Classification of Non-Referenced Continuous-Wave Terahertz Reflection Spectra for Remote Material Identification

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

Published in:
Infrared Physics and Technology

DOI (link to publication from Publisher):
10.1016/j.infrared.2024.105420

Publication date:
2024

Document Version
Publisher's PDF, also known as Version of record

Infrared Physics & Technology 140 (2024) 105420

Contents lists available at ScienceDirect

Infrared Physics and Technology

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

Research paper

Classification of non-referenced continuous-wave terahertz reflection spectra for remote material identification

Mathias Hedegaard Kristensen *, Paweł Piotr Cielecki, Esben Skovsen

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

ARTICLE INFO

Dataset link: 10.5281/zenodo.5079558

Keywords:
Terahertz reflection spectroscopy
Terahertz screening
Terahertz remote sensing
Material identification
Continuous-wave terahertz
Terahertz frequency-domain spectroscopy
Machine learning
Dimensionality reduction
Linear discriminant analysis
Principal component analysis

ABSTRACT

Commonly, terahertz spectra (both continuous-wave and pulsed) are deconvoluted by reference spectra to remove the water vapor absorption lines and other system related responses. However, in real-life applications obtaining reference spectra can be problematic and adds to the complexity of the system. Thus, a reference-free method for classification of terahertz spectra could be a welcomed advance for remote sensing applications. In this paper, we study how simple machine learning algorithms perform as a reference-free method for terahertz stand-off identification of materials. The algorithms are trained using spectra measured under controlled humidity conditions and tested by a completely independent data set measured under ambient conditions. We apply three different classification algorithms; namely a Gaussian Bayes model, the k nearest neighbors, and a support vector machine. We found that, if the terahertz spectra are processed using a supervised algorithm (Regularized Linear Discriminant Analysis), very high classification scores (>98.6%) can be retained for the non-referenced spectra. Moreover, the high accuracy is obtained meanwhile the dimensionality is reduced by a factor larger than 160, which further reduces the computational requirements. Hence, we have demonstrated that simple supervised machine learning algorithms can serve as a highly accurate reference-free method for THz material identification. This could be of great importance for real-world remote sensing applications based on terahertz spectroscopy.

1. Introduction

Terahertz (THz) spectroscopy has proven to be a promising technology for security, defense, safety, and quality control applications [1,2] since many compound materials exhibit unique spectroscopic characteristics. Including hazardous substances such as explosives, commercial and illicit drugs, and toxic gasses, *e.g.*, ammonia or carbon monoxide. Recently, researchers have demonstrated THz spectroscopy to be a powerful tool for identification of black plastics in the scope of recycling [3], where other optical techniques fall short. Concurrently, THz radiation allows for non-invasive screening as many non-polar and non-metallic materials are transparent within the THz frequency band. In addition, the low photon energy inherent to THz radiation results in a non-ionizing nature and is thus harmless to biological samples at low power levels. Hence, THz spectroscopy is not only desired but also safe for screening of personnel and objects.

For many real-world applications of THz spectroscopy, it is necessary to be able to distinguish different substances in an efficient and reliable manner. Additionally, many applications require the THz spectroscopic measurements to be done in a reflection scheme. The weak and broad spectroscopic characteristics of substances inherent to reflection spectra, caused by the dependence on the refraction index [4], complicates things further.

Several rather complex machine learning (ML) techniques including Bayesian models [5,6], artificial neural networks [7–9], support vector machine [10–12], and random forests [5,7,11] have previously been utilized for classification of THz spectra. Generally, ML algorithms are said to be supervised or unsupervised if class membership information of the data is included or not in the training of the algorithm. Additionally, ML algorithms can be utilized to reduce the number of features, while preserving relevant information, which can facilitate the classification task. Since THz spectra are multivariate data consisting of 100's or 1000's discrete frequencies, this is a very pertinent capacity. Thus, such dimensionality reduction (DR) methods are typically applied to make identification algorithms more efficient. This lowers the computational requirements, increases the learning speed of the ML algorithm, and allows for visualization of the data for easier interpretation. Previously, we compared the performance of two such unsupervised and supervised linear DR methods, respectively, under different conditions [13]. We showed that even simple ML algorithms

* Corresponding author.
E-mail address: mhkr@icloud.com (M.H. Kristensen).

https://doi.org/10.1016/j.infrared.2024.105420
Received 22 January 2024; Received in revised form 13 May 2024; Accepted 21 June 2024
Available online 27 June 2024

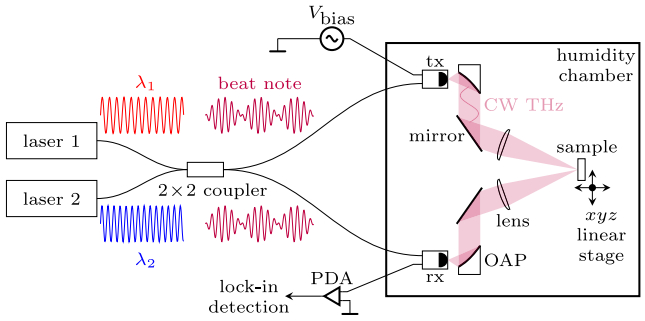

Fig. 1. Illustration of the experimental setup.

are sufficient for highly accurate classification of THz reflection spectra. Most recently, Park et al. [14] have reviewed the use of ML algorithms for THz time-domain spectroscopy (TDS) and THz imaging, underpinning the importance of ML in THz applications.

All the above-mentioned results, ours included, relied on referenced THz measurements, *i.e.* THz spectra that have been deconvoluted by proper reference spectra. However, the need for reference spectra can be very inconvenient or even impossible to fulfill in applications outside the lab. Nevertheless, only very few attempts have previously been made to develop reference-free methods. Zhang et al. [15] have reported on reference-free phase imaging in transmission for identification of three explosive materials. Here, they demonstrated identification of the samples by their absorption signatures extracted from the first-order derivative of the instantaneous THz phase divided by the frequency. But the presented methodology is limited to transmission measurements. Later, Zhong et al. [16] extended the technique to be used in THz reflection TDS. Similarly, the absorption signatures of the materials are extracted from the second-order derivative of the THz phase with respect to the frequency. However, this extended method is only reference-free in the ideal case, where atmospheric absorption can be ignored, and thus, it is not applicable for stand-off applications in real-world scenarios. Nonetheless, the need for a reliable and truly reference-free methodology remains, and its importance for the advancement of THz remote sensing applications is immediately evident.

