

Detection and identification of three types of kaolin using terahertz time-domain spectroscopy

Qi Sun(孙齐), Shuguang Miao(苗曙光), Qiuyue Zhang(张秋悦), Shuangwu Shi(施双伍), Xiaonuo Hou(侯晓诺)
Citation: Chin. Phys. B. 35, 094101 (2026). DOI: 10.1088/1674-1056/ae1c1d | CSTR: 32038.14.CPB.ae1c1d
http://cpb.iphy.ac.cn; http://iopscience.iop.org/cpb

What follows is a list of articles you may be interested in

Exploring protein conformations by cluster-guided iterative multiple independent molecular dynamics simulations

Chengtao Ding(丁成涛), Guanglin Chen(陈光临), Qingguo Gong(龚庆国), Zhiyong Zhang(张志勇)
Chin. Phys. B. 35, 058702 (2026). DOI: 10.1088/1674-1056/ae0898 | CSTR: 32038.14.CPB.ae0898

Time delay interferometry and principal component analysis for noise cancellation of TianQin

Xinchun Hu(胡新春), Yan Wang(王炎)
Chin. Phys. B. 35, 050401 (2026). DOI: 10.1088/1674-1056/ae12d4 | CSTR: 32038.14.CPB.ae12d4

Phase transition extracted by principal component analysis in the disordered Moore–Read state

Na Jiang(江娜), Shuaixin Fu(付帅鑫), Zhengzhi Ma(马正直), Lian Wang(王莲)
Chin. Phys. B. 34, 047306 (2025). DOI: 10.1088/1674-1056/adb25f | CSTR: 32038.14.CPB.adb25f

---

Detection and identification of three types of kaolin using terahertz time-domain spectroscopy

Qi Sun(孙齐)[1,2], Shuguang Miao(苗曙光)[1,2,†], Qiuyue Zhang(张秋悦)[1,2], Shuangwu Shi(施双伍)[1,2], and Xiaonuo Hou(侯晓诺)[1,2]

[1] *School of Physics and Electronic Engineering, Huaibei Normal University, Huaibei* 235000, *China*
[2] *Anhui Province Key Laboratory of Intelligent Computing and Applications, Huaibei Normal University, Huaibei* 235000, *China*

(Received 31 July 2025; revised manuscript received 16 October 2025; accepted manuscript online 6 November 2025)

Kaolin, as a versatile non-metallic mineral resource, is widely used in ceramics, casting, construction, and other fields. To rapidly and accurately identify and classify three polymorphs of kaolin, while addressing the challenge of aligning resource characteristics with application requirements, a novel method for kaolin identification and classification has been proposed. This method employs terahertz time-domain spectroscopy (THz-TDS) in conjunction with cluster analysis (CA) and principal component analysis (PCA). It facilitates the extraction of spectral data, the calculation of Euclidean distances, and effective dimensionality reduction of the dataset. Terahertz spectroscopy, a non-destructive analytical technique, was employed to obtain terahertz spectra from three kaolin specimens using a transmissive terahertz spectrometer. The refractive index, dielectric constant, and absorption coefficient of the samples within the frequency range of 0.5 THz to 2.75 THz were calculated from the THz-TDS data, utilizing established equations, including the fast Fourier transform. Using all available refractive index and absorption coefficient data within this frequency band as input variables, CA and PCA were conducted separately to determine the Euclidean distance and the first principal component (PC1) for the corresponding samples. The findings reveal significant variations in the refractive index and absorption coefficient among the analyzed samples. PCA identified three primary components for both the refractive index and absorption coefficient, which accounted for cumulative contribution rates of 99.99% and 97.26%, respectively. CA categorized the samples into three distinct groups based on Euclidean distance metrics, with the CA clusters closely aligning with the PCA results. The sandy kaolin (SZGLT-GX325) exhibited the highest PC1 score and the greatest Euclidean distance from the second group, measuring 156.16. Utilizing the CA-PCA model, the three types of kaolin were accurately identified and classified based on inter-sample variations, achieving a classification accuracy of 99.99%. This study underscores the utility of THz-TDS combined with chemometric techniques, providing a robust framework for the rapid, precise, and non-destructive analysis of kaolin, thereby demonstrating significant potential for practical applications.

Keywords: terahertz time domain spectroscopy technology, identification of three types of kaolin, principal component analysis, cluster analysis

PACS: 41.90.+e, 29.30.–h, 33.20.–t, 33.20.Kf

DOI: 10.1088/1674-1056/ae1c1d **CSTR:** 32038.14.CPB.ae1c1d

1. Introduction

China was the earliest country in the world to discover and utilize kaolin, and has proven reserves of 3.496 billion tons. Kaolin, also known as china clay, is an important non-renewable non-metallic mineral resource. The chemical formula of kaolin is $Al_2O_3{\cdot}2SiO_2{\cdot}2H_2O$. It is mainly composed of kaolinite clay minerals, quartz, and feldspar, and includes three types: soft, sandy, and hard. Soft kaolin reserves are approximately 209.76 million tons, mainly distributed in Jiangxi and Hunan provinces. Sandy kaolin reserves are approximately 2.0976 billion tons, primarily distributed in Guangdong and Fujian provinces. Hard kaolin reserves are approximately 174.8 million tons, mainly distributed in Inner Mongolia Autonomous Region and Anhui province of China. With the expansion of its applications, different types of kaolin have been shown to have their own characteristics and methods of development and utilization. Therefore, the accurate identification and classification of kaolin can avoid resource misallocation and waste, greatly improving the utilization efficiency of kaolin resources, ensuring national resource security and efficient utilization, and promoting the sustainable and high-quality development of China's economy.

