Invited Paper

Terahertz spectroscopy combined with machine-learning models for crude oil classification

Shanzhe Zhang [1, 2*], Dongyu Zheng [1, 2], Xiaorong Sun [1, 2*], Cuiling Liu [1, 2], Jingzhu Wu [1, 2], and Sining Yan [1, 2]

[1] School of Artificial Intelligence, Beijing Technology and Business University, Beijing 100048, China

[2] Beijing Key Laboratory of Big Data Technology for Food Safety, Beijing Technology and Business University, Beijing 100048, China

[*2] Email: zhangsz@btbu.edu.cn; sxrchy@sohu.com

(Received March 2023)

Abstract: The classification of crude oils plays an important role in the petroleum transportation and production. In this paper, terahertz time-domain spectroscopy (THz-TDS) is used to assess seven various crude oils combined with machine-learning algorithms. From THz-TDS, frequency, refractive index and absorption coefficient are used to set models, which are based on Extreme Gradient Boosting (XGBoost), Random Forest (RF) and K-Nearest Neighbors (KNN), respectively. In order to evaluate the accuracy of each model, the confusion matrix and the Area under the curve (AUC) are introduced to access the classification ability, and 5-fold cross-validation are used to compare the generalization ability and robustness. Compared to other models, the classification accuracy of XGBoost reaches the maximum 0.9622. Meanwhile, the test 5-fold cross-validation F1-score and the AUC of XGBoost model are higher than other models, which indicates the high consistency and robustness. Experimental results suggests that terahertz time-domain spectroscopy may be a powerful tool for the identification of various crude oils.

Keywords: Terahertz spectroscopy, Crude oil classification

Doi:

1. Introduction

As the blood of the industry and a kind of non-renewable resource, petroleum is becoming more valuable and scarcer than before. In the modern oil industry, classification for crude oil plays an important role in the process of petroleum transportation and transaction. Before being processed and refined, crude oil is transported to refineries through long pipelines. Due to the wide range of oil source, many kinds of crude oils are transported alternately in the pipeline. In addition, transmission system is very complicated because of numerous trunk lines. Therefore, it is of great practical significance to identify various crude oils. Moreover, oil spills have grown up to be a global problem, especially in industrialized countries.[1] Oil tracing plays a key role in the investigation and accountability of oil spills.

Generally, the crude oil consists of paraffin, aromatics, naughtiness and asphaltene.[2] The composition of crude oils is different, which is closely related to the geographical location. Most crude oil analysis methods are built on standard methods, which are related to the American Society for Testing and Materials International and Institute of Petroleum. Though these detection methods have the advantages of high precision and high sensitivity,[3] recently, with the development of technology, optical methods provides credible tools for oil analysis. Applied gas chromatography (GC) and mass spectrometry (MS) in crude oil analysis can improve sensitivity, selectivity, resolution.[4-9] Inductively coupled plasma (ICP) and nuclear magnetic resonance (NMR) have also shown a great potential in oil characterization.[10-14, 15-17] In order to ensure the accuracy of the experiment, crude oil must be detected many times. However, most of standard test methods are rather time consuming, environmental non-friendly, elaborate and expensive.[18]

In recent years, terahertz (THz) spectroscopy has been used in the analysis of crude oil. The effect of asphaltenes on different crude oils had been researched based on terahertz time-domain spectroscopy (THz-TDS). Results suggest that the refractive index spectra of the asphaltene show variation in the low THz frequencies and comparable spectra in the higher frequencies.[19] Furthermore, the morphology and structure of wax crystals in crude oils is characterized by THz-TDS. Dynamic processes of the clusters in the model oils are analyzed and identified based on the measured absorption and extinction coefficients in the THz region.[20] However, the size and complexity of the problem increase with the number of classes. It becomes important to use highly effective approach for the crude oil classification. As the performance of machine learning has improved, it has been widely used in the THz field.[21] Combined with deep learning method, THz has been used to analyze polluted sand particles and monitor crude oil spills in real time.[22]

In this study, seven crude oils are measured by THz-TDS which is used to realize the crude oil classification combined with machine learning methods. Frequency, refractive index and absorption coefficient are selected as the inputs of various models. Extreme Gradient Boosting (XGBoost), Random Forest (RF) and K-Nearest Neighbors (KNN) are used to set different models.[23-26] Moreover, the classification accuracy of different models is compared and both the generalization ability and robustness are considered. Results suggest that THz-TDS can be used to characterize crude oils from different oil fields.

