Spectrochimica Acta Part A: Molecular and Biomolecular Spectroscopy 332 (2025) 125806

Contents lists available at ScienceDirect

Spectrochimica Acta Part A: Molecular and Biomolecular Spectroscopy

journal homepage: www.journals.elsevier.com/spectrochimica-acta-part-a-molecular-and-biomolecular-spectroscopy

Machine learning-integrated surface-enhanced Raman spectroscopy analysis of multicomponent dye mixtures

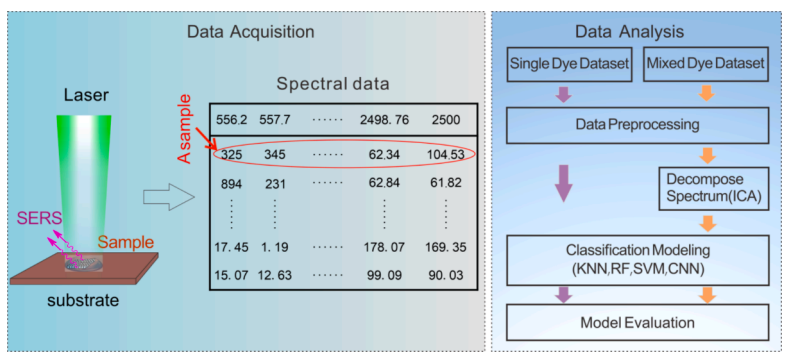
Yan Yu [a,b], Wenjing Lu [b], Xiaobin Yao [c,*], Yurui Jiang [b], Junhui Li [b,d], Meng Yang [b], Xingjiu Huang [b], Xianghu Tang [a,b,*]

[a] School of Energy Materials and Chemical Engineering, Hefei University, Hefei 230601, China
[b] Institute of Solid State Physics, HFIPS, Chinese Academy of Sciences, Hefei 230031, China
[c] Department of Chemistry, The Hong Kong University of Science and Technology, Clear Water Bay, Hong Kong, China
[d] Anhui Sci-rule Analysis and Studying Technology Co., Ltd., Hefei 230088, China

HIGHLIGHTS

- Accurate identification of multicomponent dye mixtures using machine learning-assisted SERS analysis.
- Qualitative analysis of single dyes using SERS technique combined with classification models (KNN,SVM,RF,CNN).
- Qualitative analysis of multi-component dyes using ICA signal separation combined with classification models.
- Signal separation combined with classification modeling has great potential for SERS in the qualitative analysis of complex systems.

GRAPHICAL ABSTRACT



ARTICLE INFO

Keywords:
surface-enhanced Raman spectroscopy (SERS)
Machine learning
Dye mixtures
independent principal component analysis (ICA)
convolutional neural network (CNN)

ABSTRACT

Surface-enhanced Raman spectroscopy (SERS) is broadly used in the detection and analysis of materials with its fingerprint-like specificity and high sensitivity. However, resembling signals of analytes highly affect the identification and assignment of spectra, which has become a long-term issue to be solved. In this study, various models of machine learning are utilized and compared to support data analysis of complex SERS spectra. Silver-coated gold core–shell nanocubes (Au@AgNCs) are optimized as SERS substrates for the detection of four common dyes – methylene blue (MB), crystal violet (CV), rhodamine B (RhB) and malachite green (MG). Independent principal component analysis (ICA) was utilized to isolate the signals from the SERS spectra of the dye mixtures, and the isolated signals were further classified by commonly used classification models including K Nearest Neighbors (KNN), Support Vector Machines (SVM), Random Forests (RF), and Convolutional Neural Networks (CNN). The results show that the CNN model achieved an accuracy of 98% in the classification of single dyes and an accuracy of 97% in the classification of dye mixtures, which is significantly better than other models. Based on these findings, we propose ICA combined with CNN-assisted SERS spectroscopy as an effective analytical tool for analyzing dye mixtures.

* Corresponding authors.
E-mail addresses: xiaobinyao@ust.hk (X. Yao), tangxh2011@iim.ac.cn (X. Tang).

https://doi.org/10.1016/j.saa.2025.125806
Received 6 December 2024; Received in revised form 20 January 2025; Accepted 24 January 2025
Available online 26 January 2025
1386-1425/© 2025 Elsevier B.V. All rights are reserved, including those for text and data mining, AI training, and similar technologies.

1. Introduction

Nowadays, industrial dyes play a significant role in many industries such as textiles, food and cosmetics while the overuse of dyes lead to water pollution, ecosystem imbalance, and especially potential harm to human health[1,2]. Vanhinsbergh *et al.* reported a case of severe neonatal hemolytic anemia (Heinz corbosome hemolytic anemia) caused by the excessive use of methylene blue (MB)[3]. Mani *et al.* studied crystal violet (CV) and found that it can cause a variety of health problems including skin and eye irritation, digestive problems, liver and kidney damage, and promote tumor growth in some fish, posing a serious threat to the environment[4]. Li *et al.* showed that malachite green (MG) exposure would reduce fish intestinal microbial diversity, increase intestinal permeability, trigger immune response, destroy microbial community structure and stability, increase the risk of infection, and pose a threat to the environment and fish health[5]. Moreover, dyes are often used as mixtures which further deteriorate the situation. Therefore, efficient detection methods are necessary and critical in the precaution and supervision of dye uses. The common dye detection methods mainly include high performance liquid chromatography (HPLC), ultraviolet–visible spectrometer, thin layer chromatography (TLC), mass spectrometry (MS) and other detection methods[6–8], but these methods have limitations in separation ability, poor selectivity, low resolution and high cost, which limit their applications in the analysis of complex mixtures.

In recent years, surface-enhanced Raman spectroscopy (SERS), attracts extensive attentions with its high sensitivity, non-destructive and unique fingerprint-like specificity[9,10]. As an advanced Raman technique, SERS spectroscopy integrates nanomaterials with Raman spectroscopy to overcome the issue of low Raman cross section of molecules [11,12]. By utilizing the localized surface plasmon oscillation of nanoparticles, the detection limit is lift up to the single-molecule level which makes SERS applicable in trace analysis and practical applications[13]. Notably, interfered by resembling signals, direct assignments of SERS spectra are insufficient to distinguish similar analytes while the identification of multicomponent analytes is highly needed in practical applications. Benefiting with the development of machine learning, this issue is gradually overcome by the combination of SERS detection with data analysis methods and the accuracy and reliability of spectral assignment in SERS are apparently improved.

Fang *et al.* prepared TLC-AG substrate by assembling single-layer silver nanoparticles on TLC plate and obtained spectral data of all target analytes through one-dimensional and scanning. Convolutional neural network (CNN) and spectral angular similarity (SAS) machine learning algorithm were used to distinguish the spectral information of mixed pesticides[14]. Weatherston *et al.* prepared a novel "nanopaper" substrate on glass microfiber filter paper through a silver mirror reaction, where they used as a paper chromatography substrate (PC-SERS) for separating and identifying mixed organic dyes[15]. Lu *et al.* used a TLC plate composed of diatomaceous earth as the stationary phase, and directly added real beef juice drops onto the diatomaceous earth plate for separation and detection. Through principal component analysis and backpropagation neural network (PCA-BPNN) model, the concentration of ofloxin was quantitatively predicted[16]. The strategy used by the above-mentioned works is based on the procedure of analyte separation by experiments, and then product identification.

