Analytical Methods

ROYAL SOCIETY OF CHEMISTRY

PAPER

View Article Online
View Journal | View Issue

Cite this: *Anal. Methods*, 2023, **15**, 1649

Ultraviolet-induced fluorescence of oil spill recognition using a semi-supervised algorithm based on thickness and mixing proportion–emission matrices

Bowen Gong,[ab] Hongji Zhang,[a] Xiaodong Wang,[a] Ke Lian,[c] Xinkai Li,[a] Bo Chen,*[a] Hanlin Wang[ab] and Xiaoqian Niu[ab]

In recent years, marine oil spill accidents have been occurring frequently during extraction and transportation, and seriously damage the ecological balance. Accurate monitoring of oil spills plays a vital role in estimating oil spill volume, determination of liability, and clean-up. The oil that leaks into natural environments is not a single type of oil, but a mixture of various oil products, and the oil film thickness on the sea surface is uneven under the influence of wind and waves. Increasing the mixed oil film thickness dimension and the mix proportion dimension has been proposed to weaken the effect of the detection environment on the fluorescence measurement results. To preserve the relationships between the data of oil films with different thicknesses and the relationships between the data of oil films with different mixing proportions, the three-dimensional fluorescence spectral data of mixed oil films on a seawater surface were measured in the laboratory, producing a thickness–fluorescence matrix and a proportion–fluorescence matrix. The nonlinear variation of the fluorescence spectra was investigated according to the fluorescence lidar equation. This work pre-processes the data by sum normalization and two-dimensional principal component analysis (2DPCA) and uses the dimensionality reduction results as two feature-point views. Then, semi-supervised classification of collaborative training (co-training) with K-nearest neighbors (KNN) and a decision tree (DT) is used to identify the samples. The results show that the average overall accuracy of this coupling model can reach 100%, which is 20.49% higher than that of the thickness-only view. Using unlabeled data can reduce the cost of data acquisition, improve the classification accuracy and generalization ability, and provide theoretical significance and application prospects for discrimination of spectrally similar oil species in natural marine environments.

Received 31st October 2022
Accepted 23rd February 2023

DOI: 10.1039/d2ay01776h

rsc.li/methods

1. Introduction

With increasing oil demand, oil spill accidents are occurring more frequently during oil extraction, transportation, and shipping incidents.[1] In 2010, the "Deepwater Horizon" drilling rig in the Gulf of Mexico exploded. The oil spill lasted 87 days and caused at least 2500 square kilometers of seawater to be covered with oil. In August 2020, the Japanese cargo ship "MV Wakashio" ran aground and spilled fuel oil when it struck an island near Mauritius, severely damaging one of the world's primary nature reserves. In April 2021, the Panamanian general cargo ship "Sea Justice" collided with the Liberian tanker "A Symphony" in the waters off Qingdao, Shandong Province, resulting in a spill of about 9400 tons of cargo oil into the sea. It will take more than ten years for the natural fishery resources in the area to recover to the pre-pollution level.

Light oil has low viscosity, high volatility and significant acute toxicity, and diffuses quickly, but its ability to cause sustained pollution is weak and it is easily weathered by natural processes. Therefore, it is classed as a non-persistent oil by the industry. The risk of fire and explosions is generally higher during leakage of light oil. For example, in 2018, a collision between the Panamanian tanker "SANCHI" and the "CF CRYSTAL" (from Hong Kong, China) caused a fire that completely burned the tanker "SANCHI", which was carrying 13.6 tons of condensate. However, light oil is less polluting to the environment.

Although some safety-enhancing measures have been introduced, maritime accidents are still a primary concern, because severe accidents continue to happen frequently and

[a]*Changchun Institute of Optics, Fine Mechanics and Physics, Chinese Academy of Sciences, Changchun, Jilin Province, 130033, China. E-mail: ciomp@ciomp.ac.cn; gongbowen1997@126.com;*

[b]*University of Chinese Academy of Sciences, Beijing, 100049, China. E-mail: @mails.ucas.ac.cn*

[c]*Shanghai Institute of Spacecraft Equipment, Shanghai, 200240, China*

have led to serious consequences in recent years.[2] Oil spills can be advected by wind, waves, and tides,[3] and compressed by waves and ocean currents into narrow oil slicks, while also diffusing, dissolving, emulsifying and evaporating. Oil spills can disrupt the ecological balance by affecting the exchange of gases between seawater and air, causing the death of aquatic organisms, such as floating algae and fish, and seabirds.[4,5] There are carcinogens in the oil, which accumulate in organisms and can eventually reach humans, causing harm to health.[6] Oil is also prone to fires and explosions, resulting in more serious economic losses and casualties. Reducing the risk of oil spill disasters is essential to protect ecosystems and minimize economic damage. To some extent, accurate and timely acquisition of information on the location, type, and size of oil spills is the basis for accountability of oil spills and oil spill accident reduction.[7]

Currently, several methods are used in offshore oil spill detection, such as thermal infrared,[7,8] ultraviolet (UV),[9] microwave radiometer,[10] visible light,[11] laser acoustic,[12,13] synthetic aperture radar and UV-induced fluorescence[14,15] remote sensing. Among the above detection methods, only the UV-induced fluorescence method can accurately and conveniently distinguish the type of oil spill and algae from oil spills and is one of the most prevalent oil spill detection methods, with high sensitivity and strong resistance to interference.[16] Fluorescence-based techniques feature relatively simple instrumentation, fast response speed and easy sample preparation, and are less affected by oil weathering. Therefore, UV-induced fluorescence is selected in this study to classify oil spills on the seawater surface.

Offshore hydrocarbon spills are disturbed by the complex marine environment, which will impact on the detection and identification results to a certain extent. Due to the influence of wind, tidal forces, and seabed activities, the sea surface is constantly moving, and oil slicks do not have an even thickness. Emulsification, which leads to changes in the chemical properties of the oil film, will occur. These factors will change the oil's fluorescence spectrum. Oil sample fluorescence detection and identification technologies mainly include simultaneous fluorescence spectroscopy (SFS),[17] excitation–emission fluorescence spectroscopy (EEM),[18] time-resolved fluorescence spectroscopy (TRS),[19] relative fluorescence intensity (RFI) analysis,[20] and so on.[21] Nevertheless, these methods ignore the change in the fluorescence characteristics of the oil film with its state.

Wang *et al.* proposed increasing the concentration dimension of the fluorescence spectrum and used Gabor wavelet analysis combined with a support vector machine (SVM) algorithm for spectral classification.[22] The scanning equipment is very slow, but suitable for measuring samples showing small changes over time.[9] Loh *et al.* designed a portable laser-induced fluorescence (LIF) oil spill classifier. Further, they validated the prediction performance and robustness with classification models such as partial least squares discriminant analysis (PLS-DA) and support vector machine-discriminant analysis (SVM-DA).[23,24] A novel method for oil pollution identification based on excitation–emission matrix fluorescence spectroscopy and parallel factor framework-clustering analysis (PFFCA), improving upon parallel factor analysis (PARAFAC), was presented.[2] These models are challenging to use for detecting and identifying oil spills in the natural environment, since the classification depends on a stable experimental environment. Further exploration of interference-resistant detection and identification algorithms is necessary to improve the accuracy and robustness of monitoring.

Classification algorithms do not usually take into account the variation of fluorescence with oil film thickness. In this work, the classification algorithm is based on the variation of fluorescence with thickness. Heavy oil possesses a high concentration of fluorophores, resulting in a high collisional energy transfer rate. The nonlinear variation of heavy oil fluorescence with thickness is caused by fluorescence burst and reabsorption processes.[25,26] Although the components in oil products are complex, and the nonlinear change in fluorescence cannot be quantitatively analyzed, such processes can assist in distinguishing different oil substances using ordinary fluorescence spectroscopy. We investigate the nonlinear variation of the fluorescence spectrum through the fluorescence lidar equation. The fluorescence spectra of mixed oils are similar, and the mixing proportion also affects the spectrum. These factors make classification and identification difficult. Hence, mixed oil samples are selected to test the effect of the classification algorithm.

