Measurement 265 (2026) 120429

Contents lists available at ScienceDirect

Measurement

journal homepage: www.elsevier.com/locate/measurement

AI-assisted methodology for robust digital measurements by Raman spectroscopy: Quantification of inorganic pollutants in water

Antonio Nocera [a], Lorenzo Luciani [b], Gianluca Ciattaglia [a], Michela Raimondi [a], Laura Burattini [a], Susanna Spinsante [a], Ennio Gambi [a], Rossana Galassi [b,*]

[a] Dipartimento di Ingegneria dell'Informazione, Università Politecnica delle Marche, Ancona, Italy
[b] Scuola di Scienze e Tecnologie, Chemistry Division, Università di Camerino, Camerino, Italy

ARTICLE INFO

Keywords:
Raman spectra processing
Machine learning
Repeatability
Traceability
Water analysis

ABSTRACT

Raman spectroscopy is a versatile analytical tool, yet it often struggles with low sensitivity, hardware noise, and environmental interference. To address these limitations, this study presents an automated, Artificial Intelligence (AI)-assisted methodology to convert noisy optical signals into robust digital measurements.

The process involves acquiring high-dimensional, noisy spectral data from analyte solutions. A grid search across various algorithms identifies the optimal pre-processing pipeline to minimize noise variance and ensure metrological repeatability. Instead of relying on raw sensor feeds, the method fits a Gaussian curve combined with a polynomial baseline to the data, extracting precise measurements from the peak of this mathematical model. Supported by AI, the method successfully separates multiple optical signals and their shifts originating from interactions among analytes, proving itself capable to compensate also for possible hardware misalignment and thermal drift. As such, it can be used to quantify the concentration of selected inorganic pollutants in a mixture of analytes.

The primary application addressed in this work is quantifying inorganic pollutants in water, to enable in situ analysis without continuous expert supervision. Tests on binary and ternary mixtures of inorganic pollutants in pure water demonstrated that the Mean Absolute Percentage Error (MAPE) for nitrate was consistently below 10% in the concentration range between 0 mg/L to more than 15 000 mg/L, dropping to under 5% for concentrations exceeding 1000 mg/L. For concentrations below 1000 mg/L, the Mean Absolute Error (MAE) values were 67 mg/L for nitrate, 1475 mg/L for sulfate, and 736 mg/L for nitrite, respectively.

1. Introduction

The contamination of natural waters due to agricultural [1] or industrial activities [2], along with the increasing demand for drinking water driven by climate change [3], has intensified the need for efficient, fast, and user-friendly methods for water analysis [4–6]. Raman Spectroscopy (RS) is a promising technique for water analysis, as it is capable of detecting molecular vibrational modes of analytes in water samples with rather good resolution, but low sensitivity [7]. RS shows potential for waste or surface water analysis, being a non-destructive and non-invasive technique, and because it does not require any sample preparation or alteration, making it ideal for analyzing water samples without incurring in contamination or interference.

The massive development of optical technology, and the possibility to perform RS directly in water, without sample collection or complex preparation, make this technique suitable for *in situ* and real-time analysis. In fact, aqueous samples can be analyzed locally, without complicated multistep pretreatments, except for the removal of suspensions that may cause light dispersion compromising species detection [8]. Moreover, RS enables rapid detection methods [9,10], also thanks to the advent of special technical supplies that facilitate rapid sampling, such as the immersion probes [11]. Based on the above, RS makes real-time monitoring of water quality possible, which is especially useful in environmental or industrial applications [12].

The low sensitivity of the technique appears to be the most impeding factor for its wide application [13]. RS provides data for both qualitative and quantitative analysis; the position of the Raman shift can characterize the analyte and discriminate it from others, while the intensity of the peak is related to the concentration of the analyte. When two or more analytes are dispersed within the same sample, the presence of one Raman signal (e.g., sulfate) shifts the frequency of another (e.g., nitrate), and methods based on raw spectrum intensity

* sCorresponding author.
E-mail address: rossana.galassi@unicam.it (R. Galassi).

https://doi.org/10.1016/j.measurement.2026.120429
Received 31 July 2025; Received in revised form 15 December 2025; Accepted 12 January 2026
Available online 12 January 2026

analysis may fail. Although water has a strong Raman signal concentrated at specific regions in the spectrum, thus leaving a wavelength range for detecting other compounds without substantial overlap, it is necessary to improve the sensitivity of the RS technique, to make the analysis of contaminants and pollutants in water effective.

Several approaches have been proposed in the literature. As an example, Surface Enhanced Raman Spectroscopy (SERS) [14] and Fiber Enhanced Raman Spectroscopy (FERS) [15] techniques have been used to enhance Raman signals. Another approach, mainly consisting of sample pretreatment procedures, has been applied to enrich certain species and make them more detectable [16]. As the Raman data need to be processed to correct for instrumental factors (baseline, laser power variability, shot light filtering, and optical path corrections) and accidental interference, such as laser-induced fluorescence, chemometrics approach and data elaboration are needed to obtain useful outputs [17–19]. This paper addresses the limitation of low sensitivity inherent in Raman sensors by implementing an Artificial Intelligence (AI)-assisted gain mechanism, which relies on a summation of spectra by averaging repeated scans, to mathematically boost the Signal-to-Noise Ratio (SNR) of the process.

AI has recently gained attention in RS thanks to its ability to automatically process complex and dynamic Raman spectra, aiding in both quantitative and qualitative analysis [20,21]. Traditionally, recognizing specific analytes involves manual pattern matching in existing databases, requiring high expertise and being prone to errors. AI offers a solution for automated, objective substance recognition by training on extensive Raman spectra datasets. This way, thanks to the integration of AI in the processing pipeline, RS has seen applications in the medical field, in food safety monitoring, in drug monitoring, in pathogen screening, and environmental monitoring [22].

The most common AI approaches include Machine Learning (ML), which is simpler and quicker to train, but requires significant preprocessing, and Deep Learning (DL), which can automatically handle raw spectra, learning preprocessing, and feature extraction, while needing large amounts of training data. For these reasons, a common trend in the recent scientific literature is to employ transfer learning of neural networks, first trained on large datasets and then fine-tuned on experimental recordings [19]. Despite several studies that leverage AI in RS, quantitative Raman spectroscopy through the use of AI is still in early phases. Only a few papers in the current state-of-the-art show it is possible to obtain the quantification of analytes' concentration from Raman spectra by ML, or DL [23]. Most of the methods focus solely on the classification of the analytes.

Raman analysis is, by itself, a qualitative technique that needs an expert user, who interprets the patterns in the spectrum to identify typical analytes. In this perspective, the use of AI enables improving the qualitative performance of Raman spectroscopy, by providing objective algorithms that can discriminate with a given performance the desired class of analytes. A recent example in measurement science showed the use of a micro-Raman spectrometer applied to the identification of microplastics in soil, with an accuracy up to 73.92% when paired to traditional ML, and up to 99.01% when using a CNN model, for seven classes [24]. The next step to advance Raman measurement science is adding the capability to quantify the concentration of a given analyte. Recently, interest in the quantification of substances has increased, with approaches from the traditional feature-based ML [25] and from DL, without a particular need for pre-processing steps [26].

Referring to water analysis, when combined with ML and enhancement techniques like SERS, RS has proven to be a reliable approach for overcoming challenges related to detecting water contaminants with high sensitivity [27]. Some of the pollutants that have already been detected with the applications of ML and DL are pathogens, such as bacteria, viruses, and nucleic acids, or organic contaminants, such as polyaromatic hydrocarbons, organophosphorus pesticides, or inorganic pollutants such as heavy metals, anions, and microplastics [27]. Despite the promising results, in all these works, no quantification of the concentration is provided. For inorganic pollutants, a large corpus of research focuses on investigating the application of RS for microplastic concentration quantification [28] and classification [24,29]. On the other hand, anion contaminants, such as nitrate, sulfate, and nitrite, have seen limited analysis with ML-assisted RS, and their concomitant detection is relevant to reveal biological material degradation [30]. Traditional techniques for spectra dimensionality reduction, such as Principal Component Analysis (PCA), have been used for ML-assisted SERS nitrate detection [31]. Two recent examples employed a monodimensional Convolutional Neural Network, pre-trained on the RUFF mineral dataset [32], to discriminate a group of substances, including nitrate and other pharmaceuticals, with detection limits of a few μg/L [33,34]; nevertheless, the application of ML was still limited to the classification of substances and not their quantification. Another recent trend in water and wastewater analysis is using ML-based regression models applied to RS. For instance, marine water samples were analyzed with the support of traditional ML techniques, including support vector regression, decision trees, gradient-boosted regression, k-nearest neighbors, multi-layer perceptron, and multivariate linear regression, for the quantification of glycine, glucose, and butyric acid, all organic substances [35]. Additionally, Lange et al. [26] analyzed eight substances, dissolved in water, including the inorganic analytes nitrate and sulfate, by means of traditional Raman spectroscopy in conjunction with machine learning and deep learning to obtain the quantification of their concentrations with performances ranging from 10 mg/L to 50 mg/L for nitrate and 190 mg/L to 410 mg/L for sulfate.

From the literature analysis, most of the studies leverage ML and DL methods for classifying analytes in the RS spectra. In this context, this work shows the development of a chemometric method using algorithms for the pre-processing of Raman spectra, and ML for quantifying the concentrations of analytes in a mixture of nitrates, sulfates, and nitrites, in water samples. The integration of AI in the processing pipeline of raw Raman spectral data leads to robust digital measurements, by providing dimensionality reduction, as specific fingerprint regions are isolated to focus the measurement window, and Gaussian fitting, so that instead of relying on raw intensity, which is susceptible to noise, the system fits a Gaussian curve combined with a polynomial baseline to the data. Despite the complexity of the raw data, the proposed method demonstrates that once the Gaussian features are extracted, a simple Linear Regression (LR) model outperforms more complex nonlinear models in making the final output, which relates concentration to spectrum peak (and area), a reliable linear function.

This is one of the first studies in the field of quantitative traditional RS applied to anion contaminants quantification based on ML. The approach relies on data collected from various chemical experiments and measurements; hence, the first step of this work is the production of a dataset, followed by its elaboration by trained ML methods. The training data of the ML model is restricted to single analyte cases, which are easily replicable in laboratory settings, while testing data is provided on more complex binary and ternary mixtures of analytes, showcasing the generalization capabilities of the proposed approach. The ML pipeline is treated not as a *black box* predictor, but as a signal processing stage designed to rectify determining instrumental errors, and minimize measurement uncertainty. This is attained by pursuing the minimization of the Mean Replicate Variation (MRV) and Mean Replicate Coefficient of Variation (MRCV) over all the concentration values, which effectively calibrates the system to minimize noise variance between repeated measurements of the same input, a core requirement for metrological repeatability. The proposed method significantly reduces prediction errors for low-concentration inputs, effectively extending the linear dynamic range of the RS system. For example, nitrate prediction error at low concentrations drops from 32.8% to 15.3% purely through the proposed algorithmic enhancement.

The paper is organized as follows: Section 2 highlights the measurement context and metrological relevance of the proposed study. All the details about the acquisition protocol can be found in Section 3.

The processing steps can be divided into pre-processing, where the best baseline correction, smoothing, and normalization are chosen to minimize the variability in the measurements, as explained in Section 4, followed by a feature extraction based on Gaussian fitting explained in Section 5, and regression based on a linear model, the presentation of which can be found in Section 6. Results and discussion can be found in Section 7, where the spectra following the pre-processing can be visualized, a feature correlation analysis is performed, the cross-validation results on the training and validation sets are commented, and the results on the mixture of analytes test sets are presented and discussed, before and after summation of spectra. Section 8 concludes the paper.

2. Measurement context and metrological relevance of the study

From a measurement perspective, Raman spectroscopy systems function as optical transducers that map molecular vibrations (playing the role of the measurand) to spectral intensity (the indication). By their inherent nature, these systems feature low sensitivity and significant susceptibility to influence quantities — specifically, baseline drift caused by fluorescence, and instrumental instability. The raw indication provided by a Raman spectrometer is affected by both systematic effects (baseline wander) and random effects (photon shot noise), rendering the direct relationship between signal intensity and analyte concentration non-linear and poorly repeatable, which also limits the reliability of the quantitative estimation derived from the spectral data. The challenging research problem is, therefore, to develop a calibration chain that minimizes the variability of repeated indications (thus leading to a reduced Type A uncertainty), corrects for systematic spectral shifts caused by environmental or hardware instability, and extends the measuring interval (dynamic range) into low-signal regions dominated by the noise floor.

In this paper, all of the above steps are tackled by an AI-assisted computational method, which abandons the direct use of raw spectral intensity in favor of a parametric model. The measurement result Y is derived via a calibration function f (Linear Regression) applied to a feature vector X (Gaussian peak), according to the model:

$$Y = f(X_{peak}) + \epsilon \tag{1}$$

where ϵ represents the residual error, due to measurement deviations the proposed ML pipeline helps to correct. The peak of the Gaussian curve and the area under the same curve are linearly correlated, as it happens in the ideal mathematic model, as the standard deviation of the Raman spectrum does not vary with concentration.

Systematic effects (bias) are corrected by employing an exhaustive grid search of baseline correction algorithms, to identify the optimal correction function, as detailed in Section 5. This computational step removes the systematic bias introduced by fluorescence, effectively zeroing the RS instrument before signal evaluation. The wavelength shift (drift of the x-axis), that arises in the presence of several analytes, but can be caused also by thermal expansion or optical misalignment, is investigated, testing the system's sensitivity to this influence quantity by introducing artificial shifts of $4.3\,\mathrm{cm}^{-1}$ to $13.3\,\mathrm{cm}^{-1}$ (Section 7.4). While the Linear Regression model based on raw intensity fails in the presence of drift, the model based on Gaussian Feature Extraction remains stable, decoupling the measurement result from slight instabilities of the spectrometer.

