of the substances in the mixture. Considering the baseline and noise to be nonlinear with the change of frequency, the wavelet transform is used to deal with these in the THz spectra. The wavelet transform has good time–frequency localization character and decorrelation, which can remove the baseline and noise with different scales and retain the effective information in the different frequency regions of the THz spectrum at the same time [14]–[16]. The Mallat algorithm [17] is used for wavelet decomposition as follows:

$$f(u) = C_J\varphi_J(u) + \sum_{j=1}^{J} d_j\psi_j(u). \tag{2}$$

The mother wavelet $\varphi_J(u)$ is orthogonal to the scaling function $\psi_j(u)$, C_J is a coefficient in the $(J + 1)$th level of the low-frequency component, and d_j is a coefficient in the *j*th level ($1 \leq j \leq J$) of the high-frequency component. With this equation, different parts and frequencies of the signal can be analyzed [18]. Based on the mother wavelet Daubechies 9, the spectra of the mixtures were decomposed at level six. The coefficients in levels 1–6 contain both noise and useful information. We dropped the high-frequency components from levels 1–3, which represented noise, and retained the low-frequency components from levels 4–6, which represented useful information. Then, the polynomial fitting method [19] was used to correct the baseline of the THz spectrum. For each spectrum after wavelet transform, we chose minimums at different frequencies as series reference points, then fit them by using different order polynomials based on the least square. The obtained curve is used as the absorption baseline, which will be removed from the initial spectrum.

C. SVR Identification

SVR was used in our qualitative and quantitative identification [20]. A given set of train data (X_i, Y_i), $i = 1, 2, 3, \ldots, N$, $X_i \in R_n$ are used to establish the SVR model, X_i is the absorption spectral data of the mixture sample, which is the input data of SVR model. $X_i = (x_1, x_2, \ldots, x_m)$, *m* is the number of the discrete sampling frequency in spectral data, and *N* is the number of samples. Y_i is the concentration of target components in a mixture sample, which is the output data of the SVR model. As our quantitative identification method for certain components is a situation of multioutputs (number of target substance in mixture), it is necessary to establish multiple SVR models where each SVR corresponds to a certain component in the mixture. Then, this model is employed to predict the concentration of the target component for testing set samples. The SVR estimation can be described as follows:

$$Y_n = w^T \cdot \varphi(X) + b. \tag{3}$$

Y_n is the predicted result of the SVR model, *w* and *b* are the weight and bias parameters of the regression function, respectively, and $\varphi(X)$ is the function that transforms the input *x* to the separating space [21]. Thus, the risk function of the SVR model is as follows:

$$|Y_n - Y_i| = \begin{cases} 0, & \text{if } |Y_n - Y_i| \leq \varepsilon \\ |Y_n - Y_i| - \varepsilon, & \text{otherwise} \end{cases} \tag{4}$$

where ε is the loss parameter based on Vapnik's ε-insensitive [22]. Then, *w* and *b* can be derived by the SVR objective function as follows:

$$\min\left[\frac{1}{2}w^T w + c\left(\sum_{n=1}^{N}\varepsilon_n^{\wedge}+\sum_{n=1}^{N}\varepsilon_n^{\vee}\right)\right] \tag{5}$$

subject to

$$\begin{cases}-\varepsilon-\varepsilon_n^{\vee}\le Y_n-w^T\cdot\varphi(X)-b\le\varepsilon+\varepsilon_n^{\wedge}\\ \varepsilon_n^{\vee}\ge 0,\varepsilon_n^{\wedge}\ge 0\end{cases} \tag{6}$$

where *c* is the regularization parameter that represents the degree of the penalty loss, $\varepsilon_n^{\wedge}$ and $\varepsilon_n^{\vee}$ are the slack variables of the loss parameter ε.

