



Unraveling surface-enhanced Raman spectroscopy results through chemometrics and machine learning: principles, progress, and trends

Diego P. dos Santos¹ · Marcelo M. Sena^{2,3} · Mariana R. Almeida² · Italo O. Mazali¹ · Alejandro C. Olivieri⁴ · Javier E. L. Villa¹

Received: 24 November 2022 / Revised: 2 February 2023 / Accepted: 20 February 2023 / Published online: 3 March 2023
© Springer-Verlag GmbH Germany, part of Springer Nature 2023

Abstract

Surface-enhanced Raman spectroscopy (SERS) has gained increasing attention because it provides rich chemical information and high sensitivity, being applicable in many scientific fields including medical diagnosis, forensic analysis, food control, and microbiology. Although SERS is often limited by the lack of selectivity in the analysis of samples with complex matrices, the use of multivariate statistics and mathematical tools has been demonstrated to be an efficient strategy to circumvent this issue. Importantly, since the rapid development of artificial intelligence has been promoting the implementation of a wide variety of advanced multivariate methods in SERS, a discussion about the extent of their synergy and possible standardization becomes necessary. This critical review comprises the principles, advantages, and limitations of coupling SERS with chemometrics and machine learning for both qualitative and quantitative analytical applications. Recent advances and trends in combining SERS with uncommonly used but powerful data analysis tools are also discussed. Finally, a section on benchmarking and tips for selecting the suitable chemometric/machine learning method is included. We believe this will help to move SERS from an alternative detection strategy to a general analytical technique for real-life applications.

Keywords Nanomaterials · Vibrational spectroscopy · Plasmonics · Data analysis · Supervised methods · PCA

Introduction

In the last decades, nanomaterials have become very important because of their useful properties, such as a large surface area to volume ratio and enhanced electrical and optical

responses, mainly related to the quantum size effect [1]. By exploiting these properties, outstanding advances have been achieved in science and technology applications. From colorimetric-based assays to ultrasensitive fluorescent sensors, the optical properties of nanomaterials have made it possible to design a wide variety of analytical platforms [2]. In this regard, an efficient sensing platform should be capable to characterize samples with complex matrices, demanding high sensitivity (to monitor relevant changes in composition at trace levels), and selectivity and chemical specificity (to capture and unequivocally identify target molecule(s) present(s) in the system).

Surface-enhanced Raman spectroscopy (SERS) is a nanoscience-based technology that comprises high sensitivity and chemical specificity for molecules adsorbed on nanostructured metallic surfaces. Since the discovery and correct interpretation of SERS in the 1970s, advances in nanotechnology have driven its rapid development and analytical applications [3]. This can be corroborated by the continuous increase in the number of SERS-related publications every year as shown in Fig. 1. In contrast to other highly sensitive techniques (e.g., fluorescence spectroscopy), SERS provides an

Published in the topical collection *Young Investigators in (Bio-) Analytical Chemistry 2023* with guest editors Zhi-Yuan Gu, Beatriz Jurado-Sánchez, Thomas H. Linz, Leandro Wang Hantao, Nongnoot Wongkaew, and Peng Wu.

✉ Javier E. L. Villa
jelv@unicamp.br

¹ Instituto de Química, Universidade Estadual de Campinas (UNICAMP), Campinas, SP 13083-970, Brazil

² Departamento de Química, Universidade Federal de Minas Gerais (UFMG), Belo Horizonte, MG 31270-901, Brazil

³ Instituto Nacional de Ciência e Tecnologia em Bioanalítica (INCT Bio), Campinas, SP 13083-970, Brazil

⁴ Departamento de Química Analítica, Facultad de Ciencias Bioquímicas y Farmacéuticas, Universidad Nacional de Rosario, Instituto de Química Rosario (IQUIR-CONICET), Suipacha 531, 2000 Rosario, Argentina

additional spectral fingerprint that stems from the vibrational modes of the molecules. Therefore, structural and chemical information can be obtained with high sensitivity. The selective adsorption of target molecules is nevertheless unusually observed in the analysis of real-world samples, in which the matrix molecules may be much more concentrated than the analyte and competitive adsorption for the available surface is recurrent. Consequently, at the core of every SERS-based analytical method is the challenge to selectively adsorb target molecules onto the metallic surface before the measurements.

Own to the lack of selectivity, a complex overlapping of SERS signals is often observed. To deal with this issue, one can modify the metallic surface with capture agents that possess a high affinity for the analyte [4, 5]. However, the need for additional fabrication steps and specific capture agents are the main limitations of this approach. Advanced signal processing methods based on chemometrics and machine learning can alternatively be implemented. The incorporation of these methods has been gaining prominence in SERS, mainly due to improvements in selectivity and automatization toward the development of “smart” sensors. As can be seen in Fig. 1 (inset), the number of publications using SERS along with chemometrics and machine learning has rapidly increased in the last 5 years. These multivariate approaches are devoted to extracting key information contained in the data and, therefore, have great potential to boost the development of automated SERS sensors for real-world applications.

In this review, we start by defining some theoretical and practical concepts of SERS, chemometrics, and machine learning to provide a useful background in the context of analytical chemistry. We then survey recent advances in the combination of SERS and a range of different multivariate

models for both qualitative and quantitative applications. Limitations, challenges, and trends of this coupling to achieve the much-desired SERS selectivity via advanced data analysis are also discussed. Finally, benchmarking and insights about the selection of the suitable multivariate method for SERS applications are presented.

Surface-enhanced Raman spectroscopy/scattering (SERS)

The intensity of Raman scattering of molecules adsorbed onto nanostructured metals (e.g., Ag, Au, and Cu) is substantially enhanced if compared to the normal Raman measurements. This enhancement effect, known as SERS, stems from chemical and electromagnetic mechanisms that will be described in this section.

The Raman scattering intensity of a vibrational mode of frequency ω_{vib} excited with radiation laser of frequency ω_0 can be expressed as [6]:

$$I_{Raman} \propto (\omega_0 \pm \omega_{vib})^4 N |E_0|^2 \sum_{\rho, \sigma} |\alpha_{\rho\sigma}|^2 \quad (1)$$

where $\omega_0 \pm \omega_{vib}$ is the frequency of the scattered radiation (+ for anti-Stokes and – for Stokes), N is the number of scattering molecules, E_0 is the incident radiation electric field amplitude, and $\alpha_{\rho\sigma}$ are the elements of the Raman polarizability tensor. For molecules with small polarizability, such as water, the Raman and SERS intensities are very weak. This constitutes an important advantage for studying molecules in an aqueous solution, unlike infrared absorption spectroscopy. The chemical mechanism is related to modifications of the polarizability tensor upon molecular adsorption, especially in chemisorption that leads to strong interactions between molecular and metal electronic states [7]. On the other hand, the electromagnetic mechanism stems from modifications of the local field probed by a molecule at the metal surface (distance < 10 nm) and modifications in the dipole radiative emission rate [6]. As the electromagnetic mechanism is mainly responsible for signal enhancement, we will focus on it to highlight analytical applications. It should be noted, however, that important effects due to chemical mechanisms can be observed under specific experimental conditions, such as in spectroelectrochemical studies [8].

For nanostructured materials made of Ag, Au, and Cu, the incident radiation fields can strongly couple with excitations of surface plasmon polariton modes in the metal (collective oscillation of charge density). The general result of such excitations is a considerable increase in the electric field near the metallic surface, especially in the case of coupled nanoparticle systems [9, 10]. As an example, Fig. 2A (black

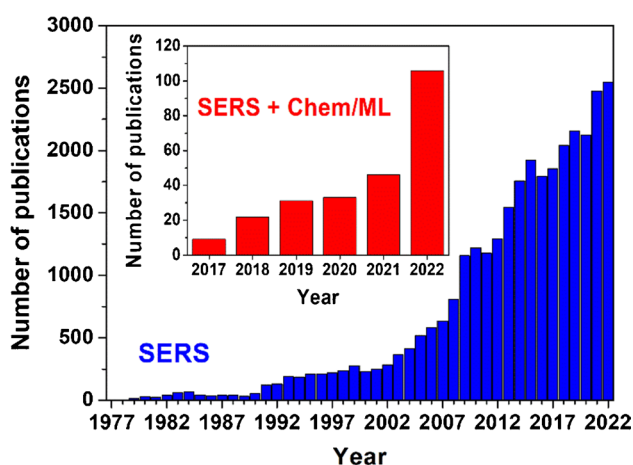


Fig. 1 Published SERS papers since its discovery. The number of recently published papers related to SERS coupled to chemometrics (Chem) and machine learning (ML) is shown in inset. Source: Web of Science—January 2023 (keywords: SERS, Chemometrics and Machine Learning)

line) presents a numerical simulation based on the generalized Mie theory for the near-field enhancement $(|E|/|E_0|)^2$ in the gap of two Ag nanospheres of 50 nm in diameter. At such regions, known as hot spots, it is possible to achieve very large field enhancements, making it even possible to experimentally detect signals from single molecules [11]. For this reason, it is a common analytical practice to promote nanoparticle cluster/aggregate formation before SERS measurements. Another important aspect of Raman scattering and SERS is the induced dipole in a molecule oscillating at a frequency $\omega_0 \pm \omega_{vib}$ (see Eq. 1). For such an oscillating dipole, the relaxation may occur either via a radiative decay, giving rise to an inelastically scattered field, or via a non-radiative decay pathway. The key point here is that the radiative decay rate (scattering probability) can be strongly enhanced in the presence of metallic nanomaterials [6]. Figure 2A (red lines) shows the enhancement in the radiative decay rate, i.e., the ratio between rates in the presence and absence of nanoparticles, Γ/Γ_0 . The plasmon modes increase the scattering intensity associated to an oscillating dipole. Interestingly, the scattering enhancement is dependent on dipole orientation relative to the metal surface. Figure 2A also displays a comparison of the radiative decay rate enhancement for a dipole perpendicular to the metal surface and after 45° rotation. In SERS, such dipole configurations simulate orientations of vibrational modes, and the largest increase in scattering is for vibrations oscillating perpendicularly to the metal surface.

This result leads to differential enhancements in the Raman signals among vibrational modes depending on orientations relative to the surface. Therefore, the SERS spectrum may display significant changes in terms of relative band intensities when compared to normal Raman measurements [6].

The discussion about the electromagnetic mechanism leads to the following figure of merit describing the expected enhancement in SERS measurements relative to normal Raman:

$$F = \left(\frac{|E_{loc}|}{|E_0|} \right)_{\omega_0}^2 \cdot \left(\frac{\Gamma}{\Gamma_0} \right)_{\omega_0 \pm \omega_{vib}} \tag{2}$$

where F is the so-called SERS enhancement factor, considering only the electromagnetic mechanism. The first term represents the near-field enhancement contribution (black line in Fig. 2A), whereas the second term describes the modification in dipole spontaneous emission (red lines in Fig. 2A). Note that, for a dipole perpendicular to the surface, the scattering probability and near-field enhancements are virtually the same, which allows approximating:

$$F \approx \left(\frac{|E_{loc}|}{|E_0|} \right)_{\omega_0}^2 \cdot \left(\frac{|E_{loc}|}{|E_0|} \right)_{\omega_0 \pm \omega_{vib}}^2 \approx \left(\frac{|E_{loc}|}{|E_0|} \right)_{\omega_0}^4 \tag{3}$$

where the second approximation, a.k.a. E^4 approximation [12], is due to a small difference between incident and inelastic field frequencies (see Raman shift scale in Fig. 2A).

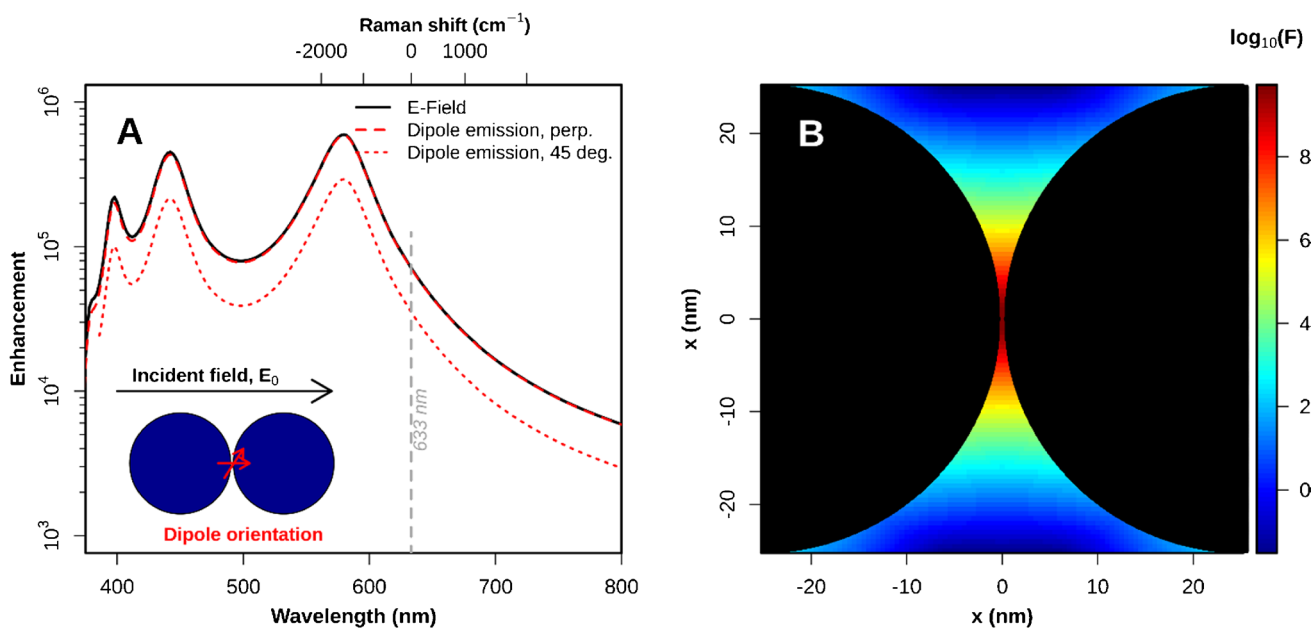


Fig. 2 Mie theory simulations for a dimer of Ag nanospheres. **A** Local field enhancement (computed as the squared ratio between local and incident electric fields, black line), and dipole radiative decay rate enhancement (ratio between decay rates in the presence and absence of nanoparticles, red lines). Two dipole orientations are

considered: perpendicular and 45° rotation (inset scheme). The field and radiative enhancements are computed as a function of wavelength (incident and scattered radiation). **B** E^4 approximation simulation of SERS enhancement factor spatial distribution for a 633 nm excitation source

The largest SERS enhancements can be achieved for excitation source in resonance with plasmon excitations tuned via nanoparticle/hot spot engineering. Plasmonic nanomaterial fabrication is a very active research area, and they can be obtained by bottom-up or top-down approaches [3, 12]. Lithographic techniques are excellent examples of the latter, in which shapes are crafted on a solid platform. The main limitation with such techniques is the not cost-effective spatial resolution the patterns can be created. Alternatively, one can produce SERS substrates by adsorbing metal nanoparticles onto rigid surfaces to generate hot spots with large enhancement factors [13–15]. Wet chemistry synthesis is widely employed as a bottom-up methodology for obtaining colloidal nanoparticles with different shapes (spheres, rods, prisms, stars, cubes, etc.) and sizes. Furthermore, hot spot engineering and self-assembly strategies allow the fabrication of ultrasensitive rigid SERS substrates from these nanoparticles [3, 12]. SERS measurements can be performed by direct analyte addition to the colloid or onto the rigid SERS substrate. In colloidal solution, nanoparticle aggregation yields strong SERS signals from the analyte because of the large number of hot spots created. Nevertheless, uncontrolled aggregation and natural shape dispersion of particles may decrease SERS repeatability. Controlled aggregation and self-assembly of nanoparticles have been reported in the literature to mitigate this problem [16–19].