In our opinion, the step of analyte separation using chromatography experiments in SERS-based research is unnecessary and time-consuming. Instead, machine learning is powerful and sensitive enough to separate the spectral data of mixtures. For example, Bajomo *et al.* used combined the characteristic peak extraction (CaPE) a data compression algorithm in combination with signal separation techniques (ICA, NICA, SICA, NSNMF, NMF) to achieve good separation results by deconvolution of mixed SERS spectra of polycyclic aromatic

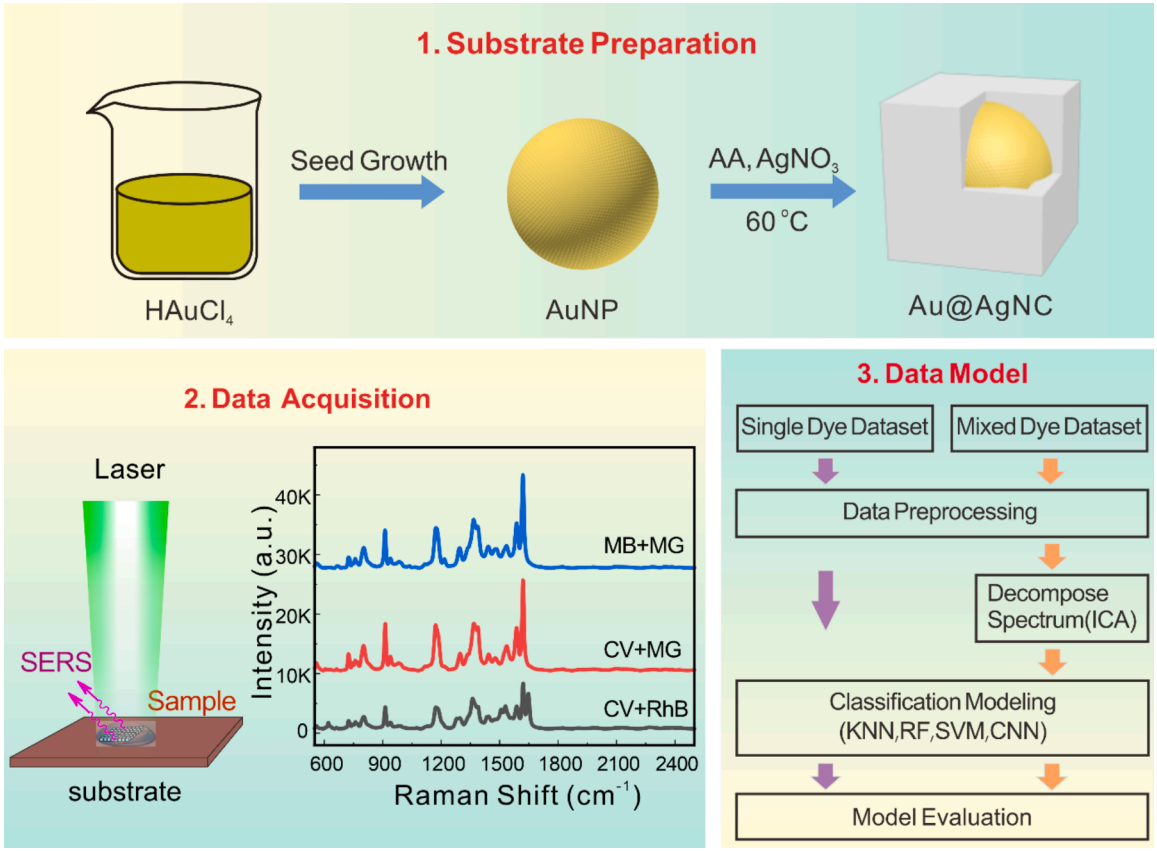


Fig. 1. Illustrated scheme for the detection and identification of dyes using Au@AgNCs SERS substrates assisted with machine learning models.

hydrocarbons (PAHs)[17]. Wang *et al.* innovatively combined three SERS substrates (AgC NPs, Ag_7Au_3 NPs, AgA NPs) with the 1D-CNN model for the identification of roasted tobacco varieties, the origins of green tea and purple rice, and the adulteration of agricultural products, and the experimental results show that the method performs well in all kinds of identification tasks, with a correct identification rate of at least 94.8 %[18]. Fang et al. synthesized Au@Ag-Au PPFs substrates through a multi-step process and functionalized them with L-cysteine. The modified Au@Ag-Au PPFs were then assembled onto filter paper to construct a mask-type SERS chip. Using a CNN algorithm, highly sensitive and specific quantitative detection of gaseous benzaldehyde was achieved, which providing a potential effective tool for the early diagnosis of lung cancer[19]. Luo et al. combined SERS with CNN technology to achieve qualitative analysis of six common microplastics (PE, PC, PS, PVC, PP, PET), achieving an accuracy of up to 99.54 %. This strategy provides a new method and approach for microplastic detection[20]. The above shows that it is feasible to introduce machine learning into the data analysis of SERS spectra. In the qualitative analysis of multi-component SERS spectra, it is believed that the difficulty in characterizing multi-components due to the spectral cross-interference can be addressed by solving the problem of separation of the mixed signals and the identification/classification of the separated spectra. Currently, the use of signal separation techniques combined with machine learning classification models is yet to be systematically explored.

In this study, Independent Component Analysis (ICA) is adopted as the signal separation technique. Compared to Principal Component Analysis (PCA), which focuses on directions of maximum variance, ICA yields components with clearer physical significance. In contrast to Non-negative Matrix Factorization (NMF), which assumes non-negativity, ICA can handle signals containing negative values and is insensitive to changes in signal amplitude and scale. The inherent characteristic of ICA, based on the assumption of statistical independence of signals, makes it particularly suitable for separating spectral signals composed of different substances stacked together[21–23]. For the classification of spectra, we employ four classification models: K-Nearest Neighbors (KNN), Support Vector Machine (SVM), Random Forest (RF), and Convolutional Neural Network (CNN). Overall, this research aims to address the complex challenge of qualitative analysis of multi-component dyes through an innovative integration of signal processing techniques and classification models. As shown in Fig. 1, silver coated gold core–shell nanocubes (Au@AgNCs) were prepared and optimized as the SERS substrate. Four dyes MB, CV, RhB and MG and their mixtures were selected as model molecules. The spectra datasets of dyes were preprocessed with smoothing, noise reduction and baseline deduction. Compared to the datasets of single dyes, ICA was applied to decompose and extract the spectral information of dye mixtures before they were imported into the trained classification models for signal identification. Finally, to confirm the validity of spectral information extraction and classification model construction, the confusion matrix of each machine learning model was used for evaluation.

2. Materials and methods

2.1. Chemicals

Chloroauric acid ($HAuCl_4{\cdot}4H_2O$), cetyltrimethylammonium chloride (CTAC), sodium hypochlorite (NaClO), crystal violet (CV), methylene blue (MB), malachite green (MG) and Rhodamine B (RhB) were purchased from Sinopod Chemical Reagents (Shanghai) Co., LTD. Silver nitrate ($AgNO_3$) and ascorbic acid (AA) were purchased from Shanghai Aladdin Reagent Co., Ltd., China, and sodium borohydride ($NaBH_4$) and p-nitrothiophenol (p-NTP) were purchased from Sigma-Aldrich China. All chemicals were used directly without pretreatment, and 18.25 MΩ·cm of Milli-Q deionized water was used.

2.2. Instrumentations

Ultraviolet–visible spectrophotometer (UV-2550, Shimadzu, Japan) was used to check the absorption wavelengths of nanocubes. Transmission electron microscope (JEM-2100, Nippon Electronics Corporation) and scanning electron microscope (JSM-6700F, Nippon Electronics Corporation) were used to measure the morphologies and structures. Confocal micro-Raman spectrometer with a 532 nm laser (LabRAM-HR800, HORIBA-JobinYvon, France) was used to acquire all Raman data.

2.3. Fabrication of Au@AgNCs

Referred to the reference[24], Au@AgNCs with different sizes were fabricated.1) Preparation of gold nanoparticles (AuNPs): Briefly, AuNPs were prepared by a common seed growth method[25]. At room temperature, 206 μL of 0.025 M $HAuCl_4$ was mixed with10 mL of 0.1 M CTAC solution and stirred for 10 min, and then 400 μL of 0.02 M $NaBH_4$ solution was added and stirred vigorously for 5 min to form the nuclei of gold nanospheres. After standing for 10 min, 5 mL of the solution was taken and added to 45 mL of 0.1 M CTAC solution, mix well and left for 2 h to promote the growth of gold nanospheres. Subsequently, 3.6 mL of the solution was added into 10 mL of 0.1 M CTAC and 30 mL of water which was slightly stirred for 10 min, and then 160 μL of 0.1 M AA solution was added and stirred for 5 min, and finally 410 μL of 0.025 M $HAuCl_4$ solution was added and stirred for 5 min which was left for 1 h to complete the preparation of AuNPs. 2) Synthesis of Au@AgNCs: 1 mL of AuNPs was added into 4 mL 0.02 M CTAC solution and the mixture was stirred at 60 ℃ for 30 min. To adjust the thickness of the Ag shell, different volumes of silver nitrate (0.25 mL, 0.5 mL, 1 mL, 2 mL and 4 mL) were explored. An injection pump was used to simultaneously inject the same volume of 0.002 M silver nitrate, 0.05 M ascorbic acid and 0.04 M CTAC solution at a rate of 0.2 mL/min. The color of the solutions changed from red to brownish yellow during injection. The solutions were stirred for 4 h continuously and then cooled in ice baths to obtain Au@AgNCs sols for following measurements.