2. Materials and methods

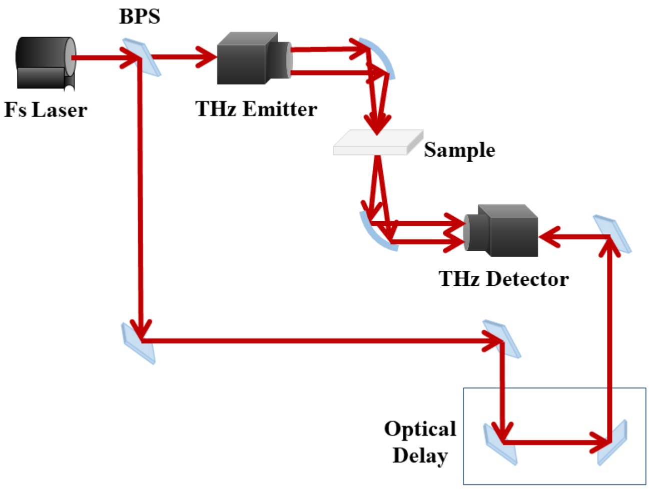

Fig. 1 Schematic diagram of the THz-TDS system.

Crude oil samples are obtained from Wenmi oilfield, Russia, Changqing oilfield, Hongtai oilfield, Jiudong oilfield, Santanghu oilfield and Daqing oilfield, which named Oil A, Oil B, Oil C, Oil D, Oil E, Oil F and Oil G, respectively. Moreover, crude oils are dehydrated to avoid effect of water molecules. The quartz cuvettes are used to accommodate crude oils with the geometry of 20 *mm* × 45 *mm* and a thickness of 10 *mm*. The experimental equipment is TeraPulse 4000 which is produced in UK. A THz spectrometer in the transmission mode is used to acquire the optical information of crude oils. As shown in Figure 1, laser pulses are split into two beams: one beam is employed to irradiate the oil sample, and the other is applied for THz detection. The air environment is measured in the first as a reference signal. For accuracy, TeraPulse4000 are performed 3 times for a single sample. And then, the average value is used to calculate the optical parameters. Furthermore, all of the measurements are performed at room temperature.

The confusion matrix is usually used to evaluate the performance of the models. Furthermore, confusion matrix can also compare the performance of different models according to the evaluation indicators. The matrix consists of two parts, which is corresponded to the results of model classification results and training data. True Positive (TP)，True Negative (TN)，Fales Positive (FP) and Fales Negative (FN) are the four important indicators of the confusion matrix. Accuracy, recall, precision and F1 score are calculated by equation (1) to equation (4). True Positive Rate (TPR) represents the percentage of correctly classified data in the total number of predictions.

F1 score is expressed by the harmonic mean of accuracy and TPR.

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \tag{1}$$

$$Precision = \frac{TP}{TP + FP} \tag{2}$$

$$TPR = \frac{TP}{TP + FN} \tag{3}$$

$$F1 = \frac{2 \times Precision \times Recall}{Precision + Recall} \tag{4}$$

Receiver operating characteristic (ROC) curve is used to select the best model and set the best threshold. Moreover, Area under the curve (AUC) is an evaluation indicator for classifying models, which is calculated by integration. A large value of the AUC represents a high accuracy of the model classification. In this study, ROC method is used to solve multi-classification problems by conducting a binary classification analysis. The macro-average of each binary analysis is chosen as the final evaluation basis.

