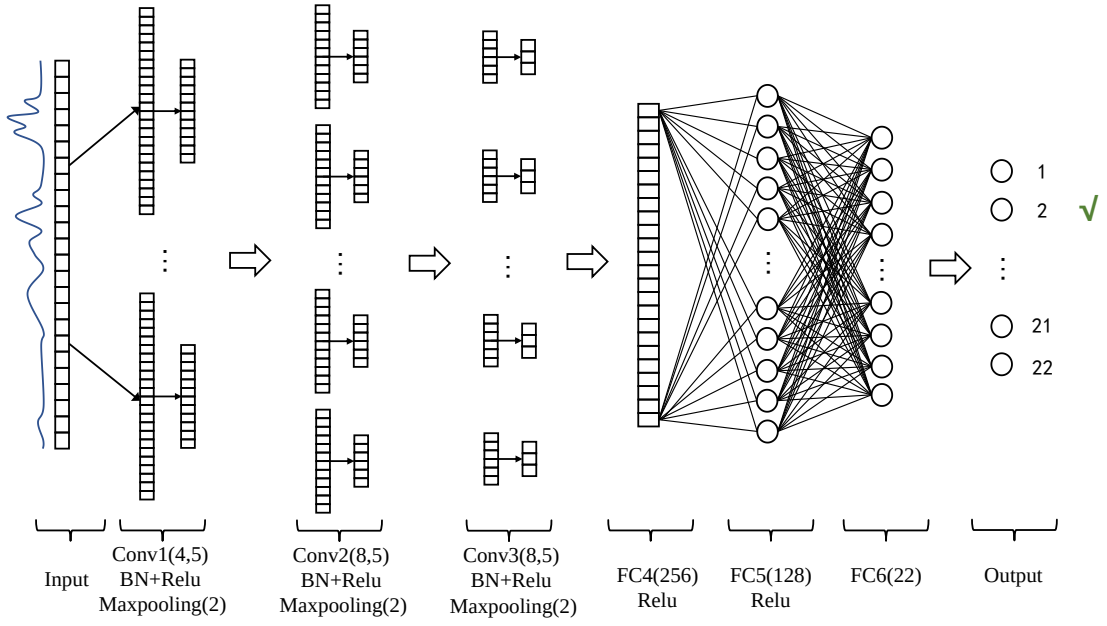
Industrial Wastewater Source Tracing: The Initiative of SERS Spectral Signature Aided with One-dimensional Convolutional Neural Network

Yuting Huang, Bingxue Yuan, Xueqing Wang, Yongsheng Dai, Dongmei Wang, Zhengjun Gong, Junmin Chen, Li Shen*, Meikun Fan* and Zhilin Li

Faculty of Geosciences and Environmental Engineering, Southwest Jiaotong University, Chengdu, 610031, China. E-mail: lishen@swjtu.edu.cn, meikunfan@gmail.com

The architecture of 1D-CNN:

The architecture of the proposed 1D-CNN model is illustrated in Scheme S1. It contains an input layer, three convolutional units, three fully connected layers, and an output layer. Each convolutional unit consists of a convolutional layer, followed by a batch normalization (BN) layer with a rectified linear unit (ReLU) layer.



Scheme S1 Model architecture of the one-dimensional convolutional neural network. FC: fully connected layer; Conv: convolutional layers; BN: batch normalization layer; Relu: rectified linear unit layer.

(1) **Input layer**: Since the randomness of the enhancement effect, the signal intensity of the SERS spectra varies greatly. Therefore, the SERS spectra were normalized by linearly rescaling the intensity values to the range of [0,1], which can reduce the training time and improve the accuracy of the model. The SERS spectrum is processed with the one-dimensional convolution kernel and each spectrum data is divided into multiple batches, and the batch size is set as 4.

(2) **Convolutional layers**: The normalized and vectorized SERS spectra are inputted into the convolutional layer for feature learning. Three one-dimensional convolution layers are established to extract the spectral feature information. The convolution kernel size is 5×1. The number of output channels is 4, 8 and 8, respectively.

(3) **BN layers**: The BN layer can standardize the output from the previous layer. The calculation is conducted using the following Equation 1.

$$z^i = \gamma \frac{y^i - \mu}{\sqrt{(\sigma^2 + \varepsilon)}} + \beta \qquad (1)$$

where z^i is the result processed by BN, y^i is the i^{th} input data, μ and σ are the mean and variance of the input array respectively, ε is a constant that improves numerical stability when the variance is very small. γ and β are used for rescaling and shifting the vector containing values from the previous operations. The introduction of the BN layer can improve the gradient of the network, acquiring a better learning rate. At the

same time, it can completely disrupt the training data and reduce the dependence on the initial data, avoiding the overfitting of the model.

