DISCRIMINATION OF MANUFACTURERS ORIGIN OF OXYTETRACYCLINE USING TERAHERTZ TIME-DOMAIN SPECTROSCOPY WITH CHEMOMETRIC METHODS**

J. Guo [1,2], H. Deng [1], Q. Ch. Liu [1], L. Y. Chen [1], Zh. G. Xiong [1,3], L. P. Shang [1*]

[1] School of Information Engineering at Southwest University of Science and Technology, Mianyang 621010, China; e-mail: 121992278@qq.com
[2] TianFu College at Southwestern University of Finance and Economics, Mianyang 621000, China
[3] School of Mechanical Engineering at Guilin University of Aerospace Technology, Guilin 541004, China

We propose an effective approach for determining the manufacturer of oxytetracycline based on terahertz time-domain spectroscopy and chemometrics. The applied method allowed detection and distinction of oxytetracycline tablets produced by four Chinese manufacturers. Absorption spectra within the frequency range of 0.2–1.6 THz were acquired for analysis, and principal component analysis (PCA) and t-distributed stochastic neighbor embedding (t-SNE) were implemented for dimensionality reduction. The obtained data were input into different classifiers, namely, grid-search support vector machines (SVM), genetic algorithm-SVM, particle swarm optimizations (PSO-SVM), decision trees, and random forests. The results indicated that the absorption spectra of oxytetracycline produced by the four manufacturers could be differentiated, and that t-SNE outperformed PCA in terms of dimensionality reduction. The optimal classifier was obtained by combining t-SNE and PSO-SVM, and this approach demonstrated the highest cross validation accuracy (98.33%), an average training set accuracy of 97.72%, and an average test set accuracy of 94.25%. The novel approach described herein achieved rapid classification of oxytetracycline and allowed accurate tracing of the manufacturers.

Keywords: *Terahertz time-domain spectroscopy, oxytetracycline, principal component analysis, t-distributed stochastic neighbor embedding, support vector machines, genetic algorithm, particle swarm optimization.*

ОПРЕДЕЛЕНИЕ ПРОИЗВОДИТЕЛЕЙ ОКСИТЕТРАЦИКЛИНА С ИСПОЛЬЗОВАНИЕМ ТЕРАГЕРЦОВОЙ СПЕКТРОСКОПИИ ВО ВРЕМЕННОЙ ОБЛАСТИ И ХЕМОМЕТРИЧЕСКИХ МЕТОДОВ

J. Guo [1,2], H. Deng [1], Q. Ch. Liu [1], L. Y. Chen [1], Zh. G. Xiong [1,3], L. P. Shang [1*]

УДК 543.42

[1] Школа информационной инженерии Юго-Западного университета науки и техники, Мяньян 621010, Китай; e-mail: 121992278@qq.com
[2] Колледж ТяньФу при Юго-Западном финансово-экономическом университете, Мяньян 621000, Китай
[3] Школа машиностроения Гуйлиньского университета аэрокосмических технологий, Гуйлинь 541004, Китай

(Поступила 17 июля 2020)

Для определения производителя окситетрациклина предложен эффективный подход на основе терагерцовой спектроскопии во временной области и хемометрии. Метод позволяет обнаружить и отличить таблетки окситетрациклина, произведенные четырьмя китайскими производителями. Зарегистрированы спектры поглощения в диапазоне 0.2–1.6 ТГц, для уменьшения размерности использованы анализ главных компонент (PCA) и стохастическое вложение соседей с t-распреде-

**Full text is published in JAS V. 88, No. 4 (http://springer.com/journal/10812) and in electronic version of ZhPS V. 88, No. 4 (http://www.elibrary.ru/title_about.asp?id=7318; sales@elibrary.ru).

лением (t-SNE). Полученные данные вводились в различные классификаторы: метод опорных векторов (SVM), генетический алгоритм-SVM, оптимизация роя частиц (PSO-SVM), деревья решений и случайные леса. Показано, что спектры поглощения окситетрациклина разных производителей могут быть дифференцированы, а t-SNE превосходит PCA с точки зрения уменьшения размерности. Оптимальный классификатор получен путем объединения t-SNE и PSO-SVM. Подход демонстрирует высокую точность перекрестной валидации (98.33%), среднюю точность обучающего набора 97.72% и среднюю точность тестового набора 94.25%. Метод позволяет быстро классифицировать окситетрациклин и точно отслеживать производителей.

