Article

Detection and Identification of Organic Pollutants in Drinking Water from Fluorescence Spectra Based on Deep Learning Using Convolutional Autoencoder

Jie Yu , Yitong Cao, Fei Shi, Jiegen Shi, Dibo Hou * , Pingjie Huang, Guangxin Zhang and Hongjian Zhang

State Key Laboratory of Industrial Control Technology, College of Control Science and Engineering, Zhejiang University, Hangzhou 310027, China; yu_jie@zju.edu.cn (J.Y.); supercyt@zju.edu.cn (Y.C.); shifei_zju@zju.edu.cn (F.S.); sdragon@zju.edu.cn (J.S.); huangpingjie@zju.edu.cn (P.H.); gxzhang@zju.edu.cn (G.Z.); hj_zhang@zju.edu.cn (H.Z.)

* Correspondence: houdb@zju.edu.cn

Citation: Yu, J.; Cao, Y.; Shi, F.; Shi, J.; Hou, D.; Huang, P.; Zhang, G.; Zhang, H. Detection and Identification of Organic Pollutants in Drinking Water from Fluorescence Spectra Based on Deep Learning Using Convolutional Autoencoder. *Water* **2021**, *13*, 2633. <https://doi.org/10.3390/w13192633>

Academic Editor:
Alfred Paul Blaschke

Received: 2 August 2021

Accepted: 22 September 2021

Published: 25 September 2021

Publisher's Note: MDPI stays neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Copyright: © 2021 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (<https://creativecommons.org/licenses/by/4.0/>).

Abstract: Three dimensional fluorescence spectroscopy has become increasingly useful in the detection of organic pollutants. However, this approach is limited by decreased accuracy in identifying low concentration pollutants. In this research, a new identification method for organic pollutants in drinking water is accordingly proposed using three-dimensional fluorescence spectroscopy data and a deep learning algorithm. A novel application of a convolutional autoencoder was designed to process high-dimensional fluorescence data and extract multi-scale features from the spectrum of drinking water samples containing organic pollutants. Extreme Gradient Boosting (XGBoost), an implementation of gradient-boosted decision trees, was used to identify the organic pollutants based on the obtained features. Method identification performance was validated on three typical organic pollutants in different concentrations for the scenario of accidental pollution. Results showed that the proposed method achieved increasing accuracy, in the case of both high-(>10 µg/L) and low-(≤10 µg/L) concentration pollutant samples. Compared to traditional spectrum processing techniques, the convolutional autoencoder-based approach enabled obtaining features of enhanced detail from fluorescence spectral data. Moreover, evidence indicated that the proposed method maintained the detection ability in conditions whereby the background water changes. It can effectively reduce the rate of misjudgments associated with the fluctuation of drinking water quality. This study demonstrates the possibility of using deep learning algorithms for spectral processing and contamination detection in drinking water.

Keywords: contaminant detection; convolutional autoencoder; drinking water; fluorescence excitation-emission matrix (EEM); gradient-boosted decision tree

1. Introduction

Drinking water is a critical resource that affects all aspects of our life. The safety and security of drinking water have been global concerns as well as key priorities for many countries for quite a while [1–3]. Accidental and intentional water pollution may occur for a range of reasons, such as the malfunction of water treatment systems or the discharge of urban and industrial pollutants [4]. Organic pollution is one of the most common and harmful types of pollution in drinking water which is characterized by long-term toxicity and bioaccumulation. It has an adverse impact on both the health of humans and ecosystems [5–8]. Hence, a rapid and repeatable method to detect organic pollutants is an essential goal to both alert and enable users to enact follow-up responses to water pollution incidents.

Spectral analysis has served as a critical technology in water-pollutant detection systems. Compared to other conventional analytical techniques, like nuclear magnetic resonance [9], high-performance liquid chromatography [10], and gas chromatography-

mass spectrometry [11], spectral analysis is free of complex and time-consuming operations. It can be rapidly and easily implemented since sample enrichment, chemical reagent addition, and other pretreatments are not necessarily needed. Fluorescence spectroscopy, as an absorbance spectroscopy, has been used increasingly to detect pollutants in water treatment systems [12–15] due to its multiple advantages.

Fluorescence is the light emitted by a substance that has absorbed incident light or other electromagnetic radiation, and fluorescence spectroscopy has been utilized to determine the degree of organic pollution in sewage water, as well as in rivers, lakes, and urban water supplies [16–19]. Many studies have been conducted and confirmed that fluorescence spectroscopy can be used for the rapid and convenient identification and quantitative of aquatic organic pollutants. Results from fluorescence spectroscopy experiments are provided in the form of the excitation-emission matrix (EEM), which is multidimensional with the advantages of high comprehensive characterization and abundant fluorescence information. However, EEM is hard to analyze directly since it is high-dimensional.

The traditional methods to reduce the EEM dimensionality include extracting fluorescence intensity from the expert-defined spectral region (peak) [20] and multi-way methods. Principal component analysis (PCA) and parallel factor analysis (PARAFAC) [21] are some of the typical multi-way approaches. PCA is a linear transformation that enables users to project data onto the orthogonal direction of maximizing variance. Implementation of this analysis reduces EEM dimensions while retaining the features that make the most contribution to the difference. PARAFAC is a trilinear decomposition method that allows EEM to be decomposed into different components. Results from various studies indicate that PARAFAC is an effective method to process fluorescence spectroscopy data and determine the concentration of organic compounds in drinking water [22]. Yixiang Zhang analyzed the EEM obtained from fluorescence data in the downstream of watersheds by PARAFAC, and assessed fluorescence properties as proxy indicators for non-point source (NPS) pollution and labor-intensive routine water quality indicators [23]. Heibati analyzed the four independent fluorescent components of PARAFAC in drinking water and showed that dissolved organic matter (DOM) fluorescence was a sensitive indicator of drinking water quality, as long as potential interferences were taken into consideration [24]. Boehme used PCA to assess changes in fluorescence bandwidth and shifts in wavelength throughout the peak region, and it was able to monitor the seasonal and regional variations of colored dissolved organic matter in the Gulf of Mexico [25]. Yu used the ATLD method, which is an improvement on the PARAFAC method, to analyze the features of normal drinking water samples. Through the residual matrix, the threshold method enabled researchers to determine whether the drinking water was polluted or not, although they could not identify the specific types of pollutants [26].