Even under controlled conditions or the use of high-quality rigid SERS substrates, intensity fluctuations may happen because of the hot spot spatial localization, especially at low analyte concentrations [20]. Figure 2B shows a 2D map simulation for the enhancement factor distribution in a hot spot at 633 nm excitation. A molecule in the hot spot region should display a large enhancement in the scattering intensities, making it possible to detect spectroscopic signals at ultra-low concentrations (e.g., below 10^{-12} mol L⁻¹). However, moving a few nanometers from the hot spot center is enough to decrease F by various orders of magnitude. This means that only molecules approaching the hot spot can produce significant SERS intensities to be detected at ultra-low concentrations, imposing a spatial limitation to the SERS response. Moreover, once F is not homogeneously distributed on the metal surface, the SERS intensities do not show a linear behavior with the increase molecular surface coverage [20]. Therefore, SERS intensities do not linearly scale with analyte concentrations at trace levels.

The electromagnetic mechanism predicts the same enhancement regardless of the molecular structure, though the affinity for the metal surface can dramatically change from one molecule to another (e.g., mercaptobenzoic acid displays a much higher affinity for metallic surfaces than glucose). Since the electromagnetic mechanism is dominant in most of the experiments, SERS is not considered a

selective technique and depends on the competitive diffusion/adsorption of the molecules onto the metal surface. Thus, a SERS spectrum may contain a highly complex mixture of bands associated to the components of the sample. One strategy to selectively adsorb an analyte present in a complex matrix is by modifying the metal surface with capture agents that possess strong interactions with the analyte [4, 5, 21]. However, this procedure is subjected to additional experimental steps and might require specific capture agents when interferences with similar chemical structures to the analyte are present.

All these pieces of evidence lead to the need for alternative metrics and protocols for using SERS in real-world applications. In this context, chemometrics and machine learning can be useful tools to extract relevant qualitative and quantitative information with minimum sample preparation.

Multivariate data analysis: chemometrics and machine learning

In this section, we describe the fundamentals of multivariate data analysis, with an emphasis on chemometrics and machine learning methods that have already been implemented along with SERS. Although no sharp distinction can be assumed between chemometrics and machine learning, we adopt a terminology compatible with the literature: (1) we refer to chemometrics as those methods based on principal component and factor analysis, related to dimensionality reduction (e.g., principal component analysis, partial least squares, and multivariate curve resolution), and (2) we refer to machine learning as a set of advanced algorithms that learn from data and are applicable for non-linear/high complexity modelling (e.g., support vector machine, artificial neural networks, and random forest). Furthermore, most of the papers discussed in this review are focused on supervised methods, in which the datasets are labeled/tagged with actual answers/values (unlike unsupervised methods).

Data preprocessing

Preprocessing methods are useful to compensate for undesired external sources of signal fluctuations apart from the chemical components of the sample [22]. They can be applied on rows (objects/samples) or columns (variables/Raman shifts) of a data matrix. The most common artifacts in SERS are baseline deviations, fluorescence, misalignment, and noise. Despite centering and autoscaling being very popular preprocessing methods, autoscaling is not always recommended since it might give

equal weights for spectral bands of different intensities and increase noise. Savitzky-Golay smoothing is a well-known method for filtering noise and consists of a moving-average filter, in which a polynomial function is fitted [23]. Filtering noise should be carefully used because important chemical information can be lost when using a too-large window size (e.g., more than 15 points per window). Moreover, this might introduce correlations into the noise structure, leading to worst model predictions when compared to the use of raw data [24]. SERS data may also contain linear baseline deviations, evidenced by a constant offset of instrumental variations, and/or non-linear ones, such as drifts caused by other physical effects. In offset correction, the baseline is estimated, for example, by asymmetric least squares or penalized least squares [25], and then subtracted from the spectrum. Derivatives are also useful preprocessing methods to improve the visualization of spectral differences [26], but they might also increase the spectral noise. The first and second derivatives remove linear and non-linear sample-dependent background signals, respectively. Although derivatives change the spectral format and might difficult the interpretation of informative vectors (e.g., spectral loadings and regression coefficients), they can still be associated to the original spectra.

Principal component analysis (PCA)

PCA is the most popular unsupervised chemometric method for pattern recognition/exploratory analysis, and the basis of various chemometric models [22, 27]. In SERS, one often handles data matrices containing hundreds or thousands of Raman shifts. PCA decomposes the data matrix by means of linear combinations of correlated original variables (Raman shifts) to create a reduced number of variables (principal components, PCs) containing valuable vibrational information. Figure 3A describes this dimensionality reduction strategy. Hence, a

SERS dataset $X_{n,m}$ (n samples and m Raman shifts) can be expressed by matrices of scores $T_{n,A}$ and loadings $P_{m,A}$:

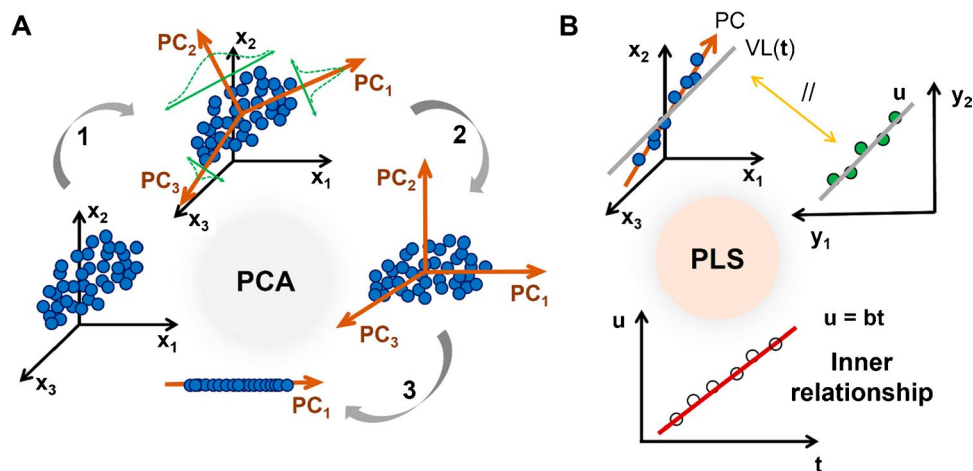
$$X = t_1 p_1^T + t_2 p_2^T + \dots + t_A p_A^T + E = TP^T + E \quad (4)$$

where A is the selected number of PCs, t_1 and p_1 are the score and loading vectors of the first PC, respectively, and E contains the residual variance not described by the model. Scores (T -matrix) provide the composition of each PC related to the samples, whereas loadings describe this composition in terms of variables (P -matrix). For SERS data, one can plot the loadings for each PC as a function of Raman shifts to correlate them with differences or similarities among samples visualized in the scores plot. Moreover, the PCs are modeled in descending order of explained variance, the columns of T are mutually orthogonal, and the columns of P are orthonormal, i.e., uncorrelated. The suitable number of PCs can be determined by assessing the explained/captured variance and the random behavior of the residuals. Occasionally, one can be interested in the information representing a small part of the data variance (e.g., 10% or even less), but should be careful not to include spectral noise information in the model. Outliers can be identified in PCA models by using Hotelling's T^2 statistics and high Q residues. While the Q statistic measures the unexplained variance of each sample (sum of the squares of the residual vectors), Hotelling's T^2 statistic represents the variation in each sample within the model [27]. For objective decisions in outlier detection, confidence limits for these two parameters are estimated at a certain probability level. Nevertheless, outliers should carefully be handled and interpreted.

Soft independent modelling of class analogy (SIMCA)

SIMCA is a linear class modelling method introduced by S. Wold in 1976 [28] and used for supervised authentication

Fig. 3 **A** Dimensionality reduction using PCA in three steps: (1) estimate of PCs, (2) subspace representation, and (3) projection on the PCs of interest. **B** Schematic representation of PLS by maximizing the correlation between X and Y matrices



and classification purposes. SIMCA may be considered as consisting of PCA models individually developed for predefined target classes with the best numbers of PCs selected by cross-validation. As a class modelling algorithm, SIMCA defines multivariate enclosed class spaces at a predefined confidence level. Class distances are estimated as statistical probabilities and SIMCA allows for a straightforward detection of outliers as samples not fitting to any of the modeled classes. Moreover, new samples can be assigned to more than one class when there is class overlapping. SIMCA is also suitable for one-class modelling, required for authentication purposes, mainly in food analysis. In this case, information from any non-targeted class should be avoided during modelling because it may introduce bias related to specific patterns of non-authenticity [29]. Thus, a rigorous one-class modelling considers only the target class information. Criteria for sample assignment in SIMCA are based on both Q and Hotelling's T^2 statistics, which are useful for outlier detection and were introduced from the multivariate statistical process control [30].

Partial least squares (PLS) and partial least squares discriminant analysis (PLS-DA)

PLS was introduced in the 1980s [31] and has become the most popular multivariate calibration method. PLS models can quantify analytes or properties in the presence of interferences, which are introduced in the calibration samples. PLS decomposes the dataset X in scores and loadings, but with the scores estimated in such a way as to maximize the correlation with a dependent variable y (or Y in the case of more than one). This is accomplished by the simultaneous decomposition of the two data blocks (see Fig. 3B). The factors in PLS are thus constrained and might deviate from orthogonality, giving origin to the so-called latent variables (LVs). An additional advantage of PLS is that a relationship between X and y may be detected even if the key variables have a small contribution to the first LV. The implementation of PLS in situations in which more than one dependent variable is predicted (Y), a.k.a. PLS2, is advisable only when there is a strong correlation between predicted variables [22]. The determination of the proper number of LVs is crucial for developing robust PLS models and is performed by internal cross-validation based on minimizing the root mean square error of cross-validation (RMSECV). A very small number of LVs may result in underfitting because important information/variance is left out of the model. On the other hand, a very large number of LVs may render an overfitted model unable to provide accurate predictions for external validation/test samples. An indicative of overfitting is a much larger root mean square error of prevision (RMSEP) than the root mean square error of calibration (RMSEC). Some inadequate practices include the biased choice of the

calibration/validation samples, the absence of an external validation set, or the expansion of the data set by incorporating replicates/virtual samples instead of true independent samples. An accurate PLS model should explain most of the variance of both X and y blocks.

PLS-DA is based on PLS regression and is suitable for discrimination/classification purposes [22]. Nevertheless, it is worth emphasizing that there is a key difference between PLS-DA and SIMCA. SIMCA is suitable for one-class modelling, when having a well-defined class, training a model, producing a prediction rule, and proceeding to include future samples in this specific class, or a universe of less well-defined target classes [32]. PLS-DA on the other hand is suitable for discrimination, when we have two or more well-defined classes, train a model, produce a discrimination rule, and proceed to assign future samples to any of these previously defined samples. In the PLS-DA model, the response vector y (or matrix Y) is constituted by values of 0 and 1 (or more depending on the number of classes) indicating whether a sample belongs or not belongs to a defined class. The suitable threshold value for the class separation can be calculated using Bayes' theory or the receiver operating characteristic curve (ROC).

Artificial neural networks (ANN)

Inspired by the human brain's capacity to perform very complex cognitive, perceptual, and control tasks, ANN was first proposed by McCulloch and Pitts in the 1940s [33]. ANN is a powerful mathematical tool used to accurately solve complex classification and regression problems [34]. ANN transforms a vector/SERS spectrum containing m input variables/Raman shifts, $x \in R^m$, into an output signal y using an activation/transfer function f , a threshold b , and the weight vector $w \in R^m$:

$$y = f(wx - b) = f\left(\sum_{i=1}^m w_i x_i - b\right) \begin{cases} 1, & \text{if } \sum_{i=1}^m w_i x_i - b \geq 0 \\ 0, & \text{if } \sum_{i=1}^m w_i x_i - b < 0 \end{cases} \quad (5)$$

The general idea of ANN is mathematical data processing, which is performed by the weighted sum and application of an activation function. Alternative functions can also be considered, for instance, the rectified linear unit. The ANN model estimates the weight vector w and the threshold value b , such as they separate the classes or calculate the actual concentration/property value correctly. The predictive accuracy can be optimized by minimizing an error function of desired/actual and predicted values for all the calibration/training samples. During the learning process, the error value is minimized by gradually updating the parameters w and b .

A simple artificial neuron (a.k.a. perceptron) is shown in Fig. 4A and can solve only a very limited number of linearly

separable problems. However, the use of specific activation functions, the backpropagation method, and more elaborated ANN architectures enable solving of highly complex problems. The basic architecture of an ANN consists of input, hidden, and output layers, and the number of hidden neurons directly affects the model performance [34]. For example, Fig. 4B shows a fully connected ANN with two hidden layers. A very large number of hidden neurons may ensure correct learning but poor predictions, whereas a very small number of them may imply a poor learning process or the error may not achieve an acceptable value. Modelling ANN with several hidden layers is sometimes required to solve highly complex qualitative and quantitative problems, thereby giving origin to deep learning. ANN architectures can be divided into four main groups [34]: (1) single-layer feedforward networks, (2) multilayer feedforward networks, (3) recurrent networks, and (4) mesh networks.

Support vector machine (SVM) and support vector regression (SVR)

SVM is a machine learning algorithm originally proposed by Vapnik and consists of finding a suitable hyperplane to maximize class separation [35]. Such hyperplane can be estimated using a set of n samples (e.g., one SERS spectrum per sample) and m measurable variables/Raman shifts. The dataset of n individual samples $x_i \in R^m$ with labels $y_i \in \{-1, +1\}$ is assumed to be separable by a hyperplane ($w x + b = 0$, b is a constant) using the decision rule:

$$f(x_i) \begin{cases} wx_i + b \geq 0 \text{ for } y_i = +1 \\ wx_i + b \leq 0 \text{ for } y_i = -1 \end{cases} \text{ or } y_i f(x_i) \geq 1 \quad (6)$$

A geometrical representation of SVM for linearly separable classes is shown in Fig. 5A. Assuming a separation margin M between the class boundary and the training samples, the objective of SVM is to find a hyperplane that maximizes M . This is an optimization problem that can be solved by using Lagrange functions. Additionally, in the case of not fully separable classes, a slack variable ξ is introduced:

$$y_i(w x_i + b) - 1 + \xi_i \geq 0 \text{ where } \xi_i \geq 0 \forall i \quad (7)$$

A SVM soft margin is built, and the misclassified data points have a penalty that increases with the distance. To minimize the number of misclassified points, a new parameter C is introduced to balance ξ and M . Hence, the objective is to minimize the Lagrangian with respect to w, b , and ξ considering the Lagrangian multipliers. The parameters are calculated, and the model can then be tested for new objects/samples.

