Optik - International Journal for Light and Electron Optics 193 (2019) 163017

Contents lists available at ScienceDirect

Optik

journal homepage: www.elsevier.com/locate/ijleo

Original research article

Quantitative analysis of alum based on Terahertz time-domain spectroscopy technology and Support vector machine

Ai-hong Guan*, Yong-yang Chao

Henan University of Technology, College of Information Science and Engineering, Henan, Zhengzhou, 450001, China

ARTICLE INFO

Keywords:
THz-TDS
Alum
Quantitative analysis
PLS
SVM

ABSTRACT

In this paper, the Terahertz time-domain spectroscopy (THz-TDS) technique was used to quantitatively analyze the alum proportion in sweet potato starch. Terahertz time-domain spectra and frequency-domain spectra of pure sweet potato starch and pure alum and their mixtures were obtained. The absorption coefficient spectrum and refractive index spectrum of the samples were calculated. Results show that alum exhibit unique absorption peaks in the terahertz range, and can thus be identified with its absorption fingerprints in the detection from Sweet potato starch. Differences in the refractive index can also be observed between the alum and the sweet potato starch. The Partial Least Squares (PLS) and the Support vector machine (SVM) algorithm were used to establish the mathematical model between the absorption coefficient and the proportion. Results show that the prediction accuracy can reach 99.8% and 99.9%. Comparing the two predictive models, SVM has a better prediction effect. In all, the THz-TDS system plus SVM method is very promising for the further quantitative analysis of alum in sweet potato starch, with the characteristics of being nondestructive and laborsaving compared to other analytical tools.

1. Introduction

Alum ($KAL(so_4)_2.\ 12H_2o$) is a common food additive, which is widely used in the manufacture of starch food. Alum can achieve the effect of reinforcement in the process of vermicelli processing. Alum contains aluminum. Excessive intake of aluminum will directly harm people's health. According to China's national food safety standards for the use of food additives (GB2760-2011), the use of aluminum-containing food additives has been strictly regulated, and the ultimate aluminum residue in food should not exceed 100 mg/kg [1]. Alum can produce AL^{+3} when dissolved in water. Excessive intake of AL^{+3} will lead to Alzheimer's disease, osteoporosis, dyspepsia and other diseases. So, it is particularly important to find a scientific and rapid method to detect alum in sweet potato starch.

Terahertz spectroscopy causes little damage to the target material due to its low photon energy, and can provide rich intermolecular and low-frequency intramolecular modes of the chemicals. Terahertz pulse width is in picosecond range and has high time resolution. Comparing with infrared detection, X-ray detection, UV detection, Terahertz detection has many advantages [2–4]. Due to its uniqueness, this technique has been qualitatively and quantitatively used in many fields [5–8]. YF Hua et al. qualitative and quantitative detected pesticides and achieved the relative error of less than 5% predicting the weight ratio [9]. Quantitative analysis of the effective content of 2-mercaptobenzothiazole was carried out by Yin X et al. using THz-TDS technology. [10] Zhang F et al. used THz-TDS technology to achieve quantitative detection of plasticizers [11]. Fang H et al. achieved quantitative analysis of

* Corresponding author.
E-mail address: oe_haut@126.com (A.-h. Guan).

https://doi.org/10.1016/j.ijleo.2019.163017
Received 11 April 2019; Accepted 26 June 2019