Furthermore, kernel function $k(x,y)=\varphi(X)\cdot\varphi(Y)$ is very important for constructing the SVR model. It was mainly used to map the input data onto a higher dimensional space that enables the solution of the nonlinear optimization problem linear separable. The three most commonly used kernel functions are described by the following equations:

$$\text{Linear kernel}:\ k(x,y)=x\cdot y \tag{7}$$

$$\text{Polynomial kernel}:\ k(x,y)=[(x\cdot y)+1]^d \tag{8}$$

$$\text{RBF kernel}:k(x,y)=\exp(-|x-y|^2/g^2). \tag{9}$$

The kernel parameter *g* and the regularization parameter *c* should be defined using the best fitting value, which can be obtained by trial and error. Leave-one-out cross validation was used to find the optimal parameters by which the model can achieve the best predicted results.

To evaluate the performance of the established models, RMSE and correlation coefficient were employed to evaluate the developed model's accuracy. RMSE represents the dispersion degree of predicted results, while the correlation coefficient represents the relevancy. They are, respectively, calculated as follows:

$$\text{RMSE}=\sqrt{\frac{\sum_{i=1}^{N}\left(Y_i-\hat{Y}_i\right)^2}{N}}\times 100\% \tag{10}$$

$$R=\sqrt{1-\frac{\sum_{i=1}^{N}\left(Y_i-\hat{Y}_i\right)^2}{\sum_{i=1}^{N}\left(Y_i-\bar{Y}_i\right)^2}}\times 100\% \tag{11}$$

where *N* represents the number of samples in the training set, Y_i^{-} is the average of Y_i, and Y_i and $\overset{\wedge}{Y_i}$ are the actual value of the *i*th sample in the dataset and the predicted value of the *i*th sample in the developed model, respectively.

III. Results and Discussion

For the training spectral data, these seven components were mixed randomly at ten different concentrations (mg/mg) under a constant total mass, where the masses of all tablets are controlled to 130 ± 5 mg (the PE is 30 ± 1 mg) and the concentrations of target components (NE and NAA) are within 0%–12% with random distribution. As the input label of our method needs only the concentration of the target substance and the concentrations of all components in the mixture are irrelevant with each other, only the masses of the target substances (NAA and NE) are recorded. The specific value of masses in ten mixture samples is given in Table I.

TABLE I
Ten Mixture Samples Under a Constant Total Mass (130 ± 5 mg), the Concentrations of Target Components (NE and NAA) are Within 0%–12% with Random Distribution

sample	TC	L-glu	NAA(mg)	GABA	NE(mg)	D-MI	CMH
1	--	--	3.02	--	1.13	--	--
2	--	--	0	--	7.02	--	--
3	--	--	11.99	--	0	--	--
4	--	--	0	--	3.82	--	--
5	--	--	1.2	--	9.98	--	--
6	--	--	4.13	--	8.91	--	--
7	--	--	15.2	--	12.03	--	--
8	--	--	12.87	--	0	--	--
9	--	--	4.86	--	4.8	--	--
10	--	--	7.02	--	13.03	--	--

The corresponding THz spectra of these ten mixture samples are shown in Fig. 1(a), which obviously have separate and sharp absorption peaks but with small noise and baseline drifting. To improve the precision and accuracy of identification, it is necessary to reduce the noise and correct the baseline of spectra before applying the algorithm. Fig. 1(b) shows the spectra after removing the noise by wavelet transform. We can see that the small vibrations near the absorption peaks have all been smoothed. Then, the baselines of spectra are eliminated by the polynomial fitting method. The final processed THz absorption spectra are shown in Fig. 1(c). It can be clearly seen that the baselines have been well corrected, i.e., their base values are almost close to zero.

Considering the spectrum of each pure substance in mixture to be unknown, we need to build and train an SVR model to find out the regularity between the spectra of mixture samples and the concentration of substances. The open-source software LIBSVM [19] was used in our method to identify the concentration of target substances (NAA and NE) in mixtures. Due to the small number of samples in this experiment, all samples were divided into five equal parts randomly. One part (two samples) was used as the prediction set and the remaining four parts (eight samples) were used as the training sets to construct the model. The loop iteration was repeated five times. Subsequently, the obtained results of prediction sets were pooled and utilized to estimate the RMSE and correlation coefficient.

