Classification of Terahertz Reflection Spectra using Machine Learning Algorithms

Kristensen, Mathias Hedegaard; Cielecki, Pawel Piotr; Skovsen, Esben

Published in:
IRMMW-THz 2022 - 47th International Conference on Infrared, Millimeter, and Terahertz Waves.

DOI (link to publication from Publisher):
10.1109/IRMMW-THz50927.2022.9895909

Publication date:
2022

Document Version
Accepted author manuscript, peer reviewed version

Analysis and Classification of Frequency-domain Terahertz Reflection Spectra Using Supervised and Unsupervised Dimensionality Reduction Methods

Paweł Piotr Cielecki, Mathias Hedegaard Kristensen and Esben Skovsen*

Department of Materials and Production, Section for Physics and Mechanics, Aalborg University, Skjernvej 4A, DK-9220 Aalborg East, Denmark

*es@mp.aau.dk

ABSTRACT

The unique properties of terahertz (THz) spectroscopy shows a great potential for security and defense applications such as safe screening of persons and objects. However, a successful implementation of THz screening systems requires a development of reliable and efficient identification algorithms. Dimensionality reduction (DR) methods aim to reduce the dimensionality of the multivariate data and are therefore commonly used as a preprocessing step for classification algorithms and as an analytical tool allowing data visualization. In this paper, we compare the use of unsupervised and supervised DR methods for analysis and classification of THz reflection spectra based on their most widespread linear representatives, namely Principal Component Analysis and Linear Discriminant Analysis, respectively. To this end, both methods were applied to more than 5000 THz reflection spectra acquired from six active materials mixed at three different concentrations with polyethylene and measured at various humidity conditions. While considering scenarios with different level of complexity, we found that supervised approach provide better results because it enables efficient grouping despite intra-class variability. Furthermore, we showed that manipulating labels introduced into the supervised DR algorithm allows conditioning the data for a desired classification task such as security screening. Presented classification results show that simple machine learning algorithms are sufficient for highly accurate classification (>98.6%) of THz spectra, which will be suitable for many real-life applications of THz spectroscopy based on material identification.

KEYWORDS

THz spectroscopy, THz screening, frequency-domain, dimensionality reduction, classification, machine learning

DECLARATIONS

Funding: This work was financed by the Innovation Fund Denmark Grand Solutions program with grant no. IFD-7076-00017B

Conflicts of interest/Competing interests: The authors have no relevant financial or non-financial interests to disclose.

Availability of data and material: The data is available upon request from the corresponding author.

Code availability: the data processing as well as dimensionality reduction and classification algorithms have been implemented in MATLAB (MathWorks, version R2018a).

Authors' contributions:

Paweł Piotr Cielecki: Formal Analysis, Investigation, Methodology (lead), Software, Validation, Visualization, Writing – original draft, Writing – review & editing.

Mathias Hedegaard Kristensen: Investigation, Methodology, Writing – review & editing.

Esben Skovsen: Conceptualization, Funding acquisition, Methodology, Supervision, Writing – review & editing.

1 INTRODUCTION

Terahertz (THz) spectroscopy exhibit a unique potential for security and defense applications [1–6], which, along with wireless communication [7–10] and quality control applications [6,11–14], has been the main stimuli for a rapid development of THz technology during the past decades. Many hazardous substances including pure and military-grade explosives [1–4,15,16], illicit drugs [1,2,17,18], and toxic gasses [19–21] possess distinctive spectral fingerprints in the THz region allowing their identification. Furthermore, THz radiation penetrates most nonpolar dielectric covering materials such as paper, cardboard, plastic and textiles (e.g. cotton and polyester) with only moderate absorption losses enabling identification of concealed objects [4,22–24]. Finally, THz radiation is nonionizing. Therefore, at reasonable intensities, it is considered

safe for scanning persons and objects [5,24]. This exceptional set of features makes it possible to implement a nondestructive and noninvasive THz security screening, which can be used to improve safety in public places such as airports and subway stations [25–27] and monitor the content of mails and parcels [4,5,28,29].

Despite being overshadowed by the more popular pulsed time-domain technology, THz continuous-wave (CW) frequency-domain spectroscopy (FDS) offers numerous advantages desired in security screening applications. Spectral selectivity associated with this technology enables measurements within water transmission windows. This limits water absorption losses that are unavoidable for broadband time-domain systems. Consequently, CW technology has a potential for measurements at stand-off distances. Furthermore, a high spectral resolution of FDS makes it suitable for detection of gasses, which usually have very narrow absorption lines [19,30,31]. Modern THz CW spectrometers based on photomixing technology and highly reliable distributed feedback 1.5 μm lasers offer a wide range of frequency tuning and high signal-to-noise ratio even at relatively short exposure times [32–34]. In comparison to time-domain systems that require femtosecond lasers and delay lines, CW spectrometers are cheaper and more robust. Furthermore, their compact size and low weight allows development of portable systems, which can be mounted on vehicles or drones enabling a myriad of new applications [35,36].

Nevertheless, a successful implementation of THz screening systems requires a development of reliable and efficient identification algorithms. A variety of machine learning techniques has been fostered for classification of THz spectra. This includes Bayesian models [37,38], artificial neural networks [39–41], support vector machines [42–44] and random forests [37,39,43]. Dimensionality reduction (DR) methods play an important role in that process [37,43,45–49]. They transform the data into a lower dimensional space, while preserving most of the relevant information. This allows lowering the computational requirements of the machine learning algorithm and increasing the speed of learning. Therefore, DR is commonly used as a preprocessing step for classification [37,43–50]. Furthermore, DR often allows visualization of the data, which is an important analytical tool that facilitate the interpretation of the data.

In this work, we compare the use of unsupervised and supervised DR methods in the analysis of THz spectra. To this end, we used Principal Component Analysis (PCA) and Linear Discriminant Analysis (LDA), which are the most commonly-used linear representatives of each approach, respectively [51,52]. We analyzed spectra of six compounds, measured in a reflection configuration using a THz CW FDS. While analyzing the spectra, we focused on two main applications of DR, namely data visualization and preprocessing the data for classification. To test the visualization ability of the featured methods, we have reduced the dimensionality of the data to at most three dimensions. Furthermore, we have classified the DR-processed data using three different classifiers, namely: k-Nearest Neighbors (k-NN), Bayesian classifier and support vector machines (SVM). We have considered four different scenarios with different levels of complexity. In section 3.1, we have focused on samples with only a single concentration of each active material. In section 3.2, we have analyzed spectra of the samples that contain three different concentrations of the active material (80%, 50% and 20%). Section 3.3 investigates an impact of atmospheric conditions, namely water vapor absorption. For that purpose, we analyzed THz spectra measured under controlled humidity conditions at a relative humidity of 90%, 50% and 10%. Finally, we considered a real-life example related to THz security screening (in section 3.4). For that purpose, we arbitrarily divided the samples according to the threat they pose. This included RDX that is an active component in several military explosives [3]. This approach allowed us to test, whether the DR methods are able to properly condition the data for this specific classification task.

This paper explains differences between the use of unsupervised and supervised DR methods in the analysis of THz spectra. It shows that DR methods are a useful tool that allows the visualization of multidimensional spectra and that they are able to ensure high classification accuracy when used even with simple classification algorithms. Our results are highly relevant for applications that rely on the classification or identification of THz spectra such as security screening.

2 EXPERIMENTAL DETAILS AND METHODS

2.1 Sample fabrication

We selected six compounds with discernible spectral features in the THz regime, namely: Galactitol, L- Tartaric Acid (L-TA), 4-Aminobenzoic acid (PABA), Hexogen (RDX), Theophylline and α-Lactose monohydrate. All the materials were purchased from Sigma Aldrich except of L-TA that was purchased from MERCK and RDX that was supplied by the courtesy of Danish Ministry of Defence Acquisition and Logistics Organisation. We mixed each of the materials with polyethylene (PE) powder, which functions as a binder [53], at weight percentages of 80%, 50% and 20% of the active material, respectively. Since PE does not have any features in the observed spectral range, it does not affect the measured spectrum other than providing a constant background [54]. To facilitate easy handling of the samples, the mixtures were subsequently compressed into cylindrical pellets with a diameter of 25 mm and weight of 7 g using a hydraulic press. The applied pressure was approximately 4 tons. To prevent interference arising between the front and the back-surface reflections, we fabricated

an inset to the hydraulic press that made the back surface at an angle of 15° relative to the front surface. As a result, we produced truncated cylindrical samples as shown in the inset of Fig. 1. After pressing the samples, no additional surface treatment such as polishing was performed. We fabricated two samples for each material composition and two additional samples of a pure PE. Table 1 provides an overview of all samples used in this study. Due to the relatively large crystal size of Galactitol and L-TA, we were unable to fabricate satisfactory pellets, especially for the compositions with a high content of the active material. The samples were brittle and had a tendency to break when removed from the press. Furthermore, their surface was rough, following the crystal size, which promoted light scattering. Therefore, before mixing with PE, Galactitol and L-TA were ground into a fine powder using mortar and pestle.

Table 1 Overview of the samples used in our studies. Each filled circle represents a single sample with a given concentration of the active material.