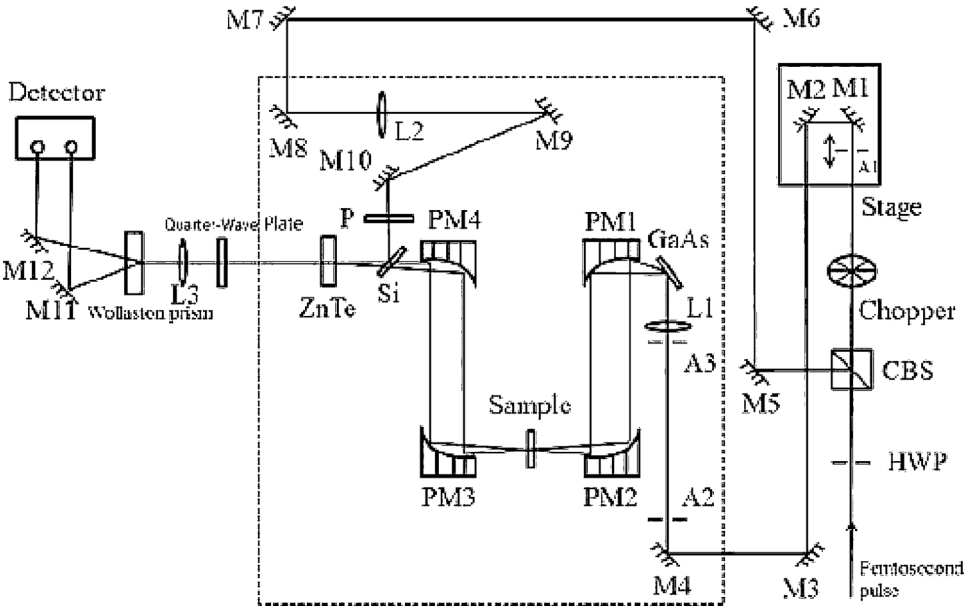

Fig. 1. Schematic diagram of the THz-TDS system.

additives in wheat by THz-TDS [12]. Fu X et al. proved that the THz-TDS technology can be used to detect wheat with different amount of talc powder quickly and effectively [13]. In this paper, Terahertz time-domain spectra of alum, sweet potato starch and their mixture were obtained, and the frequency-domain spectra were obtained by Fourier transform. Then the frequency-domain waveform data is processed to obtain the absorption coefficient spectra and refractive index spectra of different samples. The relationship model between the percentage of alum in sweet potato starch and the absorption coefficient are established. The quantitative analysis of alum in sweet potato starch was achieved by those models.

2. Experimental setup and sample preparation

The experimental setup is comprised of a THz-TDS system Z3-XL from Zomega Terahertz Corporation (Z3, USA) and a femtosecond fiber laser. Fig. 1 shows the schematic diagram of the THz-TDS system in the transmission mode. The femtosecond fiber laser was supplied by Toptica (FFPRO780, Germany) with average power 150 mw, pulse duration 100 fs, and central wavelength 780 nm. The input laser beam is split into a pump beam and a detection beam. After transmitting through the delay line, the pump beam elicits a terahertz (THz) beam at the emitter which is composed of a photo-conductive antenna. Then the THz beam is focused onto the sample by two parabolic mirrors. Again, with two parabolic mirrors, the sample-characteristic-carried terahertz beam meets the detection beam at the ZnTe crystal and is detected through the electro-optic effect.

Alum and sweet potato starch were supplied by Aladdin Reagent (Shanghai) Co. Ltd. To avoid moisture absorption in the air, the THz-TDS system was filled with pure dry nitrogen during the experiments, to maintain the humidity less than 1%. All the mixtures were carefully mixed by a vortex mixer and ground using a pestle and mortar, and were then pressed into about 1-mm-thick disks with a hydraulic press under the pressure of approximately 8 MPa. Five varieties of samples with alum percentage of 0%, 10%, 20%, 30% and 100% were prepared. Three samples were made for each weight ratio, totaling 15 samples whose parameters were gave on Table 1.

Provided that there is no light reflection, scattering or fluorescence but absorption effect in the transmission of THz pulses, and the incidence of the terahertz beam is 90°. The time-domain reference and sample spectra were denoted as R (t) and S (t) respectively. According to Lambert-Beer's law and the physical model of optical parameter extraction proposed by TD. Dorney and L. Duviaret [14], we can calculate the $\rho(\omega)$ and $\varphi(\omega)$ by Eq. (1).

Table 1
Sample parameters.