In the case of SVR, the SVM algorithm is adapted to perform regression and it is necessary to predict a real output $y_i \in R$:

$$y_i = w x_i + b \quad (8)$$

SVR uses sophisticated penalty functions and there is no penalty if the predicted value \hat{y}_i is similar (tolerance value of ϵ) to the actual value $y_{ref,i}$ (i.e., $|\hat{y}_i - y_{ref,i}| < \epsilon$). The region $y_i \pm \epsilon$ is called the ϵ -insensitive tube and the output variables that are outside have a slack variable penalty above

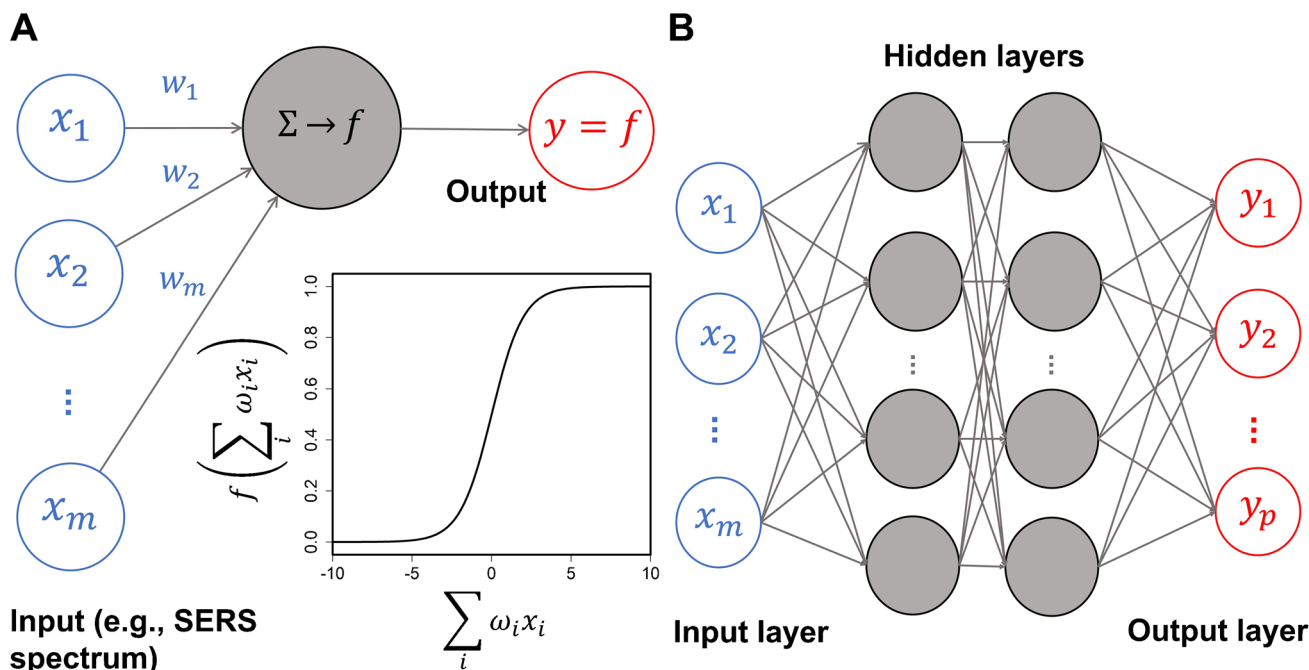


Fig. 4 A Representation of signal processing by ANN. B ANN architecture with multiple hidden layers, which is the basis of deep learning

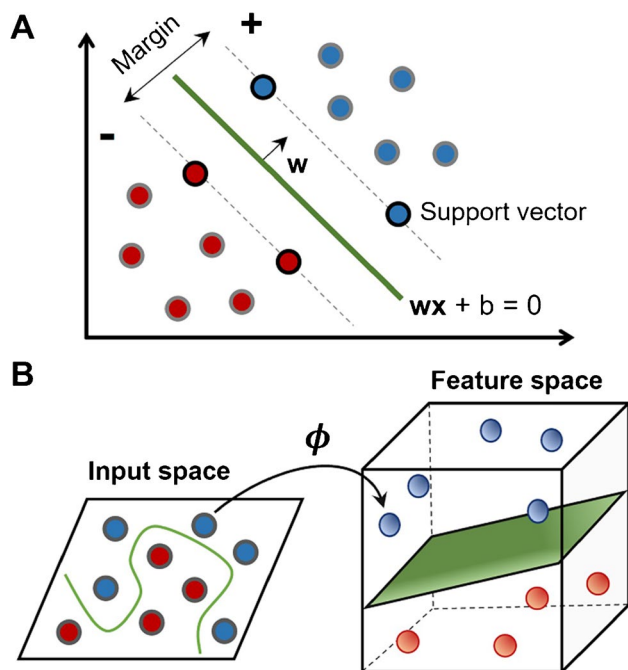


Fig. 5 **A** Basic elements in SVM for classification. **B** Geometrical representation of SVM for non-linearly separable classes

(ξ^+) or below (ξ^-) the tube ($\xi^+ > 0, \xi^- > 0 \forall i$). After calculating the parameters, new predictions are estimated.

In general, to apply SVM for non-linear data, a new dot product is defined by using a family of functions called kernel functions [36]. The direct calculation of the dot product $\phi(x_i)\phi(x_j)$ without the need for converting each object x to the higher dimension object $\phi(x)$ is called the kernel trick. It is very useful because several systems are not linearly separable in the original space (see Fig. 5B). Some of the most popular kernel functions are linear, polynomial, radial/Gaussian, and sigmoidal functions.

Analytical validation

Though analytical validation of multivariate methods is not thoroughly well-established, especially for quantitative models, it is essential to show that the developed method fulfills its purpose by estimating proper figures of merit (FOM). For qualitative methods [37, 38], FOM are based on evaluating predicted discrete categorical/dummy variables including the rates of false positive (FP) and false negative (FN) predictions. Important FOM are sensitivity (STR) rate, complementary to FN rate, and specificity (SPR) rate, complementary to FP rate. Predictive accuracy and F1-score are also important FOM for global comparison of qualitative methods. Although accuracy is a more intuitive value, the F1 score deals better with imbalanced datasets (classes with very different numbers of samples). ROC curves can

also be used to evaluate the classification model capability and the area under the curve (AUROC) is calculated for each class after plotting STR versus SPR rates. This approach searches for a threshold value that simultaneously minimizes the number of FP and FN predictions.

For multivariate calibration models [22, 39], it is necessary to corroborate whether the method has adequate FOM for its specific application. FOM evaluated for multivariate calibration methods include accuracy (trueness and precision), linearity, analytical sensitivity, confidence intervals, bias, residual prediction deviation, and limits of detection (LOD) and quantification (LOQ). Furthermore, any model must be rigorously validated by carefully and systematically using calibration and internal and external/test validation sets. Trueness indicates the agreement between reference and predicted values, and RMSEP is the main parameter to express it. Relative errors for calibration and validation samples, as well as precision, should also be estimated and mentioned. For quantitative multivariate methods, the fit is evaluated through the correlation coefficient (R) of a plot of reference/actual versus predicted values. To assess linearity, an R value close to one and the random behavior of the residuals should be verified by several tools, such as visual inspection ANOVA, or appropriate statistical tests (checking for normality, homoscedasticity, and independency).

Qualitative data analysis in SERS applications

Multivariate methods and qualitative analysis have established a vibrant relationship in the last years, overcoming quantitative applications in the literature [40]. Taking advantage of the high sensitivity of SERS, qualitative multivariate applications have focused on the discrimination and detection of pathogens [41–59], cancer studies and medical diagnosis [60–83], and food and forensic analysis [84–96]. To systematically validate the multivariate models, external sets of independent samples should be used based on clear, systematic, and representative criteria. Only internal cross-validation or permutation tests are not sufficient to establish a reliable validation, though it may seem tempting in some situations, providing a performance that can look better for a non-specialist at first sight [97]. Unfortunately, a significant number of papers have not adopted an external validation. Another key point often absent in the literature is the spectral interpretation of the model to understand the underlying chemical phenomena. Particularly for PLS-DA, the so-called informative vectors, such as regression coefficients and variable importance in projection (VIP) scores [98], contain important model information that should be interpreted. Authors using multivariate analysis should not be restricted to the question “what is there?” but rather should search

to answer other queries: “what is its/their relation to?” and “what is the difference between?” [99].

Discrimination and detection of pathogens

An example of proper PLS-DA model interpretation has been reported for the rapid discrimination of bacteria using filter paper-based SERS substrates [50]. PLS-DA was built with 180 samples (120 for training and 60 for validation), providing 100% of accuracy for discriminating genera. Uncertainties for model predictions were calculated by a resampling bootstrap strategy, and an in-deep model interpretation was carried out using VIP scores. This allowed the association of discriminant information to specific SERS bands of purine bases and the identification of possible metabolic pathways. Finally, a potentially new species of bacteria, not present in the training dataset, was correctly identified as an outlier using Q residues and Hotelling's T^2 statistic. The combination of SERS, PLS-DA, and suitable model validation to discriminate gram-positive and gram-negative bacteria [42] and sensitive and resistant *E. coli* strains [46] have also been reported. Spectral interpretations were performed using the loadings of the PCA models and PLS-DA results were discussed by comparing ROC curves, which is a very useful tool in qualitative analysis. Score plots of the first two LVs were used to show PLS-DA results, like in unsupervised PCA models. However, some authors have argued this is not the most reliable way to represent the results of supervised models [97, 100]. The interpretation of these scores can also be misleading if the number of variables far exceeds the number of samples, which is a common situation in SERS.

SERS analysis of bacteria using Au nanoparticles (NPs) oligomers and ANN has been reported for rapid antimicrobial susceptibility testing [52]. The authors proposed the use of a variational autoencoder, an interesting approach inspired by natural language processing that projected the data in a latent space to improve the visualization of the results. Analogously to PCA, this latent space represents the distribution of the samples and allows the identification of SERS fingerprints of the main metabolites. Following this idea, discrimination of multidrug-resistant bacteria has been performed by combining SERS and autoencoder-SVM (see Fig. 6) [41]. The autoencoder consisted of feedforward neural networks with non-linear activation functions to introduce the data in the SVM model. In contrast to the linear combination established in PCA-based dimensionality reduction, this method can directly adjust non-linear data. However, a much higher computational cost is required.

SERS and multivariate analysis have also been applied to the detection of viruses. Serum samples containing hepatitis B virus (HBV) from clinically diagnosed patients have been discriminated from controls with excellent accuracy (98%) and proper spectral interpretation of biomarkers [43]. A limitation of this method was the small number of samples analyzed (30 HBV-positive and 11 healthy individuals) that might compromise the reliability and robustness of the proposed PLS-DA model. Additional multivariate methods used along with SERS to discriminate and detect pathogens are highlighted in Table 1. This table includes, for example, the use of linear discriminant analysis (LDA) and decision trees (DT). In LDA, a linear function is calculated by maximizing and minimizing the interclass and intraclass variances,

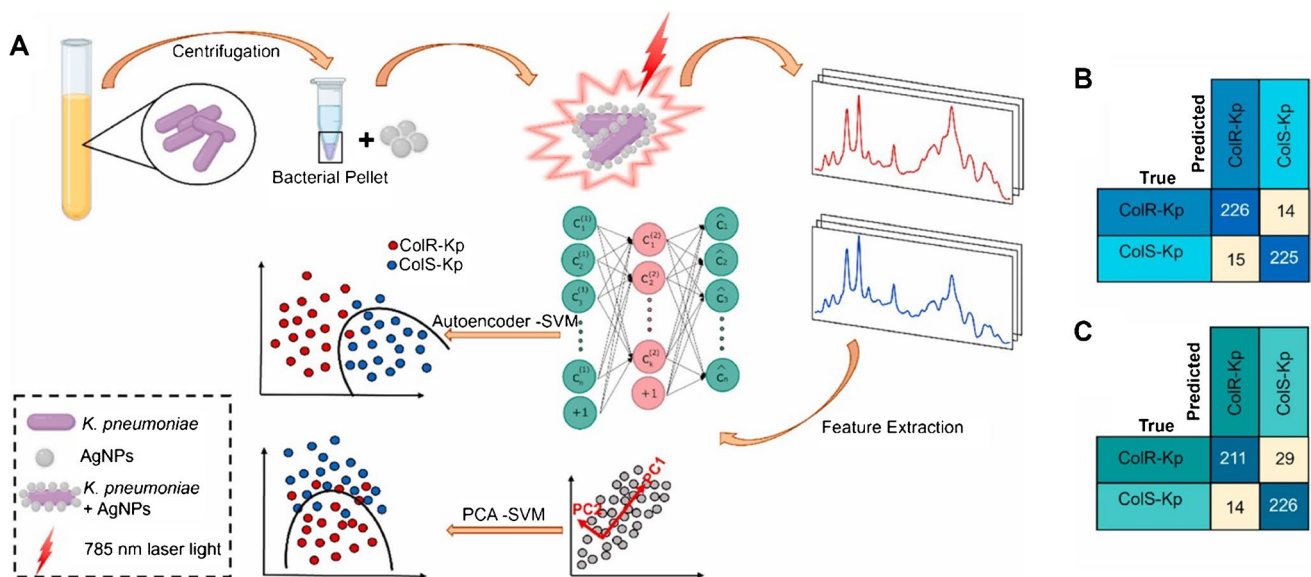


Fig. 6 A Schematic description of a SERS-based protocol for bacterial resistant assessment using SVM. Figures of merit for the discrimination analysis using **B** autoencoder-SVM and **C** PCA-SVM. Reprinted from [41], with permission from Elsevier

respectively, thus dividing the space into regions/classes. DT is a non-parametric supervised algorithm that uses a flowchart-like tree model and is constituted by a root node, internal nodes (tests), branches (outcomes of the tests), and leaf nodes (class labels) [101].

Cancer studies and medical diagnosis

Portability is an important trend in spectroscopy for rapid and in situ sample analysis. For example, incorporating portable methods in consumer devices, point-of-care tests, and smart appliances to become part of the internet of things [102]. As can be seen in Fig. 7A–C, a portable SERS-based breathalyzer has been designed for the rapid diagnosis of COVID-19 using PLS-DA [71]. Breath samples were collected for 501 participants, 15% of them COVID-positive, and the method displayed an excellent performance (sensitivity and specificity greater than 96%). It is worth mentioning that such an imbalanced ratio of classes may lead to reliability problems when the models are built using only accuracy as the response. The portable SERS device successfully detected breath volatile organic compounds, whose concentrations were affected by the immune response and metabolic changes of COVID-19 patients. Orthogonal PLS-DA (OPLS-DA) is a variant of PLS-DA widely used in metabolomics that allegedly provides better prediction ability and easier model interpretation [103], although this is not a consensus. OPLS-DA has been applied to detect cervical and breast cancer in serum by SERS, jointly with LDA and PCA-SVM [82]. The proposed methods presented excellent accuracies (in the range of 93–98%) and allowed for non-invasive cancer diagnosis.

One can also take advantage of the excellent predictive accuracy of ANN in medical diagnosis. By using ANN and ultrasensitive Ni-NiO SERS substrates, it has been possible to identify the type of cancer (lung, breast, and colorectal and brain cancer), and even the stage and location of brain cancer with accuracies of 100% and 96%, respectively (see Fig. 7D) [78]. Importantly, the liquid biopsy of a serum instead of conventionally invasive and time-consuming tissue biopsy highlights the capability of SERS to characterize trace levels of cancer biomarkers in complex matrices. Following a similar idea, breast, lung, and colorectal cancer have been discriminated with 100% accuracy by using ANN and a tridimensional SERS nanosensor. Although ANN provided better predictions than PLS-DA, associating discriminant results to chemically meaningful SERS bands is easier in PLS-DA models (e.g., by evaluating the VIP scores). A DNA-target functionalized surface has been combined with SERS in a decision system for detecting a specific oligonucleotide sequence identical to a gene fragment responsible for β -lactam antibiotic resistance [72]. The decision system was composed of a Siamese neural network (a class of neural network architectures containing two or more identical subnetworks) coupled with robust

statistics and Bayes decision theory. This strategy was complex but very sensitive, being able to detect target oligonucleotides at a picomolar level.

SVM has also shown excellent performance in medical diagnosis using SERS. SVM-based applications have included the diagnosis of lung and colon [74], and colorectal cancer [75], with accuracies greater than 80%. In the latter application, differences between the mean SERS spectra of healthy individuals and colorectal cancer patients were assessed by PCA before supervised modelling. The preliminary PCA approach allowed the identification of differences in the SERS spectra of the classes by evaluating scores and loading plots. It is worth empathizing that, in addition to the advisable unsupervised exploratory analysis (e.g., by PCA), the correlation between the discriminant results of the supervised model (e.g., SVM) and SERS bands should be performed to validate the method. A comparison of SVM with other machine learning algorithms has been performed for the diagnosis/prognosis of coronary artery disease at different stages with 92% of overall accuracy [66]. The authors used a Ti/Au SERS substrate to enhance the signal of small extracellular vesicles (EVs) from plasma samples, a new trend in liquid biopsy-based diagnosis. The use of simple and efficient supervised methods for discriminant analysis, such as LDA and decision trees, is a parsimonious alternative to SVM and ANN. For example, the comparison among SVM, RF, DT, and LDA has indicated the superior accuracy of LDA for detecting breast cancer in serum (83%), whereas DT presented the best accuracy in urine (89%) [81]. Importantly, the authors found that the SERS spectra of serum and urine are dominated by the purine metabolites, and not sugars, lipids, or coenzymes, as previously suggested in the literature.