Though widely used in dimension reduction of three-dimensional fluorescence data, PCA and PARAFAC have some limitations; for instance, their extracted features are linear. This linearity may reduce detection accuracy as a result of a loss in feature information. In recent years, many scholars have proposed other fluorescence analysis methods to compensate for it. Yang compared the applicability of three methods, and results from the relevant study demonstrated that U-PLS/RBL has obvious advantages when estimating the concentrations of six polycyclic aromatic hydrocarbons in the case of compounds displaying low fluorescence intensity [27]. Huang used a 2-D Gabor wavelet to extract features from three-dimensional fluorescence spectra, and combined it with SVM to identify pollutants present in water [28]. Peleato utilized the artificial neural network to reduce the dimensionality of high-dimensional fluorescence data, and by this approach was able to increase the accuracy of the measured concentration of disinfection by-products [29]. In addition, the growing sophistication of deep learning in image recognition has also suggested new ideas for realizing spectral feature extraction.

However, these methods hardly mention the adaptability of the model under the change of water quality background. Tap water quality is determined by the different untreated water inputs into a sewage treatment plant and these inputs also determine the concentrations of disinfection by-products. Between leaving the sewage treatment plant and reaching the consumer's tap, drinking water enters the distribution network typically for hours to days which may result in the change of drinking water quality as well. In the detection of organic pollutants in drinking water, the change of background drinking water and the low intensity of fluorescence peaks associated with the presence of low-concentration analytes lead to the failure of linear features extracted by PCA and PARAFAC according to our experiments.

The present study aims to introduce a novel method for the detection of organic pollutants in drinking water based on three-dimensional fluorescence spectra, which is applicable to the case of weak spectral signals from low-concentration analytes, in the context of background fluctuations of water quality. The deep convolutional autoencoder (CAE) is designed and used to reduce the dimensions of and extract multi-layer features from EEMs. The convolution neural network is introduced into the algorithm, which can effectively extract the neighborhood features of the three-dimensional fluorescence spectrum from the local field of view and pool layer. It guarantees the feature invariance of the organic pollutant spectrum under the background change and the automatic learning of the nonlinear features of the organic pollutant spectrum with generalization. In this study, the classifier XGBoost, a gradient boosting method, is applied to identify the organic pollutants present in drinking water. A few organic pollutants (phenol, rhodamine B, and salicylic acid) were tested to validate the mentioned approach. Two conditions of pollutants in high concentration and low concentration along with the fluctuation of water background are designed and experimented with in testing process. The results obtained applying the method thus developed were compared with those obtained via the conventional dimension reduction methods PCA and PARAFAC, so as to prove the advantages of the newly developed approach.

The remainder of this paper is organized as follows. Section 2 presents the details of our proposed method. In Section 3 results were evaluated utilizing several indicators and reported on validation dataset rather than entire dataset to provide an accurate assessment, along with discussion and explanation in detail. In Section 4, we summarize our conclusions.

2. Methodology

2.1. Architecture of Model

In the detection of organic pollutants in drinking water, the change of background drinking water and the low intensity of fluorescence peaks associated with the presence of low-concentration analytes lead to the failure of linear features extracted by PCA and PARAFAC; this problem can be solved by employing the method introduced in this paper. The detection of organic pollutants in drinking water involves a few analytical stages, including spectral preprocessing, feature extraction, and recognition model training. The overall process of the analytical method we developed is schematically depicted in Figure 1. In detail, after pretreatment, the original spectrum was divided into a training set and a testing set, according to the time of data acquisition. In the modeling phase, the training set was used to train the network parameters in CAE, with the aim of reproducing the input spectrum itself and saving the network parameters. The feature spectra of multiple channels from the coding layer were flattened as training features to train the XGBoost model, and the model parameters thus obtained were saved. At this time, the trained CAE as feature extraction model and the trained XGBoost as a discriminant model had been obtained. In the testing phase, the testing data were used as input into the trained CAE model to obtain the testing feature spectra. The trained XGBoost model was then applied to predict the category of the testing data, and the prediction results were evaluated using statistical indicators. Three typical organic pollutants (phenol, rhodamine B, and salicylic

acid) were tested to validate the approach. These three chemicals have been frequently detected organic pollutants in drinking water in recent report [30–32], and some of their fluorescence spectrum in low concentration is similar to drinking water.

Figure 1. Flow chart of the proposed method to identify and measure organic pollutants in water samples.

2.2. Spectral Pretreatment

When performing the fluorescence-based analysis of a solution, the data collected usually include signals resulting from Rayleigh and Raman scattering by the solvent, the Tyndall effect by colloidal particles, and light scattering on the surface of the container. The cubic interpolation method was employed in the present study to preprocess the original spectral data so as to reduce the influence of Rayleigh scattering and eliminate the signals due to Raman scattering was based on the method of subtracting the blank solvent background [33].

2.3. Convolutional Autoencoder

An autoencoder is a typical self-supervised learning algorithm, which is divided into two parts: encoder and decoder [34]. The encoder converts the high-dimensional input data x into a low-dimensional coded representation h ; the decoder restores the low-dimensional code to the high-dimensional original input $\hat{x}$:

$$h = f(Wx + b) \quad (1)$$

$$\hat{x} = f(W'h + b') \quad (2)$$

Here, f is a nonlinear activation function; W and W' are the weights between two layers; b and b' are the bias.

