Infrared Physics & Technology 141 (2024) 105452

Contents lists available at ScienceDirect

Infrared Physics and Technology

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

Development of a semi-supervised machine learning based noise filter for quantum cascade laser-coupled mid-infrared spectrometer

Soumyadipta Chakraborty, Indrayani Patra, Ardhendu Pal, Koushik Mondal, Manik Pradhan *

Department of Chemical and Biological Sciences, S. N. Bose National Centre for Basic Sciences, Salt Lake, JD Block, Sector-III, Kolkata 700106, India

ARTICLE INFO

Keywords:
Quantum cascade laser
Machine learning
Principal component analysis
Spectroscopy
Noise reduction
Mid-IR

ABSTRACT

The rapid and accurate quantitative determination of molecular rovibrational spectral features is highly desirable for applications in fundamental high-resolution spectroscopy, trace chemical sensing, and environmental monitoring. But significant spectral noise can impede the accuracy and precision of such spectral measurements as well as the precise classification of fine and hyperfine spectral fingerprints. Nevertheless, the denoising of weak rovibrational spectral data has long been a computational and experimental challenge. Here, we develop a new semi-supervised machine learning (SSML) denoising method combining unsupervised spectral eigenvector space dimensionality reduction of the spectral data matrix exploiting principal component analysis with supervised Fourier domain residual analysis for enhancing high-resolution rovibrational spectral signal quality. We discussed the detailed implementation of the SSML algorithm on the real experimental mid-infrared rovibrational spectral features of nitrogen dioxide (NO_2) in the gas-phase acquired from a quantum cascade laser-coupled cavity ring-down spectroscopic method. Our results confirm the robustness and feasibility of the SSML approach and could lead to a wide range of pertinent applications in infrared spectroscopy including precise spectral signal analysis, concentration retrieval and performance enhancement of laser-coupled spectroscopic sensors.

1. Introduction

Optical spectroscopy relies on light-matter interaction, which can provide valuable insights on the composition and properties of materials. Absorption, emission, and scattering of light at various wavelengths lead to important signatures of the characteristics of molecules and their behaviour with the environment. Rovibrational spectroscopy, an integral part of optical molecular spectroscopy, studies the interaction of the rotation and vibration of the molecule in the infrared (IR) and mid-infrared (mid-IR) molecular fingerprint regions. It is important to understand the molecular structure and dynamics by interpreting the rovibrational spectra of the molecules. Furthermore, various fundamental aspects of molecular physics like molecular collision dynamics, interactions, fine and hyperfine structures can be manifested using rovibrational spectroscopy. Gas-phase IR rovibrational spectroscopy can be crucial for various applications, such as trace gas sensing, environmental analysis, the study of atmospheric sciences as well as non-invasive disease detection via exhaled breath analysis [1–5]. In all cases, the accuracy and precision of the measurements become critical factors.

However, the precision and accuracy of the spectroscopic measurements may be hindered by the noise present in the spectral data. Such noisy fluctuations can originate from electronic interferences, detector glitches, scattered stray light, imperfections, or misalignments in the optical setups, as well as other environmental conditions such as vibrations and temperature fluctuations. These noise fluctuations can be in the form of white noise variation or as pink noise. Although the use of high frequency lock-in detection can suppress the pink fluctuations, the inherent instrumental noise may still persist. Spectroscopic investigations such as concentration retrieval and gas analysis utilizing rovibrational absorption transitions, the study of specific rovibrational profiles, the examination of hyperfine splitting, investigations into spectral narrowing, speed dependent phenomena, etc. may be significantly affected by the presence of large spectral noise. Analytical results could be impaired by the noise interference. Therefore, over the decades, denoising in molecular spectroscopic data has become an important challenge and has not yet been fully implemented in high-resolution rovibrational spectral features in the gas-phase, enabling a

* Corresponding author at: Department of Chemical and Biological Sciences, S. N. Bose National Centre for Basic Sciences, Salt Lake, JD Block, Sector-III, Kolkata 700106, India.
E-mail address: manik.pradhan@bose.res.in (M. Pradhan).

https://doi.org/10.1016/j.infrared.2024.105452
Received 7 May 2024; Received in revised form 8 July 2024; Accepted 16 July 2024
Available online 16 July 2024

wide range of applications in physics, chemistry, biology, and medical diagnostics. In the recent times, a flourishing interest is observed in development of machine learning and deep learning-based noise reduction pipelines [6–11]. In a recent work, principal component analysis was used to reduce the noise of solid state NMR spectra [12]. To address the challenge of noise mitigation, we provide here a unique proof-of-concept semi-supervised machine learning (SSML) denoising method combining unsupervised spectral eigenvector space dimensionality reduction of the spectral data matrix with supervised Fourier domain residual analysis for optimization of the spectral filter conditions. Our aim was to explore the potential of Principal Component Analysis (PCA), a dimensionality reduction technique, in reducing the rovibrational spectral noise of an IR spectrometer. For instance, here we have used a quantum cascade laser (QCL)-operated cavity ring-down spectrometer (CRDS) working in the mid-IR region. PCA[13,14] is a vital unsupervised multivariate analysis method extensively employed in chemometrics[15–17] for data exploration, visualization, and dimensionality reduction. In chemometrics, PCA is utilized to handle complex datasets with numerous interrelated variables, commonly encountered in fields such as spectroscopy[18–26] and chromatography [27–31]. PCA serves as an invaluable tool to extract useful insights from the rovibrational spectra of materials, which can have important applications in various fields. It can be used for the classification of IR vibrational spectral bands [32]. A recent work shows that argan oil adulteration was investigated by identifying samples based on the rovibrational spectra in the MIR and NIR regions using PCA and other chemometric algorithms [33]. A PCA-coupled backpropagation neural network-based algorithm was developed for multi-gas sensing applications based on the rovibrational mid-IR spectra of CH_4 and C_2H_6 [34]. A recent review article shows the application of chemometric methods like PCA on the rovibrational FTIR-based mid-infrared spectral features for widespread usage in the food industry, medicine, quality control, etc [35]. The SSML approach is important in applied spectroscopy, and recent investigations show the application of semi-supervised deep learning-based multi-component spectral calibration modeling for UV−vis and NIR spectroscopy[36], quantification of milk adulteration using a semi-supervised deep learning-based regression framework based on NIR spectral data[37], identification of explosives based on laser-induced breakdown spectroscopy data with accuracy enhancement of the K-nearest neighbor model by a supervised learning approach[38], etc. Though the SSML approach presents multifaceted usage in applied spectroscopy, according to the best of our knowledge, its investigation on the efficacy of noise-mitigation of gas phase rovibrational spectral data remains to be explored.