2.4. Preparation of dye mixtures

$1x10^{-3}$ M stock solutions of CV, RhB, MB and MG were first prepared with ethanol, and then diluted to different concentrations with Milli-Q water. Three different dye mixtures were prepared, including CV + RhB, CV + MB and MB + MG. For example, 2 mL CV and 2 mL RhB with different concentrations (10^{-5} M, 10^{-6} M, 10^{-7} M, 10^{-8} M) were added and thoroughly mixed in a 15 mL centrifuge tube for following SERS measurements. The same procedure was applied to other two mixtures.

2.5. SERS experiments

1 mL of Au@AgNCs sol was centrifugally rinsed twice and concentrated to 100 μL and then mixed up with 100 μL dye stock solution by ultrasonic dispersion which was incubated at room temperature for 10 min and then dropped on silicon wafers to dry naturally. For each sample, SERS spectra were acquired from 10 random sites at a laser power of 2 mW and an integration time of 3 s. Totally, 330 SERS spectra were collected, including 210 and 120 spectra of single and mixed dyes, respectively

2.6. Signal separation and machine learning models

In this study, independent principal component analysis (ICA) is used for signal separation, and four machine-learning classification models such as K Nearest Neighbors (KNN), Support Vector Machines (SVM), Random Forests (RF), and Convolutional Neural Networks (CNN) are used for classification. In addition, in order to prevent the effects of factors such as noise and baseline drift, the collected data are

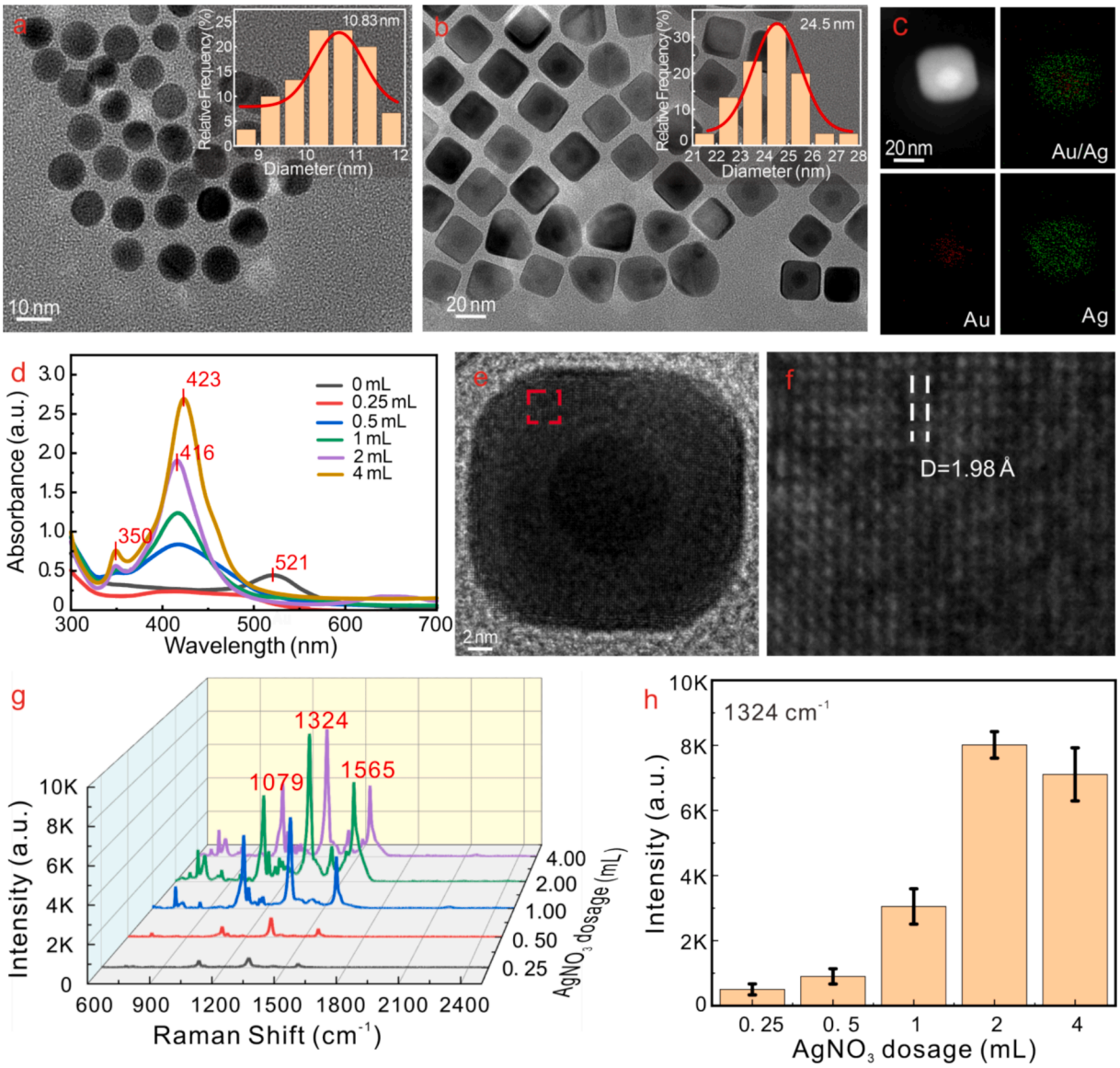


Fig. 2. A) TEM image of AuNPs. b) TEM image of Au@AgNCs. c) TEM image and EDS mapping. d) UV–vis spectra of different Au@AgNCs. e-f) HRTEM image of Au@AgNCs with 7 nm Ag shell. g) SERS spectra of p-NTP detected from different Au@AgNCs substrates. h) Intensity diagram of p-NTP at 1324 cm^{-1} acquired from different Au@AgNCs regulated by different amounts of $AgNO_3$.

preprocessed before performing the classification modeling; noise removal is performed using a nine-point third-order Savitzky-Golay filter[26], baseline correction is performed using adaptive iterative reweighted penalized least squares (airPLS)[27], and dimensionality reduction of Raman data is performed using the ReliefF feature selection algorithm. Independent component analysis (ICA) is a statistical technique used to separate statistically independent non-Gaussian source signals from multivariate observational data. It assumes that the observed data is composed of multiple independent source signals through an unknown linear mixing process. Thus, the original independent source signal can be recovered from the observed data[28,29]. KNN is an intuitive and easy to implement supervised learning method, widely used in classification and regression tasks, but its performance largely depends on the balance of training data, using grid search and other technologies to find the optimal K value[30,31]. SVM (Support Vector Machine) is an efficient supervised learning method, mainly used for binary classification. It constructs decision boundaries through key data points (support vectors) to achieve high accuracy, but performance deteriorates when the feature dimension is much larger than the number of samples, and it is sensitive to missing values[32,33]. RF is a supervised learning method that performs classification or regression by constructing multiple decision trees and synthesizing their results[34]. It not only has high accuracy and good resistance to overfitting, but also can handle high-dimensional data, recognize important features, and is robust to noise and missing values in the data[35]. CNN is a deep learning method designed for processing data with grid structure (such as images), which automatically extracts multi-level feature representations through multi-layer convolutional layers[36]. CNN has shown excellent performance in image recognition, classification, object detection and other tasks, and can effectively capture spatial hierarchy and local patterns, while having a certain degree of translation invariance[37,38]. In addition, CNN are able to reduce the number of parameters, thereby reducing the risk of overfitting and improving the generalization ability of the model. In this study, we construct a two-layer one-dimensional convolutional network in the CNN model, and use ReLU and Softmax as activation functions, Adam algorithm as optimization algorithm, and “Dropout” as regularization parameters to reduce the risk of overfitting in the model. In the aspect of support vector