(4) **Activation function layers**: The most important advantage of an activation function is its ability to apply non-linear transformations to the input data, making it capable to handle more complex non-linear tasks. Training a deep network with ReLU tends to converge much more quickly and reliably, thus, ReLU function is used here as the activation function.

(5) **Pooling layers**: The pooling layer is used to reduce the dimensionality of feature maps while retaining the prominent features. This processing step can reduce the number of parameters, which both save the training time and prevent overfitting. Generally, the pooling layer is attached after the convolutional layers to downsample the output feature maps. In this work, each convolutional layer is followed by a max-pooling layer, and the maximum value in the sampling window of size 2 is taken as the output feature value.

(6) **Fully connected layers**: After the operations of the convolutional layer, BN, activation function and pooling layer, the neural network effectively extracts the multi-level features from the original data. Because these features have a depth greater than one, a flattening processing is used to concatenate them to a single feature vector which can then be fed into the fully connected layer. It should be noted that the fully connected layers are exactly same as the traditional artificial neural networks and perform the same mathematical operations.

(7) **Output layer**: Output layer can be used as classifiers to identify wastewaters

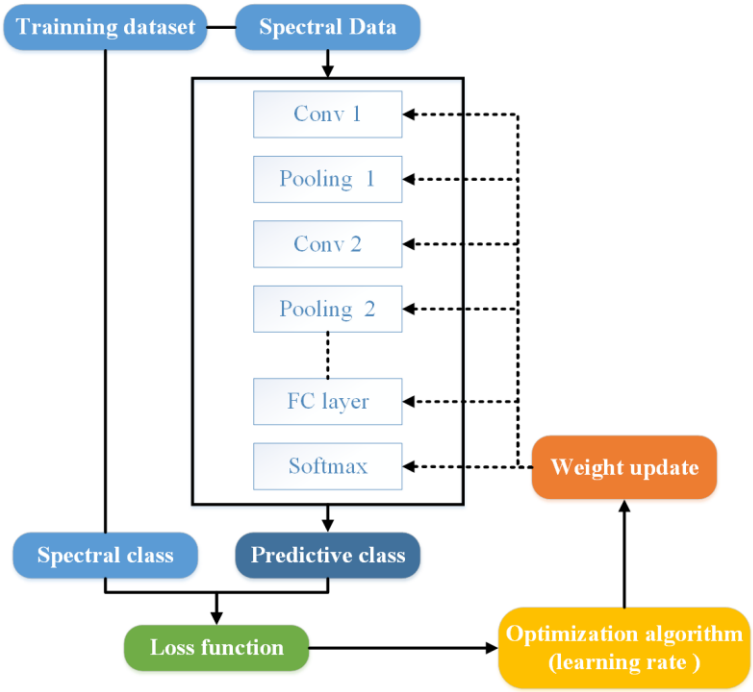
according to the features extracted by the pervious layers from SERS spectral data. The output layer uses the softmax function to normalize the output of the final fully-connected layer to a probability distribution. To be specific, this transform can be described as Equation 2:

$$S_j = \frac{e^{a_j}}{\sum_{k=1}^{T} e^{a_k}} \quad (2)$$

Here, a_j is the j^{th} value of the output vector of the final fully-connected layer, T represents the total number of classes, S represents the values from the neurons of the output layer. Thus, when a vector of K real values passes through the softmax function, a probability vector, whose K elements range between 0 and 1 and sum up to 1, can be obtained.

Taining of 1D-CNN:

The training of the 1D-CNN contains two processes: forward propagation and backward propagation. The predicted value can be acquired thought forward propagation processing shown in Figure 3. At the same time, the loss between the true value and the predicted value is calculated by the loss function. Based on the loss, the optimization function will update the weights and biases of each layer according to the chain rule. The updated parameters continue to participate in the training iteratively, and the training is finished until the loss function is reaching a global minimum. The overall training process is shown in Schema S2.



Schema S2 The training process of CNN.