SVM, RF, and convolutional neural networks have been compared in cancer studies using a magnetically assisted sandwich assay with ultrasensitive SERS nanotags made of Au@Ag cages [70]. Convolutional neural networks displayed the highest predictive accuracy and were capable to recognize specific miRNA cancer biomarkers extracted from tissue samples for potential head and neck cancer diagnosis. Although RF displayed a lower predictive accuracy than SVM and ANN, it provided a lower risk of overfitting. SERS has also been employed to differentiate and characterize EVs derived from pancreatic tissue and bone marrow using machine learning [68]. Logistic regression provided better sensitivity (89%) and specificity (88%) than SVM, random forest (RF), and naïve Bayes and CN2 rule induction. Logistic regression is a linear method that estimates class membership probability, whereas RF (an algorithm that combines several DT) and naïve Bayes (a family of classifiers based on Bayes' theory) are non-linear. Moreover, naïve Bayes works better for an uncorrelated and small number of variables, which is not the case with SERS data. Though no external validation was performed, the authors suggested the development of a platform

Table 1 Complementary recent and relevant applications combining SERS and qualitative multivariate methods

Scientific field	Multivariate methods*	Analytical scope and highlights	Maximum accuracy (%)	Reference
Discrimination and detection of pathogens	SVM, LDA, QDA	Lectin-modified cellulose nanocrystals and Au NPs to differentiate 19 bacterial strains	88	[45]
	Least squares SVM, PCA-KNN	Colloidal Ag NPs to discriminate bacteria with phylogenetic analysis as reference method	100	[48]
	KNN, SVM, DT, naïve Bayes	Discrimination of antibiotic-resistant and susceptible bacteria using colloidal Ag NPs	98	[49]
	SVM, naïve Bayes, KNN, logistic regression	2D arrays of Au NPs to discriminate wild and mutant cyanobacteria	97	[51]
	ANN, SVM, DT, KNN, LDA	Direct analysis of methicillin-resistant and sensitive bacteria using colloidal Ag NPs	98	[54]
	LDA	Ultrasensitive S-protein expressed coronavirus sensor based on Au nanoneedles array trap	100	[57]
	SVM, ANN, KNN	Odor SERS sensor based on self-assembled Au NPs for <i>Escherichia coli</i> detection	92	[58]
Cancer studies and medical diagnosis	Lasso-DA, PLS-DA, PCA-LDA	Diagnosis of an immune-related complication using a capillary tube SERS substrate (AUROC value of 0.908)	-	[61]
	PLS-DA	Discriminating drug-resistant and susceptible tuberculosis patients by a PCR-SERS assay	95	[62]
	SVM, XGBoost, logistic regression, DT, RF	Spectral interpretation and discrimination of receptor-binding domains of SARS-CoV-2 spike proteins	99	[63]
	PLS-DA, SIMCA	Si/Ag SERS platforms for the diagnosis of sexually transmitted diseases	100	[65]
	PCA-SVM	Papers-based SERS substrate for potential point-of-care diagnosis of prenatal diseases	93	[69]
	ANN, PLS-DA	3D titanium nanosensor to differentiate cancer stem cells EVs by liquid biopsy	100	[76]
	SVM, LDA	Gastric cancer detection using Au nanopyramids and tissue, blood, and saliva EVs	90	[79]
PLS-DA	SERS provided greater AUROC values than Raman spectroscopy (0.94 vs 0.83) for breast cancer diagnosis in serum	-	[80]	
Food and forensic analysis	KNN, DT, SVM, AdaBoost	Discrimination of pesticides using filter paper coated with Ag dendritic nanostructures	92	[84]
	SVM	Self-assembled Ag nanocubes modified with capture agents for discrimination of wine flavors	100	[87]
	PCA-SVM	Colloidal Ag NPs as SERS substrate to detect a psychedelic drug in wines	100	[91]
	DD-SIMCA	2D-correlation SERS discrimination of structural changes in edible bird's nest samples	100	[95]

*The supervised qualitative method with the best accuracy is placed first

for detecting cancer-derived EVs. Additional complementary papers using multivariate models in cancer studies and medical diagnosis are highlighted in Table 1.

Food and forensic analysis

Another important field of SERS is food analysis, with the potential to implement these methods in portable sensors for

real-life end-user applications. Although discriminant classification models (PLS-DA and LDA) have been commonly used, class modelling by SIMCA and its variants were also implemented. This is in connection with the controversy in the literature, mainly in food application, about discriminant and one-class modelling [32, 104]. SERS and data-driven-SIMCA (DD-SIMCA) have also been combined to detect the pesticide thiabendazole in water extracts of mango peels

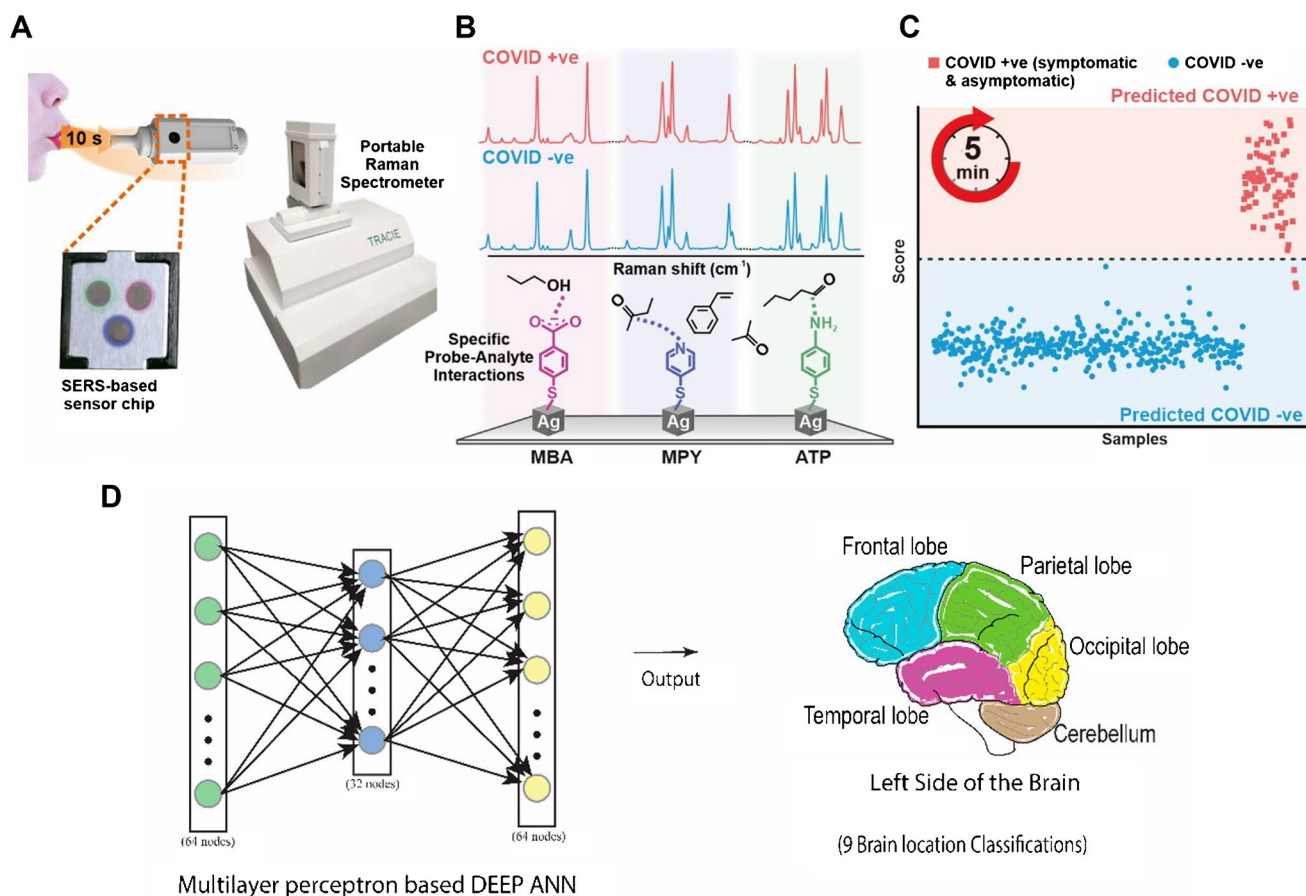


Fig. 7 **A** SERS-based sensor for breath analysis, **B** detection mechanism by surface modification, and **C** discriminant analysis of healthy and infected individuals. **D** Deep learning architecture used for SERS

diagnostic of brain cancer location. Figures reproduced with permission. Adapted with permission from [71] and [72]. Copyright 2022 American Chemical Society

using Au NPs-coated paper as substrate [90]. The one-class DD-SIMCA model was rigorously validated by incorporating an outlier detection step and showed good selectivity (95%) and sensitivity (92%) in classifying samples according to the limit established by the Brazilian Health Regulatory Agency. SIMCA and SERS have also been applied in forensic analysis in which a preliminary liquid–liquid extraction step compensated for the matrix effect in the detection of clonazepam in beer, juice, and energy drink samples [94]. The estimated sensibility and efficiency were 92 and 96%, respectively. In another forensic application, eleven anti-inflammatory and analgesic drugs typically used in China as adulterants in anti-rheumatic health foods have been detected by thin-layer chromatography coupled with portable SERS and hybrid multivariate methods (PC-LDA, PCA-KNN, and PCA-SVM) [89]. The PC-LDA model presented the best performance in terms of predictive accuracy (100%). However, replicated samples were used in the construction of discriminant models, which might lead to biased results, and the chemometric discussion was not extensive.

Residues of four different pesticides have been detected in tea by SERS and a one-dimensional convolution neural network (1D-CNN) [85]. The 1D-CNN model showed a superior predictive accuracy (100%) in comparison to SVM, PLS-DA, and KNN. However, an in-deep model interpretation was not presented, probably because of the limited capacity of 1D-CNN to associate outputs with chemically meaningful variables. Alternatively, the trade-off between computational cost and interpretability of hybrid methods (e.g., PCA-ANN or PCA-SVM) may help to partially compensate for such limitations. For example, urine samples have been analyzed to detect drugs by combining SERS and various architectures of ANN (deep learning networks) with a maximum accuracy of 98% [96]. Although CNN displayed a greater predictive accuracy than PCA-ANN hybrid model, the latter provided an easier chemical interpretation of the results. Additional papers about the use of SERS and multivariate methods in food and forensic analysis are highlighted in Table 1.

In general, machine learning methods display excellent performances to solve classification problems. Nevertheless,

the criteria and real need to use them instead of simpler chemometric methods are still unclear and, in most of the cases, not systematically justified. Therefore, investigations beyond conventional comparison of predictive accuracy will help to clarify these points.

Quantitative data analysis in SERS applications

Currently, one of the main challenges in SERS applications is performing reliable and reproducible quantitative analysis [3]. In the univariate strategy, the height of a SERS band or the ratio between two bands is used; nevertheless, this is limited by the need for identifying and isolating pure signals. When univariate analysis fails, multivariate analysis arises as an excellent alternative to circumvent this issue [105]. Quantitative SERS methods using multivariate analysis have been mainly focused on the determination of organic compounds, including pesticides [106–116], drugs [96, 117–125], biomolecules [19, 126–130], and metallic ions [131–137].

PLS has been the most widely used multivariate calibration method in the literature for pesticide residue detection, probably because of its simplicity, speed, good performance, and easy accessibility. Moreover, some modified PLS regression models are available for non-linear calibration purposes. In a recent publication, dynamic SERS was employed to detect acephate, an organophosphate pesticide, in rice samples [115]. Measurements were collected during the transition from wet to dry state of colloidal Au nanorods modified with cysteamine dropped on a silicon chip. PLS, RF, and SVM algorithms were tested for building multivariate calibration models, with 82 samples randomly split into calibration and internal and external validation sets. Alternatively, systematic methods to split the most representative samples are also advisable [138]. PLS displayed the best performance, evidenced by the lowest RMSEP value. Therefore, machine learning/non-linear algorithms should not always be the first option for data analysis. As a good practice, preference should be given to the most straightforward and parsimonious model. Furthermore, linearity should preliminarily be evaluated in the concentration range of interest, for example, by using ANOVA and plotting the X scores against y .

The non-linear relationship between the SERS intensity and thiabendazole pesticide concentration has been demonstrated by Li et al. [108], as the signal depends on the number of nanoparticle clusters and molecules captured in the hot spots. Taking this into consideration, the self-assembly of Au NPs by supramolecular host molecules (acting as bridges) was employed to selectively capture methylxanthine compounds [120]. The advantage of this approach is the generation of hot spots with selective affinity for molecules with similar

chemical structures (e.g., theobromine, theophylline, and caffeine). Quantitative analyses were performed by using a non-linear ANN model with excellent predictive accuracy (in terms of RMSE) and LODs down to 50 nmol/L. A bootstrapping random resampling procedure with 1000 iterations was performed to partially compensate for the small number of calibration samples. Machine learning and SERS have also been implemented for indirect monitoring of pH by measurement changes in the SERS spectra of 4-mercaptopyridine (see Fig. 8) [139]. As non-linear dependence of the pH against the analytical SERS signal was observed, SVM, regression trees, and the Gaussian regression process were assessed, with the latest displaying the best performance (lowest RMSE). Nevertheless, the number of samples analyzed was relatively small, which is a limitation for method generalization.

As an alternative to PLS, multivariate curve resolution with alternating least squares (MCR-ALS) has also been used in quantitative SERS applications [19, 110, 126, 128]. MCR-ALS is a factor analysis method that aims at solving mixture signals, thus providing information from the components of the sample [140]. MCR-ALS decomposes the data matrix (X) into scores associated to relative SERS intensities and loadings associated to the pure SERS spectra. Combining MCR-ALS and SERS results in versatile and chemically meaningful pseudo-univariate models, which are compatible with the standard addition strategy and require a small number of samples for calibration [129]. Additionally, similar predictive accuracies in terms of RMSE have been reported for MCR-ALS and PLS-DA models built with SERS data [19].

Variable selection is a valuable strategy to improve the model performance by removing uninformative/interfering variables. In a recent SERS application, genetic algorithm (GA), competitive adaptive reweighted sampling (CARS), and ant colony optimization (ACO) were combined with PLS to determine the pesticide thiabendazole in citrus using Au nanorods [109]. The GA-PLS model showed better performance (in terms of RMSEP) than CARS-PLS, ACO-PLS, and full spectra PLS model, and the LOD was 0.33 $\mu\text{g}/\text{mL}$. The coffee-ring effect was proposed to build label-free SERS methods to quantify patulin and alternariol, primary mycotoxin contaminants, in fruits (see Fig. 9A) [118]. The coffee-ring effect is mostly employed to preconcentrate the analyte and nanoparticles at the edge of the ring by capillarity without any external force [141, 142]. Synergy interval-PLS (Si-PLS), GA-PLS, and uninformative variable elimination-PLS (UVE-PLS) models were evaluated. Si-PLS and GA-PLS yielded the best performances (greatest R values) for patulin and alternariol, respectively. Note that there is not an absolute best model for selecting variables, as this will depend on the interaction between the method and the data properties. Highlights of additional interesting SERS applications using multivariate and variable selection methods are shown in Table 2.