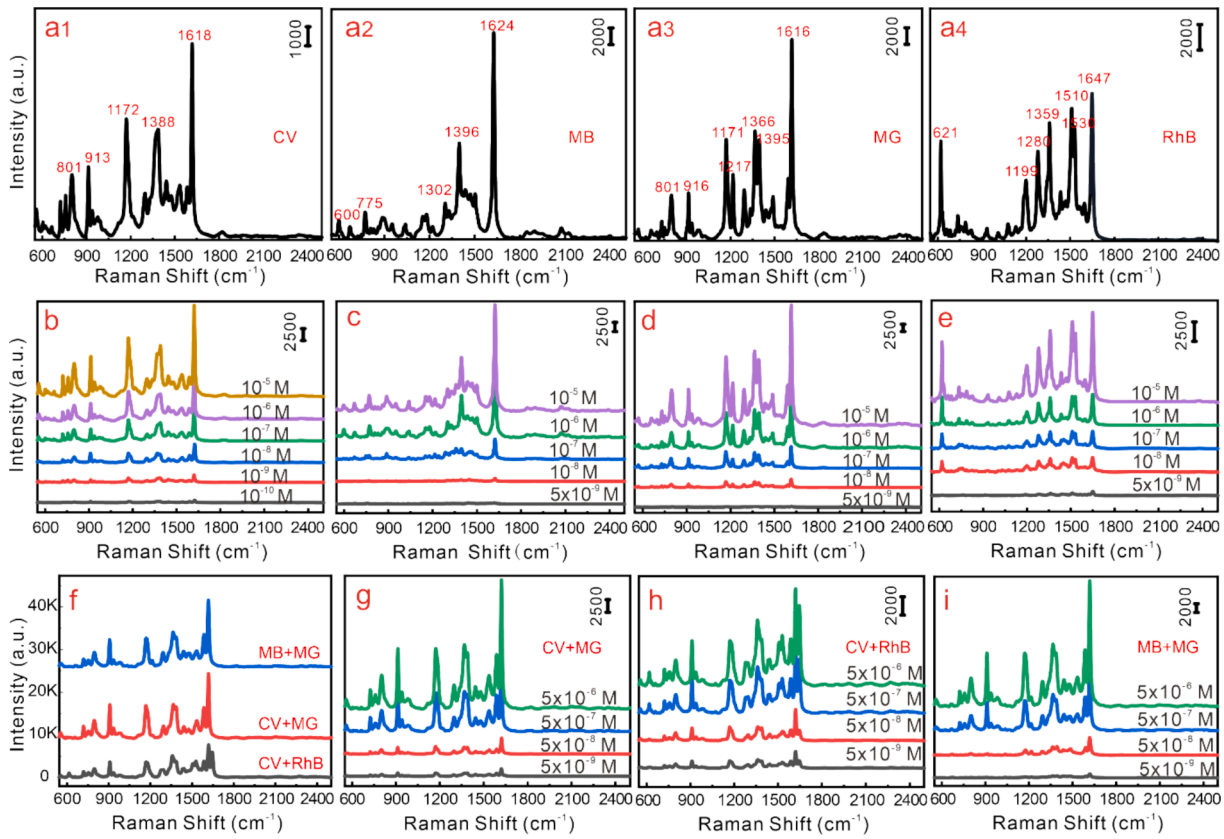


Fig. 3. a1-a4) SERS spectra of CV, MB, MG and RhB. b-e) SERS spectra of CV, MB, MG and RhB at different concentrations. f) SERS spectra of MB + MG, CV + MG and CV + RhB mixtures. g-i) SERS spectra of MB + MG, CV + MG and CV + RhB mixtures at different concentrations.

machine, radial basis function with better results is used for classification. In the random forest model, the optimal parameters were selected by the cross-validation method, where the number of decision trees was set to 1000 and the maximum depth of each tree was artificially limited to 4. All data processing and modeling in this study was done in Python-3.7.1 for Windows.

2.7. Assessment methods

When evaluating classification model, accuracy rate, recall rate and F1 score are the three core evaluation indicators[39,40]. Their mathematical expressions are shown as followed Equations (1)-(4), where true positive (TP) refers to the number of samples accurately predicted by the model as the target category, true negative (TN) refers to the number of samples accurately predicted by the model as the non-target category, and false positive (FP) indicates that the model incorrectly predicts the sample as the target category. False negatives (FN) represent samples of the target class that the model failed to recognize.

$$\text{Accuracy Rate}: \frac{TP+TN}{TP+TN+FP+FN}*100\% \quad (1)$$

$$\text{Recall Rate}: \frac{TP}{TP+FN}*100\% \quad (2)$$

$$\text{F1 score}: \frac{2TP}{2TP+FP+FN}*100\% \quad (3)$$

$$\text{Precision}: \frac{TP}{TP+FP}*100\% \quad (4)$$

3. Results and discussion

3.1. The structural characterization of Au@AgNCs

Fig. 2a shows AuNPs prepared by seed growth method, most of which are regular and uniform. The size distribution of AuNPs is shown in Fig. 2a, which has the average size of 10.83 ± 0.74 nm. In the TEM image (Fig. 2b), Au@AgNCs have cubic structures with spherical cores. The clear color contrast indicates a clear core–shell structure which is also confirmed by the TEM mapping in Fig. 2c where Au cores are coated by Ag shells. Fig. 2d is the UV–vis spectrum of these colloids, with the main absorption peaks of 350 nm, 416 nm and 521 nm. With the increase of $AgNO_3$, the ultraviolet absorption peak of Au nucleus at 521 nm was buried gradually and the absorption peak of Ag shell at 416 nm gradually increased and red-shifted, and the absorption peak at 350 nm also gradually increased, which is assumed to be caused by the growing of Ag shell. Fig. 2f is a zoom-in HRTEM image of the red-box area of Fig. 2e. The lattice spacing is 1.98 Å corresponding to the (200) crystal face of Ag. To obtain the best SERS effect, several sizes (14.9 ± 0.89, 17.3 ± 1.02, 20.7 ± 0.74, 24.5 ± 1.19, 28.8 ± 1.09 nm) of Au@AgNCs were synthesized by adjusting the amount of $AgNO_3$ (0.25 mL, 0.5 mL, 1 mL, 2 mL, 4 mL), whose morphologies and size distributions are shown in Fig. S1 a1-a5 and S1 b1-b5, respectively. With increasing the amount of $AgNO_3$, the sizes of particles gradually grew up and the structures changed from spheres to cubes. Fig. S1 c1-c5 are the corresponding EDS spectra of Au@AgNCs at each dose. With the increase of $AgNO_3$, the proportion of Ag element also increased from 71.93 % to 96.73 %. As a result, nanocubes with different shell thicknesses of Au@AgNCs were successfully prepared on the surface.