Table S1 Summary of the water quality parameters of the industrial wastewaters

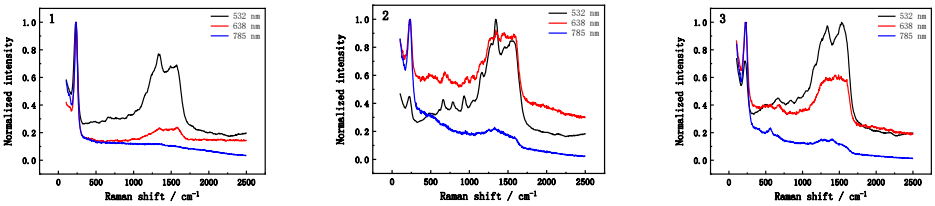
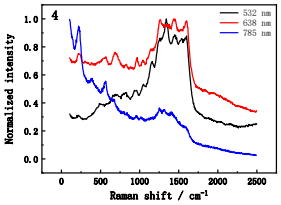
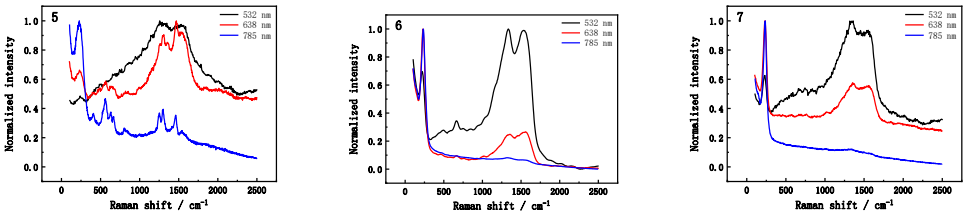
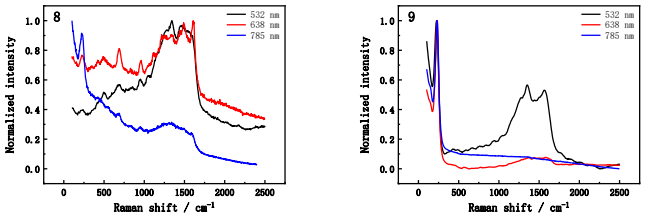
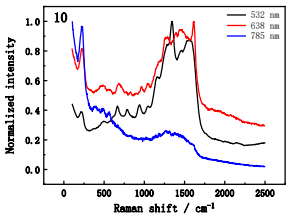
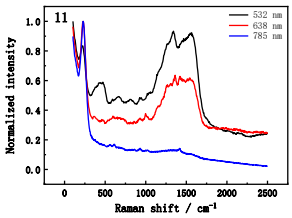
Industrial category	Sample No.	pH	SS (mg/L)	BOD (mg/L)	NH_3-N(mg/L)	COD_{Cr} (mg/L)
R	1	7.31	16	21.8	14.8	77
	2	6.56	47	135	30.4	246
	3	7.18	52	62.3	26.2	248
Pa	4	7.61	32	78.6	42.5	252
F	5	5.1	160	188	73	886
	6	6.52	74	102	16.3	247
	7	6.45	60	98	12.7	259
Pe	8	7.03	84	165	40.9	259
	9	7.12	92	138	36.3	184
T	10	7.02	30	90.8	43.9	148
Fu	11	7.19	38	50.1	21.8	154
Ca	12	6.84	36	5.6	1.49	73
	13	6.74	43	136	13.8	462
	14	7.26	96	134	42.6	237
Ph	15	2.74	40	7.8	11	80
D	16	7.04	43	91.3	39.4	294
	17	7.1	13	3.4	16.4	26
	18	6.24	122	144	11.7	437
S	19	6.92	52	95.1	65.8	351
	20	7.05	50	130	44.3	348
	21	6.87	66	96.2	39.2	308
	22	6.74	43	85.2	106	379