In this paper, we present a robust and highly accurate reference-free identification methodology based on supervised ML algorithms for THz reflection measurements, that is readily available for out-of-the-lab THz remote sensing applications. To demonstrate, we present the classification results of non-referenced vs. referenced continuous-wave (CW) THz reflection spectra processed by a supervised DR method (Regularized Linear Discriminant Analysis, RLDA) for three different classifiers: Bayesian, k nearest neighbors (k-NN), and support vector machines (SVM). These classifiers represent a probabilistic, a non-parametric, and a finely tuned linear algorithm, respectively. We compare the results with those obtained following an unsupervised DR method (Principal Component Analysis, PCA). Noteworthy, our results are achieved using the simple, yet powerful, and well-established RLDA to reduce the dimensionality of the multivariate data by a factor larger than 160. Our results clearly shows that, in the scope of THz remote material identification, supervised algorithms are superior to their unsupervised counterparts. The method could readily be adapted for pulsed THz reflection spectra as well.

2. Methods and materials

2.1. Experimental setup, samples, and measurements

The samples were characterized from 0.09 to 1.19 THz in a reflection geometry by CW THz frequency-domain spectroscopy (THz-FDS) using a TeraScan 1550 system manufactured by Toptica Photonics. The angle of incidence was approx. 11°. The THz path was enclosed by a custom-built humidity chamber that could be purged with either dry or water vapor saturated nitrogen to achieve relative humidity (RH) levels between 5% and 95% within a ±2% points accuracy. A sketch of the experimental setup is seen in Fig. 1 Five compound materials with spectral characteristics in the frequency range of the TeraScan 1550 were selected for the study. Including galactitol, L-tartaric acid (L-TA), 4-aminobenzoic acid (PABA), theophylline, and alpha-lactose monohydrate. Six pellets from each compound were fabricated in pairs at weight percentages of 20%, 50%, and 80% of active material mixed with polyethylene (PE) powder as well as two pellets of pure PE. The flat response of PE in the spectral band of interest makes it a convenient binder matrix for the active compounds [17]. The sample pellets were shaped as a 15° wedge to avoid interference between the front and the rear surface reflections (*i.e.* reflections from rear surface never reaches the detector). Moreover, the sample holder was designed with a clear aperture much larger than the THz spot size. Each sample was measured over the entire spectral range in 80 MHz increments integrated for 3 ms at different positions, picked randomly within the sample surface for each measurement. Thus, each measurement is performed at a new and uncorrelated spot on the sample surface. Two different data sets were recorded for the training and testing of the ML algorithms, respectively. The training data set consisted of 1920 spectra, that is each sample was measured 20 times under controlled conditions of 10%, 50%, and 90% RH, respectively. Hence, each material was represented by 360 spectra (120 spectra in the case of pure PE) A reference spectrum was recorded for every 40 measurements replacing the sample with an aluminum mirror. For the testing data set, the same samples were each measured 80 times under ambient conditions giving a total of 2560 spectra, *i.e.* 480 spectra of each material plus 160 of pure PE. Here, a reference spectrum was recorded for every 20 measurements. Noteworthy, the entire data set (training and test) included almost 4500 spectra and the ratio between the training and test sets was 3:4. All the data is available online as a part of the *"Database of frequency-domain terahertz reflection spectra for the DETRIS project"* [18]. A thorough description of the experimental setup and the measuring procedure can be found in Refs. [13,18].

2.2. Data preprocessing

The CW THz-FDS setup operates in a coherent detection scheme, which causes phase oscillations in the recorded photocurrent $I_{ph}(\nu)$ as the THz frequency ν is scanned. In this study, the instantaneous amplitude $A(\nu)$ and instantaneous phase $\phi(\nu)$ were calculated by applying the Hilbert transformation $\mathcal{H}$ to the oscillating photocurrent proportional to the THz electric field [19,20]. Let us note, that using the Hilbert transform, the spectral resolution is equal to the optical frequency step size and independent of the THz optical path length. The resulting complex-valued analytic signal

$$I_a(\nu) = I_{ph}(\nu) + i\mathcal{H}\left\{I_{ph}(\nu)\right\} = A(\nu)\exp\left[i\phi(\nu)\right]$$

was then Fourier transformed into the time-domain, in which it was filtered for any reflections in the experimental setup causing Fabry–Pérot interference, and inversely Fourier transformed back into the frequency-domain [21]. However, not all reflections could be filtered as it would degrade the effective spectral resolution. Subsequently, the data was cropped to an interval from 0.4 to 1.05 THz to include only the spectral region containing spectroscopic characteristics of samples and achieving a proper signal to noise ratio. In coherence to our previous study [13], we used only the spectral amplitude $A(\nu)$ of the THz field in the further processing. Nonetheless, the same unique material information is contained in the spectral phase $\phi(\nu)$ as for THz-TDS [19,20]. The reflection coefficient can be obtained by deconvoluting the spectrum with an appropriate reference spectrum:

$$r(\nu) = A_{sample}(\nu)/A_{reference}(\nu). \quad (1)$$

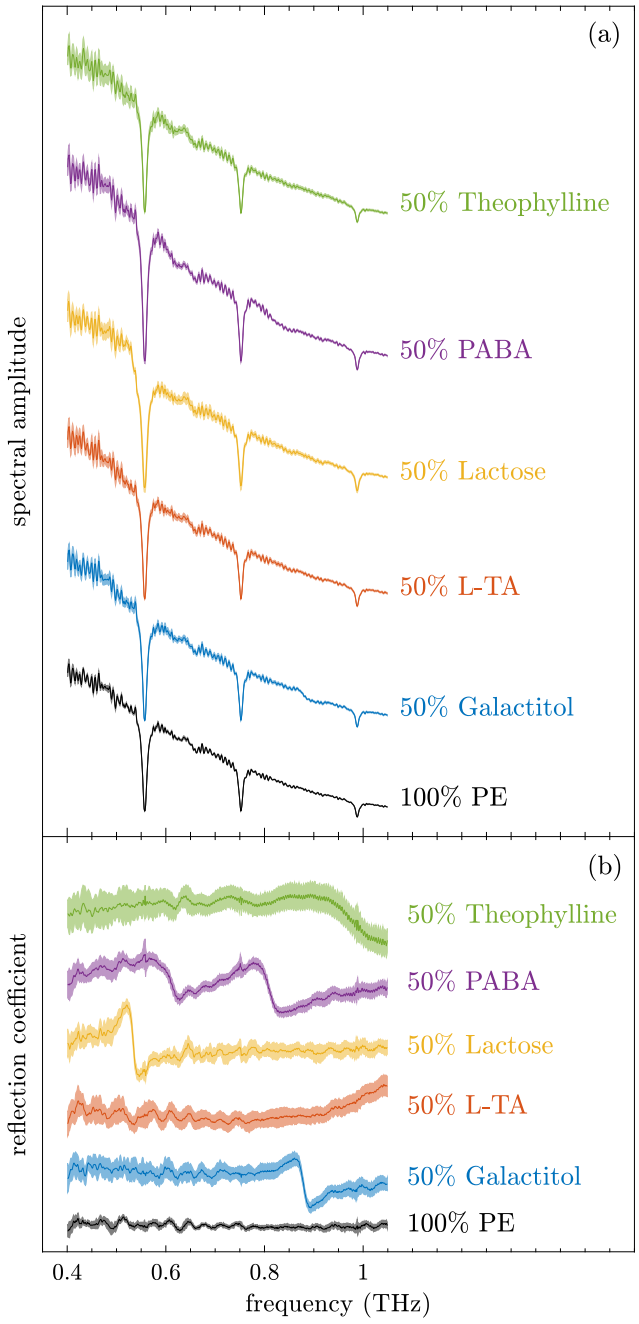

Fig. 2. Non-referenced (a) and referenced (b) THz reflection spectra of the samples with 50% of active material and pure PE measured under ambient conditions. The dark colored center line of each curve is the mean of 160 measurements, while the light colored fill represents the standard deviation. The curves of each material are shifted vertically for a better readability.

Prior to the calculation of $r(\nu)$, the data was interpolated onto integer GHz-frequencies to ensure a proper deconvolution of the individual spectral components. Each spectrum in the data sets spanned 649 discrete frequencies. In the remaining part of this paper, we categorize the spectral data $A(\nu)$ and $r(\nu)$ as *non-referenced* and *referenced*, respectively.

The non-referenced CW THz reflection spectra of the samples with 50% active material and of pure PE, measured under ambient conditions, are seen in Fig. 2(a). The center line of each curve is the mean of 160 measurements, while the filled area represents the standard deviation. The curves are shifted vertically for a better readability. Each spectrum shows clear absorption lines of atmospheric water vapor around 0.55, 0.75, and 0.99 THz, while the spectral characteristics intrinsic to the studied materials are much harder to recognize. The referenced reflection spectra are seen in Fig. 2(b). Evidently, the material specific characteristics are now easily recognized. However, when the weight percentage of the active material in a sample drops to from 50% to 20%, it becomes difficult to distinguish or identify materials like theophylline, L-TA and PE [13]. It should be emphasized that both the non-referenced [Fig. 2(a)] and the referenced data [Fig. 2(b)] originate from the exact same data, and that the latter have only been deconvoluted with appropriate reference spectra to obtain the reflection coefficients of the samples.

2.3. Machine learning

Let us start by defining some technical terms that are pertinent, when discussing ML algorithms. If the task of the algorithm is to assign each input to one of several discrete categories, it is said to be a *classification* problem. By contrast, the task is a *regression* problem, if the output is one or several continuous variables. When dealing with classification of THz spectra, the term *class* refers to the categories that the spectra can be divided into based on their spectroscopic characteristics. Particularly for material identification, each type of material would constitute a class. The data points in every class can be further specified using *labels*. However, if there are no particular differences between the spectra within each class, the labels within a given class are simply the class label. In classification tasks, the model will map each input onto a class label. When dealing with classification of THz spectra, we shall regard each discrete frequency component constituting the THz spectra as a random variable. Finally, such individual measurable properties of an observed phenomenon are referred to as *features* in machine learning terminology. However, we will in this paper restrict ourselves to solely use the term when referring to the variables in the lower-dimensional feature spaces of the DR methods. That is, the new variables in the linear discriminant and/or principal component feature space.

In the following, we will let X be our data matrix of size $N \times M$ such that each row n represents a THz spectrum S_n (observation), and each column m a discrete frequency component ν_m (variable), *i.e.*

$$X = \begin{bmatrix} S_1(\nu_1) & S_1(\nu_2) & \cdots & S_1(\nu_M) \\ S_2(\nu_1) & S_n(\nu_m) & \cdots & S_2(\nu_M) \\ \vdots & \vdots & \ddots & \vdots \\ S_N(\nu_1) & S_N(\nu_2) & \cdots & S_N(\nu_M) \end{bmatrix} \tag{2}$$

Linear Discriminant Analysis (LDA) is a common supervised DR method for transforming data into a lower-dimensional feature space. During the training phase, the algorithm includes class information to calculate distances between class means and within-class variances, aiming to maximize class separation and minimize intra-class spread in the projection space. However, we shall now give a more rigorous introduction of LDA. Let $\vec{s}_i$ represent the ith row of the data matrix X (the ith observation/spectrum). The class information is utilized to label each spectrum $\vec{s}_i$ such that the data matrix X can be partitioned into K classes C_j of n_j spectra:

$$X = \begin{bmatrix} C_1 \\ \vdots \\ C_K \end{bmatrix} \quad \text{with} \quad C_1 = \begin{bmatrix} \vec{s}_1 \\ \vdots \\ \vec{s}_{n_1} \end{bmatrix}, \quad C_2 = \begin{bmatrix} \vec{s}_{n_1+1} \\ \vdots \\ \vec{s}_{n_2} \end{bmatrix}, \quad \ldots \tag{3}$$