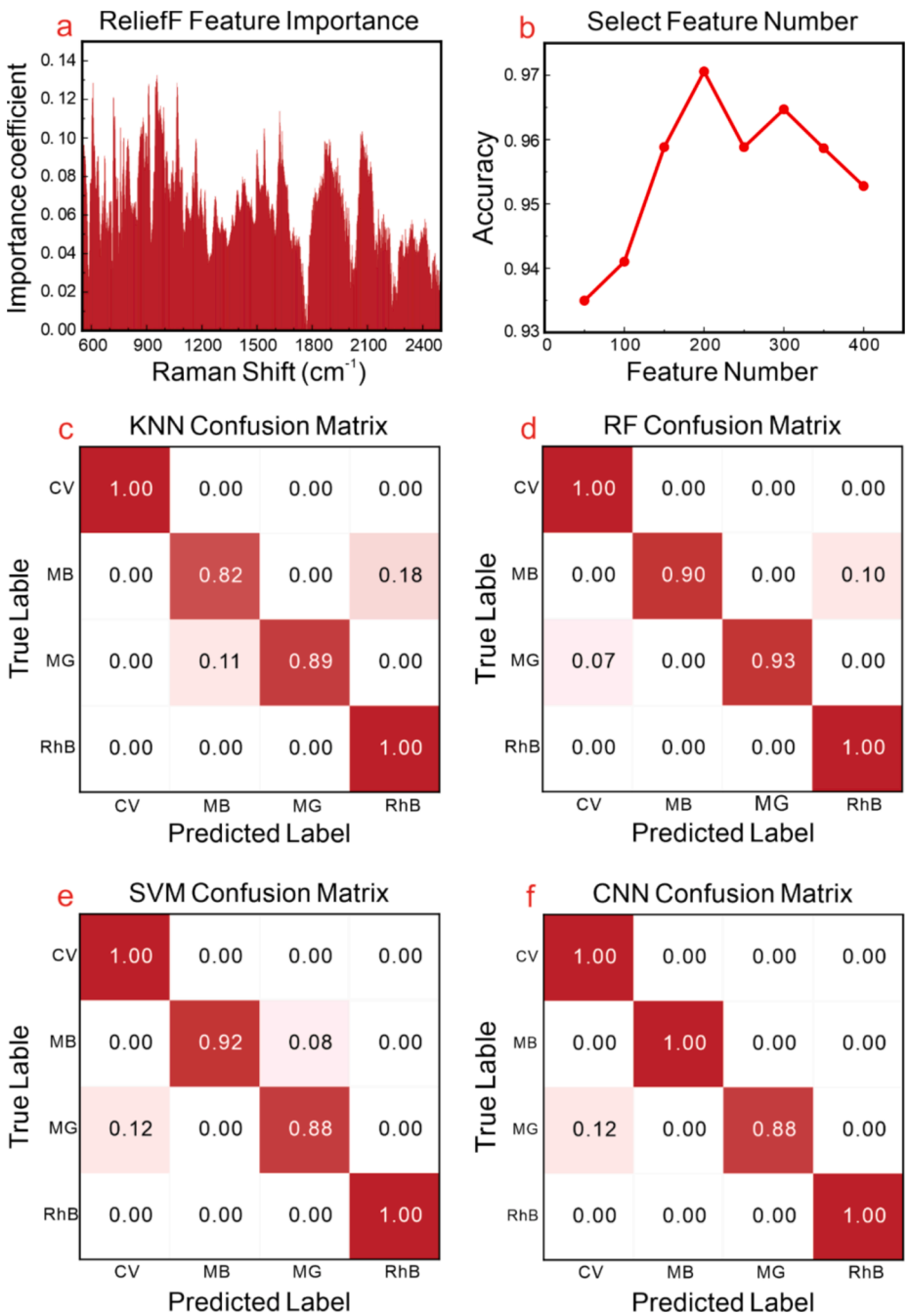


Fig. 4. A) importance coefficient of relief f algorithm. b) cross-validation to select the best number of features. c-f) the confusion matrices of knn, RF, SVM and CNN models for classifying single dyes, respectively.

3.2. SERS activity and uniformity of Au@AgNCs

Ag shell thicknesses of Au@AgNCs (2 nm, 3 nm, 5 nm, 7 nm, 9 nm) correspond to the different amounts of $AgNO_3$ (0.25 mL, 0.5 mL, 1 mL, 2 mL, 4 mL) are shown in Fig. S2 a1-a5. These Au@AgNCs were further prepared as SERS substrates for following experiments. For SERS substrates made by core–shell nanomaterials, it is known that the thickness of the shell can affect its SERS performance[41]. Therefore, to select the best Au@AgNCs substrate, p-NTP was used as the probe molecule. First, 1 mL Au@AgNCs sol was centrifugally rinsed twice and concentrated to 100 μL, and then ultrasonically mixed with 200 μL of $1x10^{-5}$ M p-NTP solution. Subsequently, 5 μL of the mixed solution was dropped onto a silicon chip and dried naturally for following SERS measurements. The SERS spectra of 10^{-5} M p-NTP collected from above-prepared five Au@AgNCs substrates and the corresponding intensity diagram extracted at 1324 cm^{-1} (N=O stretching) are shown in Fig. 2g and 2 h, respectively. Clearly, the Au@AgNCs with 7 nm Ag shell made from 2 mL $AgNO_3$ showed the strongest SERS signals. To check the reproducibility and reliability of the substrate, SERS spectra were acquired from 15 random points (Fig. S2 b). The intensities at 1079 cm^{-1} (ring breathing mode), 1324 cm^{-1} and 1565 cm^{-1} (ring stretching mode) showed relative standard deviations (RSD) of 11.83 %, 4.93 % and 11.85 %, respectively, which reflected the feature of a good SERS substrate (Fig. S2 c-e).

3.3. SERS spectra of dyes

SERS spectra of single dyes were collected separately, as shown in

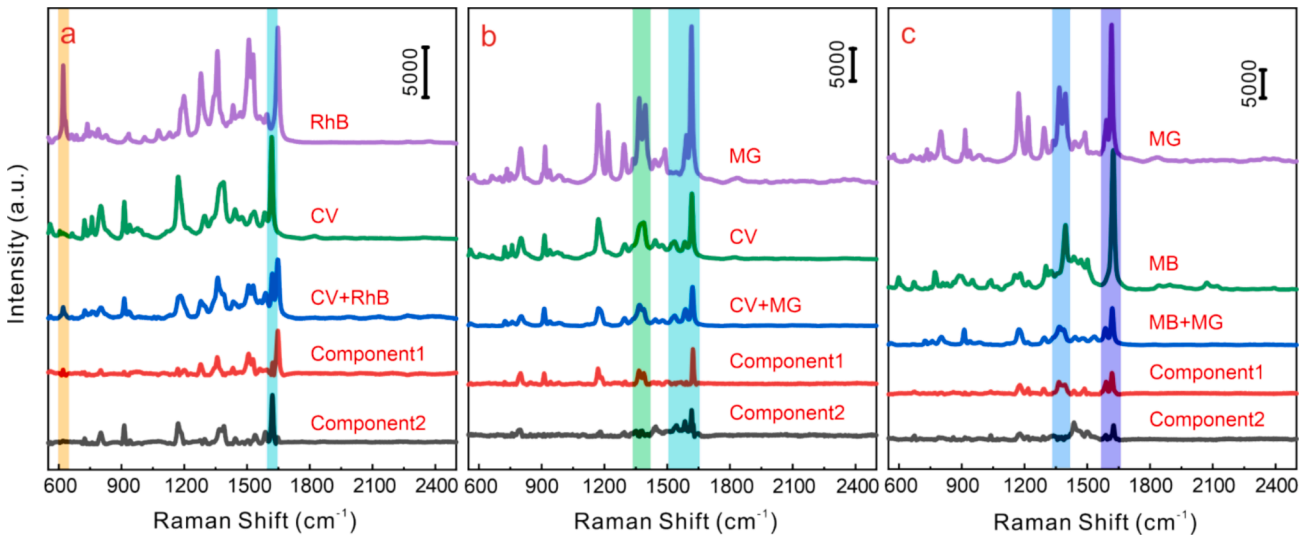


Fig. 5. Decomposed spectra of a) CV + RhB, b) CV + MG, and c) MB + MG.

Fig. 3 a1-a4, corresponding to CV, MB, MG and RhB respectively. It can be seen from Fig. 3a that the characteristic peaks of the four dyes were mainly located in the range of 550–1700 cm^{-1}. The main Raman characteristic peaks of CV are shown in Table S1, which are 801, 913, 1172, 1388, and 1618 cm^{-1}, respectively. These characteristic peaks are assigned as C–H bending vibration (801 cm^{-1}), C–C stretching vibration in aromatic ring (913 cm^{-1}), C-N stretching vibration (1172 cm^{-1}), C–H symmetrical bending vibration (1388 cm^{-1}), and C=C-stretching vibration (1618 cm^{-1})[42,43]. The main Raman characteristic peaks of MB are shown in Table S2, which are 600, 775, 1302, 1396, and 1624 cm^{-1}, among which 600 cm^{-1} belongs to the out-of-plane bending vibration of the aromatic ring, 775 cm^{-1} is the C–H out-of-plane bending vibration of the aromatic ring, 1302 cm^{-1} and 1396 cm^{-1} are the C–H in-plane bending vibration of the aromatic ring, and 1624 cm^{-1} are C=C stretching vibrations [44,45]. The main Raman characteristic peaks of MG are shown in Table S3, which are 801, 916, 1171, 1217, 1366, 1395 and 1616 cm^{-1}. Its vibration modes are C–H out-of-plane bending (801 cm^{-1}), C–H in-plane bending vibration (916 cm^{-1} and 1217 cm^{-1}), C-N stretching vibration (1171 cm^{-1}), C–H symmetrical bending vibration (1366 cm^{-1} and 1395 cm^{-1}), and C=C stretching vibration (1616 cm^{-1}) [46–48]. The main Raman shifts of RhB are shown in Table S4, which are 621, 1199, 1280, 1359, 1510, 1530, and 1647 cm^{-1}. The corresponding vibration modes are C–H out-of-plane bending vibration (621 cm^{-1}), C-N stretching vibration (1199 cm^{-1}), C–C stretching vibration (1280 cm^{-1}), C–H symmetrical bending vibration (1359 cm^{-1}), C=C stretching vibration (1510 cm^{-1} and 1530 cm^{-1}), C=O stretching vibration (1647 cm^{-1})[49–51].

In order to improve the accuracies of the machine learning models, SERS spectra of dyes at different concentrations ranging from 10^{-10} to 10^{-5} M were collected (Fig. 3 b-e). The SERS spectra of specific dye mixtures (CV + MG, CV + RhB and MB + MG) at different concentration were also acquired, respectively (Fig. 3 f-i). Apparently, these spectra are highly similar which are difficult to be distinguished and identified visually. To such cases, the assistance of machine learning is significant.