Terahertz time-domain spectroscopy (THz-TDS) is a non-destructive spectroscopic technique[1–4] with significant potential across various fields, including medicine, manufacturing, and communications. Terahertz radiation is characterized by electromagnetic waves that span frequencies from 0.1 THz to 10 THz, corresponding to wavelengths of 0.03 mm to 3 mm, and occupies the spectral region between the microwave and infrared bands.[5–7] The terahertz spectra of various substances possess distinct fingerprint characteristics. Analyzing these spectra can yield valuable information regarding the absorption coefficient, refractive index, and other properties of these substances. Additionally, different types of kaolin

†Corresponding author. E-mail: msgmcu@126.com

© 2026 Chinese Physical Society and IOP Publishing Ltd. All rights, including for text and data mining, AI training, and similar technologies, are reserved.

demonstrate varying responses within the terahertz frequency range. This study examines terahertz data from three types of kaolin using two multivariate statistical analysis techniques: cluster analysis (CA)[8,9] and principal component analysis (PCA).[10–13] The analysis includes a quantitative assessment of the Euclidean distance[14,15] diagram and the first principal component[16–19] score data to determine similarities and differences among the kaolin samples, facilitating their classification.

2. Experimental details

2.1. Terahertz time domain spectroscopy system

The THz-TDS[20] system is a crucial experimental apparatus employed to investigate the interaction of terahertz band electromagnetic radiation with matter. This system utilizes ultrashort laser pulses[21] to both excite and detect the time-domain waveform of terahertz waves, thus enabling the acquisition of the sample's frequency-domain spectral characteristics. Based on the experimental setup, these configurations are typically classified into several types, including transmissive, reflective, differential, and ellipsometric.[22,23] Among these, transmissive and reflective configurations represent the two fundamental architectures for THz-TDS and are the most widely applied. This study employs the Advantest TAS7400SU, a high-performance transmissive[24,25] terahertz spectroscopy system. Its spectral range is 0.5–7.0 THz, with a frequency resolution of 7.6 GHz. This article uses its transmission module. The device's architecture is illustrated in Fig. 1.

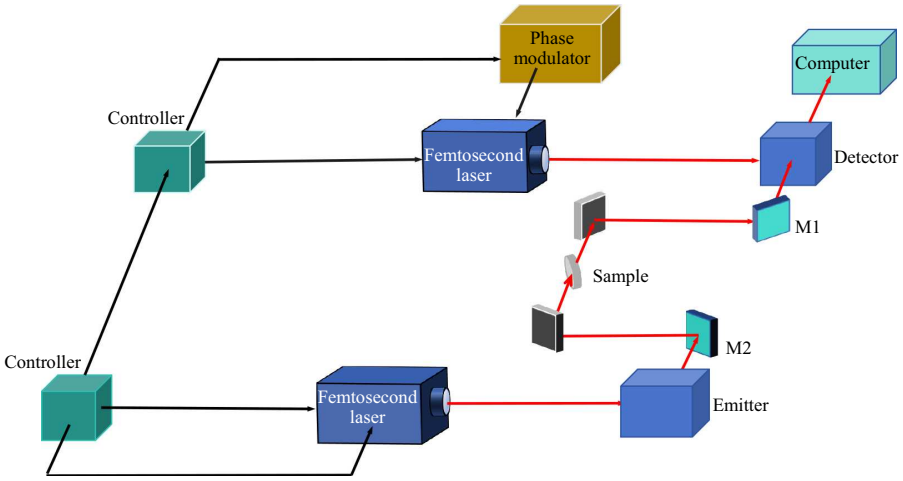


Fig. 1. Structural diagram of the THz-TDS system.

2.2. Materials

The three types of kaolin used in sample preparation for the experiment were sourced from Guangxi (soft kaolin, RZGLT-GX325; sandy kaolin, SZGLT-GX325) and Inner Mongolia Autonomous Region of China (hard kaolin, YZGLT-NM325). Their chemical compositions and physical properties are presented in Table 1. To enhance the accuracy and reproducibility of detection, and to mitigate tablet fracture caused by insufficient sample thickness during tableting, high-density polyethylene powder[26] with minimal terahertz wave absorption is incorporated before compressing the kaolin sample powder into tablets using a tablet press. High-density polyethylene is commonly used as a diluent in terahertz spectroscopy measurements. It exhibits no characteristic absorption and causes no chemical or physical interference. It possesses a certain degree of light transmittance in the terahertz frequency band and does not affect the subsequent data interpretation during feature extraction and CA. This facilitates the accurate detection of sample information carried by terahertz waves. We combine kaolin powder and polyethylene in a 1:5 ratio. Subsequently, we place the mixture into a tablet mold and compress it at a pressure of 20 MPa for a duration of 3 min to produce a sample with a thickness of 1.2 mm. This method effectively mitigates issues related to tablet brittleness, cracking, and agglomeration. The experiment should be conducted at 25 °C. The optical path section must be purged with dry gas to reduce the relative humidity to below 5% RH and ensure environmental desiccation. The entire process requires the samples to be air-dried, pulverized, ground, and sieved using a 200-mesh screen, and must be consistent with the reference samples. Simultaneously, three independent samples were selected for each material, resulting in a total of nine samples. Multiple measurements were taken for each sample, and their averages constituted independent data points for subsequent CA. The nine experimental samples are designated as follows: RZGLT A1, RZGLT A2, RZGLT A3, SZGLT B1, SZGLT B2, SZGLT B3, YZGLT C1, YZGLT C2, and YZGLT C3.

Table 1. Basic chemical composition of three kaolin types and their respective parameter contents.