The proposed study explicitly adopts GUM [36] guidelines to evaluate the dispersion of measurement results. The experimental standard deviation for N repeated observations is calculated, to quantify the variability of the instrument's indication at specific wavelengths. To quantify and minimize random measurement error, the study introduces the Mean Replicate Coefficient of Variation (MRCV), that normalizes the standard deviation σ against the signal mean, providing a dimensionless quantity to assess the relative measurement repeatability

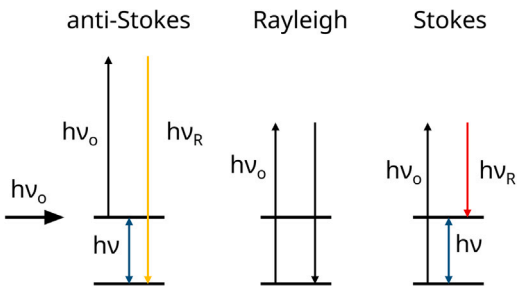


Fig. 1. Energy diagram showing the components of scattered light in a Raman spectrum. The laser light (ν_0) excites the molecules to a virtual state; the energy release occurs through the elastic Rayleigh light (ν_R) and the inelastic components due to the different vibrational populations [37].

across the entire spectral band. Then, the pre-processing pipeline is optimized specifically to minimize this value (Section 4). Random effects (noise) are addressed through signal averaging (summation of spectra), that leads to an increase in SNR. The final measurement accuracy is reported as the deviation from the reference standard, and quantified by the MAPE figure. The study transparently reports outliers, where the algorithm fails to converge, representing the system's statistical failure rate.

In the experimental setup, focused on water samples analysis, the measurand is represented by the mass concentration of inorganic anions (nitrate, sulfate, nitrite) in an aqueous matrix, and traceability is established using analytical standard salts with >99% purity, gravimetrically prepared in ultrapure water, to establish reference values.

3. Acquisitions of Raman spectra

This section is divided into paragraphs that describe the process and the instrumentation used to obtain the Raman spectra.

Raman spectroscopy. RS is an analytical spectroscopic technology based on the scattering of weak infrared components of light after a strong excitation through very intense light, with wavelengths that may be in the UV-vis-FIR (Ultra Violet, visible, Far InfraRed) range (ν_0). The scattered light is mainly laser light with weak infrared light components as a result of the vibrational modes of the molecules. The Raman effect, on which the technique is based, is very weak and requires intense light excitations to be detected. The large availability of lasers as powerful light sources, assembled with high-performance spectral connections and a spectrograph coupled with CCDs (Charge Coupled Devices) capable to separate the light components and to generate electric signals accordingly, yields outputs consisting of plots of Raman peaks against wavelength scans. In this plot, the Raman spectrum, the vibrational components, Raman shifts, derived from the subtraction or addition of laser light (generally Stokes), are drawn. Light combinations arise from the thermal distribution of the molecular population in vibrational levels, which upon absorption of laser light gives as a result the components denoted as $h\nu_0+h\nu$ and $h\nu_0-h\nu$, respectively, as shown in Fig. 1. The analysis of these vibrational components is diagnostic of the compounds' nature, and the intensities of the peaks are linearly correlated to their concentration, making this spectroscopy valuable for qualitative and quantitative analytical chemistry.

Instrumentation. In this set of analyses, the Raman optical spectroscopy of the stock solutions was performed by a micro-Raman system consisting of two Olympus microscopes that allow different experimental configurations to be addressed, and laser sources to be used depending on the material to be investigated: 532 nm, DPSS laser, 633 nm, He-Ne laser, and 785 nm, solid-state laser. Thanks to the use of optical fibers, the system can operate with laser beam dimensions in the range of 2 µm to 5 µm. The scattered photons are collected in backscattering geometry

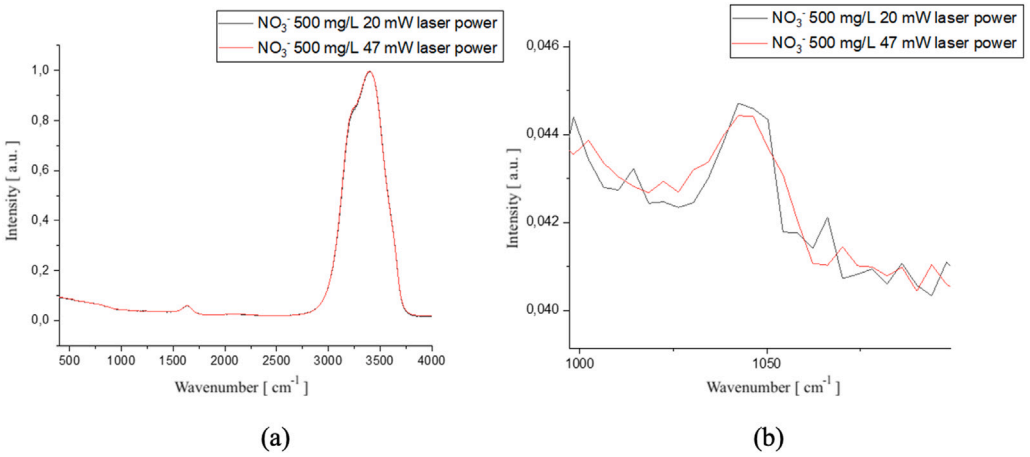


Fig. 2. Raman spectra acquired for a 500 mg/L solution of NO_3^-: (a) overlapped Raman spectra acquired with 532 nm laser powers at 20 mW (black curve) and 47 mW (red curve), with the nitrate peak at 1044 cm^{-1}; (b) magnification of the nitrate peak at 1044 cm^{-1}. Typical Raman peaks of water are at 3350 cm^{-1} and 1620 cm^{-1}.

Table 1
Instrumental parameters for the acquisition of the datasets.

Parameter	Value
Number of spectra	10
Number of accumulations per spectrum	5
Acquisition time for each accumulation [s]	10
Total acquisition time [s]	50
Laser wavelength [nm]	532
Laser power output [mW]	20

and analyzed by a spectrometer (Horiba iHS320) in Czerny-Turner geometry equipped with a Horiba Sincerity CCD7 camera detector. The spectrum is collected in the Stokes region in a range of 70 cm^{-1} to 6700 cm^{-1} wavenumbers, and the elastic peak of the output signal is eliminated through a low-pass filter (edge type).

Solution preparation. The solutions were prepared in standard glassware (Wheaton glass vials) using ultrapure water (MilliQ) and sodium or potassium anhydrous salts purchased from Merck. All the salts used in these assays are analytical standards with the highest commercial purity, over 99%. These salts were used without further purification and stored in a desiccator to prevent humidity. Stock solutions were prepared by dissolving the needed amount of solid samples in ultrapure water. Further dilutions of the stock solutions were performed by pipetting volumes and adding water to prepare a set of analytical samples with concentrations of the anions ranging from 15 000 mg/L to 200 mg/L. Such a process of solution preparation allows to establish reference values for traceability purposes of the proposed method.

Raman spectra acquisition. The water solutions were analyzed after transfer in an open Wheaton vial, to facilitate optical focus at 800 µm below the upper surface. The laser used for the set of analysis is a green light at 532 nm with a laser output power of 20 mW. At such power, Raman spectrum shown in Fig. 2(a) exhibits a peak at 1044 cm^{-1}, while an increase of laser power to 47 mW slightly affects the peak's intensity (Fig. 2b).

After an explorative set of acquisitions with different acquisition times, the analyses were performed by recording 10 scans with an integration time of 5 seconds for each spectrum, as reported in Table 1, repeating 10 times at any given concentration, to obtain the data sets. Averaging 10 repeated scans significantly reduces the random error component.

Spectra of solutions of the single analyte or mixtures of them were checked after a routine baseline correction through the peak analyzer method, according to the spline function. An example of the effect of applying the peaks analyzer method is provided comparing the raw spectra of Fig. 3(a) to those obtained with baseline correction, in Fig. 3(b).

The characteristic peaks of the analytes were normalized by considering the Raman peaks due to the symmetric stretching mode of water, centered at 3378 cm^{-1}. Raman spectra analysis of mixtures of inorganic salts is complicated by the interionic interactions between the cations and the anions in solutions, influencing the shape and the Raman shift of the analytes' peaks. Therefore, the recognition of an analyte by the peak detection might be compromised by the presence of other interacting chemical species, which may cause the peak's shift [38]. It will be shown later in the paper that the proposed AI-assisted model learns to decouple these mixed signals, proving itself resilient to wavelength shifts, whatever the cause that originated them. For example, in the solutions herein considered, the nitrate ion peak is shifted from the single-analyte case to the mixture of analytes as can be seen in Fig. 3(b). The spectra acquired for a single analyte nitrate are shown in Fig. 4(a), and for a mixture of nitrate, sulfate, and nitrite, they are reported in Fig. 4(b).

The concentration of the three anions is equal in this solution, but the signals are different both in terms of intensity and shape. The concentration of the single analyte primarily determines the absolute peak intensity; however, a highly polarizable molecule yields more intense peaks, thereby differentiating the relative intensities of the peaks. The peak position of an analyte recorded in a mixture may be shifted from that recorded in a single analyte solution; this effect is due to the interionic interactions among different analytes. Cations and anions may form ionic couples by intermolecular electrostatic attractions affecting the overall molecular symmetry, and the energy associated with the vibrational modes of the ions [8].

Different classes of raw Raman data were collected. Firstly, it was confirmed that a linear correlation exists between the intensity of a reference Raman peak due to an analyte and its concentration in the solution. For example, to build the calibration curve, a typical vibrational mode of the NO_3^- ion in water was considered, centered at 1046 cm (stretching NO_3^-, ν_1, (Ag symmetry), as a function of the ion concentration. In Fig. 5(a), the Raman shifts attributed to ν_1 nitrate vibrational mode in solutions with variable concentration (in the range 122 mg/L to 30 667 mg/L) are shown, and the linear regression between the peak areas and the concentration values (mg/L) is reported in Fig. 5(b).

Afterward, sets of data were generated by analyzing single analytes in a range of concentrations to determine the LODs (Limits Of Detection) (it was 578.85 mg/L for nitrate, according to the instrumental

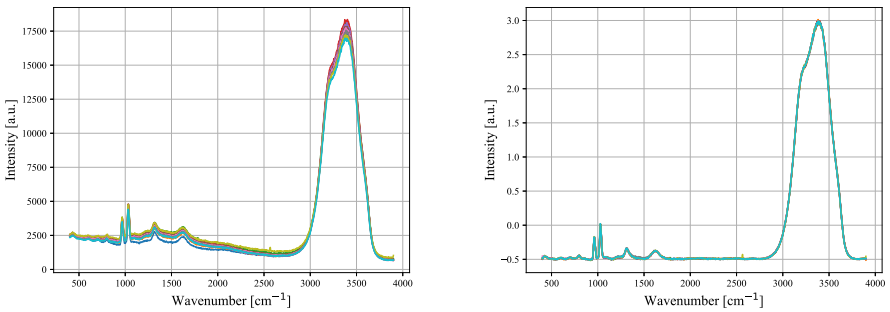


Fig. 3. Comparison between the Raman spectra recorded for a solution made of NO_3^-, NO_2^- SO_4^{-2} in equal concentration at 15 333 mg/L: (a) raw spectra, (b) spectra after baseline correction and SNV normalization.

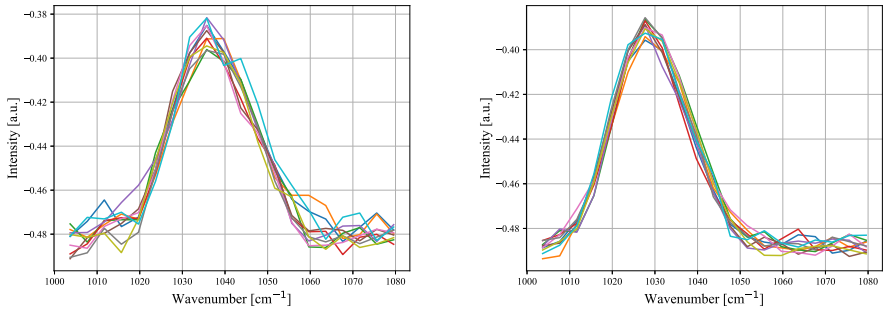


Fig. 4. Spectra in the range of wavenumber 1000 cm^{-1} to 1080 cm^{-1} of (a) a nitrate solution, and (b) a mixture of NO_3^-, NO_2^- SO_4^{-2}. The nitrate peak is located between 1030 cm^{-1} and 1040 cm^{-1} in the single analyte solution, and it is slightly shifted towards 1020 cm^{-1} in the mixture.