3.4. Training of classification models using single dye spectra

Before training the classification models, it is necessary to analyze the spectral bands to avoid excessive redundancy information, which may lead to overfitting of the model and hinder classification, as shown in Fig. 4a. The ReliefF algorithm was used to analyze the importance index of the four dye wavelengths to select useful spectral wavelengths, as shown in Fig. 4b. The cross-validation method was used to select the optimal number of features, and the top 50, 100, 150, 200, 250, 300, 350, and 400 features were selected for cross-validation. When the number of features was 200, the best results were obtained. Therefore, the top 200 spectral wavelengths with the highest feature importance coefficients were selected for model training. A total of 210 spectra of different concentration gradients of each dye were collected, and the spectral data was randomly divided into a training set (80 %) and a test set (20 %). In order to eliminate the effect of dataset partitioning, the whole process (including dataset partitioning, model training and validation) was repeated 10 times and the average value was taken as the final result. The evaluation metrics and confusion matrices for the four machine learning models are shown in Fig. 4 c-f and Fig. S3. The KNN model had an overall accuracy of 92 %, a recall rate of 93 %, an accuracy rate of 92 %, and an F1 score of 92 %. The recognition accuracies of KNN for CV and RhB were 100 %, and the recognition accuracy of MG was

89 %, among which 11 % was incorrectly identified as MB, and the worst recognition effect was that only 82 % of MB was accurately identified and 18 % was incorrectly identified as RhB. However, SVM model had an overall accuracy of 95 %, a recall rate of 95 %, an accuracy rate of 94 %, and an F1 score of 94 %. SVM had accuracies of 100 % for CV and RhB, an accuracy of 92 % for MB, where 8 % was identified as MG, and an accuracy of 88 % for MG where 12 % was misidentified as CV. In contrast, the RF model had an overall accuracy rate of 95 %, a recall rate of 96 %, an accuracy rate of 95 % and an F1 score of 95 %. RF had accuracies of 100 % for CV and RhB, an accuracy of 90 % for MB where 10 % was identified as RhB, and an accuracy of 93 % for MG, where 7 % was misidentified as CV. Compared to others, the CNN model has an overall accuracy of 98 %, a recall rate of 97 %, an accuracy rate of 97 %, and an F1 score of 97 %. CNN had accuracies of 100 % for CV, MB, and RhB, an accuracy of 88 % for MG where12% was misidentified as CV. Considering the comprehensive evaluation of accuracy rate, recall rate, accuracy rate and recall rate, CNN had the best classification effect for single dyes among the four classification models.

3.5. Training of ICA using the SERS spectra of dye mixtures

In this study, ICA was used to process the Raman spectra of several dye mixtures in order to effectively distinguish their components. 120 spectra of dye mixtures were imported for the training of ICA. Fig. 5 a-c shows the Raman spectra of three groups of dye mixtures (CV + RhB, CV + MG, and MB + MG) and the results obtained by ICA. For the CV + RhB mixed dye, ICA processing was able to clearly distinguish two independent components, Component 1 and Component 2, which were highly consistent with the characteristic peaks of RhB and CV. Similarly, after ICA treatment, the independent components of CV + MG and MB

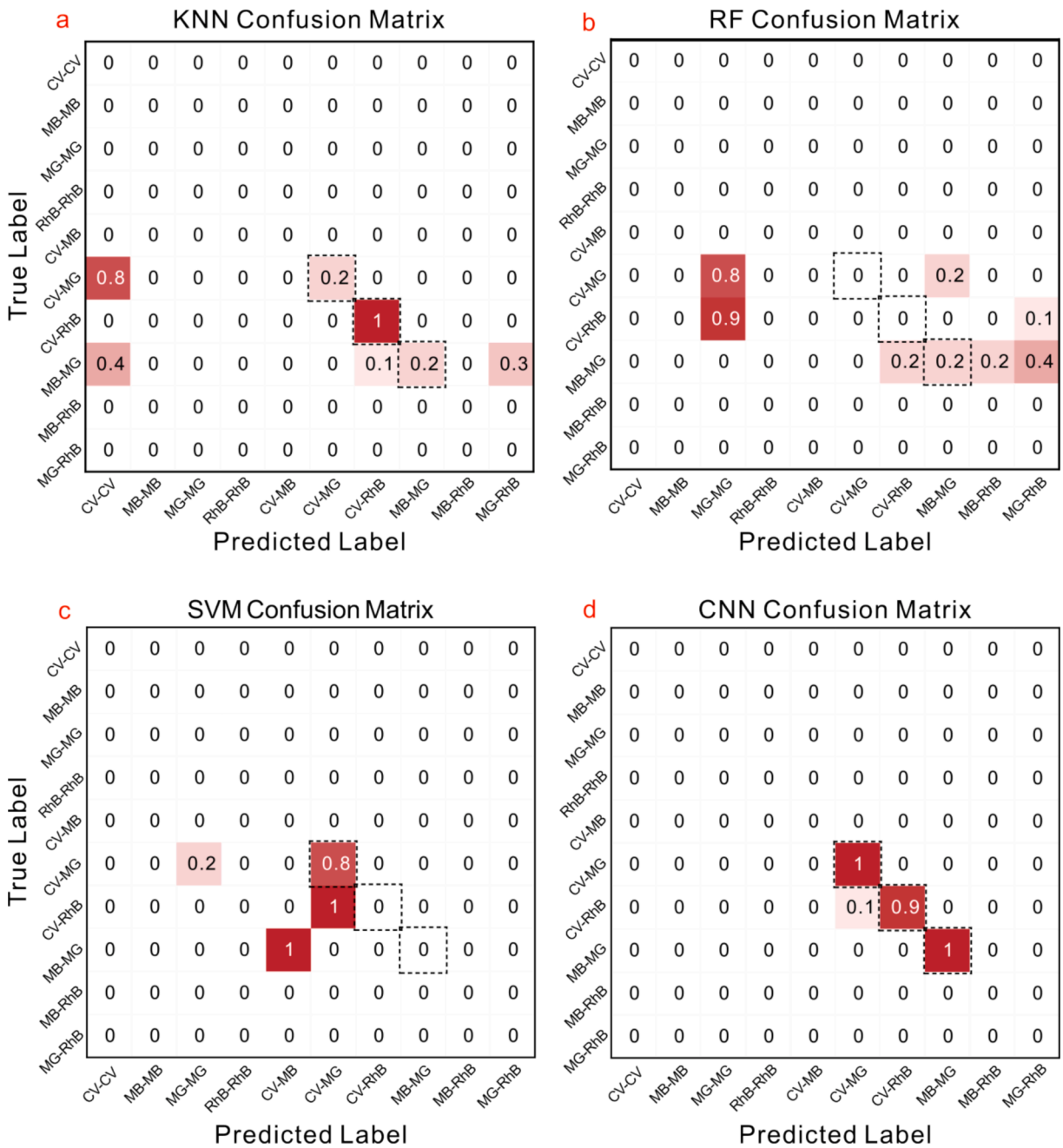


Fig. 6. Confusion matrices for classification of dye mixtures by a) KNN, b) RF, c) SVM and d) CNN models.

+ MG mixed dyes also well matched with the characteristic peaks of CV/MG and MB/MG, respectively. The above results indicate the feasibility and accuracy of ICA in separating Raman spectra of complex dye mixtures.

3.6. Identification of dye mixtures using ICA and classification models

The construction of the dye mixture dataset is shown in Fig. S4: For the three dye mixtures, CV + RhB, CV + MG, and MB + MG, different concentration gradients (5×10^{-6} M – 5×10^{-9} M) were collected for each dye, and each concentration gradient was used as an ICA separation dataset, with a total of 12 datasets containing 10 SERS spectra per dataset. In addition, all concentration gradients for each dye mixture were used as ICA separation datasets, with 3 large groups containing 40 SERS spectra each, for a total of 15 for all subgroups. In order to reduce the bias caused by random factors in a single separation and thus to evaluate the separation and classification performance more accurately, two separation operations by using ICA were performed for each subgroup, with the initialization parameters set randomly each time, and two spectra (Component1,Component2) were separated for each subgroup. 30 groups of data separated by ICA totaling 60 spectral data were imported into the four previously trained models, and the evaluation metrics and confusion matrices obtained from 10 predictions are shown in Fig. 6 and Fig. S5. KNN model had an overall accuracy of 47 % with a recall rate of 23 %, an accuracy rate of 58 %, and an F1 score of 32 %. KNN correctly identified 100 % of CV + RhB while only 20 % of CV + MG were identified and 80 % were incorrectly identified as CV + CV, and only 20 % of MB + MG were identified and the rest were incorrectly identified as CV + CV (40 %), CV + RhB (10 %) and MG + RhB (30 %) (Fig. 6a). The RF model had the worst performance with an overall accuracy of 7 %, a recall rate of 3 %, an accuracy rate of 8 %, and an F1 score of 5 %. The accuracy of RF identification of CV + MG and CV +