Ключевые слова: *терагерцовая спектроскопия во временной области, окситетрациклин, анализ главных компонент, стохастическое вложение соседей с t-распределением, метод опорных векторов, генетический алгоритм, оптимизация роя частиц.*

Introduction. The quality of pharmaceuticals is a global concern because the production and sale of counterfeit drugs can damage public health and severely violate the interests of legal enterprises. Any differences in the quality of the raw materials or the manufacturing processes can manifest as variations in the quality of the drugs produced by different manufacturers. Therefore, it is crucial to develop a qualitative drug identification method to trace the manufacturer and assist in drug supervision. Oxytetracycline, which is also known as oxytetracycline hydrochloride, belongs to a larger family of tetracycline antibiotics and was the object of identification in the present study. It induces an antibacterial effect against many cocci and bacilli and has been applied extensively in animal husbandry [1]. As a result, oxytetracycline occupies an important position in the current antibacterial drug market.

Common strategies for identifying and characterizing antibiotics mainly include infrared spectroscopic analysis, Raman spectral analysis, and liquid chromatography [2]. However, all of these techniques involve complex detection procedures, relatively long detection times, and low resolution. The terahertz (THz) wavelength is between 3 μm and 3 mm, and the frequency domain is 0.1–10 THz, so THz waves are between microwaves and infrared waves on the electromagnetic spectrum. In addition to its ultrawide spectrum, a THz wave has a transient nature, exhibits penetrability, is safe (single-photon energy = 4.1 meV), and is conducive to fingerprinting (the THz frequency can be matched with the rotational energy level of biomacromolecules, making it suitable for fingerprint analysis of such matter) [3]. Given the aforementioned benefits, THz time-domain spectroscopy (THz-TDS) has already been applied for biological tissue recognition [4, 5], food and drug detection [6, 7], and explosives detection [8].

Drug detection based on THz spectroscopy has several unique advantages. For example, it can be used to investigate the inner structure and organizational features of drugs because the vibrational and rotational energy levels of macromolecules, such as antibiotics and traditional Chinese medicines, are in the THz waveband. Zeitler et al. [9] applied THz spectroscopy to identify different crystal forms of sulfathiazole, and Limwikrant et al. [10] obtained the THz spectrum of an ofloxacin-oxalic acid complex. Additionally, Zhang [11] determined the molecular vibration modes of piracetam and 3-hydroxybenzoic acid. Zhang et al. [12] analyzed the THz absorption spectra of metronidazole, tinidazole, and ornidazole and obtained their fingerprints. Finally, Xie et al. [13] used DFT (Density Functional Theory) calculations to show that tetracycline had a definitive THz absorption spectrum at certain frequencies.

Many studies have demonstrated the feasibility of applying THz spectroscopy for drug detection. Liu [14], Luo [15], Qin [16], and Zhan [17] et al. have used pattern recognition algorithms and reported higher recognition accuracy in the classification or identification of food and drugs based on their THz absorption spectra. However, the same drug produced by different manufacturers only differs slightly, so the THz time-domain spectra only contain minor distinctions, making direct qualitative identification difficult. Herein, we propose a method for tracing the manufacturers of oxytetracycline by integrating THz-TDS and analyzing SVM results. The THz time-domain spectra were obtained from experimental samples, and the THz absorption spectra were calculated. Principal component analysis (PCA) and *t*-distributed stochastic neighbor embedding (*t*-SNE) were used for dimensionality reduction. Next, various pattern recognition algorithms (i.e., GS-SVM, GA-SVM, PSO-SVM, DT, and RF) were applied to establish qualitative models, and the optimal comparative classifier was determined. This novel approach for differentiating the manufacturers of oxytetracycline was validated based on the experimental results.

Experimental section. Forty oxytetracycline tablets (from the same batch) produced by four manufacturers were detected using the developed system. First, the drug samples were ground to avoid particle heterogeneity that would scatter the THz waves and to increase the signal-to-noise ratio. Then an electronic balance (precise to 0.1 mg) was used to weigh a designated amount of the sample, which was pressed into a sample tablet using a tablet maker. A digital caliper (precise to 0.02 mm) was used to measure the thickness of the sample tablets. Overall, 40 drug samples were prepared from each of the four manufacturers, and THz absorption spectra were analyzed to evaluate the detection performance. Finally, the 160 samples were randomly divided into a calibration set (30×4) and a testing set (10×4).