Sample name	SiO_2 (%)	Al_2O_3 (%)	Fe_2O_3 (%)	CaO (%)	K_2O (%)
RZGLT-GX325	52.15	32.88	0.33	0.18	2.65
SZGLT-GX325	57.78	37.41	0.37	0.08	0.57
YZGLT-NM325	50.75	33.78	0.32	0.19	2.55

2.3. Terahertz optical parameter extraction and algorithmic principles

Due to the differential absorption of terahertz waves by various materials, when terahertz electromagnetic waves are incident upon samples within the terahertz frequency range, the constituent molecules, atoms, or crystal lattices absorb the energy of the terahertz electromagnetic waves. The transmitted terahertz waves convey information about the sample, generating a terahertz spectrum. This spectrum is subsequently transformed into the sample's frequency-domain signal via a fast Fourier transform. Finally, the refractive index, absorption coefficient, and extinction coefficient of the kaolin sample are computed using the equations.[27–30] The calculation equations are presented below:[31,32]

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

$$\alpha(\omega) = \frac{4\pi k(\omega)v(\omega)}{c}, \tag{2}$$

$$k(\omega)) = \frac{c}{\omega d}\ln\left(\frac{4n(\omega)}{\rho(\omega)\left[n^2(\omega)+1\right]^2}\right), \tag{3}$$

where $\Delta\varphi(\omega)$ denotes the phase difference between the sample signal and the reference signal; d represents the sample thickness (in meters); c is the speed of light (in meters per second); ω is the angular frequency (in radians per second); and $\rho(\omega)$ is the amplitude ratio of the sample signal to the reference signal; where $v(\omega) = \omega/2\pi$ is the frequency.

Following acquisition of the requisite terahertz data, encompassing absorption coefficients and refractive indices, PCA and CA were employed to conduct an objective, multivariate[35,36] data analysis via dimensionality[33,34] reduction, with the aim of investigating the similarities and distinctions among the three kaolin samples. PCA is a statistical dimensionality reduction technique that uses linear transformation to convert multiple original variables into a smaller set of mutually independent principal components. This process transforms the data from the original space to the principal component space. The resulting principal components preserve as much of the original data variance as possible. Calculation of the data also yields the eigenvalues and eigenvectors.[37] The weights of the principal components are determined by the magnitude of the eigenvalues, with the first principal component capturing the largest proportion of the sample variance. Plotting the scores of the first principal component in a histogram allows for clear observation of the similarities and differences between samples, thus enabling classification. CA is an unsupervised algorithm that classifies samples within a dataset into distinct clusters[38] without any prior categorization of the dataset. Samples within the same cluster exhibit high similarity, whereas those in different clusters demonstrate significant differences. This facilitates automated grouping and clustering. In sample analysis, the Euclidean distance between different feature samples is computed. For the three kaolin samples, those with varying distances are grouped into separate clusters. For the three samples measured three times, the three samples with the closest distances are grouped into one cluster, and this process is iterated with the remaining samples. This process is repeated until all samples are grouped into a single cluster. Both CA and PCA rely on the examination of the sample data's inherent characteristics. Through the construction of simplified models, these methods can streamline data processing and reduce complexity, thereby enhancing analytical efficiency.

3. Results and discussion

3.1. Qualitative analysis of terahertz time-domain spectroscopy

Measurements were conducted on the experimental samples in accordance with the experimental protocol. Each sample was measured in triplicate, yielding three data sets. To minimize experimental error, the average of these three sets was selected as the reference analysis data for THz-TDS[39] of several kaolin samples. This terahertz time-domain spectrometer has a measurement range of 0–8 THz. For this experiment, the frequency band selected was 0.5–2.75 THz, yielding the waveforms shown in Fig. 2(a). Significant differences in terahertz spectra were observed between different kaolin samples, as were variations in peak values. Of these, SZGLT-GX325 exhibited its main peak at the earliest time, 17.398 ps, with an amplitude of 0.216 V. YZGLT-NM325 had the highest main peak, appearing at 17.544 ps with an amplitude of 0. For different sample types, the refractive indices also varied, resulting in different terahertz wave speeds as they penetrated the sample. This led to differences in the timing of the main peak's appearance. Fig. 2(b) illustrates the power variation trend with frequency for the three samples. Within the 0.5–2.75 THz frequency range, the energy consumed by terahertz pulses passing through the three kaolin samples decreases in the order: YZGLT-NM325 > SZGLT-GX325 > RZGLT-GX325. Analysis of Figs. 2(a) and 2(b) reveals significant differences in spectral information and delays between the three kaolin samples. Due to the varying energy consumption of the terahertz pulses in the three samples, the amplitude attenuation and waveform characteristics also differ. Therefore, terahertz time-domain spectroscopy can be used to identify and classify the three kaolin samples.

The refractive index and absorption coefficient of the kaolin sample can be determined by substituting the time-domain spectral data obtained according to the experimental principle into Eqs. (1) and (2). As illustrated in Fig. 3(c), the refractive spectrum of the sample reveals notable variations among the three kaolin samples within the 0.5–2.75 THz

frequency range. YZGLT-NM325 exhibits the highest refractive index, followed by RZGLT-GX325 and then SZGLT-GX325 in descending order. Based on the chemical composition percentages of the various kaolin samples in Table 1, SZGLT-GX325 is shown to have the highest SiO_2 content at 57.78%, followed by RZGLT-GX325 and YZGLT-NM325 with SiO_2 contents of 52.15% and 50.75%, respectively; YZGLT-NM325 has the lowest SiO_2 content. As illustrated in Fig. 3(c), a negative correlation exists between the SiO_2 content and the refractive indices of several kaolin samples. The sample with the highest SiO_2 content (SZGLT-GX325) exhibits the lowest refractive index, whereas the sample with the lowest SiO_2 content (YZGLT-NM325) displays the highest. Although the refractive index difference between SZGLT-GX325 and RZGLT-GX325 is considerable, the Fe_2O_3 content difference is minimal, at only 0.04%.

Furthermore, the Al_2O_3 and K_2O content differences between the two samples are significant. In conclusion, the refractive index of kaolin samples is primarily governed by their overall compositional makeup, with SiO_2 being a key determinant.