The performance of the SVR model depends on a proper setting of several parameters such as the regularization parameter *c*, kernel function parameter *g*, and the loss parameter *e* [23]–[25]. The parameter *c* determines the tradeoff between the training error and the model complexity, the parameter *g* determines the distribution of support vectors in the new feature space, and the parameter *e* affects the number of support vectors used for the construction of the regression model. Here, RBF kernel function and the leave-one-out cross validation were used to determine the optimum value of *c*, *g*, and *e* [26], [27]. Through circular

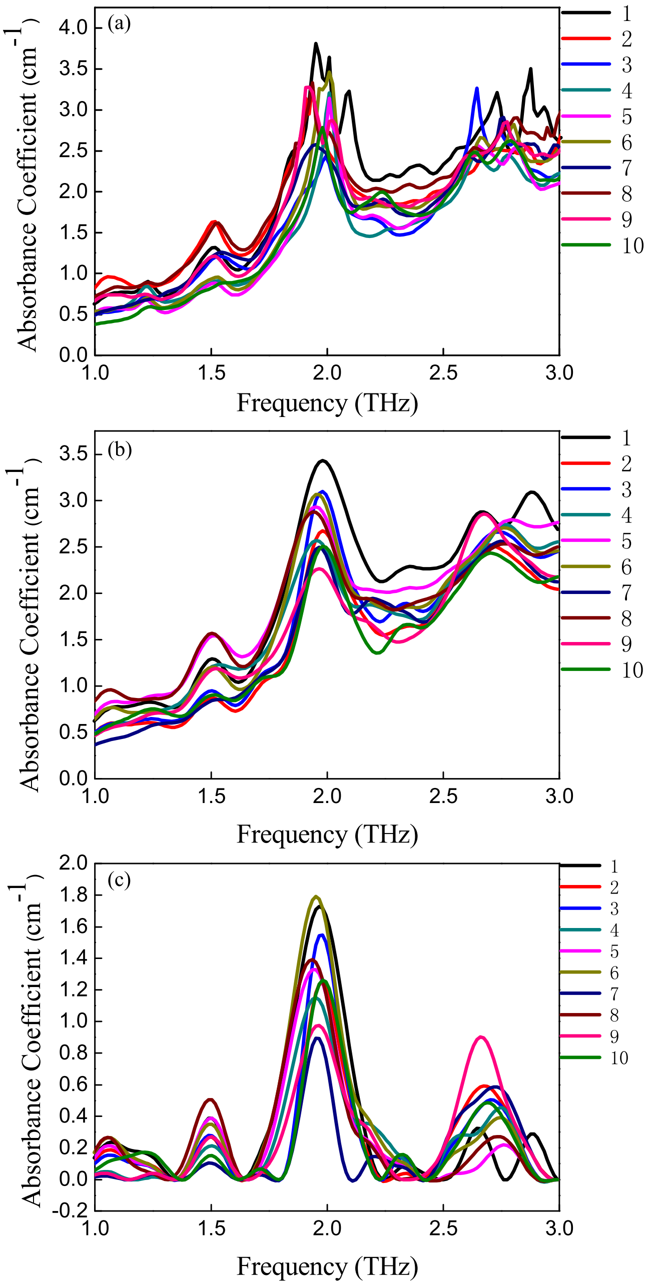

Fig. 1. THz absorption spectra of ten mixture samples. (a) Original spectra. (b) Spectra after wavelet transform. (c) Baseline correction by polynomial fitting.