In this study, two-dimensional principal component analysis (2DPCA) is used as the data dimensionality reduction algorithm, which preserves the inter-row relationships of the spectral data matrix and improves the speed of dimensionality reduction. We propose a semi-supervised method based on co-training, K-nearest neighbors (KNN) and a decision tree (DT) to identify the 3D UV-induced fluorescence spectral dataset of mixed oil films with varying thicknesses and proportions. The results show that the recognition accuracy of this algorithm can reach 100%. Moreover, the precision, recall, and F1-score of the coupling classifier can all reach 100%. This method achieves higher classification accuracy using less labeled data. It improves the generalization ability of the classification model and enables timely, fast and accurate discrimination of spectrally similar oil species.

2. Experimental and data processing

2.1 Experimental materials and procedures

UV light-emitting diodes (LEDs) have high stability, low noise and low energy consumption, and are low-cost. However, due to their weak light intensity, the elimination of the fluorescence background is very demanding.[27] Despite this, a UV LED is used as the excitation light source in the laboratory experiments.

Certain polycyclic aromatic hydrocarbon (PAH) compounds in petroleum absorb UV light, becoming electronically excited and emitting longer-wavelength fluorescence. It is well known that different oils have different types and proportions of PAHs.[28] Therefore, different oils have distinct fluorescence emission features that allow reliable oil classification.

Most of the oil spills in the natural sea environment are heavy oil, but it is difficult to degrade and obtain. In addition to

heavy oil, light oil spills can also occur in shallow sea areas, such as gasoline and diesel fuel leaks from motorboats at scenic spots. Acquiring fluorescence spectra of these oil samples is critical for monitoring and identifying oil spills in shallow seas. Four light oil products and two crude oils were selected as classified test samples: 95# gasoline (95# G), −35# diesel oil (−35# D), −20# diesel oil (−20# D), light crude oil, medium crude oil, and Mobil lubricant oil 20W-40 (Lube). The physical parameters of these six test oil samples are shown in Table 1. All of these light oils were bought from the local gas station. The light crude oil sample is a light oil moisture standard and the medium crude oil sample is obtained from the Changchun Oil Recovery Plant.

Oil spills resulting from shipboard accidents are often not just one type of oil but a mixture of fuel and lubricants. For example, bilge oil contains a mixture of fuel oil and lubricant.[29] Lube was mixed with light crude oil, medium crude oil, 95# G, −35# D, and −20# D in different proportions to obtain mixed oil samples with different compositions. In addition, a mixture of −35# D and −20# D was used as the interference term.

Traditional laser-induced fluorescence systems employ a UV laser operating between 308 nm and 355 nm as a source of excitation.[30] The fluorescence signal from vegetation can overlap with the fluorescence from oil samples at the excitation wavelength of 355 nm. This causes false positives for laser sensors used in oil leak control.[31] Comparing 254 nm and 310 nm LED modules, we found that the short-wavelength source is more sensitive to thin oil slicks. We note that the scattering peak of the 310 nm LED shows a weak red-shift, which can be used as a feature in PCA. But the scattering peak and the fluorescence spectrum induced by the 310 nm light source appear to overlap, which causes difficulties in the classification. The difference in the fluorescence spectra of −20# D using these two light sources is shown in Fig. 1. Fluorescence with 254 nm LED excitation starts at 275 nm (in the solar-blind UV range) with much less effect from background noise. These are the reasons why an LED module with a central wavelength of around 254 nm is used as the UV excitation light source.

An FX2000-EX optical fiber spectrometer was used to collect the fluorescence spectra; its detection range is 196–1170 nm. The spectrometer operates with a slit width of 50 μm, yielding a spectral resolution of 1.54 nm.

A semi-circular slide with an inner diameter of 136 mm was designed, with the head of the optic fiber and the light source

Table 1 Physical parameters of test oil samples

Oil samples	Density (20 °C, g mL^{-1})	API (°)	Viscosity (40 °C, mm^2 s^{-1})
95# G	0.74	60.3	0.72
−35# D	0.82	40.9	2.1
−20# D	0.83	38.8	2.8
Lube	0.89	27.3	121
Light crude oil	0.81	42.3	4.3
Medium crude oil	0.87	31.1	16.7

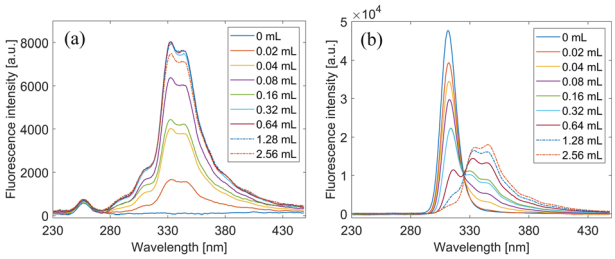

Fig. 1 Fluorescence spectra of −20# D at different LED excitation wavelengths: (a) 254 nm and (b) 310 nm.

both fixed on the slider. The workpiece sizes were selected based on the principle that the excitation light should be completely unobstructed by the probe slider. The angle of the slider and the distance between the optic fiber and light source can be adjusted as needed (Fig. 2(a)).

The experiments were performed in a dark room. As illustrated in Fig. 2(c), the incident LED light source is at 90° and the optical fiber probe of the spectrometer receives the fluorescence signal at an angle of 45°.[32] This angular configuration allows effective reception of the fluorescence signals and reduction of the effect of scattered signals. Different excitation reception angles only affect the light intensity, which can be weakened by the normalization method in spectral preprocessing.

In the first step, 60 mL of seawater (collected from Liaodong Bay) was added to a Petri dish with a diameter of 90 mm. Afterward, various volumes of mixed oil samples were dripped into Petri dishes and allowed to diffuse freely until evenly distributed on the seawater surface. The volumes of the oil samples were 0.02 mL, 0.04 mL, 0.08 mL, 0.16 mL, 0.32 mL, 0.64 mL, 1.28 mL and 2.56 mL. The used volumetric measuring tool was a dropper with a range of 1 mL and a division value of 0.02 mL. Notice that the same seawater volume and diffusion time were maintained during the experiments.

The integration time was 5000 ms. A 1 : 1 volume mixture of the two types of oil was used to measure the thickness-

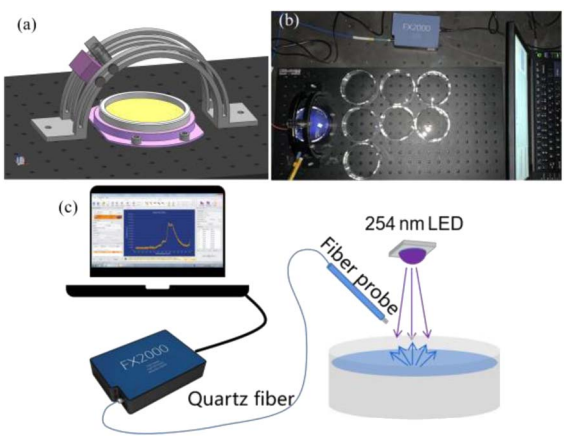

Fig. 2 Schematic (a) and photograph (b) of the water surface oil film detection system. (c) Schematic diagram of the oil film fluorescence measurement system.

dependent fluorescence spectra of the mixed oil films. Here, the new dimension of thickness is introduced.