Figure 3(b) presents the absorption coefficient waveforms for the three samples. Employing the refractive index analysis method within the 0.5–2.75 THz frequency range reveals that the absorption coefficient increases with increasing frequency. SZGLT-GX325 demonstrates the highest absorption coefficient, followed by RZGLT-GX325; YZGLT-NM325 exhibits the lowest. Within the 0.5–2.25 THz frequency range, the three samples display similar waveforms with only minor variations. However, within the 2.25–2.75 THz frequency range, significant differences emerge in the absorption coefficient waveforms for the three samples, allowing for their identification based on the magnitude of these differences. By comparing the chemical composition and parameter content of the samples in Table 1 with the absorption coefficient waveforms of the three samples, we determined which chemical components in kaolin affect the absorption coefficient. Specifically, the order of Fe_2O_3 content in the samples is completely consistent with the order of absorption coefficients. Comparing these three kaolin samples, as the Fe_2O_3 content increases, the absorption coefficient also increases. This indicates that the terahertz absorption coefficient of kaolin is influenced by the Fe_2O_3 content in the sample.

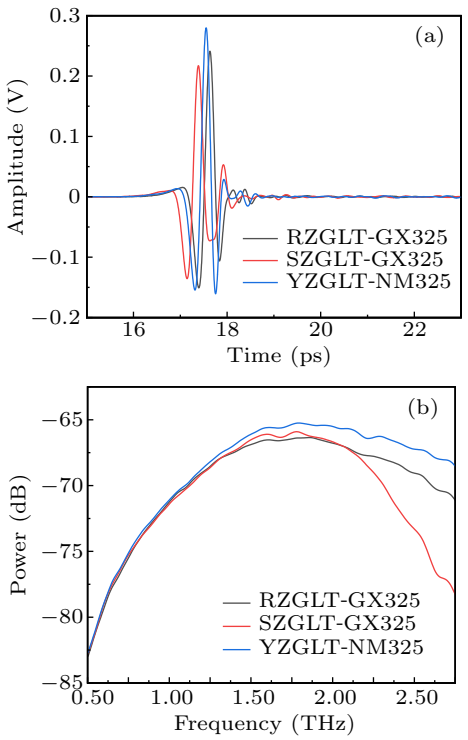


Fig. 2. (a) Time-domain spectra and (b) power spectra of the three samples.

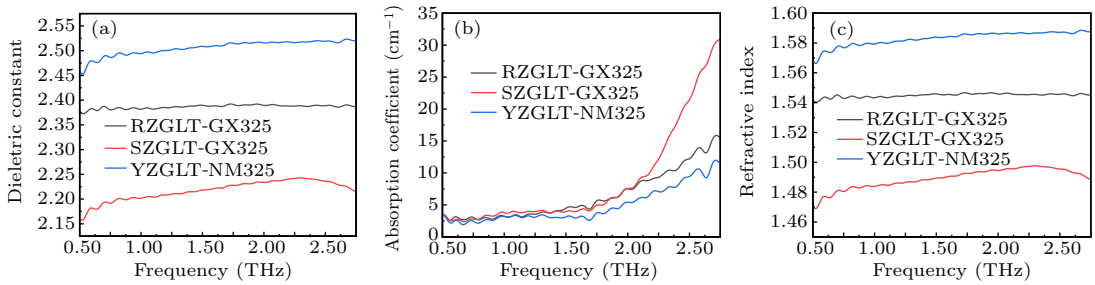


Fig. 3. (a) The dielectric constants, (b) absorption coefficients, (c) refractive indices of the three samples.

The dielectric constant, as a physical quantity describing the response characteristics of materials and samples in an electric field, is typically represented by a complex number closely related to frequency. Its real and imaginary parts respectively describe the material's ability to polarize in an electric field and the energy loss associated with it. The dielectric constant of a sample can be calculated using the sample's extinction coefficient and refractive index with the equation

$$\varepsilon = \varepsilon_1 + \mathrm{i}\varepsilon_2, \quad \varepsilon_1 = n^2 - k^2, \quad \varepsilon_2 = 2nk, \tag{4}$$

where k is the extinction coefficient, ε_1 is the real part of the dielectric constant, and ε_2 is the imaginary part of the dielectric constant.

The dielectric constant waveforms of the three kaolin samples obtained from Eq. (4) are shown in Fig. 3(a). As

is evident from Fig. 3(a), the dielectric constants and refractive indices of the three samples demonstrate a similar trend of change with frequency. The dielectric constant of YZGLT-NM325 is higher than that of the other two samples, thus ranking first. Conversely, the dielectric constant of SZGLT-GX325 is lower than that of RZGLT-GX325, thus placing it at the bottom. As the terahertz frequency band is approached, a close relationship becomes evident between the refractive indices and dielectric constants of the three samples. It is evident that alterations in the refractive index also have a significant impact on the dielectric constant. Consequently, factors influencing the magnitude of the refractive index also influence the magnitude of the dielectric constant, indicating that the SiO_2 content is also one of the factors affecting the magnitude of the dielectric constant.

3.2. Classification identification

In order to undertake a qualitative analysis of the similarities and differences among the nine sets of data for the three kaolin samples, two methods were used: PCA and CA. These methods were employed to analyze and classify the three kaolin samples separately. Clustering analysis, as an unsupervised learning method, aims to quantitatively uncover the inherent structure within data, enabling the automated grouping of physical entities and abstract objects. During the clustering process, precise classification criteria need not be predefined; the algorithm can directly begin with sample data. It calculates the Euclidean distance between samples within each category and designates the two groups with the closest Euclidean distances as a new category. The Euclidean distance between this new category and the remaining samples is then computed, and the two closest categories are merged into a new cluster. In conclusion, merging all samples from the measurement data resulted in the establishment of a single category. The CA method generated a cluster dendrogram, visually illustrating the similarities and differences among the nine data sets. The calculation of the Euclidean distance is given by

$$d^2 = (X_m - X_n)(X_m - X_n)', \tag{5}$$

where X_m and X_n represent the spectral data feature vectors of the m-th and n-th kaolin samples, respectively.