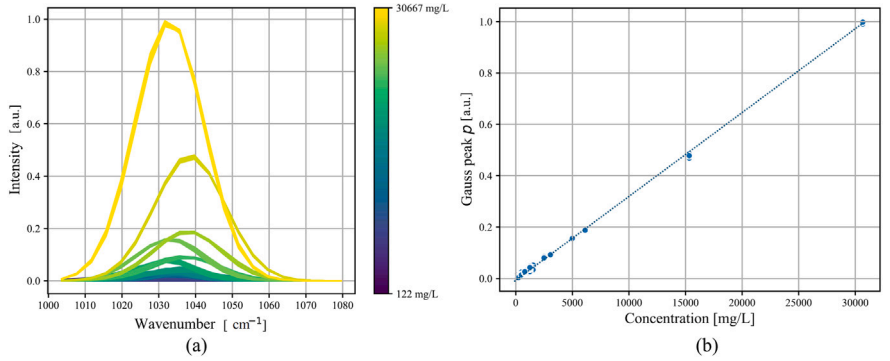


Fig. 5. Nitrate solutions at different concentrations: (a) overlapped spectra in the range of wavenumbers 1000 cm^{-1} to 1080 cm^{-1}, (b) linear regression of the peak and the concentration values, in mg/L.

setup herein considered), and data sets at different concentrations. Secondly, data were collected on solutions containing a mixture of two analytes in different proportions, and finally, data were collected on solutions resulting from the mixing of the three analytes NO_3^-, NO_2^- and SO_4^{2-}. The solutions and the corresponding concentrations in mg/L are reported in Tables 2, 3, and 4, respectively. The collected raw data were used for the data processing described in this paper.

Table 2
Solutions of single analyte [mg/L].

	A NO_3^-	B SO_4^{2-}	C NO_2^-
1	30 667	5072	5072
2	15 333	2536	2536
3	6133	1692	1692
4	3066	1014	1014
5	1533	845	845
6	613	507	507
7	306	253	253
8	122	–	–

4. Minimization of the mean replicate variation

The pre-processing steps are necessary when dealing with a Raman spectrum in a problem of ML, as they can drastically affect the performance of the models. Nevertheless, the ability to choose the proper combination of pre-processing steps necessitates expertise. To simplify this selection process, the pre-processing steps are considered as parameters to be tuned in an exhaustive grid search protocol, where different combinations of steps are tested, to minimize the average sample standard deviation of all the Raman shifts in a specific set of concentrations [39].

Table 3
Mixtures of two analytes [mg/L].

	D SO_4^{2-}/NO_3^-	E NO_2^-/NO_3^-	F NO_2^-/NO_3^-	G NO_2^-/NO_3^-
1	16 907/16 907	15 333/15 333	3067/15 333	15 333/3067
2	3381/3381	7655/7655	1533/7666	7666/1533
3	1690/1690	3066/3066	767/3833	3833/767
4	845/845	511/1533	383/1916	1916/383
5	422/422	511/511	–	–
6	3381/676	–	–	–
7	8453/1690	–	–	–
8	16 097/3381	–	–	–

For each concentration c_k of a given analyte, N observations (Raman spectra) are collected. Each Raman spectrum that describes intensity values as a function of wavelength is numerically represented as a one-dimensional vector S_j^k $(j = 1 \dots N)$ of length L. Each vector element $S_j^k[i]$ is identified by an index i, with $i = 1 \dots L$. Index i corresponds to each wavelength value for which an intensity value is present in the Raman spectrum, so that $S_j^k[i] = S_j^k(\lambda_i)$.

Taking N observations at a parity of concentration c_k, it is possible to compute the sample standard deviation σ_i^k of the intensity value S^k at wavelength λ_i, denoted here as $S^k(\lambda_i)$, across all the N realizations of the repeated Raman intensity acquisition, as (2):

$$\sigma_i^k = \sqrt{\frac{1}{N-1}\sum_{j=1}^{N}(S_j^k(\lambda_i) - \overline{S^k(\lambda_i)})^2} \tag{2}$$

where

$$\overline{S^k(\lambda_i)} = \frac{1}{N}\sum_{j=1}^{N} S_j^k(\lambda_i) \tag{3}$$

The experimental standard deviation σ_i^k, defined according to the Guide to the Expression of Uncertainty in Measurement (GUM) [36], is used to represent the variability of the detected Raman intensity at a single wavelength λ_i, over N repeated acquisitions performed at a parity of the analyte concentration. This way, a vector of standard deviation values σ_i^k is obtained, for $i = 1 \dots L$. By taking the average of these standard deviation values across a wavelength fingerprint region going from λ_A $(i = A)$ to λ_B $(i = B)$, $1 \leq A, B \leq L$, calculated as $\frac{1}{B-A}\sum_{i=A}^{B}\sigma_i^k$, it is possible to define a metric of variability for a specific concentration c_k over a given wavelength range. Then, taking the average over M different concentrations, a metric for repeatability, namely the Mean Replicate Variation (MRV), is defined as (4):

$$MRV = \frac{1}{M}\sum_{k=1}^{M}\left(\frac{1}{B-A}\sum_{i=A}^{B}\sigma_i^k\right) \tag{4}$$

The wavelength interval, from λ_A to λ_B, is the one in which it is expected to find the peak of the Raman spectrum intensity, for a specific analyte of interest, based on previous knowledge.

The problem with such a definition of MRV is that it is affected by the peak intensity values of each observation within the set of N. For this reason, a different variability metric is introduced, accounting for a normalization factor equal to the mean value of the intensity, at wavelength λ_i. Therefore, instead of the sample standard deviation σ_i^k, the sample coefficient of variation CV_i^k is considered, as (5):

$$CV_i^k = \frac{\sigma_i^k}{S^k(\lambda_i)} \tag{5}$$

Thereby, a Mean Replicate Coefficient of Variation (MRCV) is obtained by averaging over all the concentration values, as (6):

$$MRCV = \frac{1}{M}\sum_{k=1}^{M}\left(\frac{1}{B-A}\sum_{i=A}^{B}CV_i^k\right) \tag{6}$$

The normalization step applied to MRCV computation provides a dimensionless quantity to assess the relative measurement repeatability across the entire spectral band. This way, the proposed pipeline to be optimized is composed of the following steps: baseline correction, smoothing, normalization, region crop, and feature extraction.

The baseline correction is chosen from a set of state-of-the-art algorithms provided by the RamanSPy framework [40]. The alternatives used are Asymmetrically Reweighted Penalized Least Squares (ARPLS), Asymmetric Least Squares (ASLS), Improved Asymmetric Least Squares (IASLS), Adaptive Iteratively Reweighted Penalized Least Squares (AIRPLS), Doubly Reweighted Penalized Least Squares (DR-PLS), Improved Asymmetrically reweighted Penalized Least Squares (IARPLS), and Adaptive Smoothness Penalized Least Squares (ASPLS). Regarding the smoothing step, the Savitzky–Golay filter is commonly applied to Raman spectra as a denoising algorithm. Still, it necessitates specifying a sample window and a polynomial order. The window length and polynomial order are two parameters to set. The window length is selected considering all the values between 5 and 7, while the poly-order is selected considering all the values between 2 and 5 [41]. For normalization, four different algorithms are tested: the maximum intensity normalization, the norm vector of the entire spectra normalization, the area under the curve of the spectra normalization, and the standard variate normalization by removing the mean spectrum and dividing by the sample standard deviation of the total spectrum intensities.

All the combinations of the previously listed algorithms are exhaustively checked to find, for each analyte, the best pre-processing pipeline in terms of MRCV in the fingerprint related to the analyte.

5. Region crop and feature extraction

The region crop is necessary to focus the regressor on the wavelength region where the analytes are known to appear, and it performs a data dimensionality reduction that helps to simplify the training process of the regressor. Nitrate, nitrite, and sulfate analytes appear in a Raman spectrum as a deviation from the baseline at specific wavelength bands, as can be seen in Fig. 3(b). In the first step of our analysis, a baseline correction was identified as the ARPLS [42]. Then, the spectra can be cropped in the region associated with the specific analyte. Knowing that the nitrate analyte has a peak from 1030 nm to 1040 nm a fingerprint region is proposed from 1000 nm to 1080 nm. For the sulfate regression, knowing that a peak will appear from 970 nm to 980 nm, a fingerprint region from 930 nm to 1010 nm is proposed. For the nitrite, a fingerprint region from 1280 nm to 1420 nm is proposed.

Instead of relying on raw spectrum intensity, which is susceptible to noise, the proposed approach fits a Gaussian curve combined with a first-order polynomial baseline to the data in the fingerprint region, using the expression (7) below:

$$G_f(\lambda) = q + m\lambda + pe^{\frac{-(\lambda-\mu)^2}{2\sigma^2}} \tag{7}$$

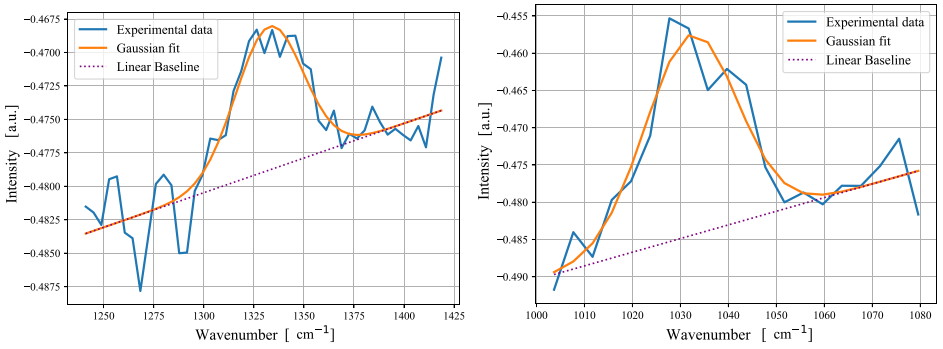
where q is the intercept of the linear function, m is the slope of the linear function, p, μ, and σ are the peak, the mean, and the standard deviation of the Gaussian curve, respectively. The first-order polynomial fitting is added to the Gaussian curve to address the issue of possible local baseline wander that could not be solved by the previous baseline correction. Examples of local baseline wanders can be found in Figs. 6(a) and 6(b).

The extracted peak, standard deviation, and the product between the peak and standard deviation of the Gaussian curve are used as parameters for the regression (Fig. 7). Even though any Gaussian curve is perfectly identified by its peak, standard deviation, and mean, only the first two parameters are considered as input features for the regression of the analyte concentration. The mean of the Gaussian curve is related to the type of analyte. The area is added as another possible feature, and computed as:

$$A = \sqrt{2\pi}\sigma p \tag{8}$$

Table 4
Mixtures of three analytes [mg/L].

	H $SO_4^{2-}/NO_3^-/NO_2^-$	I $SO_4^{2-}/NO_3^-/NO_2^-$	L $SO_4^{2-}/NO_3^-/NO_2^-$	M $SO_4^{2-}/NO_3^-/NO_2^-$
1	15 333/15 333/15 333	15 333/3067/3067	3067/15 333/3067	3067/3067/15 333
2	7665/7665/7665	7666/1533/1533	1533/7666/1533	1533/1533/7666
3	3066/3066/3066	3833/766/766	766/3833/766	766/766/3833
4	1533/1533/1533	1916/383/383	383/1916/383	383/383/1916
5	1022/1022/1022	–	–	–
6	511/511/511	–	–	–
7	255/255/255	–	–	–



(a) nitrite fit example (b) nitrate fit example

Fig. 6. Example of the first-order polynomial fit and Gaussian curve: (a) nitrite at a concentration of 756 mg/L, (b) nitrite at a concentration of 1000 mg/L.

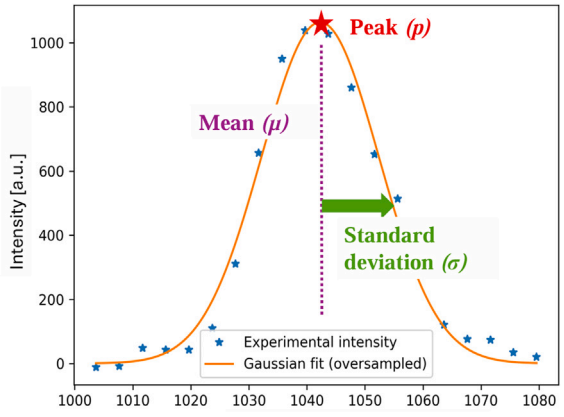


Fig. 7. Example of Gaussian fit on the experimental data of a nitrate analyte with a concentration of 6133 mg/L. Peak, mean, and standard deviation of the Gaussian fit are extracted as parameters.

The average intensity over the considered fingerprint region is also considered. Moreover, the raw experimental intensity measured by the Raman system without the feature extraction step is employed directly as an input feature vector. To choose the best input feature vector, a correlation analysis is performed to exclude the uncorrelated input, and then, based on cross-validation performances, as explained in the next section, the final features are found. All the computational steps described in the previous paragraphs help removing the systematic bias affecting the system, which is, for example, introduced by fluorescence, thus effectively zeroing the instrument before signal evaluation.

6. Machine learning approach

The obtained datasets containing spectra of a single analyte are employed as a development set, divided into training and validation sets following a three-fold partition. The examples in the validation sets are related to concentrations that do not appear in the training set.

Linear Regression (LR), Random Forest Regressor (RFR), and XGBoost Regressor (XGBR) are chosen as models to be trained, validated, and tested. The motivation behind choosing these models is that they have recently been employed for a similar purpose on sulfate concentration regression [43]; therefore, they produce good benchmarks on the newly acquired experimental data. The default hyperparameters are employed. The performance metrics computed on the validation sets can be used to understand the most suitable model for testing on unseen data. Moreover, as another discriminant, a robustness test for light wavenumber shifts in the Gaussian peak is used to understand the model and the input features with the best generalization capabilities in more complex scenarios.

Even though the previous analysis is useful for understanding the best-performing models in a controlled scenario, it is necessary to take a step towards a generalized scenario, where multiple analytes could appear in the spectra. For this reason, the best model found in the single-analyte case is trained for each analyte and then tested in the newly recorded examples with mixtures of analytes. Each single-analyte model is trained to quantify the concentration of a specific analyte, by analyzing a predetermined fingerprint region where the Raman peak will appear.