Material	Weight percentage of the active material			
	20%	50%	80%	100%
Galactitol	● ●	● ●	● ●	
Lactose	● ●	● ●	● ●	
L-TA	● ●	● ●	● ●	
PABA	● ●	● ●	● ●	
RDX	● ●	● ●	● ●	
Theophylline	● ●	● ●	● ●	
PE				● ●

2.2 Terahertz setup and measurements

We characterized the samples by means of THz-FDS using the reflection setup shown in Fig. 1. The setup is based on a TeraScan 1550 system (Toptica Photonics), which operates in a coherent detection scheme. A combined output of two tunable distributed-feedback diode lasers creates a heterodyne (beat pattern), which illuminates two InGaAs photomixers (emitter and receiver, respectively) modulating their conductivity. Applying a bias voltage to the emitter induces a photocurrent oscillating at the difference frequency of the lasers, which is then outcoupled into free space by an integrated antenna. The emitted THz radiation covers the spectral range from 0.09 to 1.19 THz. A 1″ diameter off-axis parabolic mirror collimates the THz beam, which is then reflected towards the sample at an angle of approximately 11° and focused onto the sample's surface by a TPX lens. A symmetrical optical system collects the reflected THz signal and focuses it onto the receiver. The total optical path length of the THz beam is approximately 1 m. The detected THz signal superimposed with the optical beat induces a photocurrent in the receiver. The photocurrent is amplified by a lock-in amplifier, recorded by the DLC smart controller for the TeraScan system and analyzed by a computer.

First, we performed measurements under ambient conditions. Each sample were measured 80 times over the entire spectral range using a frequency step of 80 MHz and 3 ms integration time. Using these parameters, the time per scan was around 45 s. A computer-controlled two-axis translation stage moved the sample after each measurement to a random position within a 7 x 7 mm scanning area with a step resolution of 0.5 mm. Every 20 measurements, we replaced the sample with an aluminum mirror and recorded its spectrum as a reference.. Subsequently, we performed measurements under controlled humidity conditions. For this purpose, a custom-built humidity chamber enclosing the terahertz setup was purged with either dry or water vapor saturated nitrogen depending on, whether the intended humidity was lower or higher than the ambient air. Based on a relative humidity readout from a DHT22 sensor, a custom-built humidity controller manages a solenoid valve, which opens and closes the nitrogen flow. A fan installed next to the nitrogen inlet ensures a uniform humidity distribution inside the chamber. The applied methodology allowed us to keep the relative humidity (RH) at the intended level with a stability better than ±0.3 percent points. We measured each sample 20 times at the RH of 90%, 50% and 10%, respectively using the same parameters as before. Due to a long purging time, the samples were measured in groups of two with a single reference measurement at RH of 50% used for both samples. Despite a good performance of our humidity controller, we observed a small variation in the intensity of water absorption peaks. We attribute this to changes in the room temperature, which were approximately ±0.7 °C during the measurements.

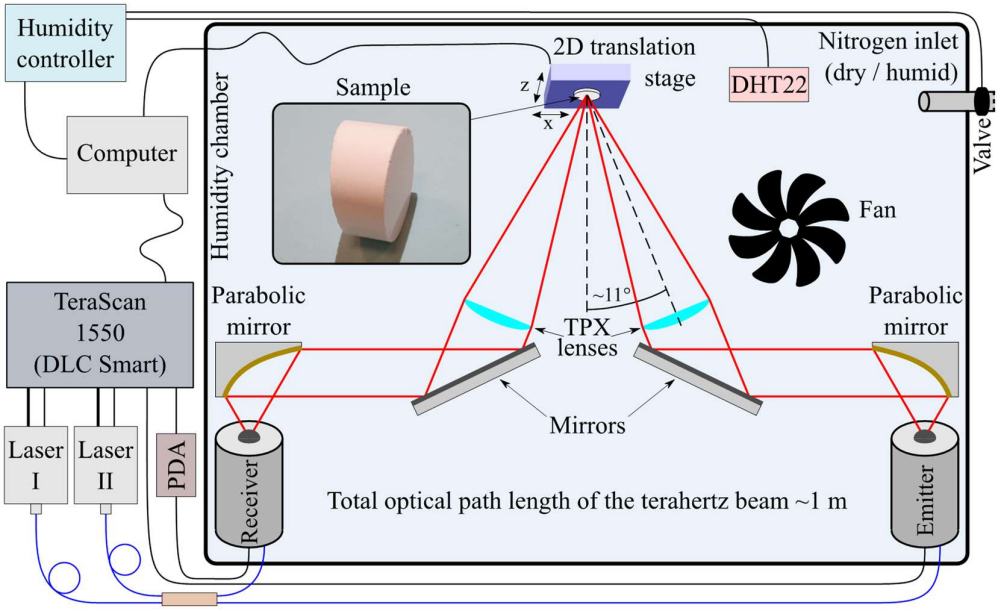

Fig. 1 Schematic drawing of the terahertz reflection setup. The inset of the figure shows a typical truncated cylindrical sample used in the experiments.

2.3 Data processing

In the CW coherent detection scheme, the photocurrent, I_{ph}, depends on the THz electric field amplitude, E_{THz}, and the phase shift, $\Delta\varphi$, between the THz wave and the optical beat via $I_{ph} \propto E_{THz}\cos(\Delta\varphi = 2\pi\Delta L\nu/c)$ [55–57]. The phase difference depends, in turn, on the frequency of the THz signal, ν, and the optical path difference $\Delta L = L_{THz} + L_E - L_R$, where L_{THz} is THz beam path, and L_E and L_R are optical beat paths to emitter and receiver, respectively. For setups with a fixed path length difference, which is the case here, it is possible to separate the magnitude and phase of the detected THz signal by analyzing an interference pattern (fringes) in the frequency scan [55–58]. To achieve this, we employed a well-known approach of finding fringe extrema, where the THz amplitude is directly proportional to the absolute value of the photocurrent and the THz phase is a multiple of π. This approach neglects all other measurement points. Consequently, the spacing between adjacent extrema defines the effective measurement resolution, which for our setup was around 0.9 GHz. To eliminate the possible offset in the photocurrent, instead of treating the extrema separately, the photocurrent was obtained from two adjacent extrema and used at the average frequency. The coherent detection scheme allows extraction of the phase information, but in this study, we used only the amplitude of the THz field. Therefore, the methodology presented in this paper can be also applied to THz systems using non-coherent detection. To compensate for spectral shifts of fringe extrema, the data was interpolated onto integer GHz frequencies. Finally, we determined the reflection coefficient of the sample by dividing the THz spectrum with the corresponding reference spectrum. Due to the standing wave patterns and lack of spectral features in the low frequency regime, we only used the data in the spectral range from 0.3 to 1.15 THz.

We recorded 3040 terahertz reflection spectra of six compounds with three different weight percentages of the active material under ambient conditions. Four out of these 3040 measurements were excluded due to processing issues and not analyzed further. These spectra are used in sections 3.1, 3.2 and 3,4. Additionally, we collected 2280 spectra under controlled humidity conditions, which are analyzed in section 3.3. A dataset of this size constitutes a good basis for developing various machine learning classifiers, which is an important step towards a real-time identification of illegal substances, e.g. explosives or drugs, using terahertz spectroscopy. Before publishing online, the data will be available upon request from the corresponding author.

The datasets were arranged in n x d matrices, where n is the number of observations (measurements, also referred to as data points) and d is the number of features (dimensions). Before DR, the dataset under consideration was randomly divided into a training set and a test set at ratio of 4:1 using stratified random sampling. Furthermore, the features were standardized so that they had a zero mean and a variance equal to unity, which is crucial for a correct operation of the PCA algorithm [59,60]. All presented algorithms (feature standardization, DR and classification) were only trained on the training set, and the test set was used solely for the final evaluation. We performed the data processing and implemented all DR and classification algorithms used in this study in MATLAB (MathWorks, version R2018a).

2.4 Dimensionality reduction

As the name suggests, dimensionality reduction algorithms aim to lower the dimensionality of the multivariate data, while maintaining most of the information it contains [61–63]. They can be divided into two categories, namely: feature selection and feature extraction. In machine learning, the term 'feature' refers to an individual measurable property of the observed object or phenomenon [64]. In our case, features correspond to discrete frequencies in THz spectra. Feature selection methods strive to select a subset of features that contain the most relevant information (for a desired task) while eliminating redundant, noisy, and irrelevant features. Common approaches for feature selection include filters such as Relief and Information Gain, and wrappers, which utilize a preselected machine learning model to evaluate relevance of considered features [61–63]. Additionally, some machine learning algorithms, such as random forests and models based on LASSO regularization, have an embedded feature selection step [61–63]. Feature extraction methods construct a new feature space with reduced dimensionality, where new features are obtained as a combination of the original features. In this work, we focus on two well-known linear feature extraction methods Principal Component Analysis and Linear Discriminant Analysis, which represent unsupervised and supervised approach, respectively [51,52]. Therefore, when referring to dimensionality reduction, we explicitly refer to feature extraction.

Principal Component Analysis is a widespread multivariate analysis and dimensionality reduction method, which has found application in various fields including image processing, pattern recognition and chemometrics [47,49,62,65,66]. As an unsupervised method, PCA does not process class membership information (labels); therefore, its outcome depends only on the hidden patterns in the data. It transforms original features into a set of new uncorrelated (orthogonal) features, called principal components, designed to maximize the variance of the data. This should lead to separation of observations having different properties (i.e., belonging to different classes). In practice, the variance maximization is achieved by eigendecomposition of the data's covariance matrix (or correlation matrix [59]). The obtained eigenvectors are ordered by the amount of explained variance, which is proportional to their eigenvalues, so that the first principal component represent the highest variability. Dimensionality reduction is obtained by selecting a limited number of most relevant principal components. In many cases, projecting the initial data onto a relatively small number of principal components is able to explain the vast majority of the variance. This allows for a significant reduction in dimensionality [45,67,68].