The traditional autoencoder ignores the neighborhood features of the image, and also the input layer and the hidden layer are fully connected, introducing too many redundant parameters. However, CAE solves the problem associated with the fact that the autoencoder cannot effectively extract local features through a convolutional layer that is locally connected. This convolutional structure directly processes the two-dimensional image, extracts features on the overlapping blocks, and preserves the neighborhood features of the image. Multi-layered CAE overlays form a deep CAE that can be used to extract deep spectral features [35].

Let a convolutional layer have H feature maps and the k -th feature graph has a weight matrix of W^k , an offset of b^k , and an activation function is f . Training convolution layer neurons using a three-dimensional fluorescence spectrum matrix as input x to obtain the k -th ($k = 1, 2, \dots, H$) feature map:

$$h^k = f(x * W^k + b^k) \quad (3)$$

where x is the EEM data of substances samples, $*$ represents a two-dimensional convolution, and h^k represents the k -th ($k = 1, 2, \dots, H$) feature map. The reconstruction of the feature graph is then obtained by decoder:

$$\hat{x} = f\left(\sum_{k=1}^H h^k * \bar{W}^k + c\right) \quad (4)$$

where $\bar{W}^k$ denotes the transposition of the weight matrix W^k of the k -th feature map, and c is the corresponding offset.

Notably, the purpose of this function is to minimize the value of the reconstruction error function $E(W, b)$:

$$E(W, b) = \|\hat{x} - x\|^2 \quad (5)$$

An error back-propagation algorithm similar to the BP neural network is used to calculate the gradient of the objective function:

$$\frac{\partial E(W, b)}{\partial W^k} = x * \delta h^k + \tilde{h}^k * \delta \hat{x} \quad (6)$$

where δh and $\delta \hat{x}$ represent residuals of the convolution layer and reconstruction layer, respectively.

$$W = W - \alpha \cdot \nabla_W J(W; x^{(i)}) \quad (7)$$

$$\nabla_W J(W; x^{(i)}) = x^{(i)} * \delta h^k + \tilde{h}^k * \delta \hat{x}^{(i)} \quad (8)$$

Here, α represents the learning rate and $\nabla_W J(W; x^{(i)})$ represents the gradient value obtained for a single sample.

As can be observed from Figure 2, inputs for this network are the original spectrums of the samples. Each encoder layer has a corresponding decoder layer, and each encoder layer consists of a convolution layer, a ReLU activation function, and a maximum pooling layer. For the maximum sampling layer in each encoder, the index of the maximum value on the feature map is stored. The upsampling layer in the decoder uses the location stored by the corresponding encoder to sample the feature map and reconstructs the input spectrum through the convolution layer in the decoder.

Figure 2. Representation of the procedure based on convolutional autoencoder aimed at excitation-emission matrix feature extraction.

The coder and decoder networks used in the present study consist of three layers, each with a convolution core size of 16, 8, and 6 channels, respectively. The final output of the decoder is reconstructed via the convolution layer and Sigmoid activation function.

$$f_{ReLU}(a) = \max(0, a) \quad (9)$$

$$f_{Sigmoid}(a) = \frac{1}{1 + e^{-a}} \quad (10)$$

In these equations, a is the linear output value of the neuron before it passes through the activation function. The SGD method is used to update the parameters of a single training image once at a time.

2.4. XGBoost Classifier

XGBoost is an extensible boosting tree machine learning method introduced in 2016 [36]. Gradient boosting, the original model on which XGBoost is based, is a decision tree algorithm based on iterative accumulation, and it constructs a set of weak decision trees and accumulates the results of multiple decision trees as the final predictive output. Gradient boosting uses residuals to correct previous predictors in each iteration [37]. XGBoost adds a regularization term to the objective function of gradient boosting to control the complexity of the model [38]. Its objective function is as follows:

$$J(\Theta) = L(\Theta) + \Omega(\Theta) \quad (11)$$

where Θ represents the model training parameters. For instance, L , the loss function, the mean square error or cross-entropy, measures the fit degree of the model on the training data. Ω is a regularization term that allows us to strike a balance between model complexity and accuracy. Simpler models are often hard to overfit. As the base classifier is the decision tree, the model output $\hat{y}_i$ is the vote or average of the set F of the K -tree:

$$\hat{y}_i = \sum_{k=1}^K f_k(x_i), f_k \in F \quad (12)$$

The objective function after the t -th iteration can be turned into:

$$J^{(t)} = \sum_{i=1}^n L(y_i, \hat{y}_i^{(t)}) + \sum_{k=1}^t \Omega(f_k) \quad (13)$$

where n is the number of training samples given and $\hat{y}^{(t)}$ is the prediction in the t -th iteration:

$$\hat{y}_i^{(t)} = \sum_{k=1}^t f_k(x_i) = \hat{y}_i^{(t-1)} + f_t(x_i) \quad (14)$$

The regularization term $\Omega(f_k)$ is defined as:

$$\Omega(f_k) = \gamma T + \frac{1}{2} \lambda \sum_{j=1}^T w_j^2 \quad (15)$$

where γ is the complexity of the leaf nodes in the decision tree, T denotes the number of the leaf index in each decision tree, λ is the penalty coefficient, and w is the weight of the leaf nodes.

3. Results and Discussion

3.1. Fluorescence and Sample Description

All fluorescence measurements were performed on a Hitachi F-4600 fluorescence spectrophotometer to obtain the EEM, whereby the emission and excitation wavelengths were stepped in 5 nm from 200 nm to 700 nm. Samples were kept in a 1 cm × 1 cm × 4.5 cm quartz cell, the slit width was 10 nm. The R3788 photomultiplier tube voltage with 185–750 nm spectral response range, 420 nm peak wavelength and 1200 A/lm anode luminous sensitivity, was kept constant at 700 V.