6.1. Performance metrics

Given a set c of concentration values on either the test or validation set, and a set c' of values predicted by a trained model, and using i to identify a generic example ($i = 1, \dots, N$), it is possible to define the Absolute Error (AE), and the APE as follows:

$$AE = |c_i - c'_i| \tag{9}$$

$$APE = \frac{|c_i - c'_i|}{c'_i} \cdot 100 \tag{10}$$

Then, Mean Absolute Error (MAE) and MAPE are given as follows:

$$MAE = \frac{1}{N} \sum_{i=1}^{N} |c_i - c'_i| \tag{11}$$

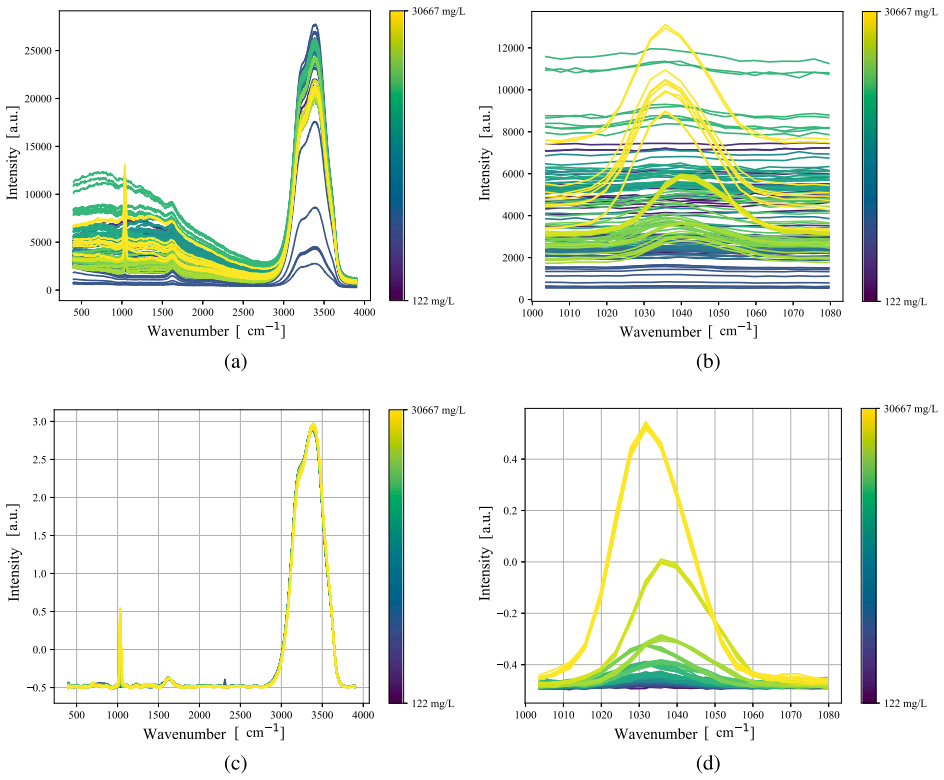


Fig. 8. Stacked plots of nitrates (a) without processing, (b) without processing and cropped, (c) after processing, (d) after processing and cropped.

$$MAPE = \frac{1}{N}\sum_{i=1}^{N}\left(\frac{|c_i - c_i'|}{c_i'} \cdot 100\right) \tag{12}$$

For the evaluation of the errors as performance metrics on the cross-validation for the best pre-processing, and the evaluation of the best performances on the final test set, MAPE is employed as it normalizes the error by the given concentration. MAE values are reported for completeness and to allow comparisons with other approaches in the scientific literature, only on the test set.

7. Results and discussion

In the next paragraphs, the results of the experiments are discussed, starting from the results related to the pre-processing steps, together with a qualitative analysis of the appearance of the spectra, by showing plots after the application of the selected pre-processing (Section 7.1). An analysis of the Gaussian feature extraction is proposed to understand the best set of features to employ in the ML regression. A correlation analysis is displayed in Section 7.2. Thereafter, in Section 7.3 three common ML algorithms are tested in the training and validation set, corresponding to the single analyte cases, to choose the best performing option, followed by a robustness test to slight wavelengths shifts in Section 7.4, to understand if the model is overfitting to specific peak locations. Thus, in Section 7.5, the best model is used for further validation on an external test set including various mixtures of analytes, to understand its generalization capabilities in more complex cases. Finally, in Section 7.6, to further improve the results, a novel definition of the Signal-to-Noise Ratio of the spectra is employed, a summation of spectra is introduced, and its effects are evaluated on the same test set, showing promising improvements.

7.1. Pre-processing results

In the training set of single analytes, the pre-processing steps, consisting of baseline correction, normalization, and smoothing, are chosen by an exhaustive grid search, to minimize the MRCV of the spectra. The best baseline correction for the nitrates and sulfates is the ARpls, followed by an SNV normalization. In contrast, a second-order polynomial is the best to correct the baseline of nitrites based on our experiments, followed by an SNV normalization. For nitrites and sulfate, a Savitzky–Golay smoothing with a window of 5 samples and a second-order polynomial kernel is employed, whereas for nitrates, a Savitzky–Golay smoothing filter with a window length of five samples and a third-order polynomial is used. The results, clustered on a single plot for nitrates, sulfates, and nitrites, can be seen in the panels of Figs. 8, 9, and 10, respectively. In each group of panels, it is possible to observe: the behavior of the spectra without processing, in Fig. 8(a), Fig. 9(a), and Fig. 10(a), respectively; the behavior of the spectra without processing cropped, in Fig. 8(b), Fig. 9(b), and Fig. 10(b), respectively; the behavior of the spectra with great changes in the baseline and variations of the peak, and the minimization of these variations after processing, in Figs. 8(c) and 8(d); Figs. 9(c) and 9(d); Figs. 10(c) and 10(d), respectively.

7.2. Feature correlation analysis

A correlation analysis can further explore the features extracted from the fingerprint Gaussian region to determine whether they are useful for regression. Table 5 shows a correlation matrix for each analyte and the extracted features.

It is possible to observe that the standard deviation (σ) of the Gaussian curve is not linearly correlated with the concentration and the other features. On the other hand, p, A, and the average intensity of the fingerprint are highly correlated. This is expected because the area is computed after the Gaussian fit, as the multiplication between σ and p. Thus, given that σ can be considered a constant at different concentrations, the area and the peak of the Gaussian curve give the same information. The average intensity over the fingerprint region, computed as the discrete mean of the intensity values, can be considered a discrete integral, and it is therefore correlated with the area and the peak of the Gaussian curve.

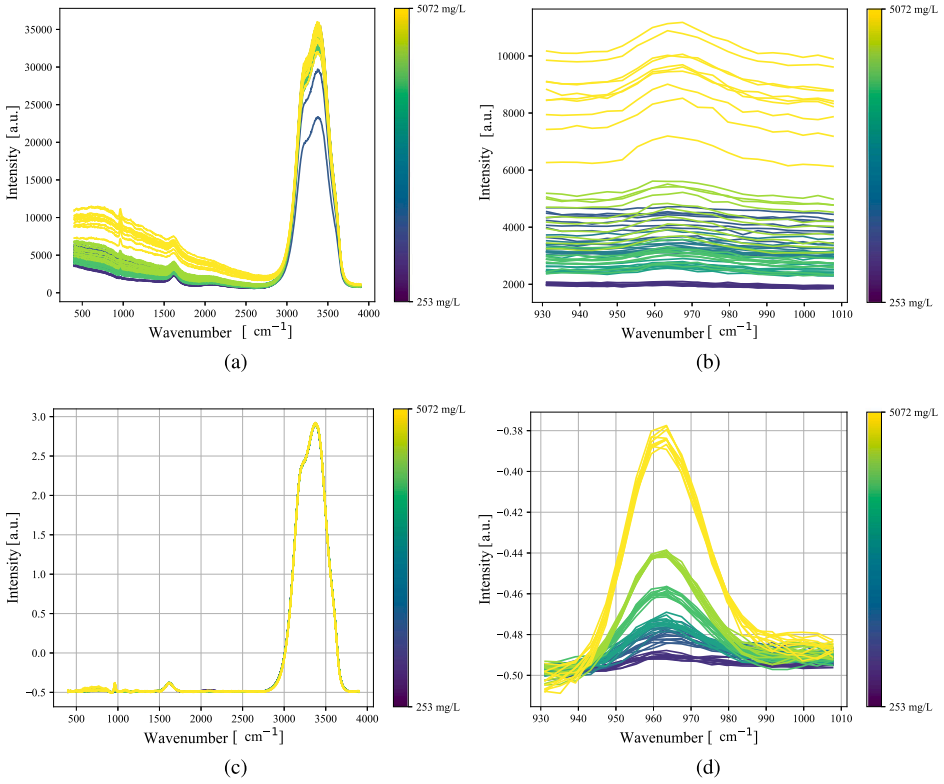


Fig. 9. Stacked plots of sulfates (a) without processing, (b) without processing and cropped, (c) after processing, (d) after processing and cropped.

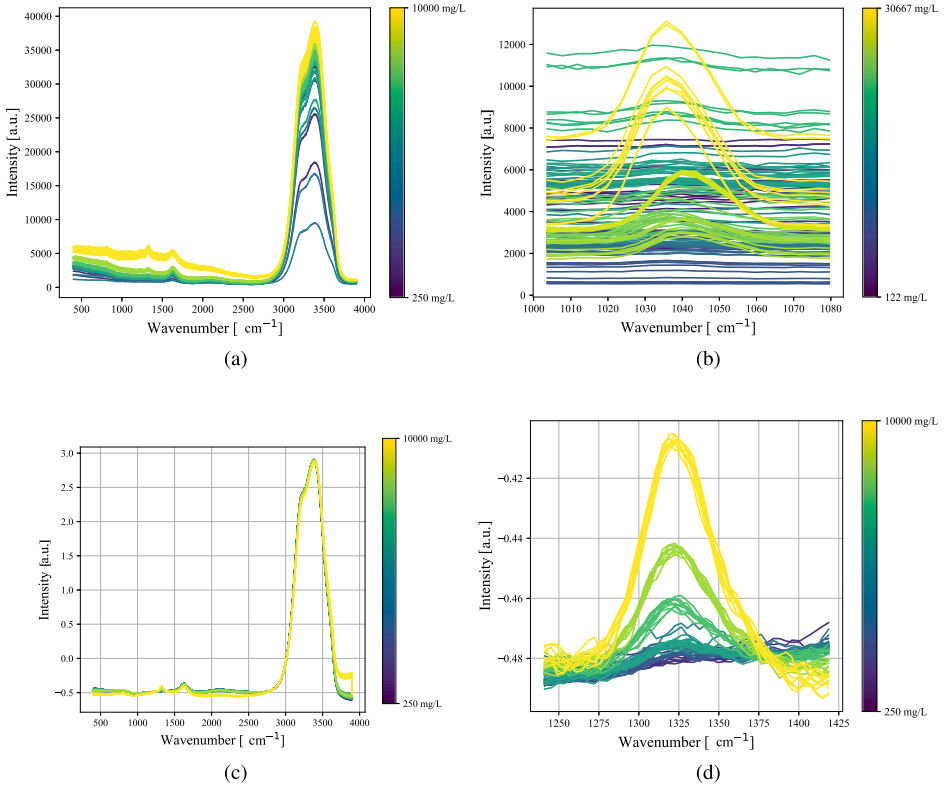


Fig. 10. Stacked plots of nitrites (a) without processing, (b) without processing and cropped, (c) after processing, (d) after processing and cropped.

Table 5
Correlation (C) matrix for the employed features in each analyte.

Analyte	Features	p	A	σ	*average intensity*	C
nitrate	p	1.00	1.00	−0.08	1.00	1.00
	A	1.00	1.001.00	−0.07	1.00	1.00
	σ	−0.0750	−0.07	1.00	−0.07	−0.07
	average intensity	1.00	1.00	−0.07	1.0000	1.00
sulfate	p	1.0000	1.00	−0.14	1.00	0.9981
	A	0.9972	1.00	−0.08	1.00	1.00
	σ	−0.14	−0.08	1.0000	−0.08	−0.14
	average intensity	1.00	1.00	−0.08	1.0000	1.00
nitrite	p	1.00	0.98	0.10	0.99	0.99
	A	0.98	1.00	0.2689	1.00	0.95
	σ	0.10	0.27	1.0000	0.20	0.02
	average intensity	0.9929	1.0	0.20	1.00	0.97

7.3. The best model on the development set of single analytes

Figs. 11(a), 11(c), and 11(e) show the distribution of the AE with boxplots for nitrate, sulfate, and nitrite, respectively. Instead, Figs. 11(b), 11(d), and 11(f) show the distribution of the APE with boxplots for nitrate, sulfate, and nitrite, respectively. The prediction errors are computed over a twenty-five-times repeated three-fold cross-validation. LR, RFR and XGBR are the models used. The parameters are peak, area, discrete mean, and raw intensity. It can be seen that the performances for the linear models are better than the RFR and XGBR models. The peak and raw intensity are the best-performing input features among the ones employed.

7.4. Robustness to slight wavelength shifts in raw intensity and peak features

The linear regression using peak or raw intensity as features represents, in both cases, a lightweight model if compared with the state-of-the-art approach, consisting of a deep one-dimensional Convolutional Neural Network. Nevertheless, the peak model is composed of only one weight and a bias. In contrast, the raw intensity model weights each sample in the chosen fingerprint region. The major issue is that increasing the complexity of the model, while maintaining a small number of examples in the training dataset, could impair the generalization capabilities of a slightly more complex model. In a real aqueous matrix, the interaction between different analytes could cause a slight wavelength shift in the position of the peak. Therefore, two linear regression models are trained: one with peak features and another one with raw intensity features, on all the examples of single analytes. Then, the training examples are shifted to the left and to the right by a number of samples from 1 to 3. Considering that the wavelength resolution is $4.3\,\mathrm{cm}^{-1}$ in our experiments, this robustness test shifts the spectra by a minimum of $4.3\,\mathrm{cm}^{-1}$ to a maximum of $13.3\,\mathrm{cm}^{-1}$. Figs. 12(a), 12(b), and 12(c) show the effect on the quantification error of slight shifts in the chosen fingerprint region, between the two input parameters for a linear regression model.