Linear Discriminant Analysis is a commonly used dimensionality reduction technique, which utilizes a supervised approach. In comparison to unsupervised methods, using class membership information offers a range of new possibilities for formulating transformation criteria. LDA seek a transformation that maximize the distance between classes, while minimizing the scatter within each class [69]. This should provide a large inter-class spacing and small intra-class separation in the new reduced feature space. Satisfying these criteria requires calculating between-class S_b and within-class S_w scatter matrices defined as:

$$S_w = \frac{1}{n}\sum_{j=1}^{m}\sum_{x\in X_j}\left(x - c^{(j)}\right)\left(x - c^{(j)}\right)^T$$

$$S_b = \frac{1}{n}\sum_{j=1}^{m} n_j\left(c^{(j)} - c\right)\left(c^{(j)} - c\right)^T$$

where x is a vector representing a specific observation, c is a global centroid and $c^{(j)}$ is a centroid of j-th class, n and m are the number of observations and classes, respectively. Alternatively, a total scatter matrix S_t calculated as

$$S_t = \frac{1}{n}\sum_{i=1}^{n}(x_i - c)(x_i - c)^T = S_b + S_w$$

can be used instead of S_w.It has been shown that these two approaches yield equivalent solution [52,70,71]. In this work, we used second approach (calculating S_t). The desired optimization task is then reduced to eigendecomposition of ${S_t}^{-1}S_b$ (or ${S_w}^{-1}S_b$) [52,69]. Since the rank of the S_b matrix is limited to $m-1$ (S_b is the sum of m matrices of rank 1), the obtained solution contains at most $m-1$ independent eigenvectors that are associated with non-zero eigenvalues [51,71]. Consequently, LDA can project the data onto at most $m-1$ dimensions. Since in most cases the number of original features is larger than the number of considered classes, this property alone imposes a reduced dimensionality of the new feature space. However, it is also possible to select a smaller number of most relevant features to further reduce the dimensionality.

2.5 Machine learning - classification algorithms

The Bayes classifier is a probabilistic classification model that utilizes Bayes' theorem [51,72,73]. According to the Bayes' formula, a posterior probability $p(C_i|x)$, which is the probability that the observation with value x belongs to the i-th class, can be calculated as

$$p(C_i|x) = \frac{p(x|C_i)\,p(C_i)}{p(x) = \sum_i p(x|C_i)\,p(C_i)} \qquad \text{posterior} = \frac{\text{likelihood} \times \text{prior}}{\text{evidence}}$$

where $p(C_i)$ is called the prior and it is the probability that the observation, regardless of its value, belongs to the class i, and the likelihood $p(x|C_i)$ is a probability that the member of the i-th class take an observation value x. The expression in the denominator $p(x)$, called the evidence, is a marginal probability that the observation x occurs. The evidence is merely a normalization factor, which ensures that the posteriors sum up to one, and can be neglected. The prior and the class likelihood are estimated from the training data and used for computing the class-specific posterior probability. Finally, the Bayesian classifier assigns the observation to the class with the highest posterior probability. In our work, we assumed a multivariate normal (Gaussian) distribution of the class likelihood function. This limited the likelihood estimation to finding the two unknown parameters of the multivariate Gaussian namely, the mean vector and the covariance matrix. Henceforth, we refer to the method as the Gaussian Bayes classifier.

k-Nearest Neighbors is a simple, yet effective, non-parametric classification algorithm. Despite being considered a machine learning algorithm, k-NN has no model learning (training) phase (a so-called lazy learning algorithm) [74,75]. Instead, the algorithm stores all training data and classify a new observation based on its similarity to the training instances. Here, similarity is expressed in terms of the geometrical distance, such as the Euclidean, Minkowski, Manhattan or Chebyshev distance, between the data points [75,76]. The algorithm assigns the new observation to the class possessed by the majority of the k closest data points called nearest neighbors. The choice of k, the number of considered nearest neighbors, is crucial as it allows for controlling the algorithm's behavior. For small k the decision boundaries are more flexible, but the algorithm becomes prone to outliers, while higher k results in more robust boundaries [77,78]. In this study, we used the Euclidean distance as a similarity metric in the k-NN algorithm. For each considered case, we used k comparable to the number of training examples in the smallest class (k = 125 in chapters 3.1 and 3.2, k = 95 in chapter 3.3 and k = 383 in chapter 3.4, as indicated in the corresponding classification accuracy tables). This approach should provide a good estimate of intra-class distribution and separation between classes.

Support Vector Machine was introduced by Vapnik in 1995 and has been one of the most widely used classification algorithms ever since [79]. The algorithm is based on the maximal margin classification concept. It searches for a hyperplane, $h(x) = x^T\beta + \beta_0 = 0$, that separates the observations belonging to two classes and has the largest value of the minimum geometrical distance to class representatives – the margin M. The classification of new observations is based on which side of the hyperplane they are on. The maximal margin hyperplane is found by solving the following optimization problem [80–82]:

$$\max_{\beta,\beta_0,\|\beta\|=1} M$$

$$subject\ to \quad y_i({x_i}^T\beta + \beta_0) \geq M, \quad i = 1, \dots, n,$$

where β and β_0 are parameters of the optimal hyperplane, $y \in \{-1, 1\}$ are labels associated with two considered classes and n is the number of observations. The maximal margin classifier offers the solution with $M > 0$ only when the classes are linearly separable. For non-linearly separable data, the algorithm utilizes a soft margin approach. It allows some data points to violate the margin or even be on the opposite side of the hyperplane in exchange for imposing a penalty, called the slack variable ξ_i, on the objective function. The optimization problem then becomes:

$$\max_{\beta,\beta_0,\|\beta\|=1} M$$

$$y_i({x_i}^T\beta + \beta_0) \geq M(1 - \xi_i), \quad \xi_i \geq 0, \quad \sum_i^n \xi_i \leq C, \quad i = 1, \dots, n,$$

where C is a non-negative tuning parameter that controls the balance between the margin width and the slack (the bias-variance trade-off). In practice, finding an optimal hyperplane only requires the consideration of a small number of observations that lie on or violate the margin (so called support vectors). Since the hyperplane has a dimensionality of $d - 1$, where d is the dimensionality of the input data, the obtained decision boundary is linear. However, SVM can also create non-linear decision boundaries. This is achieved by mapping the data into higher-dimensional space using non-linear kernel functions. In this study, we used SVMs with a linear kernel. Originally, SVMs were defined for a binary classification problem. In order to perform multiclass classification, we adopted a one-versus-one approach in which a separate SVM is

constructed for each pair of classes [80]. The final classification is a majority decision of all $\binom{m}{2}$ pairwise SVM classifiers. To determine the optimal value of C, we performed a 10-fold cross-validation on the training set.

3 RESULTS AND DISCUSSION

3.1 Samples with single concentration of active materials

Some publications on material identification using THz spectroscopy consider samples with only a single concentration of active materials within each material class [40,83]. Therefore, we first compared the performance of DR methods in this simple scenario. To this end, we used the samples with the highest content of the active material (80%), which exhibit the most pronounced spectral features, and two samples of pure PE. As shown in Fig. 2, all six active materials have distinctive spectral features in the measured frequency range, while PE is inert and exhibits an almost featureless spectrum. Since THz spectroscopy measurements were performed in a reflection configuration, the spectra depend on the refractive index of the materials [84] having smooth and relatively broad features. Consequently, there is significant spectral overlap among materials and the spectral features cover almost the entire frequency range. Two prominent absorption lines of atmospheric water vapor, located at around 1100 and 1115 GHz, respectively [4], reduce the signal measured from the sample (but not the reference) to the level comparable to the noise floor. This leads to the formation of two narrow, noisy peaks. By excluding the noisy data in the range 1086-1119 GHz (blue line in Fig. 2), we obtained 1117 spectra containing 817 discrete frequencies (dimensions / features), which we divided into training and test sets (4:1) and used for evaluation of the DR methods.

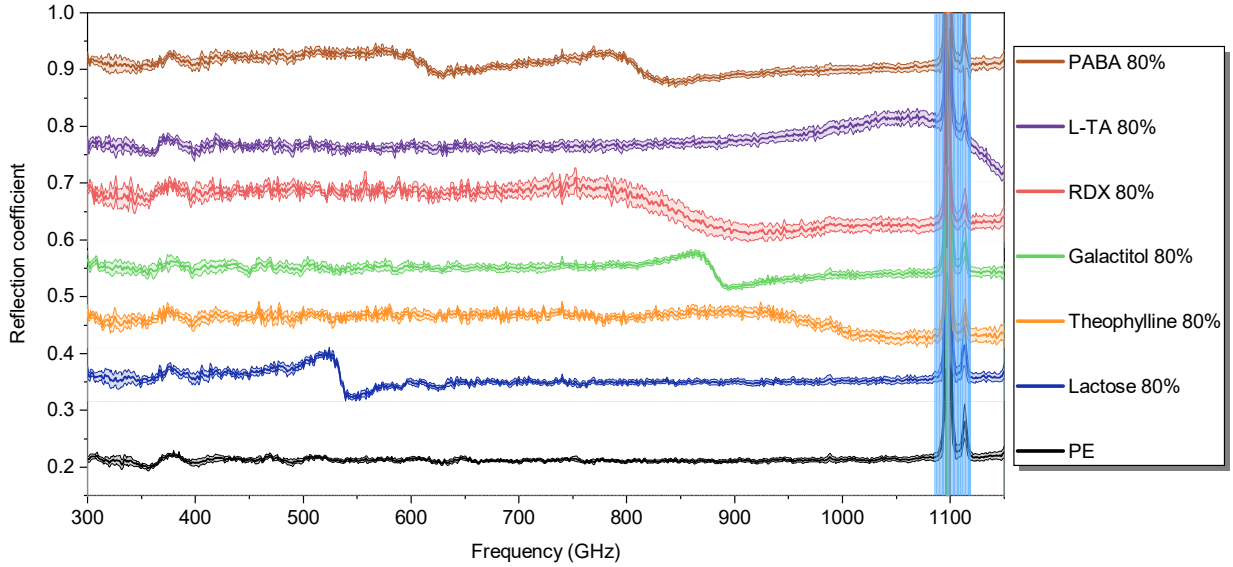

Fig. 2 THz reflection spectra of the samples with 80% of the active material and samples of pure PE. The central lines represent the mean of 160 measurements performed on two different samples, while the error bars (outer lines) represent a standard deviation of the measured spectra. For clarity, each consecutive spectrum, except PE, has been shifted upwards by 0.1. The blue line represents a spectral region excluded from simulations due to significant water absorption-related noise.