A Zomega THz-TDS system (Z3) was used for the detection. Pulses generated by the femtosecond laser system passed through the beam splitter, which split them into a pump beam and probe beam. The pump beam was fed into the 50-μm dipole antenna (after being reflected) to generate THz waves. After reflecting and focusing, the THz waves penetrated the samples. Then, the THz waves carrying the drug sample information and the probe beam that had been passed through the delay system were focused onto a photo transistor. Finally, they passed through the lock-in amplifier to obtain the THz time-domain spectra. The key parameters of the system were set as follows: wavelength of the femtosecond laser system = 800 nm, frequency = 80 MHz, intensity of pump light = 100 mW, intensity of probe light = 20 mW, scan stroke = 50 ps, useful spectral range ≈0.1–2 THz, and dynamic range >70 db. The entire detection system was enclosed in a hood to control the ambient temperature and humidity and to minimize water evaporation. The temperature was maintained at 23ºC, and the humidity was maintained below 2%.

Calculations. *Data processing*. The THz time-domain spectra of the samples were acquired using the experimental system described above. Denoising was performed for the spectra using the Savitzky–Golay smoothing filter. The reflection peaks were eliminated via empirical mode decomposition [18], and the spectral information was extracted using a Fourier transform. The model was extracted based on the optical parameters proposed by Dorney [19] and Duvillaret et al. [20] and considering the relationships shown below:

$$\rho(\omega) = A_s/A_r,$$

$$\varphi(\omega) = \varphi_s - \varphi_r,$$

where $\rho(\omega)$ is the amplitude ratio, A_s is the signal amplitude of the sample, A_r is the signal amplitude of the reference, $\varphi(\omega)$ is the phase difference, φ_s is the phase of the sample, and φ_r is the phase of the reference.

The index of refraction, $n(\omega)$, and absorption coefficient, $\alpha(\omega)$, were calculated for the samples,

$$n(\omega) = \varphi(\omega)\frac{c}{\omega d} + 1,$$

$$\alpha(\omega) = \frac{2}{d}\,\mathrm{In}\left\{\frac{4n(\omega)}{\rho(\omega)\left[n(\omega)+1\right]^2}\right\},$$

where c is the speed of light in vacuum, ω is the angular frequency, and d is the sample thickness.

Algorithm flow chart. The overall workflow of the algorithm is presented in Fig. 1, which includes dimensionality reduction and classifier construction.

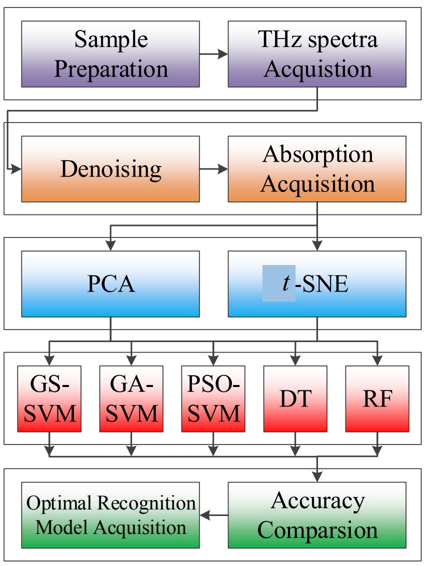


Fig. 1. Flow chart describing the applied procedure.

Stage 1: Pressing drug sample tablets and detection using THz-TDS.
Stage 2: Data processing involved denoising of the THz time-domain spectra and calculation of the THz absorption spectra.
Stage 3: Feature selection (i.e., dimensionality reduction) was performed using PCA and *t*-SNE [21].
Stage 4: Various oxytetracycline classification models were constructed.
Stage 5: The classifiers were assessed based on the prediction accuracy, and the optimal model was determined.

Results and discussion. *Spectral analysis.* The developed THz-TDS system was used to acquire the THz time-domain spectra of oxytetracycline samples derived from drugs produced by four manufacturers, as shown in Fig. 2a. The absorption spectral windows were truncated to the frequency range of 0.2–1.6 THz, as shown in Fig. 2b. The oxytetracycline tablets produced by the four manufacturers had the same absorption peaks at 0.887 and 1.207 THz. Therefore, considering the spectral features alone was not sufficient to differentiate between the manufacturers. To address this problem, classifiers were built using various pattern recognition algorithms.