RhB was 0 among which CV + MG was incorrectly identified as MG + MG (80 %) and MB + MG (20 %), CV + RhB was incorrectly identified as MG + MG (90 %) and MG + RhB (10 %). The recognition accuracy of MB + MG was 20 %. Others were misidentified as CV + RhB (20 %), MB + RhB (20 %), and MG + RhB (40 %) (Fig. 6b). SVM model had an overall accuracy of 27 %, a recall rate of 16 %, an accuracy rate of 9 % and an F1 score of 11 %. SVM correctly identified 80 % of CV + MG and 20 % were incorrectly identified as MG + MG, the correct recognition rate of CV + RhB and MB + MG was 0 among which CV + RhB was incorrectly identified as CV + MG, and MB + MG was incorrectly identified as CV + MB (Fig. 6c). In contrast, CNN had an overall accuracy of 97 %, a recall rate of 97 %, an accuracy rate of 97 % and an F1 score of 97 %. The recognition rate of CV + MG and MB + MG was 100 % while the recognition rate of CV + RhB was 90 % and 10 % were incorrectly identified as CV + MG (Fig. 6d). Referring to the confusion matrices, it is found that KNN, RF and SVM had very poor classification effect for multi-component dyes, while CNN model presented the best classification performance.

4. Conclusion

In summary, ICA combined with different classification models were applied and investigated to assist the SERS study of dye mixtures. Silver coated gold nanocubes (Au@AgNCs) were prepared as the SERS substrate. Various dyes including CV, RhB, MB and MG were selected as model molecules. To mimic practical applications, the mixtures of CV + RhB, CV + MG and MB + MG were introduced. Classification models including KNN, SVM, RF and CNN were applied and investigated in the identification of single and mixed dyes. The spectra of single dyes were used for the training and test of classification models where the results showed that CNN is the best model with an accuracy of 98 %. For the identification of dye mixtures, ICA was firstly trained to separate the spectra of dye mixtures where the separated signals showed good consistencies to the analytes. Combined with the trained ICA, the four trained classification models were applied in the classification of dye mixtures. CNN also showed excellent performance, with an overall accuracy of 97 %. This study demonstrates that the analysis of multi-component surface-enhanced Raman scattering (SERS) spectra assisted with the signal separation techniques and machine learning classification models efficiently solving the characterization problems caused by cross-interference of multi-component spectra or interference of signals in complex environments, which is expected to be applied in the fields of medical and pharmaceutical analyses, characterization of environmental pollutants, identification of agricultural products, *etc*.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgements

This work was supported by the National Natural Science Foundation of China (No. 22204166), the National Key R&D Program of China (No. 2021YFB3201400), The Youth Innovation Promotion Association of CAS (No. 2023469), the Key R&D Program of Anhui Province (No. 202104i07020011), the HFIPS Director's Fund (No. YZJJ202102, YZJJ-GGZX-2022-01).

Appendix A. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.saa.2025.125806.

Data availability

Data will be made available on request.

References

[1] J. Zhang, F. Li, Q. Sun, Rapid and selective adsorption of cationic dyes by a unique metal-organic framework with decorated pore surface, *Appl. Surf. Sci.* 440 (2018) 1219–1226.
[2] T. Islam, M.R. Repon, T. Islam, Z. Sarwar, M.M. Rahman, Impact of textile dyes on health and ecosystem: a review of structure, causes, and potential solutions, *Environ Sci Pollut Res Int* 30 (4) (2023) 9207–9242.
[3] L. Vanhinsbergh, S. Uthaya, B.J. Bain, Methylene blue-induced heinz body hemolytic anemia in a premature neonate, *Am J Hematol* 93 (5) (2018) 716–717.
[4] S. Mani, R.N. Bharagava, Exposure to crystal violet, Its toxic, genotoxic and carcinogenic effects on environment and Its degradation and detoxification for environmental safety, *Rev Environ Contam Toxicol* 237 (2016) 71–104.
[5] T. Li, D. Tian, Z. Zhu, W. Jin, S. Wu, H. Li, The gut microbiota: a new perspective on the toxicity of malachite green (MG), *Appl. Microbiol. Biotechnol.* 103 (23–24) (2019) 9723–9737.
[6] Q. Sun, Y. Li, Y. Su, M. Wei, H. Li, J. Liu, Determination of restricted dyes in textile raw material solid wastes by ultra-high performance liquid chromatography-tandem mass spectrometry, *J. Chromatogr. A* 1711 (2023).
[7] T. Tomić, S. Babić, M. Biošić, N.U. Nasipak, A.-M. Čižmek, Determination of the solvent blue 35 dye in diesel fuel by solid phase extraction and high-performance liquid chromatography with ultraviolet detection, *Dyes Pigm.* 150 (2018) 216–222.
[8] W. Schwack, E. Pellissier, G. Morlock, Analysis of unauthorized sudan dyes in food by high-performance thin-layer chromatography, *Anal Bioanal Chem* 410 (22) (2018) 5641–5651.
[9] X. Wang, C. Zhu, X. Hu, Q. Xu, H. Zhao, G. Meng, Y. Lei, Highly sensitive surface-enhanced Raman scattering detection of organic pesticides based on ag-nanoplate decorated graphene-sheets, *Appl. Surf. Sci.* 486 (2019) 405–410.
[10] Li, Z.; Sun, K.; Du, Z.; Chen, B.; He, X., Galvanic-Cell-Reaction-Driven Deposition of Large-Area Au Nanourchin Arrays for Surface-Enhanced Raman Scattering. *Nanomaterials (Basel)* **2018,** *8* (4).
[11] K. Sun, Z. Tang, Q. Zhao, Q. Deng, X. Wang, H. Tang, Improved SERS detection of anionic pollutant with Ag Nanoparticles-decorated ZnAl layered double hydroxide nanosheets array, *Mater. Lett.* 314 (2022).
[12] B. Yan, S. Wang, M. Muhammad, C. Zhu, K. Sun, Q. Huang, Preparation of transparent SERS substrates based on the stepwise anodization potential AAO-template approach for rapid detection of trace pesticide residues, *Opt. Mater.* 149 (2024).
[13] S. Chao, C. Valsecchi, J. Sun, H. Shao, X. Li, C. Tang, M. Fan, Highly sensitive Surface-Enhanced Raman Scattering detection of hydroxyl radicals in water microdroplets using phthalhydrazide/Ag nanoparticles nanosensor, *Environ Sci Technol* 58 (37) (2024) 16497–16506.
[14] G. Fang, W. Hasi, X. Lin, S. Han, Automated identification of pesticide mixtures via machine learning analysis of TLC-SERS spectra, *J. Hazard. Mater.* 474 (2024).
[15] J.D. Weatherston, R.K.O. Seguban, D. Hunt, H.-J. Wu, Low-Cost and simple fabrication of nanoplasmonic paper for coupled chromatography separation and surface enhanced raman detection, *ACS Sensors* 3 (4) (2018) 852–857.
[16] X. Lu, Y. Ma, S. Jiang, Z. Wang, Q. Yu, C. Ji, J. Guo, X. Kong, Quantitative monitoring ofloxacin in beef by TLC-SERS combined with machine learning analysis, Molecular and Biomolecular Spectroscopy, Spectrochimica Acta Part A, 2024, p. 308.
[17] M.M. Bajomo, Y. Ju, J. Zhou, S. Elefterescu, C. Farr, Y. Zhao, O. Neumann, P. Nordlander, A. Patel, N.J. Halas, Computational chromatography: A machine learning strategy for demixing individual chemical components in complex mixtures, *Proc Natl Acad Sci U S A* 119 (52) (2022) e2211406119.
[18] X. Wang, F. Li, L. Wei, Y. Huang, X. Wen, D. Wang, G. Cheng, R. Zhao, Y. Lin, H. Yang, M. Fan, Rapid and precise differentiation and authentication of agricultural products via deep learning-assisted multiplex SERS fingerprinting, *Anal Chem* 96 (11) (2024) 4682–4692.
[19] G. Fang, X. Lin, J. Wu, W. Xu, W. Hasi, B. Dong, Porous Nanoframe Based Plasmonic Structure With High-Density Hotspots for the Quantitative Detection of Gaseous Benzaldehyde, *Small* (2025) e2408670.
[20] Y. Luo, W. Su, D. Xu, Z. Wang, H. Wu, B. Chen, J. Wu, Component identification for the SERS spectra of microplastics mixture with convolutional neural network, *Sci Total Environ* 895 (2023) 165138.
[21] Z. Liu, M. Huang, Q. Zhu, J. Qin, M.S. Kim, A packaged food internal Raman signal separation method based on spatially offset Raman spectroscopy combined with FastICA, *Spectrochim Acta A Mol Biomol Spectrosc* 275 (2022) 121154.
[22] M. Szymańska-Chargot, P.M. Pieczywek, M. Chylińska, A. Zdunek, Hyperspectral image analysis of Raman maps of plant cell walls for blind spectra characterization by nonnegative matrix factorization algorithm, *Chemom. Intel. Lab. Syst.* 151 (2016) 136–145.
[23] A. de Juan, R. Tauler, Multivariate Curve Resolution: 50 years addressing the mixture analysis problem - A review, *Anal Chim Acta* 1145 (2021) 59–78.
[24] Y. Ma, W. Li, E.C. Cho, Z. Li, T. Yu, J. Zeng, Z. Xie, Y. Xia, Au@Ag core-shell nanocubes with finely tuned and well-controlled sizes, shell thicknesses, and optical properties, *ACS Nano* 4 (11) (2010) 6725–6734.
[25] C. Lin, K. Tao, D. Hua, Z. Ma, S. Zhou, Size effect of gold nanoparticles in catalytic reduction of p-nitrophenol with NaBH4, *Molecules* 18 (10) (2013) 12609–12620.