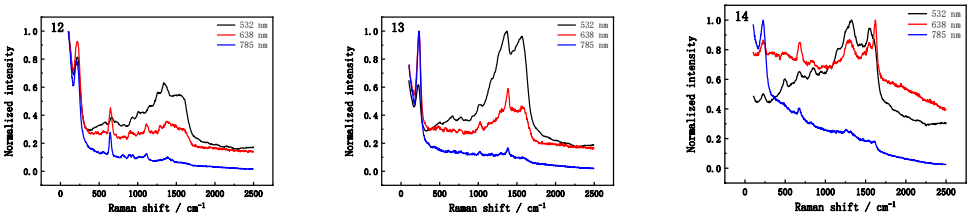
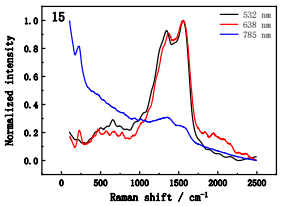
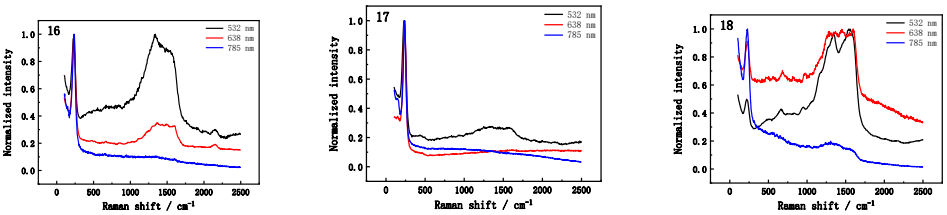
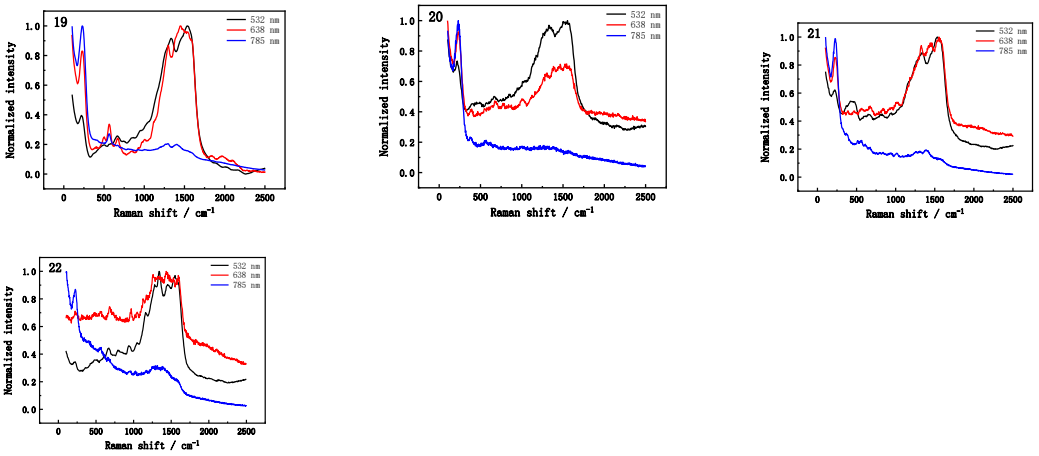
R: Resin and plastic; Pa: Papermaking; F: Food; Pe: Petrochemical; T: Textile dying; Fu: Furniture; Ca: Car manufacture; Ph: Pharmaceutical; D: Domestic sewage; S: Steel.

Table S2. Variance Analysis: Variables and Results

	pH	SS (mg/L)	BOD (mg/L)	NH_3-N(mg/L)	COD_{Cr} (mg/L)
Mean	6.67	58.59	93.55	34.11	270.7
Mean Sq	2.127	1117	2461	645.1	25137
Statistic F	11.25	0.869	0.856	1.186	0.663
P-value	0.000138	0.575	0.584	0.383	0.728

Table S3. SERS spectra for each industrial category

Industrial category[a]	SERS
R	
Pa	
F	
Pe	
T	
Fu	

Ca	
Ph	
D	
S	

[a] Refer to Table S1 for full names of industrial categories

Table S4 Summary of major SERS vibration bands[a]

Raman shift (cm^{-1})	Vibration[b]	Possible functional groups or compounds[b]
305-480	skeletal modes CC, CCC ring def	Methyl and methylene groups, lipids, proteins, carbohydrates
530-580	C-O-C glycosidic ring def; S-S str	Polysaccharides, ethers, proteins
720-735	COO^- def; CH_2 rock	Carboxylates, methylene groups, adenine
760	ring breath Trp	tryptophan
828	ring breath Tyr; O-P-O str	Tyrosine, phosphate
855-899	C-C str, C-O-C 1,4-glycosidic link	Alkanes, ethers
1000-1020	$C\text{-}CH_3$ def	Methyl groups, alkanes
1025-1060	C-O, C-C C-N str	Carbonate, amide, alkanes
1080-1085	C-O-H bend, CH bend	Hydroxyl and carboxyl groups
1090-1095	C-C str, C-O-C glycosidic link; ring breath, sym; PO^{2-} str, sym	Polysaccharides, aromatic hydrocarbon, phosphate,
1135-1145	NH_2 twist	Amide
1155	C-N str,	Amides, adenine
1230-1295	NH_2	Amide III
1320-1340	Aromatic ring	Aromatic modes, adenine,
1440-1460	CH_2 def.	Lipids, proteins, carbohydrates
1520-1550	C–H bend or C=C str	Olefinic
1575-1590	C–N str	Proteins, amides
1620	C=C str	Olefinic
2800-3100	CH_3, CH_2 and $=CH_2$ str	Carbohydrates, lipids, proteins, olefins

[a] bend, bending; breath, breathing; def, deformation; rock, rocking; str, stretching.

[b]References: Chisanga M et al. *Applied spectroscopy* 2018, 72(7): 987-1000; Ivleva N P et al. *The journal of physical chemistry B* 2010, 114(31): 10184-10194; Yang L et al. *Chemosphere* 2015, 127: 222-228.