PCA transforms the basis of the multidimensional data into a new hyperspace spanned by the eigenvectors, which are known as the principal components. These principal components (PCs) can be expressed as linear combinations of the previous basis set. The PCs can identify the dominant patterns of the data and explain the variances. They are ordered in terms of the amount of variance elucidated by them. By creating a subspace of the actual eigenvector space, retaining the most significant PCs, a dimensionality reduction of the PC space is facilitated, which can effectively reduce the complexity of the data set while preserving the essential information. This dimensionality reduction aids in the visualization, interpretation, and analysis of chemometric applications. PCA leverages the inherent structure within the data to identify and extract relevant signal components while suppressing noise contributions. A typical cavity ring-down spectrometer utilizes complex electronics. In field measurements, dealing with electronic glitches and laser intensity fluctuations that introduce noise into the signal can be challenging. A CRD spectrometer employs mode-matching phenomena to excite TEM_{00} mode. This requires fine optical alignment. A slight deviation from the optimal alignment condition can excite other higher-order modes, and this can cause noise insertion in the optical signal. Implementing hardware-based noise reduction for in-situ measurements is difficult. Therefore, our motivation was to devise a simple and highly accurate SSML noise filtering method that would work on the experimental data from the cavity ring-down spectrometer in both laboratory-based benchtop and portable configurations. To accomplish this, we utilized our laboratory-developed mid-IR CRD spectrometer and artificially infused white noise into the obtained data to increase the noise level of the signal and simulate a faulty instrumental condition. Our work investigates the efficacy of a PCA-based filter for signal revival and enhancement of the signal-to-noise ratio (SNR) of the rovibrational spectral signatures obtained from a CRDS method. Therefore, a coupling of the SSML technique with analytical spectroscopy can be pivotal in enhancing the sensitivity of spectroscopic measurements, precise classification of fine spectral splitting, trace gas sensing and consequently this approach may be significant in diverse applications due to its simple and robust implementation.

2. Methods

2.1. A brief overview of the spectroscopic technique

Cavity ring-down spectroscopy (CRDS) is used for measuring the absorption properties of samples. It depends on the principle of trapping laser radiation within a high-finesse optical cavity formed by two or more highly reflective mirrors. In CRDS, the ring-down time serves as a fundamental measurement parameter, representing the lifetime of a photon inside the cavity. In an empty cavity, the ring-down time (τ_0) signifies the time required for the leaked intensity from the cavity to decay to 1/e times of the incident intensity, providing a measure of the intrinsic losses within the cavity. When a sample is introduced into the cavity, altering the absorption characteristics, the ring-down time (τ) decreases due to additional losses caused by the sample. The decay rate, inversely proportional to the ring-down time, quantifies the rate at which the leaked intensity decreases over time and is influenced by both the cavity properties and the sample absorption. The decay rate is given by the following equation[39]:

$$\alpha = \frac{1}{c}\left(\frac{1}{\tau} - \frac{1}{\tau_0}\right) = \frac{\Delta k}{c} = \sigma_\lambda [X] \tag{1}$$

Here, α is the absorption coefficient of the sample, Δk corresponds to the change in decay rate,c is the speed of the light, σ_λ is the wavelength-dependent absorption coefficient, and $[X]$ is the sample concentration. A change in wavelength indicates a change in the absorption coefficient, which is manifested in the alteration of decay rates. The concentration of an analyte can be calculated by employing the area under the curve (AUC) of the decay rate versus wavelength variation.

2.2. Computational overview

Let us consider that we have n observations for decay rates as we have captured n wavenumbers. Suppose, we took k CRD scans. Therefore, our data has a $n \times k$ dimension. Let us define our spectral data set as the matrix D. We standardize our data by subtracting the mean of each column (μ_k) from the respective columns of D and normalized by the standard deviation (σ_k) of that column:

$$\overline{D_{ij}} = \frac{D_{ij} - \mu_j}{\sigma_j} \tag{2}$$

This creates a standardized CRD dataset $\overline{D}$. Then we compute the covariance matrix Σ using the standardize data set $\overline{D}$. The covariance matrix can be calculated as:

$$\Sigma = \frac{\overline{D}^T \overline{D}}{n} \tag{3}$$

Subsequently, we performed eigenvalue decomposition of the covariance matrix:

$$\Sigma v_i = \lambda_i v_i \tag{4}$$

where, λ_i is the i-th eigenvalue and v_i represents the i-th eigenvector or the i-th principal component. The eigenvectors follow the orthonormal condition:

$$v_i^T v_j = \delta_{ij} \tag{5}$$

Let, a matrix V stores the principal components (eigenvectors). It can be considered as a collection of the individual eigenvectors and can be represented as:

$$V = [v_1, v_2, v_3, \cdots, v_k] \tag{6}$$

The total number of eigenvectors for the covariance matrix is same as the number of the scans taken for the cavity ring-down measurements. The transformed data can be obtained by using the following projection operation:

$$Z = \overline{D}\, V \tag{7}$$

Z contains the projected data points in the new space defined by the principal components. An element $Z_{i,j}$ can be considered as the transformed data point for the i-th wavenumber and j-th eigenvector. After projecting the data into the new space defined by the principal components, it is important to reconstruct the data by again projecting it into the previous basis. This reconstruction can be achieved by the following operation with a necessary rescaling of the data:

$$\widehat{D} = ZV^T diag(\sigma) + 1\mu^T \tag{8}$$

If we consider the total number of eigenvectors for the data reconstruction, then the dimensions of the various matrices are: $D, \overline{D}, Z, \widehat{D} \rightarrow n \times k$ and $\Sigma, V \rightarrow k \times k$, respectively.

Now, suppose we select a subset of the eigenvectors. Let us consider that we are reconstructing the data after selecting the first d principal components such as $d < k$. Then the dimensions of the various matrices are: $D, \overline{D}, \widehat{D} \rightarrow n \times k$ *and* $\Sigma \rightarrow k \times k,\ V \rightarrow k \times d,\ Z \rightarrow n \times d$, respectively. While the dimension of the reconstructed matrix is the same as that of the initial data matrix, it only retains the information captured by the first d eigenvectors.