	Sample 1	Sample2	Sample3	Sample4	Sample5
thickness(mm)1	0.84	0.849	0.932	1.008	1.07
thickness(mm)2	1.2	1.14	1.09	1.14	1.135
thickness(mm)3	1.132	1.169	1.136	1.15	1.198
percentage (%)	0%	10%	20%	30%	100%

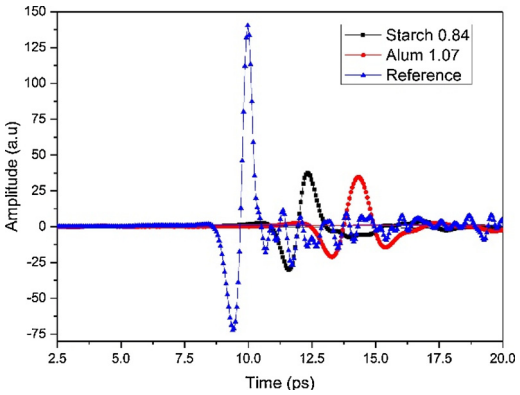

Fig. 2. THz time-domain spectra of the reference and samples.

$$\frac{S(\omega)}{R(\omega)} = \rho(\omega)e^{-j\varphi(\omega)} \tag{1}$$

R(ω) and S(ω) are respectively obtained by the Fourier transform of R(t) and S(t). Then by Eq.s (2),(3),(4) the refractive index and absorption coefficient of the samples can be calculated.

$$n_b(\omega) = \Phi(\omega)\frac{c}{\omega d} + 1 \tag{2}$$

$$k_b(\omega) = \frac{c}{\omega d}\ln\frac{4n_b(\omega)}{\rho(\omega)[n_b(\omega)+1]^2} \tag{3}$$

$$\alpha_b(\omega) = \frac{2k_b(\omega)\omega}{c} = \frac{2}{d}\ln\frac{4n_b(\omega)}{\rho(\omega)[n_b(\omega)+1]^2} \tag{4}$$

Φ(ω) is the phase difference between the samples and the reference signal, ρ(ω) is the amplitude ratio of mixed samples and reference signal, d is sample thickness, c is the light velocity, and ω is the angular frequency.

3. Experimental results

To reduce the influence of errors, each sample was tested three times. Fig. 2 gives the THz time-domain spectra of pure starch with thickness of 0.84, pure alum with thickness of 1.07 mm and air (reference signal). Because the refractive index of starch and alum is greater than that of air, it can be seen from Fig. 2 that the time domain spectrum of alum and starch has a certain delay relative to the reference signal.

Fig. 3 gives the frequency-domain spectrum of samples after Fourier transformation of Fig. 2. Due to the strong absorption of these samples, the effective frequency range is reduced to 0.2–1.6 THz. Fig. 4 shows the absorption coefficient of Alum and sweet potato starch. From this figure, we can see that alum has obvious absorption peaks at 0.9523 THz, 1.0474 THz, and 1.1179 THz but starch is not so obvious. So, alum can thus be easily discriminated from sweet potato starch.

Fig. 5 shows the absorption coefficient spectra of different alum mass fraction. The fig shows obvious absorption peaks at 0.9523 THz, 1.0474 THz, and 1.1179 THz, and the position of the absorption peak does not change with the increase of alum content,

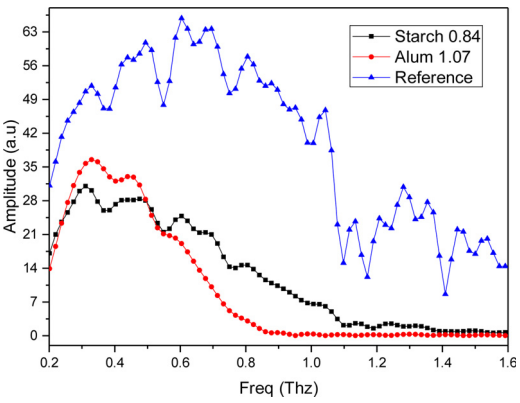

Fig. 3. THz Frequency-domain spectra of the reference and samples.