Differentiation of the wastewaters with PCA

Dimensionality reduction was implemented by the PCA to obtain more visualized differences among the spectra from different effluents. Figure S1 shows the PCA results of SERS spectra from 22 wastewaters at different excitation wavelengths. The first two principal components of the recorded SERS spectra were plotted to visualize the characteristic distribution of the wastewater samples. Each wastewater data cluster containing 29 spectra is encapsulated within a 95% confidence ellipse. The distribution at the excitation wavelength of 638 nm is the most scattered, indicating that the spectra at 638 nm contain more characteristic information, followed by 532 nm, and 785 nm is the most barren. However, there are still many overlapping areas of the PCA distribution even at 638 nm, which shows that the classical PCA method is inefficient to realize the discrimination of different wastewaters.

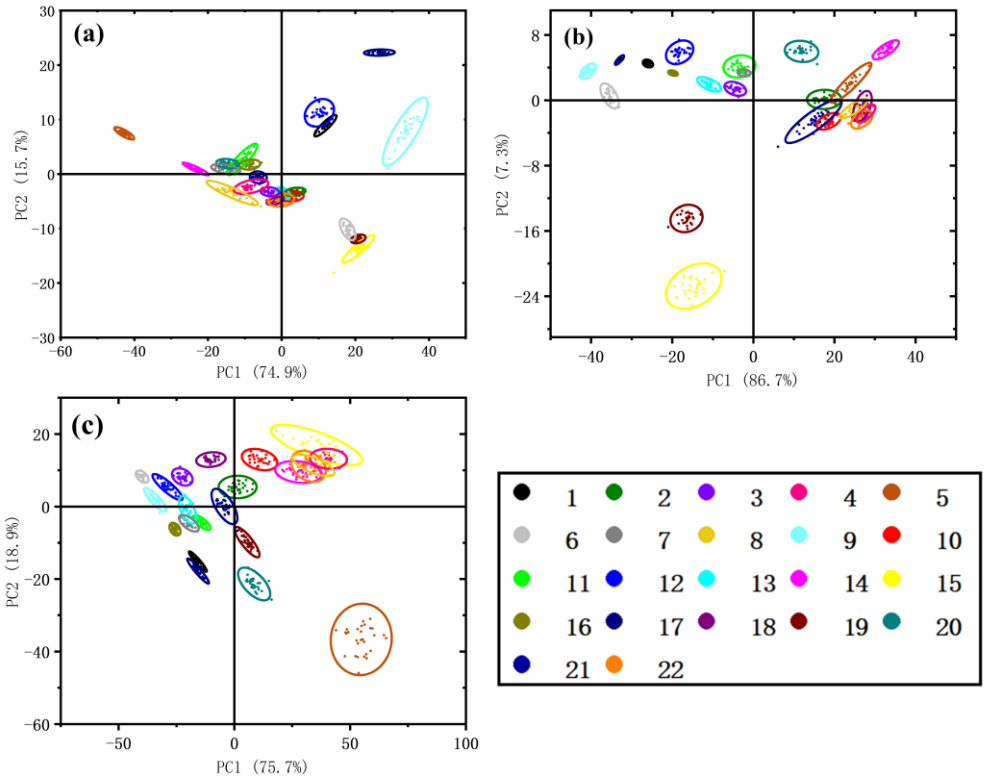
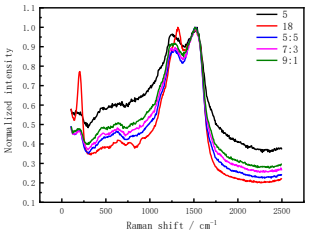
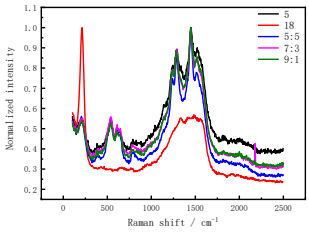
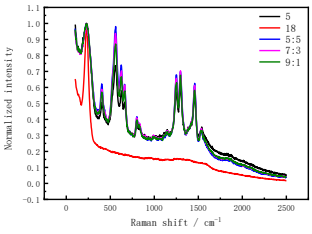
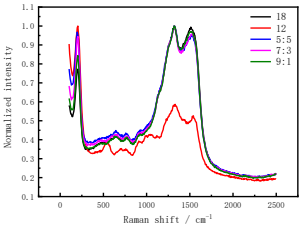
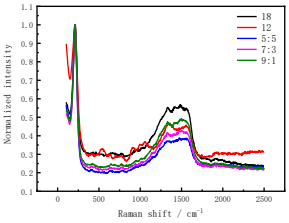
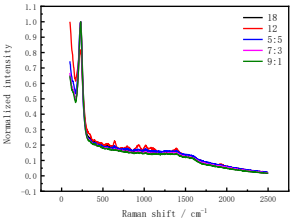


Figure S1 Scatter plot of the first principal component (PC1) versus the second

principal component (PC2) from SERS spectra of the wastewaters at different laser wavelengths. (a) 532 nm; (b) 638 nm; (c) 785 nm.