2.3. Experimental Section

In this study, we employed the CRDS technique in conjunction with a continuous-wave (cw) external-cavity quantum cascade laser (cw EC-QCL). The experimental setup, which can be found elsewhere[40], was developed in our laboratory. Briefly, we utilized a highly versatile water-cooled cw EC-QCL (Model: 41062-MHF; Manufacturer: Daylight Solutions, USA) with a mode-hop-free (MHF) tuning capability covering the range of 5.88 μm to 6.50 μm (equivalent to 1701–1538 cm^{-1}). This laser was characterized with a high-power output (>120 mW) and a narrow linewidth (~0.0003 cm^{-1}). Wavenumber measurements were conducted using a wavemeter (Model: 621B-MIR; Manufacturer: Bristol Instruments) with an accuracy of 0.001 cm^{-1}. To construct the high-finesse optical cavity, we utilized a 50 cm long quartz-coated cylindrical ring-down cell (RDC), fitted with two highly reflective (HR) mirrors (Reflectivity > 99.98 %, Manufacturer: CRD Optics Inc., USA) at each end. To achieve periodic resonance between the laser and cavity modes, it was necessary to employ a cavity-length modulation technique. This was accomplished by applying a ramp voltage to three piezo-electric transducers (PZT, Manufacturer: Thorlabs PE4) attached to the HR mirror of the RDC. The optical signal leaking out from the cavity was directed onto a mercury-cadmium-telluride (MCT) detector (Model: PVI-4TE-8-1X1, Manufacturer: Vigo Systems S.A.), followed by amplification through a preamplifier (Model: SR560; Manufacturer: Stanford Research Systems). Data acquisition was performed using a high-speed data acquisition card (Model: PCI 5122, 14-bit, 100 MHz bandwidth, Manufacturer: National Instruments), with subsequent analysis conducted using a custom LabVIEW program. Pressure inside the optical cavity was monitored using a pressure gauge from Pfeiffer Vacuum (Model: CMR 361).

3. Results and Discussions

We introduced a dilute mixture of nitrogen dioxide (INTERGAS, Newfield Industrial Estate, 99.5 % Grade) into the optical cavity of the CRD spectrometer. A wavenumber scan was conducted from 1620.0413 cm^{-1} to 1620.1410 cm^{-1}, while maintaining the pressure inside the optical cavity at 1 Torr (0.0013 atm). A variation in the decay rate was observed, corresponding to changes in the wavenumber. Surges in decay rate values were observed for peaks in the absorption coefficients at the transition wavenumbers. Fig. 1(a) illustrates the experimental variation of decay rate with wavenumber obtained from CRDS.

Synthetic Gaussian noise, with a standard deviation of 10^4 s^{-1} and a zero mean, was added to identical experimental CRD spectra to elevate the noise level, mimicking the behaviour of a noisy CRD spectrometer and reducing the SNR. Therefore, noisy CRD scans were generated by adding random noise to the actual experimental CRD spectra. Fig. 1(b) illustrates the resulting noisy CRD spectrum for a particular scan. A multi-peak Gaussian function was used to fit the CRD scans. The multivariate data from various CRD scans were obtained and collected in the form of an input spectral data matrix. A Python script was developed to implement PCA on the CRD data and reconstruct the noise-mitigated spectra. The PCA operations were performed using the scikit-learn library[41], which is a widely used toolkit for ML in Python. The eigenvalues of the covariance matrix were computed to analyse the main variance of the signal patterns. It was observed that the eigenvalues exhibited a steady decrease as the number of eigenvectors increased (Fig. 2), indicating the incorporation of noise behaviours into the main signal variances. The maximum signal variance was shown by the first eigenvector, or the first principal component (PC_1).

The signal retrieval was assessed in terms of SNR defined as[11]:

$$Retrived\ SNR(dB) = 10log_{10}\left(\frac{\sigma\left(Signal_{Experiment}\right)}{\sigma\left(Signal_{Experiment} - Signal_{Reconstructed}\right)}\right) \tag{9}$$

Here, σ refers to the standard deviation, $Signal_{Experiment}$ refers to the initial CRD scan obtained experimentally, and $Signal_{Reconstructed}$ refers to the reconstructed CRD scan obtained after noise filtering achieved by PCA. An increased value of the retrieved SNR indicates a more accurate recovery of the true experimental signal from the noisy data. Eventually, we investigated the effect of the number of scans on the signal recovery capacity of this method, while keeping a fixed principal component. Here, we chose PC_1 for the spectra reconstruction as it preserves the maximum variance. Fig. 3(a) depicts the variation of the retrieved SNR with the number of spectral scans for reconstruction corresponding to PC_1. Similar variations can be observed for spectra reconstruction using higher PCs, albeit with lower SNR values.

The variation of the retrieved SNR with the number of CRD scans corresponding to a signal reconstruction up to a particular principal component can be modelled by a single exponential model:

$$Retrieved\ SNR = A_{PC_j}\exp\left(\frac{-Scan\ no.}{t_{PC_j}}\right) + SNR_O \tag{10}$$

An increase in the number of scans leading to a higher retrieval SNR may be attributed to the PCA filter capturing a larger number of variances in the data from the high-dimensional input. The term SNR_O in equation (10) refers to an offset that can account for the retrieved SNR achieved for a very high number of CRD scans. For signal reconstruction with PC_1,