It can be noticed that the linear model taking as input the raw intensity does not generalize well to different positions of the Gaussian peak inside the chosen fingerprint region, whereas the peak model is not affected, as long as a large shift does not compromise the Gaussian fitting. Thus, the test for the mixture of analytes is conducted with the simpler model, which not only has a higher degree of interpretability, but it is also more robust to peak shifts.

7.5. Mixture of analytes analysis

A linear model for each analyte is trained on the single-analyte datasets. The regression models are composed of an angular coefficient that multiplies the extracted peak (p) in the given fingerprint, plus an added bias. The results have the following analytical expressions:

$$y_{NO_3^-} = 31334.40p + 130.40 \tag{13}$$

$$y_{SO_4^{2-}} = 44490.18p + 23.20 \tag{14}$$

$$y_{NO_2^-} = 135884.43p - 96.15 \tag{15}$$

Figs. 13(a), 13(b), and 13(c) show how the test examples are placed with respect to the predicted concentrations, based on the line plots.

Tables 6 report the MAE and MAPE for all the test examples, stratified for each analyte, for all the concentration c ranges, and for a given threshold of 1000 mg/L used to discriminate between low and high concentrations. It is possible to observe that the best results are obtained for the high concentration class, whereas the prediction error increases for the lower concentrations. This behavior is due to errors in the baseline correction and Gaussian fitting procedure, as analyzed in detail in the examples provided in Appendix.

7.6. Summation of spectra to improve signal-to-noise ratio

One of the ways to improve the results for these edge cases, where the baseline correction is not properly working or where the signal-to-noise ratio is low, is to sum or average the different observation spectra of the same sample. For each experiment and concentration, ten different recordings of 10 s are performed; thereby, by averaging these spectra, it is possible to obtain a single spectrum, with a higher signal-to-noise ratio. The SNR is defined in the fingerprint region of the considered analyte. After Gaussian fitting $G_f(\lambda)$ and local baseline removal, the power of the Gaussian signal is computed as the sum of the squared intensities. The noise signal $n(\lambda)$ is obtained by subtracting the Gaussian fitting and the first-order local baseline $(q + m\lambda)$ from the experimental intensities. The noise power is then computed as a summation of the squared noise intensity; eventually, the SNR can be calculated in logarithmic form.

$$R(\lambda) = G_f(\lambda) + n(\lambda) \tag{16}$$

$$P_{e(\lambda)} = \sum_{i=0}^{L} (R(\lambda) - G_f(\lambda))^2 = \sum_{i=0}^{L} (n(\lambda_i))^2 \tag{17}$$

$$P_{G(\lambda)} = \sum_{i=0}^{L} (G(\lambda_i) - (q + m\lambda))^2 \tag{18}$$

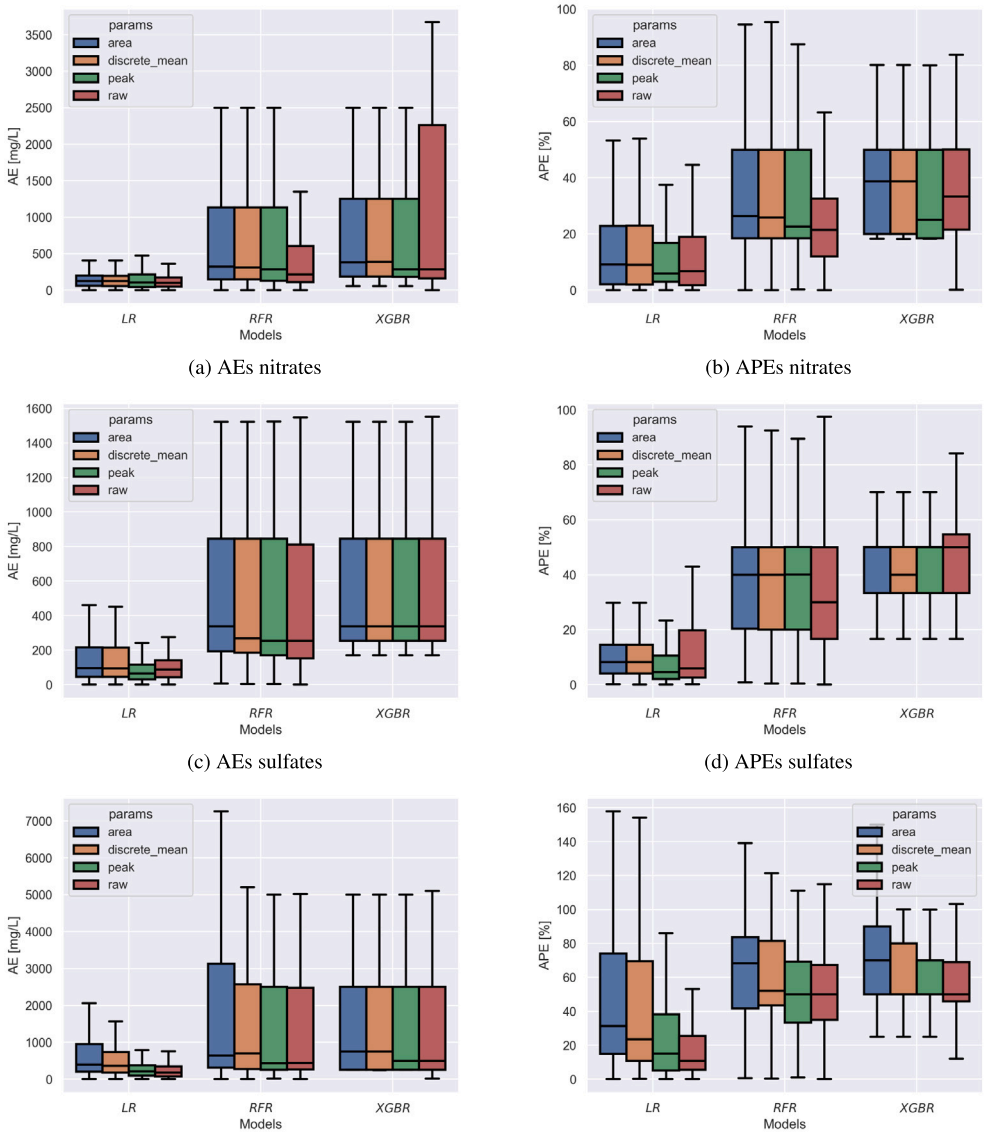
$$SNR = 10\ log_{10}\left(\frac{P_{G(\lambda)}}{P_{n(\lambda)}}\right) \tag{19}$$

Figs. 14(a), 14(b), and 14(c), show how a simple summation improves the SNR in the experiments considered. The scatter plots show the median SNR for the initial cases and the new SNR, after the spectra average.

It can be observed that there is a general improvement in SNR, spread between all the examples, with a highly positive jump for the lower concentrations.

Table 7 the results in terms of MAE and MAPE are reported after the summation of spectra.

It is possible to observe an improvement over the entire concentration range of the test example, and in particular for concentrations above 1000 mg/L. Considering all the concentration range, the nitrate

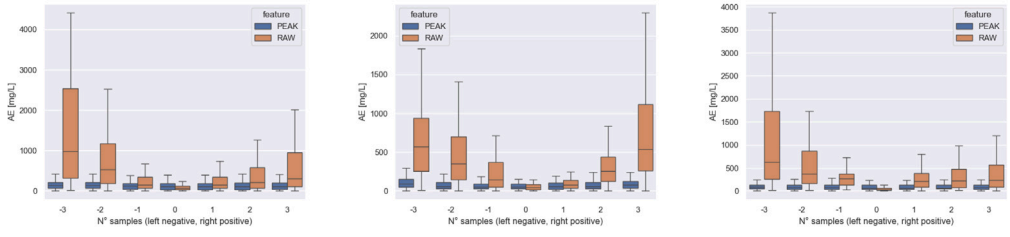


(a) AEs nitrates (b) APEs nitrates

(c) AEs sulfates (d) APEs sulfates

(e) AEs nitrites (f) APEs nitrites

Fig. 11. Boxplots of AEs and APEs for different models and parameters (params), in the single analyte case.

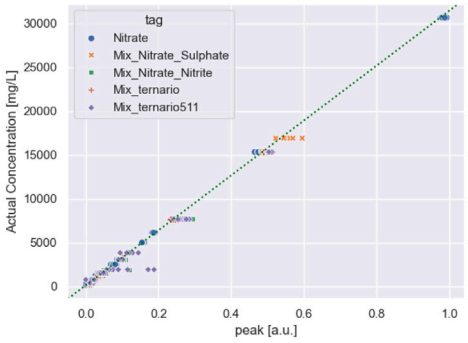


(a) (b) (c)

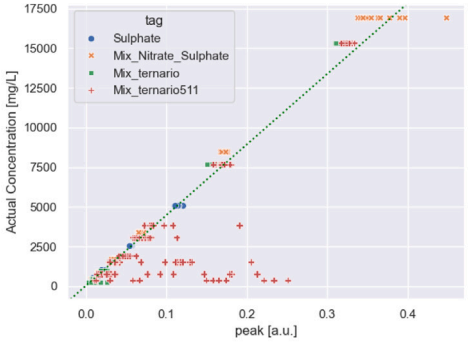
Fig. 12. Quantification errors due to slight shifts on a linear model, with input peak or raw intensity: (a) nitrates, (b) sulfates, (c) nitrites.

Table 6
Summary statistic of the prediction error for the analyzed analytes across all the experiments and concentration (c) ranges.

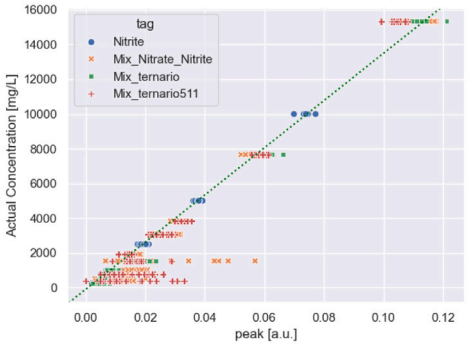
	All c		c above 1000 mg/L		c below 1000 mg/L	
	MAE [mg/L]	MAPE [%]	MAE [mg/L]	MAPE [%]	MAE [mg/L]	MAPE [%]
nitrate	244.0	15.7	281.0	8.6	154.0	32.8
sulfate	951.0	107.1	746.0	23.4	1432.0	304.2
nitrite	684.0	75.2	574.0	25.4	977.0	208.1



(a)



(b)



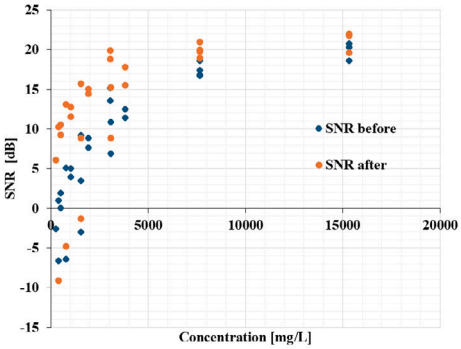
(c)

Fig. 13. Line plots and concentrations for the considered test examples for each analyte: (a) nitrate, (b) sulfate, (c) nitrite.

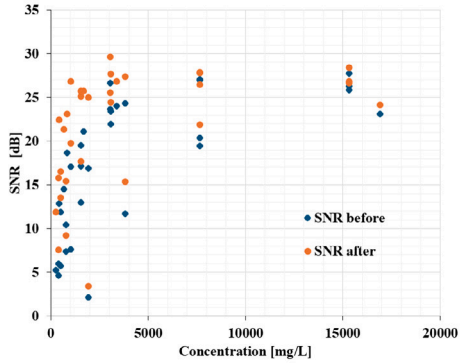
Table 7
Summary statistic of the prediction error for the analyzed analytes across all the experiments and concentration (c) ranges, after spectra summation.

	All c		c above 1000 mg/L		c below 1000 mg/L	
	MAE [mg/L]	MAPE [%]	MAE [mg/L]	MAPE [%]	MAE [mg/L]	MAPE [%]
nitrate	126.0	7.5	151.0	4.3	67.0	15.3
sulfate	902.0	106.9	660.0	20.1	1475.0	313.0
nitrite	510.0	56.1	425.0	17.3	736.0	159.5

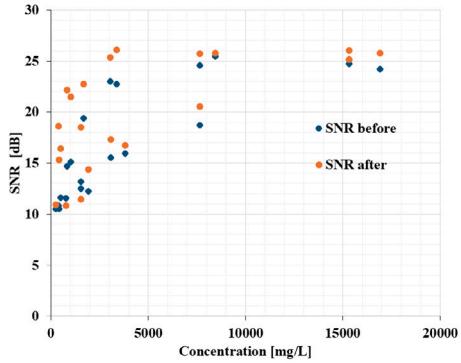
MAE decreases from 244 mg/L to 126 mg/L, the sulfate MAE from 951 to 902 mg/L and the nitrite MAE goes from 684 to 510 mg/L; the MAPE reduces accordingly, and for the nitrate, a MAPE of 7.5% is reached. For concentrations above 1000 mg/L, the prediction error tends to decrease; the MAE for nitrate lowers from 281 to 151 mg/L, the MAE for sulfate goes from 746 to 660 mg/L, and the MAE for nitrite goes from 574 to 425 mg/L. Concerning the MAPE, the nitrate reaches a value of 4.3%, the sulfate a value of 20.1%, and the nitrite a value of 17.3%. Concentrations below 1000 mg/L are associated with the highest prediction errors, as the signal of interest is so low that errors may occur in either the baseline correction procedure or the Gaussian fitting feature extraction. It is possible to observe, for the nitrate, a large reduction in prediction errors, with the MAE decreasing from 154 mg/L to 67 mg/L, and the MAPE from 32.8% to 15.3%; the sulfate did not seem to improve the results for this concentration range, due to the presence of some outliers as detailed in Appendix. Finally, the nitrite



(a)



(b)



(c)

Fig. 14. SNR improvements for (a) nitrate, (b) nitrite, and (c) sulfate.

improved from a MAE of 977 to 736 mg/L, and a MAPE going from 208.1% to 159.5%.

