Peak correlation classifier (PCC) applied to FTIR spectra: a novel means of identifying toxic substances in mixtures

Robert French, Vesna Simic, Mathieu Thevenin

Peak correlation classifier (PCC) applied to FTIR spectra: a novel means of identifying toxic substances in mixtures

Robert M. French[1], Vesna Simic[2], Mathieu Thevenin[3]

1 CNRS LEAD Université de Bourgogne, F21000 Dijon, France, robert.french@u-bourgogne.fr

2 CEA, LIST, Bd 516, F91191 Gif-sur-Yvette, France

3 CEA-CNRS, Université Paris Saclay, IRAMIS SPEC, Bd 772, F91191 Gif-sur-Yvette, France

Abstract

Fourier transform infrared (FTIR) spectrometry is commonly used for the identification of reference substances (RSs) in solid, liquid, or gaseous mixtures. An expert is generally required to perform the analysis, which is a bottleneck in emergency situations. This study proposes a support vector machine (SVM)-based algorithm, the peak correlation classifier (PCC), designed to rapidly detect the presence of a specific threat or reference substance in a sample. While SVM has been used in various spectrographic contexts, it has rarely been used on FTIR spectra. The proposed algorithm discovers correlation similarities between the FTIR spectrum of the RS and the test sample and then uses SVM to determine whether or not the RS is present in the sample. The study also shows how the additive nature of FTIR spectra can be used to create 'synthetic' substances that significantly improve the detection capability and decision confidence of the SVM classifier.

1 Introduction

Emergency services are equipped with mobile laboratories to react as rapidly as possible in the event of a terrorist attack, an industrial disaster etc. These mobile labs are able to immediately analyse samples collected on-site. Available analytical tools include various types of spectroscopy equipment (Fourier transform infrared (FTIR), Raman, X-ray fluorescence etc.). A number of these tools have been in use for a number of years [1] for the identification of threats. The problem is that, in general, these analyses produce data that must be interpreted by an expert, often a time-consuming process creating a bottleneck for rapid, subsequent action. This problem is particularly acute for samples taken from a crime scene [2]. In addition, some substances can be dangerous for the security forces. For this reason, there is a pressing need to be able to identify, rapidly and accurately, prohibited, toxic or explosive substances contained in mixtures of other substances. Although data-analysis software usually comes packaged with the spectroscopy device used by the police laboratories, it is generally difficult to detect specific threats contained in mixtures, especially at low concentrations.

In the present paper, we focus on data obtained by FTIR spectroscopy, a widely used analytical tool that produces spectra of chemical substances based on the interaction of infrared (IR) light with the chemical bonds that compose the substance being tested. In FTIR spectra, frequency (or wave number) is indicated on the *x*-axis and IR absorbance or transmittance on the *y*-axis. This paper presents a novel algorithm designed to analyse IR spectra of sampled materials, typically those taken from a crime scene. A typical spectrum – in this case of 2,4-dinitrotoluene (DNT), a derivative substance of the explosive

trinitrotoluene – is shown in Fig. 1, top. All spectra used here were obtained using FTIR in the mid-IR frequency range (4000–400 cm^{-1}).