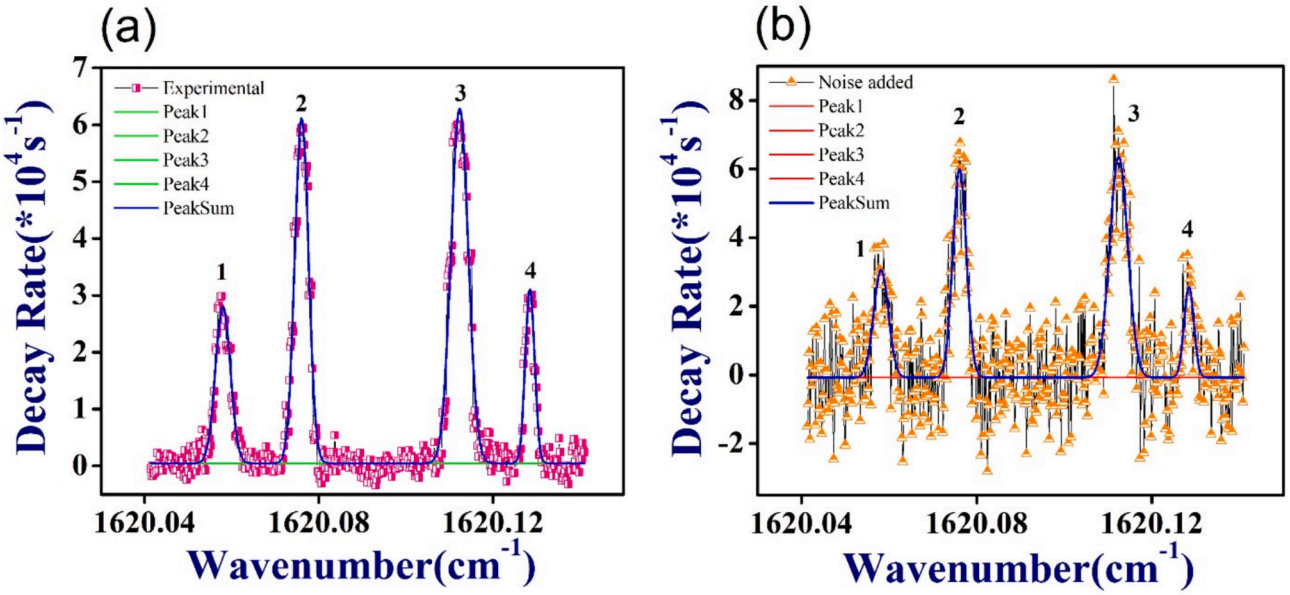

Fig. 1. (a) Experimental cavity ring-down spectra of nitrogen dioxide molecule. (b) Synthetic noise added cavity ring-down spectra.

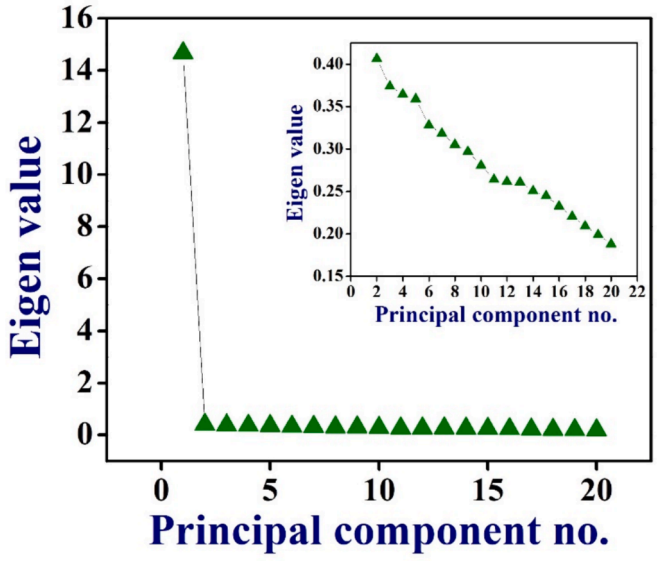

Fig. 2. A scree plot showing the variation of eigenvalue with the principal component number. The inset section shows the variation excluding the first principal component.

the values of the obtained regressor parameters are: $A_{PC_1} = -6.52233(fiterror : 0.1624)$, $t_{PC_1} = 12.56783(fiterror : 0.93527)$, and $SNR_O = 9.73947(fiterror : 0.13699)$. Next, we investigated the contribution of noise insertion by reconstructing spectra with higher principal components. The retrieved SNR was calculated for the reconstructed spectra, revealing that the introduction of noise became more prominent with the inclusion of higher principal components. Consequently, the SNR decreased with an increase in the number of principal components. Fig. 3(b) illustrates a decrease in the SNR with an increase in the principal component number for a particular number of scans (in this case, 20 scans). A maximum limit of PC_3 can be chosen for the reconstruction because beyond that point, the SNR diminishes significantly. A difference between the actual CRD spectra and the reconstructed spectra can be considered as residuals. It can be observed that the standard deviation of the residuals increases with an increase in the number of principal components for a particular number of scans, indicating the contribution of noise insertion in the spectra reconstruction. Fig. 4(a) and (b) visually illustrate this phenomenon of noise insertion, respectively.

The frequency spectrum of the residual was investigated to observe any systematic variations resulting from the filtering process mediated by PCA. Fig. 4(c) and (d) illustrate the frequency domain analysis of the PCA filter for different configurations. Frequency domain residual analysis can be utilized to explain the phenomenon of inaccuracies in spectral information retrieval resulting from possible signal drifts. It can be observed from Fig. 4(c) that the residual of the noisy CRD spectra showed large magnitude fluctuations, but these fluctuations were approximately uniform across all frequency ranges. This may be attributed to the randomness of the signal variations. For a particular PC, as the SNR increases with an increase in the number of scans. Therefore, the magnitude of the fluctuations decreases with a higher number of scans (20 and 50, depicted here). There are no high amplitudes at the higher frequency components, indicating the absence of any fast-varying noise in the filtered CRD spectra. However, it is interesting to note that a slight increase in the amplitude of the fluctuation at the zeroth frequency can be observed for the reconstructed spectra obtained from 50 scans, as shown in Fig. 4(c). This increment becomes more prominent with an increase in the number of scans, as seen in Fig. 4(d), where increasing the number of scans to 100 results in a high amplitude of fluctuation at the zeroth frequency. This can be attributed to an inaccuracy in spectral information retrieval, despite the increase in retrieval SNR. A spike at the zeroth frequency may indicate the drift of the baseline component of the signal. This issue may be addressed at the cost of retrieval SNR by reconstructing the spectra using higher PCs. In Fig. 4(d), it can be seen that PC_2 shows a decrease in amplitude at the zeroth frequency. However, it is possible that this spike occurs for higher PCs immediately following PC_1 as well. Hence, Fourier analysis must be conducted for various PCA filter conditions to address the issue of spectral information retrieval reliability, albeit at the expense of retrieval SNR. Additionally, the number of spectral scans acquired depends on the experimental feasibility. Based on the analysis above, an algorithm for reducing noise in CRDS data using PCA is proposed, with the condition that we are restricted in the number of experimental CRD scans we can take, and that we have an actual low-noise experimental CRD scan available to compare. A detailed flow chart of the proposed algorithm showing the optimization of the PCA filter conditions based