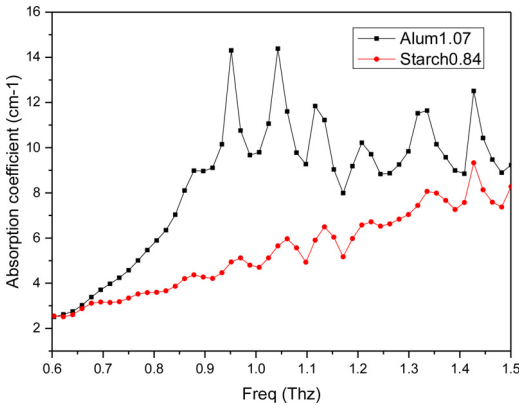

Fig. 4. Absorption coefficient spectrum of samples.

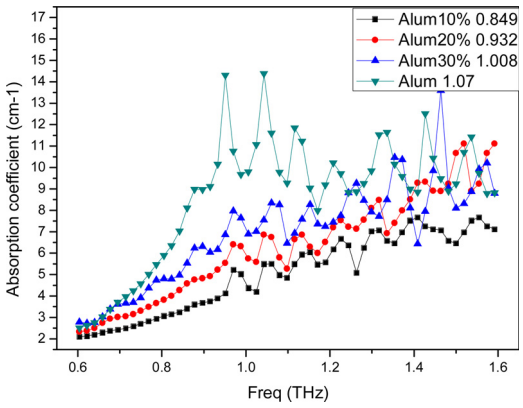

Fig. 5. Absorption coefficient spectrum of percentage of alum.

but the amplitude increases with alum percentage. Therefore, the corresponding prediction model can be established according to the absorption spectrum of the sample in this frequency band to quantitatively analyze the content of alum in the sample [15].

Fig. 6 gives the refractive index for different percentage of alum. It shows the refractive index of alum is higher than the starch's, and the refractive index increases with alum percent. So, it can be known whether the sweet potato starch contains alum or not qualitatively.

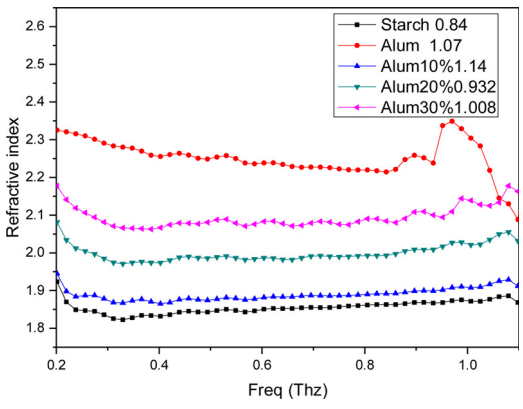

Fig. 6. The refractive index of different percentage of alum.

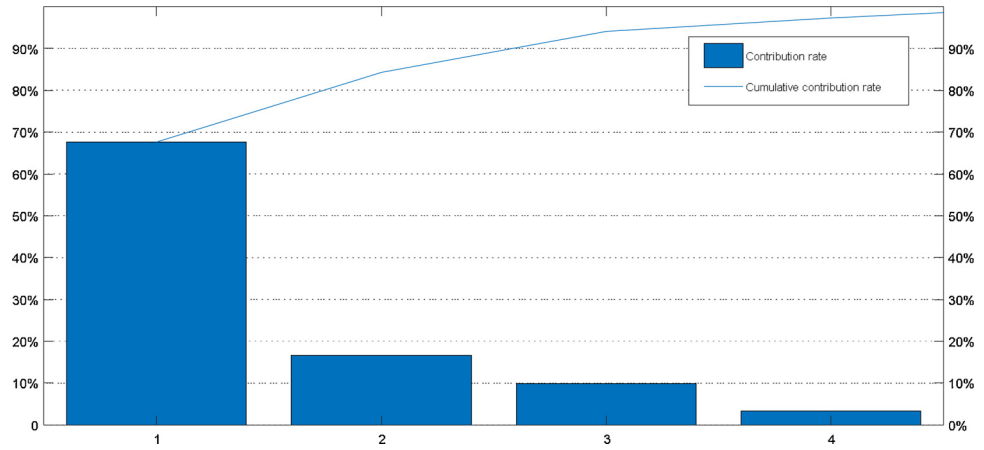

Fig. 7. Cumulative contribution rate of each principal component.