[26] J. Huang, T. Shi, B. Gong, X. Li, G. Liao, Z. Tang, Fitting an optical fiber background with a weighted Savitzky–Golay smoothing filter for Raman spectroscopy, *Appl. Spectrosc.* 72 (11) (2018) 1632–1644.
[27] Z.M. Zhang, S. Chen, Y.Z. Liang, Baseline correction using adaptive iteratively reweighted penalized least squares, *Analyst* 135 (5) (2010) 1138–1146.
[28] A. Hyvarinen, E. Oja, Independent component analysis: algorithms and applications, *Neural Netw* 13 (4–5) (2000) 411–430.
[29] Y.-H. Chen, S.-P. Wang, Low-Cost implementation of independent component analysis for biomedical signal separation using Very-Large-Scale integration, *IEEE Trans. Circuits Syst. Express Briefs* 67 (12) (2020) 3437–3441.
[30] S. Zhang, Challenges in KNN classification, *IEEE Trans. Knowl. Data Eng.* 34 (10) (2022) 4663–4675.
[31] R.L. MacArthur, E. Teye, S. Darkwa, Predicting adulteration of palm oil with Sudan IV dye using shortwave handheld spectroscopy and comparative analysis of models, *Vib. Spectrosc* 110 (2020).
[32] S. Niell, F. Jesús, R. Díaz, Y. Mendoza, G. Notte, E. Santos, N. Gérez, V. Cesio, H. Cancela, H. Heinzen, Beehives biomonitor pesticides in agroecosystems: Simple chemical and biological indicators evaluation using support vector machines (SVM), *Ecol. Ind.* 91 (2018) 149–154.
[33] W. Ye, T. Yan, C. Zhang, L. Duan, W. Chen, H. Song, Y. Zhang, W. Xu, P. Gao, Detection of pesticide residue level in grape using hyperspectral imaging with machine learning, *Foods* 11 (11) (2022).
[34] S. Moosavi, O. Manta, Y.A. El-Badry, E.E. Hussein, Z.M. El-Bahy, N.F.B. Mohd Fawzi, J. Urbonavicius, S.M.H. Moosavi, A study on machine learning methods' application for dye adsorption prediction onto agricultural waste activated carbon, *Nanomaterials (basel)* 11 (10) (2021).
[35] S. Zhang, Z. Tan, J. Liu, Z. Xu, Z. Du, Determination of the food dye indigotine in cream by near-infrared spectroscopy technology combined with random forest model, *Spectrochim Acta A Mol Biomol Spectrosc* 227 (2020) 117551.
[36] C. Jiang, A.J.X. Guo, Y. Li, Y. Wang, J. Sun, Z. Chen, X. Chen, Q. Ma, Inspired by human olfactory system: Deep-learning-assisted portable chemo-responsive dye-based odor sensor array for the rapid sensing of shrimp and fish freshness, *Chem. Eng. J.* 484 (2024).
[37] A. Manda-Handzlik, K. Fiok, A. Cieloch, E. Heropolitanska-Pliszka, U. Demkow, Convolutional neural Networks-Based image analysis for the detection and quantification of neutrophil extracellular traps, *Cells* 9 (2) (2020).
[38] S. Stefa, M. Griniezaki, M. Dimitropoulos, G. Paterakis, C. Galiotis, G. Kiriakidis, E. Klontzas, M. Konsolakis, V. Binas, Highly porous Thin-Layer g-C3N4 nanosheets with enhanced adsorption capacity, *ACS Appl. Nano Mater.* 6 (3) (2023) 1732–1743.
[39] R. Ghezelbash, A. Maghsoudi, M. Shamekhi, B. Pradhan, M. Daviran, Genetic algorithm to optimize the SVM and means algorithms for mapping of mineral prospectivity, *Neural Comput Appl* 35 (1) (2023) 719–733.
[40] Z. Feng, L. Zheng, J. Liu, Classification of household microplastics using a multi-model approach based on Raman spectroscopy, *Chemosphere* 325 (2023) 138312.
[41] L. Yin, T. You, H.R. El-Seedi, I.M. El-Garawani, Z. Guo, X. Zou, J. Cai, Rapid and sensitive detection of zearalenone in corn using SERS-based lateral flow immunosensor, *Food Chem* 396 (2022) 133707.
[42] W. Wei, Q. Huang, Preparation of cellophane-based substrate and its SERS performance on the detection of CV and acetamiprid, *Spectrochim. Acta A Mol. Biomol. Spectrosc.* 193 (2018) 8–13.
[43] Y.-C. Lu, W.-H. Chiang, C.-Y. Liu, J.P. Chu, H.-C. Ho, C.-H. Hsueh, Wafer-scale SERS metallic nanotube arrays with highly ordered periodicity, Chemical, Sensors and Actuators B, 2021, p. 329.
[44] M. Verma, T.K. Naqvi, S.K. Tripathi, M.M. Kulkarni, P.K. Dwivedi, Paper based low-cost flexible SERS sensor for food adulterant detection, *Environ. Technol. Innov.* 24 (2021).
[45] E. Murugan, S. Santhoshkumar, S. Govindaraju, M. Palanichamy, Silver nanoparticles decorated g-C(3)N(4): An efficient SERS substrate for monitoring catalytic reduction and selective Hg(2+)ions detection, *Spectrochim Acta A Mol Biomol Spectrosc* 246 (2021) 119036.
[46] Y. Liu, H. Guan, S. Lin, H. Dong, W. Hasi, B. Dong, Plasmonic nanosensor based on Ag nanocubes of high purification by extraction filtration strategy for SERS determination of malachite green in aquaculture water, Chemical, Sensors and Actuators B, 2022, p. 358.
[47] Y.N. Liu, J.J. Li, G.J. Weng, J. Zhu, J.W. Zhao, Reliable detection of malachite green by self-assembled SERS substrates based on gold-silicon heterogeneous nano pineapple structures, *Food Chem* 451 (2024) 139454.
[48] Y. Zhang, Y. Huang, Y. Kang, J. Miao, K. Lai, Selective recognition and determination of malachite green in fish muscles via surface-enhanced Raman scattering coupled with molecularly imprinted polymers, *Food Control* 130 (2021).
[49] X. Lin, L.-J. Li, H.-Y.-X. Guo, R. Li, J. Feng, Preparation of 3D nano silver trees/sea urchin-like gold and SERS detection of uric acid, Molecular and Biomolecular Spectroscopy, Spectrochimica Acta Part A, 2024, p. 305.
[50] S. Kamal, T.C.K. Yang, A silver trimesate organic framework as an ultrasensitive surface-enhanced Raman scattering substrate for detection of various organic pollutants, *Microchem. J.* 163 (2021).
[51] D. Xu, H. Jiang, S. Zhang, W. Yang, Y. Zhang, Z. Wang, J. Chen, High roughness gold nanoparticles/silver nanowires composites: Fabrication, characterization and ultrasensitive SERS detection towards Rhodamine B, *Microchem. J.* 158 (2020).