We used PCA and LDA algorithms to reduce dimensionality of the THz spectra. To numerically verify the performance of the DR methods in terms of grouping and separating different materials, we classified the data using three different classifiers, namely a Gaussian Bayes classifier, a k-NN and an SVM. Table 2 summarizes the obtained classification accuracies. As shown in Fig. 3, projecting the spectra onto a two-dimensional space provides good separation between materials in the training set (filled circles). For PCA (Fig. 3a), there is a small overlap between Lactose, Theophylline and Galactitol, while the other materials are well separated. Classification of PCA-processed data with the SVM and the Bayes classifier yielded similar training classification accuracies of 0.9888 and 0.9843, respectively. The k-NN algorithm showed slightly lower performance of 0.9742. As expected, all misclassification cases originate from confusing two of the overlapping materials. For LDA (Fig. 3b), all the materials seem to be spatially separated. Although Theophylline is located close to L-TA, there is no overlap between these materials. The performed classification confirms that observation as all classifiers yielded a perfect accuracy on the training set.

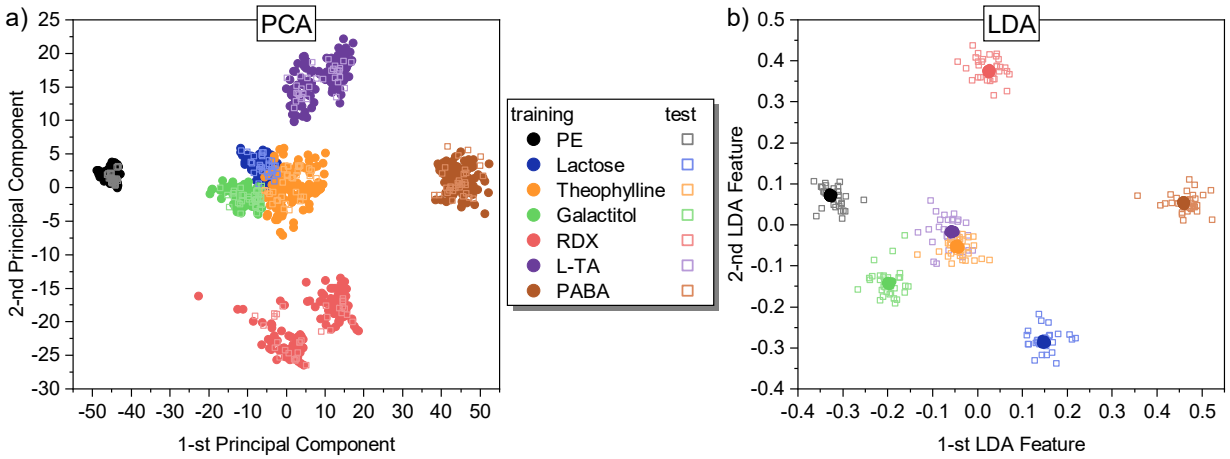

Fig. 3 THz reflection spectra (shown in Fig. 2) projected onto a two-dimensional feature space using two DR methods a) PCA and b) LDA. Filled circles correspond to the training set, while empty squares to the test set.

Subsequently, we performed DR on the test set (empty squares in Fig. 3) using the trained PCA and LDA algorithms to verify their performance on unseen data, a so-called generalization. For PCA (Fig. 3a), the test set overlap distinctly well with the training set and each cluster corresponding to different materials show similar distribution for both datasets. Furthermore, the obtained test classification accuracies are comparable to these obtained on the training set, which proves a good generalization of the PCA algorithm. However, as shown in Fig. 3b, LDA does not generalize as well as PCA. Data points corresponding to the specific material are more scattered in the test set than it is the case for the training set. Classifying the LDA-processed test data, we obtained a relatively poor classification accuracy of around 0.9509 and 0.9464 for Bayes classifier and k-NN, and for SVM, respectively. Most cases of misclassification arose from confusing Theophylline and L-TA due to the increased dispersion of the test data. The observed lack of generalization together with a perfect performance on the training set suggests overfitting of the LDA algorithm [85,86]. It has been reported that LDA is prone to overfitting if the dimensionality of the data is comparable to the size of the training set [87–89] (poorly-posed problem), which applies to the considered case (817 features and 893 training spectra). The eigenvectors obtained during the training confirms that the algorithm follows the noise instead of patterns (spectral features) in the data (See Fig. S1 in the supplementary material). Consequently, the algorithm performs poorly on the unseen data, which exhibit a different noise pattern.

Table 2 Classification accuracy of DR-processed THz spectra. Only the samples with 80% of active material and samples of pure PE have been considered.

	PCA		***LDA***		***RLDA***	
	Train	**Test**	**Train**	**Test**	**Train**	**Test**
Bayes	0.9843	0.9955	1.0000	0.9509	1.0000	1.0000
125-NN	0.9742	0.9732	1.0000	0.9509	0.9966	1.0000
SVM	0.9888	0.9732	1.0000	0.9464	1.0000	1.0000

Overfitting is a well-known problem in machine learning and is closely related to the bias / variance trade-off, where the trained model exhibits a high variance [51,85]. Increasing the size of the training set constrains the variability of the model and thus prevents overfitting. However, performing a great number of additional measurements is usually time-consuming and costly, and in some applications, it can be impractical or even impossible. Another approach is to reduce the number of parameters to train. It can be achieved by feature selection methods, which eliminate redundant features and those that contain the least amount of information [61]. However, feature selection is a complex task and sometimes may remove essential features leading to inferior performance of the algorithm [88]. Regularization is a simple, yet effective, solution to the overfitting problem. It has been implemented in many machine learning algorithms [51,81,85]. It allows constraining the variance by imposing a penalty on complex models. In LDA, the simplest form of regularization is implemented by adding a regularization parameter λ to the diagonal elements of the total scatter matrix S_t following the formula: $\hat{S}_t = S_t + \lambda I$, where $\lambda > 0$ and I is the identity matrix [52,69,71]. However, more sophisticated forms of regularization have also been proposed [52,71,86,90]. Choosing λ is crucial for a correct operation of regularized LDA (RLDA) because it directly controls

the balance between bias and variance of the model (See Fig. S2 in the supplementary information). To estimate an optimal value of λ, we performed 10-fold stratified cross validation on the training set [81,87,91]. Using the classification accuracy of the Gaussian Bayes classifier as a validation criterion, we found the optimal value of λ to be 0.5. Finally, we trained the RLDA algorithm on the entire training set using the previously estimated optimal value of λ and applied the trained model to the test set.

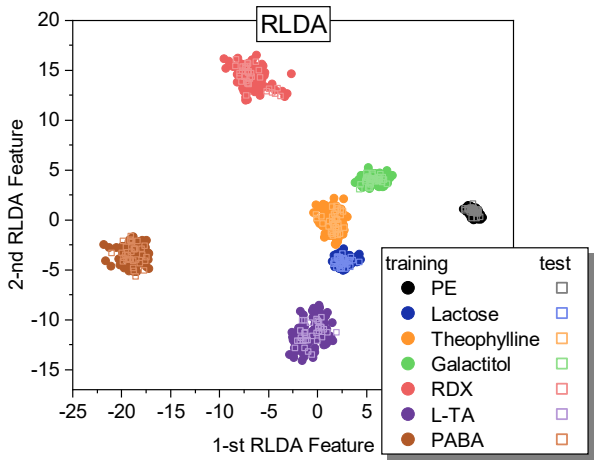

Fig. 4 THz reflection spectra (shown in Fig. 2) projected onto a two-dimensional feature space using RLDA. Filled circles correspond to the training set, while empty squares to the test set.

As shown in Fig. 4, RLDA provides significantly better generalization than the classical LDA algorithm, while maintaining a perfect separation between materials. There is a perfect alignment between the training and test data and material-specific clusters show similar distribution in both datasets. The SVM and the Bayes classifier had no problems with the RLDA-processed data yielding a unity accuracy on both training and test set. The k-NN also achieved a perfect test accuracy but misclassified a few training examples yielding a training accuracy of 0.9966. Our results shows that RLDA slightly outperforms PCA in terms of preprocessing data for classification purposes. However, reducing the dimensions of the data from 817 to two is very strict. Projecting the THz spectra onto three- instead of two-dimensional space allows obtaining a perfect separation of materials also for PCA as confirmed by the Gaussian Bayes classifier and SVM (k-NN misclassified a few training examples resulting in a training accuracy of 0.9955). In comparison to PCA, RLDA provides more efficient spatial grouping of the materials. For PCA, data points corresponding to RDX are scattered and forms two separate clusters (a similar behavior can be observed for L-TA). We found that two nominally identical samples of RDX exhibit slightly different spectra, which we attribute to non-uniform distribution of the active material. Since PCA is an unsupervised method, it considers only the patterns in the data. Therefore, spectral differences between RDX samples result in spatial separation in the reduced feature space. In case of RLDA, the provided class membership information, which is a key property of supervised learning methods, enables efficient grouping of data despite the spectral differences. A poor grouping ability of unsupervised DR methods may have a negative impact on the performance of some classification algorithms.