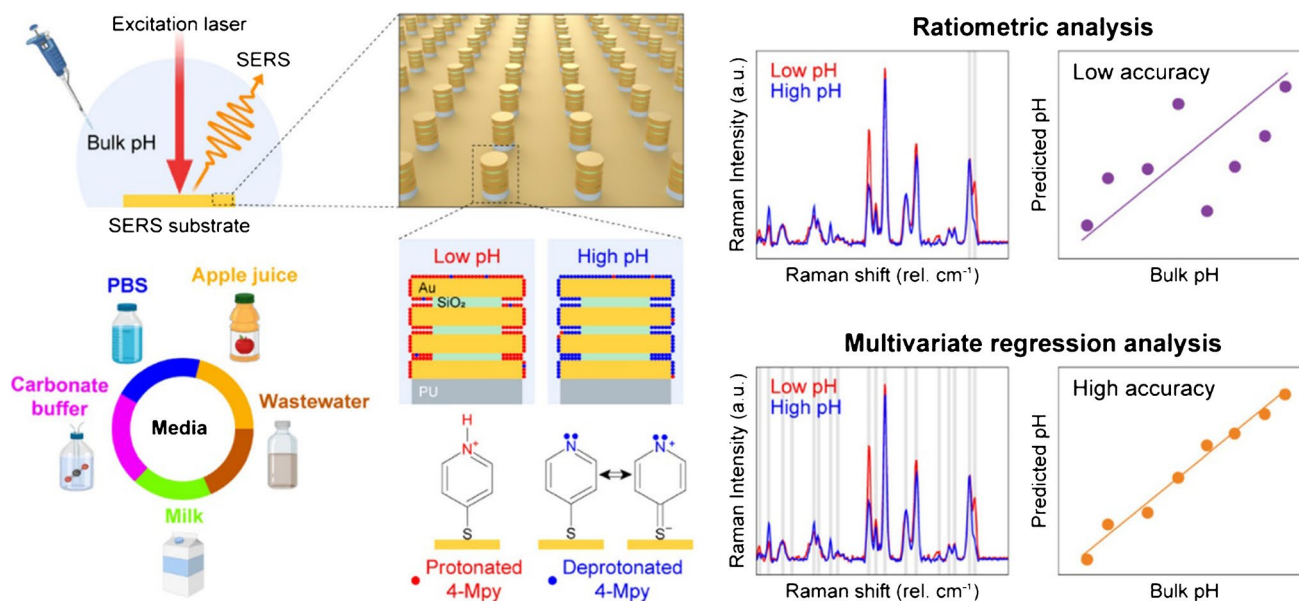


Fig. 8 Schematic representation of the pH sensing by the modified SERS substrate and the use of non-linear calibration strategy. Adapted with permission from [139]. Copyright 2022 American Chemical Society

A methodology employing thin-layer chromatography with surface-enhanced Raman scattering (TLC-SERS) has been proposed to quantify histamine, a seafood allergen, in tuna samples [125]. TLC plates coated with gold nanoparticles served simultaneously as SERS substrate and as the stationary phase to separate histamine from the sample matrix. A PCA-SVR hybrid model showed a better performance than PLS, demonstrated by the higher R value and lower RMSEP in the concentration range of interest. Advanced mathematical modelling based on machine learning algorithms seems to fit better a complex non-linear SERS signal behavior. However, it should not be generalized because an in-depth discussion is necessary to provide systematic criteria to select the suitable machine learning method.

SERS has also found applicability in the quantification of metallic ions, present in drinking water and contaminated food. Indirect SERS detection strategies have mainly been adopted for this purpose. For instance, the Au(III) reduction to produce Au NPs in the presence of aptamers and graphene oxide allowed the indirect detection of Pb^{2+} ions in black tea. SERS signals were monitored using the Raman reporter 4-mercaptobenzoic acid and PLS method with variable selection (CARS-PLS, siPLS, GA-PLS, and Si-GA-PLS) and were compared [135]. In terms of R values, the CARS-PLS model provided the best performance and a LOD of $0.1 \mu\text{g/L}$. RMSEP values were in some cases significantly lower than the RMSEC values, which might indicate a biased choice of validation samples. Interestingly, some reports have reinforced the excellent performance of CARS-PLS for the label-free determination of Cd^{2+} ions

[132] and total arsenic [131] in tea samples with LODs of $23 \mu\text{g/L}$ and 27ng/g , respectively.

Although several recently published articles have combined PLS with variable selection methods, some important aspects have not been systematically discussed. For example, the criteria for selecting a suitable number of LVs, which may lead to building underfitted or overfitted models. In a recent publication, a SERS method for quantitatively detecting Zn^{2+} ions at nmol/L level using a complexometric indicator was reported (Fig. 9B–C) [134]. In this work, the model was properly built using PLS with 10 VLs, determined by leave-one-out cross-validation. Additionally, the authors followed the ASTM recommendations, and samples were split into about two-thirds for the calibration set and one-third for the validation set. Other important aspects often absent in papers using SERS and multivariate calibration models are spectral interpretation and outlier detection. Informative vectors generated for PLS models should be critically interpreted and associated with the respective spectral bands, and the identification of outliers is crucial for the development of calibration models with good prediction ability.

Benchmarking and considerations for method selection

The main characteristics of the chemometric and machine learning methods successfully incorporated in SERS applications are summarized in Table 3. Note that non-supervised models, such as PCA, were not included because they should

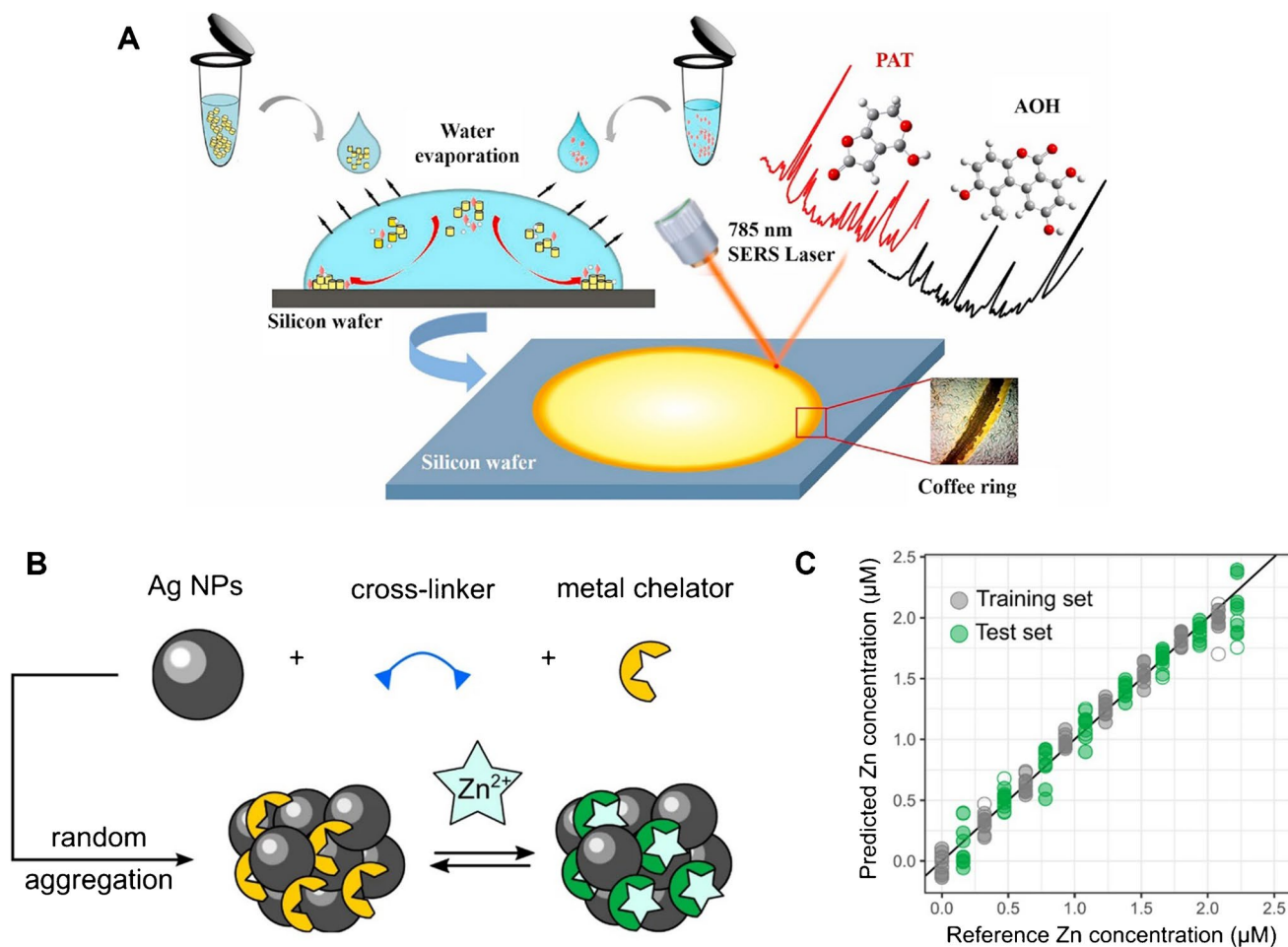


Fig. 9 **A** Utilization of the coffee-ring effect in SERS detection of mycotoxins. **B** Surface modification of AuNPs for metal ion detection and **C** comparison of actual and predicted concentrations for the

quantitative PLS model. Reprinted from references [118] and [134], with permission from Elsevier

Table 2 Complementary recent and relevant applications combining SERS and quantitative multivariate methods

Analytical determination	SERS substrate	Multivariate methods*	LOD	Reference
2,4-Dichlorophenoxyacetic acid in green tea	Paper coated with Ag NPs	UVE-PLS	0.1 ng/g	[106]
Flusilazole in oolong tea	Ag NP-coated cellulose nanofiber	PLS	0.5 mg/kg	[111]
Carbendazim in tea and rice	Ti ₂ C MXene/Au-Ag nanoshuttles	Relevance vector machine, SVM, ANN	10 nmol/L	[112]
Acetamiprid in green tea	Graphene oxide-Au nanostar composite	GA-PLS, PLS	20 ng/L	[114]
Ganciclovir, penciclovir, and valacyclovir in rat plasma and tablets	Colloidal Ag NPs	ANN, PLS	10 μmol/L	[119]
Chloramphenicol in pork	Paper coated with flower-like Ag NPs	CARS-PLS, UVE-PLS, ACO-PLS, PLS	10 ng/L	[121]
Norfloxacin in fish	Bimetallic Au@Ag NPs	GA-PLS	20 ng/L	[123]
Hg ²⁺ in fish and water samples	Au@Ag NPs modified with R6G	GA-PLS, successive projection algorithm-PLS, ACO-PLS	1 ng/g	[133]
Pb ²⁺ in water samples	NaYF ₄ @Yb,Ho,Au nanohybrid material	GA-PLS	1 μg/L	[136]

*The supervised quantitative method with the best performance is placed first

always be performed before supervised modelling. The first aspect to consider is the dimensionality of the SERS data, where the number of variables/Raman shifts frequently exceeds the number of samples. Therefore, models of low variance and dimensionality reduction are good options (e.g., PLS, PLS-DA, SIMCA, linear SVM, and MCR-ALS). On the other hand, if the number of samples is similar or higher than the number of variables/Raman shifts, high variance models may be a suitable first choice (e.g., RF, kernel SVM, SVR, and ANN). Moreover, one can consider the following key criteria: (1) predictive accuracy, (2) simplicity/parsimony, (3) interpretability, (4) linearity, and (5) computational cost. For example, the diagnosis of diseases such as cancer and COVID-19 demands highly accurate results because errors could have a dramatic impact on the population. Here, the use of flexible ANN architectures would be a good choice to fulfil such a requirement, despite the complexity (less parsimony), the poor ability to explain what is behind the outputs, and the high computational cost. Obtaining excellent accuracy is always desired; nevertheless, it is just one criterion to decide whether a model is useful or not to solve a particular problem. Accordingly, the study of potential biomarkers or biological pathways underlying medical conditions by SERS demands interpretable models (e.g., PLS-DA and RF). In this sense, great efforts are being made to improve the interpretability of machine learning algorithms aiming to change their “black box” nature (e.g., ANN and SVM) [143, 144]. In food analysis, supervised classification and authentication problems have been usually faced. Classification and discrimination analysis can be performed by using PLS-DA, LDA, or KNN with good predictive accuracies. On the other hand, rigorous authentication demands the use of SIMCA or other one-class modelling approaches, in which the authentic class is modeled without including the adulterated/non-authentic class information in the training set [29, 32]. This is because the adulterated samples can dramatically change over time and an infinite number of new types of adulteration may occur. Although it has not yet been applied in the SERS field, the use of machine learning for food authentication (e.g., one-class SVM and RF) has gained attention in recent spectroscopic applications [145, 146].

As can be seen in Table 3, SVM/SVR and PLS/PLS-DA have been the most used methods in SERS applications, probably because of their good predictive accuracies and availability in several software packages. PLS/PLS-DA outperforms SVM/SVR in terms of interpretability and simplicity/parsimony. However, the non-linear behavior of the SERS signals may negatively affect the performance of

PLS models and restrict their applicability to narrow linear ranges. It is therefore advisable to test the linearity in the concentration range of interest before going through complex SVR or ANN modelling, which is less parsimonious, reduces the interpretability, and increases the computational cost. Finally, after choosing the suitable multivariate method, only a rigorous validation step will assure the robustness and generalization of the SERS-multivariate method in real-world applications.

Conclusions and outlook

Recent works implementing chemometrics and machine learning in SERS have been described and discussed in this review. These multivariate techniques allow the performing of SERS analysis with excellent predictive accuracy, especially when using ANN. Most of the qualitative applications have been focused on the characterization of pathogens, human health care, and food and forensic analysis. Complex machine learning models with limited interpretability and high computational cost should be avoided to solve simple discrimination/classification problems (principle of parsimony). For example, PLS-DA and SIMCA are interpretable, and good options for classification and authentication, respectively, despite they might provide lower prediction accuracy for complex SERS data. In quantitative applications, most of the published SERS applications have focused on monitoring pesticides, drugs, biomolecules, and metallic ions. Methods based on PLS, SVR, and ANN provide excellent performances in terms of predictive accuracy; nevertheless, the linearity should previously be assessed for choosing a suitable/parsimonious multivariate method. Although RF is not yet frequently used in SERS, it combines important advantages, such as good prediction accuracy and interpretability. Future improvements in the interpretability of accurate machine learning algorithms, such as SVM and ANN, will certainly improve the quality of future smart SERS sensors. Moreover, the rigorous validation of multivariate models, by estimating FOM and by using external validation/test sets with independent samples, should become a general practice because it will guarantee the robustness and reproducibility of SERS methods in real-world applications. Therefore, combining SERS and multivariate analysis significantly improves selectivity and allows for extracting crucial spectroscopic information, albeit sensing of metallic ions or molecules with very low affinity for metallic surfaces is still limited and needs preliminary surface modifications.

Table 3 Comparison of multivariate supervised methods for SERS data analysis

Method	Advantages	Limitations	References
SIMCA	Simplicity, suitability for authentication, and reasonable interpretability	Discriminant information between classes is neglected	[56, 90, 93–95]
LDA	Simplicity	Restricted to linearly separable classes	[47, 54, 57, 60, 61, 64, 72, 77, 79, 81–83, 88, 89, 117]
MCR-ALS	Chemically meaningful results (pure spectra) and versatile constraints	Restricted working range and rotational ambiguity	[19, 77, 110, 126, 128, 129]
KNN*	Simplicity and easily optimized	Sensitive to irrelevant variables and not able to detect outliers or to identify discriminant variables	[47–49, 51, 53, 54, 58–60, 66, 84, 85, 88, 89, 96, 109, 117]
PLS-DA/PLS*	Dimensionality reduction and good predictive accuracy and interpretability	Restricted working range and not appropriate for authentication	[19, 42–44, 46, 50, 55, 56, 61, 62, 65, 70, 71, 73, 74, 76, 80, 82, 85, 96, 106, 107, 109, 111, 113–125, 127, 131–137]
ANN*	Excellent predictive accuracy, versatile architectures, and wide working range	Low interpretability, high complexity, many parameters to tune, prone to overfitting, and computationally expensive	[47, 52–54, 58, 66, 70, 72, 76, 78, 85, 96, 112, 115, 119, 120, 127]
SVM/SVR*	Good predictive accuracy, effective in high dimensional spaces, and wide working range	Low interpretability, high complexity, prone to overfitting, and computationally expensive	[41, 45, 47–49, 51–54, 58–60, 63, 66–70, 74, 75, 79, 81, 82, 84–87, 89, 91, 92, 96, 109, 112, 115, 125, 127, 130, 139]
DT/RF*	Good predictive accuracy and interpretability, wide working range, and low risk of overfitting	High complexity and computationally expensive	[47, 49, 53, 54, 59, 63, 68, 70, 81, 84, 85, 96, 115, 127, 139]

*Applicable for both classification and regression

Declarations

Conflict of interest The authors declare no competing interests.