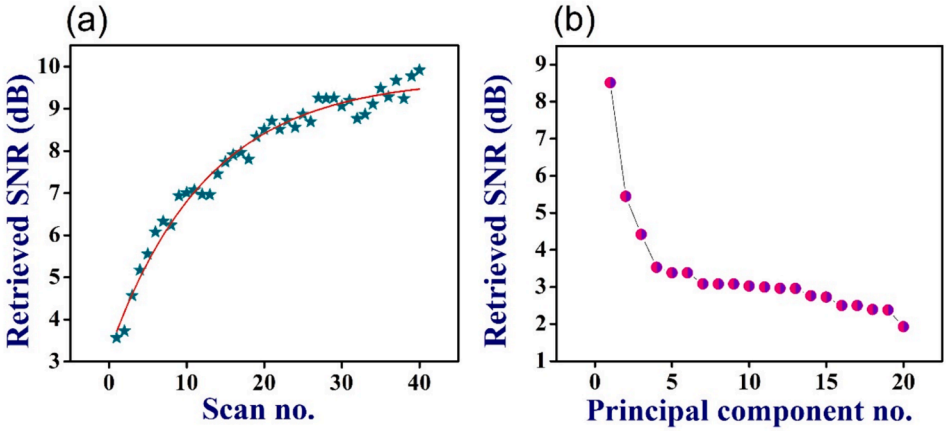

Fig. 3. (a)Variation of the retrieved SNR with the number of CRD scans. (b) Variation of the retrieved SNR corresponding to principal component numbers, for 20 CRD scans.

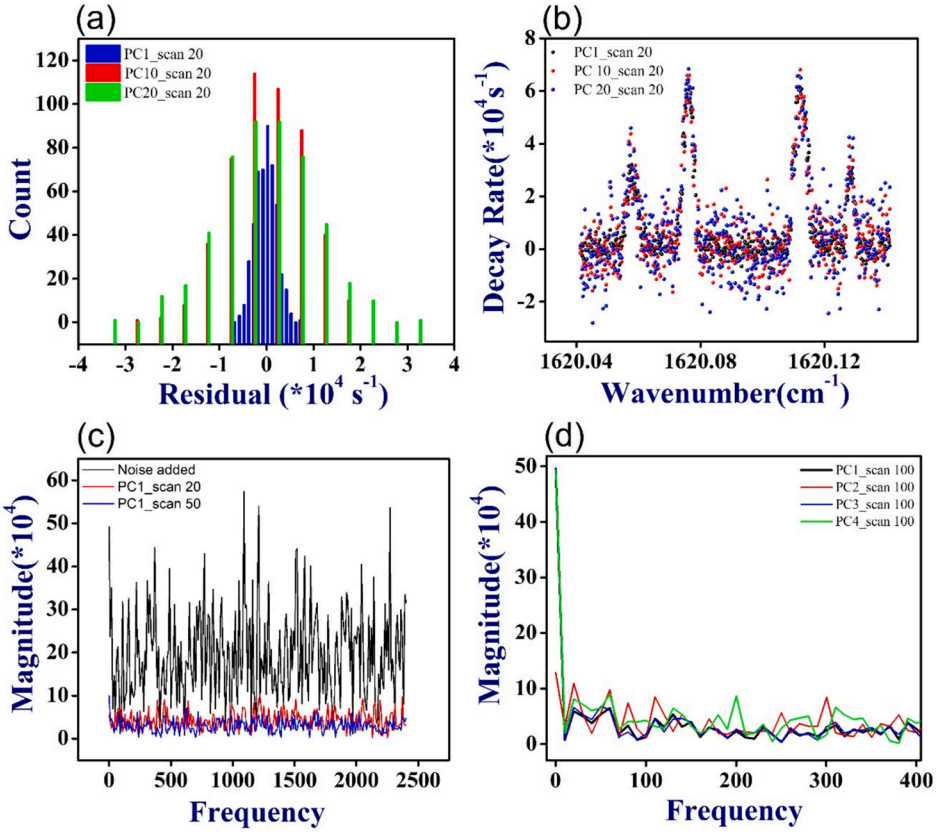

Fig. 4. (a) Distribution of residuals for the reconstructed CRD spectra for 20 scans corresponding to principal component 1, principal component 10, and principal component 20, respectively. (b) Reconstructed CRD spectra for 20 scans corresponding to principal component 1, principal component 10, and principal component 20, respectively. (c) Fourier domain residual analysis of noise added spectra and reconstructed spectra corresponding to the first principal component of 20 and 50 CRD scans. (d) Fourier domain residual analysis of the reconstructed spectra corresponding to the principal component up to 1st, 2nd, 3rd, and 4th component for 100 CRD scans.

on Fourier analysis of the spectral residuals is given in Fig. 5.

Fig. 6(a) presents the experimentally obtained CRD scan along with the reconstructed CRD scan, based on a PCA filter condition of 50 scans and PC_1. Additionally, the extracted mixing ratio of the NO_2 mixture from the third peak observed in the actual experimental CRD scan (Fig. 1 (a)) is presented in Fig. 6(b). Furthermore, we utilized the same peak of NO_2, with a line intensity of 4.036×10^{-20} cm/molecule (taken from the HITRAN online database), to retrieve the mixing ratios under various PCA filter conditions. The AUC corresponding to the third peak was employed for concentration measurements. Fig. 6(b) depicts the concentrations retrieved from the various CRD spectra. The concentration levels vary depending on the PCA filter conditions. The same figure illustrates that the mixing ratio is significantly overestimated when retrieving concentrations from noisy CRD spectra. Additionally, there is a large uncertainty in the mixing ratio due to high errors in the AUC resulting from line-shape fitting. Furthermore, these uncertainties