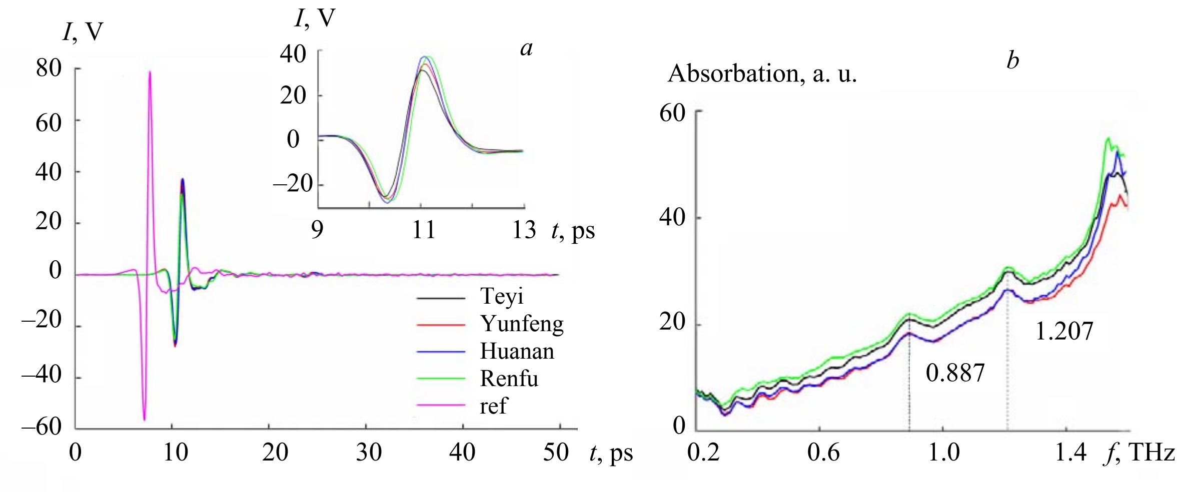
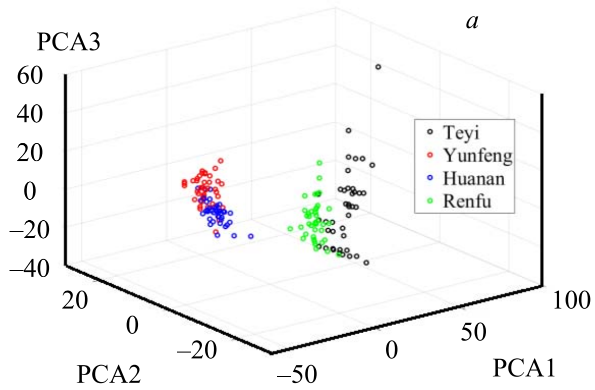


Fig. 2. Oxytetracycline samples from different manufacturers. (a) THz time-domain spectra; (b) THz absorbance spectra.

Visualization of the sample classification. Dimensionality reduction was performed for the spectral data using PCA and *t*-SNE to increase the prediction accuracy and reduce the model training time. Figure 3a shows the three-dimensional (3D) distribution of the principal components of the absorption spectra of the oxytetracycline tablets produced by the four manufacturers. PCA of the absorption spectra was performed on 160 frequencies within the range of 0.2–1.6 THz, allowing extraction of the three principal components (PC1, PC2, PC3), whose contributions were 71.78, 17.71, and 7.30%, respectively. The sum of the three principal components contributions was 96.79%. Figure 3b shows the 3D distributions plotted following *t*-SNE analysis. The sample divergence in this plot was slightly greater than that in Fig. 3a, with less overlap of samples from different manufacturers. This observation was consistent with the results from comparing different classifiers.

Classification analysis. After dimensionality reduction through PCA or *t*-SNE, the new data matrix (160 samples×3 dimensions) replaced the original spectral data matrix as the model input. These two dimensionality reduction methods formed different recognition models with SVM, DT, and RF algorithms. Next, the 160 samples were randomly divided into a training set (30×4 samples) and a testing set (10×4 samples). These models were tested and compared, and the optimal parameters and classifiers were selected.