In order to illustrate the superiority of SERS spectra in information capture of DOM, the mixed wastewater samples with different blending ratios were analyzed by SERS. Sample 5 and 18 were mixed with a volume ratio of 5:5, 7:3 and 9:1. The same mixing process was implemented to sample 18 and 12. The SERS spectra of the mixed samples are shown in Table S5. The PCA analysis was conducted to the SERS spectra, showing a trend that the data point of the mixed wastewater sample will gradually move closer to that of the dominant wastewater sample as the proportion of it increases (Figure S2). The three-dimensional PCA analysis of mixed wastewaters are shown in Figure S3 (sample 5/18) and Figure S4 (sample 18/12). From the PCA loadings included in Figure S3 and Figure S4, we can see that the discrimination mainly takes place along the PC 1 and PC 2 axis and the amount of characteristic information that can be captured at different excitation wavelengths varies greatly. Even in the case of the barren spectral feature information (Figure S4), the composition variations reflected in gradually approaching trend can still be observed. The results suggest that the SERS spectra can indeed capture subtle changes in wastewater composition, exhibiting the potential to discriminate different wastewaters and trace the pollution source. However, the trend shown by the PCA cannot give specific and valuable identification results, so the more advanced CNN algorithm was introduced.

Table S5 The SERS spectra of the mixed water samples with different excitation wavelength

	532 nm	638 nm	785 nm
5/18			
18/12			

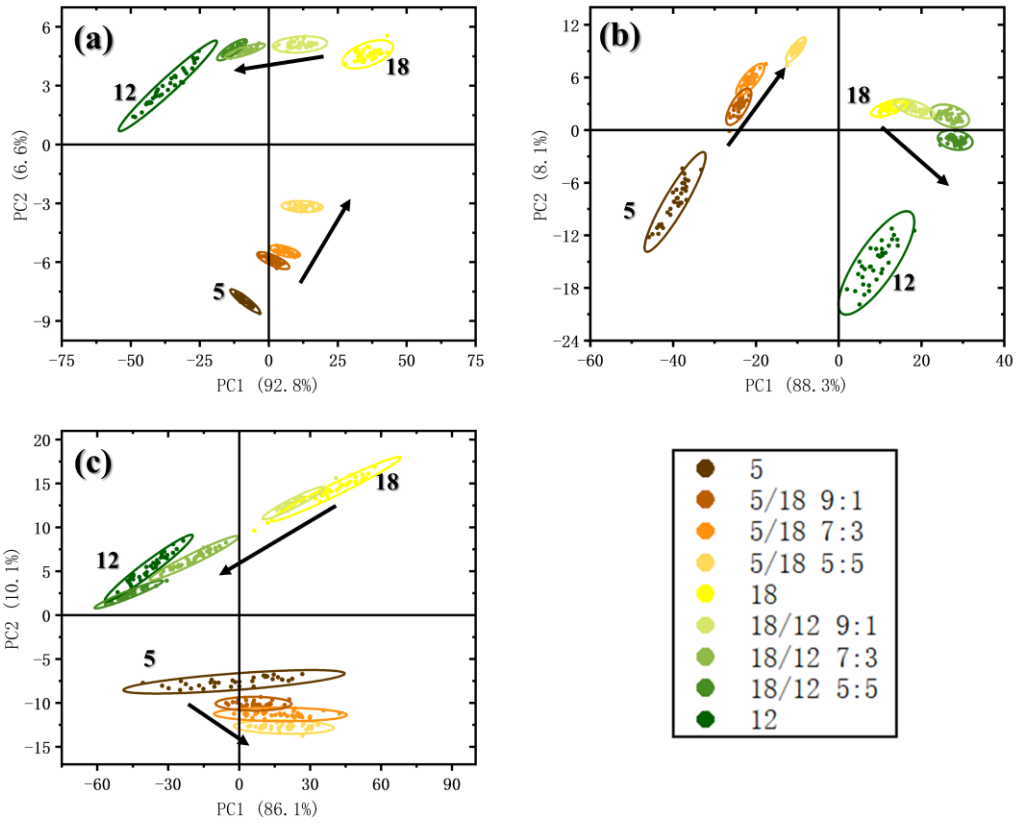


Figure S2 PCA results of mixed wastewaters (sample 5/18 and 18/12) at different excitation wavelengths. (a) 532 nm; (b) 638 nm; (c) 785 nm.