3.2 Samples with various concentrations of active materials

Subsequently, we considered a more realistic scenario, where the samples contain various concentrations of the active material to be identify. To this end, we characterized the samples with three different concentrations 80%, 50% and 20%, respectively. Additionally, we used two samples of pure PE to test, whether the samples with low active material content and therefore high PE content, can be distinguished from the background. Fig. 5 shows the spectra of all types of samples used in this study. Since the samples are mixtures of the active material and PE, the obtained reflection coefficient is an intermediate value between the spectra of the two components. As the concentration of the active material decreases, the material-specific spectral features become less pronounced. Furthermore, for materials having a significantly higher refractive index than PE e.g., PABA and Theophylline, the reflection coefficient drops over the entire spectral range with decreasing concentration. Here, we investigate if the increased variability of the data associated with material concentration affects the performance of the DR algorithms.

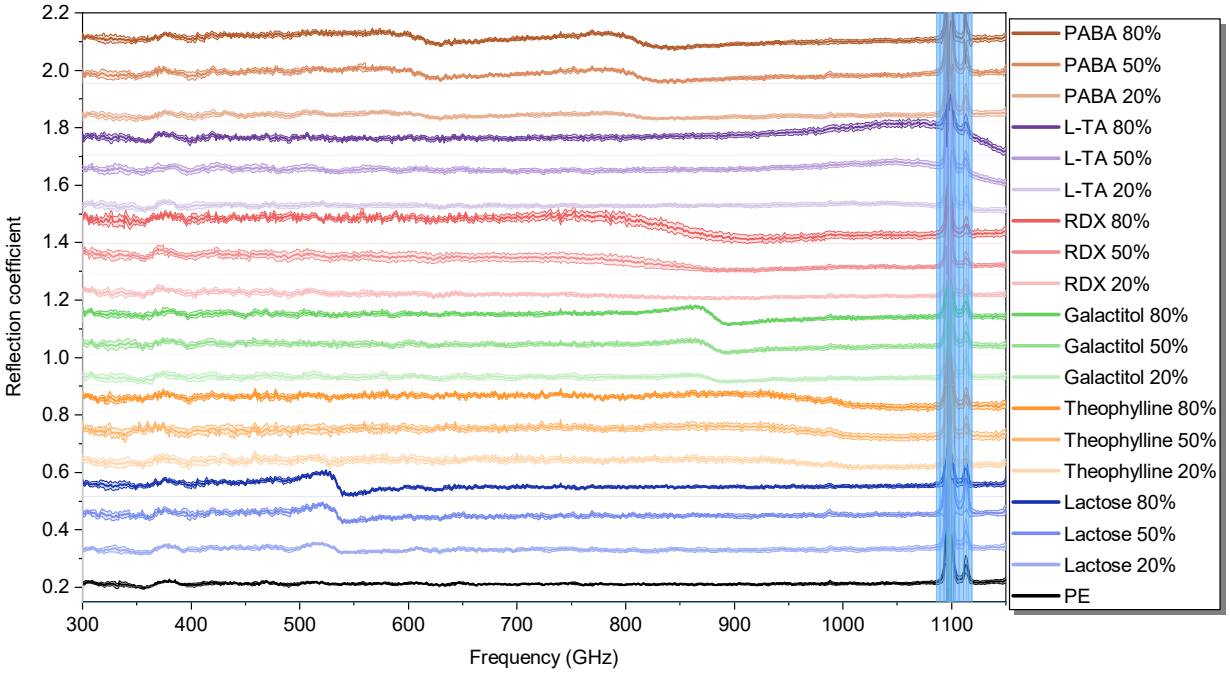

Fig. 5 THz reflection spectra of all samples used in this study. The central lines represent the mean of 160 measurements performed on two different samples, while the error bars (outer lines) represent the standard deviation of the measured spectra. For clarity, each consecutive spectrum, except PE, has been shifted upwards by 0.1. The blue line represents a spectral region excluded from simulations due to significant water absorption-related noise.

While performing DR we used the same methodology as before. First, PCA and LDA algorithms were trained on the training set containing 80% (2428) of the THz spectra. Then, we applied them to the remaining 20% (608) of the spectra, which constitute a test set. Since the aim was to identify the sample by the contained active material regardless of its concentration, the class labels used for the supervised DR contain only the information about the type of the material, while neglecting its concentration. In contrast to the simple scenario considered previously, the DR algorithms were unable to provide good separation between the materials using only two most significant features. Therefore, we projected the data onto three-dimensional space. As an unsupervised method, PCA considers only the patterns in the input data, or more precisely, its variance. Therefore, we expect that an additional variability related to material concentration is going to be transferred into a lower-dimensional space as we observed for RDX in the previous section. As shown in Fig. 6a and Fig. 7, PCA processing arranged the data into a shape resembling a conical surface. At the vertex of the cone, there are data points corresponding to the pure PE, which is a component of all the samples. The data corresponding to samples with a low content of active material, and hence with a high content of PE, is located close to the pure PE and distributed around the cone's symmetry axis depending on the contained active material (Fig. 7). In this configuration, faint spectral features of the samples with low active material content correspond to a small spatial separation between samples. As the concentration of the active material increases, the data points are projected further from the vertex. Consequently, the data corresponding to each material is divided into clusters, which correspond to different material concentrations. The materials that exhibit larger concentration-related spectral changes (globally, over the entire spectra) e.g., RDX or PABA (see Fig. 5), are more scattered in the reduced feature space. Therefore, the spatial separation between the samples with high active material content, which have more pronounced spectral features, becomes larger. In other words, the greater spectral differences between the samples the larger separation in the reduced feature space. This intuitive explanation of variance can become useful in data analysis; however, in terms of data visualization the clarity offered by PCA is rather poor. A comparable distribution of test and training data in Fig. 6a and Fig. 7 indicates that, as in the previous case, PCA provides a good generalization on unseen data. Subsequently, we classified the PCA-processed data depending on the active material contained in the sample (Table 3). The classification accuracies obtained on the training set ranged from 0.8871 for the Bayes classifier to 0.9090 for the k-NN. For test set, the classifiers yielded accuracies ranging from 0.8914 to 0.9145 for Bayes classifier and k-NN, respectively. Test results similar or, in that case, better than these obtained on the training set prove a good generalization of PCA. However, the concentration-related separation of data within the material class disturbs the operation of heavily biased classifiers used in this study. This, in turns, results in relatively poor classification accuracy.

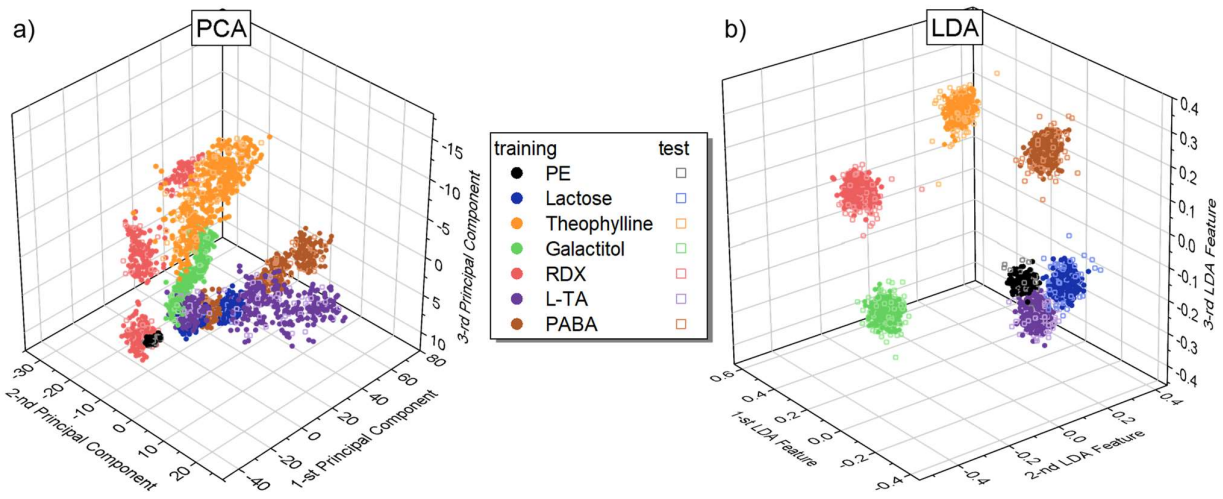

Fig. 6 THz reflection spectra of samples with various concentrations of active material (shown in Fig. 5) projected onto a three-dimensional feature space using two DR methods a) PCA and d) LDA.

The objective of LDA is to maximize the inter-class distance and minimize the scatter within each class. Fig. 6b shows that the algorithm does it very well. LDA not only provides a great separation between materials, but also a superior intra-class grouping than PCA. The class membership information, introduced into a supervised algorithm by labels, enables efficient grouping of data associated with the same active material despite having an additional variability related to the concentration. However, the distribution of the data in the lower-dimensional subspace is not as intuitive and easy to interpret as we observed for PCA. In comparison with the simple scenario presented in the previous section, LDA shows better generalization. Since we used more THz spectra for training, (2428 compared to 893 used previously) the noise is effectively averaged out and the algorithm better learns the pattern in the data. As a result, the test set shows a similar distribution compared to the training set. As shown in Table 3, all classifiers achieved a high accuracy of 0.9975 for the LDA-processed training data. For the test set, the obtained classification accuracies were 0.9868 and 0.9885 for the Bayes classifier and the SVM, and the k-NN, respectively. An inferior test performance of both classifiers may suggest that despite an increased amount of data the algorithm slightly overfits the data. It has been suggested that for the optimal performance, machine learning algorithms requires at least ten times as many training examples as the number of features [92], while in our case this ratio is only around three. Nevertheless, LDA clearly outperforms PCA in terms of preprocessing the data for classification, as the efficient grouping mechanism offered by the supervised approach results in significantly better performance.