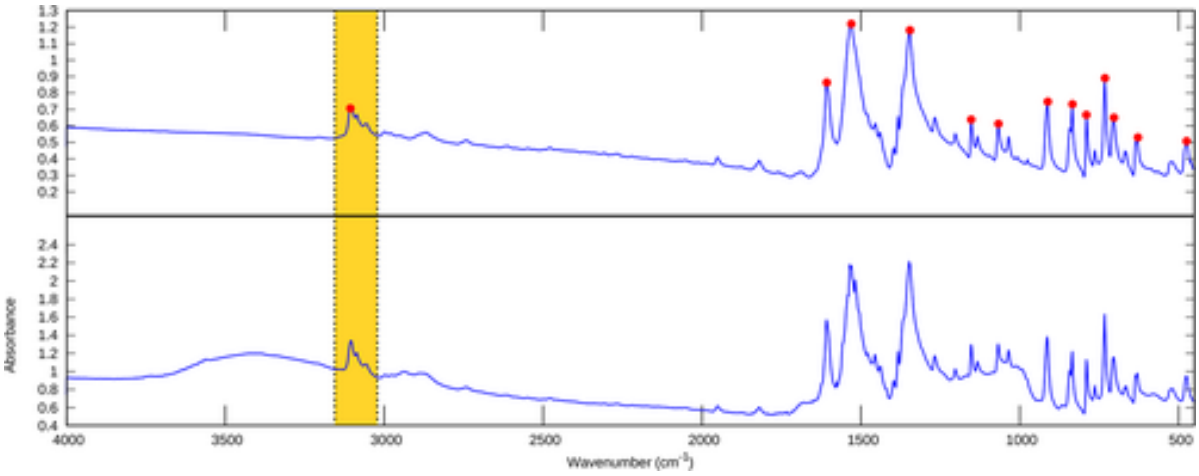

Fig. 1 *FTIR spectra for the RSS and the USS are aligned and the peak intervals from RSS are extended to USS. A peak interval is shown around a peak at 3050 cm^{-1}*

The proposed algorithm, called the peak correlation classifier (PCC), is based on the calculation of a vector of correlations between the reference substance spectrum (RSS) and the unknown substance spectrum (USS). This vector, which characterises the unknown substance (US) with respect to the reference substance (RS), is then input to a support vector machine (SVM) [3], a widely used and extremely powerful classifier [4-6].

Over the years, a significant body of related work has used SVM classifiers to discriminate data from various types of spectra and images. This work has been done, to a large extent, although not exclusively, in chemical and biomedical domains [7]. In the field of early detection of cancer, SVM classifiers have been used with mass spectroscopy data [8] and magnetic resonance (MR) spectroscopy data [8-10]. They have also been used with tandem mass spectroscopy for peptide studies or their identification [11, 12]. SVMs and other machine-learning techniques have been used to identify and classify substances in various fields, such as biology, textile science, and agronomics [8, 13-16]. However, to date, most of this body of work has involved using SVM to process data from mass spectroscopy, MR spectroscopy, or near IR spectroscopy and there has been relatively little use of SVM on data from FTIR spectroscopy [9], despite a number of encouraging results [17]. This is, arguably, related to the large size of FTIR data vectors.

The approach proposed here reduces the size of the input vectors by extracting a limited number of features from the spectrum, thereby allowing an SVM-based algorithm to classify FTIR data derived from both powder and gaseous mixtures containing dangerous or toxic substances. Crucially, the purely additive nature of FTIR spectra of non-interacting substances allowed us to create ‘synthetic’ (i.e. theoretical) spectra by combining various proportions of the RSS and the spectra from other substances. We then trained the SVM on synthetic spectra derived from a wide range of substances containing very low (sometimes even zero) to high concentrations of the RS.

There are four main contributions presented in this paper: (a) a new mean of representing the data input to the SVM by correlation-matching of peak intervals of the test and RSs; (b) the presentation of an algorithm that is able to determine whether or not a dangerous target

substance is present in the solid, liquid or gaseous mixtures being tested and to provide a measure of confidence in its determination; (c) the development of a novel training procedure for substance classification that relies on the additive nature of the spectra from non-interacting substances; and (d) finally, the algorithm can be trained off-line with a very large number of spectra, both synthetic and real, containing various concentrations of the RS, which means that no further training is necessary when testing new US for the presence or the absence of the RS.

The remainder of this paper is organised as follows: The basic PCC algorithm is described in Section 2. In Section 3, we present the experimental data used to test the PCC and describe the procedures used to acquire this data. In Section 4, we present the results of our tests of the PCC algorithm on our datasets, i.e. on both powder and gaseous mixtures, using both real and synthetic training data. We compare the performance of the PCC with two other widely used classification approaches, feedforward backpropagation neural networks [18] and linear discriminant analysis (LDA) [19]. We also demonstrate the significantly improved performance of the algorithm when it is trained on 6400 synthetic mixtures and tested on 5000 others. Finally, Section 5 concludes and introduces the perspectives for future research and applications.