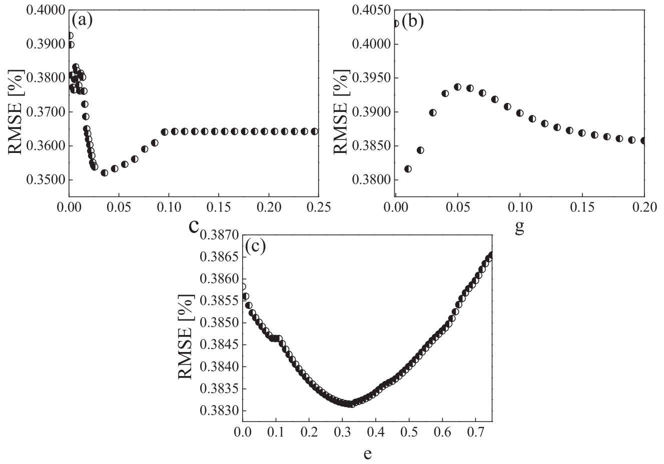

Fig. 2. RMSE of the training set based on leave-one-out cross validation versus three SVR parameters. (a) Regularization parameter c with $g = 0.01$, $e = 0.01$. (b) Kernel function parameter g with $c = 0.25$, $e = 0.01$. (c) Loss parameter e with $c = 0.25$, $g = 0.01$.

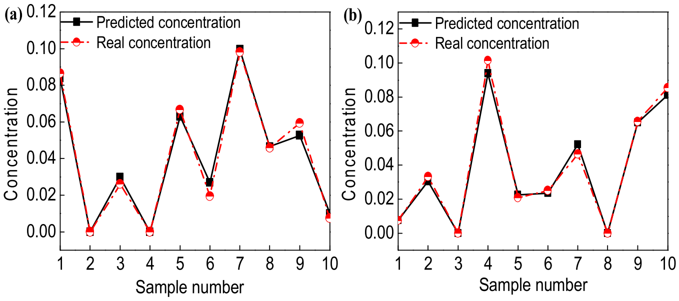

Fig. 3. Actual and predicted concentrations of the two target components in mixtures. (a) NAA. (b) NE.

TABLE II
QUANTITATIVE ANALYSES OF NAA AND NE COMPONENTS IN MIXTURES

Sample name	RMSE	Average RMSE	R	Average R
NAA	0.40%	0.40%	99.13%	99.135%
NE	0.40%		99.14%	

computations, one parameter varied with fixed step lengths in the allowable range, while the other two parameters were fixed. The selection progress and results are shown in Fig. 2(a)–(c).

The position of the minimum RMSE corresponds to the best parameter. Optimization parameters of SVR are: $c = 0.025$, $g = 0.01$, and $e = 0.33$. We can see that the fine-tuning of these values can apparently influence the predictability of our model. The rapid increase of c and g could induce incorrect fitting in the training data, while the decrease of those parameters can induce the failure of testing data prediction. As for parameter e, if the value is too large, only a small amount of support vectors can be selected, leading to a decrease of the final prediction performance; if the value is too small, excessive support vectors will lead to an over fitting of the SVR model. These optimum parameters were used to build an SVR model, and then predict the concentrations of target substances in testing data.

Fig. 3 presents the comparison between the actual and predicted results for NAA and NE components in mixtures. The SVR model using processed spectral data shows good results. Table II shows the RMSE and R of the test sets. As it can be seen, except for accurate qualitative identification, the quantitative correlation coefficients of NAA and NE are 99.13% and 99.14%, respectively, both with the RMSE of 0.40% as compared to the experimental ratios.