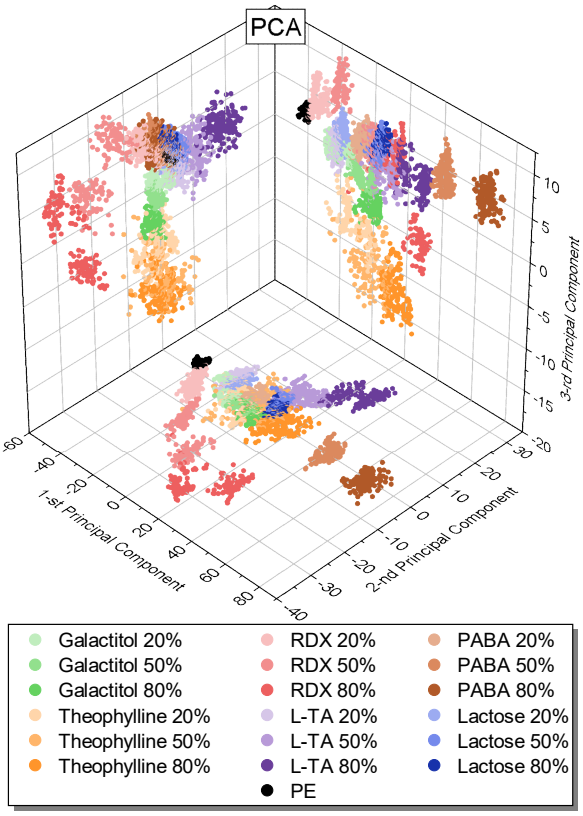

Fig. 7 PCA-processed THz spectra from Fig. 6 projected (for clarity) onto two-dimensional planes and labeled with respect to the active material and its concentration.

Table 3 Classification accuracy of DR-processed THz spectra for the samples with various concentrations of active materials.

	PCA		***LDA***	
	Training	**Test**	**Training**	**Test**
Bayes	0.8871	0.8914	0.9975	0.9868
125-NN	0.9090	0.9145	0.9975	0.9885
SVM	0.8937	0.9046	0.9975	0.9868

3.3 Samples with various concentrations of active materials under various (controlled) humidity conditions

Absorption of the atmospheric water vapor is a well-known problem in the THz technology. In section 3.1, we observed that strong water absorption lines located around 1100 GHz and 1115 GHz almost completely attenuated the THz signal, preventing any useful measurements in the adjacent spectral range (Fig. 2 and Fig. 5). For weaker absorption lines that allow resolving the remaining THz signal, the reference measurement was able to efficiently remove the spectral features related to atmospheric water vapor (Fig. S3 in supplementary materials). However, in many potential out-of-the-lab applications e.g., stand-off identification of hazardous substances, obtaining a precise reference is difficult or even impossible to achieve. Consequently, the THz spectra may possess additional spectral features related to water absorption. To verify their impact on the performance of DR algorithms, we characterized all samples at three different RH levels of 90%, 50% and 10%, respectively. The reference measurements were performed at RH of 50%. The obtained THz spectra have three narrow peaks corresponding to the water absorption lines located at around 557 GHz, 753 GHz and 988 GHz (See Fig. S3 in the supplementary materials). The peaks are positive for measurements performed at RH 10% and negative for RH 90%. At RH 50%, the peaks do not occur as they were removed by the reference.

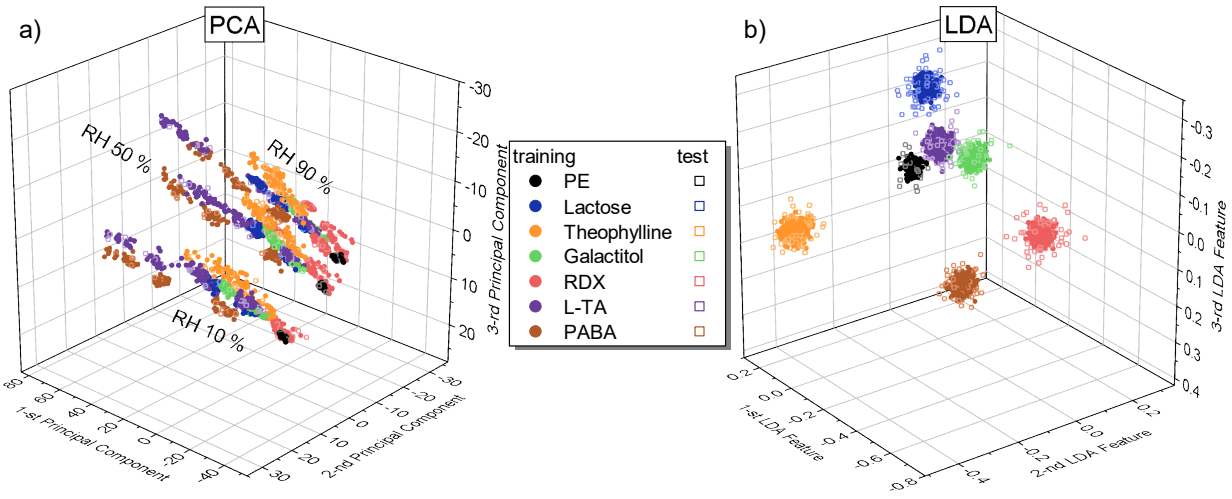

Fig. 8 Three-dimensional projections of THz spectra obtained by a) PCA and b) LDA. The THz spectra of samples with various concentrations of active materials were measured under various humidity conditions.

Using an unaltered methodology, we applied DR algorithms to project the THz spectra onto a three-dimensional subspace. As shown in Fig. 8a, PCA arranged the data into three distinctive clusters that correspond to RH level during the measurements. Each RH-related cluster contains smaller clusters corresponding to the active material contained in the sample and its concentration. We conclude that the additional variability related to the water absorption has an effect similar to the material concentration in section 3.2 and leads to further separation (grouping) of data in the principal components space. This had a significant impact on classification results of PCA-processed data, presented in Table 4. The k-NN algorithm, which showed the best performance, provided an accuracy of only around 0.70 and 0.69 for training and test set, respectively. In turn, training and test accuracies obtained with SVM were as low as 0.6321 and 0.6053. Fig. 8b shows that despite the additional variability, LDA provides an excellent separation between materials and great intra-class grouping. Classification results of the LDA-threated data summarized in Table 4 proves a superior performance of the supervised method. All classification algorithms yielded a unity accuracy on the training set and a test accuracy better than 0.986. The obtained results are similar to these from the previous section showing that the additional variance originating from the water absorption does not adversely affect LDA performance.

Table 4 Classification accuracy of DR-processed THz spectra for the samples with various concentrations of active materials measured under various humidity conditions.

	PCA		***LDA***	
	Training	**Test**	**Training**	**Test**
Bayes	0.6612	0.6316	1.0000	0.9868
95-NN	0.7007	0.6908	1.0000	0.9912
SVM	0.6321	0.6053	1.0000	0.9868

3.4 Manipulating class label in supervised methods

Our results from previous sections show that due to using class membership information, supervised DR methods provide better separation of the materials and superior grouping in comparison to their unsupervised counterparts. However, class labels offer another important functionality. As an additional input parameter, they provide better control over the classification algorithm. To demonstrate this, we considered an example related to security screening, which is a potential application of THz spectroscopy. For security screening, it is not necessary to identify an exact material composition of the investigated object. Instead, it is more relevant to recognize whether the object pose any threat, specify a type and magnitude of the threat and determine the proper course of action in response. Therefore, we re-labeled the spectra used in section 3.2 based on the type of hazard that the corresponding samples represent. RDX is a common explosive compound that has been used in numerous bomb plots including terrorist attacks. It is widely used as a stand-alone explosive but also as a part of explosive mixtures such as Composition C-4, Torpex and Semtex H [3]. Due to posing an imminent threat to multiple

persons, detection of RDX require immediate actions. Therefore, we labeled RDX as 'Danger'. Theophylline is a pharmaceutical compound used to treat respiratory diseases such as asthma. It has a narrow therapeutic range and is toxic if ingested at higher doses. Nevertheless, in comparison to RDX, the threat it constitutes is minor and does not require such drastic measures. Consequently, Theophylline was assigned with a 'Warning' label. The rest of investigated materials (Galactitol, Lactose, L-TA, PABA and PE) does not pose any hazard; hence, they were labeled as 'Safe'.

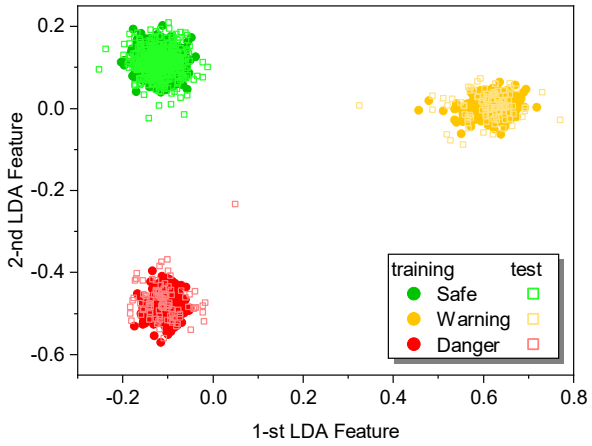

Fig. 9 THz reflection spectra projected onto a two-dimensional feature space using LDA. The labels used in supervised DR correspond to the hazard represented by the samples. Filled circles correspond to the training set, while empty squares to the test set.