2 Peak correlation classifier

The main idea of the proposed PCC approach consists of detecting the presence of a specific threat in a sample which is the mixture of various substances. It takes the form of an algorithm that must be executed for each RS that must be identified. As with most machine-learning algorithms, the PCC relies on the extraction and transformation of features from the data to form a database. Training sets are used to train a classifier on labelled data, which is used to test the new US. This training is performed each time a new RS is added to the database.

2.1 Reference substance

The RS produces an FTIR spectrum that is normalised to create the RSS. We then apply *x*-axis re-scaling (up-scaling and down-scaling are performed by B-spline interpolation) to obtain a 'scale' *s* of the RSS, denoted by RSS_s. A peak-finding algorithm first finds the *x*-value of each peak of the RSS_s. This is done by a 1D variant of the Shen and Castan algorithm [20]. We also filter the peaks obtained using various parameters, such as minimum peak separation, minimum peak height etc., to get an appropriate set of peaks. We obtain a list of the *x*-values corresponding to the peaks in the spectrum for a set of parameters given by ζ_s, for example: $\zeta_s = \{s, x_{min}, x_{max}, b, \ldots\}$, where the minimum and maximum *x*-values $x_{min}$and $x_{max}$of the spectrum are expressed in wave number and *b* is the number of bins. Let $\boldsymbol{P}$ be the complete list of the peaks for the whole set of parameters. It is formed by the union of all the peaks *P* for the *N* scales considered, $\boldsymbol{P} = \bigcup_{s=1}^{N} P_{\zeta_s}$

2.2 Peak correlation feature as substance representation

To perform comparisons between spectra, we chose to measure the Pearson correlation [21] of the segments of the RSS and USS inside each of the peak intervals (see Fig. 1), because we are interested in the *similarity of the shapes* of the two spectra within each peak interval. For this reason, measures of how much two distributions differ, such as Kullback–Leibler, are not

appropriate, simply because we are not interested in how much the spectra differ, but rather how closely their shapes match within the intervals around the peaks defining the RSS.

To obtain the peak correlation vector of values given to the SVM, we perform the following calculations: for each scale *s* we calculate a transform of the RSS using the function *t ( )* and of the USS, thereby obtaining t (RSS_s)and t (USS_s). In the PCC, this transform is the derivative. For each of the peak values given by the peak detection algorithm we calculate a similarity score S c ($P_{\zeta s}$) between t (RSS_s)and t (USS_s). For this, we chose the Pearson correlation [21], which is calculated between the $x = t(\mathrm{RSS}_s)_{P-k}^{P+k}$and the $y = t(\mathrm{USS}_s)_{P-k}^{P+k}$on an interval of *2.k+1* points around the peak *x*-coordinate, called a 'peak interval'. The advantage of calculating the Pearson correlation between these transforms is that this value is independent of the value of the points in the USS. It is a measure of how closely the shapes of the two spectra resemble each other over each of the peak intervals. The result of this calculation is a vector of correlation scores for each peak of a given scale. Thus, the representation of the substance is the union of all the *Sc* scores for a spectrum and can be written as a vector ***CC*** = *Sc* (***P***). We use this set (***CC***) of correlation values as a representation of the US with respect to the RS.

2.3 Choice of a classifier

The higher the correlation values between the RSS and a given USS, the more likely it is that the latter resembles the former, and the higher the probability that the US contains the RS. However, in some cases, in particular, when the RSS contains many narrowly spaced peaks, the average correlation does not provide a good estimate of the amount of the RS in the US. For this, we need a classifier to classify the peak-interval correlation vectors. Among the variety of standard classifiers, the most widely used for the classification of chemical substances are the LDA [22] and artificial neural networks (ANNs). The ANN provides good results for chemiometrics [23].