Let us also introduce the linear combination of each spectrum's discrete frequency components

$$y_i = \sum_{m=1}^{M} w_m S_i(\nu_m) = \vec{w}^{\mathsf{T}} \vec{s}_i^{\mathsf{T}}, \tag{4}$$

where $\vec{w}$ is a row vector of constants $w_1, \ldots, w_M$. T denotes transpose. That is, y_i is the 1-dimensional projection of the ith spectrum. Then,

the separation of the different classes in the projection space can be quantified by the between-class scatter

$$\sum_{j=1}^{K} n_j (m_j - m)^2, \tag{5}$$

where

$$m_j = \frac{1}{n_j} \sum_{\vec{s}_i \in C_j} y_i = \vec{w}^{\mathsf{T}} \vec{\mu}_j \quad \text{with} \quad \vec{\mu}_j = \frac{1}{n_j} \sum_{\vec{s}_i \in C_j} \vec{s}_i^{\mathsf{T}} \tag{6}$$

is the jth class centroid in the projection space and

$$m = \frac{1}{N} \sum_{j=1}^{K} n_j m_j = \vec{w}^{\mathsf{T}} \vec{\mu} \quad \text{with} \quad N = \sum_{j=1}^{K} n_j \tag{7}$$

is the weighted mean of the projected centroids. $\vec{\mu}_j$ and $\vec{\mu}$ are the corresponding jth class and global centroids, respectively, in the original space. When these expressions are inserted in Eq. (5), the between-class scatter can be written as

$$\sum_{j=1}^{K} n_j (m_j - m)^2 = \vec{w}^{\mathsf{T}} S_{\mathrm{B}} \vec{w}, \tag{8}$$

where the between-class scatter matrix

$$S_{\mathrm{B}} = \sum_{j=1}^{K} n_j (\vec{\mu}_j - \vec{\mu})(\vec{\mu}_j - \vec{\mu})^{\mathsf{T}} \tag{9}$$

is calculated in the original space. Similarly, the spread of each projected class j can be quantified by the within-class scatter

$$\sigma_j = \sum_{\vec{s}_i \in C_j} (y_i - m_j)^2 \tag{10}$$

Hence, it is easy to show that the total within-class scatter of the projected classes is

$$\sum_{j=1}^{K} \sigma_j = \vec{w}^{\mathsf{T}} S_{\mathrm{W}} \vec{w} \tag{11}$$

with the total within-class scatter matrix

$$S_{\mathrm{W}} = \sum_{j=1}^{K} \sum_{\vec{s}_i \in C_j} (\vec{s}_i^{\mathsf{T}} - \vec{\mu}_j)(\vec{s}_i^{\mathsf{T}} - \vec{\mu}_j)^{\mathsf{T}}. \tag{12}$$

Let us recall that LDA seeks to maximize the distance between the projected class means and minimize the within-class variance. This is equivalent to maximizing the ratio

$$\arg\max_{\vec{w}} \left\{ \frac{\vec{w}^{\mathsf{T}} S_{\mathrm{B}} \vec{w}}{\vec{w}^{\mathsf{T}} S_{\mathrm{W}} \vec{w}} \right\}. \tag{13}$$

The arg max denotes that we seek arguments that maximize the function. This optimization problem is not bound, so we require the weights $\vec{w}$ to be of unit length. Moreover, we introduce the constraint $\vec{w}^{\mathsf{T}} S_{\mathrm{W}} \vec{w} = 1$, since we are only concerned with directions. Thus, we need to solve

$$\arg\max_{\vec{w}^{\mathsf{T}} S_{\mathrm{W}} \vec{w} = 1} \left\{ \vec{w}^{\mathsf{T}} S_{\mathrm{B}} \vec{w} \right\}. \tag{14}$$

This is a generalized eigenvalue problem:

$$S_{\mathrm{B}} \vec{w} = \lambda S_{\mathrm{W}} \vec{w}. \tag{15}$$

Given S_{W} is not singular, we can write

$$S_{\mathrm{W}}^{-1} S_{\mathrm{B}} \vec{w} = \lambda \vec{w}, \tag{16}$$

which is easily solved by eigendecomposition of $S_{\mathrm{W}}^{-1} S_{\mathrm{B}}$. The optimal solution is the eigenvector $\vec{v}_1$ associated with the largest eigenvalue λ_1, and so forth. However, at most $K - 1$ non-zero eigenvectors exist because $\mathrm{rank}(S_{\mathrm{B}}) \leq K - 1$ (its columns are linearly dependent). Furthermore, $S_{\mathrm{W}}^{-1} S_{\mathrm{B}}$ is not necessarily symmetric, thus, the eigenvectors are not generally orthogonal. Consequently, the dimensionality of the original data matrix X is reduced to $k \leq K - 1$ features by projecting it onto the eigenvectors. The projections $\vec{y}_k = X\vec{v}_k$ are the *linear discriminants* (LDs) comparable to the principal components (PCs) of PCA. However, the assumption that S_{W} is non-singular is often not true. A common solution is to regularize S_{W} by

$$S'_{\mathrm{W}} = S_{\mathrm{W}} + \beta I \tag{17}$$

with regularization parameter β and identity matrix I. Then by eigendecomposition, we have

$$S'_{\mathrm{W}} = Q \Lambda Q^{\mathsf{T}} + \beta I = Q(\Lambda + \beta I) Q^{\mathsf{T}}. \tag{18}$$

Here Q is the square matrix containing the eigenvectors and Λ the diagonal matrix of the eigenvalues. This is known as Regularized-LDA (RLDA). To obtain the optimal regularization value β, we applied a 10-fold stratified cross-validation on the training set.