3. Results and discussion

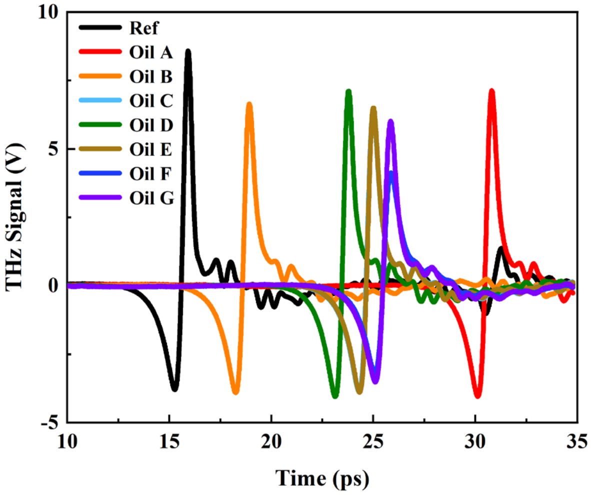

Fig. 2 Terahertz time-domain spectrum of seven crude oil samples.

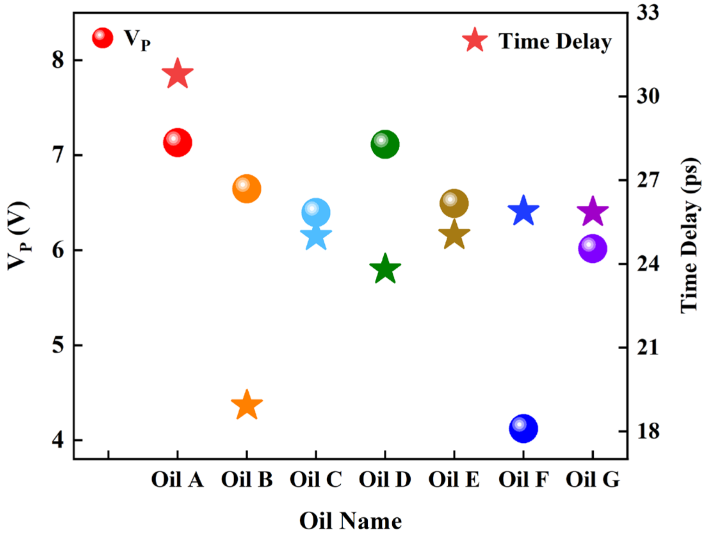

Fig. 3 V_P and time delay of crude oil samples.

The THz pulse waveforms are displayed in **Figure 2**, which are transmitted through the empty quartz cuvette and different crude oils. The THz response of quartz cuvette are selected as the reference pulse. A phase shift relative to the reference pulse occurrs for THz pulses transmitted through the crude oil samples, and a significant decrease in the amplitude is also observed. The time delay and THz signal peaks (V_P) are extracted from THz pulse waveforms, which are shown in **Figure 3**. The V_P of Oil F achieves maximum value at 6.49 *V*, while the V_P of Oil G achieves minimum value at V. The time delay values of Oil A, Oil B, Oil C, Oil D Oil E and Oil G are different, which are 30.79 *ps*, 18.92 *ps*, 24.99 *ps*, 23.80 *ps*, 25.02 *ps*, 25.88 ps, and 25.85 *ps*, respectively. The V_P and time delay are different from each other, which indicates that various crude oils can be identified using THz spectra.

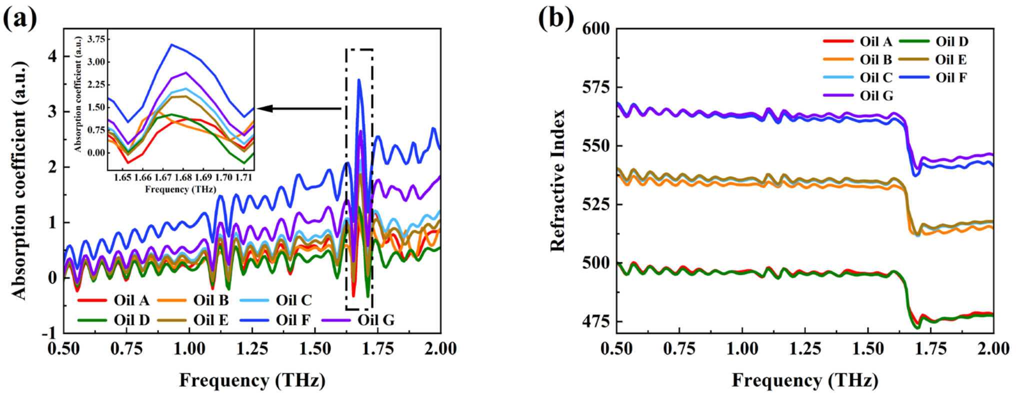

Fig. 4 Frequency dependence of (a) absorption coefficient and (b) refractive index spectra.