We calculated the standard thickness (h) of the oil film according to the below formula:

$$h = v/s \quad (1)$$

where v is the volume of the oil sample, and s is the area of the oil film.

The six types of mixed light oil used in the experiment are: Lube + −35# D, Lube + −20# D, Lube + 95# G, Lube + light crude oil, Lube + medium crude oil and −20# D + −35# D. Each mixed oil sample was mixed in the ratios of 10%, 20%, 30%, 40%, 50%, 60%, 70%, 80% and 90% by volume, with a total volume of 0.16 mL. The fluorescence spectral data of the four mixed oils with different mixing proportions were measured using the same experimental procedure introduced above. Then, we introduced mixing proportions as a new dimension and obtained a 3D fluorescence matrix for mixing proportions.

Finally, the spectra of each oil sample with different thicknesses and proportions were separately integrated into a two-dimensional matrix.

2.2 Data processing

Spectral preprocessing consists of three parts: denoising, smoothing, and normalization. Denoising is achieved by subtracting the background noise spectrum from the measured fluorescence spectrum. The moving average method is used as the smoothing algorithm.

In the field of machine learning, different evaluation indicators have different dimensions and units, which will affect the results of data analysis. In order to eliminate the dimensional influence of the indicators, it is necessary to perform normalization processing to eliminate the adverse effects caused by the individual samples.

With the increase of oil film thickness, the fluorescence intensity of different oil species increases and reaches saturation, and the saturated thickness and saturated fluorescence intensity of different oil species can be distinguished. Hence, we want to preserve the discrepancies in the saturated fluorescence intensity. Data preprocessing for the previous classification used normalization, but normalization is no longer appropriate in this case. We use the summation normalization algorithm for data preprocessing, which can weaken the influence of large variable values on the model while retaining the peak characteristics.

2.2.1 2DPCA. Principle component analysis (PCA) is a commonly used data dimensionality reduction algorithm. It realizes principal component space mapping of samples by solving the eigenvectors corresponding to the first N most prominent features of the target covariance matrix to form a feature mapping matrix. After adding thickness and proportion dimensions, respectively, the data becomes a two-dimensional matrix. Traditional PCA requires conversion of the two-dimensional matrix into row vectors and combination of multiple row vectors corresponding to multiple sample images into a large-scale covariance matrix. The eigenvalues and eigenvectors of this covariance matrix are solved to construct the mapping matrix. The disadvantages of PCA are the long preprocessing time and the inability to take into account the relationship between row vectors.

2DPCA is based on the PCA algorithm to reduce the features of the two-dimensional matrix.[33] 2DPCA directly calculates the overall covariance matrix of the sample and performs dimensionality reduction in one dimension. 2D2DPCA performs 2DPCA on both sides and executes dimensionality reduction processing in two dimensions. The thickness and proportion dimension data added in this paper are relatively few. Consequently, 2DPCA is selected as the feature extraction algorithm, which retains the relationships between the thickness and

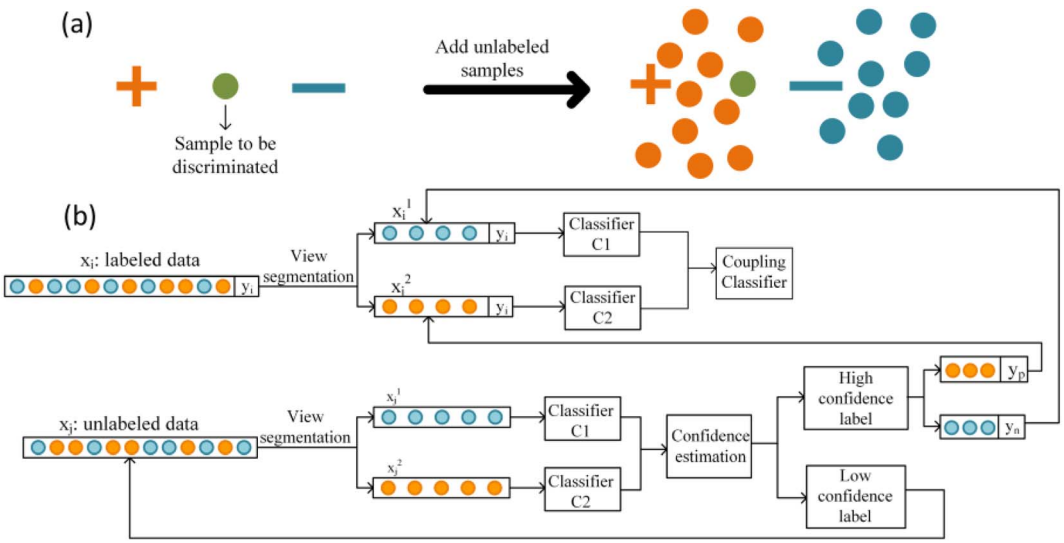

Fig. 3 Schematic of a classification algorithm. (a) Example of using unlabeled samples. (b) Schematic diagram of the improved co-training algorithm.[34]

proportion dimension data and improves the processing speed of feature extraction.

2.2.2 Semi-supervised classification. Semi-supervised learning is a machine learning algorithm used in training data that are a mixture of labeled and unlabeled data. Data with exact categories as labels are referred to as labeled samples; correspondingly, data without category attributes are unlabeled data. The principle of using unlabeled data to improve the identification accuracy is shown in Fig. 3(a). If there are only a small number of labeled samples, the distribution of feature points will not be complete, causing difficulties in classification. If the distribution of unlabeled samples is included, the low classification accuracy caused by insufficient samples will be solved.

In an actual oil spill scenario, different oil types may be present at different locations and changes may also occur, such as oil spill emulsification and diffusion. Thickness and mixing proportions will seriously affect the oil spill fluorescence spectrum. Obtaining labeled samples requires a large workforce and a lot of material resources. Using unlabeled samples as a training set can reduce the high cost of labeled data acquisition, improve the generalization ability of the classification model, and enhance the classification accuracy to a certain extent.

A schematic diagram of the improved co-training algorithm is shown in Fig. 3(b). The 3D fluorescence spectral data with increasing thickness and mixing proportion dimensions are taken as two views: View1 and View2.

(1) Part of the labeled data X_i from View1 and View2 is selected as the training set of labeled samples to train two single-view classifiers, C1 and C2.