The SVM [24] is an extremely powerful classification algorithm that seems to have been largely overlooked in chemiometrics. SVM finds the mathematically optimal hyperplane separating the data to be classified. The distance of each of the classified data points from this separating hyperplane provides a measure of SVM's confidence in its classification. The further from the hyperplane, the more confident the SVM is of the correctness of its classification. Neither an ANN nor an LDA does this. A backpropagation network, arguably the most widely used ANN, stops changing its weights as soon as a separating boundary is found (i.e. when each of the outputs of the network fall below a certain error criterion), and this boundary may or may not be the optimal separating hyperplane.

We have tested the SVM in a number of other contexts [25, 26] and found it to be clearly superior in its classification performance to ANNs and LDA. However, one of the key factors in the use of the SVM is that it is more robust than the other standard classifiers tested. The robust nature of SVM classification has also been demonstrated elsewhere [27].

As we will discuss later in this article, we trained the SVM, not only on the spectra from real, known substances but also on artificially created ('synthetic') training spectra. The number of these synthetic training spectra used to train the SVM was inversely proportional to the concentration of the RSS that they contained. In other words, the SVM was trained on a proportionately greater number of spectra in which the concentration of the RSS was low.

This significantly increased the classifier's sensitivity to mixtures with very low concentrations of the RS.

2.4 Choice of the training set

We train an SVM to classify each of the spectra of the USs as either containing the RS or not. The distance from the SVM separating hyperplane provides a way of measuring the algorithm's confidence in the accuracy of its classification.

The standard approach for creating training data involves, first, deriving the RSS from pure RS. Then, a set of USS is obtained from FTIR spectrometry of real substances. Some of these substances are interferents, i.e. substances chosen for their close resemblance to the RS. However, using spectra derived from real samples involves considerable time and effort, and is subject to human error. Initially, we relied on this approach for training the SVM. One problem we encountered was that, for the SVM to correctly classify substances containing very small concentrations of the RS, it needed to have been trained on a large number of samples of this type. Also, these samples were not readily available. We, therefore, developed a technique of training the SVM on 'synthetic' substances.

It turns out that FTIR spectra have a singular property that allows these synthetic spectra to be created. The Beer–Lambert law allows the individual FTIR spectra of pure, non-interacting substances in a mixture to simply be added together to produce the FTIR spectrum of the mixture. Thus, for each RS, we generate mixtures with random concentrations of random pure substances. We then use the additivity of the individual spectra of the component substances to obtain the spectrum of the mixture. We relied on public or proprietary chemiometry databases to obtain the spectra of the pure substances making up these mixtures. Mixtures not containing the RS are tagged as negative while mixtures containing the RS are tagged as positive. The number of positive mixtures generated is a function of $1/\log(c_{RS})$, where c_{RS} denotes the concentration of the RS. These synthetic mixtures spectra are used to train the SVM classifier. The training is done once when an RS is added to the database for a given set of parameters ζ_s. Then each USS is classified against an RS to determine if the RS is present or not in the US. The distance from the SVM separating hyperplane provides a measure of the algorithm's confidence in the accuracy of its classification.

2.5 Parameterisation of the PCC algorithm

The parameters of the algorithm, most importantly the parameters used by the peak-finding routine, must be adjusted for each type of spectra under consideration. For example, the parameters for the spectra of gaseous mixtures and those of powder mixtures were very different. In the former, there are a great many peaks with small separations between them; in the latter, the peaks are considerably smaller, rarer, and more widely separated. For this reason, parameters, such as the minimum peak prominence, the minimum peak distance, the minimum peak height, the width of the correlation intervals, and the amount of smoothing of the spectra, need to be adjusted using a sample of known mixtures in which some contain the substance-to-be-detected and others do not.