Principal Component Analysis projects the data onto a lower-dimensional space, aiming to maximize the variance of the full data set disregarding any class information. Hence, it is an unsupervised method, which relies solely on patterns in the data. First, we assume that the variables $\vec{x}_m$ (columns of the data matrix X) have been standardized, *i.e.* each column has zero mean and unit variance [22], which is crucial for PCA to operate correctly [23,24]. Then, we construct a linear combination $\vec{y}$ of the individual variables, *i.e.*

$$\vec{y} = \sum_{m=1}^{M} w_m \vec{x}_m = X\vec{w}, \tag{19}$$

where $\vec{w}$ is a vector of constants $w_1, \ldots, w_M$. The PCA algorithm searches for the optimal weights $\vec{w}$ that maximizes the variance $\mathrm{var}(\vec{y})$, *i.e.*

$$\arg\max_{\|\vec{w}\|=1} \left\{ \mathrm{var}(\vec{y}) \right\}. \tag{20}$$

The constraint $\|\vec{w}\| = \vec{w}^{\mathsf{T}} \vec{w} = 1$ secures that the weights are normalized. Otherwise, the variance could attain an arbitrary large value for an optimal $\vec{w}$. The variance is calculated in the usual way:

$$\mathrm{var}(\vec{y}) = \frac{1}{N-1} \sum_{n=1}^{N} (y_n - \bar{y})^2 = \frac{\vec{y}^{\mathsf{T}} \vec{y}}{N-1}. \tag{21}$$

Here, it is implied that the data matrix X and, hence, all linear combinations $\vec{y}$ are mean-centered ($\bar{y} = 0$). The optimization task can then be restated as

$$\arg\max_{\|\vec{w}\|=1} \left\{ \vec{w}^{\mathsf{T}} \Sigma \vec{w} \right\} \tag{22}$$

using the covariance matrix $\Sigma = X^{\mathsf{T}} X/(N-1)$. This is a standard problem in linear algebra solved by eigendecomposition [23,24]. The resulting orthogonal eigenvectors $\vec{v}_m$ are ranked in descending order according to the eigenvalues λ_m. The associated linear combinations $\vec{y}_m = X\vec{v}_m$ are the so-called *principal components* (PCs) of the data. The eigenvalues equal the variances of the PCs $\vec{y}_m$. The eigenvector $\vec{v}_1$ yields the optimal solution to Eq. (8), resulting in principal component $\vec{y}_1$ with maximal variance λ_1. Equivalently, the successive eigenvectors represent the next orthogonal (uncorrelated) PCs along which the maximal proportion of the remaining variance in the data is captured, respectively. The overall variance can, hence, be calculated by summing all the eigenvalues. The quality of each PC can therefore be quantified by the amount of total variance it explains, *i.e.*

$$\frac{\lambda_m}{\sum_{m=1}^{M} \lambda_m} \tag{23}$$

Finally, the dimensionality of the data can be reduced while maintaining most of the information by retaining only a relative few of the most significant PCs.

Before turning to the classification task, a natural question arises: How many features should be kept? In the case of PCA, one typically looks at the proportion of explained variance [Eq. (9)] vs. the number

of features kept to aid this decision [24]. However, for LDA and RLDA this quantity is not equivalent to the proportion of explained variance, since the eigenvalues are related to the between-class and within-class variances, and hence, reflect the robustness and the ability to discriminate between different classes [25]. Instead as a common frame of reference, we can calculate the captured variance (*i.e.* the cumulative variance along each new feature normalized to the total variance) to aid the decision of how many features should be kept [24].

For classification, three different algorithms were applied. First, a probabilistic classification model based on Bayes' theorem. The Bayes classifier calculates the posterior probability $p(C_j|\vec{s})$ of an observation with value $\vec{s}$ to belong to the jth class C_j as the product of the prior probability that the observation belongs to the jth class $p(C_j)$ and the likelihood $p(\vec{s}|C_j)$ of an observation $\vec{s}$ given class j normalized by the marginal probability of $\vec{s}$. That is [26],

$$p(C_j|\vec{s}) = \frac{p(\vec{s}|C_j)p(C_j)}{p(\vec{s})}, \qquad p(\vec{s}) = \sum_j^K p(\vec{s}|C_j)p(C_j)$$

Accordingly, the Bayes classifier appoints the observation to the class with the highest posterior probability. Here, we assumed the class likelihood function to follow a multivariate normal distribution. Second, the simple non-parametric k Nearest Neighbors algorithm. This classification algorithm stores all training data, calculates the geometrical distance to a new observation, and subsequently, classifies the new observation identically to the majority of the k nearest neighboring data points. In our study we utilized the Euclidean distance metric and $k = 746$ equal to the average number of observations in each class. Choosing such a large value of k means that the k-NN algorithm will be an indicator of how well LDA/RLDA and PCA are at intra-class grouping. The final approach was the support vector machine. The algorithm searches for a hyperplane separating observations of two classes with maximal margin, i.e. the maximum geometrical distance to both classes. A new observation is classified according to the half-space it is belonging. Real-world data is often linearly inseparable, wherefore a soft margin is often applied. It allows data points to violate the hyperplane at the cost of a penalty. This results in a wider margin that generalizes better to unseen data. We utilized a linear kernel and a soft margin approach together with a 10-fold cross-validation on the training set. The data processing was performed in MATLAB (MathWorks, R2020b).

3. Results and discussion

The LDA algorithm was separately trained on the non-referenced and the referenced THz data using the data sets based on 1920 spectra measured at controlled RHs (10%, 50% and 90% RH). For comparison, the PCA algorithm was trained in an identical manner. The generalization of both algorithms were then investigated in the testing phase based on 2560 spectra measured under ambient conditions in their non-referenced and referenced form, respectively.