Herein, we provide a detailed description of how SVM was used to construct the classifier. The overall performance depended on the penalty factor (c) and the kernel parameter (g). The penalty factor prevents the training error of SVM from being equal to zero, which helps achieve a balance between the minimum experience loss and the minimum risk loss. The kernel parameter affects the nonlinear transformation function, such that, if g is large, it will cause overfitting, and if g is small, it will cause underfitting. To attain the desired performance, it is crucial to optimize the process and select the appropriate model parameters. First, parameters *c* and *g* were optimized using a grid search (GS) function, and the GS-SVM model was built. Then, after fivefold cross validation (CV), the optimal values for *c* and *g*, CV-accuracy, training set accuracy, and test set accuracy were determined using PCA and t-SNE, as shown in Table 1. Figure 4a shows the 3D results of parameter selection for PCA-GS-SVM, where the CV-accuracy was 95%, and Fig. 4b shows the 3D results of parameter selection for *t*-SNE-GS-SVM, where the CV-accuracy was 96.67%.



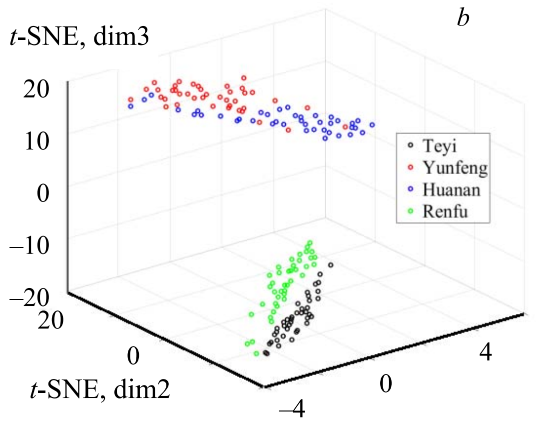
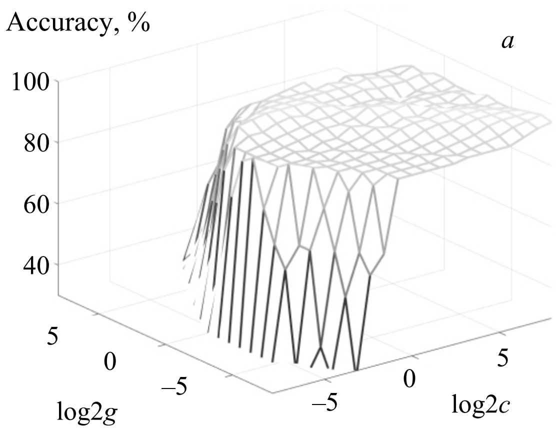


Fig. 3. The visualization plots of different manufacturers of oxytetracycline by PCA and *t*-SNE. (a) Three-dimensional map of the first three PCs; (b) the visualization plots by *t*-SNE.

TABLE 1. Parameters of GS, GA, and PSO SVMs

Chemometric method	Features selection method	Best *c*	Best *g*	Population	Iteration	Cross validation	CV accuracy, %	Trainings set accuracy, %	Test set accuracy, %
GS-SVM	PCA	5.278	0.32988	–	–	5	95	95.83	92.50
	t-SNE	1	1.7411	–	–	5	96.67	98.33	95.00
GA-SVM	PCA	7.2563	3.72	20	50	5	95.83	96.67	97.50
	t-SNE	46.8369	11.0758	20	50	5	97.50	98.33	97.50
PSO-SVM	PCA	7.6777	1.0215	20	50	5	97.50	98.33	97.50
	t-SNE	65.3903	6.9798	20	50	5	98.33	99.17	97.50



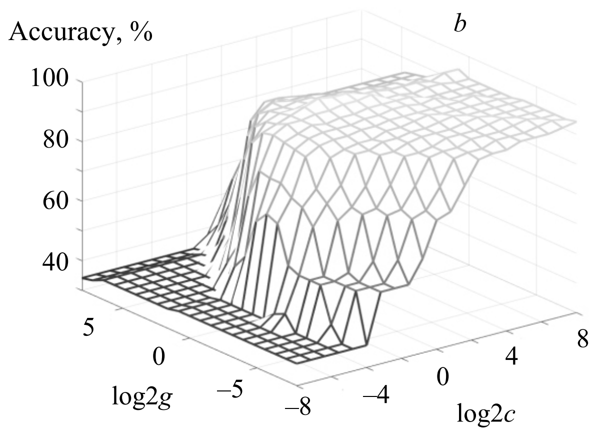


Fig. 4. Result of GS-SVM parameter selection: (a) PCA-GS-SVM and (b) *t*-SNE-GS-SVM.