Subsequently, we used the spectra with new labels to train the LDA algorithm. Since we divided the data into three classes, LDA can project the data onto a two-dimensional subspace at most. Despite this limitation, LDA provides a perfect separation between newly established classes and an efficient intra-class grouping, as shown in Fig. 9. Furthermore, LDA shows a good generalization on the test data. As expected, none of the classifiers had any problems with LDA-processed data, yielding unity accuracy on both training and test set. This shows that by altering the class labels, it is possible to project the data into different groups that are more suitable for the desired application. For comparison, we classified PCA-processed data using new class labels. Since the output of unsupervised methods does not depend on the class labels, we refer to the results from the section 3.2 (PC1-PC2 plane in Fig. 7). Obtained classification results, summarized in Table 5, show that PCA was not up to the task. Bayes classifier and k-NN correctly predicted only around 0.77 and 0.75 of test data, respectively. On the other hand, SVM yielded a higher, but still far from satisfactory, test classification accuracy of approximately 0.83.

Table 5 Classification accuracy of DR-processed THz spectra for the samples with various concentrations of active materials, labeled according the type of hazard they represent.

	PCA		***LDA***	
	Training	**Test**	**Training**	**Test**
Bayes	0.7648	0.7747	1.0000	1.0000
383-NN	0.7467	0.7500	1.0000	1.0000
SVM	0.8361	0.8322	1.0000	1.0000

4 CONCLUSIONS

In summary, we compared the use of Principal Component Analysis and Linear Discriminant Analysis – two linear methods that represent an unsupervised and a supervised approach, respectively, for analysis of THz CW reflection spectra. We focused on two common applications of dimensionality reduction, namely data visualization and preprocessing the data for classification. In the simple scenario, where we considered only samples with a single concentration of active materials, both approaches were able to compress 817-dimensional spectra down to a two-dimensional subspace, while providing good discrimination between materials.

However, as we increased the complexity of the data by using samples with various concentrations of active materials, the performance of the unsupervised approach decreased significantly. We observed a formation of separate clusters for samples having different active material concentrations in PCA-processed data. The same effect was observed while evaluating measurements performed under various humidity conditions. Since the unsupervised methods depend only on the patterns in the data, the additional intra-class variance is transferred into the reduced lower-dimensional space. This may have a negative impact on the classification accuracy, especially for highly biased classifiers as those used in this study. On the other hand, using class membership information (labels) allowed LDA to effectively group the samples with the same active material despite differences in concentration and RH. As a result, LDA provides better visualization clarity and higher classification accuracy. Therefore, if the considered application provides the labels, we strongly recommend using supervised methods, as the use of unsupervised methods in this case involves neglecting information. Furthermore, using THz security screening as an example, we demonstrated that manipulating the labels allows tuning the output of the algorithm for the desired classification task.

Despite showing a better performance and more flexibility than PCA, LDA has some drawbacks that need to be considered. When the number of observations is comparable to the number of features, a so-called poorly-posed problem, LDA tends to overfit the data. This can result in a significant drop in classification accuracy. Therefore, LDA requires more measurements than its unsupervised counterpart does, which in some applications may be an obstacle. We demonstrated that regularization is an effective solution for the overfitting problem; however, tuning the regularization parameter requires an additional user intervention. Since LDA aims to maximize the distance between classes and minimize the intra-class scatter, the distribution of the data in the lower-dimensional subspace is not as intuitive as it is the case in PCA, which aims to maximize the variance.

Furthermore, our experiments show that simple algorithms (linear dimensionality reduction and basic classifiers such as Bayesian classifier and k-NN) are sufficient for visualization and classification of THz reflection spectra. In all presented scenarios, the obtained classification accuracy was better than 98.6% using LDA. However, we recognize that in more complex cases, such as measurements through covering materials and diffused reflection measurements, which are relevant for implementation of THz screening systems, the use of more sophisticated algorithms, e.g. neural networks, may be beneficial.

ACKNOWLEDGEMENTS

This work was financed by the Innovation Fund Denmark Grand Solutions program with grant no. IFD-7076-00017B. The project was performed in collaboration with MyDefence. The authors wish to express their gratitude to the Danish Defence, and Danish Ministry of Defence Acquisition and Logistics Organisation for knowledge and assistance in measurements of explosives. Finally yet importantly, the authors are thankful to Associate Professor Torben Tvedebrink for fruitful conversations.

REFERENCES

1. A. G. Davies, A. D. Burnett, W. Fan, E. H. Linfield, and J. E. Cunningham, Mater. Today **11**, 18 (2008).

2. J. F. Federici, B. Schulkin, F. Huang, D. Gary, R. Barat, F. Oliveira, and D. Zimdars, Semicond. Sci. Technol. **20**, S266 (2005).

3. M. R. Leahy-Hoppa, M. J. Fitch, and R. Osiander, Anal. Bioanal. Chem. **395**, 247 (2009).

4. H.-B. Liu, H. Zhong, N. Karpowicz, Y. Chen, and X.-C. Zhang, Proc. IEEE **95**, 1514 (2007).

5. A. Luukanen, R. Appleby, M. Kemp, and N. Salmon, in *Terahertz Spectrosc. Imaging. Springer Ser. Opt. Sci.*, edited by K. . Peiponen, A. Zeitler, and M. Kuwata-Gonokami (Springer Berlin Heidelberg, 2012), pp. 491–520.

6. P. U. Jepsen, D. G. Cooke, and M. Koch, Laser Photon. Rev. **5**, 124 (2011).

7. S. Koenig, D. Lopez-Diaz, J. Antes, F. Boes, R. Henneberger, A. Leuther, A. Tessmann, R. Schmogrow, D. Hillerkuss, R. Palmer, T. Zwick, C. Koos, W. Freude, O. Ambacher, J. Leuthold, and I. Kallfass, Nat. Photonics **7**, 977 (2013).

8. I. F. Akyildiz, J. M. Jornet, and C. Han, Phys. Commun. **12**, 16 (2014).

9. T. Kleine-Ostmann and T. Nagatsuma, J. Infrared, Millimeter, Terahertz Waves **32**, 143 (2011).

10. D. M. Mittleman, J. Appl. Phys. **122**, 230901 (2017).

11. F. Rutz, M. Koch, S. Khare, M. Moneke, H. Richter, and U. Ewert, Int. J. Infrared Millimeter Waves **27**, 547 (2007).

12. A. A. Gowen, C. O'Sullivan, and C. P. O'Donnell, Trends Food Sci. Technol. **25**, 40 (2012).

13. A. I. Hernandez-Serrano, S. C. Corzo-Garcia, E. Garcia-Sanchez, M. Alfaro, and E. Castro-Camus, Appl. Opt. **53**, 7872 (2014).

14. F. Ellrich, M. Bauer, N. Schreiner, A. Keil, T. Pfeiffer, J. Klier, S. Weber, J. Jonuscheit, F. Friederich, and D. Molter, J. Infrared, Millimeter, Terahertz Waves **41**, 470 (2020).

15. J. Chen, Y. Chen, H. Zhao, G. J. Bastiaans, and X.-C. Zhang, Opt. Express **15**, 12060 (2007).

16. M. R. Leahy-Hoppa, M. J. Fitch, X. Zheng, L. M. Hayden, and R. Osiander, Chem. Phys. Lett. **434**, 227 (2007).

17. A. D. Burnett, W. Fan, P. C. Upadhya, J. E. Cunningham, M. D. Hargreaves, T. Munshi, H. G. M. Edwards, E. H. Linfield, and A. G. Davies, Analyst **134**, 1658 (2009).

18. K. Kawase, Y. Ogawa, Y. Watanabe, and H. Inoue, Opt. Express **11**, 2549 (2003).

19. F. Hindle, A. Cuisset, R. Bocquet, and G. Mouret, Comptes Rendus Phys. **9**, 262 (2008).

20. J. Qin, B. Zhu, Y. Du, and Z. Han, Opt. Fiber Technol. **52**, 101990 (2019).

21. N. Shimizu, T. Ikari, K. Kikuchi, K. Matsuyama, A. Wakatsuki, S. Kohjiro, and R. Fukasawa, in *2011 IEEE MTT-S Int. Microw. Symp.* (IEEE, 2011), pp. 1–4.

22. U. Puc, A. Abina, M. Rutar, A. Zidanšek, A. Jeglič, and G. Valušis, Appl. Opt. **54**, 4495 (2015).

23. C. Baker, T. Lo, W. R. Tribe, B. E. Cole, M. R. Hogbin, and M. C. Kemp, Proc. IEEE **95**, 1559 (2007).

24. M. C. Kemp, in *2007 Jt. 32nd Int. Conf. Infrared Millim. Waves 15th Int. Conf. Terahertz Electron.* (IEEE, Cardiff, 2007), pp. 647–648.

25. D. Zimdars, J. S. White, G. Stuk, A. Chernovsky, G. Fichter, and S. Williamson, Insight - Non-Destructive Test. Cond. Monit. **48**, 537 (2006).

26. G. Tzydynzhapov, P. Gusikhin, V. Muravev, A. Dremin, Y. Nefyodov, and I. Kukushkin, J. Infrared, Millimeter, Terahertz Waves **41**, 632 (2020).

27. K. Nawata, Y. Takida, Y. Tokizane, T. Notake, Z. Han, and H. Minamide, in *2019 44th Int. Conf. Infrared, Millimeter, Terahertz Waves* (IEEE, 2019), pp. 1–2.

28. H. Hoshina, Y. Sasaki, A. Hayashi, C. Otani, and K. Kawase, Appl. Spectrosc. **63**, 81 (2009).

29. K. Yamamoto, M. Yamaguchi, F. Miyamaru, M. Tani, M. Hangyo, T. Ikeda, A. Matsushita, K. Koide, M. Tatsuno, and Y. Minami, Jpn. J. Appl. Phys. **43**, L414 (2004).

30. C. Hepp, S. Luttjohann, A. Roggenbuck, A. Deninger, S. Nellen, T. Gobel, M. Jorger, and R. Harig, in *2016 41st Int. Conf. Infrared, Millimeter, Terahertz Waves* (IEEE, 2016), pp. 1–2.