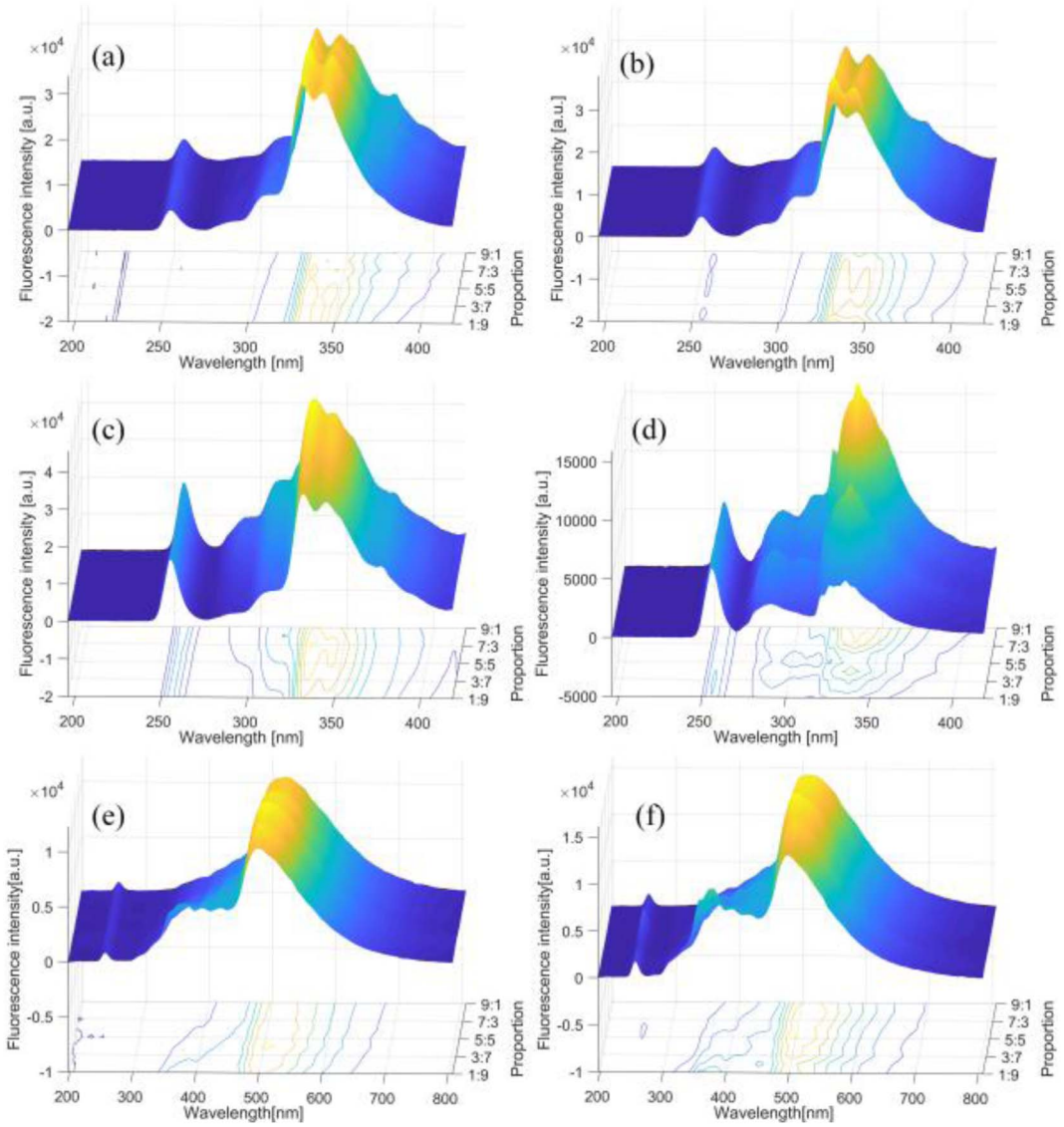

Fig. 4 Fluorescence spectra of oil films with different mixing proportions. (a) −35# D + −20# D, (b) Lube + −20# D, (c) Lube + −35# D, (d) Lube + 95# G, (e) Lube + light crude oil, (f) Lube + medium crude oil.

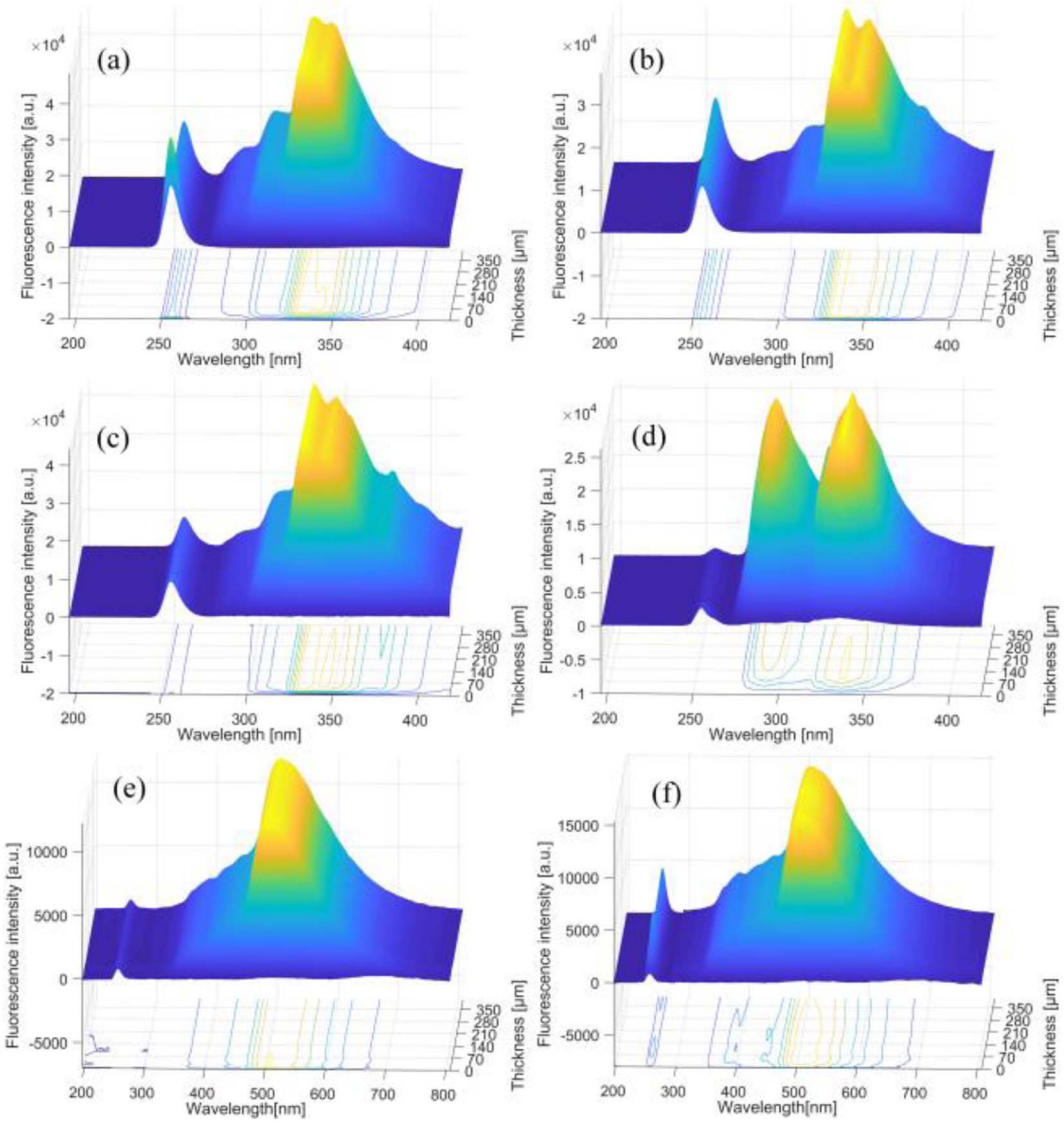

Fig. 5 Fluorescence spectra of oil films with different thickness. (a) −35# D + −20# D, (b) Lube + −20# D, (c) Lube + −35# D, (d) Lube + 95# G, (e) Lube + light crude oil, (f) Lube + medium crude oil.

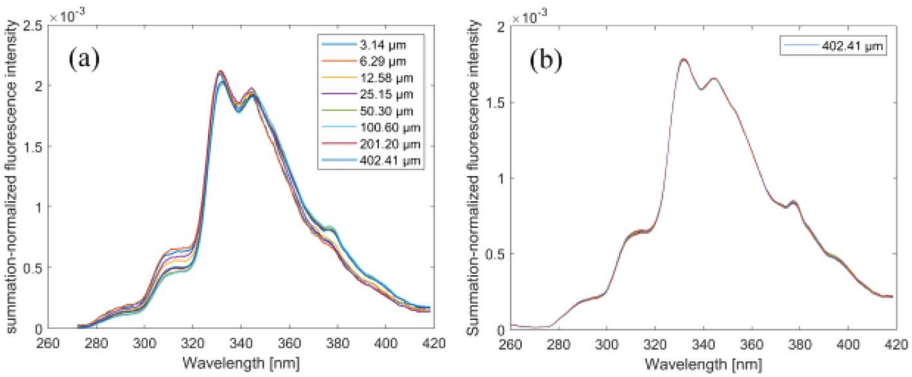

Fig. 6 The comparison of summation-normalized spectra of (a) different-thickness oil films and (b) multiple measurements of a single thickness of Lube + −35# D oil film.