4. Quantitative analyses

4.1. PLS model

Partial least squares (PLS) is a typical multivariate statistical analysis method, especially for small sample data [16,17]. The whole frequency range 0.6˜1.6 THz is selected for PLS modeling. Five types of samples with alum percentage of 0%, 10%, 20%, 30%, and 100% were prepared respectively. Three samples were made for each type and a total of fifteen samples were prepared. Each sample was measured 3 times, so there are 45 sets of data in all. Among them, 2 sets were randomly selected from the 3 sets of data of each sample as the calibration set, and the remaining 15 sets of data were used as the prediction set.

Data preprocessing and principal component selection are carried out first. The pretreatment is realized by the standardized processing method as formula (5)

$$X^{'} = \frac{(X - \tilde{X})}{\sigma} \tag{5}$$

Where $X^{'}$、 X、 $\tilde{X}$ and σ are the original data after standardization, the original data, the averaged values of the original data, and the standard variance of the original data.

After pretreatment, the data obeys the standard normal distribution, and has the mean and variance of the standard normal distribution data, which is convenient for data processing. In addition, the data is standardized, which avoid the impact of different dimensions between the data. After standardization, principal component analysis was used to analyze the data to reduce the number of principal components. Fig. 7 shows the contribution rate of each principal component after principal component analysis. Table 2 gives cumulative contribution rate of each principal component.

The cumulative contribution rate of the first four principal components reached 97.35%, which contain almost all the information of the original data. So, the first four principal components were extracted as the processing data in the experiment, and X, Y and B, which are mapped to the principal component space, are calculated by formula (6)–(8).

$$X_c = X^{*}CX \tag{6}$$

$$Y_c = Y^{*}CY \tag{7}$$

$$Y_c = X_c{}^{*}B_c + E_c \tag{8}$$

Where X_c、 Y_c、 B_c are absorption coefficient matrices mapped to the principal component space respectively, CX and CY are the load matrices of X and Y.

Fig. 8 shows the results predicted by the PLS model, where the vertical axis is the predicted percentage of alum in the sample, and the horizontal axis is the actual percentage. The predicted result is closely distributed near the actual values, which shows that the prediction results are very close to the actual values.

Table 2
Cumulative contribution rate of each principal component.

Principal components	First	Second	third	Fourth
Cumulative contribution rate	67.56%	84.25%	94.12%	97.35%

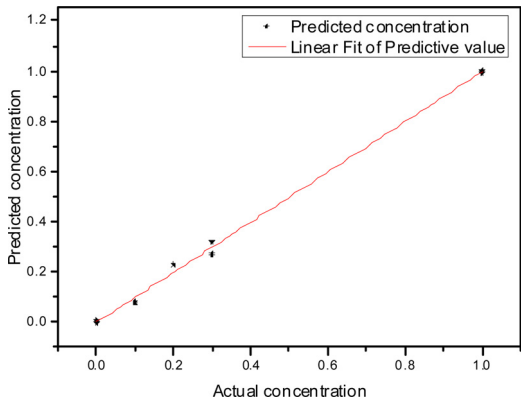

Fig. 8. The fitted curve between predicted and actual values.

Table 3
The model parameters under different kernel functions.

Kernel functions	Linear	Polynomial	RBF	Sigmoid
RMSE	0.0068	0.3517	0.1961	0.2464
R	0.9998	0.9480	0.9688	0.9306

Table 4
Main parameters of the SVR model.

Model	Kernel function	c	g
SVR	RBF	16	0.0313

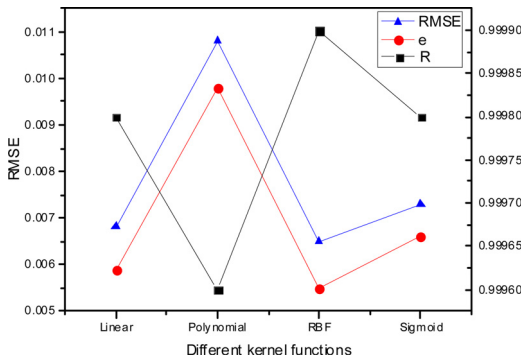

Fig. 9. Model parameter values under different kernel functions.