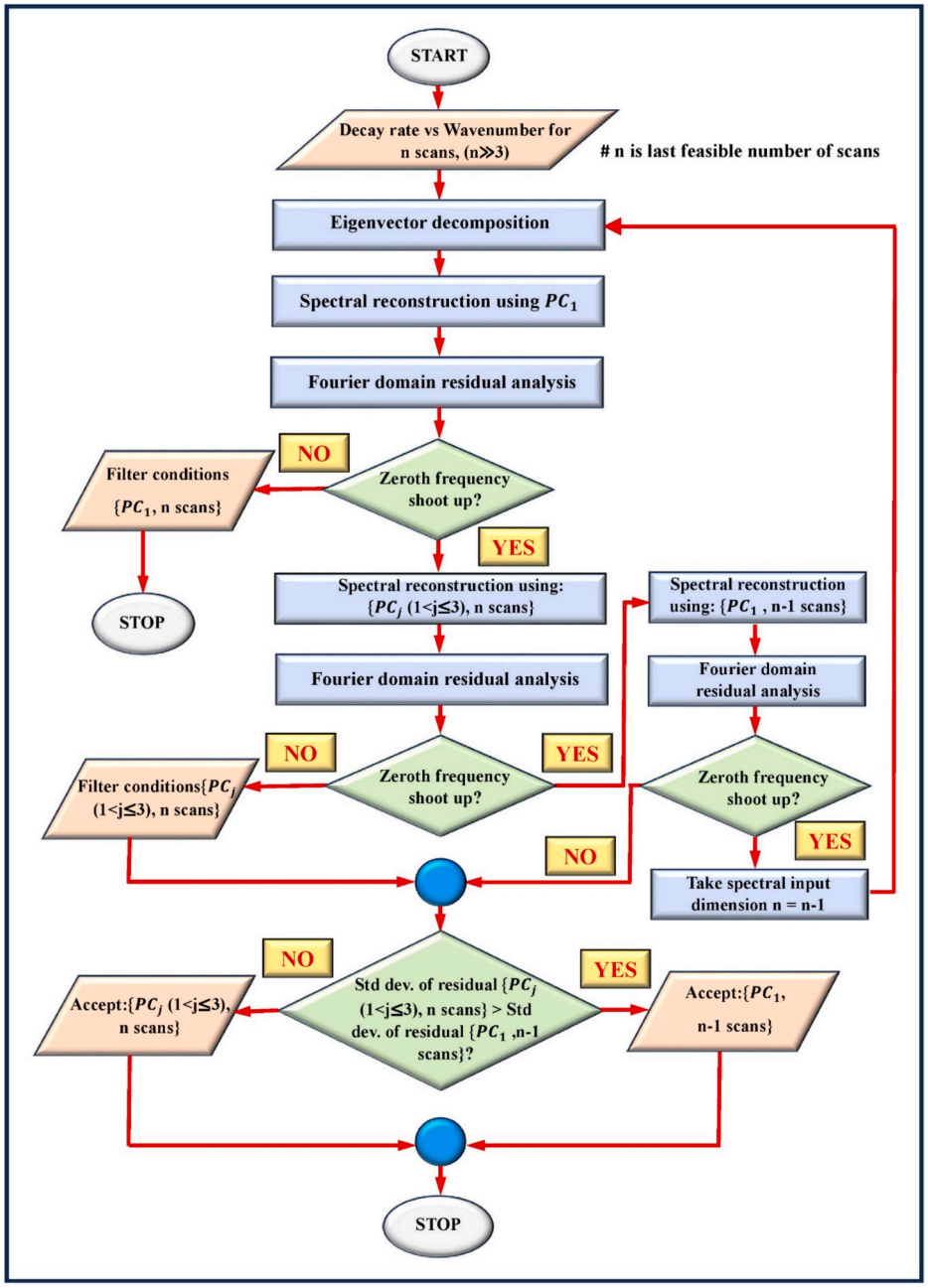

Fig. 5. A flowchart depicting the proposed SSML algorithm.

decrease when reconstructing CRD spectra using a high dimensional spectral input matrix and a lower PC number. This decrease can be attributed to the increased correlation in profile fitting, which leads to reduced uncertainties in the AUC, owing to the improved retrieval SNR. Conversely, for reconstructions with fewer scans, such as 20 scans, the mixing ratio is underestimated. Concentrations retrieved from CRD spectra reconstructed using a higher number of scans, such as 100 scans, tend to overestimate the mixing ratio compared to the actual experimental spectra. This discrepancy can be attributed to the shifting of the baseline component of the signal, as addressed by frequency domain analysis. A higher number of spectral scans can refer to a high-dimensional eigenvector space. Therefore, the PCA-spectral filter gets a chance to study the variance from a high dimensional input, and the captured variances that can be represented as eigenvalues are usually larger compared to the variances captured for a lower dimensional input data matrix. Spectral reconstruction using only the first few eigenvectors for a high dimensional eigenvector space can lead to a leak of spectral variance in the form of noise. This may attribute the phenomena of baseline shifting and reduction in spectral signal retrieval accuracy, though the precision of measurement may increase. Hence, in practical scenarios, one can apply the proposed algorithm, which employs PCA and Fourier domain residual analysis for noise reduction and spectral information retrieval.

4. Conclusions

In summary, we have explored the performance of a principal component analysis (PCA)-based denoising method in a rovibrational infrared spectrometer operated under simulated faulty conditions. To judge the efficacy of our proposed method, we retrieved NO_2 gas mixture concentrations using both the original experimental data and reconstructed data obtained from various PCA filter conditions. Thus, to

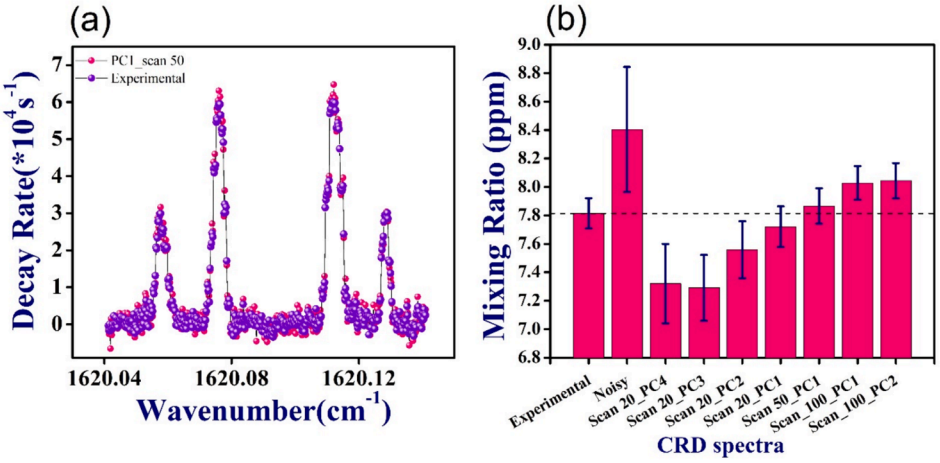

Fig. 6. (a) Actual experimental CRD scan and reconstructed CRD scan based on a PCA filter condition of 50 scans and PC_1. (b) Concentrations retrieved from various CRD spectra.

ensure the reliability of spectral signal retrieval and minimize spectral information loss, we have introduced an algorithm that utilizes PCA in collaboration with Fourier domain residual analysis to optimize the PCA filter conditions in real scenarios. Our findings are not only relevant to CRDS but also have broader implications for *in-situ* spectral measurements, highlighting the potential of machine learning methods for enhancing spectral signal quality. This investigation underscores the versatility and effectiveness of the PCA-based SSML technique in addressing noise challenges in spectroscopic data, paving the way for more robust and reliable spectral analysis methodologies for future applications in real-word infrared optical sensor development.