7.7. Comparison with state-of-the-art approaches

The typical application of ML to RS is to enhance its qualitative capabilities to identify and classify a set of substances, as recently performed by Post et al. [33,34]. North et al. [35] applied Raman spectroscopy to the quantification of organic substance in water. Nevertheless, as mentioned in our recent review paper on the topic of RS for nitrate detection [23], there are still limited contributions related to ML-assisted qualitative detection joint with quantification of nitrate and inorganic substances in water samples. The typical machine learning approach is composed of a pre-processing pipeline with baseline correction and peak removal, and smoothing, followed by an optional dimensionality reduction step, which could be a Principal Component Analysis or a feature extraction protocol, as in our case, followed by the machine learning model. On the other hand, the deep learning architectures can work without the need for pre-processing steps, on the condition of enough input data, which enables the model to learn an adequate pre-processing algorithm. The common deep learning models have a backbone based on one-dimensional convolutional operations. Even though the deep learning models have been shown to increase the performance over traditional machine learning approaches, they are based on the availability of thousands of spectra for the training process; additionally, they lack interpretability. In the presented experiments, the available number of spectra did not enable to build a complex deep learning architecture, which could suffer an issue of overfitting and low generalization capabilities with a small number of training examples. Thus, data scarcity was overcome by employing a simpler set of models, taken from traditional machine learning, with a feature protocol that enables the interpretability of the results.

When dealing with the quantification of nitrate with ML or DL applied to Raman spectroscopy, the existing contributions, to the authors'

best knowledge, are a preprint [44], dealing with SERS assisted with deep learning, and a recent publication by Lange et al. [26] on the quantification of eight substances, including nitrate and sulfate, with traditional Raman spectroscopy assisted by eleven machine learning and deep learning models. Given the difference in the sensor technology from the first study using SERS, a better comparison can be made against the second study employing standard Raman spectroscopy. There are some strengths and limitations in both the approach proposed by Lange et al. [26] and the approach herein presented, and some key differences that distinguish the two contributions.

Firstly, the acquisition campaigns are different, focusing overall on different analytes, with different concentration ranges; we included nitrate, sulfate, and nitrite, the latter being also the most challenging analyte, whereas Lange et al. considered nitrate, sulfate, and six other substances. A key difference is also in the concentration ranges: Lange et al. considered a concentration range between 0 and 1000 mg/L for nitrate, while our experiments extended the range to 15 000 mg/L, and focused more on the range above 1000 mg/L. Concerning sulfate, Lange et al. considered concentrations between 0 and 6000 mg/L, similar to our range in the training set (between 254 and 5066 mg/L). Secondly, the training, validation, and testing split choice is different. Lange et al. chose to perform a common 70/20/10 split with training, validation, and testing sets drawn randomly from the same experimental conditions. Thus, the authors chose to have both in training and testing a mixture of substances. In their training set, 832 spectra contained nitrate mixed with other substances, including sulfate; in their test set, a total of 114 examples of nitrate spectra were considered, of which 86 included nitrate and sulfate. Instead, our analysis used as a training set a simpler case with the presence of a single analyte with 80 spectra of nitrate, and, as test sets, different experiments with mixtures of analytes, with up to 330 spectra, including nitrate mixed with sulfate and nitrite, to show the generalization capabilities of the employed models. Our motivation is that the model used to quantify the target substance can be built through a simple training procedure on lab-controlled single analyte cases, and then generalized to more complex scenarios, similar to real-world matrices, where the analyte will appear mixed with other substances. Thirdly, our approach is similar to the proposed machine learning pipelines in Lange et al. with the addition of a feature extraction protocol that enhances the interpretability of the results, as we obtain a linear regression curve relating the spectra peak to the concentration of the substances.

Concerning a quantitative comparison of the results, [26] reported as best results a mean AE of 10 mg/L over a concentration range between 0 and 1000 mg/L of nitrate, and 190 mg/L over a concentration range between 0 and 6000 mg/L of sulfate. On the other hand, their worst results were obtained for a machine learning pipeline, including a baseline correction and a smoothing, with a mean AE of 50 mg/L for the nitrate, and 380 mg/L for the sulfate, in the previous concentration ranges. A fair comparison of our approach and theirs is complex because of the differences in the way the spectra are obtained, the data set cardinality and concentration ranges, and the choice of training, validation, and testing set splits. Moreover, our data set appears to be noisier when observed using the same SNR figure employed by Lange et al.: the authors reported peak values of the SNR distribution equal to 6500 and 8800, whereas our distributions, shown in Figs. 15(a), 15(b) and 15(c), have a major peak value of around 3000.

Given the presence of many outliers and the low SNR in our dataset, a fair comparison can be achieved by limiting the results to a high SNR threshold of 6000, as in Table 8. In this case, the spectra summation step is not performed. In total, 39 examples of sulfates and 28 examples of nitrate are analyzed, given the concentration and SNR limitations.

Lange et al. [26] managed to obtain better results in terms of MAE for nitrate for the "low" concentration range between 0 and 1000 mg/L, with 10 mg/L of error against our MAE of 96 mg/L; concerning sulfate, their best MAE amounts to 190 mg/L in a range between 0 and 6000 mg/L, and in the same concentration range our best MAE is 160 mg/L.

Our lower performance for the nitrate over the concentrations below 1000 mg/L could be explained by:

Table 8
Comparisons of MAE [mg/L] with a recent state-of-the-art approach.

	our MAE	[26] best MAE	[26] worst MAE
nitrate (<1000 mg/L)	96	10	50
sulfate (<6000 mg/L)	160	190	410

1. the fact that Lange et al. considered only "low" concentration (0 to 1000 mg/L) for their training, and thus their models showed better performances for this range;
2. Lange et al. had a greater number of samples in training with a distribution of SNR shifted towards higher quality signals.
3. a training set and a test set drawn from the same experimental conditions, being a mixture of analytes. Our study, on the other hand, aims at showing the generalization from the single-analyte to the multi-analyte case.

In fact, our results for the nitrate with signals of SNR higher than 6000 are similar across a large concentration range with 96 mg/L, 93 mg/L, and 98 mg/L for concentrations lower than 1000 mg/L, between 1000 and 6000 mg/L and over 6000 mg/L respectively.

8. Conclusions

There is a necessity for fast, efficient, and simple methods for water analysis to compensate for the spread of water contaminants due to agricultural and industrial activities and allow in situ verification. Raman spectroscopy has been demonstrated to be a tool for the qualitative evaluation of pollutants in water matrices. Nevertheless, before it can be used effectively for the purpose of contaminants quantification in water, the physical limitations of the low-sensitivity RS optical transducer need to be addressed. The method proposed in this paper shows how advanced signal processing can compensate for hardware noise, environmental interference, and sensor drift, thus leading to an advance in the RS measurement science.

In this work, Machine Learning acts as a virtual metrology layer, translating a low-fidelity optical indication into a traceable measurement result. This is achieved by correcting systematic bias (baseline removal), reducing random uncertainty (spectral summation and smoothing), ensuring robustness against instrumental drift (Gaussian parameterization). Then, as an application case of interest, the proposed ML-assisted RS is employed for inorganic pollutant detection and quantification. Nitrate, nitrite, and sulfate ions in water solutions have been chosen as models for water pollutants and taken as case studies to produce Raman data sets to be submitted to ML data elaboration. The pre-processing and training are developed on training and validation sets composed of single-analyte cases. Then, the developed approach is tested on multi-analyte cases, featuring the displacement of Raman peaks due to interionic interactions, allowing the generalization of the approach in more complex cases. The use of ML allows to stabilize the RS system output, by handling multi-analyte response interactions and decoupling mixed Raman signals. This translates into improved specificity in quantifying ternary mixtures (nitrate, nitrite, sulfate) from a single optical channel. Outlier errors could appear for certain concentrations depending on a failure of the pre-processing steps to properly address the baseline wander, the presence of noise, and to properly extract the Gaussian features of the analyte's fingerprint. For concentrations higher than 1000 mg/L of nitrate in a mixture of analytes, the absolute percent prediction error is below 10%, reaching values less than or equal to 5% at 15 333 mg/L. The absolute percent prediction errors for the nitrite are below 10% for concentrations over 1533 mg/L. Concerning the sulfates, the absolute percent prediction errors are below or close to 10% for most examples after 1533 mg/L.

These good results are confined within the borders of this study, but the proposed approach does not simply clean chemical data: it establishes an automated protocol to stabilize the Raman system output.

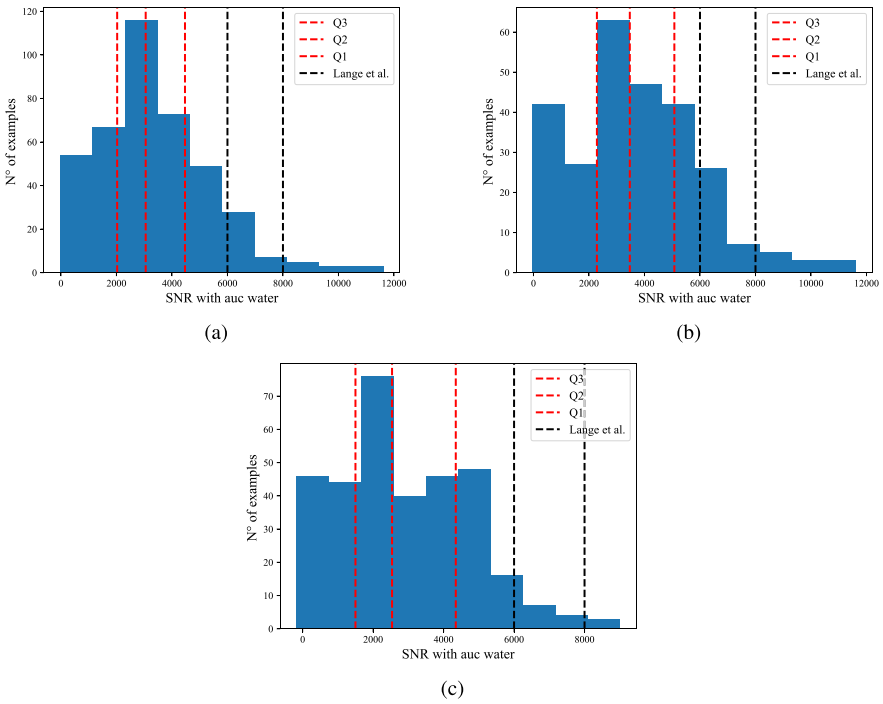


Fig. 15. Distribution of SNR values computed as in [26]. For our distribution, the first (Q1), second (Q2), and third quantiles (Q3) are used to better understand the difference with the other dataset. The distributions are divided per analyte: (a) nitrate, (b) sulfate, and (c) nitrite.

The methods herein applied appear to be promising in terms of AI-driven soft-calibrating sensors, leading to optimization of the Raman acquisitions, which pivots the technological development of low-cost and portable Raman instruments, to be applied to water analysis and beyond.

CRediT authorship contribution statement

Antonio Nocera: Writing – original draft, Methodology, Formal analysis, Data curation, Conceptualization, Investigation, Software, Validation, Visualization, Writing – review & editing. **Lorenzo Luciani:** Validation, Formal analysis, Data curation, Conceptualization, Investigation, Visualization, Writing – original draft, Writing – review & editing. **Gianluca Ciattaglia:** Formal analysis, Conceptualization, Validation, Writing – original draft, Writing – review & editing. **Michela Raimondi:** Writing – original draft, Formal analysis, Data curation, Conceptualization, Software, Validation, Visualization, Writing – review & editing. **Laura Burattini:** Data curation, Conceptualization, Formal analysis, Validation, Writing – original draft, Writing – review & editing. **Susanna Spinsante:** Formal analysis, Conceptualization, Data curation, Supervision, Validation, Visualization, Writing – original draft, Writing – review & editing. **Ennio Gambi:** Writing – review & editing, Supervision, Funding acquisition, Formal analysis, Data curation, Conceptualization, Investigation, Methodology, Project administration, Resources, Software, Validation, Visualization, Writing – original draft. **Rossana Galassi:** Writing – review & editing, Writing – original draft, Supervision, Project administration, Funding acquisition, Data curation, Conceptualization, Formal analysis, Investigation, Resources, Validation, Visualization.

Declaration of competing interest

The authors declare that they do not present any conflict of interest on the publication of this article.

Acknowledgments

The work described in this article was partially supported by the Italian Ministry for Enterprise and Made in Italy - DM 31/12/2021 and Agreements for Innovation 16/11/2023, Project “AAIWAS: Application of Artificial Intelligence to the Water and Air quality Sensing” - F/350142/01-03/X60, CUP (UNIVPM): B39J24000570005 and CUP (UNICAM): B19J24000430005.