4.2. SVM model

Another statistical method SVM was used to build a more accuracy model to quantitatively analyze alum in starch. In our study, the absorption coefficients of different samples were taken as the independent variable X. In the effective frequency band of 0.6˜1.6 THz, 55 frequency points were taken, so the dimension of the independent variable X is 55. The percentage of the alum was the dependent variable Y and was used as the label of the sample. The same data preprocessing and principal component selection are carried out as in PLS.

The most important thing in SVR regression analysis is the selection of the kernel function and the values of the main parameters c and g. SVR provides four kinds of kernel functions including linear kernel function, polynomial kernel function, radial basis kernel function (RBF) and sigmoid kernel function. In order to compare the effects with different kernel functions under the same parameters, default parameter values of c and g are selected. Table 3 shows the parameters of the model are obtained by training with different kernel functions.

It can be seen from Table 3 that the mean square error of the linear kernel function and the radial basis kernel function regression prediction is smaller and the correlation coefficient is larger, when c and g are the default values of the system. So, the linear kernel

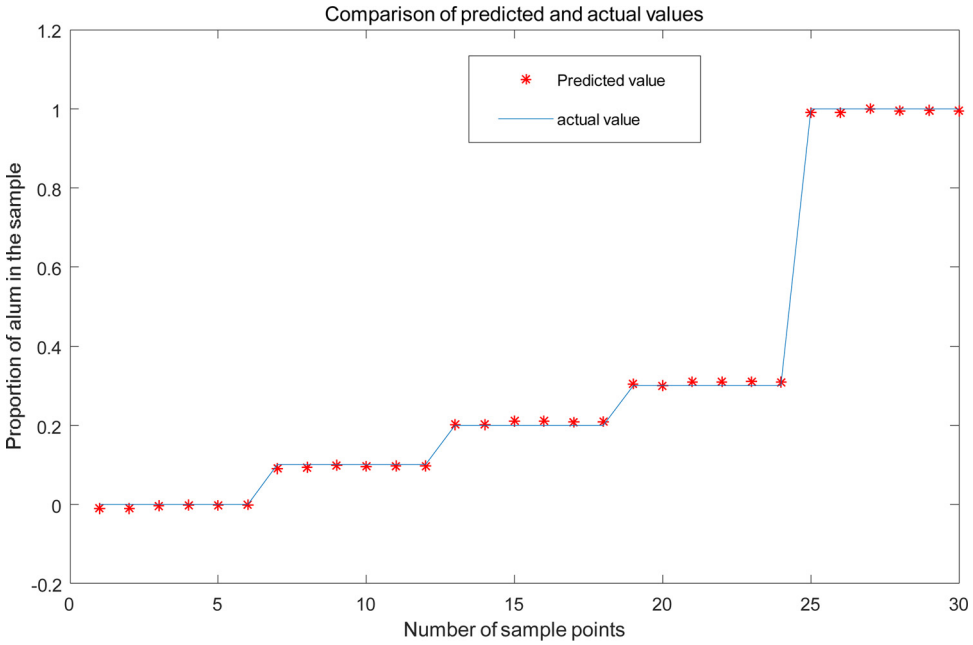

Fig. 10. the prediction results of the training set predicted by the SVM model.