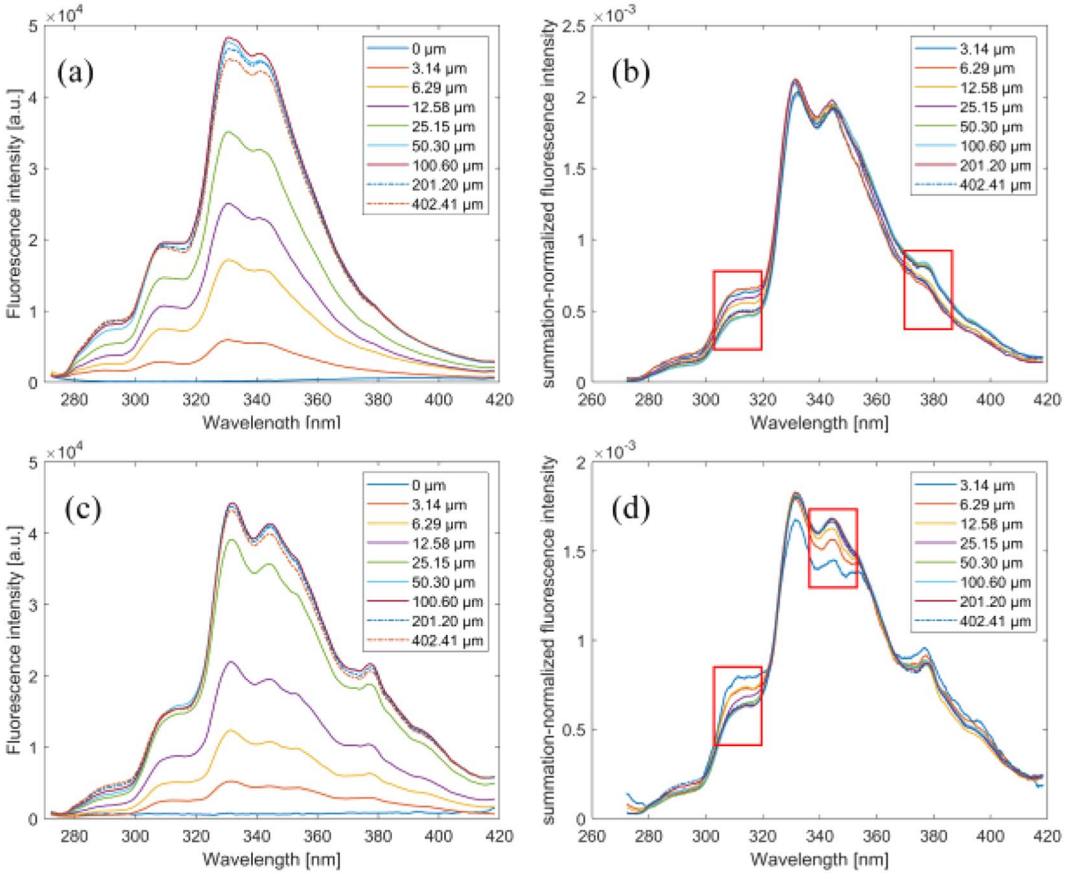

Fig. 7 −20# D + −35# D: original spectra (a) and summation-normalized fluorescence spectra (b) of different-thickness oil films. Lube + −35# D: original spectra (c) and summation-normalized fluorescence spectra (d) of different-thickness oil films.

(2) The single-view classifier is used to label the unlabeled samples. The unlabeled samples (X_j) with the highest confidence are selected as pseudo-labeled samples.

(3) The classifier is retrained with the labeled and pseudo-labeled samples.

(4) Steps 2 and 3 are looped until the accuracy of the classifier no longer improves.

(5) The prediction results of C1 and C2 are compared and coupled, and the discrimination labels are output.

3. Results and discussion

In this study, film thickness and oil proportions were used to increase the spectral characteristics of the spilled oil film. The fluorescence spectra of a series of oil films on a seawater surface with different thicknesses and mixing proportions were measured according to the experimental procedures in Section 2.1.

3.1 Fluorescence measurement results

This work measured nine different thicknesses and nine mixing proportions, and we performed 50 repeated measurements of the oil film for each thickness and each mixing proportion. Consequently, the data of each view is a 2048 × 9 × 50 3D matrix, where 2048 is the number of bands. The measured 3D spectral matrices for different mixing proportions and thicknesses are shown in Fig. 4 and 5, respectively.

3.2 Fluorescence characteristics

The normalized fluorescence spectra do not entirely overlap, as shown in the experimental results. The ratio of long-wavelength fluorescence intensity to short-wavelength fluorescence

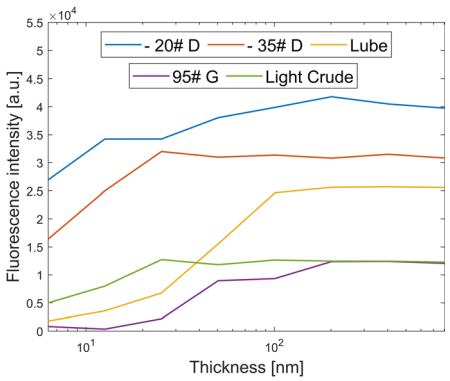

Fig. 8 Different oil fluorescence saturation values corresponding to the different thicknesses.

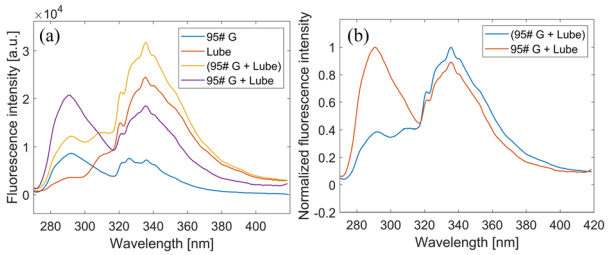

Fig. 9 Fluorescence varies nonlinearly with the mixing proportions. (a) The single oil spectra of 95# G and Lube, the arithmetic sum of these two spectra ((95# G + Lube)), and the fluorescence spectrum of the mixed oil (95# G + Lube). (b) The normalized fluorescence spectrum of mixed oil and sum of single-oil normalized fluorescence spectra .

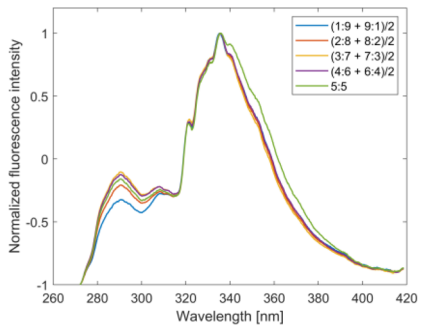

Fig. 10 Change in fluorescence spectra with the Lube and 95# G mixed oil proportions.

intensity gradually increases with the increase of oil film thickness. The reason for this is discussed below.