Appendix. Results stratified by concentration

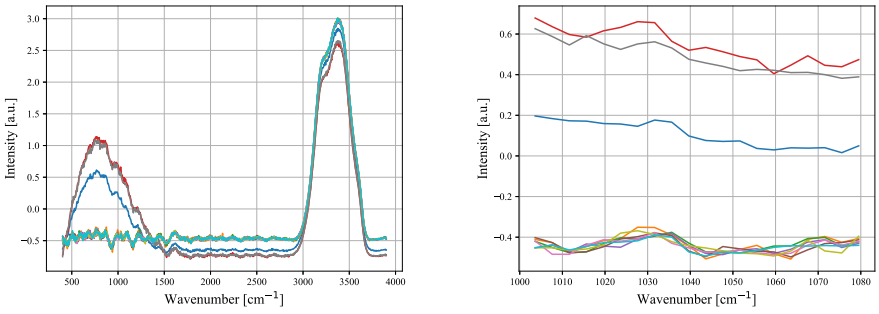
Table A.9 reports the test results on mixtures of analytes in terms of MAE and MAPE stratified by concentrations. When the mean APE goes under 10%, the cells are colored green, when the MAPE is between 10 and 20%, the cells are colored yellow, and for MAPE higher than 20%, the cells are colored red.

It is possible to observe a general pattern of decreasing error as the concentration increases for all the experimental setups. The lower concentrations pose a major problem because the signal has a lower intensity, impairing a proper Gaussian fitting over the fingerprint region. Analyzing the nitrate, it can be seen that for most of the cases, the MAPE goes under 10% (green highlights) when the concentration is above 1000 mg/L. This is clear in the experimental scenario H, where above the threshold of 1022 mg/L, it is possible to observe MAPE that go from 6.4% to 1.4%. A similar pattern can be seen for experiment D (nitrate and sulfate together), where the MAPE starts to decrease under the 10% threshold after 676 mg/L, going from MAPE of 8.2% at 845 mg/L to MAPE of 4.0% at 16.907 mg/L. Also, for the experiments E, F corresponding to the nitrate and nitrite mixture, and the experiments I, L, and M corresponding to the ternary mix with a ratio of 5:1:1 for the concentrations, it can be noticed that a behavior of declining error as the concentrations rise, with 1000 mg/L being the threshold above which the model performs better. The issues for these experiments are the appearance of some outlier examples that are associated with greater MAPE, such as the 1916 mg/L and 7666 mg/L concentrations.

These outlier results are caused mainly by baseline shifts that are not properly corrected by the proposed pipeline, as can be observed by the examples in Figs. A.16(a) and A.16(b).

Table A.9
All results on test sets stratified by concentrations and analytes; MAPEs are colored based on a threshold of less than 10% (green), less than 20% (yellow), more than 20% (orange).

Experiments		nitrate		sulfate		nitrite	
id	c [mg/L]	MAE [mg/L]	MAPE [%]	MAE [mg/L]	MAPE [%]	MAE [mg/L]	MAPE [%]
E/F/G	383	177	46.1	–	–	1032	269.4
E/F/G	511	250	49.0	–	–	1008	197.2
E/F/G	767	176	23.0	–	–	677	88.2
E/F/G	1022	209	20.5	–	–	1046	102.4
E/F/G	1533	149	9.7	–	–	1621	105.8
E/F/G	1916	354	18.5	–	–	214	11.2
E/F/G	3066	102	3.3	–	–	244	8.0
E/F/G	3067	208	6.8	–	–	754	24.6
E/F/G	3833	137	3.6	–	–	344	9.0
E/F/G	7655	118	1.5	–	–	271	3.5
E/F/G	7666	955	12.5	–	–	274	3.6
E/F/G	15 333	158	1.0	–	–	579	3.8
D	422	88	20.9	66	15.7	–	–
D	676	75	11.1	107	12.7	–	–
D	845	69	8.2	111	6.6	–	–
D	1690	81	4.8	384	11.4	–	–
D	3381	102	3.0	808	9.6	–	–
D	16 907	679	4.0	1321	7.8	–	–
H	255	193	75.8	171	67.2	290	113.5
H	511	166	32.4	69	13.5	179	35.0
H	1022	118	11.5	60	5.9	164	16.1
H	1533	97	6.4	139	9.1	901	58.8
H	3066	139	4.5	293	9.6	192	6.3
H	7665	158	2.1	690	9.0	584	7.6
H	15 333	216	1.4	1403	9.1	326	2.1
I/L/M	383	155	40.5	3337	871.2	1676	437.5
I/L/M	767	167	21.7	2117	276.4	1126	147.0
I/L/M	1533	133	8.7	2069	134.9	590	38.5
I/L/M	1916	1242	64.8	308	16.1	185	9.6
I/L/M	3067	87	2.8	339	11.1	394	12.8
I/L/M	3833	372	9.7	919	24.0	524	13.7
I/L/M	7666	805	10.5	257	3.4	263	3.4
I/L/M	15 333	754	4.9	928	6.1	1125	7.3



(a) (b)

Fig. A.16. The outlier in the experiments I, L, and M at concentration 1916 mg/L for nitrate: (a) spectra after baseline correction, (b) spectra after baseline correction and fingerprint crop.

Analyzing the sulfate in the same table, it can be observed that a similar pattern of declining error with the increase in concentration, even though the linear model does not reach the same performance as in the case of the nitrate regression, which can go near 1% for the highest concentration scenarios. The best performances for the sulfate linear model reach a MAPE between 3.4 and 9.6%. The outlier performances in yellow between 10 and 20% also appear for relatively large concentrations, such as 3833 mg/L, for experiments I, L, and M. In these cases, as before, the pre-processing pipeline does not compensate for the baseline shifts, leading to great errors in the final predictions.

Finally, the nitrite has a MAPE that goes under 10% after 1533 mg/L for most of the considered cases. The issue with the nitrite is the shape of its Gaussian curve, which is twice as large as the curve for sulfates and nitrates and is more prone to noise. For the highest concentration of 15 333 mg/L, the MAPE varies from 2.1 to 7.3%. It can be noticed that there are many cases of orange highlights with metrics of error going over 20%. The I, L, M, and E, F experiments have large errors until the 1533 mg/L. An outlier appears at 3067 mg/L and it can be observed in Figs. A.17(a) and A.17(b). Even though it appears to have a large linear baseline shift, the Gaussian fitting is working, and the high degree of error observed is primarily due to the high noise in the signal.

Table A.10 reports the new metrics AE and APE, along with the relative Δ, representing the difference between the error in the improved SNR case and the MAPE for the previously analyzed cases. Major improvements in the errors can be found in the lower concentrations.

Starting from the nitrate, for the H experiments, the orange highlights corresponding to 255 and 511 mg/L have an improvement of 39% and 8.7%, respectively; the MAPE for the 1022 mg/L goes under 10%. In the I, L, and M experiments, the previous outlier at 1916 mg/L has an improvement of 53%, reaching an APE of 11.9%. For experiments D and E, F, most of the previous orange highlights in Table A.9 improve to yellow or green highlights. The 422 mg/L in experiment D has an improvement of 4.3%, reaching an APE of 16.6%, and the next concentration at 676 mg/L has an improvement of 10.9%, reaching a APE of 0.2%.

Concerning nitrite, in the previous experiments, the green highlights appeared only after concentrations of 1533 mg/L. With the summation

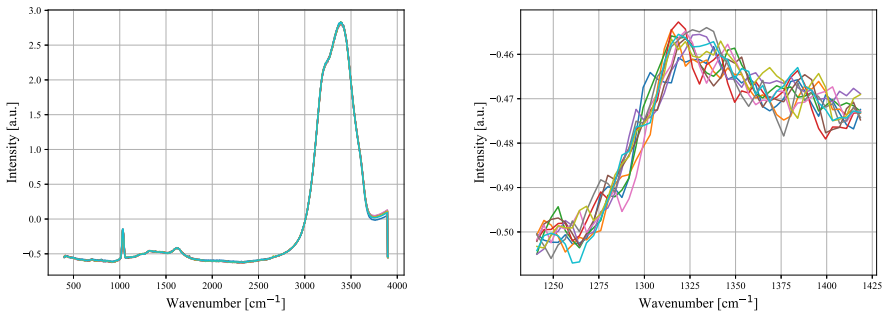


Fig. A.17. The outlier in the experiments I, L, and M at a concentration of 3067 mg/L for nitrite: (a) spectra after baseline correction, (b) spectra after baseline correction and fingerprint crop.

Table A.10
Results after summation of spectra; the deltas of variation for the APE are included, and when negative are related to an improvement; in italic are highlighted the increase in errors.

Experiments		nitrate		sulfate		nitrite	
id	c [mg/L]	AE [mg/L]	APE [%]	AE [mg/L]	APE [%]	AE [mg/L]	APE [%]
E/F/G	383	59	15.5 (−30.7)	–	–	866	226.1 (−43.3)
E/F/G	511	139	27.2 (−21.8)	–	–	629	123 (−74.1)
E/F/G	767	36	4.7 (−18.3)	–	–	604	78.7 (−9.5)
E/F/G	1022	60	5.9 (−14.6)	–	–	944	92.3 (−10.1)
E/F/G	1533	89	5.8 (−3.9)	–	–	904	59 (−46.8)
E/F/G	1916	360	*18.8 (0.3)*	–	–	153	8 (−3.2)
E/F/G	3066	81	2.6 (−0.7)	–	–	204	6.7 (−1.3)
E/F/G	3067	181	5.9 (−0.9)	–	–	692	22.6 (−2)
E/F/G	3833	22	0.6 (−3)	–	–	304	7.9 (−1)
E/F/G	7655	140	*1.8 (0.3)*	–	–	257	3.4 (−0.2)
E/F/G	7666	532	6.9 (−5.5)	–	–	142	1.9 (−1.7)
E/F/G	15 333	101	0.7 (−0.4)	–	–	566	3.7 (−0.1)
D	422	70	16.6 (−4.3)	77	18.1 (2.4)	–	–
D	676	2	0.2 (−10.9)	119	14 (1.4)	–	–
D	845	57	6.7 (−1.5)	118	7 (0.4)	–	–
D	1690	56	3.3 (−1.5)	390	11.5 (0.2)	–	–
D	3381	123	*3.6 (0.6)*	816	9.7 (0.1)	–	–
D	16 907	58	0.3 (−3.7)	1264	7.5 (−0.3)	–	–
H	255	94	36.9 (−39)	73	28.6 (−38.5)	53	20.6 (−92.9)
H	511	121	23.7 (−8.7)	53	10.3 (−3.2)	0	0.1 (−35)
H	1022	78	7.6 (−3.9)	16	1.5 (−4.4)	70	6.9 (−9.2)
H	1533	72	4.7 (−1.7)	5	0.3 (−8.8)	833	54.3 (−4.5)
H	3066	115	3.8 (−0.8)	287	9.4 (−0.2)	129	4.2 (−2.1)
H	7665	56	0.7 (−1.3)	675	8.8 (−0.2)	554	7.2 (−0.4)
H	15 333	37	0.2 (−1.2)	1422	9.3 (0.1)	106	0.7 (−1.4)
I/L/M	383	89	23.3 (−17.1)	3580	934.6 (63.4)	1542	402.7 (−34.9)
I/L/M	767	21	2.7 (−19)	2159	281.9 (5.5)	694	90.6 (−56.4)
I/L/M	1533	67	4.4 (−4.3)	2008	131 (−3.9)	243	15.8 (−22.7)
I/L/M	1916	227	11.9 (−53)	177	9.3 (−6.8)	113	5.9 (−3.8)
I/L/M	3067	35	1.2 (−1.7)	236	7.7 (−3.4)	236	7.7 (−5.1)
I/L/M	3833	82	2.1 (−7.6)	141	3.7 (−20.3)	466	12.2 (−1.5)
I/L/M	7666	616	8 (−2.5)	127	1.7 (−1.7)	186	2.4 (−1)
I/L/M	15 333	727	4.7 (−0.2)	960	6.3 (0.2)	1153	7.5 (0.2)

of spectra, it is possible to observe the appearance of "green" highlights for concentrations of 511 and 1022 mg/L in experiment H, where even for the lower concentration of 255 mg/L, the prediction error is slightly over 20%; nevertheless, in the other experiments, the concentrations below or equal to 1533 mg/L still have high prediction errors even with a great improvement of in the APE. The issue of the low sensitivity for the nitrite remains for the lower concentrations, and the summation enhances the results for the higher concentrations, reaching errors below 5% for concentrations higher than 1533 mg/L, for experiments E, F, G, and H.

For the sulfate examples, the major improvements can be found for the concentrations in between 1533 mg/L and 15 333 mg/L for I, L, and M experiments, where the prediction errors for the "yellow" outliers in Table A.9 disappear, and overall, the prediction errors are declining. However, in many cases, the averaging of the spectra does not seem to produce a benefit. In the H experiment, the concentration 15 333 mg/L experiences a slight increase in APE while maintaining a prediction error below 10%; a similar behavior can be found for experiment D with slight worsening of the results. The reason why the summation of the spectra does not produce an improvement for the higher concentration is probably based on the lack of examples of high concentrations of sulfates in the training and validation set, which would have biased the sulfate model towards a reduced amount of prediction error for the lower concentrations. Moreover, some of the issues with baseline correction could be amplified after spectra summation.

Data availability

Data will be made available on request.

References

[1] G. Mancuso, G.F. Bencresciuto, S. Lavrnić, A. Toscano, Diffuse water pollution from agriculture: A review of nature-based solutions for nitrogen removal and recovery, Water 13 (14) (2021) http://dx.doi.org/10.3390/w13141893.

[2] S. Madhav, A. Ahamad, A.K. Singh, J. Kushawaha, J.S. Chauhan, S. Sharma, P. Singh, Water Pollutants: Sources and Impact on the Environment and Human Health, Springer, Singapore, 2020, pp. 43–62, http://dx.doi.org/10.1007/978-981-15-0671-0_4.