The optimal combinations of g and c were investigated using a genetic algorithm (GA) and particle swarm optimization (PSO) to improve the prediction accuracy of the model [22, 23]. Again, fivefold cross validation was used after dimensionality reduction via PCA or *t*-SNE. The initial population size was set to 20, and the number of iterations was 50. The fitness curves for the GA-SVM using the two methods of dimensionality reduction are shown in Fig. 5. Specifically, Fig. 5a presents the fitness curve for PCA-GA-SVM, where the CV-accuracy was 95.83%, and Fig. 5b shows the fitness curve for *t*-SNE-GA-SVM, where the CV-accuracy was 97.50%. Unlike in the GA, PSO does not include crossover and mutation operators, and the global optimum is ultimately determined by tracking the current optimal value; as a result, the accuracy of the PSO approach is higher. Figure 6a shows the fitness curve for PCA-PSO-SVM, where the CV-accuracy was 97.5%, and Fig. 6b presents the fitness curve for t-SNE-PSO-SVM, where the CV-accuracy was the highest (98.33%).

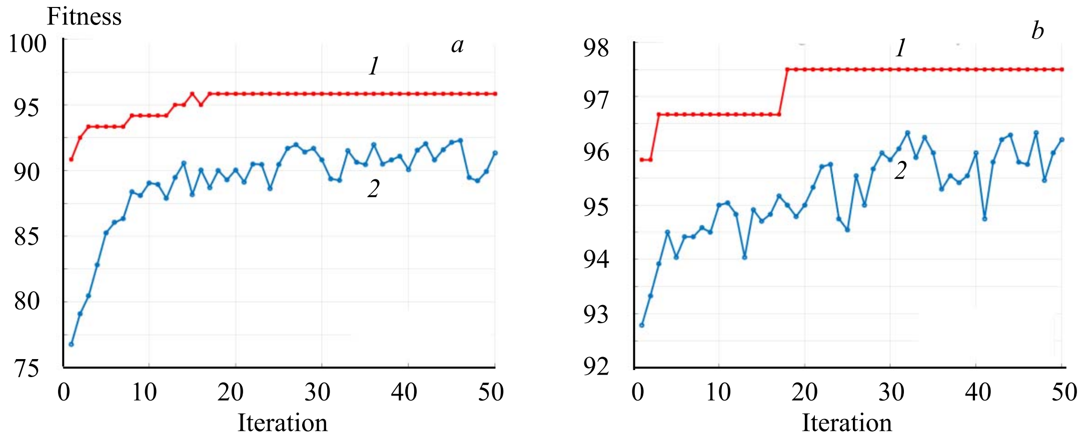


Fig. 5. GA fitness curves (a) PCA-GA-SVM; (b) *t*-SNE-GA-SVM;
1 – best fitness, *2* – Average fitness.

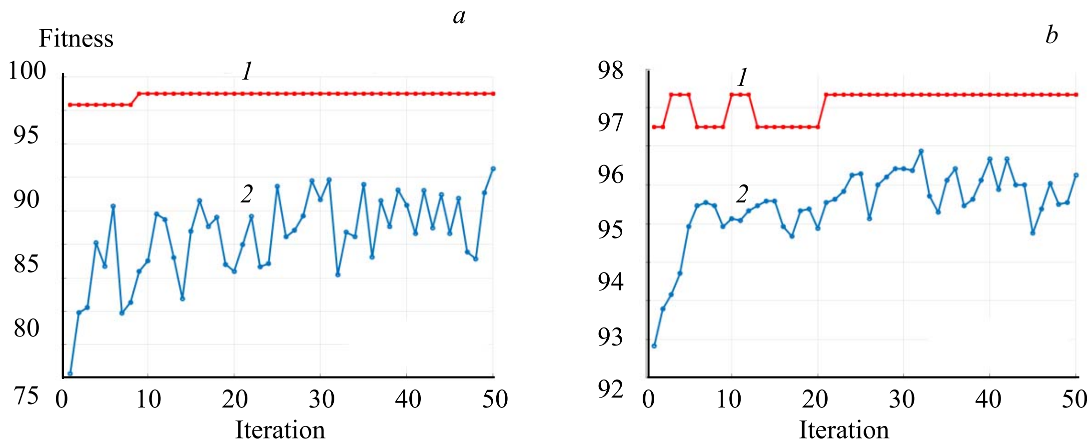


Fig. 6. PSO fitness curves (a) PCA-PSO-SVM; (b) *t*-SNE-PSO-SVM.
1 – best fitness, *2* – Average fitness.