In this work, the distinct feature of the EEM was obtained as part of the search for organic pollutants in drinking water, and we utilized phenol, rhodamine B, and salicylic acid as test compounds to achieve the mentioned goal. In particular, phenol is the compound routinely checked for drinking water quality detection; rhodamine B is widely used as a dye in the cosmetic industry; salicylic acid is the main raw material in the pharmaceutical industry, and it is also used as a preservative for cosmetics. In fact, these three chemicals are frequently detected organic pollutants in drinking water [30–32].

Figure 3 shows the spectral images of four samples after pretreatment (drinking water, rhodamine B, salicylic acid, phenol in 20 µg/L, respectively). The vertical ordinate is excitation wavelength, and the horizontal ordinate is emission wavelength.

Two characteristic peaks of drinking water can be seen in Figure 3a at excitation/emission wavelengths 260–290/280–320 nm and 220–250/330–370 nm which might be from fluorescent dissolved organic matter such as protein-like and humic-like matter [39]. Although these organic components are mainly present in untreated raw water, it is hard to remove them completely by conventional processes, so this organic matter can be detected in drinking water. The characteristics of the spectra reported in Figure 3b–d indicate that two characteristic peaks of drinking water are present in the samples containing organic pollutants. Other fluorescence peaks are, however, also clearly visible. The excitation/emission wavelengths of the characteristic peak of rhodamine B are 545–555/570–580 nm; the excitation/emission wavelengths of the characteristic peak of salicylic acid are 290–300/400–410 nm. Notably, the excitation/emission wavelengths of the characteristic peak of phenol and drinking water coincide at 270–280/305–315 nm.

Figure 3. Preprocessed spectra: (a) drinking water spectrum; (b) 20 µg/L rhodamine B spectrum; (c) 20 µg/L salicylic acid spectrum; (d) 20 µg/L phenol spectrum.

3.2. Spectral Feature Extraction Based on CAE

After the three-dimensional fluorescence spectrum ($100 \times 100 \times 1$) is provided as output by the last encoder layer, the feature spectra ($13 \times 13 \times 6$) are obtained, which can be regarded as the six-channel feature map whereby the size of each channel is 13×13 . Figure 4a reports the EEM data obtained for a 20 µg/L rhodamine B solution after pre-processing. Figure 4b–h reports the feature spectra extracted from the mentioned data on six channels and the superposition of the feature spectrum on six channels, respectively. The number of feature points in one dimension is reduced to 13 from 100, the original number of wavelengths. The abscissa and ordinate in Figure 4b–h represent the feature index converted from the excitation wavelengths and the emission wavelengths respectively. Combining Figure 4a with Figure 4b–h, it is indicated that CAE seeks both high contribution and texture features from EEM. This ability and the advantages brought are presented and discussed in the following sections of the paper.

Figure 4. (a) Preprocessed EEM of rhodamin B in 20 $\mu\text{g/L}$, and its feature spectra; (b–g) feature spectrum on channel 1–6; (h) superimposed spectrum on each channel.

3.3. Qualitative Identification Results Based on XGBoost

The conventional method for the fluorescence-based identification of pollutants consists in reducing the dimension of the training spectra by PARAFAC or PCA and in training the feature matrix employing the classifier to obtain the identification model. Dimension reduction via PCA can first reduce the dimension of the training data and project the testing data into the same dimension reduction coordinate system as the training data. However, the PARAFAC-based reduction method cannot be used to apply the same load matrix as the training data to decompose the testing data, so it is not suitable for online detection. Due to the only discussion on the discriminant effect of the algorithm, and for PARAFAC, the training data and the testing data were decomposed simultaneously the concentration factor of the training data was taken as the training set, and the concentration factor of the testing data was taken as the testing set.

The fluorescence of the three organic pollutants taken into consideration in this work is more obvious when their concentrations are higher than 10 µg/L; when their concentrations are below or equal to this threshold, on the other hand, identifying some specific fluorescence peaks by macrography is difficult to achieve. Therefore, in this paper, analyte samples characterized by concentrations higher than 10 µg/L are defined as high-concentration samples, whereas those with concentrations equal or below 10 µg/L are defined low-concentration samples.

3.3.1. Detection of High-Concentration Organic Pollutants in Drinking Water

As can be evinced from Table 1, the values reported for the recall rate indicate that both multi-way methods and the deep learning method can be used to correctly identify rhodamine B, salicylic acid, and phenol in drinking water at concentrations higher than 10 µg/L. The reason for this observation is that the fluorescence peaks characteristics of these three compounds are quite different from that of drinking water. All three methods can, therefore, be utilized to extract the effective features for identification. However, the values reported for the precision rate indicate that the use of PARAFAC and PCA leads to false positives, which in turn explains the lower F1-Scores measured for PARAFAC and PCA than for CAE.

Table 1. Comparison of detection results of high concentration organic pollutants ^a.

Method	Precision Rate			Recall Rate			F1-Score		
	RhB	SA	Phenol	RhB	SA	Phenol	RhB	SA	Phenol
PARAFAC + XGBoost	100%	88%	100%	100%	100%	100%	100%	94%	100%
PCA + XGBoost	100%	88%	100%	100%	100%	100%	100%	94%	100%
CAE + XGBoost	100%	100%	100%	100%	100%	100%	100%	100%	100%

^a RhB: Rhodamine B; SA: Salicylic acid.