CRediT authorship contribution statement

Soumyadipta Chakraborty: Writing – original draft, Visualization, Validation, Methodology, Investigation, Formal analysis, Data curation, Conceptualization. **Indrayani Patra:** Writing – review & editing, Visualization, Validation, Formal analysis, Data curation. **Ardhendu Pal:** Writing – review & editing, Visualization, Validation, Data curation. **Koushik Mondal:** Writing – review & editing, Validation, Data curation. **Manik Pradhan:** Writing – review & editing, Supervision, Project administration, Funding acquisition.

Declaration of competing interest

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

Data availability

Data will be made available on request.

Acknowledgements

Prof. Manik Pradhan gratefully acknowledges the financial support from the S.N. Bose National Centre for Basic Sciences (SNBNCBS). Soumyadipta Chakraborty and Ardhendu Pal acknowledge SNBNCBS for the Senior Research Fellowship (SRF). Indrayani Patra acknowledges University Grant Commission (India) for the PhD research fellowship. Koushik Mondal thanks the Advanced Postdoctoral Research Programme (APRP) of SNBNCBS for the postdoctoral fellowship.

References

[1] M.W. Sigrist, R. Bartlome, D. Marinov, J.M. Rey, D.E. Vogler, H. Wächter, Trace gas monitoring with infrared laser-based detection schemes, Appl. Phys. B 90 (2008) 289–300, https://doi.org/10.1007/s00340-007-2875-4.
[2] P.L. Hanst, Infrared spectroscopy and infrared lasers in air pollution research and monitoring, Appl Spectrosc 24 (1970) 161–174, https://doi.org/10.1366/000370270774371930.
[3] T.J. Johnson, R.L. Sams, S.W. Sharpe, The PNNL quantitative infrared database for gas-phase sensing: a spectral library for environmental, hazmat, and public safety standoff detection, in: A.J. Sedlacek Iii, R. Colton, T. Vo-Dinh (Eds.), Providence, RI, 2004, 10.1117/12.515604.
[4] C. Wang, P. Sahay, Breath analysis using laser spectroscopic techniques: Breath biomarkers, spectral fingerprints, and detection limits, Sensors 9 (2009) 8230–8262, https://doi.org/10.3390/s91008230.
[5] B. Panda, A. Pal, S. Chakraborty, M. Pradhan, An EC-QCL based dual-species (CH4/N2O) detection method at 7.8 μm in mid-IR region for simultaneous applications of atmospheric monitoring and breath diagnostics, Infrared Phys. Technol. 125 (2022) 104261, https://doi.org/10.1016/j.infrared.2022.104261.
[6] L. Pan, P. Pipitsunthonsan, P. Zhang, C. Daengngam, A. Booranawong, M. Chongcheawchamnan, Noise reduction technique for raman spectrum using deep learning network, in: 2020 13th International Symposium on Computational Intelligence and Design (ISCID), IEEE, Hangzhou, China, 2020, pp. 159–163, 10.1109/ISCID51228.2020.00042.
[7] X. Fan, Y. Zeng, Y. Zhi, T. Nie, Y. Xu, X. Wang, Signal-to-noise ratio enhancement for Raman spectra based on optimized Raman spectrometer and convolutional denoising autoencoder, J Raman Spectroscopy 52 (2021) 890–900, https://doi.org/10.1002/jrs.6065.
[8] Y. Zeng, Z. Liu, X. Fan, X. Wang, Modified denoising method of Raman spectra-based deep learning for Raman semi-quantitative analysis and imaging, Microchem. J. 191 (2023) 108777, https://doi.org/10.1016/j.microc.2023.108777.
[9] L. Zhang, G. Tian, J. Li, B. Yu, Applications of absorption spectroscopy using quantum cascade lasers, Appl Spectrosc 68 (2014) 1095–1107, https://doi.org/10.1366/14-00001.
[10] Y. Li, J. Xia, J. Guo, D. Zou, T. Ma, H. Nie, J. He, B. Zhang, A neural network filter based high-sensitive MIR CO2 sensor, Measurement 224 (2024) 113896, https://doi.org/10.1016/j.measurement.2023.113896.
[11] S. Zhou, N. Liu, C. Shen, L. Zhang, T. He, B. Yu, J. Li, An adaptive Kalman filtering algorithm based on back-propagation (BP) neural network applied for simultaneously detection of exhaled CO and N2O, Spectrochim. Acta A Mol. Biomol. Spectrosc. 223 (2019) 117332, https://doi.org/10.1016/j.saa.2019.117332.
[12] Y. Kusaka, T. Hasegawa, H. Kaji, Noise reduction in solid-state NMR spectra using principal component analysis, J. Phys. Chem. A 123 (2019) 10333–10338, https://doi.org/10.1021/acs.jpca.9b04437.
[13] Principal Component Analysis for Special Types of Data, in: Principal Component Analysis, Springer-Verlag, New York, 2002: pp. 338–372. Doi: 10.1007/0-387-22440-8_13.
[14] S. Wold, K. Esbensen, P. Geladi, Principal Component Analysis, (n.d.).
[15] R. Bro, A.K. Smilde, Principal component analysis, Anal. Methods 6 (2014) 2812–2831, https://doi.org/10.1039/C3AY41907J.
[16] K. Kumar, Principal component analysis: Most favourite tool in chemometrics, Reson 22 (2017) 747–759, https://doi.org/10.1007/s12045-017-0523-9.
[17] M. Hubert, P.J. Rousseeuw, S. Verboven, A fast method for robust principal components with applications to chemometrics, Chemom. Intel. Lab. Syst. 60 (2002) 101–111, https://doi.org/10.1016/S0169-7439(01)00188-5.