We performed a parameter space exploration by generating random parameter values to cover the space and then selecting the parameters that provide the best results, i.e. minimum rate of false positives and negatives. Fig. 2 clearly shows that, for mixtures containing DNT, a peak interval of $k = 8$ or 10 provides the best results and is largely independent of the value of the

peak separation. These parameters could also be set by using machine-learning techniques, such as genetic algorithms [28] or by hill-climbing etc.

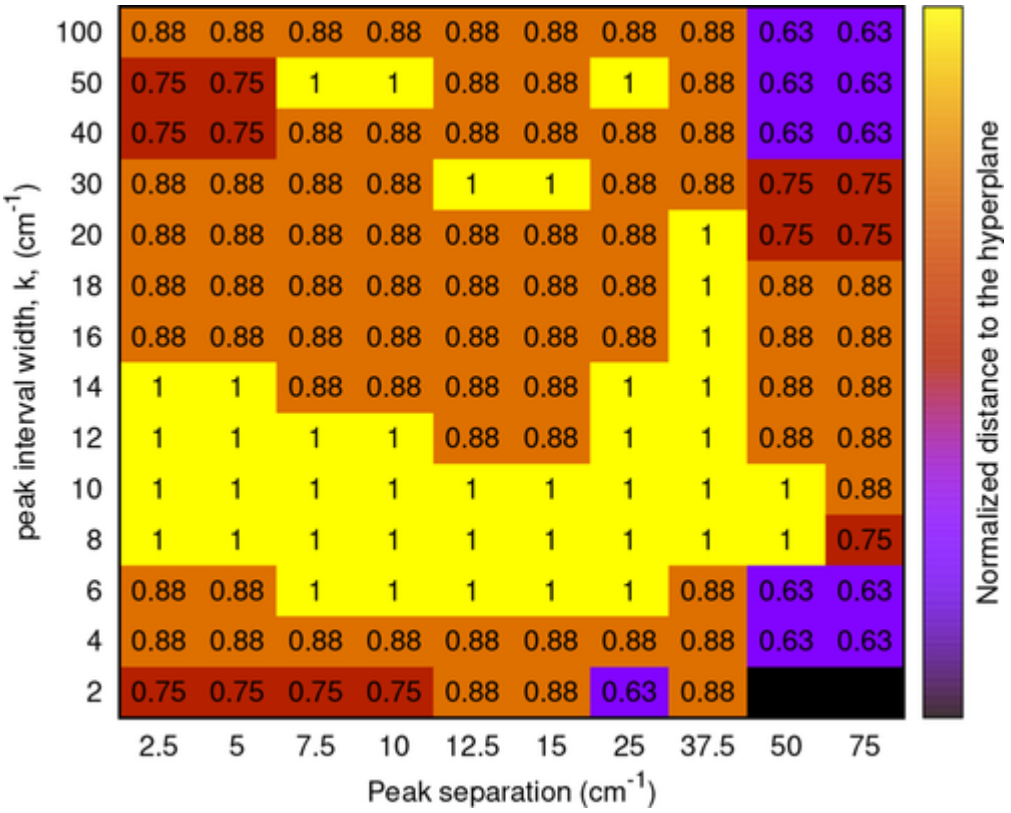

Fig. 2 *Fraction of correct classification of unidentified substances with respect to the width of the peak interval (k parameter, expressed here in wave number) and the minimal separation between any two peaks. X and Y units are expressed in wave number, the spectrum resolution is 0.5 cm^{-1} per sample*

3 Data sets used to test the PCC

This section describes how the real and the synthetic data sets used to test how the PCC algorithm performs were obtained. There are solid samples, gaseous samples, and synthetic data.

3.1 Solid samples

The initial samples tested consisted of 40 pellets containing various amounts of DNT. This nitroaromatic was mixed with various percentages of substances with structures either similar to DNT (toluene, musk ketone, 4-nitrophenol, nitrobenzene, polystyrene, 2-hydroxybenzoic acid, and hydroquinone) or significantly different from that of DNT (flour, sucrose, acetone, sodium bicarbonate ($NaHCO_3$)).