Figure 5 shows the main feature vectors of the three pollutant testing samples using multi-way decomposition methods, with a comparison of drinking water, respectively. As can be observed from the data reported in Figure 5, the features of the three solutions after dimensionality reduction by PARAFAC and PCA are such that, to a large extent, they are still distinguishable from those of drinking water. However, in the detection of salicylic acid, some drinking water samples are likely to be misjudged as containing salicylic acid, resulting in false positives, while it does not occur in the proposed CAE method. Figure 6 depicts the superimposed feature spectra (the numbers on the x-axis and y-axis are the index of excitation feature and emission feature extracted from CAE) on each channel of phenol, salicylic acid, and rhodamine B solutions in 20 µg/L. The characteristic peaks of rhodamine B (excitation/emission wavelengths: 545–555/ 570–580 nm) and salicylic acid (excitation/emission wavelengths: 290–300/400–410 nm) can correspond at relative positions in the characteristic spectra (the convolution-processed feature image becomes smaller, and the absolute position of each point in the original spectra are lost). Although the characteristic peaks of phenol coincide with those of drinking water (excitation/emission wavelengths: 270–280/305–315 nm), data following CAE-based dimension reduction still show the peak superposition caused by phenol at characteristic peaks and the spectral differences at non-coincidence peaks. Therefore, high-concentration samples of the three organic compounds can still be distinguished from drinking water, after performing CAE-based dimension reduction.

Figure 5. Identification of high concentration testing samples using multi-way decomposition methods: rhodamine B using (a) PARAFAC and (b) PCA; salicylic acid using (c) PARAFAC and (d) PCA; phenol using (e) PARAFAC and (f) PCA.

Figure 6. Cont.

Figure 6. Feature spectrum of 20 µg/L (a) rhodamine B, (b) salicylic acid, and (c) phenol.

3.3.2. Detection of Low-Concentration Organic Pollutants in Drinking Water

In the case of low-concentration samples of organic pollutants, the methods presented in this paper performed outstandingly, as can be evinced from the data reported in Table 2. The values of the precision rate indicate that none of the three methods for dimension reduction leads to misreporting of the water samples when identifying rhodamine B, and they can be used to successfully detect rhodamine B in low-concentration samples. However, for identifying the other two pollutants, CAE achieved better performance. CAE resulted in a zero-misreporting rate of water samples, and can identify salicylic acid except for the samples in 1 µg/L. In the case of low-concentration phenol samples, data dimension reduction with PARAFAC and PCA only allows detecting few pollutant samples in the 3–10 µg/L concentration range, whereas CAE performs much better, and it allows detecting phenol in almost all samples, with the exception of the phenol samples in the 1–2 µg/L concentration range.

Table 2. Comparison of detection results of low concentration organic pollutants ^a.

Method	Precision Rate			Recall Rate			F1-Score		
	RhB	SA	Phenol	RhB	SA	Phenol	RhB	SA	Phenol
PARAFAC + XGBoost	100%	75%	100%	100%	88%	50%	100%	81%	67%
PCA + XGBoost	100%	79%	100%	100%	88%	61%	100%	83%	76%
CAE + XGBoost	100%	100%	100%	100%	88%	78%	100%	94%	88%

^a RhB: Rhodamine B; SA: Salicylic acid.

As can be evinced from the data in Figure 7, the features of low-concentration rhodamine B samples extracted by multi-way methods maintain good discrimination for both training set and testing set, so a low-concentration rhodamine B sample can easily be distinguished from (“pure”) drinking water. In the case of low-concentration salicylic acid samples, the features extracted by PARAFAC and PCA are such that it is difficult to distinguish some drinking water samples from low-concentration salicylic acid samples, which may, in turn, lead to identifying false positives. In the case of phenol, as already mentioned, an overlap exists between this compound’s characteristic fluorescence peak and those of drinking water. Therefore, the lower the phenol concentration, the harder it is for PARAFAC and PCA to extract features that allow clear discrimination between “clean” water samples and polluted ones. Although the training features of phenol and drinking water extracted by the two methods display substantial differences, the discriminant model based on the training samples has poor adaptability to the testing samples. As can be evinced from the data in Figures 7 and 8, the classification boundary of training samples is not suitable for phenol identification in the testing samples. The main reason for this observation is that the two multi-way methods only extract the linear features of the spectrum and are insensitive to the change in background water quality. Compared with the training data, with the two methods it is difficult to extract the features that are helpful in detecting phenol after a change in background water quality. Therefore, although PARAFAC and

PCA can extract features that enable users to distinguish a phenol solution from a drinking water sample, the linear features extracted from samples at different times cannot be used in the detection of feature solution samples.

Figure 7. Identification of low-concentration testing samples using multi-way decomposition methods: rhodamine B using (a) PARAFAC and (b) PCA; salicylic acid using (c) PARAFAC and (d) PCA; phenol using (e) PARAFAC and (f) PCA.

Figure 8. Distinction of low-concentration training samples after dimension reduction of (a) PARAFAC and (b) PCA.

Figure 9 reports the feature spectra of channel 4, which indicate that the largest difference between phenol and drinking water is observed in the case of a phenol solution at 4 $\mu\text{g/L}$ concentration; at this concentration level, no false positives are produced. At the same time, compared with the feature spectrum of channel 4 in the training set, the similarity between the two feature spectrums is very high, as well as in other channels. Therefore, the discriminant model based on the training set provides data that afford the accurate detection of phenol in the testing set. This also explains why feature extraction by CAE is a far more effective approach to detecting phenol in solution than those based on conventional methods.

Figure 9. Feature spectrum in channel 4: (a) 4 $\mu\text{g/L}$ testing phenol; (b) testing drinking water; (c) 4 $\mu\text{g/L}$ training phenol.

3.3.3. Influence of Background Fluctuations in Drinking Water Quality

Affected as it is by the activity of water treatment plants and the change of substances in the transportation process, the quality of drinking water often shows fluctuations. Based on the previously established detection model of organic pollutants, we recorded the fluorescence spectra of drinking water sampled at uniform time intervals over a period of 3 months. Figure 10 reports the fluorescence spectra of drinking water samples 1–4, which cover the mentioned 3-month interval. As can be evinced from this figure, the water quality fluctuates only slightly between samples 1 and 2 and between samples 3 and 4; however, the water quality changes drastically between samples 2 and 3. It was preliminarily analyzed in the last section that in the low-concentration case, one of the reasons that the proposed method performed better than PCA and PARAFAC is that it can adapt to the water quality background fluctuations between training samples and testing samples. In this section, the above 200 drinking water samples were added into previous testing samples to further analyze and prove that the proposed method can effectively reduce the impact of water background fluctuations.