[18] A.G. Ryder, Classification of narcotics in solid mixtures using principal component analysis and raman spectroscopy, J. Forensic Sci. 47 (2002) 15244J, https://doi.org/10.1520/JFS15244J.
[19] J.R. Beattie, F.W.L. Esmonde-White, Exploration of principal component analysis: Deriving principal component analysis visually using spectra, Appl Spectrosc 75 (2021) 361–375, https://doi.org/10.1177/0003702820987847.
[20] P. Pořízka, J. Klus, E. Képeš, D. Prochazka, D.W. Hahn, J. Kaiser, On the utilization of principal component analysis in laser-induced breakdown spectroscopy data analysis, a review, Spectrochim. Acta B At. Spectrosc. 148 (2018) 65–82, https://doi.org/10.1016/j.sab.2018.05.030.
[21] J.F. Villa-Manríquez, J. Castro-Ramos, F. Gutiérrez-Delgado, M.A. Lopéz-Pacheco, A.E. Villanueva-Luna, Raman spectroscopy and PCA-SVM as a non-invasive diagnostic tool to identify and classify qualitatively glycated hemoglobin levels *in vivo*, J. Biophotonics 10 (2017) 1074–1079, https://doi.org/10.1002/jbio.201600169.
[22] G. Toscano, Å. Rinnan, A. Pizzi, M. Mancini, The use of near-infrared (NIR) spectroscopy and principal component analysis (PCA) to discriminate bark and wood of the most common species of the pellet sector, Energy Fuels 31 (2017) 2814–2821, https://doi.org/10.1021/acs.energyfuels.6b02421.
[23] V.K. Unnikrishnan, K.S. Choudhari, S.D. Kulkarni, R. Nayak, V.B. Kartha, C. Santhosh, Analytical predictive capabilities of laser induced breakdown spectroscopy (LIBS) with Principal Component Analysis (PCA) for plastic classification, RSC Adv. 3 (2013) 25872, https://doi.org/10.1039/c3ra44946g.
[24] M. Fontalvo-Gómez, J.A. Colucci, N. Velez, R.J. Romañach, In-line near-infrared (NIR) and raman spectroscopy coupled with principal component analysis (PCA) for in situ evaluation of the transesterification reaction, Appl Spectrosc 67 (2013) 1142–1149, https://doi.org/10.1366/12-06729.
[25] G. Giubileo, F. Colao, A. Puiu, Identification of standard explosive traces by infrared laser spectroscopy: PCA on LPAS data, Laser Phys. 22 (2012) 1033–1037, https://doi.org/10.1134/S1054660X12060035.
[26] Y. He, X. Li, X. Deng, Discrimination of varieties of tea using near infrared spectroscopy by principal component analysis and BP model, J. Food Eng. 79 (2007) 1238–1242, https://doi.org/10.1016/j.jfoodeng.2006.04.042.
[27] Y. Hou, C. Jiang, A.A. Shukla, S.M. Cramer, Improved process analytical technology for protein a chromatography using predictive principal component analysis tools, Biotech & Bioengineering 108 (2011) 59–68, https://doi.org/10.1002/bit.22886.
[28] M.E. Pate, M.K. Turner, N.F. Thornhill, N.J. Titchener-Hooker, Principal component analysis of nonlinear chromatography, Biotechnol Progress 20 (2008) 215–222, https://doi.org/10.1021/bp034133a.
[29] K.M. Pierce, J.L. Hope, K.J. Johnson, B.W. Wright, R.E. Synovec, Classification of gasoline data obtained by gas chromatography using a piecewise alignment algorithm combined with feature selection and principal component analysis, J. Chromatogr. A 1096 (2005) 101–110, https://doi.org/10.1016/j.chroma.2005.04.078.
[30] K. Wiberg, M. Andersson, A. Hagman, S.P. Jacobsson, Peak purity determination with principal component analysis of high-performance liquid chromatography–diode array detection data, J. Chromatogr. A 1029 (2004) 13–20, https://doi.org/10.1016/j.chroma.2003.12.052.
[31] M.R. Euerby, P. Petersson, Chromatographic classification and comparison of commercially available reversed-phase liquid chromatographic columns using principal component analysis, J. Chromatogr. A 994 (2003) 13–36, https://doi.org/10.1016/S0021-9673(03)00393-5.
[32] M. Lasalvia, V. Capozzi, G. Perna, A comparison of PCA-LDA and PLS-DA techniques for classification of vibrational spectra, Appl. Sci. 12 (2022) 5345, https://doi.org/10.3390/app12115345.
[33] M. El Maouardi, K. De Braekeleer, A. Bouklouze, Y. Vander Heyden, Comparison of near-infrared and mid-infrared spectroscopy for the identification and quantification of argan oil adulteration through PCA, PLS-DA and PLS, Food Control 165 (2024) 110671, https://doi.org/10.1016/j.foodcont.2024.110671.
[34] Y. Yang, J. Jiang, J. Zeng, Z. Chen, X. Zhu, Y. Shi, CH4, C2H6, and CO2 multi-gas sensing based on portable mid-infrared spectroscopy and PCA-BP algorithm, Sensors 23 (2023) 1413, https://doi.org/10.3390/s23031413.
[35] P. Koczoń, J.T. Hołaj-Krzak, B.K. Palani, T. Bolewski, J. Dąbrowski, B.J. Bartyzel, E. Gruczyńska-Sękowska, The analytical possibilities of FT-IR spectroscopy powered by vibrating molecules, IJMS 24 (2023) 1013, https://doi.org/10.3390/ijms24021013.
[36] F. Cheng, C. Yang, H. Zhu, Y. Li, L. Lan, K. Wang, Semi-supervised deep learning-based multi-component spectral calibration modeling for UV–vis and near-infrared spectroscopy without information loss, Anal. Chem. 95 (2023) 13446–13455, https://doi.org/10.1021/acs.analchem.3c01132.
[37] M. Said, A. Wahba, D. Khalil, Semi-supervised deep learning framework for milk analysis using NIR spectrometers, Chemom. Intel. Lab. Syst. 228 (2022) 104619, https://doi.org/10.1016/j.chemolab.2022.104619.
[38] Q. Wang, G. Teng, C. Li, Y. Zhao, Z. Peng, Identification and classification of explosives using semi-supervised learning and laser-induced breakdown spectroscopy, J. Hazard. Mater. 369 (2019) 423–429, https://doi.org/10.1016/j.jhazmat.2019.02.015.
[39] A. Maity, S. Maithani, M. Pradhan, Cavity ring-down spectroscopy: recent technological advancements, techniques, and applications, Anal. Chem. 93 (2021) 388–416, https://doi.org/10.1021/acs.analchem.0c04329.
[40] Sanchi Maithani, Santanu Mandal, Abhijit Maity, Mithun Pal, Manik Pradhan, High-resolution spectral analysis of ammonia near 6.2 μm using a cw EC-QCL coupled with cavity ring-down spectroscopy, Analyst (2018), https://doi.org/10.1039/c7an02008b.
[41] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, Scikit-learn: Machine learning in python, The Journal of Machine Learning Research (2011) 2825–2830. http://scikit-learn.sourceforge.net.