According to the fluorescence lidar equation,[35] the fluorescence signal $N_1(\lambda_e,\lambda_f)$ at wavelength λ_e and excitation intensity N_0, and from distance H, can be described as follows:

$$N_1(\lambda_e,\lambda_f) = N_0 T^2 \frac{r^2}{4H^2}\frac{\alpha_e}{k_e k_f}\eta(\lambda_e,\lambda_f)\{1-\exp[-(k_e+k_f)L]\} \quad (2)$$

T is the transmission coefficient at the oil–water interface; r is the aperture of the detection lens; k_e and k_f are the extinction coefficients of the oil at excitation wavelength λ_e and emission wavelength λ_f of the fluorescence. $\eta(\lambda_e,\lambda_f)$ is the fluorescence conversion efficiency of the oil film; L is the thickness of the oil film. According to eqn (2), when $\lambda_1 < \lambda_2$, the fluorescence wavelength proportion of oil films with different thicknesses is:

$$R(\lambda_{f1},\lambda_{f2},L) = \frac{N(\lambda_e,\lambda_{f2})}{N(\lambda_e,\lambda_{f1})} \quad (3)$$

The thickness of the oil film when the fluorescence reaches saturation is expressed as $L = \infty$, and when the oil film is infinitely thin it is expressed as $L = 0$. The proportion of the fluorescence wavelength at these two thicknesses is defined as $R_{\infty/0}$:

$$R_{\infty/0} = \frac{R(\lambda_{f1},\lambda_{f2},\infty)}{R(\lambda_{f1},\lambda_{f2},0)} = \frac{k_e+k_1}{k_e+k_2} \quad (4)$$

Since the extinction coefficient at short wavelengths is longer than that at long wavelengths, *i.e.* $k_2 < k_1 < k_e$, from eqn (4) we can obtain:

$$1 < R_{\infty/0} < 2 \quad (5)$$

The experimental data show that the fluorescence proportion in the range of 280–420 nm satisfies eqn (5), so the nonlinear change in light oil fluorescence is caused by the change in the extinction coefficient with wavelength.

3.2.1 Different thicknesses. The spatial distribution and temporal distribution of oil film thickness are vital factors in

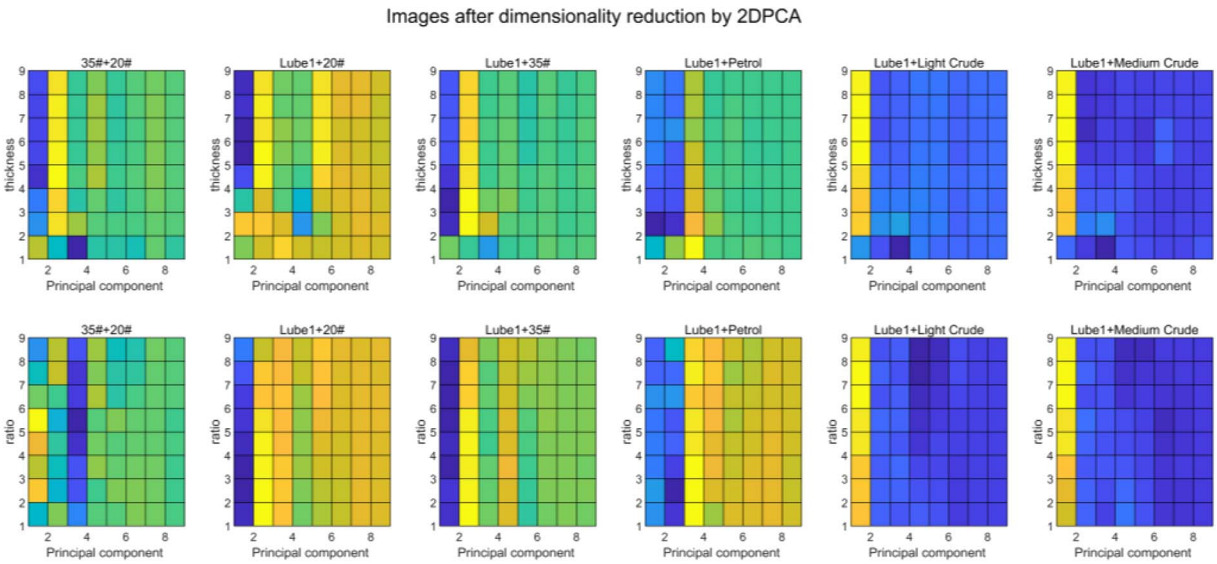

Fig. 11 Spectral images after 2DPCA dimensionality reduction.

evaluating seawater–atmosphere interaction. Only a gradual increase in fluorescence intensity can be seen from the original spectrum. The summation-normalized fluorescence spectra of Lube + −35# D oil film at multiple thicknesses do not coincide (Fig. 6(a)). However, the fluorescence spectra for multiple measurements of a single thickness are essentially coincident (Fig. 6(b)). The trends in the change of the summation-normalization spectra at long and short wavelengths are precisely opposite to each other (Fig. 7(b) and (d)).

Furthermore, as the thickness of the oil film on the seawater surface gradually increases, the fluorescence intensity of the oil film also increases, until the thickness is greater than the penetration depth of the light source, and the fluorescence reaches a saturated state.

The fluorescence saturation intensity and saturation thickness vary (Fig. 8), even for similar oil species. Due to the experimental conditions, such as the intensity of the light source in this work, the measurable thickness range of the light oil samples is 0.1–100 μm, but the thickness range of crude oil is 0.1–20 μm. This is mainly because the absorption coefficient of crude oil is greater than those of light oils. We use this feature to measure the fluorescence of oil films with different thicknesses, introducing thickness as a new dimension and obtaining 3D fluorescence spectral data. The relationships between the

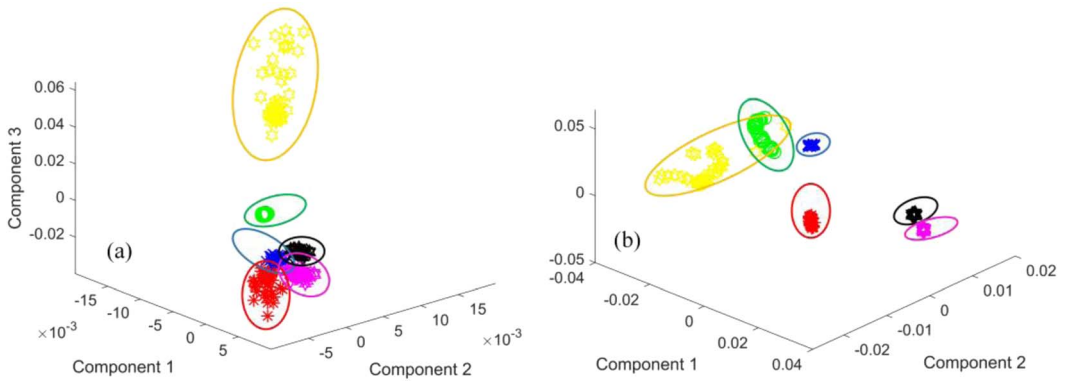

Fig. 12 Principal component analysis similarity maps of the emission spectra of mixed oils. (a) Feature points of the thickness view. (b) Feature points of the proportion view.