[3] W. Rosińska, J. Jurasz, K. Przestrzelska, K. Wartalska, B. Kaźmierczak, Climate change's ripple effect on water supply systems and the water-energy nexus – A review, Water Resour. Ind. 32 (2024) 100266, http://dx.doi.org/10.1016/j.wri.2024.100266.
[4] S. Cinti, V. Mazzaracchio, G. Öztürk, D. Moscone, F. Arduini, A lab-on-a-tip approach to make electroanalysis user-friendly and decentralized: Detection of copper ions in river water, Anal. Chim. Acta 1029 (2018) 1–7, http://dx.doi.org/10.1016/j.aca.2018.04.065.
[5] H. Lee, W. Woo, Y.S. Park, A user-friendly software package to develop storm water management model (SWMM) inputs and suggest low impact development scenarios, Water 12 (9) (2020) http://dx.doi.org/10.3390/w12092344.
[6] S.D. Richardson, S.Y. Kimura, Water analysis: Emerging contaminants and current issues, Anal. Chem. 92 (1) (2020) 473–505, http://dx.doi.org/10.1021/acs.analchem.9b05269.
[7] A. Chandra, V. Kumar, U.C. Garnaik, R. Dada, I. Qamar, V.-K. Goel, S. Agarwal, Unveiling the molecular secrets: A comprehensive review of Raman spectroscopy in biological research, ACS Omega 9 (51) (2024) 50049–50063, http://dx.doi.org/10.1021/acsomega.4c00591.
[8] Yao-Hui Wang, Shisheng Zheng, Wei-Min Yang, Ru-Yu Zhou, Quan-Feng He, Petar Radjenovic, Jin-Chao Dong, Shunning Li, Jiaxin Zheng, Zhi-Lin Yang, Gary Attard, Feng Pan, Zhong-Qun Tian, Jian-Feng Li, In situ Raman spectroscopy reveals the structure and dissociation of interfacial water, Nature 600 (7887) (2021) 81–85, http://dx.doi.org/10.1038/s41586-021-04068-z.
[9] Z. Li, J. Wang, D. Li, Applications of raman spectroscopy in detection of water quality, Appl. Spectrosc. Rev. 51 (4) (2016) 333–357, http://dx.doi.org/10.1080/05704928.2015.1131711.
[10] S. Almaviva, F. Artuso, I. Giardina, A. Lai, A. Pasquo, Fast detection of different water contaminants by Raman spectroscopy and surface-enhanced Raman spectroscopy, Sensors 22 (21) (2022) http://dx.doi.org/10.3390/s22218338.
[11] M.K. Nieuwoudt, S.E. Holroyd, C.M. McGoverin, M.C. Simpson, D.E. Williams, Screening for adulterants in liquid milk using a portable Raman miniature spectrometer with immersion probe, Appl. Spectrosc. 71 (2) (2017) 308–312, http://dx.doi.org/10.1177/0003702816653130.
[12] A.A. Gowen, R. Tsenkova, M. Bruen, C. O'donnell, Vibrational spectroscopy for analysis of water for human use and in aquatic ecosystems, Crit. Rev. Environ. Sci. Technol. 42 (23) (2012) 2546–2573, http://dx.doi.org/10.1080/10643389.2011.592758.
[13] Y. Tang, Y. Zhuang, S. Zhang, Z.J. Smith, Y. Li, X. Mu, M. Li, C. He, X. Zheng, F. Pan, T. Gao, L. Zhang, Azo-enhanced Raman scattering for enhancing the sensitivity and tuning the frequency of molecular vibrations, ACS Central Sci. 7 (5) (2021) 768–780, http://dx.doi.org/10.1021/acscentsci.1c00117.
[14] T. Murphy, S. Lucht, H. Schmidt, H.D. Kronfeldt, Surface-enhanced Raman scattering (SERS) system for continuous measurements of chemicals in sea-water, J. Raman Spectrosc. 31 (10) (2000) 943–948, http://dx.doi.org/10.1002/1097-4555(200010)31:10<943::AID-JRS626>3.0.CO;2-X.
[15] H. Kerdoncuff, L.C. Deleebeeck, M. Lassen, Quantitative fiber-enhanced Raman sensing of inorganic nitrogen species in water, Chemosensors 9 (2) (2021) 29, http://dx.doi.org/10.3390/chemosensors9020029.
[16] S. Damrongsiri, Y. Hawangchu, A dual-mechanism pretreatment idea for improving the derived Raman spectrum for (micro)plastic analysis, Int. J. Environ. Anal. Chem. 0 (0) (2024) 1–20, http://dx.doi.org/10.1080/03067319.2024.2407052.
[17] Y. Qi, D. Hu, Y. Jiang, Z. Wu, M. Zheng, E.X. Chen, Y. Liang, M.A. Sadi, K. Zhang, Y.P. Chen, Recent progresses in machine learning assisted Raman spectroscopy, Adv. Opt. Mater. 11 (14) (2023) 2203104, http://dx.doi.org/10.1002/adom.202203104.
[18] R. Zhang, M. Guo, M. Li, H. Tang, T. Zhang, H. Li, Surface-enhanced Raman spectroscopy combined with chemometrics for quantitative analysis and carcinogenic risk estimation of polycyclic aromatic hydrocarbons in water with complex matrix, Chemometr. Intell. Lab. Syst. 257 (2025) 105293, http://dx.doi.org/10.1016/j.chemolab.2024.105293.
[19] Z. Shi, J. Wang, Y. Su, Z. Liang, J. Zi, C. Wang, H. Bi, X. Xiang, Transfer contrastive learning for Raman spectra data of urine: Detection of glucose, protein, and prediction of kidney disorders, Chemometr. Intell. Lab. Syst. 261 (2025) 105384, http://dx.doi.org/10.1016/j.chemolab.2025.105384.
[20] M. Wu, et al., Simulation and quantitative analysis of raman spectra in chemical processes with autoencoders, Chemometr. Intell. Lab. Syst. 248 (2024) 105119, http://dx.doi.org/10.1016/j.chemolab.2024.105119.
[21] M. Wu, U. Di Caprio, O. Van Der Ha, B. Metten, D. De Clercq, F. Elmaz, S. Mercelis, P. Hellinckx, L. Braeken, F. Vermeire, M.E. Leblebici, ConInceDeep: A novel deep learning method for component identification of mixture based on Raman spectroscopy, Chemometr. Intell. Lab. Syst. 234 (2023) 104757, http://dx.doi.org/10.1016/j.chemolab.2023.104757.
[22] H. Jabbar, I.A. Zgair, K. Heydaryan, S.A. Kadhim, S. Mehmandoust, V. Eskandari, H. Sahbafar, Applications of artificial intelligence and machine learning in combination with surface-enhanced Raman spectroscopy (SERS), Chemometr. Intell. Lab. Syst. (2025) 105445, http://dx.doi.org/10.1016/j.chemolab.2025.105445.
[23] L. Luciani, A. Nocera, M. Raimondi, G. Ciattaglia, S. Spinsante, E. Gambi, R. Galassi, Raman spectroscopy for nitrate detection in water: A review of the current state of art, ACS Meas. Sci. Au 5 (4) (2025) 443–460, http://dx.doi.org/10.1021/acsmeasuresciau.5c00016.
[24] Y. Zhu, S. Wang, J. Lv, J. Yin, D. Kong, Study on the identification of microplastics in agricultural soils using segmented GASF combined with deep learning based on confocal micro-Raman spectroscopy, Measurement (2025) 118226, http://dx.doi.org/10.1016/j.measurement.2025.118226.
[25] J. Zhu, J. Deng, F. Meng, Y. Jia, E. Tian, H. Jiang, Quantitative analysis of petroleum derivatives in vegetable oils based on vibration spectrum data fusion and multi-space optimized LASSO feature selection model, Measurement (2025) 118010, http://dx.doi.org/10.1016/j.measurement.2025.118010.
[26] C. Lange, M. Altmann, D. Stors, S. Seidel, K. Moynahan, L. Cai, S. Born, P. Neubauer, M.N.C. Bournazou, Deep learning for Raman spectroscopy: Benchmarking models for upstream bioprocess monitoring, Measurement (2025) 118884, http://dx.doi.org/10.1016/j.measurement.2025.118884.
[27] S. Srivastava, W. Wang, W. Zhou, M. Jin, P.J. Vikesl, Machine learning-assisted surface-enhanced Raman spectroscopy detection for environmental applications: a review, Environ. Sci. Technol. 58 (47) (2024) 20830–20848, http://dx.doi.org/10.1021/acs.est.4c06737.
[28] R. Luo, J. Popp, T. Bocklitz, Deep learning for Raman spectroscopy: a review, Analytica 3 (3) (2022) 287–301, http://dx.doi.org/10.3390/analytica3030020.
[29] Y. Luo, W. Su, X. Xu, D. Xu, Z. Wang, H. Wu, B. Chen, J. Wu, Raman spectroscopy and machine learning for microplastics identification and classification in water environments, IEEE J. Sel. Top. Quantum Electron. 29 (4) (2023) 1–8, http://dx.doi.org/10.1109/JSTQE.2022.3222065.
[30] M.J. Moorcroft, J. Davis, R.G. Compton, Detection and determination of nitrate and nitrite: a review, Talanta 54 (5) (2001) 785–803, http://dx.doi.org/10.1016/S0039-9140(01)00323-X.
[31] T. Küster, G.D. Bothun, In situ SERS detection of dissolved nitrate on hydrated gold substrates, Nanoscale Adv. 3 (14) (2021) 4098–4105, http://dx.doi.org/10.1039/D1NA00156F.
[32] B. Lafuente, R.T. Downs, H. Yang, N. Stone, The power of databases: the RRUFF project, in: Highlights in Mineralogical Crystallography, vol. 1, 2015, p. 25, http://dx.doi.org/10.1515/9783110417104-003.
[33] C. Post, S. Brülisauer, K. Waldschläger, W. Hug, L. Grüneis, N. Heyden, S. Schmor, A. Förderer, R. Reid, M. Reid, R. Bhartia, Q. Nguyen, H. Schüttrumpf, F. Amann, Application of laser-induced, deep UV Raman spectroscopy and artificial intelligence in real-time environmental monitoring—solutions and first results, Sensors 21 (11) (2021) 3911, http://dx.doi.org/10.3390/s21113911.
[34] C. Post, N. Heyden, A. Reinartz, A. Foerderer, S. Bruelisauer, V. Linnemann, W. Hug, F. Amann, Possibilities of real time monitoring of micropollutants in wastewater using laser-induced Raman & fluorescence spectroscopy (LIRFS) and artificial intelligence (AI), Sensors 22 (13) (2022) 4668, http://dx.doi.org/10.3390/s22134668.
[35] N.M. North, J.B. Clark, A.A.A. Enders, A.J. Grooms, S.G. Wairegi, K.A. Duah, E.I. Palassis-Naziri, A. Badu-Tawiah, H.C. Allen, Multi-analyte concentration analysis of marine samples through regression-based machine learning, ACS Earth Space Chem. 8 (8) (2024) 1549–1559, http://dx.doi.org/10.1021/acsearthspacechem.4c00018.
[36] BIPM, et al., Evaluation of Measurement Data — Guide to the Expression of Uncertainty in Measurement, vol. 100, Joint Committee for Guides in Metrology, JCGM, 2008, https://www.bipm.org/documents/20126/2071204/JCGM_100_2008_E.pdf/cb0ef43f-baa5-11cf-3f85-4dcd86f77bd6.
[37] P. Hildebrandt, F. Siebert, Vibrational Spectroscopy in Life Science, Wiley-VCH Verlag GmbH, 2007, 10.1002/9783527621347.
[38] M.D. Fontana, K. Ben Mabrouk, T.H. Kauffmann, Raman spectroscopic sensors for inorganic salts, 2013, http://dx.doi.org/10.1039/9781849737791-00040.
[39] K.H. Liland, A. Kohler, N.K. Afseth, Model-based pre-processing in Raman spectroscopy of biological samples, J. Raman Spectrosc. 47 (6) (2016) 643–650, http://dx.doi.org/10.1002/jrs.4886.
[40] D. Georgiev, S.V. Pedersen, R. Xie, A. Fernández-Galiana, M.M. Stevens, M. Barahona, RamanSPy: An open-source python package for integrative Raman spectroscopy data analysis, Anal. Chem. 96 (21) (2024) 8492–8500, http://dx.doi.org/10.1021/acs.analchem.4c00383.
[41] T. Bocklitz, A. Walter, K. Hartmann, P. Rösch, J. Popp, How to pre-process Raman spectra for reliable and stable models?, Anal. Chim. Acta 704 (1–2) (2011) 47–56, http://dx.doi.org/10.1016/j.aca.2011.06.043.
[42] S. Baek, A. Park, Y. Ahna, J. Choo, Baseline correction using asymmetrical reweighted penalized least squares smoothing, Analyst 140 (1) (2015) 250–257, http://dx.doi.org/10.1039/C4AN01061B.
[43] S. Ma, C. Xu, Research on the quantitative analysis for detection sulfate using laser Raman spectroscopy, in: 3rd International Conference on Laser, Optics, and Optoelectronic Technology, LOPET 2023, vol. 12757, SPIE, 2023, pp. 410–418, http://dx.doi.org/10.1117/12.2690267.
[44] C. Laia, X. Chena, X. Jiang, J. Xiang, H. Tanga, Quantitative analysis of nitrides in water by Raman spectroscopy based on deep learning and relative position matrix, 2025, http://dx.doi.org/10.2139/ssrn.5275013, Available at SSRN 5275013.