These mixtures were used to test both the sensitivity and specificity of the PCC algorithm, i.e. to determine its ability to detect the presence of DNT when it was, in fact, present in a mixture (sensitivity) and to determine that it was absent when it was not present in a sample (specificity). The amount of DNT in these test mixtures ranged from very low (2%) to up to 71% DNT by weight. The algorithm was also tested on a number of pure substances not containing DNT – namely, pure toluene, salicylic acid, $NaHCO_3$, flour, hydroquinine, musk

ketone, and nitrophenol, in addition, we also conducted a 'blind test' of the PCC algorithm on a set of eight real powder mixtures for which no prior information was given as to whether or not a given substance contained DNT, see Section 4.4.

3.2 Gaseous samples

The Laboratoire Central de la Préfecture de Police de Paris (LCPP) prepared a number of gaseous samples to be tested using the PCC algorithm. These samples consisted of ten mixtures containing various amounts of ammonia (NH_3) and air. Two of the samples consisted of air only, i.e. they contained no NH_3. In the samples containing NH_3, the amount of NH_3 was decreased by half for each successive sample. The first sample contained ~250 ppm NH_3, the second, 125 ppm, the third, 62.5 ppm etc. The final sample contained <1 ppm of NH_3. The LCPP provided us with the mid-range FTIR spectra corresponding to all of these samples. The NH_3 reference spectrum (NH_3 543 ppm, 750 Torr, 1 M) was taken from a reference library provided by Thermo Scientific. The LCPP also provided a number of spectra from samples of NO + air and NO_2 + air. These were also tested using the PCC algorithm. The FTIR spectra of gaseous samples were recorded on a Thermo Scientific model Nicolet iZ10 FTIR spectrophotometer equipped with a 2 m optical path gas cell in the range of 550 and 4000 cm^{-1} with a resolution of 0.5 cm^{-1}. The gas mixtures were prepared with an AlyTech model Gasmix/Liqmix LG4CA gas diluter and sampled in 3 l SKC Tedlar gas sampling bags. To further demonstrate the sensitivity of the PCC algorithm, we also modified the NH_3 spectrum by eliminating parts of it that had initially made it easy for the algorithm to distinguish it from the spectrum for air, as illustrated by Fig. 3.

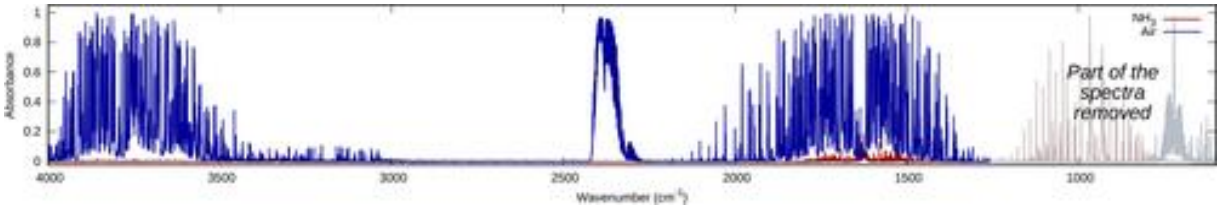

Fig. 3 *Spectra for NH_3 (red) and a mixture of air and 250 ppm NH_3 (blue) to challenge the algorithm, the right part of the spectra is truncated since the NH_3-related peaks are mostly present under 1300 cm^{-1}*

3.3 Synthetic mixtures

It is a delicate and time-consuming process to create spectra from real substances. In addition, this process is a potentially dangerous one, depending on the substance being processed (e.g. Sarin nerve gas etc.). The problem is that, to appropriately train an SVM, we need a relatively large database of mixtures containing various amounts of the RS. We, therefore, generated synthetic spectra using reference spectra of pure threat or non-threat substances taken from various chemiometry databases. The 'theoretical' mixtures were obtained by applying the Beer–Lambert law, which says the spectrum of a composite of non-interacting substances is a linear combination of the individual spectra of the pure substances making up the composite. For each RS, we generated spectra derived from a random concentration of the RS and the spectra of random concentrations of other substances. Mixtures not containing the RS are tagged as negative while mixtures containing the RS were tagged as positive. The synthetic spectra created in this way were then used to train the classifier.