Figure 10. Cont.

Figure 10. Drinking water spectra for (a) day 1, (b) day 2, (c) day 3, (d) day 4.

Table 3 presents the misreporting rate of recognizing drinking water as pollutants, and Table 4 presents the misreporting rate of treating pollutants as normal water samples, using the three methods respectively. As can be observed from the data in Tables 3 and 4, CAE achieved a much lower misreporting rate compared with PCA and PARAFAC, especially in phenol recognition. The reason was further investigated by looking at the features extracted by the three methods in both training and setting samples, shown in Figures 11 and 12. As shown in Figure 11b, the feature of drinking water extracted by PCA in the training samples is not well adapted to the testing samples when facing the change of background water quality, which leads to a much higher misreporting rate. Though PARAFAC achieved a relatively low misreporting rate of drinking water as pollutants, it still missed 28.1% phenol samples as normal water samples. The main reason for this observation is that the linear features extracted by PARAFAC do not contain enough information to effectively distinguish phenol from drinking water, given that the characteristic fluorescence peak of phenol overlaps with the background drinking water spectrum, a problem exacerbated by changes in drinking water quality and the weakness of the fluorescence signal at low phenol concentration; by contrast, feature extraction by CAE greatly reduces the gravity of the described problem.

Table 3. Misreporting rate of drinking water samples to relative pollutant.

Method	Rhodamine B	Salicylic Acid	Phenol
PARAFAC + XGBoost	2%	3%	2%
PCA + XGBoost	2%	2%	14%
CAE + XGBoost	0%	0%	0%

Table 4. Misreporting rate of organic pollutants to drinking water.

Method	Rhodamine B	Salicylic Acid	Phenol
PARAFAC + XGBoost	0%	7%	28.1%
PCA + XGBoost	0%	7%	21.9%
CAE + XGBoost	0%	7%	14.3%

As can be evinced from the feature spectra reported in Figure 12, CAE enables users to extract the feature spectrum of drinking water in the training set and testing set. The characteristic peaks in the feature spectrum are clearly visible, and the interference caused by the background change is removed. In contrast with the cases of PCA and PARAFAC, the CAE-based feature extraction method proposed in the present study achieves a zero misreporting rate. Based on the CAE model with trained parameters, the algorithm can be directly used for online pollutant detection.

Figure 11. Drinking water features extracted by (a) PARAFAC and (b) PCA.

Figure 12. Drinking water feature spectrum of (a) training set and (b) test set using CAE.

4. Conclusions

A new procedure utilizing CAE to process high-dimensional fluorescence data and XGBoost to classify features for organic pollutants in drinking water was presented and shown to be advantageous over traditional approaches. Results of this work indicate that, although the traditionally utilized dimension reduction methods PCA and PARAFAC can, to some extent, extract distinguishing features from EEM, their use is associated with some limitations. For example, the identification performance of PCA and PARAFAC is poor when the concentration of the pollutant is relatively low because of linear features. In addition, PCA and PARAFAC are easier to be disturbed. The linear features from the above methods are better distinguishable mainly when the characteristic fluorescence peaks of the organic pollutants do not overlap with the peaks of “unpolluted” drinking water and when the concentration of the pollutant is relatively high. By contrast, CAE implementation can afford the acquisition of multi-layer convolutional features and reduce information loss, due to the deep layer construction associated with this method. Consequently, CAE enables users to collect both high contribution and texture features from spectra, resulting in a better pollutant identification performance, especially in the context of low pollutant concentrations and substantial background water quality fluctuations.

The results presented in this paper suggest the applicability of the convolutional autoencoder to the interpretation of fluorescence results, as well as the applicability of deep learning approaches to the detection of pollutants in water. The method developed as part of this work is also well suited to processing large amounts of high-dimension datasets. As online spectrometers are being rapidly developed and online monitoring sites quickly grow in number, the novel approach described herein may find potential application in online monitoring and as part of early warning systems for drinking water contamination.

Author Contributions: The project was conceived by J.Y. and D.H. Methodology development, investigation, and analysis were performed by J.Y., Y.C., F.S. and J.S. Sample collection was performed by Y.C. and F.S. This manuscript was written by J.Y. and Y.C. with editing and review contribution was performed by all other authors. The overall supervision was performed by D.H., P.H., G.Z. and H.Z. All authors have read and agreed to the published version of the manuscript.

Funding: This work was funded by the National Natural Science Foundation of China (No.61803333; U1509208), the Key Technology Research and Development Program of Zhejiang Province (No. 2021C03177), and the National Key R&D Program of China (2017YFC1403801).

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: No new data were created or analyzed in this study. Data sharing is not applicable to this article.

Conflicts of Interest: The authors declare no conflict of interest.