As demonstrated in Fig. 4(a), the application of CA to the refractive index data of samples within the 0.5–2.75 THz frequency range results in the generation of a cluster tree diagram based on the refractive index spectrum. As demonstrated in the accompanying Fig. 4(b), corresponding to the nine sample test data, cluster them into various sample types based on the dissimilarity of Euclidean distance. When the inter-class Euclidean distances range from 0.68 to 1.59, the kaolin samples are clustered into two categories. Within the initial category, soft kaolin (RZGLT) and hard kaolin (YZGLT) are grouped into a subcategory, while sandy kaolin (SZGLT) is grouped into a separate category. When the clustering distance falls between 0.39 and 0.68, it is possible to completely cluster kaolin samples into three categories. The three categories are: Category 1, which corresponds to hard kaolin; Category 2, which corresponds to soft kaolin; and Category 3, which corresponds to sandy kaolin. It is within this range that optimal classification results can be attained. The three clusters of samples correspond to the SiO_2 content of each sample type, thereby confirming that SiO_2 content is also one of the factors influencing the refractive index of the samples.

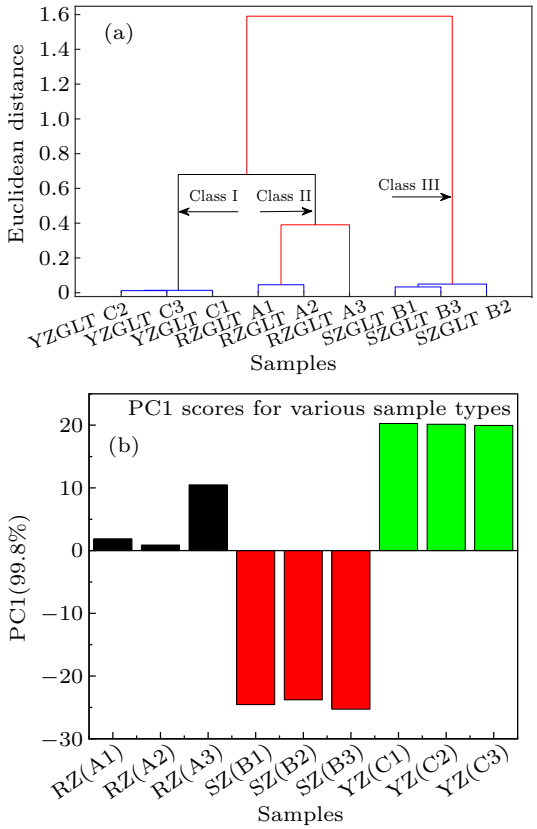


Fig. 4. (a) Dendrogram of refractive index values for different sample types. (b) Histogram of PC1 scores corresponding to different sample types.

The application of PCA to the analysis of refractive index data from samples within the 0.5–2.75 THz frequency range results in the identification of the eigenvalues of the sample data, in addition to the contribution rates and cumulative contribution rates of the principal components. The initial three principal component data features of the THz refractive index are preserved. As demonstrated in Table 2, the contribution rates of the first three principal components to the original data reach 99.99%, with the first principal component contributing approximately 99.83%. In PCA, the reflection of sample information is closely related to the contribution rate; the higher the contribution rate of a principal component, the more sample information it reflects. Consequently, the analysis results of the first three principal components can fundamentally reveal the similarities and differences between different types

of kaolin. The extent of similarity between samples is determined by the differential PC1 scores between said samples. The smaller the difference in PC1 scores between samples, the more similar they are; conversely, the larger the difference in PC1 scores, the greater the differences between samples. The PC1 score histograms for the refractive indices of various sample groups are presented in Fig. 4(b). The PC1 scores for the RZGLT-GX325 group range from −10 to 10, those for the SZGLT-GX325 group range from 10 to 20, while the PC1 scores of the YZGLT-NM325 group range from −10 to −20. A comparison of the dendrogram and PC1 histogram in Fig. 4 reveals that the score differences between the C1, C2, and C3 groups are minimal, resulting in the smallest Euclidean distances between these three groups in the dendrogram. Consequently, these groups are clustered into a single category. In a similar manner, the A1, A2, and A3 groups, as well as the B1, B2, and B3 groups, are also clustered into their respective categories based on the score differences between groups. This finding serves to substantiate the hypothesis that the PC1 scores are consistent with the sample similarity as indicated by the dendrogram. Therefore, it can be concluded that, in view of the findings relating to the Euclidean distances between different types of samples and the varying PC1 score differences between samples, the identification and classification of kaolin samples is indeed a possibility.

Table 2. Refractive index characteristic values for the three samples.

Factor	K	Contribution rate (%)	Cumulative contribution rate (%)
1	392.32271	99.82766	99.82766
2	0.63034	0.16039	99.98805
3	0.03665	0.00932	99.99738
4	0.00534	0.00136	99.99874
5	0.00228	0.00089	99.99963

In order to minimize experimental errors, we use the mean of the data from three independent samples of the corresponding material as the input set.

CA and PCA were applied separately to analyze the data, yielding a dendrogram of sample refractive indices as shown in Fig. 5(a) and a PC1 score plot as shown in Fig. 5(b). As demonstrated in Fig. 5(a), when the Euclidean distance ranges from 0.68 to 1.59, samples are categorized into two distinct groups: the first category comprises samples RZGLT-GX325 and YZGLT-NM325, while sample SZGLT-GX325 is distinctly separated and classified into the second category. When the Euclidean distance is between 0 and 0.68, the kaolin samples are also clustered into three categories, corresponding to soft kaolin, hard kaolin, and sandy kaolin. Within this distance range, kaolin samples can also be identified and classified. Comparison of the PC1 score histograms of the samples' refractive indices reveals that the PC1 score of RZGLT-GX325 ranges from 0 to 5, while that of SZGLT-GX325 ranges from −10 to −20. Furthermore, the PC1 score of YZGLT-NM325 is between 10 and 20. The disparity in PC1 scores between RZGLT-GX325 and YZGLT-NM325 is minimal, resulting in their Euclidean distance in the dendrogram being minimal. Consequently, they are grouped into a single category. This finding also corroborates the conclusion that the sample similarity exhibited in the dendrogram and the PC1 score chart is consistent.