In Fig. 3(a)–(d), the data is projected onto the two most prominent features of the new feature spaces of LDA and PCA, respectively. The training and test sets are respectively plotted with light colored squares and dark colored dots. For the non-referenced [Fig. 3(a) and (c)] data, both LDA and PCA show a clear discrepancy between the training and the test data. For LDA [Fig. 3(c)], the training data is tightly clustered according to the respective classes, while the test data are loosely spread around these clusters. A similar trend is seen for the referenced data [Fig. 3(d)], but to a lesser extent. Thus, the LDA algorithm seems to overfit the data in the training phase leading to a somewhat poor generalization to the new unseen data in the test phase. This conclusion is also backed up when inspecting the LDA eigenvectors (not shown here), which overall fit to the noise pattern instead of the spectroscopic characteristics of the samples. Overall, the PCA algorithm [Fig. 3(a)–(b)] clusters the data according to the sample material and further by the material concentration as the algorithm

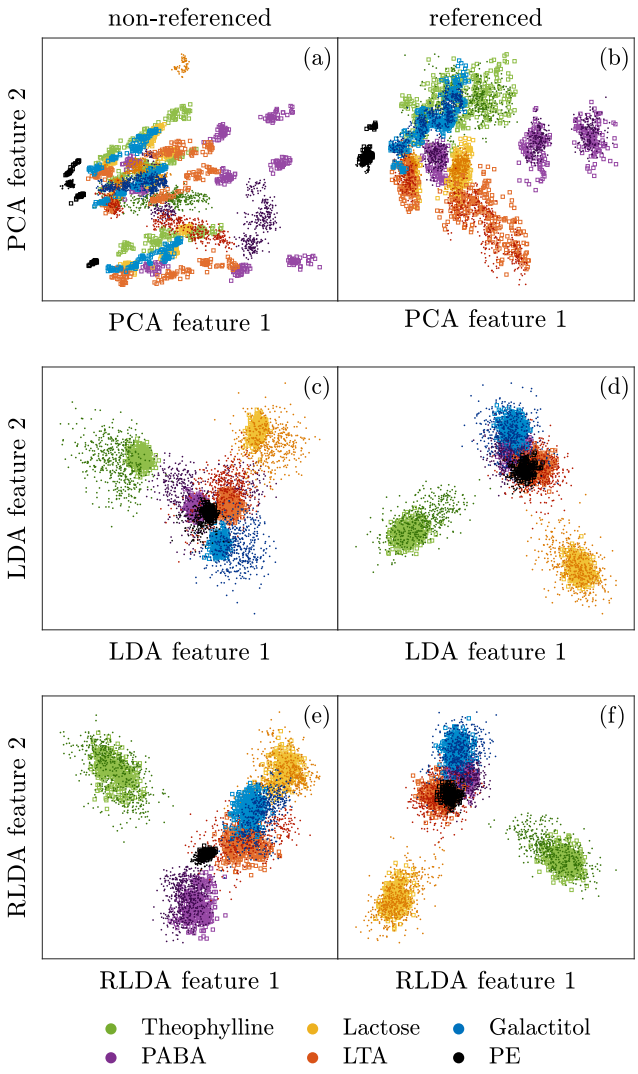

Fig. 3. Projections of the non-referenced (left panels) and referenced (right panels) spectra to a two-dimensional feature space processed by PCA (a)–(b), LDA (c)–(d), and RLDA (e)–(f), respectively. The training data is marked by light colored squares, while the test data is marked by dark colored dots.

aims to maximize the overall variance. Thus, PCA results in poor intra-class grouping in contrast to LDA, which cluster each type of material regardless of concentration. For the non-referenced data [Fig. 3(a)], the PCA algorithm even separated the data in different planes according to the RH conditions. Hence, as the test data was recorded under ambient conditions, it is generally located in a plane between the data recorded at 10% and 50% RH. This is not seen for the referenced data [Fig. 3(b)], as the water vapor absorption signatures are removed by reference spectra. These findings are in excellent agreement with our previous study in Ref. [13], where the data were referenced using non-ideal reference spectra. To overcome the issue of overfitting with LDA, we introduced a regularization parameter β [Eq. (17)]. The RLDA algorithm was trained similarly to LDA and PCA with the addition of a 10-fold stratified cross-validation to obtain the optimal regularization value β In Fig. 3(e)–(f), we see a very good clustering of the different materials using RLDA, which is less tight compared to LDA, for both the non-referenced and the referenced data. More importantly, RLDA shows a excellent agreement between the test and the training data with a slightly larger spread for the non-referenced data. Let us note, that even though materials like LTA, Galactitol and Lactose seem to overlap in the two-dimensional feature space, all materials are well-separated

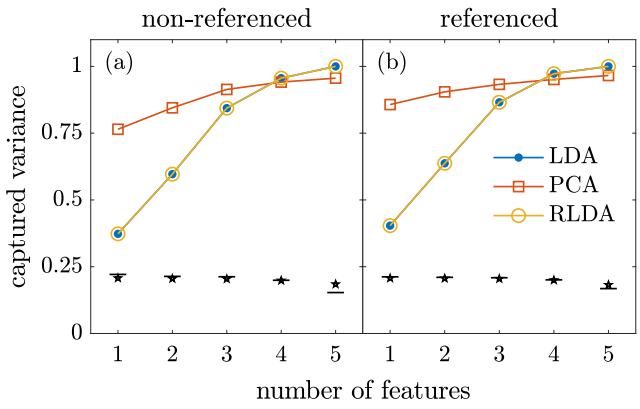

Fig. 4. Cumulative captured variance of the non-referenced (a) and referenced (b) training data vs. the number of included features in the reduced feature space of LDA, PCA, and RLDA. The normalized eigenvalues of LDA (pentagon) and RLDA (horizontal bar) are plotted for comparison.

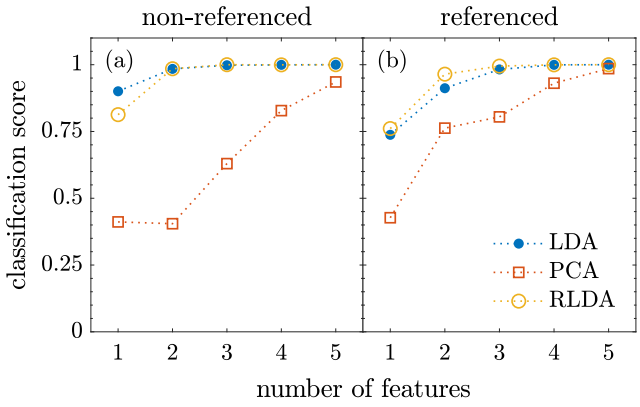

Fig. 5. Bayes classification scores of the non-referenced (a) and referenced (b) training data vs. the number of included features in the reduced feature space of LDA, PCA, and RLDA.

when projected onto a higher-dimensional feature space. By contrast to LDA, the RLDA eigenvectors (not shown here) fit to the spectroscopic characteristics of the materials and not the noise. Therefore, we can conclude that RLDA generalizes exceptionally to the unseen test data, which notably were recorded under ambient conditions, even in the non-referenced case.