Abbreviations

The following abbreviations are used in this manuscript:

PCA	Principal component analysis
PARAFAC	Parallel factor analysis
CAE	Convolutional autoencoder
RhB	Rhodamine B
SA	Salicylic acid

References

1. Ahmed, M.; Mokhtar, M.B.; Majid, N. Household Water Filtration Technology to Ensure Safe Drinking Water Supply in the Langat River Basin, Malaysia. *Water* **2021**, *13*, 1032. doi:10.3390/w13081032. [[CrossRef](#)]
2. Hou, D.; Zhang, J.; Yang, Z.; Liu, S.; Huang, P.; Zhang, G. Distribution water quality anomaly detection from UV optical sensor monitoring data by integrating principal component analysis with chi-square distribution. *Opt. Express* **2015**, *23*, 17487–17510. [[CrossRef](#)]
3. Ido, A.; Hiromori, Y.; Meng, L.; Usuda, H.; Nagase, H.; Yang, M.; Hu, J.; Nakanishi, T. Occurrence of fibrates and their metabolites in source and drinking water in Shanghai and Zhejiang, China. *Sci. Rep.* **2017**, *7*, 45931. [[CrossRef](#)] [[PubMed](#)]
4. Xiang, R.; Xu, Y.; Liu, Y.Q.; Lei, G.Y.; Liu, J.C.; Huang, Q.F. Isolation distance between municipal solid waste landfills and drinking water wells for bacteria attenuation and safe drinking. *Sci. Rep.* **2019**, *9*, 17881. [[CrossRef](#)] [[PubMed](#)]
5. Persichetti, G.; Testa, G.; Bernini, R. High sensitivity UV fluorescence spectroscopy based on an optofluidic jet waveguide. *Opt. Express* **2013**, *21*, 24219–24230. [[CrossRef](#)] [[PubMed](#)]
6. Zahn, D.; Frömel, T.; Knepper, T.P. Halogenated methanesulfonic acids: A new class of organic micropollutants in the water cycle. *Water Res.* **2016**, *101*, 292–299. [[CrossRef](#)]
7. Gowland, D.; Robertson, N.; Chatzisyneon, E. Photocatalytic Oxidation of Natural Organic Matter in Water. *Water* **2021**, *13*, 288. [[CrossRef](#)]
8. Wee, S.Y.; Aris, A.Z.; Yusoff, F.M.; Praveena, S.M. Occurrence of multiclass endocrine disrupting compounds in a drinking water supply system and associated risks. *Sci. Rep.* **2020**, *10*, 17755. [[CrossRef](#)]
9. Majumdar, R.D.; Blumkin, L.; Lane, D.; Soong, R.; Simpson, M.; Simpson, A.J. Analysis of DOM phototransformation using a looped NMR system integrated with a sunlight simulator. *Water Res.* **2017**, *120*, 64–76. [[CrossRef](#)]
10. Zhao, Y.; Yuan, Y.; Chen, J.; Li, M.; Pu, X. Chemometrics-enhanced high performance liquid chromatography strategy for simultaneous determination on seven nitroaromatic compounds in environmental water. *Chemom. Intell. Lab. Syst.* **2018**, *174*, 149–155. [[CrossRef](#)]
11. Kalscheur, K.N.; Penskar, R.R.; Daley, A.D.; Pechauer, S.M.; Kelly, J.J.; Peterson, C.G.; Gray, K.A. Effects of anthropogenic inputs on the organic quality of urbanized streams. *Water Res.* **2012**, *46*, 2515–2524. [[CrossRef](#)]
12. Go, R.J.; Yang, H.L.; Kan, C.C.; Ong, D.C.; Segura, S.G.; de Luna, M.D.G. Natural Organic Matter Removal from Raw Surface Water: Benchmarking Performance of Chemical Coagulants through Excitation-Emission Fluorescence Matrix Spectroscopy Analysis. *Water* **2021**, *13*, 146. [[CrossRef](#)]