References

- Baig N, Kammakam I, Falath W. Nanomaterials: a review of synthesis methods, properties, recent progress, and challenges. *Mater Adv.* 2021;2:1821–71. <https://doi.org/10.1039/D0MA00807A>.
- Zhou J, Yang T, Chen J, Wang C, Zhang H, Shao Y. Two-dimensional nanomaterial-based plasmonic sensing applications: advances and challenges. *Coord Chem Rev.* 2020;410:213218. <https://doi.org/10.1016/j.ccr.2020.213218>.
- Langer J, Jimenez de Aberasturi D, Aizpurua J, Alvarez-Puebla RA, Auguie B, Baumberg JJ, Bazan GC, Bell SEJ, Boisen A, Brolo AG, Choo J, Cialla-May D, Deckert V, Fabris L, Faulds K, Garcia de Abajo FJ, Goodacre R, Graham D, Haes AJ, Haynes CL, Huck C, Itoh T, Käll M, Kneipp J, Kotov NA, Kuang H, le Ru EC, Lee HK, Li J-F, Ling XY, Maier SA, Mayerhöfer T, Moskovits M, Murakoshi K, Nam J-M, Nie S, Ozaki Y, Pastoriza-Santos I, Perez-Juste J, Popp J, Pucci A, Reich S, Ren B, Schatz GC, Shegai T, Schlücker S, Tay L-L, Thomas KG, Tian Z-Q, van Duyn RP, Vo-Dinh T, Wang Y, Willets KA, Xu C, Xu H, Xu Y, Yamamoto YS, Zhao B, Liz-Marzán LM. Present and future of surface-enhanced Raman scattering. *ACS Nano.* 2020;14:28–117. <https://doi.org/10.1021/acsnano.9b04224>.
- Sharma B, Bugga P, Madison LR, Henry A-I, Blaber MG, Greenelth NG, Chiang N, Mrksich M, Schatz GC, van Duyn RP. Bisboronic acids for selective, physiologically relevant direct glucose sensing with surface-enhanced Raman spectroscopy. *J Am Chem Soc.* 2016;138:13952–9. <https://doi.org/10.1021/jacs.6b07331>.
- Villa JEL, Garcia I, Jimenez de Aberasturi D, Pavlov V, Sotomayor MDPT, Liz-Marzán LM. SERS-based immunoassay for monitoring cortisol-related disorders. *Biosens Bioelectron.* 2020;165:112418. <https://doi.org/10.1016/j.bios.2020.112418>.
- Le Ru EC, Etchegoin PG. Introduction to plasmons and plasmonics. In: *Principles of Surface-Enhanced Raman Spectroscopy*. Elsevier Science; 2009. pp. 121–183. <https://doi.org/10.1016/B978-0-444-52779-0.X0001-3>.
- Jensen L, Aikens CM, Schatz GC. Electronic structure methods for studying surface-enhanced Raman scattering. *Chem Soc Rev.* 2008;37:1061. <https://doi.org/10.1039/b706023h>.
- Lombardi JR, Birke RL. A unified approach to surface-enhanced Raman spectroscopy. *The Journal of Physical Chemistry C.* 2008;112:5605–17. <https://doi.org/10.1021/jp800167v>.
- Xu Y. Electromagnetic scattering by an aggregate of spheres. *Appl Opt.* 1995;34:4573. <https://doi.org/10.1364/AO.34.004573>.
- Xu Y. Electromagnetic scattering by an aggregate of spheres: far field. *Appl Opt.* 1997;36:9496. <https://doi.org/10.1364/AO.36.009496>.
- Le Ru EC, Etchegoin PG. Single-molecule surface-enhanced Raman spectroscopy. *Annu Rev Phys Chem.* 2012;63:65–87. <https://doi.org/10.1146/annurev-physchem-032511-143757>.
- Xia Y, Xiong Y, Lim B, Skrabalak SE. Shape-controlled synthesis of metal nanocrystals: simple chemistry meets complex physics? *Angew Chem Int Ed.* 2009;48:60–103. <https://doi.org/10.1002/anie.200802248>.
- Bär J, de Barros A, de Camargo DHS, Pereira MP, Mercés L, Shimizu FM, Sigoli FA, Bufon CCB, Mazali IO. Silicon microchannel-driven Raman scattering enhancement to improve gold nanorod functions as a SERS substrate toward single-molecule detection. *ACS Appl Mater Interfaces.* 2021;13:36482–91. <https://doi.org/10.1021/acsaami.1c08480>.
- de Barros A, Shimizu FM, de Oliveira CS, Sigoli FA, dos Santos DP, Mazali IO. Dynamic behavior of surface-enhanced Raman spectra for rhodamine 6G interacting with gold nanorods: implication for analyses under wet versus dry conditions. *ACS Appl Nano Mater.* 2020;3:8138–47. <https://doi.org/10.1021/acsnm.0c01530>.
- Santinom A, da Silva MA, Villa JEL, Poppi RJ, Mazali IO, dos Santos DP. Surface-enhanced Raman scattering (SERS) as probe of plasmonic near-field resonances. *Vib Spectrosc.* 2018;99:34–43. <https://doi.org/10.1016/j.vibspec.2018.07.007>.
- Lee S, Sim K, Moon SY, Choi J, Jeon Y, Nam J, Park S. Controlled assembly of plasmonic nanoparticles: from static to dynamic nanostructures. *Adv Mater.* 2021;33:2007668. <https://doi.org/10.1002/adma.202007668>.
- Lee A, Ahmed A, dos Santos DP, Coombs N, Park J, il, Gordon R, Brolo AG, Kumacheva E. Side-by-side assembly of gold nanorods reduces ensemble-averaged SERS intensity. *The Journal of Physical Chemistry C.* 2012;116:5538–45. <https://doi.org/10.1021/jp3010597>.
- Camargo PHC, Rycenga M, Au L, Xia Y. Isolating and probing the hot spot formed between two silver nanocubes. *Angew Chem Int Ed.* 2009;48:2180–4. <https://doi.org/10.1002/anie.200806139>.
- Villa JEL, Afonso MAS, dos Santos DP, Mercadal PA, Coronado EA, Poppi RJ. Colloidal gold clusters formation and chemometrics for direct SERS determination of bioanalytes in complex media. *Spectrochim Acta A Mol Biomol Spectrosc.* 2020;224:117380. <https://doi.org/10.1016/j.saa.2019.117380>.
- dos Santos DP, Temperini MLA, Brolo AG. Intensity fluctuations in single-molecule surface-enhanced Raman scattering. *Acc Chem Res.* 2019;52:456–64. <https://doi.org/10.1021/acs.accounts.8b00563>.
- Wang F, Cao S, Yan R, Wang Z, Wang D, Yang H. Selectivity/specificity improvement strategies in surface-enhanced Raman spectroscopy analysis. *Sensors.* 2017;17:2689. <https://doi.org/10.3390/s17112689>.
- Sena MM, Almeida MR, Braga JWB, Poppi RJ. Multivariate statistical analysis and chemometrics. In: Franca AS, Nollet LML, editors. *Spectroscopic Methods in Food Analysis*. 1st ed. Boca Raton: Taylor & Francis Group; 2017. pp. 273–314.
- Abraham S, Golay MJE. Smoothing and differentiation of data by simplified least squares procedures. *Anal Chem.* 1964;36:1627–39. <https://doi.org/10.1021/ac60214a047>.
- Brown CD, Wentzell PD. Hazards of digital smoothing filters as a preprocessing tool in multivariate calibration. *J Chemom.* 1999;13:133–52. [https://doi.org/10.1002/\(SICI\)1099-128X\(199903/04\)13:2%3c133::AID-CEM533%3e3.0.CO;2-C](https://doi.org/10.1002/(SICI)1099-128X(199903/04)13:2%3c133::AID-CEM533%3e3.0.CO;2-C).
- Zhang Z-M, Chen S, Liang Y-Z. Baseline correction using adaptive iteratively reweighted penalized least squares. *Analyst.* 2010;135:1138. <https://doi.org/10.1039/b922045c>.
- Phillip B. Smoothing and differentiation by an adaptive-degree polynomial filter. *Anal Chem.* 1995;67:2758–62. <https://doi.org/10.1021/ac00113a006>.
- Bro R, Smilde AK. Principal component analysis. *Anal. Methods.* 2014;6:2812–31. <https://doi.org/10.1039/C3AY41907J>.
- Wold S, Sjöström M. SIMCA: a method for analyzing chemical data in terms of similarity and analogy. In: Kowalski BR, editor. *Chemometrics: Theory and Application*. American Chemical Society; 1977. pp. 243–282. <https://doi.org/10.1021/bk-1977-0052.ch012>.
- Rodionova OYe, Oliveri P, Pomerantsev AL. Rigorous and compliant approaches to one-class classification. *Chemom Intell Lab Syst.* 2016;159:89–96. <https://doi.org/10.1016/j.chemolab.2016.10.002>.
- Nomikos P, MacGregor JF. Multivariate SPC charts for monitoring batch processes. *Technometrics.* 1995;37:41. <https://doi.org/10.2307/1269152>.
- Wold S, Ruhe A, Wold H, Dunn WJ III. The collinearity problem in linear regression. The Partial Least Squares (PLS) Approach to Generalized Inverses. *SIAM J Sci Stat Comput.* 1984;5:735–43. <https://doi.org/10.1137/0905052>.
- Rodionova OYe, Titova A v., Pomerantsev AL. Discriminant analysis is an inappropriate method of authentication. *TRAC,*

- Trends Anal Chem. 2016;78:17–22. <https://doi.org/10.1016/j.trac.2016.01.010>.
33. McCulloch WS, Pitts W. A logical calculus of the ideas immanent in nervous activity. *Bull Math Biophys.* 1943;5:115–33. <https://doi.org/10.1007/BF02478259>.
 34. da Silva IN, Hernane Spatti D, Andrade Flauzino R, Liboni LHB, dos Reis Alves SF. *Artificial Neural Networks*. 1st ed. AG Switzerland: Springer Nature; 2017.
 35. Boser BE, Guyon IM, Vapnik VN. A training algorithm for optimal margin classifiers. In: Haussler D, chairman. *Proceedings of the fifth annual workshop on Computational learning theory*. New York: Association for Computer Machinery; 1992. pp. 144–152. <https://doi.org/10.1145/130385.130401>.
 36. Cervantes J, Garcia-Lamont F, Rodríguez-Mazahua L, Lopez A. A comprehensive survey on support vector machine classification: applications, challenges and trends. *Neurocomputing.* 2020;408:189–215. <https://doi.org/10.1016/j.neucom.2019.10.118>.
 37. Pomerantsev AL, Rodionova OYe. New trends in qualitative analysis: performance, optimization, and validation of multi-class and soft models. *TrAC Trends Anal Chem.* 2021;143:116372. <https://doi.org/10.1016/j.trac.2021.116372>.
 38. López MI, Callao MP, Ruisánchez I. A tutorial on the validation of qualitative methods: from the univariate to the multivariate approach. *Anal Chim Acta.* 2015;891:62–72. <https://doi.org/10.1016/j.aca.2015.06.032>.
 39. Olivieri AC. Practical guidelines for reporting results in single- and multi-component analytical calibration: a tutorial. *Anal Chim Acta.* 2015;868:10–22. <https://doi.org/10.1016/j.aca.2015.01.017>.
 40. Szymańska E, Gerretzen J, Engel J, Geurts B, Blanchet L, Buydens LMC. Chemometrics and qualitative analysis have a vibrant relationship. *TrAC, Trends Anal Chem.* 2015;69:34–51. <https://doi.org/10.1016/j.trac.2015.02.015>.
 41. Ciloglu FU, Hora M, Gundogdu A, Kahraman M, Tokmakci M, Aydin O. SERS-based sensor with a machine learning based effective feature extraction technique for fast detection of colistin-resistant *Klebsiella pneumoniae*. *Anal Chim Acta.* 2022;1221:340094. <https://doi.org/10.1016/j.aca.2022.340094>.
 42. Kashif M, Majeed MI, Nawaz H, Rashid N, Abubakar M, Ahmad S, Ali S, Hyat H, Bashir S, Batool F, Akbar S, Anwar MA. Surface-enhanced Raman spectroscopy for identification of food processing bacteria. *Spectrochim Acta A Mol Biomol Spectrosc.* 2021;261:119989. <https://doi.org/10.1016/j.saa.2021.119989>.
 43. Bari RZA, Nawaz H, Majeed MI, Rashid N, Iqbal M, Akram M, Yaqoob N, Yousaf S, Mushtaq A, Almas F, Shahzadi A, Amin I. Surface-enhanced Raman spectroscopic analysis of centrifugally filtered HBV serum samples. *Photodiagnosis Photodyn Ther.* 2022;38:102808. <https://doi.org/10.1016/j.pdpdt.2022.102808>.
 44. Zhu A, Ali S, Xu Y, Ouyang Q, Wang Z, Chen Q. SERS-based Au@Ag NPs Solid-phase substrate combined with chemometrics for rapid discrimination of multiple foodborne pathogens. *Spectrochim Acta A Mol Biomol Spectrosc.* 2022;270:120814. <https://doi.org/10.1016/j.saa.2021.120814>.
 45. Rahman A, Kang S, Wang W, Huang Q, Kim I, Vikesland PJ. Lectin-modified bacterial cellulose nanocrystals decorated with Au nanoparticles for selective detection of bacteria using surface-enhanced Raman scattering coupled with machine learning. *ACS Appl Nano Mater.* 2022;5:259–68. <https://doi.org/10.1021/acsnm.1c02760>.
 46. Bashir S, Nawaz H, Irfan Majeed M, Mohsin M, Nawaz A, Rashid N, Batool F, Akbar S, Abubakar M, Ahmad S, Ali S, Kashif M. Surface-enhanced Raman spectroscopy for the identification of tetracycline-resistant *E. coli* strains. *Spectrochim Acta A Mol Biomol Spectrosc.* 2021;258:119831. <https://doi.org/10.1016/j.saa.2021.119831>.
 47. Kazemzadeh M, Hisey CL, Dauros-Singorenko P, Swift S, Zargar-Shoshtari K, Xu W, Broderick NGR. Label-free classification of bacterial extracellular vesicles by combining nanoplasmonic sensors with machine learning. *IEEE Sens J.* 2022;22:1128–37. <https://doi.org/10.1109/JSEN.2021.3131527>.
 48. Liu S, Zhu Y, Li M, Liu W, Zhao L, Ma Y, Xu L, Wang N, Zhao G, Liang D, Yu Q. Rapid identification of different pathogenic spore-forming bacteria in spice powders using surface-enhanced Raman spectroscopy and chemometrics. *Food Anal Methods.* 2022;15:2810–20. <https://doi.org/10.1007/s12161-022-02326-y>.
 49. Uysal Ciloglu F, Saridag AM, Kilic IH, Tokmakci M, Kahraman M, Aydin O. Identification of methicillin-resistant *Staphylococcus aureus* bacteria using surface-enhanced Raman spectroscopy and machine learning techniques. *Analyst.* 2020;145:7559–70. <https://doi.org/10.1039/D0AN00476F>.
 50. Villa JEL, Quiñones NR, Fantinatti-Garboggini F, Poppi RJ. Fast discrimination of bacteria using a filter paper-based SERS platform and PLS-DA with uncertainty estimation. *Anal Bioanal Chem.* 2019;411:705–13. <https://doi.org/10.1007/s00216-018-1485-9>.
 51. Gao K, Zhu H, Charron B, Mochizuki T, Dong C, Ding H, Cui Y, Lu M, Peng W, Zhu S, Hong L, Masson J-F. Combining dense Au nanoparticle layers and 2D surface-enhanced Raman scattering arrays for the identification of mutant cyanobacteria using machine learning. *The Journal of Physical Chemistry C.* 2022;126:9446–55. <https://doi.org/10.1021/acs.jpcc.2c00584>.
 52. Thrift WJ, Ronaghi S, Samad M, Wei H, Nguyen DG, Cabuslay AS, Groome CE, Santiago PJ, Baldi P, Hochbaum AI, Ragan R. Deep learning analysis of vibrational spectra of bacterial lysate for rapid antimicrobial susceptibility testing. *ACS Nano.* 2020;14:15336–48. <https://doi.org/10.1021/acsnano.0c05693>.
 53. Tang J-W, Liu Q-H, Yin X-C, Pan Y-C, Wen P-B, Liu X, Kang X-X, Gu B, Zhu Z-B, Wang L (2021) Comparative analysis of machine learning algorithms on surface enhanced Raman spectra of clinical *Staphylococcus* species. *Front Microbiol* 12:. <https://doi.org/10.3389/fmicb.2021.696921>
 54. Ciloglu FU, Caliskan A, Saridag AM, Kilic IH, Tokmakci M, Kahraman M, Aydin O. Drug-resistant *Staphylococcus aureus* bacteria detection by combining surface-enhanced Raman spectroscopy (SERS) and deep learning techniques. *Sci Rep.* 2021;11:18444. <https://doi.org/10.1038/s41598-021-97882-4>.
 55. Cui F, Shen X, Cao B, Ji H, Liu J, Zhuang X, Zeng C, Qu B, Li S, Xu Y, Zhou Q. Bacterial identification and adhesive strength evaluation based on a mannose biosensor with dual-mode detection. *Biosens Bioelectron.* 2022;203:114044. <https://doi.org/10.1016/j.bios.2022.114044>.
 56. Almaviva S, Palucci A, Aruffo E, Rufoloni A, Lai A. *Bacillus thuringiensis* cells selectively captured by phages and identified by surface enhanced Raman spectroscopy Technique. *Micromachines (Basel).* 2021;12:100. <https://doi.org/10.3390/mi12020100>.
 57. Yang Y, Peng Y, Lin C, Long L, Hu J, He J, Zeng H, Huang Z, Li Z-Y, Tanemura M, Shi J, Lombardi JR, Luo X. Human ACE2-functionalized gold “virus-trap” nanostructures for accurate capture of SARS-CoV-2 and single-virus SERS detection. *Nanomicro Lett.* 2021;13:109. <https://doi.org/10.1007/s40820-021-00620-8>.
 58. Thrift WJ, Cabuslay A, Laird AB, Ranjbar S, Hochbaum AI, Ragan R. Surface-enhanced Raman scattering-based odor compass: locating multiple chemical sources and pathogens. *ACS Sens.* 2019;4:2311–9. <https://doi.org/10.1021/acssensors.9b00809>.
 59. Yang Y, Xu B, Murray J, Haverstick J, Chen X, Tripp RA, Zhao Y. Rapid and quantitative detection of respiratory viruses using surface-enhanced Raman spectroscopy and machine learning. *Biosens Bioelectron.* 2022;217:114721. <https://doi.org/10.1016/j.bios.2022.114721>.
 60. Li H, Zhang S, Zhu R, Zhou Z, Xia L, Lin H, Chen S. Early assessment of chemotherapeutic response in hepatocellular carcinoma based on serum surface-enhanced Raman spectroscopy. *Spectrochim Acta A Mol Biomol Spectrosc.* 2022;278:121314. <https://doi.org/10.1016/j.saa.2022.121314>.