For comparison, PLS and BP neural networks, which are widely used as qualitative identification methods that do not need to obtain the spectrum of each component in mixtures in advance, were also tested in our case [28], [29]. PLS is a multivariate statistical analysis algorithm, which can process a large number of correlated data. BP neural networks model is

TABLE III
RESULTS OF DIFFERENT ALGORITHMS FOR QUANTITATIVE IDENTIFICATION OF NAA AND NE IN MIXTURES

algorithm	Average RMSE %	Average R
PLS	2.31%	80.52%
BP neural networks	1.75%	83.53%
Our model	0.40%	99.135%

a kind of nonlinear multivariate calibration algorithm, a multiple layer feed-forward neural network that is trained by the error reverse propagation algorithm. Here, the performances of these models are evaluated using their average RMSE and correlation coefficient for the prediction of the test data set (see Table III). It can be seen clearly that the accuracy of the PLS (80.52%) and BP neural networks (83.53%) models is much lower than that of our method (99.135%), which also show the unsatisfied RMSE coefficients 2.31% and 1.75%, respectively. This is because the PLS is more adept at the analysis of linear data, while our mixtures data with "uncertain linear correlation" directly limit its effectiveness. While for BP neural networks model, except the requirement of massive database, it is easy to fall into local minimum and then it is difficult to acquire the optimized result. Therefore, for the identification and prediction of the target component concentration in mixtures, especially for the samples with unknown relationship characteristics and small training number, the proposed method is better.

IV. Conclusion

In this paper, we propose a qualitative and quantitative mixture identification method that includes the wavelet transform, baseline elimination, SVR, and loop iteration of samples, which can be used for the identification of specific substances in a mixture effectively. In the example of seven component mixtures (GABA, L-Glu, D-MI, CMH, CHO, NE, and NAA) in the human brain, our method can achieve an average RMSE and average correlation coefficient of 0.40% and 99.135%, respectively, which is much better than the usual single-spectrum-known pattern and mixture identification methods such as PLS and BP neural networks. These results are important for the identification of mixtures in real applications with unknown relationship characteristics and small training number.

References

[1] P. H. Siegel, "Terahertz technology," *IEEE Trans. Microw. Theory Techn.*, vol. 50, no. 3, pp. 910–928, Mar. 2002.

[2] A. Rogalski and F. Sizov, "Terahertz detectors and focal plane arrays," *Opto-Electron. Rev.*, vol. 19, pp. 346–404, Mar. 2011.

[3] P. U. Jepsen, D. G. Cooke, and M. Koch, "Terahertz spectroscopy and imaging—Modern techniques and applications," *Laser Photon. Rev.*, vol. 2, pp. 418–418, Feb. 2012.

[4] Y. Peng *et al.*, "Terahertz identification and quantification of neurotransmitter and neurotrophy mixture," *Biomed. Opt. Express*, vol. 7, no. 11, pp. 4472–4479, 2016.

[5] G. Liu *et al.*, "Quantitative measurement of mixtures by terahertz time-domain spectroscopy," *J. Chem. Sci.*, vol. 121, pp. 515–520, Apr. 2009.

[6] H. Ge *et al.*, "Identification of wheat quality using THz spectrum," *Opt. Express*, vol. 22, pp. 12533–12544, Oct. 2014.

[7] T. Chen *et al.*, "Quantitative analysis of mixtures using terahertz time-domain spectroscopy and different PLS algorithms," *Adv. Mater. Res.*, vol. 804, pp. 23–28, 2013.

[8] A. Matei, N. Drichko, B. Gompf, and M. Dressel, "Far-infrared spectra of amino acids," *Chem. Phys.*, vol. 316, pp. 61–71, 2005.

[9] M. Ariza *et al.*, "Neuropsychological correlates of basal ganglia and medial temporal lobe NAA/Cho reductions in traumatic brain injury," *Arch. Neurol.*, vol. 61, pp. 541–544, Apr. 2004.

[10] H. Zhan *et al.*, "Monitoring PM2.5 in the atmosphere by using terahertz time-domain spectroscopy," *J. Infrared Millim. Terahertz Waves*, vol. 37, pp. 1–10, Sep. 2016.

[11] Z. Yan, D. Hou, P. Huang, B. Cao, and G. Zhang, "Terahertz spectroscopic investigation of L-glutamic acid and L-tyrosine," *Meas. Sci. Technol.*, vol. 19, pp. 158–160, Jan. 2007.