13. Weishaar, J.L.; Aiken, G.R.; Bergamaschi, B.A.; Fram, M.S.; Fujii, R.; Mopper, K. Evaluation of Specific Ultraviolet Absorbance as an Indicator of the Chemical Composition and Reactivity of Dissolved Organic Carbon. *Environ. Sci. Technol.* **2003**, *37*, 4702–4708. [\[CrossRef\]](#)
14. Shutova, Y.; Baker, A.; Bridgeman, J.; Henderson, R.K. Spectroscopic characterisation of dissolved organic matter changes in drinking water treatment: From PARAFAC analysis to online monitoring wavelengths. *Water Res.* **2014**, *54*, 159–169. [\[CrossRef\]](#) [\[PubMed\]](#)
15. Wilske, C.; Herzsprung, P.; Lechtenfeld, O.; Kamjunke, N.; Einax, J.; von Tümpling, W. New Insights into the Seasonal Variation of DOM Quality of a Humic-Rich Drinking-Water Reservoir—Coupling 2D-Fluorescence and FTICR MS Measurements. *Water* **2021**, *13*, 1703. [\[CrossRef\]](#)
16. Sgroi, M.; Roccaro, P.; Korshin, G.V.; Vagliasindi, F.G.A. Monitoring the Behavior of Emerging Contaminants in Wastewater-Impacted Rivers Based on the Use of Fluorescence Excitation Emission Matrixes (EEM). *Environ. Sci. Technol.* **2017**, *51*, 4306–4316. [\[CrossRef\]](#)
17. Baker, A. Fluorescence Excitation-Emission Matrix Characterization of River Waters Impacted by a Tissue Mill Effluent. *Environ. Sci. Technol.* **2002**, *36*, 1377–1382. [\[CrossRef\]](#)
18. Shi, F.; Mao, T.; Cao, Y.; Yu, J.; Hou, D.; Huang, P.; Zhang, G. Morphological Grayscale Reconstruction and ATLD for Recognition of Organic Pollutants in Drinking Water Based on Fluorescence Spectroscopy. *Water* **2019**, *11*, 1859. [\[CrossRef\]](#)
19. Papageorgiou, A.; Papadakis, N.; Voutsas, D. Fate of natural organic matter at a full-scale Drinking Water Treatment Plant in Greece. *Environ. Sci. Pollut. Res.* **2016**, *23*, 1841–1851. [\[CrossRef\]](#)
20. Goldman, J.H.; Rounds, S.A.; Needoba, J.A. Applications of Fluorescence Spectroscopy for Predicting Percent Wastewater in an Urban Stream. *Environ. Sci. Technol.* **2012**, *46*, 4374–4381. [\[CrossRef\]](#)
21. Bridgeman, J.; Bierzoza, M.; Baker, A. The application of fluorescence spectroscopy to organic matter characterisation in drinking water treatment. *Rev. Environ. Sci. Bio/Technol.* **2011**, *10*, 277. [\[CrossRef\]](#)
22. Baghoth, S.; Sharma, S.; Amy, G. Tracking natural organic matter (NOM) in a drinking water treatment plant using fluorescence excitation-emission matrices and PARAFAC. *Water Res.* **2011**, *45*, 797–809. [\[CrossRef\]](#)
23. Zhang, Y.; Liang, X.; Wang, Z.; Xu, L. A novel approach combining self-organizing map and parallel factor analysis for monitoring water quality of watersheds under non-point source pollution. *Sci. Rep.* **2015**, *5*, 16079. [\[CrossRef\]](#)
24. Heibati, M.; Stedmon, C.A.; Stenroth, K.; Rauch, S.; Toljander, J.; Säve-Söderbergh, M.; Murphy, K.R. Assessment of drinking water quality at the tap using fluorescence spectroscopy. *Water Res.* **2017**, *125*, 1–10. [\[CrossRef\]](#)
25. Boehme, J.; Coble, P.; Conmy, R.; Stovall-Leonard, A. Examining CDOM fluorescence variability using principal component analysis: seasonal and regional modeling of three-dimensional fluorescence in the Gulf of Mexico. *Mar. Chem.* **2004**, *89*, 3–14. [\[CrossRef\]](#)
26. Yu, J.; Zhang, X.; Hou, D.; Chen, F.; Mao, T.; Huang, P.; Zhang, G. Detection of water contamination events using fluorescence spectroscopy and alternating trilinear decomposition algorithm. *J. Spectrosc.* **2017**, *2017*, 1485048. [\[CrossRef\]](#)
27. Yang, R.; Zhao, N.; Xiao, X.; Yin, G.; Yu, S.; Liu, J.; Liu, W. Quantifying PAHs in water by three-way fluorescence spectra and second-order calibration methods. *Opt. Express* **2016**, *24*, A1148–A1157. [\[CrossRef\]](#)
28. Huang, P.; Mao, T.; Yu, Q.; Cao, Y.; Yu, J.; Zhang, G.; Hou, D. Classification of water contamination developed by 2-D Gabor wavelet analysis and support vector machine based on fluorescence spectroscopy. *Opt. Express* **2019**, *27*, 5461–5477. [\[CrossRef\]](#)
29. Peleato, N.M.; Legge, R.L.; Andrews, R.C. Neural networks for dimensionality reduction of fluorescence spectra and prediction of drinking water disinfection by-products. *Water Res.* **2018**, *136*, 84–94. [\[CrossRef\]](#) [\[PubMed\]](#)
30. Rahmani, M.; Kaykhai, M.; Sasaki, M. Application of Taguchi L16 design method for comparative study of ability of 3A zeolite in removal of Rhodamine B and Malachite green from environmental water samples. *Spectrochim. Acta Part A Mol. Biomol. Spectrosc.* **2018**, *188*, 164–169. [\[CrossRef\]](#)
31. Hu, R.; Zhang, L.; Hu, J. Study on the kinetics and transformation products of salicylic acid in water via ozonation. *Chemosphere* **2016**, *153*, 394–404. [\[CrossRef\]](#)
32. Ren, L.F.; Adeel, M.; Li, J.; Xu, C.; Xu, Z.; Zhang, X.; Shao, J.; He, Y. Phenol separation from phenol-laden saline wastewater by membrane aromatic recovery system-like membrane contactor using superhydrophobic/organophilic electrospun PDMS/PMMA membrane. *Water Res.* **2018**, *135*, 31–43. [\[CrossRef\]](#)
33. Zhu, G.; Bian, Y.; Hursthouse, A.S.; Wan, P.; Szymanska, K.; Ma, J.; Wang, X.; Zhao, Z. Application of 3-D Fluorescence: Characterization of Natural Organic Matter in Natural Water and Water Purification Systems. *J. Fluoresc.* **2017**, *27*, 2069–2094. [\[CrossRef\]](#)
34. Goodfellow, I.; Bengio, Y.; Courville, A. *Deep Learning*; MIT Press: Cambridge, MA, USA, 2016.
35. Masci, J.; Meier, U.; Cireşan, D.; Schmidhuber, J. Stacked Convolutional Auto-Encoders for Hierarchical Feature Extraction. In *Artificial Neural Networks and Machine Learning—ICANN 2011*; Honkela, T., Duch, W., Girolami, M., Kaski, S., Eds.; Springer: Berlin/Heidelberg, Germany, 2011; pp. 52–59. [\[CrossRef\]](#)
36. Chen, T.; Guestrin, C. XGBoost: A Scalable Tree Boosting System. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD '16)*, San Francisco, CA, USA, 13–17 August 2016; ACM: New York, NY, USA, 2016; pp. 785–794. [\[CrossRef\]](#)

-
37. Friedman, J.H. Greedy Function Approximation: A Gradient Boosting Machine. *Ann. Stat.* **2001**, *29*, 1189–1232. [[CrossRef](#)]
 38. Zhou, K.B.; Zhang, Z.X.; Liu, J.; Hu, Z.X.; Duan, X.K.; Xu, Q. Anode effect prediction based on a singular value thresholding and extreme gradient boosting approach. *Meas. Sci. Technol.* **2018**, *30*, 015104. [[CrossRef](#)]
 39. Xiaoli, C.; Guixiang, L.; Xin, Z.; Yongxia, H.; Youcai, Z. Fluorescence excitation–emission matrix combined with regional integration analysis to characterize the composition and transformation of humic and fulvic acids from landfill at different stabilization stages. *Waste Manag.* **2012**, *32*, 438–447. [[CrossRef](#)] [[PubMed](#)]