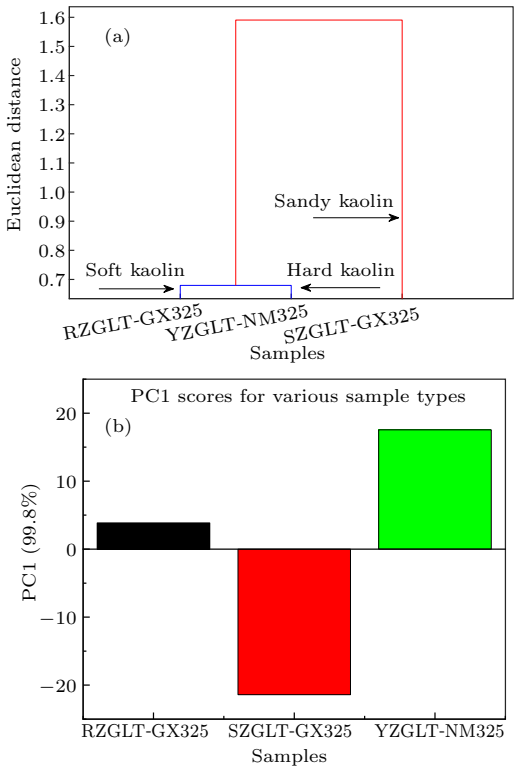


Fig. 5. (a) Dendrogram of refractive indices and (b) histogram of PC1 scores for three different kaolinite samples.

In accordance with the aforementioned analytical methodologies, the absorption data of the samples were subjected to analysis using both CA and PCA. The absorption data within the frequency range of 0.5–2.75 THz were utilized as input variables for CA, resulting in a cluster tree diagram of the absorption coefficient spectrum, as illustrated in Fig. 6(a). As demonstrated in the accompanying Fig. 6(a), the nine sets of measurement data are initially categorized into two primary classifications. The B1, B2, and B3 groups from SZGLT-GX325 constitute the primary category, with a Euclidean distance of 11.449. The remaining data samples comprise the secondary category, with a Euclidean distance of 52.782. Within the second category, the A1, A2, and A3 groups form the RZGLT-GX325 group with a Euclidean distance of 30.131. When compared to the C1, C2, and C3 groups forming the YZGLT-NM325 group with a Euclidean distance of 24.8494, the latter is slightly larger. It is evident that, in accordance with the disparities observed in the Euclidean distances, the nine groups of samples can be precisely categorized into the same category. This finding serves to illustrate the efficacy of CA in the context of sample absorption coefficients, which

has been demonstrated to facilitate the intuitive grouping of test data samples, thereby achieving the classification of the three kaolin types. Furthermore, a comparison of the Fe_2O_3 content among the kaolin samples indicates that the terahertz absorption coefficients of kaolin are influenced by the Fe_2O_3 content of the samples.

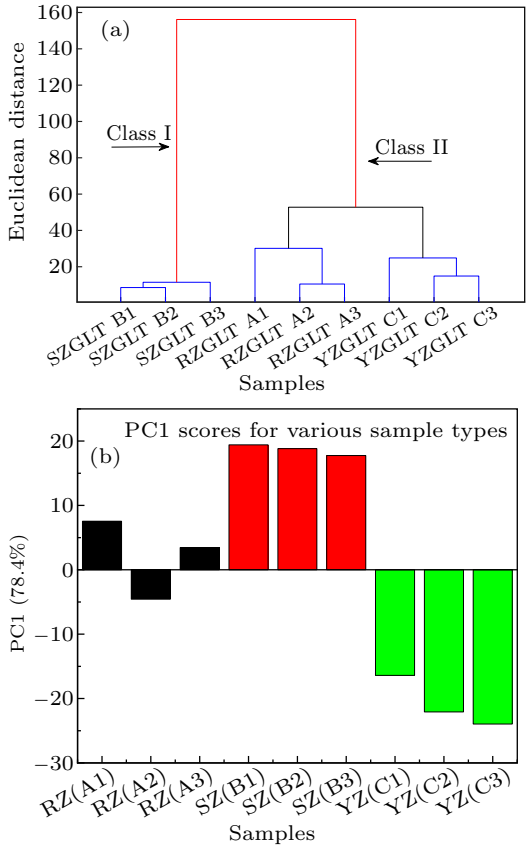


Fig. 6. (a) Dendrogram of absorption coefficient variations among different samples. (b) Histogram of PC1 scores among different samples.

Employing PCA to analyze the absorption coefficients of samples within the 0.5–2.75 THz frequency range allows for the determination of sample data eigenvalues, contribution rates, and cumulative contribution rates of the principal components, while preserving the data characteristics of the initial three principal components of the terahertz absorption coefficients. As demonstrated in Table 3, the first principal component's contribution rate is 78.48%, and the cumulative contribution rate of the first three principal components to the original data reaches 97.27%, indicating that the analysis results of these three principal components effectively represent the sample information. By observing the score differentials in the PC1 score histogram in Fig. 6(b), the measurements from the nine samples were initially categorized into two main groups.

Among the SZGLT-GX325 samples, the PC1 score differentials between the B1, B2, and B3 groups were the smallest, corresponding to the smallest Euclidean distances in the dendrogram. It can be observed that SZGLT-GX325 exhibits uniqueness in both PCA and CA; YZGLT-NM325 has the largest negative PC1 score, and the PC1 score differentials within the RZGLT-GX325 group are similar to those within the YZGLT-NM325 group, indicating a certain degree of similarity between the two. Combining CA and PCA, the test data from the nine groups of samples were intuitively clustered into three groups: SZGLT-GX325, RZGLT-GX325, and YZGLT-NM325, effectively reflecting the similarities and differences among the three kaolin types.