31. C. Bray, A. Cuisset, F. Hindle, G. Mouret, R. Bocquet, and V. Boudon, J. Quant. Spectrosc. Radiat. Transf. **203**, 349 (2017).

32. D. Stanze, A. Deninger, A. Roggenbuck, S. Schindler, M. Schlak, and B. Sartorius, J. Infrared, Millimeter, Terahertz Waves **32**, 225 (2011).

33. A. J. Deninger, A. Roggenbuck, S. Schindler, and S. Preu, J. Infrared, Millimeter, Terahertz Waves **36**, 269 (2015).

34. S. Preu, G. H. Döhler, S. Malzer, L. J. Wang, and A. C. Gossard, J. Appl. Phys. **109**, 061301 (2011).

35. J. R. Demers, F. Garet, and J.-L. Coutaz, IEEE Sensors Lett. **1**, 1 (2017).

36. J. R. Demers, J.-L. Coutaz, and F. Garet, in *Terahertz, RF, Millimeter, Submillimeter-Wave Technol. Appl. XI*, edited by L. P. Sadwick and T. Yang (SPIE, 2018), p. 20.

37. C. Cao, Z. Zhang, X. Zhao, and T. Zhang, Opt. Quantum Electron. **52**, 225 (2020).

38. M. R. Nowak, K. Nowak, M. Grzelczak, B. Szlachetko, L. Sterczewski, E. F. Plinski, P. Swiatek, M. Strzelecka, S. Plinska, and W. Malinka, in *2017 42nd Int. Conf. Infrared, Millimeter, Terahertz Waves* (IEEE, 2017), pp. 1–2.

39. W. Liu, C. Liu, J. Yu, Y. Zhang, J. Li, Y. Chen, and L. Zheng, Food Chem. **251**, 86 (2018).

40. H. Zhong, A. Redo-Sanchez, and X.-C. Zhang, Opt. Express **14**, 9130 (2006).

41. J. Zhang, Y. Yang, X. Feng, H. Xu, J. Chen, and Y. He, Front. Plant Sci. **11**, 1 (2020).

42. K. Wang, D. W. Sun, and H. Pu, Trends Food Sci. Technol. **67**, 93 (2017).

43. W. Liu, C. Liu, X. Hu, J. Yang, and L. Zheng, Food Chem. **210**, 415 (2016).

44. A. I. Knyazkova, A. V. Borisov, L. V. Spirina, and Y. V. Kistenev, J. Infrared, Millimeter, Terahertz Waves **41**, 1089 (2020).

45. A. Pohl, N. Deßmann, K. Dutzi, and H.-W. Hübers, J. Infrared, Millimeter, Terahertz Waves **37**, 175 (2016).

46. H. Zhang, Z. Li, T. Chen, and J. Liu, Optik (Stuttg). **138**, 95 (2017).

47. J. Liu, Optik (Stuttg). **131**, 885 (2017).

48. S. Yamaguchi, Y. Fukushi, O. Kubota, T. Itsuji, T. Ouchi, and S. Yamamoto, Sci. Rep. **6**, 30124 (2016).

49. J. Bou-Sleiman, J.-B. Perraud, B. Bousquet, J.-P. Guillet, N. Palka, and P. Mounaix, in *Millimetre Wave Terahertz Sensors Technol. VIII*, edited by N. A. Salmon and E. L. Jacobs (2015), p. 965109.

50. G. Chao, Y. Luo, and W. Ding, Mach. Learn. Knowl. Extr. **1**, 341 (2019).

51. E. Alpaydin, *Introduction to Machine Learning*, 3rd ed. (MIT Press, Cambridge, MA, 2014).

52. J. Ye and S. Ji, in *Biometrics* (John Wiley & Sons, Inc., Hoboken, NJ, USA, 2009), pp. 1–19.

53. H. Namkung, J. Kim, H. Chung, and M. A. Arnold, Anal. Chem. **85**, 3674 (2013).

54. Yuefang Hua and Hongjian Zhang, IEEE Trans. Microw. Theory Tech. **58**, 2064 (2010).

55. A. Roggenbuck, H. Schmitz, A. Deninger, I. C. Mayorga, J. Hemberger, R. Güsten, and M. Grüninger, New J. Phys. **12**, 043017 (2010).

56. D. W. Vogt and R. Leonhardt, Opt. Express **25**, 16860 (2017).

57. A. Roggenbuck, K. Thirunavukkuarasu, H. Schmitz, J. Marx, A. Deninger, I. C. Mayorga, R. Güsten, J. Hemberger, and M. Grüninger, J. Opt. Soc. Am. B **29**, 614 (2012).

58. D.-Y. Kong, X.-J. Wu, B. Wang, Y. Gao, J. Dai, L. Wang, C.-J. Ruan, and J.-G. Miao, Opt. Express **26**, 17964 (2018).

59. I. T. Jolliffe and J. Cadima, Philos. Trans. R. Soc. A Math. Phys. Eng. Sci. **374**, 20150202 (2016).

60. R. Bro and A. K. Smilde, Anal. Methods **6**, 2812 (2014).

61. J. Tang, S. Alelyani, and H. Liu, in *Data Classif. Algorithms Appl.*, edited by C. C. Aggarwal, 1st ed. (Chapman and Hall/CRC, 2014), pp. 37–64.

62. S. Khalid, T. Khalil, and S. Nasreen, in *2014 Sci. Inf. Conf.* (IEEE, 2014), pp. 372–378.

63. Z. M. Hira and D. F. Gillies, Adv. Bioinformatics **2015**, 1 (2015).

64. G. Chandrashekar and F. Sahin, Comput. Electr. Eng. **40**, 16 (2014).

65. T. Chen, Z. Li, X. Yin, F. Hu, and C. Hu, Spectrochim. Acta Part A Mol. Biomol. Spectrosc. **153**, 586 (2016).

66. J. El Haddad, B. Bousquet, L. Canioni, and P. Mounaix, TrAC Trends Anal. Chem. **44**, 98 (2013).

67. W. Xu, L. Xie, Z. Ye, W. Gao, Y. Yao, M. Chen, J. Qin, and Y. Ying, Sci. Rep. **5**, 11115 (2015).

68. Y. Xie and P. Sun, Opt. Quantum Electron. **50**, 46 (2018).

69. A. Tharwat, T. Gaber, A. Ibrahim, and A. E. Hassanien, AI Commun. **30**, 169 (2017).

70. W.-K. Ching, D. Chu, L.-Z. Liao, and X. Wang, Pattern Recognit. **45**, 2719 (2012).

71. J. Ye, T. Xiong, Q. Li, R. Janardan, J. Bi, V. Cherkassky, and C. Kambhamettu, in *Proc. 15th ACM Int. Conf. Inf. Knowl. Manag. - CIKM '06* (ACM Press, New York, New York, USA, 2006), p. 532.

72. Siuly, X. Yin, S. Hadjiloucas, and Y. Zhang, Comput. Methods Programs Biomed. **127**, 64 (2016).

73. Charu C. Aggarwal, in *Data Classif. Algorithms Appl.*, edited by C. C. Aggarwal (Chapman & Hall/CRC, 2014), pp. 1–36.

74. O. Boiman, E. Shechtman, and M. Irani, in *2008 IEEE Conf. Comput. Vis. Pattern Recognit.* (IEEE, 2008), pp. 1–8.

75. P. Mulak and N. Talhar, Int. J. Sci. Res. **4**, 2101 (2015).

76. G. Bhattacharya, K. Ghosh, and A. S. Chowdhury, Pattern Recognit. Lett. **33**, 356 (2012).

77. R. O. Duda, P. E. Hart, and D. G. Stork, *Pattern Classification*, 2nd ed. (Wiley, New York, 2000).

78. M. Kubat, *An Introduction to Machine Learning*, 2nd ed. (Springer International Publishing, 2017).

79. C. Cortes and V. Vapnik, Mach. Learn. **20**, 273 (1995).

80. G. James, D. Witten, T. Hastie, and R. Tibshirani, *An Introduction to Statistical Learning* (Springer New York, New York, NY, 2013).

81. T. Hastie, R. Tibshirani, and J. Friedman, *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*, 2nd ed. (Springer, New York City, 2009).

82. P.-W. Wang and C.-J. Lin, in *Data Classif. Algorithms Appl.*, edited by C. C. Aggarwal (Chapman & Hall/CRC, 2014), pp. 187–204.

83. J. Huang, J. Liu, K. Wang, Z. Yang, and X. Liu, Spectrochim. Acta Part A Mol. Biomol. Spectrosc. **198**, 198 (2018).

84. N. Palka, Acta Phys. Pol. A **120**, 713 (2011).

85. C. M. Bishop, *Pattern Recognition and Machine Learning*, 1st ed. (Springer-Verlag New York, 2006).

86. K. Z. Mao, Feng Yang, and Wenyin Tang, in *2011 IEEE Symp. Comput. Intell. Bioinforma. Comput. Biol.* (IEEE, 2011), pp. 1–7.

87. J. H. Friedman, J. Am. Stat. Assoc. **84**, 165 (1989).

88. E. Neto, F. Biessmann, H. Aurlien, H. Nordby, and T. Eichele, Front. Aging Neurosci. **8**, 273 (2016).

89. C. Su, F. Tu, X. Zhang, B. Shia, and T. Lee, J. Data Sci. **17**, 1 (2019).

90. Y. Guo, T. Hastie, and R. Tibshirani, Biostatistics **8**, 86 (2007).

91. G. James, D. Witten, T. Hastie, and R. Tibshirani, in *An Introd. to Stat. Learn.* (Springer, New York, NY, 2013), pp. 175–201.

92. R. Liu and D. F. Gillies, Pattern Recognit. **53**, 73 (2016).