As shown in Figs. 4 and 5, training on synthetic data significantly improved the performance of the PCC algorithm. This was expected since a large number of synthetic spectra, tagged as

either containing DNT or not, can be generated and used to train the SVM. This allows the algorithm to be trained on far more spectra than had we been using spectra from real samples alone.

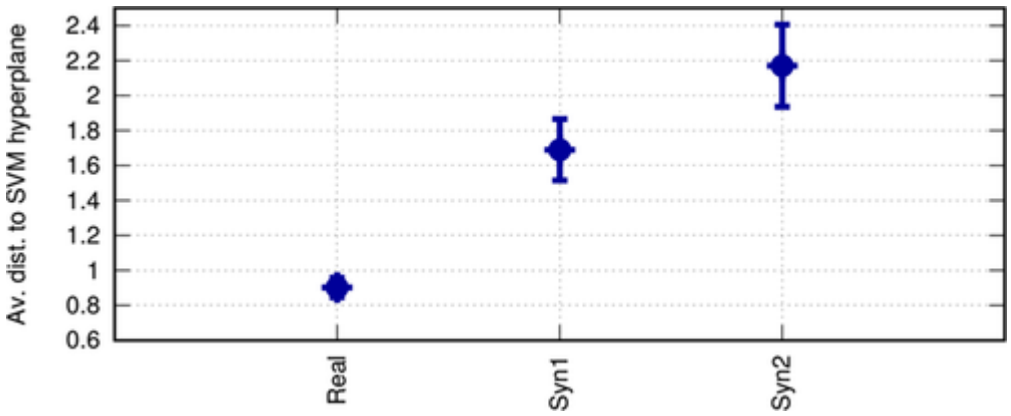

Fig. 4 *Comparative average distances to the SVM hyperplane using only real training data (with a LOOCV protocol) and with two groups of 120 synthetic mixtures with different proportions of low-DNT concentration spectra. (Standard error of the mean error bars, $F(2, 54) = 24.9$, $p < 0.0001$)*