Following the fast Fourier transform (FFT) of the reference and sample spectra, the THz-TDS are obtained in **Figure 4(a)**. It is noticed that the there is a peak at ~1.68 *THz* for different crude oils. The composition of crude oil is complex. It is difficult to verify which vibration of the chemical bonds result in the peak at ~1.68 *THz* for crude oils. According to the THz-TDS, the frequency dependence of refractive index could be obtained and are plotted in **Figure 4(b)**, when the frequency ranges from 0.5 to 2.0 *THz*. Comparing the THz-TDS of the crude oils in Figure 2, the time delays are different from each other, indicating variant refractive index among them. There is a precipitous decay at 1.625 *THz* in **Figure 4(b)**, which can be owing to the frequency dependence of dispersion. According to the refractive index, the seven crude oils are divided into three amplitude ranges. When the frequency ranges from 0.5 to 1.6 *THz*, the refractive index of Oil A and Oil D is about 498. However, the refractive index of Oil B, Oil C and Oil E is about 538 and the refractive index of Oil E and Oil F and Oil G is about 563 at the same frequency range.

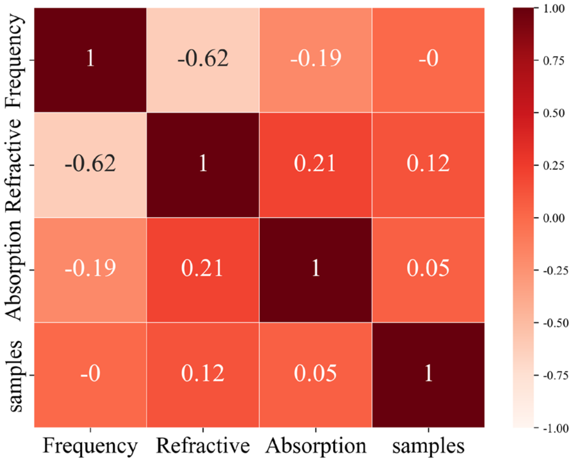

Fig. 5 Kendall correlation coefficient heat maps of features.

The spectra information dataset consists of 10472 records, which are obtained from TeraPulse4000. In order to ensure better predictability of the model, Kendall correlation analysis is used to evaluate the correlation among features. The results of Kendall correlation analysis are displayed in **Figure 5**. It is clear that correlation coefficients of features are less than 0.8, which has proved that there is no significant correlation between features. Hence, refractive index, absorption coefficient and the corresponding frequency are the independent features of oil samples, which could be used as inputs for machine-learning models.