61. Huang Z, Siddhanta S, Zheng G, Kickler T, Barman I. Rapid, label-free optical spectroscopy platform for diagnosis of heparin-induced thrombocytopenia. *Angew Chem.* 2020;132:6028–34. <https://doi.org/10.1002/ange.201913970>.
62. Dastgir G, Majeed MI, Nawaz H, Rashid N, Raza A, Ali MZ, Shakeel M, Javed M, Ehsan U, Ishtiaq S, Fatima R, Abdulraheem A. Surface-enhanced Raman spectroscopy of polymerase chain reaction (PCR) products of Rifampin resistant and susceptible tuberculosis patients. *Photodiagnosis Photodyn Ther.* 2022;38:102758. <https://doi.org/10.1016/j.pdpdt.2022.102758>.
63. Zhang K, Wang Z, Liu H, Perea-López N, Ranasinghe JC, Bepete G, Minns AM, Rossi RM, Lindner SE, Huang SX, Terrones M, Huang S. Understanding the excitation wavelength dependence and thermal stability of the SARS-CoV-2 receptor-binding domain using surface-enhanced Raman scattering and machine learning. *ACS Photonics.* 2022;9:2963–72. <https://doi.org/10.1021/acsp Photonics.2c00456>.
64. Koster HJ, Rojalín T, Powell A, Pham D, Mizenko RR, Birkeland AC, Carney RP. Surface enhanced Raman scattering of extracellular vesicles for cancer diagnostics despite isolation dependent lipoprotein contamination. *Nanoscale.* 2021;13:14760–76. <https://doi.org/10.1039/D1NR03334D>.
65. Berus SM, Adamczyk-Popławska M, Młynarczyk-Bonikowska B, Witkowska E, Szymborski T, Waluk J, Kamińska A. SERS-based sensor for the detection of sexually transmitted pathogens in the male swab specimens: a new approach for clinical diagnosis. *Biosens Bioelectron.* 2021;189:113358. <https://doi.org/10.1016/j.bios.2021.113358>.
66. Huang X, Liu B, Guo S, Guo W, Liao K, Hu G, Shi W, Kuss M, Duryee MJ, Anderson DR, Lu Y, Duan B (2022) SERS spectroscopy with machine learning to analyze human plasma derived sEVs for coronary artery disease diagnosis and prognosis. *Bioeng Transl Med.* e10420. <https://doi.org/10.1002/btm2.10420>
67. Moitra P, Chaichi A, Abid Hasan SM, Dighe K, Alafeef M, Prasad A, Gartia MR, Pan D. Probing the mutation independent interaction of DNA probes with SARS-CoV-2 variants through a combination of surface-enhanced Raman scattering and machine learning. *Biosens Bioelectron.* 2022;208:114200. <https://doi.org/10.1016/j.bios.2022.114200>.
68. Culum NM, Cooper TT, Bell GI, Hess DA, Lagugn e-Labarthe F. Characterization of extracellular vesicles derived from mesenchymal stromal cells by surface-enhanced Raman spectroscopy. *Anal Bioanal Chem.* 2021;413:5013–24. <https://doi.org/10.1007/s00216-021-03464-8>.
69. Kim W, Lee SH, Kim JH, Ahn YJ, Kim Y-H, Yu JS, Choi S. Paper-based surface-enhanced Raman spectroscopy for diagnosing prenatal diseases in women. *ACS Nano.* 2018;12:7100–8. <https://doi.org/10.1021/acsnano.8b02917>.
70. Li JQ, Dukes PV, Lee W, Sarkis M, Vo-Dinh T. Machine learning using convolutional neural networks for SERS analysis of biomarkers in medical diagnostics. *J Raman Spectrosc.* 2022;53:2044–57. <https://doi.org/10.1002/jrs.6447>.
71. Leong SX, Leong YX, Tan EX, Sim HYF, Koh CSL, Lee YH, Chong C, Ng LS, Chen JRT, Pang DWC, Nguyen LBT, Boong SK, Han X, Kao Y-C, Chua YH, Phan-Quang GC, Phang IY, Lee HK, Abdad MY, Tan NS, Ling XY. Noninvasive and point-of-care surface-enhanced Raman scattering (SERS)-based breathalyzer for mass screening of coronavirus disease 2019 (COVID-19) under 5 min. *ACS Nano.* 2022;16:2629–39. <https://doi.org/10.1021/acsnano.1c09371>.
72. Skvortsova A, Trelin A, Kriz P, Elashnikov R, Vokata B, Ulbrich P, Pershina A, Svorcik V, Guselnikova O, Lyutakov O. SERS and advanced chemometrics – utilization of Siamese neural network for picomolar identification of beta-lactam antibiotics resistance gene fragment. *Anal Chim Acta.* 2022;1192:339373. <https://doi.org/10.1016/j.aca.2021.339373>.
73. Yue X, Li H, Tang J, Liu J, Jiao J. Rapid and label-free screening of echinococcosis serum profiles through surface-enhanced Raman spectroscopy. *Anal Bioanal Chem.* 2020;412:279–88. <https://doi.org/10.1007/s00216-019-02234-x>.
74. Lin C, Liang S, Li Y, Peng Y, Huang Z, Li Z, Yang Y, Luo X. Localized plasmonic sensor for direct identifying lung and colon cancer from the blood. *Biosens Bioelectron.* 2022;211:114372. <https://doi.org/10.1016/j.bios.2022.114372>.
75. Hong Y, Li Y, Huang L, He W, Wang S, Wang C, Zhou G, Chen Y, Zhou X, Huang Y, Huang W, Gong T, Zhou Z (2020) Label-free diagnosis for colorectal cancer through coffee ring-assisted surface-enhanced Raman spectroscopy on blood serum. *J Biophotonics.* 13. <https://doi.org/10.1002/jbio.201960176>
76. Haldavnekar R, Venkatakrishnan K, Tan B. Cancer stem cell derived extracellular vesicles with self-functionalized 3D nanosensor for real-time cancer diagnosis: eliminating the roadblocks in liquid biopsy. *ACS Nano.* 2022;16:12226–43. <https://doi.org/10.1021/acsnano.2c02971>.
77. Xie Y, Wen Y, Su X, Zheng C, Li M. Label-free plasmon-enhanced spectroscopic HER2 detection for dynamic therapeutic surveillance of breast cancer. *Anal Chem.* 2022;94:12762–71. <https://doi.org/10.1021/acs.analchem.2c02419>.
78. Premachandran S, Haldavnekar R, Das S, Venkatakrishnan K, Tan B. DEEP surveillance of brain cancer using self-functionalized 3D nanoprobes for noninvasive liquid biopsy. *ACS Nano.* 2022;16:17948–64. <https://doi.org/10.1021/acsnano.2c04187>.
79. Liu Z, Li T, Wang Z, Liu J, Huang S, Min BH, An JY, Kim KM, Kim S, Chen Y, Liu H, Kim Y, Wong DTW, Huang TJ, Xie Y-H. Gold nanopillar arrays for non-invasive surface-enhanced Raman spectroscopy-based gastric cancer detection via sEVs. *ACS Appl Nano Mater.* 2022;5:12506–17. <https://doi.org/10.1021/acsnano.2c01986>.
80. Nargis HF, Nawaz H, Bhatti HN, Jilani K, Saleem M. Comparison of surface enhanced Raman spectroscopy and Raman spectroscopy for the detection of breast cancer based on serum samples. *Spectrochim Acta A Mol Biomol Spectrosc.* 2021;246:119034. <https://doi.org/10.1016/j.saa.2020.119034>.
81. Iancu SD, Cozan RG, Stefanu A, David M, Moisoiu T, Moroz-Dubenco C, Bajcsi A, Chira C, Andreica A, Leopold LF, Eniu D, Staicu A, Goidescu I, Socaciu C, Eniu DT, Diosan L, Leopold N. SERS liquid biopsy in breast cancer. What can we learn from SERS on serum and urine? *Spectrochim Acta A Mol Biomol Spectrosc.* 2022;273:120992. <https://doi.org/10.1016/j.saa.2022.120992>.
82. Gao N, Wang Q, Tang J, Yao S, Li H, Yue X, Fu J, Zhong F, Wang T, Wang J. Non-invasive SERS serum detection technology combined with multivariate statistical algorithm for simultaneous screening of cervical cancer and breast cancer. *Anal Bioanal Chem.* 2021;413:4775–84. <https://doi.org/10.1007/s00216-021-03431-3>.
83. Mert S, Sancak S, Aydın H, Fersahoğlu AT, Somay A, Özkan F, Çulha M. Development of a SERS based cancer diagnosis approach employing cryosectioned thyroid tissue samples on PDMS. *Nanomedicine.* 2022;44:102577. <https://doi.org/10.1016/j.nano.2022.102577>.
84. Sahin F, Celik N, Camdal A, Sakir M, Ceylan A, Ruzi M, Onses MS. Machine learning-assisted pesticide detection on a flexible surface-enhanced Raman scattering substrate prepared by silver nanoparticles. *ACS Appl Nano Mater.* 2022;5:13112–22. <https://doi.org/10.1021/acsnano.2c02897>.
85. Zhu J, Sharma AS, Xu J, Xu Y, Jiao T, Ouyang Q, Li H, Chen Q. Rapid on-site identification of pesticide residues in tea by one-dimensional convolutional neural network coupled with surface-enhanced Raman scattering. *Spectrochim Acta A Mol Biomol Spectrosc.* 2021;246:118994. <https://doi.org/10.1016/j.saa.2020.118994>.
86. Fang G, Lin X, Liang X, Wu J, Xu W, Hasi W, Dong B. Machine learning-driven 3D plasmonic cavity-in-cavity surface-enhanced Raman scattering platform with triple synergistic enhancement