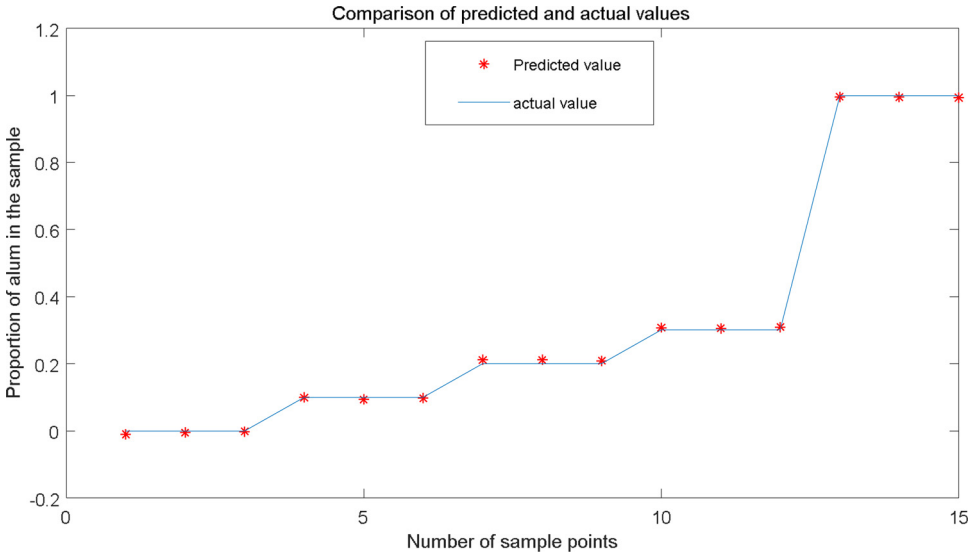

Fig. 11. the prediction results of the test set predicted by the model.

function and the radial basis kernel function can be used as the kernel function of the SVR model in this experiment.

After the parameter search is completed by grid search method, the best value of c is 16, and g is 0.0313. Put the optimal parameter values into the model and established the model by different kernel functions. According to the correlation coefficient R, e and the RMSE of the prediction result,the best kernel function was determined. Table 4 gives main parameters of the SVR model. The correlation coefficient R, e and RMSE under different kernel functions are shown in Fig. 9.

The left Y-axis is the RMSE and e, and the right Y-axis is the correlation coefficient R. It can be seen from Fig. 9 that the RFB function has the highest correlation, the smallest root means square error and the optimal prediction effect among the four kernel functions.So, the model uses the radial basis kernel function as the kernel function of the model.

After determining the parameters, the training set data is input for training to obtain the SVR regression model, then, the data in the training set is brought into the model for prediction. The prediction results are shown in Fig. 10.

It can be seen from Fig. 10 that the prediction values of the 30 sets of data in the training set are basically coincident with the actual values, which indicated the effectivity of the model. Further input 15 sets of data into the model that has been trained, and the predicted result is shown in Fig. 11. From the prediction results, the predicted values are closely distributed around the actual values,

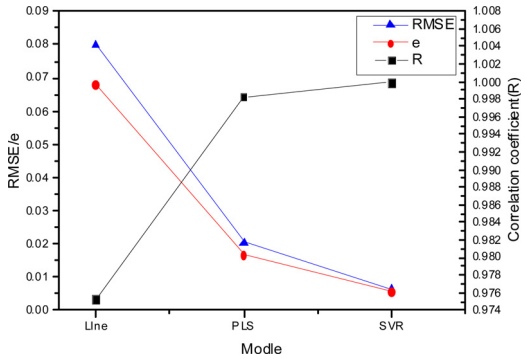

Fig. 12. Comparison of prediction results between different models.

Table 5
Results of different prediction models.

Parameter	RMSE	e	R	Fitting curve equation
PLS	0.0205	0.0169	0.9983	Y = 0.0025 + 0.9983X
SVR	0.0065	0.0055	0.9999	Y = -0.0018 + 1.0036X

which fully demonstrate the feasibility of quantitative analysis of alum in starch by SVR.

The comparison of the results obtained by training and predicting the same data in the PLS model and the SVR model are shown in Fig. 12. It can be seen from Fig. 12 the RMSE or e predicted by the SVR model are smaller than the PLS, and the correlation of the SVR model after training prediction is higher than that of the PLS model. Table 5 gives the results of different prediction models. So SVR model is more accurate in Quantitative Detection of alum in sweet potato starch.