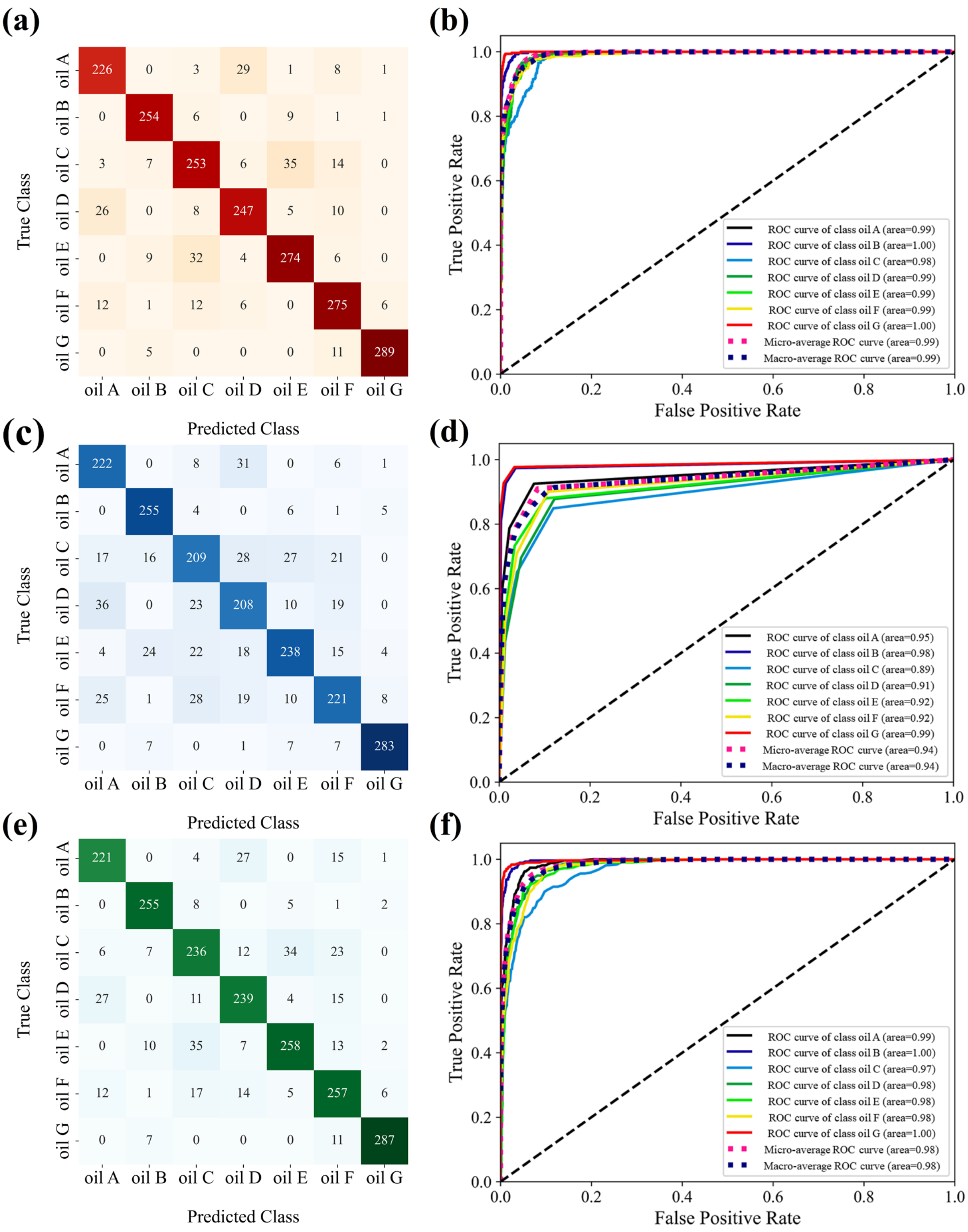

Fig. 6 The confusion matrixes of the classification results of (a) XGBoost, (c) KNN, and (e) Random Forest. The ROC curve of (b) XGBoost, (d) KNN, and (f) Random Forest.

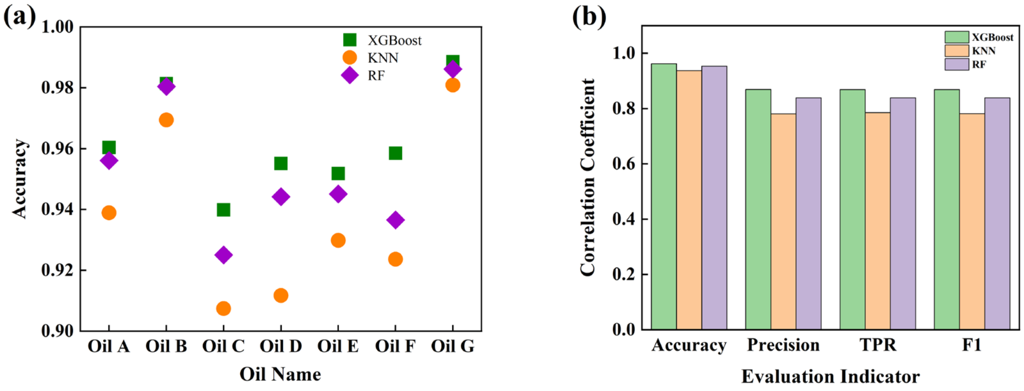

Fig. 7 (a) Classification accuracy of crude oils by various models. (b) Evaluation indicators for the classification results.