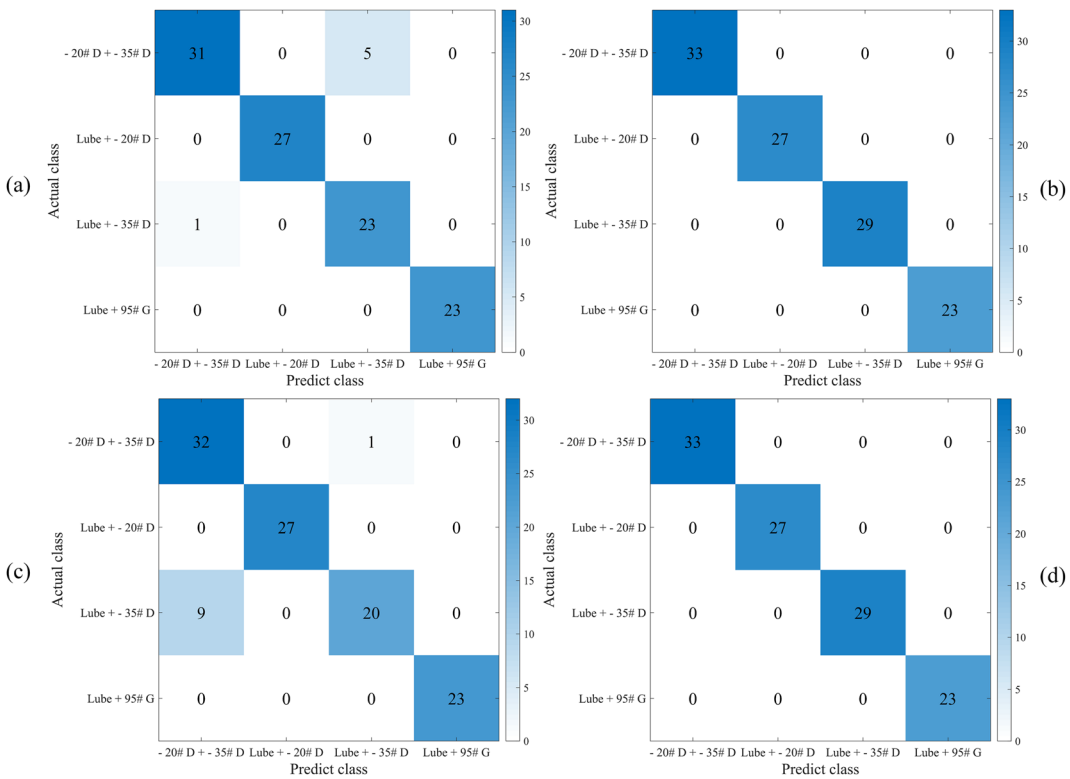

Fig. 13 The classification confusion matrices of (a) the co-training classifier with the View1 dataset, (b) the co-training classifier with the View2 dataset, (c) the KNN classifier with the View1 dataset, (d) the coupling classifier of (a) and (b).

different fluorescence intensity values when the thickness changes are retained, and the fluorescence intensity of oil films of different thicknesses is extracted as a classification feature.

3.2.2 Different mixing proportions. Obtaining the fluorescence spectrum of mixed oil products has important application value for oil spill identification, because actual oil spill accidents involve mixing oil samples of multiple oil products. The spectral similarities of mixed oil films are more prominent than those of single light oil products, and they are more challenging to identify. The fluorescence spectra of oil films with different mixing proportions at the same thickness were measured in this study. The fluorescence spectra of mixed oils with different mixing proportions are noticeably different, and the higher content oil dominates the spectrum of the mixture. Nevertheless, the spectral change of the mixed oil is not linear with the proportion changes.

The fluorescence spectra of films with different proportions of mixed oil species also change nonlinearly, and differ from the arithmetic sum of the fluorescence spectra of the separate components. The fluorescence peaks of gasoline are around 290 nm, 326 nm, and 336 nm under the same experimental conditions. In comparison, the fluorescence peak of lubricating oil is around 335 nm. The difference between the peaks of these two oil samples is enormous. We measured the fluorescence of 1.28 mL of 95# G, 1.28 mL of Lube, and 1.28 mL of a 1 : 1 mixture of these two types of oil. The oil film thickness is 402.56 μm, ensuring the fluorescence remains saturated and does not change with the thickness.

As shown in Fig. 9, there is a significant difference between the arithmetic sum of the spectral data for the single oils and the fluorescence spectral data for the blend of oils. Hence, the identification accuracy reduction caused by the mixing proportion should also be considered in the mixed oil classification dataset.

In addition, the arithmetic sum of fluorescence spectra of oils with different mixing proportions was compared with the normalized fluorescence spectrum of a 1 : 1 mixture, as shown in Fig. 10. The fluorescence is also nonlinear under ultraviolet b (UVB) radiation. Therefore, fluorescence spectra with increased thickness and proportion dimensions provide helpful features for oil species classification.

3.3 Classification results

To preserve the relationships between the thickness data and the proportion data, 2DPCA was used as a data dimensionality reduction algorithm. The first nine columns of data were selected as the classification data set for classification processing after dimensionality reduction. The spectral images of the samples after 2DPCA downscaling are shown in Fig. 11.

In this work, 30% of the data is used as the labeled training set, 20% is used as the unlabeled training set, and 50% of the data is used as the test set to verify the classification effect. This method reduces the number of sample

Table 2 The classification comparison table of KNN, co-training with KNN, and co-training with KNN and DT

	KNN				Co-training KNN			Co-training DT and KNN		
Model/indicators	Single-thickness	Single-proportion	Multi-thickness	Multi-proportion	Multi-thickness	Multi-proportion	Coupling	Multi-thickness	Multi-proportion	Coupling
Average OA	79.51%	82.42%	84.50%	85.00%	99.50%	100.00%	100.00%	95.00%	100.00%	100.00%
Max/Min OA	91.48%/50.95%	86.83%/77.04%	92.25%/63.57%	100%/78.21%	10.00%/95.61%	100.00%/100.00%	100.00%/100.00%	100.00%/89.29%	100.00%/100.00%	100.00%/100.00%
Average macro-precision	79.62%	80.03%	82.87%	83.33%	99.53%	100.00%	100.00%	95.08%	100.00%	100.00%
Average macro-recall	80.13%	87.98%	75.42%	75.98%	99.43%	100.00%	100.00%	95.27%	100.00%	100.00%
Average macro-F1-score	79.60%	80.00%	84.50%	85.00%	99.50%	100.00%	100.00%	95.00%	100.00%	100.00%
Average Kappa	75.54%	76.00%	81.23%	81.83%	99.40%	100.00%	100.00%	94.00%	100.00%	100.00%

fluorescence measurements. For example, only different mixing proportions at a certain thickness need to be measured.

We use accuracy and confusion matrix parameters to characterize the superiority of the algorithm. The fluorescence spectral feature points, considering only thickness variation, are highly similar (Fig. 12(a)). But the spectral feature points for different mixing proportions are more differentiated (Fig. 12(b)). The data from a single view cannot achieve efficient and high-precision identification, but the coupling of the classifiers from the two views can improve the accuracy, robustness, and generalization ability of the model.

The algorithm was executed one hundred times and the arithmetic mean of each evaluation parameter was calculated. The classification confusion matrices is shown in Fig. 13. The classification effect is shown in Table 2. The recognition accuracies of the KNN classifier trained with the dataset from the two views were 84.5% and 85%, respectively. The overall accuracies (OAs) of the two KNN classifiers obtained by co-training were 99.5% and 100%, respectively, and the recognition accuracy of the coupling results of the two views reached 100%. The recognition accuracies of the DT and KNN classifiers obtained by co-training were 95% and 100%, respectively. This method improves the identification accuracy and provides application prospects for oil spill detection and identification of similar oils.

4. Conclusions

Thickness and proportion factors of films of mixed oils are considered in this paper to address the effects of oil spill variations on fluorescence. Fluorescence intensity increases nonlinearly at first with the increase of oil film thickness and then reaches saturation.[26,36] The relative saturation fluorescence intensity varies for different oil products, thicknesses, and proportions. We chose to assess the fluorescence spectra of oil spills, with different thicknesses and mixing proportions, on a seawater surface in the laboratory to collect more oil spill data and simulate on-site oil spills. 2DPCA is used to extract features from the original data. This study uses the data of each factor affecting the shape of the fluorescence spectrum as an independent view, trains two classifiers utilizing the co-training method, and finally couples the two classifiers.