- toward label-free detection of antibiotics in milk. *Small*. 2022;18:2204588. <https://doi.org/10.1002/sml.202204588>.
87. Leong YX, Lee YH, Koh CSL, Phan-Quang GC, Han X, Phang IY, Ling XY. Surface-enhanced Raman scattering (SERS) taster: a machine-learning-driven multireceptor platform for multiplex profiling of wine flavors. *Nano Lett*. 2021;21:2642–9. <https://doi.org/10.1021/acs.nanolett.1c00416>.
 88. Yao-Say Solomon Adade S, Lin H, Jiang H, Haruna SA, Osei Barimah A, Zareef M, Akomeah Agyekum A, Adwoa Nkuma Johnson N, Mehedi Hassan M, Li H, Chen Q. Fraud detection in crude palm oil using SERS combined with chemometrics. *Food Chem*. 2022;388:132973. <https://doi.org/10.1016/j.foodchem.2022.132973>.
 89. Yang F, Wang C, Yu H, Guo Y, Cheng Y, Yao W, Xie Y. Establishment of the thin-layer chromatography-surface-enhanced Raman spectroscopy and chemometrics method for simultaneous identification of eleven illegal drugs in anti-rheumatic health food. *Food Biosci*. 2022;49:101842. <https://doi.org/10.1016/j.fbio.2022.101842>.
 90. Teixeira CA, Poppi RJ. Paper-based SERS substrate and one-class classifier to monitor thiabendazole residual levels in extracts of mango peels. *Spectrochim Acta A Mol Biomol Spectrosc*. 2020;229:117913. <https://doi.org/10.1016/j.saa.2019.117913>.
 91. Bao Q, Zhao H, Han S, Zhang C, Hasi W. Surface-enhanced Raman spectroscopy for rapid identification and quantification of Flibanserin in different kinds of wine. *Anal Methods*. 2020;12:3025–31. <https://doi.org/10.1039/D0AY00741B>.
 92. Dies H, Raveendran J, Escobedo C, Docoslis A. Rapid identification and quantification of illicit drugs on nanodendritic surface-enhanced Raman scattering substrates. *Sens Actuators B Chem*. 2018;257:382–8. <https://doi.org/10.1016/j.snb.2017.10.181>.
 93. Li X, Feng S, Hu Y, Sheng W, Zhang Y, Yuan S, Zeng H, Wang S, Lu X. Rapid detection of melamine in milk using immunological separation and surface enhanced Raman spectroscopy. *J Food Sci*. 2015;80:C1196–201. <https://doi.org/10.1111/1750-3841.12876>.
 94. Alves IM, Melo NO, Marinho PA, Almeida MR. Liquid–liquid extraction-assisted SERS-based detection of clonazepam in spiked drinks. *Vib Spectrosc*. 2020;110:103112. <https://doi.org/10.1016/j.vibspec.2020.103112>.
 95. Jiang Y, Du S, Xu M, Yu T, Zhou B, Yu F, Jiang H, Yang L, Su M, Liu H. Tracking structural changes of protein residues by two-dimensional correlation surface-enhanced Raman spectroscopy. *Food Chem*. 2022;382:132237. <https://doi.org/10.1016/j.foodchem.2022.132237>.
 96. Weng S, Yuan H, Zhang X, Li P, Zheng L, Zhao J, Huang L. Deep learning networks for the recognition and quantitation of surface-enhanced Raman spectroscopy. *Analyst*. 2020;145:4827–35. <https://doi.org/10.1039/D0AN00492H>.
 97. Madsen R, Lundstedt T, Trygg J. Chemometrics in metabolomics—a review in human disease diagnosis. *Anal Chim Acta*. 2010;659:23–33. <https://doi.org/10.1016/j.aca.2009.11.042>.
 98. Chong I-G, Jun C-H. Performance of some variable selection methods when multicollinearity is present. *Chemom Intell Lab Syst*. 2005;78:103–12. <https://doi.org/10.1016/j.chemolab.2004.12.011>.
 99. Trygg J, Holmes E, Lundstedt T. Chemometrics in metabonomics. *J Proteome Res*. 2007;6:469–79. <https://doi.org/10.1021/pr060594q>.
 100. Kjeldahl K, Bro R. Some common misunderstandings in chemometrics. *J Chemom*. 2010;24:558–64. <https://doi.org/10.1002/cem.1346>.
 101. Kingsford C, Salzberg SL. What are decision trees? *Nat Biotechnol*. 2008;26:1011–3. <https://doi.org/10.1038/nbt0908-1011>.
 102. Crocombe RA. Portable spectroscopy. *Appl Spectrosc*. 2018;72:1701–51. <https://doi.org/10.1177/0003702818809719>.
 103. Bylesjö M, Rantalainen M, Cloarec O, Nicholson JK, Holmes E, Trygg J. OPLS discriminant analysis: combining the strengths of PLS-DA and SIMCA classification. *J Chemom*. 2006;20:341–51. <https://doi.org/10.1002/cem.1006>.
 104. Oliveri P. Class-modelling in food analytical chemistry: development, sampling, optimisation and validation issues – a tutorial. *Anal Chim Acta*. 2017;982:9–19. <https://doi.org/10.1016/j.aca.2017.05.013>.
 105. Goodacre R, Graham D, Faulds K. Recent developments in quantitative SERS: moving towards absolute quantification. *TrAC, Trends Anal Chem*. 2018;102:359–68. <https://doi.org/10.1016/j.trac.2018.03.005>.
 106. Hassan MM, Jiao T, Ahmad W, Yi X, Zareef M, Ali S, Li H, Chen Q. Cellulose paper-based SERS sensor for sensitive detection of 2,4-D residue levels in tea coupled uninformative variable elimination-partial least squares. *Spectrochim Acta A Mol Biomol Spectrosc*. 2021;248:119198. <https://doi.org/10.1016/j.saa.2020.119198>.
 107. Zhu J, Agyekum AA, Kutsanedzie FYH, Li H, Chen Q, Ouyang Q, Jiang H. Qualitative and quantitative analysis of chlorpyrifos residues in tea by surface-enhanced Raman spectroscopy (SERS) combined with chemometric models. *LWT*. 2018;97:760–9. <https://doi.org/10.1016/j.lwt.2018.07.055>.
 108. Li H, Mehedi Hassan M, Wang J, Wei W, Zou M, Ouyang Q, Chen Q. Investigation of nonlinear relationship of surface enhanced Raman scattering signal for robust prediction of thiabendazole in apple. *Food Chem*. 2021;339:127843. <https://doi.org/10.1016/j.foodchem.2020.127843>.
 109. Pan H, Ahmad W, Jiao T, Zhu A, Ouyang Q, Chen Q. Label-free Au NRs-based SERS coupled with chemometrics for rapid quantitative detection of thiabendazole residues in citrus. *Food Chem*. 2022;375:131681. <https://doi.org/10.1016/j.foodchem.2021.131681>.
 110. Sricharoen N, Sukmanee T, Pienpinijtham P, Ekgasit S, Kitahama Y, Ozaki Y, Wongravee K. MCR-ALS with sample insertion constraint to enhance the sensitivity of surface-enhanced Raman scattering detection. *Analyst*. 2021;146:3251–62. <https://doi.org/10.1039/D1AN00069A>.
 111. Chen X, Lin H, Xu T, Lai K, Han X, Lin M. Cellulose nanofibers coated with silver nanoparticles as a flexible nanocomposite for measurement of fusilazole residues in Oolong tea by surface-enhanced Raman spectroscopy. *Food Chem*. 2020;315:126276. <https://doi.org/10.1016/j.foodchem.2020.126276>.
 112. Zhu X, Liu P, Xue T, Ge Y, Ai S, Sheng Y, Wu R, Xu L, Tang K, Wen Y. A novel graphene-like titanium carbide MXene/Au–Ag nanoshuttles bifunctional nanosensor for electrochemical and SERS intelligent analysis of ultra-trace carbendazim coupled with machine learning. *Ceram Int*. 2021;47:173–84. <https://doi.org/10.1016/j.ceramint.2020.08.121>.
 113. Zhu A, Xu Y, Ali S, Ouyang Q, Chen Q. Au@Ag nanoflowers based SERS coupled chemometric algorithms for determination of organochlorine pesticides in milk. *LWT*. 2021;150:111978. <https://doi.org/10.1016/j.lwt.2021.111978>.
 114. Hassan MM, Chen Q, Kutsanedzie FYH, Li H, Zareef M, Xu Y, Yang M, Agyekum AA. rGO-NS SERS-based coupled chemometric prediction of acetamidiprid residue in green tea. *J Food Drug Anal*. 2019;27:145–53. <https://doi.org/10.1016/j.jfda.2018.06.004>.
 115. Weng S, Zhu W, Li P, Yuan H, Zhang X, Zheng L, Zhao J, Huang L, Han P. Dynamic surface-enhanced Raman spectroscopy for the detection of acephate residue in rice by using gold nanorods modified with cysteamine and multivariate methods. *Food Chem*. 2020;310:125855. <https://doi.org/10.1016/j.foodchem.2019.125855>.
 116. Hassan MM, Li H, Ahmad W, Zareef M, Wang J, Xie S, Wang P, Ouyang Q, Wang S, Chen Q. Au@Ag nanostructure based SERS substrate for simultaneous determination of pesticides residue in tea via solid phase extraction coupled multivariate calibration. *LWT*. 2019;105:290–7. <https://doi.org/10.1016/j.lwt.2019.02.016>.
 117. Lee K-M, Yarbrough D, Kozman MM, Herrman TJ, Park J, Wang R, Korouski D. A rapid and convenient screening method for detection of restricted monensin, decoquinat, and lasalocid in animal feed by applying SERS and chemometrics. *Food Chem Toxicol*. 2020;144:111633. <https://doi.org/10.1016/j.fct.2020.111633>.
 118. Guo Z, Chen P, Wang M, Zuo M, El-Seedi HR, Chen Q, Shi J, Zou X. Rapid enrichment detection of patulin and alternariol in apple

- using surface enhanced Raman spectroscopy with coffee-ring effect. *LWT*. 2021;152:112333. <https://doi.org/10.1016/j.lwt.2021.112333>.
119. Li D, Zhang Q, Deng B, Chen Y, Ye L. Rapid, sensitive detection of ganciclovir, penciclovir and valacyclovir-hydrochloride by artificial neural network and partial least squares combined with surface enhanced Raman spectroscopy. *Appl Surf Sci*. 2021;539:148224. <https://doi.org/10.1016/j.apsusc.2020.148224>.
 120. Chio W-IK, Liu J, Jones T, Perumal J, Dinish US, Parkin IP, Olivo M, Lee T-C. SERS multiplexing of methylxanthine drug isomers via host-guest size matching and machine learning. *J Mater Chem C Mater*. 2021;9:12624–32. <https://doi.org/10.1039/D1TC02004H>.
 121. Li H, Geng W, Hassan MM, Zuo M, Wei W, Wu X, Ouyang Q, Chen Q. Rapid detection of chloramphenicol in food using SERS flexible sensor coupled artificial intelligent tools. *Food Control*. 2021;128:108186. <https://doi.org/10.1016/j.foodcont.2021.108186>.
 122. Wang L, Vendrell-Dones MO, Deriu C, Doğruer S, de B. Harrington P, McCord B., Multivariate analysis aided surface-enhanced Raman spectroscopy (MVA-SERS) multiplex quantitative detection of trace fentanyl in illicit drug mixtures using a handheld Raman spectrometer. *Appl Spectrosc*. 2021;75:1225–36. <https://doi.org/10.1177/00037028211032930>.
 123. Agyekum AA, Kutsanedzie FYH, Mintah BK, Annavaram V, Braimah AO. Rapid detection and prediction of norfloxacin in fish using bimetallic Au@Ag nano-based SERS sensor coupled multivariate calibration. *Food Anal Methods*. 2022;15:2346–56. <https://doi.org/10.1007/s12161-022-02297-0>.
 124. Lee K-M, Yarbrough D, Kozman MM, Herrman TJ, Park J, Wang R, Kurouski D. Rapid detection and prediction of chlortetracycline and oxytetracycline in animal feed using surface-enhanced Raman spectroscopy (SERS). *Food Control*. 2020;114:107243. <https://doi.org/10.1016/j.foodcont.2020.107243>.
 125. Tan A, Zhao Y, Sivashanmugan K, Squire K, Wang AX. Quantitative TLC-SERS detection of histamine in seafood with support vector machine analysis. *Food Control*. 2019;103:111–8. <https://doi.org/10.1016/j.foodcont.2019.03.032>.
 126. Durucan O, Wu K, Viehrig M, Rindzevicius T, Boisen A. Nanopillar-assisted SERS chromatography *ACS Sens*. 2018;3:2492–8. <https://doi.org/10.1021/acssens.8b00887>.
 127. Lu S, Yin Z, Zhang L, Du J, Jing C. On-site detection of multiple extracellular antibiotic resistance genes using SERS. *Sens Actuators B Chem*. 2022;369:132262. <https://doi.org/10.1016/j.snb.2022.132262>.
 128. Villa JEL, Pasquini C, Poppi RJ. Surface-enhanced Raman spectroscopy and MCR-ALS for the selective sensing of urinary adenine on filter paper. *Talanta*. 2018;187:99–105. <https://doi.org/10.1016/j.talanta.2018.05.022>.
 129. Villa JEL, Poppi RJ. A portable SERS method for the determination of uric acid using a paper-based substrate and multivariate curve resolution. *Analyst*. 2016;141:1966–72. <https://doi.org/10.1039/C5AN02398J>.
 130. Yan S, Liu C, Fang S, Ma J, Qiu J, Xu D, Li L, Yu J, Li D, Liu Q. SERS-based lateral flow assay combined with machine learning for highly sensitive quantitative analysis of *Escherichia coli* O157:H7. *Anal Bioanal Chem*. 2020;412:7881–90. <https://doi.org/10.1007/s00216-020-02921-0>.
 131. Barimah AO, Chen P, Yin L, El-Seedi HR, Zou X, Guo Z. SERS nanosensor of 3-aminobenzenesulfonic acid labeled Ag for detecting total arsenic in black tea combined with chemometric algorithms. *J Food Compos Anal*. 2022;110:104588. <https://doi.org/10.1016/j.jfca.2022.104588>.
 132. Chen P, Yin L, El-Seedi HR, Zou X, Guo Z. Green reduction of silver nanoparticles for cadmium detection in food using surface-enhanced Raman spectroscopy coupled multivariate calibration. *Food Chem*. 2022;394:133481. <https://doi.org/10.1016/j.foodchem.2022.133481>.
 133. Hassan MM, Ahmad W, Zareef M, Rong Y, Xu Y, Jiao T, He P, Li H, Chen Q. Rapid detection of mercury in food via rhodamine 6G signal using surface-enhanced Raman scattering coupled multivariate calibration. *Food Chem*. 2021;358:129844. <https://doi.org/10.1016/j.foodchem.2021.129844>.
 134. Brackx G, Guinoiseau D, Duponchel L, Gélabert A, Reichel V, Zrig S, di Meglio J-M, Benedetti MF, Gaillardet J, Charron G. A frugal implementation of surface enhanced Raman scattering for sensing Zn²⁺ in freshwaters – in depth investigation of the analytical performances. *Sci Rep*. 2020;10:1883. <https://doi.org/10.1038/s41598-020-58647-7>.
 135. Guo Z, Chen P, Yin L, Zuo M, Chen Q, El-Seedi HR, Zou X. Determination of lead in food by surface-enhanced Raman spectroscopy with aptamer regulating gold nanoparticles reduction. *Food Control*. 2022;132:108498. <https://doi.org/10.1016/j.foodcont.2021.108498>.
 136. Annaram V, Kutsanedzie YHF, Agyekum AA, Shah SA, Zareef M, Hassan MM, Waqas A, Ouyang Q, Chen Q. NaYF₄@Yb, Ho, Au/GO-nanohybrid materials for SERS applications—Pb(II) detection and prediction. *Colloids Surf B Biointerfaces*. 2019;174:598–606. <https://doi.org/10.1016/j.colsurfb.2018.11.039>.
 137. Li H, Liu S, Hassan MM, Ali S, Ouyang Q, Chen Q, Wu X, Xu Z. Rapid quantitative analysis of Hg²⁺ residue in dairy products using SERS coupled with ACO-BP-AdaBoost algorithm. *Spectrochim Acta A Mol Biomol Spectrosc*. 2019;223:117281. <https://doi.org/10.1016/j.saa.2019.117281>.
 138. Kennard RW, Stone LA. Computer aided design of experiments. *Technometrics*. 1969;11:137–48. <https://doi.org/10.1080/00401706.1969.10490666>.
 139. Kang S, Nam W, Zhou W, Kim I, Vikesland PJ. Nanostructured Au-based surface-enhanced Raman scattering substrates and multivariate regression for pH sensing. *ACS Appl Nano Mater*. 2021;4:5768–77. <https://doi.org/10.1021/acsanm.1c00549>.
 140. de Juan A, Tauler R. Multivariate curve resolution (MCR) from 2000: progress in concepts and applications. *Crit Rev Anal Chem*. 2006;36:163–76. <https://doi.org/10.1080/10408340600970005>.
 141. Brezeştean IA, Gherman AMR, Colniţă A, Dina NE, Müller Molnár C, Marconi D, Chiş V, David I-L, Cîntă-Pînzaru S. Detection and characterization of nodularin by using label-free surface-enhanced spectroscopic techniques. *Int J Mol Sci*. 2022;23:15741. <https://doi.org/10.3390/ijms232415741>.
 142. Yang M, Chen D, Hu J, Zheng X, Lin Z-J, Zhu H. The application of coffee-ring effect in analytical chemistry. *TrAC Trends Anal Chem*. 2022;157:116752. <https://doi.org/10.1016/j.trac.2022.116752>.
 143. Gentiluomo L, Roessner D, Augustijn D, Svilenov H, Kulakova A, Mahapatra S, Winter G, Streicher W, Rinnan Å, Peters GHJ, Harris P, Frieß W. Application of interpretable artificial neural networks to early monoclonal antibodies development. *Eur J Pharm Biopharm*. 2019;141:81–9. <https://doi.org/10.1016/j.ejpb.2019.05.017>.
 144. Ghaddar B, Naoum-Sawaya J. High dimensional data classification and feature selection using support vector machines. *Eur J Oper Res*. 2018;265:993–1004. <https://doi.org/10.1016/j.ejor.2017.08.040>.
 145. Kelis Cardoso VG, Poppi RJ. Cleaner and faster method to detect adulteration in cassava starch using Raman spectroscopy and one-class support vector machine. *Food Control*. 2021;125:107917. <https://doi.org/10.1016/j.foodcont.2021.107917>.
 146. de Santana FB, Borges Neto W, Poppi RJ. Random forest as one-class classifier and infrared spectroscopy for food adulteration detection. *Food Chem*. 2019;293:323–32. <https://doi.org/10.1016/j.foodchem.2019.04.073>.

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Springer Nature or its licensor (e.g. a society or other partner) holds exclusive rights to this article under a publishing agreement with the author(s) or other rightsholder(s); author self-archiving of the accepted manuscript version of this article is solely governed by the terms of such publishing agreement and applicable law.