To build a more precise model, XGBoost, RF and KNN algorithms are used to classify different crude oils. The dataset of crude oils is split into two groups. In the experiment, 80% of the data are used for training, and 20% of the data are used for testing. In **Figure 6**, the diagonal of the confusion matrix suggests the quantity of correct classification of seven oils. From **Figure 6 (a), (c)** and **(e)**, it is clearly that more features of Oil G are correctly identified followed by Oil B. As shown in **Figure 6(b), (d)** and **(f)**, the macro-average AUC of all classification algorithms exceeds 0.9, which suggests that XGBoost, RF and KNN models are suitable for crude oil classification based on THz-TDS data. The maximum macro-average AUC calculated by XGBoost is up to 0.99, which has verified the conclusion of confusion matrix.

Tab. 1 Comprehensive performance of various models.

Model	Accuracy	Precision	Recall	F1-score	macro-average AUC	5-fold cross-validation F1-score
XGBoost	0.9622	0.8690	0.8689	0.8689	0.99	0.8643
KNN	0.9374	0.7808	0.7854	0.7817	0.94	0.7634
RF	0.9534	0.8390	0.8391	0.8387	0.98	0.8334

The classification accuracy of the three machine learning models is displayed in **Figure 7 (a)**. It can be seen that the classification performances of RF and XGBoost are relatively accurate, and the overall classification effect is better than KNN. In order to verify the accuracy and stability of the model, the accuracy, recall rate, accuracy and F1 score values are calculated, respectively. From the **Figure 7 (b)**, it can be seen that four indicators of the XGBoost model are higher than those of the others, which referr to Accuracy, Precision, TPR and F1-score. The specific values of them are shown in **Table 1**. Experimental results suggest that the classification effect of XGBoost model is better than that of other models.

The performance of classification models is generally evaluated by k-fold cross validation. [27, 28] In this study, crude oils data set is first randomly divided into 5 folds, which has approximately the same number of subsets. Moreover, every fold in turn plays the role for testing the model induced from the other 4 folds. In the end, the average F1-score of the five cross-validation models is adopted as an indicator to evaluate the generalization capability of various classification models. From **Table 1**, the average F1-score of XGBoost model is higher than others, which has proved again its superiority in crude oil classification. Moreover, six indicators in **Table 1** has proved that XGBoost achieves best classification performance and generalization ability.

The optical methods have played a crucial role in the past few years. Laser technology has also been used for the monitoring of laser drilling, identification of crude oils, characterization of water content in crude oil emulsion, measurement of wax appearance temperature, characterization of viscosity change for crude oils.[29-33] Moreover, THz method has been widely used in the oil field. THz-TDS has been used to characterize oil disaggregation under magnetic field and analyze the state of oil-water two-phase flow.[20, 34, 35] In this work, THz-TDS method is used for identification of various crude oils. The V_P, time delay, refractive index and absorption coefficient are analyzed. Moreover, XGBoost, Random Forest and KNN algorithms are used for the crude oil classification. The inputs of different models are related to frequency, refractive index and absorption coefficient. Results suggest that the XGBoost model have the best classification effect among three models.

4. Conclusion

Finding the true source is a crucial step for the leakage in the oil transportation. In this work, the classification of various crude oils is realized by THz-TDS combined with machine-learning algorithms. The frequency dependence of the refractive index and absorption coefficient are analyzed based on the THz-TDS of crude oils. XGBoost, Random Forest and KNN are chosen as machine-learning models to classify various crude oils. Frequency, refractive index and absorption coefficient are selected as the inputs of different models. Results have demonstrated that the classification effect of XGBoost is the best while the KNN model is the worst. The six evaluation indicators of XGBoost model are better than those of the others. Moreover, XGBoost has the strongest consistency and robustness based on the highest test 5-fold cross-validation F1-score and the AUC value. In conclusion, combined with machine-learning algorithms, THz-TDS is a fast and simple technique for the characterization of crude oil.