Introducing spectral dimensions of the influencing factors for classification reduces sample measurement time and yields more robust and accurate classifiers. It avoids the calibration of thicknesses and proportions using conventional fluorescence spectrometry detection algorithms. The recognition accuracies of the two KNN classifiers obtained by co-training are 99.5% and 100%, respectively, and the recognition accuracy of the coupling of the two classifiers reaches 100%. The overall accuracies of the DT and KNN classifiers obtained by co-training are 95% and 100%.

The following problems have to be solved for oil slick detection and identification in the actual environment. Mixed oil is not just a simple blend of two types of oil, but often a mixture of several oils. In addition to the differences in thicknesses and mixing proportions of oil spills in different locations, the difference in emulsification degree also leads to changes in spectral characteristics.[37,38] We will build a multi-view-based co-training algorithm to further improve the accuracy of UV remote on-site classification.

Author contributions

Bowen Gong: conceptualization, methodology, software, validation, formal analysis, investigation, resources, data curation, writing – original draft, visualization.

Hongji Zhang: resources, supervision.

Xiaodong Wang: writing – review and editing, funding acquisition.

Ke Lian: writing – review and editing.

Xinkai Li: writing – review and editing.

Bo Chen: writing – review and editing, supervision, project administration.

Hanlin Wang: writing – review and editing.

Xiaoqian Niu: writing – review and editing.

Conflicts of interest

There are no conflicts to declare.

Acknowledgements

This work was partially supported by the Joint Research Fund in Astronomy (grant numbers U2031122) under cooperative agreement between the National Natural Science Foundation of China (NSFC) and Chinese Academy of Sciences (CAS).

Notes and references

1 I. A. Silva, F. C. G. Almeida, T. C. Souza, K. G. O. Bezerra, I. J. B. Durval, A. Converti and L. A. Sarubbo, *Environ. Monit. Assess.*, 2022, **194**, 143.
2 Y. Cui, D. Kong, L. Kong and S. Wang, *Spectrochim. Acta, Part A*, 2021, **253**, 119586.
3 S. Rajendran, V. M. Aboobacker, V. O. Seegobin, J. A. Al Khayat, N. Rangel-Buitrago, H. A.-S. Al-Kuwari, F. N. Sadooni and P. Vethamony, *Mar. Pollut. Bull.*, 2022, **175**, 113330.
4 K. Li, H. Yu, J. Yan and J. Liao, *IOP Conf. Ser.: Earth Environ. Sci.*, 2020, **510**, 042011.
5 J. Chenhao and X. Yupeng, *IOP Conf. Ser.: Earth Environ. Sci.*, 2021, **687**, 012070.
6 S. Ji, F. Yin, W. Zhang, Z. Song, B. Qin, P. Su, J. Zhang and D. Kitazawa, *Front. Ecol. Evol.*, 2022, **10**, 850247.
7 D. Liu, X. Luan, J. Guo, T. Cui, J. An and R. Zheng, *Sensors*, 2016, **16**, 1347.
8 G. Guo, B. Liu and C. Liu, *J. Mar. Sci. Eng.*, 2020, **8**, 135.
9 Z. Suo, Y. Lu, J. Liu, J. Ding, D. Yin, F. Xu and J. Jiao, *Opt. Express*, 2021, **29**, 13486–13495.
10 A. Dala and T. Arslan, *Micromachines*, 2022, **13**, 536.
11 C. Hu, Y. Lu, S. Sun and Y. Liu, *J. Remote Sens.*, 2021, **2021**, 1–13.

12 K. Li, J. Ouyang, H. Yu, Y. Xu and J. Xu, *IOP Conf. Ser.: Earth Environ. Sci.*, 2021, **787**, 012078.
13 C. E. Brown and M. F. Fingas, *Mar. Pollut. Bull.*, 2003, **47**, 485–492.
14 M. Fingas, *Remote Sens.*, 2018, **10**, 319.
15 M. Jha, J. Levy and Y. Gao, *Sensors*, 2008, **8**, 236–255.
16 O. Bukin, D. Proschenko, C. Alexey, D. Korovetskiy, I. Bukin, V. Yurchik, I. Sokolova and A. Nadezhkin, *Photonics*, 2020, **7**, 36.
17 E. Sikorska, T. Górecki, I. V. Khmelinskii, M. Sikorski and J. Kozioł, *Food Chem.*, 2005, **89**, 217–225.
18 W. G. Mendoza, D. D. Riemer and R. G. Zika, *Environ. Sci.: Processes Impacts*, 2013, **15**, 1017.
19 E. Hegazi and A. Hamdan, *Talanta*, 2002, **56**, 989–995.
20 Y. Hou, Y. Li, B. Liu, Y. Liu and T. Wang, *Sensors*, 2017, **18**, 70.
21 Y. Hou, Y. Li, Y. Liu, G. Li and Z. Zhang, *Mar. Pollut. Bull.*, 2019, **146**, 977–984.
22 C. Wang, X. Shi, W. Li, L. Wang, J. Zhang, C. Yang and Z. Wang, *Mar. Pollut. Bull.*, 2016, **104**, 322–328.
23 A. Loh, S. Y. Ha, D. Kim, J. Lee, K. Baek and U. H. Yim, *J. Hazard. Mater.*, 2021, **416**, 125723.
24 M. V. Bills, A. Loh, K. Sosnowski, B. T. Nguyen, S. Y. Ha, U. H. Yim and J.-Y. Yoon, *Biosens. Bioelectron.*, 2020, **159**, 112193.
25 C. H. Hidrovo, R. R. Brau and D. P. Hart, *Appl. Opt.*, 2004, **43**, 894.
26 A. B. Utkin, A. Lavrov and R. Vilar, *Evaluation of oil spills by laser induced fluorescence spectra*, ed. V. Panchenko, G. Mourou and A. M. Zheltikov, Russian Federation, Kazan, 2010, p. 799415.
27 T. Zhang, Y. Liu, Z. Dai, L. Cui, H. Lin, Z. Li, K. Wu and G. Liu, *Sensors*, 2022, **22**, 1227.
28 Y. Hou, Y. Li, G. Li, M. Xu and Y. Jia, *J. Spectrosc.*, 2021, 1–10.
29 D. Lee, J. M. Seo, K. Kooistra and H. Lee, *Environ. Res.*, 2022, **212**, 113325.
30 M. Fingas and C. Brown, *Mar. Pollut. Bull.*, 2014, **83**, 9–23.
31 Yu. V. Fedotov, M. L. Belov and V. A. Gorodnichev, *J. Opt. Technol.*, 2022, **89**, 286.
32 H. T. Kieu and A. W.-K. Law, *Int. J. Remote Sens.*, 2022, **43**, 997–1014.
33 Q. Wang and Q. Gao, in *2016 IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, IEEE, Las Vegas, NV, USA, 2016, pp. 1152–1158.
34 X. Ning, X. Wang, S. Xu, W. Cai, L. Zhang, L. Yu and W. Li, *Concurr. Comput. Pract. Exp.*, 2021, e6276.
35 T. Hengstermann and R. Reuter, *Appl. Opt.*, 1990, **29**, 3218.
36 P. Camagni, A. Colombo, C. Koechler, N. Omenetto, P. Qi and G. Rossi, *Appl. Opt.*, 1991, **30**, 26.
37 X. Zhang, B. Xie, M. Zhong and H. Hao, *Opt. Commun.*, 2022, **520**, 128492.
38 J. Jiao, Y. Lu and Y. Liu, *Mar. Pollut. Bull.*, 2022, **178**, 113640.