[12] Y. F. Hua and H. J. Zhang, "Qualitative and quantitative detection of pesticides with terahertz time-domain spectroscopy," *IEEE Trans. Microw. Theory Techn.*, vol. 58, no. 7, pp. 2064–2070, Jul. 2010.

[13] W. Chen *et al.*, "Isomers identification of 2-hydroxyglutarate acid disodium salt (2 HG) by terahertz time-domain Spectroscopy," *Sci. Rep.*, vol. 7, Jan. 2017, Art. no. 12166.

[14] P. M. Ramos and I. Ruisánchez, "Noise and background removal in Raman spectra of ancient pigments using wavelet transform," *J. Raman Spectrosc.*, vol. 36, pp. 848–856, 2005.

[15] M. AlMahamdy and H. B. Riley, "Performance study of different denoising methods for ECG signals," *Procedia Comput. Sci.*, vol. 37, pp. 325–332, 2014.

[16] R. Cohen, "Signal denoising using wavelets," Dept. Elect. Eng., Technion, Israel Inst. Technol., Haifa, Israel, Project Rep. Winter 2011/2012, Feb. 2012.

[17] J. Lu, "Parallelizing Mallat algorithm for 2-D wavelet transforms," *Inf. Process. Lett.*, vol. 45, pp. 255–259, Mar. 1993.

[18] L. Xu, D. Zhang, and K. Wang, "Wavelet-based cascaded adaptive filter for removing baseline drift in pulse waveforms," *IEEE Trans. Biomed. Eng.*, vol. 52, no. 11, pp. 1973–1975, Nov. 2005.

[19] L. Zhang, "The relationship between kernel functions based SVM and three-layer feedforward neural networks," *Chin. J. Comput.*, vol. 25, pp. 696–700, Jul. 2002.

[20] S. Knerr, L. Personnaz, and G. Dreyfus, "Single-layer learning revisited: A stepwise procedure for building and training a neural network," in *Neurocomputing*, vol. 68. Berlin, Germany: Springer, 1990, pp. 41–50.

[21] R. E. Fan, P. H. Chen, and C. J. Lin, "Working set selection using second order information for training support vector machines," *J. Mach. Learn. Res.*, vol. 6, pp. 1889–1918, Apr. 2005.

[22] X. Li *et al.*, "Using wavelet transform and multi-class least square support vector machine in multi-spectral imaging classification of Chinese famous tea," *Expert Syst. Appl.*, vol. 38, pp. 11149–11159, Sep. 2011.

[23] L. Devroye, L. Gyorfi, and G. Lugosi, *A Probabilistic Theory of Pattern Recognition*, vol. 31. New York, NY, USA: Springer-Verlag, 1996.

[24] O. Chapelle *et al.*, "Choosing multiple parameters for support vector machines," *Mach. Learn.*, vol. 46, pp. 131–159, 2002.

[25] C. Cortes and V. Vapnik, "Support-vector networks," *Mach. Learn.*, vol. 20, pp. 273–297, Mar. 1995.

[26] S. Alam *et al.*, "Performance of classification based on PCA, linear SVM, and Multi-kernel SVM," in *Proc. IEEE 8th Int. Conf. Ubiquitous Future Netw.*, 2016, pp. 987–989.

[27] G. H. Feng, "Parameter optimizing for support vector machines classification," *Comput. Appl. Eng. Edu.*, vol. 47, pp. 123–126, Mar. 2006.

[28] Q. Wang and Y. H. Ma, "Qualitative and quantitative identification of nitrofen in terahertz region," *Chemometrics Intell. Lab. Syst.*, vol. 127, pp. 43–48, Nov. 2013.

[29] T. Chen *et al.*, "Quantitative analysis of mixtures using terahertz time-domain spectroscopy and different PLS algorithms," *Adv. Mater. Res.*, vol. 804, pp. 23–28, 2013.

Yan Peng studied at the Department of Physics, East China Normal University, Shanghai, China, from 2004 to 2009.