Table 3. Absorption coefficient eigenvalues of three samples.

Factor	*K*	Contribution rate (%)	Cumulative contribution rate (%)
1	308.44873	78.48568	78.48568
2	62.69165	15.95207	94.43776
3	11.11624	2.82856	97.26632
4	4.21679	1.07297	98.33929
5	3.39765	0.86454	99.20383

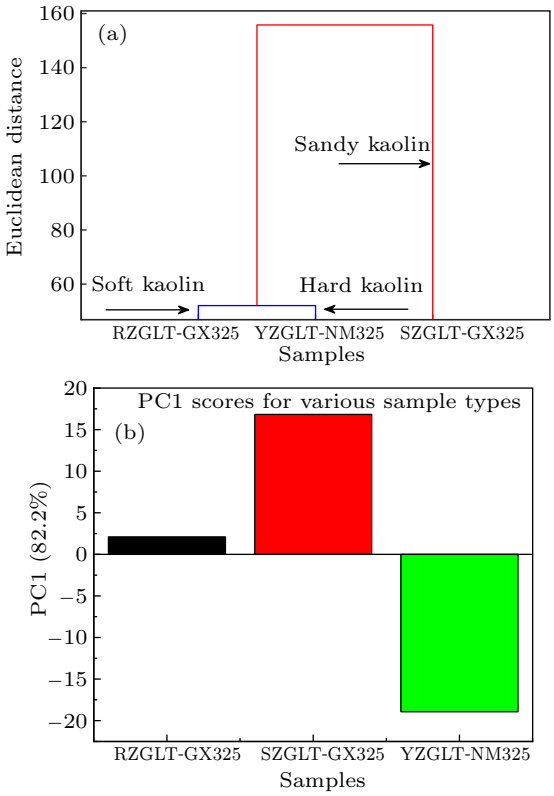


Fig. 7. (a) Dendrogram of absorption coefficients and (b) histogram of PC1 scores for three different kaolinite samples.

To mitigate experimental errors, similarly, the mean of absorption data from three independent samples of each material within the 0.5–2.75 THz frequency range is used as the input set. CA and PCA were subsequently employed to analyze the data. The resulting absorption coefficient dendrogram for the samples is presented in Fig. 7(a), and the PC1 score plot is shown in Fig. 7(b). The clustering dendrogram in Fig. 7(a) reveals that the kaolin samples are divided into two categories: SZGLT-GX325 is classified as the first category with a Euclidean distance of 155.8, while YZGLT-NM325 and RZGLT-GX325 form the second category with a Euclidean distance of 52.1. When the inter-category distance

ranges from 0 to 52.1, the tested kaolin samples can be clustered into three categories, corresponding to RZGLT-GX325, YZGLT-NM325, and SZGLT-GX325, respectively. As illustrated in the PC1 score histogram in Fig. 7(b), the PC1 score of RZGLT-GX325 ranges from 0 to 5, the PC1 score of SZGLT-GX325 ranges from 15 to 20, and the PC1 score of YZGLT-NM325 ranges from −15 to −20. The PC1 difference between RZGLT-GX325 and SZGLT-GX325 is minimal, resulting in their grouping into one category in the dendrogram. The range of Euclidean distances between the analyzed samples effectively reflects the differences and similarities among the three kaolin samples, classifying them into three categories and thus achieving the classification of kaolin.

Qualitative analysis of three independent samples (nine samples in total) was conducted using terahertz technology combined with CA and PCA. Because different types of kaolin exhibit distinct responses in the terahertz frequency band, the Euclidean distances and PC1 scores corresponding to each group also vary. Furthermore, the refractive indices and absorption coefficient spectra of the kaolin samples demonstrate significant differences. The results were visually presented through dendrograms and PC1 score histograms, illustrating the differences and similarities between categories. To minimize experimental errors, the mean values of the characteristic data for each sample category were used as the input set. The results indicate that within the 0.5–2.75 THz frequency band, the Euclidean distances and PC1 score histograms of kaolin samples also exhibit significant differences. By using CA to identify differences hidden beneath commonalities, the extracted data underwent objective multivariate analysis. Through the combination of CA-PCA models, the relationship between PC1 score differences and Euclidean distances is visually demonstrated. By combining these two models, the study achieved effective classification and identification of different types of kaolin, significantly improving the accuracy of classification. In summary, based on the differences in response of kaolin samples from different regions under terahertz time-domain spectroscopy technology, the CA-PCA model was applied for CA. The qualitative analysis results are clearly shown in the dendrogram and PC1 score plot, and a corresponding kaolin dataset was established, providing a comprehensive reference basis for the identification of kaolin from other regions.

4. Conclusion

Utilizing THz-TDS technology and machine learning methodologies, a qualitative analysis was performed on three kaolin varieties. Refractive indices and absorption coefficients of the kaolin types demonstrated notable variations in their spectral responses within the 0.5–2.75 THz frequency range. Furthermore, whether the input set comprises the mean values derived from measurements of nine samples or the mean values derived from measurements of three independent samples, CA and PCA were implemented to generate corresponding dendrograms and PC1 score plots. These analyses effectively elucidated the distinctions and similarities among soft, hard, and sandy kaolin, offering a robust method for differentiating between the kaolin types. The PCA results distinctly demonstrate the disparities among the three kaolin samples, whereas CA offers a more precise classification of the samples based on their Euclidean distance variations. By integrating CA and PCA for an objective analysis of the kaolin samples, the distinctions among the three specific categories — soft kaolin, hard kaolin, and sandy kaolin — can be comprehensively represented, thereby furnishing valuable reference for the enhancement of classification standards for the three kaolin types. Furthermore, based on the SiO_2 and Fe_2O_3 content in the samples, the corresponding trends in the refractive index and absorption coefficient of the kaolin samples were observed, thus confirming that the SiO_2 content in the samples is a factor influencing the refractive index of kaolin samples, and the Fe_2O_3 content in the samples is a factor influencing the absorption coefficient of kaolin samples. This study indicates that terahertz time-domain spectroscopy can be employed for qualitative and quantitative research on kaolin types. This research presents an effective method for the rapid detection and identification of kaolin using THz-TDS. Moreover, it provides crucial reference criteria for establishing a database in the field of kaolin identification.