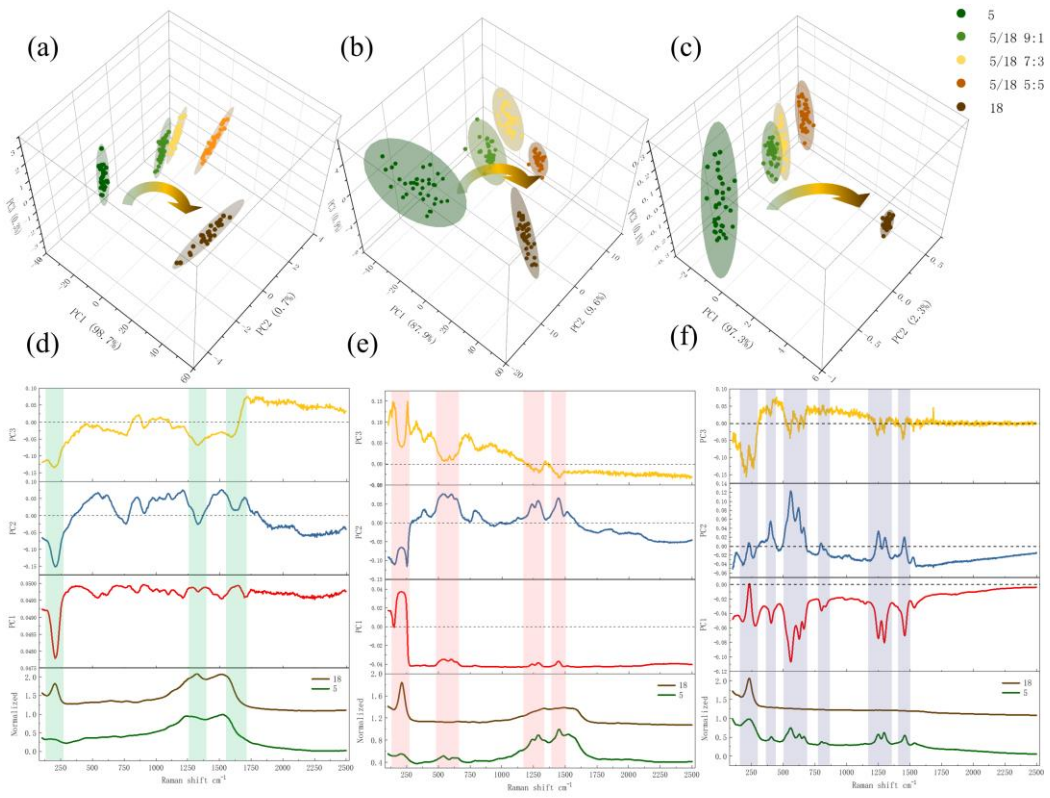


Figure S3 Three-dimensional PCA results of mixed wastewaters (sample 5/18) at (a) 532 nm, (b) 638 nm and (c) 785 nm; PC1, PC2 and PC3 loading data (top) and the normalized SERS spectra of wastewater sample 5 and 18 (bottom) at (d) 532 nm, (e) 638 nm and (f) 785 nm.