As shown in Table 1, fivefold cross validation was performed to compare the three SVM models (GS-SVM, GA-SVM, and PSO-SVM). Following dimensionality reduction using *t*-SNE, the CV-accuracy was higher than that obtained using PCA. Additionally, with *t*-SNE, the optimal combinations of c and g were 1 and 1.7411, respectively, for GS-SVM; 46.8369 and 0.081825 respectively, for GA-SVM; and 65.3903 and 6.9798, respectively for PSO-SVM. The CV-accuracy with PSO (98.33%) was slightly higher than those of GS (96.67%) and GA (97.50%), the training set accuracy with PSO (99.17%) was higher than those of GS (98.33%) and GA (98.33%), and the test set accuracy with PSO (97.50%) was the same as GA (97.50%) but higher than GS (95.00%).

Classifiers were built by combining PCA or *t*-SNE with GS-SVM, GA-SVM, PSO-SVM, C4.5 [24], and RF [25]. Each of these classifiers was run 100 times to calculate the average training accuracy and average prediction accuracy, and the results are compiled in Table 2. For the same classifier, the use of *t*-SNE for dimensionality reduction always resulted in a higher training accuracy and prediction accuracy than the use of PCA. Based on these experiments, *t*-SNE outperformed PCA in terms of dimensionality reduction performance for analyzing THz absorption spectra of oxytetracycline. The three SVM models were compared with C4.5 and RF, and the PSO-SVM performed the best; its average training set accuracy was the highest (97.72%), and its average test set accuracy was also the highest (94.25%). Therefore, the *t*-SNE-PSO-SVM model was the most suitable model for recognizing THz spectra of oxytetracycline.

TABLE 2. Classification Results of all Chemometric Methods

Chemometric method	Features selection method	Mean accuracy in training set, %	Mean accuracy in test set, %
GS-SVM	PCA	93.27	89.78
	t-SNE	96.23	91.20
GA-SVM	PCA	93.58	90.78
	t-SNE	97.02	92.75
PSO-SVM	PCA	93.40	93.58
	t-SNE	97.72	94.25
C4.5	PCA	87.91	81.90
	t-SNE	92.95	88.80
RF	PCA	88.97	85.03
	t-SNE	93.45	90.02

Conclusions. The objective of this study was to trace the manufacturers of oxytetracycline tablets based on the THz time-domain spectra. The drugs produced by different manufacturers did not differ significantly, so the THz time-domain spectra and absorption spectra only displayed minor distinctions. Differentiating between the manufacturers based on THz spectra alone was not possible, so a chemometric strategy was adopted to construct and compare various recognition models. Principal component analysis and *t*-distributed stochastic neighbor embedding methods were used for feature selection and dimensionality reduction, respectively. These two methods were then combined with GS-SVM, GA-SVM, PSO-SVM, C4.5, and RF to build the classifiers. The optimal classifier was chosen based on parameter optimization and comparative analysis. The experiments demonstrated that *t*-distributed stochastic neighbor embedding outperformed Principal component analysis in terms of dimensionality reduction. Overall, among the 10 classifiers tested, *t*-SNE-PSO-SVM had the highest average training precision (97.72%) and the highest average prediction precision (94.25%); therefore, this model was the most suitable model for differentiating oxytetracycline tablets from different manufacturers. This study has indicated that the combination of THz-TDS and chemometrics has potential applicability for food and drug detection.

Acknowledgements. This work was supported by the National Defense Basic Scientific Research Program of China (JCKY2018404C007, JSZL2017404A001, JSZL2018204C002), Sichuan Science and Technology Program of China (2019YFG0114), and Graduate Student Innovation Fund of SWUST (19ycx0104).