Acknowledgements

This work was supported by the Key Natural Science Research Project for Colleges and Universities of Anhui Province (Grant No. 2023AH050343), the Pollutant Sensitivity Monitoring and Application Innovation Team of Anhui Province (Grant No. 2023AH010043), Anhui Province Graduate Education Quality Project (Grant No. 2024jyjxggyjY204), Huaibei Normal University Quality Project (Grant Nos. 2024jgxm003 and 2025xjxqhzjd001), and Open Research Fund of Anhui Province Key Laboratory of Intelligent Computing and Applications (Grant No. AFZNJS2025KF08).

References

[1] Jiang Y, Li G, Lv M, *et al*. 2020 *Chin. Phys. B* **29** 098705
[2] Peng S, Peng Y, Yang J, Xiong X, Wan L, Liu W, *et al*. 2025 *J. Nucl. Sci. Technol.* **1** 000001
[3] Xu J, Ouyang S, Duan S, *et al*. 2024 *Chin. Phys. B* **33** 038701
[4] Hou Z, Yan B, Zhao Y, *et al*. 2025 *Chemometr Intell. Lab* **262** 105408
[5] Kojima S and Mori T 2016 *Ferroelectrics* **500** 183202
[6] Wu D M, Guo J, Sun M M, *et al*. 2023 *J. Food Process. Preserv.* **2023** 3290917
[7] Xie W, Tang Q, Xie J, *et al*. 2024 *Nat. Commun.* **15** 38
[8] Henao R and Pencina M 2025 *Eur. J. Prev. Cardiol.* zwaf164
[9] Zhan H, Wu S, Bao R, *et al*. 2015 *Fuel* **143** 189193
[10] Li H, Qin X L and Li Y 2021 *Optik* **242** 166822
[11] Zhang M, Xie X, Zhang D, *et al*. 2023 *Microw. Opt. Technol. Lett.* **65** 111721

[12] Cao B H, Fan M B and Jing S Y 2010 *Spectrosc. Spect. Anal.* **30** 1748
[13] Bardon T, May R K, Taday P F, *et al*. 2013 *Analyst* **138** 485969
[14] Cheng L, Wang L, Mei H, *et al*. 2016 *IEEE Trans. Dielectr. Electr. Insul.* **23** 2126
[15] Huang S, Deng H, Wei X, *et al.* 2023 *Front. Bioeng. Biotechnol.* **11** 1219042
[16] Bao R, Luan X, Wu S, *et al.* 2018 *ACS Omega* **3** 7798
[17] Li W, Hu B and Wang M W 2014 *Spectrosc. Spect. Anal.* **34** 3235
[18] Qi N, Zhang Z, Wang Y, *et al.* 2015 *Anal. Methods* **7** 233338
[19] Zhang H, Sfarra S, Osman A, *et al*. 2018 *IEEE Trans. Ind. Inform.* **14** 562936
[20] Zhang M, Byers J, Carnio B N, *et al*. 2024 *IEEE Trans. Terahertz Sci. Technol.* **14** 301310
[21] Merghem K, Busch S F, Lelarge F, *et al*. 2017 *J. Infrared Millim. Terahertz Waves* **38** 958962
[22] Yu J, Pu H and Sun D W 2025 *Chem. Eng. J.* **504** 158177
[23] Zou M, Su M and Yu H 2020 *Opt. Mater.* **107** 110062
[24] Zhang M, Jiang H, Qiu Y, *et al*. 2025 *Appl. Opt.* **64** 284756
[25] Li S, Shen Z, Yin W, *et al*. 2022 *J. Phys. D: Appl. Phys.* **55** 325003
[26] Murphy K N, Naftaly M, Nordon A, *et al*. 2022 *Appl. Sci.* **12** 3475
[27] Hong T, Choi K, Ha T, *et al*. 2014 *J. Korean Phys. Soc.* **64** 727
[28] Li R, Wang J, Xu Q, *et al.* 2026 *Opt. Express* **34** 17255
[29] Shi C, Ma Y, Zhang J, *et al*. 2018 *Biomed. Opt. Express* **9** 135059
[30] Yamamoto K, Kabir M H, Hayashi M, *et al*. 2005 *Phys. Chem. Chem. Phys.* **7** 194552
[31] Ma S G, Sun D, Liu Z Y, *et al*. 2022 *Spectrosc. Spect. Anal.* **42** 175560 (in Chinese)
[32] Carnio B N, Attiaoui A, Assali S, *et al*. 2022 *IEEE Trans. Terahertz Sci. Technol.* **12** 385391
[33] Song Z, Zhang S, Tu S, *et al*. 2025 *Spectrochim. Acta Part A: Mol. Biomol. Spectrosc.* **325** 125104
[34] Liu H, Zhang Z, Yang Y, *et al*. 2018 *Optik* **172** 668673
[35] Cruz J O and Carranza G E T 2026 *Maderas: Cienc. Tecnol.* **28** 3
[36] Qu F, Lin L, He Y, *et al*. 2018 *J. Infrared Millim. Terahertz Waves* **39** 114861
[37] Xu J, Wang H, Duan Y, *et al*. 2020 *Infrared Phys. Technol.* **109** 103413
[38] Zhao Y, Li Z, Liu J, *et al*. 2018 *Spectrochim. Acta Part A: Mol. Biomol. Spectrosc.* **192** 336342
[39] Ding J, Meng Q H, Shen Y, *et al*. 2023 *Chin. Phys. B* **32** 048702