Notes

The authors declare no competing financial interest.

Acknowledgement

This work was supported by the Beijing Natural Science Foundation (No. No. 4222043) and the Research Foundation for Youth Scholars of Beijing Technology and Business University (No. QNJJ2022-41).

Reference

[1] Hashemi-Nasab FS, and Parastar H. “Pattern recognition analysis of gas chromatographic and infrared spectroscopic fingerprints of crude oil for source identification”. *Microchemical Journal*. 153: 104326 (2020).

[2] Marshall AG, and Rodgers RP. “Petroleomics: the next grand challenge for chemical analysis”. *Acc Chem Res*. 37:53–9 (2004).

[3] Zhan HL, Yang YQ, Zhang Y, et al. “Terahertz for the detection of the oil bearing characteristics of shale”. *Energy Reports*. 7: 5162–5167 (2021).

[4] Prasantongkolmol T, Thongkorn H, Sunipasa A, et al. “Analysis of sulfur compounds for crude oil fingerprinting using gas chromatography with sulfur chemiluminescence detector”. *Marine Pollution Bulletin*. 186: 114344 (2023).

[5] Borisov RS, Kulikova LN, and Zaikin VG. “Mass Spectrometry in Petroleum Chemistry (Petroleomics) (Review)”. *Petroleum Chemistry*. 59: 1055-1076 (2019).

[6] Liu Y, He J, Song CW, et al. “Oil Fingerprinting by Three-Dimensional (3D) Fluorescence Spectroscopy and Gas Chromatography-Mass Spectrometry (GC-MS)”. *Environmental Forensics*. 10: 324-330 (2009).

[7] Yang BJ, Yu ZG, Yao P, et al. “Characterization of Oil by Micro-Solid-Phase Extraction and Gas Chromatography–Mass Spectrometry”. *Analytical Letters*. 48: 2493-2506 (2015).

[8] Bayona JM, Dominguez C, and Albaiges J. “Analytical developments for oil spill fingerprinting”. *Trends in Environmental Analytical Chemistry*. 5: 26-34 (2015).

[9] Sun PY, Bao MT, Li GM, et al. “Fingerprinting and source identification of an oil spill in China Bohai Sea by gas chromatography-flame ionization detection and gas chromatography-mass spectrometry coupled with multi-statistical analyses”. *Journal of Chromatography A*. 1216: 830-836 (2009).

[10] Rios SM, Barquin M, and Nudelman NS. “Characterization of oil complex hydrocarbon mixtures by HSQC-NMR spectroscopy and PCA”. 27: 352-357 (2014).

[11] Rakhmatullin IZ, Efimov SV, Tyurin VA, et al. “Application of high resolution NMR (H-1 and C-13) and FTIR spectroscopy for characterization of light and heavy crude oils”. *Journal of Petroleum science and engineering*. 168: 256-262 (2018).

[12] Silva SL, Silva AMS, Ribeiro JC, et al. “Chromatographic and spectroscopic analysis of heavy crude oil mixtures with emphasis in nuclear magnetic resonance spectroscopy: A review”. 707: 18-37 (2011).

[13] Gao GH, Cao J, Xu TW, et al. “Nuclear magnetic resonance spectroscopy of crude oil as proxies for oil source and thermal maturity based on H-1 and C-13 spectra. Fuel”. 271: 117622 (2022).

[14] Masili A, Puligheddu S, Sassu L, et al. “Prediction of physical-chemical properties of crude oils by 1H NMR analysis of neat samples and chemometrics”. *Magnetic Resonance in Chemistry*. 50: 729-738 (2012).

[15] Ogunlaja A, Ogunlaja OO, Okewole DM, et al. “Risk assessment and source identification of heavy metal contamination by multivariate and hazard index analyses of a pipeline vandalised area in Lagos State, Nigeria”. *Science of the total environment*. 651: 2943-2952 (2018).