In Fig. 4, we plot the (cumulative) captured variance computed for the LDA/RLDA and PCA of the train sets vs. the number of features. Let us recall, that while PCA can retain the dimensionality of the original data (*i.e.* 649 features), LDA/RLDA can at most project the data onto $K - 1 = 5$ features. Hence, Fig. 4 includes a maximum of five features. LDA algorithms perform almost equally well, whether the input data is referenced or not, while there is a significant difference in performance of the PCA algorithm up to three features. Interestingly, while LDA/RLDA only captures a small part of the overall variance below three features, both algorithms capture a larger part of the overall variance than PCA above three included features. In particular, 100% of the variance is captured by the five most prominent LDA/RLDA features, where PCA roughly captures 96% of the variance. For the sake of comparison, we have plotted as well the normalized eigenvalues of LDA (pentagon) and RLDA (horizontal bar). Noteworthy, each LDA/RLDA feature achieves equal discriminating power or robustness of approx. 20%. This is in stark contrast to PCA, where the first feature is exclusively appointed a quality (proportion of explained variance) of 75%–85%, and clearly demonstrates the excellence of the supervised algorithms for material identification tasks.

Table 1
Classification accuracy scores of the LDA-, PCA, and RLDA-processed data projected onto a four-dimesional feature space.

		LDA		PCA		RLDA	
		Train	Test	Train	Test	Train	Test
Bayes	Non-referenced	0.9995	0.8941	0.8281	0.7998	0.9995	0.9863
	Referenced	0.9995	0.9746	0.9307	0.9246	0.9995	0.9906
746-NN	Non-referenced	0.9375	0.8960	0.5182	0.5985	0.9370	0.9339
	Referenced	0.9375	0.9289	0.5885	0.6016	0.9375	0.9328
SVM	Non-referenced	0.9984	0.8737	0.8089	0.7819	1.0000	0.9746
	Referenced	0.9990	0.8753	0.9328	0.9398	0.9995	0.9578

However, the most captured variance is not necessarily equal to the best performance in terms of classification. Therefore, we also investigated the classification accuracy of the Bayes classifier vs. the number of features in the training phase. The results are shown in Fig. 5. Again, the supervised algorithms (LDA/RLDA) display a much better performance compared to PCA for both the non-referenced and referenced data. However, we should keep in mind that the perfect classification scores of LDA (but not RLDA) is due to overfitting, which is evident from the equivalent classification scores of the test data (not shown here). What is even more impressive is the fact that the RLDA algorithm in general achieves better classification scores using the non-referenced data. A closer look on the RLDA curves (non-referenced and referenced) reveals classification accuracies >99.94% and 100% for four and five retained features, respectively. Therefore, we chose to retain just four features onward for the classification task. Here we would like to point out the impressive reduction of dimensionality by a factor larger than 160. In this case, the optimal regularization value β of RLDA was respectively 0.002 and 0.050 for the non-referenced and referenced data.

The Bayes, k-NN, and the SVM classification algorithms are applied to quantify the performance of the DR methods on the non-referenced and referenced spectra, respectively. The classification scores is given in Table 1. At first glance, we see that the classification scores of the non-referenced and referenced data are rather similar and generally $\gtrsim$90% for all three classifiers. This clearly demonstrates that THz reflection spectra can be accurately classified without deconvolution by an precise reference measurement. Moreover, the good agreement of the train and the test scores of RLDA indicates that the algorithm generalizes very well, in contrast to the overfitting LDA. Among the three classifiers, the performance of the 746-NN classifier is inferior. Particularly, for the PCA-processed data with classification scores between 50%–60%. This is related to our choice of k equal to the total number of observations within each class, that was intended to verify the inter-class grouping. As expected, LDA/RLDA exhibit superior performance due to the clustering of the data by contrast to the spreading in case of PCA evident from Fig. 3. Finally, it is worth noting the excellent performance of the Bayes classifier that outperforms the computationally complex SVM classifier, which requires fine-tuning of the slack variable through cross-validation. Thus, the performance of the Bayes classifier is remarkable taking its simplicity into account.

4. Conclusion

In conclusion, we demonstrated a robust and highly accurate (>98.6%) reference-free method for CW THz stand-off identification of materials based on a supervised ML algorithm and the Bayes classifier. Additionally, the method allowed us to reduce the dimensionality of the data by a factor larger than 160, *i.e.* from 649 discrete frequency components to just four features, which further reduced the computational requirements. Our results clearly shows that supervised algorithms are preferable over their unsupervised counterparts for THz material identification tasks Moreover, this methodology could readily be adapted for pulsed THz reflection spectra as well. In closing, the presented methodology could be of great importance for real-world remote sensing applications based on THz spectroscopy.

Funding

This work was supported by the Innovation Fund Denmark Grand Solutions program (grant no. IFD-7076-00017B).

CRediT authorship contribution statement

Mathias Hedegaard Kristensen: Writing – review & editing, Writing – original draft, Visualization, Validation, Software, Methodology, Investigation, Formal analysis. **Paweł Piotr Cielecki:** Writing – review & editing, Methodology, Investigation. **Esben Skovsen:** Writing – review & editing, Supervision, Resources, Funding acquisition, Conceptualization.

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 underlying the results presented in this paper are publicly available at DOI: 10.5281/zenodo.5079558.

References

[1] P. Jepsen, D. Cooke, M. Koch, Terahertz spectroscopy and imaging – Modern techniques and applications, Laser Photonics Rev. 5 (1) (2011) 124–166, http://dx.doi.org/10.1002/lpor.201000011.

[2] M. Naftaly, N. Vieweg, A. Deninger, Industrial applications of terahertz sensing: State of play, Sensors 19 (19) (2019) http://dx.doi.org/10.3390/s19194203.

[3] P. Cielecki, M. Hardenberg, G. Amariei, M.L. Henriksen, M. Hinge, P. Klarskov, Identification of black plastics with terahertz time-domain spectroscopy and machine learning, Sci. Rep. 13 (1) (2023) 22399, URL https://doi.org/10.1038/s41598-023-49765-z.