REFERENCES

1. Y. Shinozuka, K. Kawai, A. Takeda, M. Yamada, F. Kayasaki, N. Kondo, Y. Sasaki, N. Kanai, T. Mukai, M. Sawaguchi, M. Higuchi, H. Kondo, K. Sugimoto, S. Kumagai, I. Murayama, Y. Sakai, K. Baba, K. Maemichi, T. Ohishi, T. Mizunuma, A. Kawana, A. Yasuda, A. Watanabe, *J. Vet. Med. Sci*., **81**, 863–868 (2019).
2. S. Sivakesava, J. Irudayaraj, *J. Dairy Sci*., **85**, 487–493 (2002).
3. S. C. Zhong, *Front. Mech. Eng*., **14**, 273–281 (2019).
4. L. H. Eadie, C. B. Reid, A. J. Fitzgerald, V. P. Wallace, *Expert Syst. Appl*., **40**, 2043–2050 (2013).
5. K. Lee, K. Jeoung, S. H. Kim, Y. B. Ji, H. Son, Y. Choi, Y. M. Huh, J. S. Suh, S. J. Oh, *Biomed. Opt. Express*, **9**, 1582–1589 (2018).
6. S. H. Baek, H. B. Lim, H. S. Chun, *J. Agric. Food Chem*., **62**, 5403–5407 (2014).

7. B. H. Cao, H. Li, M. B. Fan, W. Wang, M. Y. Wang, *Anal. Methods*, **10**, 5097–5104 (2018).
8. J. S. Melinger, N. Laman, D. Grischkowsky, *Appl. Phys. Lett.*, **93**, 3 (2008).
9. J. A. Zeitler, P. F. Taday, D. A. Newnham, M. Pepper, K. C. Gordon, T. Rades, *J. Pharm. Pharmacol.*, **59**, 209–223 (2007).
10. W. Limwikrant, K. Higashi, K. Yamamoto, K. Moribe, *Int. J. Pharm.*, **382**, 50–55 (2009).
11. H.-L. Zhang, Y. Xia, Z. Hong, Y. Du, *Spectrosc. Spectr. Anal.*, **35**, 1854–1859 (2015).
12. Z.-W. Zhang, J. Zuo, C.-L. Zhang, *Spectrosc. Spectr. Anal.*, **32**, 906–909 (2012).
13. L. J. Xie, C. Wang, M. Chen, B. B. Jin, R. Y. Zhou, Y. X. Huang, S. Hameed, Y. B. Ying, *Spectrosc. Acta A: Mol. Biomol. Spectr.*, **222**, 7 (2019).
14. W. Liu, C. Liu, J. Yu, Y. Zhang, J. Li, Y. Chen, L. Zheng, *Food Chem.*, **251**, 86–92 (2018).
15. H. Luo, J. Zhu, W. Xu, M. Cui, *Optik*, **184**, 177–184 (2019).
16. B. Qin, Z. Li, T. Chen, Y. Chen, *Optik*, **142**, 576–582 (2017).
17. H. Zhan, J. Xi, K. Zhao, R. Bao, L. Xiao, *Food Control*, **67**, 114–118 (2016).
18. R. Zhang, T. Wu, Y. Zhao, *Optik*, **183**, 906–911 (2019).
19. T. D. Dorney, R. G. Baraniuk, D. M. Mittleman, *J. Opt. Soc. Am. A, Optics, Image Science, and Vision*, **18**, 1562–1571 (2001).
20. L. Duvillaret, F. Garet, J.-L. Coutaz, *IEEE J. Select. Top. Quantum Electron.*, **2**, 739–746 (1996).
21. L. van der Maaten, *J. Mach. Learn. Res.*, **15**, 3221–3245 (2014).
22. T. Zhou, H. Lu, W. Wang, X. Yong, *Appl. Soft. Comput.*, **75**, 323–332 (2019).
23. X. K. Wang, S. Y. Guan, L. Hua, B. Wang, X. M. He, *Ultrasonics*, **91**, 161–169 (2019).
24. J. Gupta, J. Patrick, S. Poon, *Stud. Health Technol. Informatics,* **266**, 83–88 (2019).
25. F. B. de Santana, A. M. de Souza, R. J. Poppi, *Spectrosc. Acta A: Mol. Biomol. Spectr.*, **191**, 454–462 (2018).