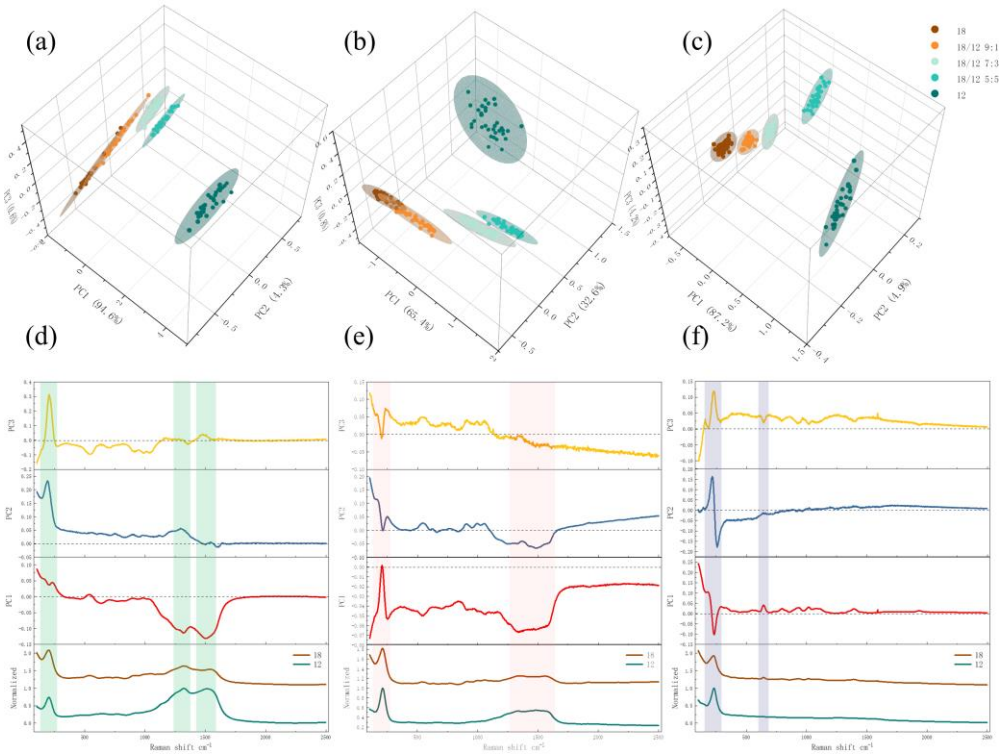


Figure S4 Three-dimensional PCA results of mixed wastewaters (sample 18/12) at (a) 532 nm, (b) 638 nm and (c) 785 nm; PC1, PC2 and PC3 loading data and the normalized SERS spectra of wastewater sample 18 and 12 at (d) 532 nm, (e) 638 nm and (f) 785 nm.

Table S6 The hyperparameters used for training 1D-CNN model

Hyperparameters	
Batch size	4
Epochs	60
Momentum	0.9
Learning rate	0.005

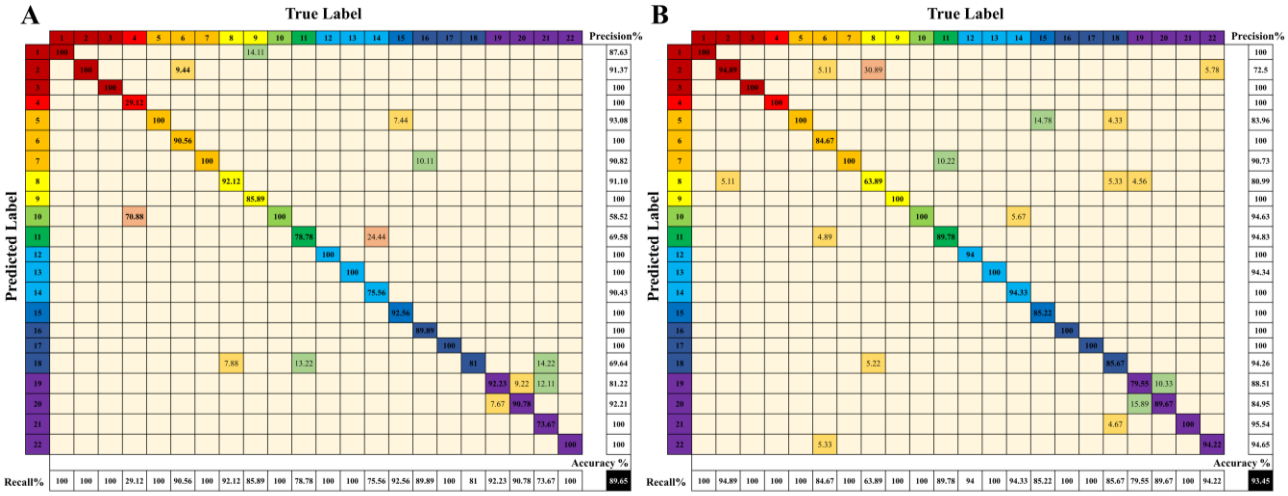


Figure S5 Confusion matrix for the prediction of 22 wastewaters based on the SERS spectra at 638 nm. (A) RF and (B) SVM.

Table S7 The mixed arrays of the mixed wastewater samples

Mixed type	Sample combination
Two-mixed	1/4, 1/8, 1/14, 1/15, 1/18, 1/19, 1/22, 9/4, 9/8, 9/14, 9/15, 9/18, 9/19, 9/22, 17/4, 17/8, 17/14, 17/15, 17/18, 17/19, 17/22
Three-mixed	1/4/8, 1/14/15, 1/18/19, 1/4/22, 9/4/8, 9/14/15, 9/18/19, 9/4/22, 17/4/8, 17/14/15, 17/18/19, 17/4/22, 8/15/19
Four-mixed	1/4/8/14, 1/15/18/19, 1/22/14/18, 9/4/8/14, 9/15/18/19, 9/22/14/18, 17/4/8/14, 17/15/18/19, 17/22/14/18
Five-mixed	1/4/8/14/15, 9/4/8/14/15, 17/4/8/14/15, 14/15/18/19/22,
Six-mixed	1/4/8/14/15/18, 1/9/17/18/19/22, 1/9/17/8/14/15