[16] Alharbi T, and El-Sorogy A. “Assessment of metal contamination in coastal sediments of Al-Khobar area, Arabian Gulf, Saudi Arabia”. *Journal of African earth sciences*. 129: 458-468 (2017).

[17] Ellis J, Rechsteiner C, Moir M, et al. “Determination of volatile nickel and vanadinum species in crude oil and crude oil fractions by gas chromatography coupled to inductively coupled plasma mass spectrometry”. *Journal*

of analytical atomic spectrometry. 26: 1674-1678 (2011).

[18] Garmarudi AB, Khanmohammadi M, Fard HG, et al. “Origin based classification of crude oils by infrared spectrometry and chemometrics”. *Fuel*. 236: 1093–1099 (2019).

[19] Matoug MM, and Gordon R. “Crude Oil Asphaltenes Studied by Terahertz Spectroscopy”. *ACS Omega*. 3: 3406-3412 (2018).

[20] Jiang C, Zhao K, Fu C, et al. “Characterization of morphology and structure of wax crystals in waxy crude oils by terahertz time-domain spectroscopy”. *Energy & Fuels*. 31: 1416-1421 (2017).

[21] Park H and Son JH. “Machine Learning Techniques for THz Imaging and Time-Domain Spectroscopy”. *Sensors*. 21: 1186 (2021).

[22] Zhan HL, Meng ZH, Ren ZW, et al. “Terahertz Spectroscopy Combined with Deep Learning for Predicting the Depth and Duration of Underground Sand Pollution by Crude Oil”. *IEEE Transactions on instrumentation and measurement*. 17: 2500108 (2022).

[23] Friedman JH. “Greedy function approximation: a gradient boosting machine”. *Annals of Statistics*. 29: 1189–1232 (2001).

[24] Yan ZZ, Chen H, Dong XH, et al. “Research on prediction of multi-class theft crimes by an optimized decomposition and fusion method based on XGBoost”. *Expert Systems with Applications*. 207: 117943 (2022).

[25] Breiman L. “Random forests”. *Machine Learning*. 45: 5–32 (2001).

[26] Cover T, and Hart P. “Nearest neighbor pattern classification”. *IEEE Transaction on Information Theory*. 1967, 13: 21–27 (1967).

[27] Wong TT, and Yeh PY. “Reliable Accuracy Estimates from k-Fold Cross Validation”. *IEEE Transactions on Knowledge and Data Engineering*. 32: 1586-1594 (2020).

[28] Fushiki T. “Estimation of prediction error by using K-fold cross-validation". *Statistics and Computing*. 21: 137-146 (2011).

[29] Zhang S Z, Huang J F, Miao X Y, et al. "Characterizing the laser drilling process of oil shale using laser-induced voltage”. *Optics and Laser Technology*. 131: 106478 (2020)

[30] Zhang S Z, Miao X Y, Zhan H L, et al. “Characterization of low- water-content crude oils under laser irradiation”.

Laser Physics Letters. 17: 106002 (2020)

[31] Zhang S Z, Sun X R, Zheng D Y, et al. “Characterization of crude oil viscosity change under laser irradiation”. *Laser Physics Letters*. 19(12): 12600 (2022).

[32] Zhang S Z, Sun X R, Yan S N, et al. “Effect of laser irradiation on a heavy crude oil sample: Changes in viscosity and implications for oil recovery and transport”. *Physics of Fluids*. 34(12): 127122 (2023).

[33] Zhang S Z, Sun X R, Liu C L, et al. “Characterization of wax appearance temperature of model oils using laser-induced voltage”. *Physics of Fluids*. 34(6): 067123 (2022).

[34] Song Y, Zhao K, Zhu J, et al. “The detection of water flow in rectangular microchannels by terahertz time domain spectroscopy”. *Sensors*, 17 2330.

[35] Song Y, Zhan H, Zhao K, et al. “Simultaneous characterization of water content and distribution in high-water-cut crude oil”. *Energy & Fuels*. 30: 3929-3933 (2016)