5. Conclusion

In this paper, the THz time-domain spectra of alum and sweet potato starch and their mixture are obtained by THz-TDS system. The frequency-domain spectra are obtained by Fourier transform. The absorption coefficient spectra and refraction index spectra of samples were calculated. Alum's absorption coefficient spectra showed obvious absorption peaks at 0.9523 THz, 1.0474 THz, and 1.1179 THz, and the position of the absorption peak does not change with the increase of alum content, but the amplitude increases with alum percentage. The refraction index showed obvious difference between alum and starch. The relationship between the absorption coefficient and the percentage of alum was established by the PLS and SVR model. Results showed that the accuracy of PLS model for the quantitative analysis of alum content in starch was over 99.8%, and the quantitative analysis accuracy of SVR model reaches 99.9%. Both models are effective to achieve quantitative analysis of alum. This paper provides two accurate methods for alum quantitative detection. Comparison results showed that the SVR model has better predictive analysis effect and reveals the potential application of THz-TDS technology in food addition quantitative detection field.

References

[1] G B 15202-2003 National Standard for Aluminum in Dietary Foods, (2003).
[2] P. Jepsen, D. Cooke, M. Koch, Terahertz spectroscopy and imaging-modern techniques and applications, Laser Photonics Rev. 5 (2011) 124–166.
[3] Q. Wang, Y.H. Ma, Qualitative and quantitative identification of nitrofen in terahertz region, Chemom. Intell. Lab. Syst. 127 (2013) 43–48.
[4] X.B. Huang, P.J. Huang, X. Li, Y.H. Ma, D.B. Hou, G.X. Zhang, Analysis of terahertz time domain spectroscopy of mixtures based on indirect hard modeling method, Spectrosc. Spectral Anal. 37 (10) (2017) 3021–3026.
[5] Z.Y. Wang, K. Kang, S.B. Wang, Determination of plane stress state using terahertz time-domain spectroscopy, Sci. Rep. 6 (2016) 36308.
[6] Zhang Huo, Li Zhi, Terahertz spectroscopy applied to quantitative determination of harmful additives in medicinal herbs, Optik 156 (2018) 834–840.
[7] Liang Jie, Guo Qijia, Chang Tianying, Reliable origin identification of Scutellaria baicalensis based on terahertz time-domain spectroscopy and pattern recognition, OPTIK 174 (2018) 7–14.
[8] Li. Zhu, Study on Food Additive Detection Technology based on Terahertz Wave, Hang Zhou: Zhe Jiang University, 2008.
[9] Y.F. Hua, H.J. Zhang, Qualitative and quantitative detection of pesticides with terahertz time-domain spectroscopy, IEEE Trans. Microwave Theory Tech. 58 (7) (2010) 2064–2070.
[10] Xian-hua Yin, Yan Jiang, Bin-chuan Lv, Quantitative study of terahertz time-domain spectra of 2-mercaptobenzothiazole, Laser Technol. 43 (01) (2019) 83–87.
[11] F. Zhang, Li-ping Liu, Mao-jiang Song, Quantitative analysis of plasticizer based on terahertz time domain spectroscopy, Laser Optoelectron. Prog. 54 (03) (2017) 308–315.
[12] H. Fang, Q. Zhang, H. Zhang, Detecting the terahertz time domain spectrum of Azo formamide in wheat flour, J. Chin. Cereals Oils Assoc. 31 (01) (2016) 107–111.
[13] X. Fu, W. Li, W. Xia, THz spectral detection of doped talc powder in wheat flour, J. Chin. Cereals Oils Assoc. 28 (03) (2013) 110–112.
[14] L. Duvillaret, F. Garet, J. Coutaz, A reliable method for extraction of material parameters in terahertz time-domain spectroscopy, IEEE J. Sel. Top. Quantum Electron. 2 (1996) 739–746.
[15] A. Guan, Z. Li, H. Ge, Qualitative and quantitative terahertz time domain spectroscopic detection of additive alum in sweet potato starch, Spectrosc. Spectral Anal. 38 (1) (2018) 267–270.
[16] J. Wang, Z. Zhang, Z. Zhang, Application of partial least squares and THz-TDS in identification of genuine rhubarb, Spectrosc. Spectral Anal. 36 (02) (2016) 316–321.
[17] H. Zhang, Z. Li, Terahertz spectroscopy applied to quantitative determination of harmful additives in medicinal herbs, Optik 156 (2018) 834–840.