[4] N. Palka, THz reflection spectroscopy of explosives measured by time domain spectroscopy, Acta Phys. Pol. A 120 (4) (2011) 713–715, http://dx.doi.org/10.12693/APhysPolA.120.713.

[5] C. Cao, Z. Zhang, X. Zhao, T. Zhang, Terahertz spectroscopy and machine learning algorithm for non-destructive evaluation of protein conformation, Opt. Quantum Electron. 52 (4) (2020) 225, http://dx.doi.org/10.1007/s11082-020-02345-1.

[6] M.R. Nowak, R. Zdunek, E. Pliński, P. Świątek, M. Strzelecka, W. Malinka, S. Plińska, Recognition of pharmacological bi-heterocyclic compounds by using terahertz time domain spectroscopy and chemometrics, Sensors 19 (15) (2019) http://dx.doi.org/10.3390/s19153349.

[7] W. Liu, C. Liu, J. Yu, Y. Zhang, J. Li, Y. Chen, L. Zheng, Discrimination of geographical origin of extra virgin olive oils using terahertz spectroscopy combined with chemometrics, Food Chem. 251 (2018) 86–92, http://dx.doi.org/10.1016/j.foodchem.2018.01.081.

[8] H. Zhong, A. Redo-Sanchez, X.-C. Zhang, Identification and classification of chemicals using terahertz reflective spectroscopic focal-plane imaging system, Opt. Express 14 (20) (2006) 9130–9141, http://dx.doi.org/10.1364/OE.14.009130.

[9] J. Zhang, Y. Yang, X. Feng, H. Xu, J. Chen, Y. He, Identification of bacterial blight resistant rice seeds using terahertz imaging and hyperspectral imaging combined with convolutional neural network, Front. Plant Sci. 11 (2020) http://dx.doi.org/10.3389/fpls.2020.00821.

[10] K. Wang, D.-W. Sun, H. Pu, Emerging non-destructive terahertz spectroscopic imaging technique: Principle and applications in the agri-food industry, Trends Food Sci. Technol. 67 (2017) 93–105, http://dx.doi.org/10.1016/j.tifs.2017.06.001.

[11] W. Liu, C. Liu, F. Chen, J. Yang, L. Zheng, Discrimination of transgenic soybean seeds by terahertz spectroscopy, Sci. Rep. 6 (1) (2016) 35799.

[12] A.I. Knyazkova, A.V. Borisov, L.V. Spirina, Y.V. Kistenev, Paraffin-embedded prostate cancer tissue grading using terahertz spectroscopy and machine learning, J. Infrared Millim. Terahertz Waves 41 (9) (2020) 1089–1104.

[13] P.P. Cielecki, M.H. Kristensen, E. Skovsen, Analysis and classification of frequency-domain terahertz reflection spectra using supervised and unsupervised dimensionality reduction methods, J. Infrared Millim. Terahertz Waves 42 (9–10) (2021) 1005–1026, http://dx.doi.org/10.1007/s10762-021-00810-w.

[14] H. Park, J.H. Son, Machine learning techniques for thz imaging and time-domain spectroscopy, Sensors (Switzerland) 21 (4) (2021) 1–25, http://dx.doi.org/10.3390/s21041186.

[15] L. Zhang, H. Zhong, C. Deng, C. Zhang, Y. Zhao, Terahertz wave reference-free phase imaging for identification of explosives, Appl. Phys. Lett. 92 (9) (2008) 091117, URL https://doi.org/10.1063/1.2891082.

[16] H. Zhong, C. Zhang, L. Zhang, Y. Zhao, X.-C. Zhang, A phase feature extraction technique for terahertz reflection spectroscopy, Appl. Phys. Lett. 92 (22) (2008) 221106, URL https://doi.org/10.1063/1.2938055.

[17] Y. Hua, H. Zhang, Qualitative and quantitative detection of pesticides with terahertz time-domain spectroscopy, IEEE Trans. Microw. Theory Tech. 58 (7) (2010) 2064–2070, http://dx.doi.org/10.1109/TMTT.2010.2050184.

[18] P.P. Cielecki, M.H. Kristensen, E. Skovsen, Database of frequency-domain terahertz reflection spectra for the DETRIS project, 2021, http://dx.doi.org/10.5281/zenodo.5079558.

[19] D.W. Vogt, R. Leonhardt, High resolution terahertz spectroscopy of a whispering gallery mode bubble resonator using Hilbert analysis, Opt. Express 25 (14) (2017) 16860–16866, http://dx.doi.org/10.1364/OE.25.016860.

[20] D.W. Vogt, M. Erkintalo, R. Leonhardt, Coherent continuous wave terahertz spectroscopy using Hilbert transform, J. Infrared Millim. Terahertz Waves 40 (5) (2019) 524–534.

[21] D.-Y. Kong, X.-J. Wu, B. Wang, Y. Gao, J. Dai, L. Wang, C.-J. Ruan, J.-G. Miao, High resolution continuous wave terahertz spectroscopy on solid-state samples with coherent detection, Opt. Express 26 (14) (2018) 17964, http://dx.doi.org/10.1364/oe.26.017964.

[22] R. Bro, A.K. Smilde, Centering and scaling in component analysis, J. Chemometr. 17 (2003) 16–33, http://dx.doi.org/10.1002/cem.773.

[23] R. Bro, A.K. Smilde, Principal component analysis, Anal. Methods 6 (2014) 2812–2831, http://dx.doi.org/10.1039/C3AY41907J.

[24] I.T. Jolliffe, J. Cadima, Principal component analysis: a review and recent developments, Phil. Trans. R. Soc. A 374 (2016).

[25] A. Tharwat, T. Gaber, A. Ibrahim, A.E. Hassanien, Linear discriminant analysis: A detailed tutorial, AI Commun. 30 (2) (2017) 169, http://dx.doi.org/10.3233/AIC-170729.

[26] E. Alpaydin, Introduction to Machine Learning, third ed., Adaptive Computation and Machine Learning, MIT Press, 2014.