In 2009, she joined the Shanghai Key Lab of Modern Optical System, University of Shanghai for Science and Technology, Shanghai. In 2013, she joined the Optical Research Center, University of Rochester, New York, NY, USA, as a Scientist. In 2014, she joined the University of Pennsylvania, as an Adjunct Professor. Since 2016, she has been working as a Professor with the University of Shanghai for Science and Technology. Her research interests include generation and regulation of terahertz waves and application of terahertz waves in biomedicine.

Chenjun Shi studied at the University of Shanghai for Science and Technology, Shanghai, China, from 2014 to 2018. He is currently working toward the master's degree in optical engineering at the Shanghai Key Lab of Modern Optical System, University of Shanghai for Science and Technology.

Mingqian Xu received the graduate degree from the Nanjing University of Posts and Telecommunications, Nanjing, China, in 2015. Since 2015, she has been working toward the master's degree in machine learning and pattern recognition at the University of Shanghai for Science and Technology, Shanghai, China.

Tianyi Kou received the B.S. degree in photoelectric information science and engineering from Zijin College, Nanjing University of Science and Technology, Nanjing, China. He is currently working toward the degree in optical engineering at the Shanghai University of Technology, Shanghai, China.

Xu Wu received the M.E. and Ph.D. degrees in material science and technology from the Zhejiang Sci-Tech University, Hangzhou, China, in 2013 and 2017, respectively.

In January 2018, she joined the Terahertz Technology Innovation Research Institute, University of Shanghai for Science and Technology, Shanghai, China, as a Postdoctoral Fellow. Her research interests include terahertz spectroscopic studies of biosolutes in human plasma at the molecular level by the combination of a variety of biophysical techniques.

Bin Song received the graduate degree from the Second Military Medical University, Shanghai, China.

He is currently an Assistant Director of pancreatic hepatobiliary surgery with the Changhai Hospital, which is affiliated with the Second Military Medical University. He is also an Associate Professor and Tutor of master's students. He is also a Doctor of clinical medicine. He has rich clinical experience and has authored or co-authored more than 50 papers in both Chinese and English, including 11 SCI papers.

Hongyun Ma received the B.S. degree from Harbin Medical University, Harbin, China, in 2009, and the M.S. degree from the Second Military Medical University, Shanghai, China, in 2014.

He is currently an Attending Doctor with the Changhai Hospital, which is affiliated with the Second Military Medical University. His research interests include surgical treatment of hepatobiliary pancreatic disease, clinical research, and basic medical research of pancreatic tumors.

Shiwei Guo received the B.S. degree in clinical medicine from the Fourth Military Medical University, Xi'an, China, in 2004, and the Ph.D. degree in general surgery from the Second Military Medical University, Shanghai, China, in 2012.

Since 2015, he has been an Attending Doctor with the HBP Department, Changhai Hospital, Second Military Medical University. His research interests include surgical treatment of pancreatic and hepatobiliary tumors, genomics research on the pathogenesis of pancreatic and hepatobiliary tumors, and individualized treatment of HBP tumors.

Lizhuang Liu received the Ph.D. degree in biomedical engineering from Xi'an JiaoTong University, Xi'an, China, in 2003.

He is currently a Professor with the Shanghai Advanced Research Institute, Chinese Academy of Sciences, Beijing, China. His research interests include computer vision, real time computing, multiple sensor data and image processing, and smart system application.

Yiming Zhu received the bachelor's degree in applied physics from Jiao Tong University, Shanghai, China, in 2002, and the Ph.D. degree in electronics engineering from the University of Tokyo, Tokyo, Japan, in 2008.

In 2003, he began working as an Assistant Researcher with the Research Center for Advanced Science and Technology, University of Tokyo. He is currently a Professor with the University of Shanghai for Science and Technology, Shanghai, China. His research interests include ultrafast optics, ultrafast electronics, terahertz technology, etc.