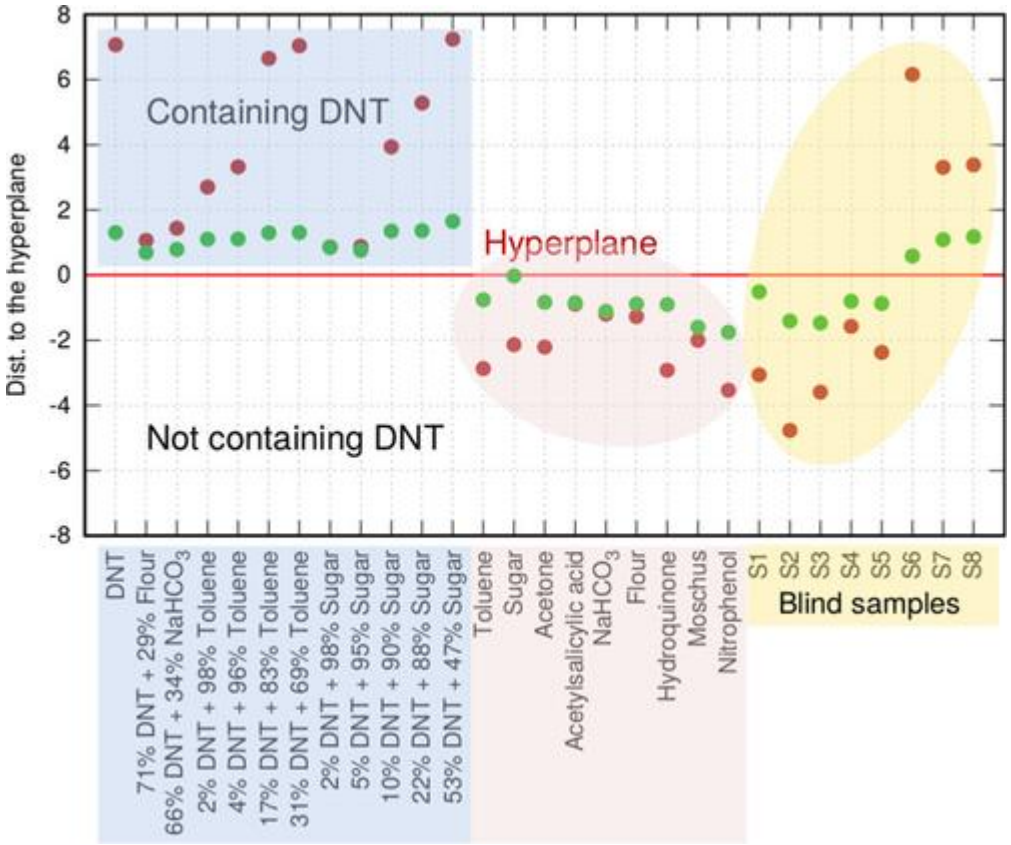

Fig. 5 *Classification of 29 real substances including eight blind USs named S1–S8 using two training modes: the green dots show the results obtained with a leave-one-out classification on real spectra while the red dots show the results obtained using 600 synthetic spectra generated from various data. In both cases, no false identification is reported, but the distance to the hyperplane is significantly improved for synthetic training, especially on samples with >5% of DNT*

4 Implementation, training, and results

After giving details of the implementation, this section then introduces the impact of synthetic training and presents the results obtained with the PCC under different conditions.

4.1 Implementation of the algorithm

The PCC algorithm was first developed and tested using Matlab. Subsequently, a 'multi-scale' C++ implementation was then developed, in which the raw spectral data were reproduced at a number of different resolutions (scales), all of which were concatenated, and these extended vectors analysed by the Dlib [29] SVM using a radial basis function (RBF) kernel. The classification decision function also returns an estimate of the probability that a given sample contains the RS, thereby, providing a good measure of the quality of the classification.

Spectra interpolation was performed using a B-spline interpolation. The probabilistic decision function returns an estimate of the probability that a given sample is in a positive class (i.e. contains the RS). The distance to the hyperplane serves as the basis of this probability estimate and provides an excellent measure of the algorithm's classification confidence.

4.2 Powder mixtures with various percentages of DNT

This subsection explains how we trained the PCC algorithm on solid substances obtained from real samples and on synthetic data.

4.2.1 Training on real data

The initial tests of the PCC algorithm were done on 20 powder mixtures containing various percentages of DNT, including samples with no DNT, but whose chemical structure in some cases closely resembled that of DNT.

We also conducted a 'blind' test of the algorithm for powder mixtures in which we were given eight samples and were not told whether or not they contained DNT. The algorithm had to determine whether or not DNT was present in each of the samples.

For the two-way classification of the data (i.e. contains DNT/does not contain DNT) an SVM, using an RBF kernel with a gamma of 0.0025 and $C = 125$ was used. The input to the SVM for each substance was its 13-value representation, where each value was the correlation of the USS with the DNT spectrum, over each of its 13 peak intervals. The training set consisted only of real FTIR spectra obtained in the laboratory. Using a standard leave-one-out cross-validation (LOOCV) methodology, the SVM learned to correctly classify all of the substances into the two DNT/no-DNT categories.

In addition, by calculating the distances from each ***CC*** corresponding to a particular substance to the SVM hyperplane, the algorithm also gives a confidence rating for each classification. The further from the SVM hyperplane, the more confident the SVM is of its classification of a particular substance. As expected, its confidence in its classification of 2% DNT + 98% toluene (very close to the SVM hyperplane) is considerably poorer than its classification of 31% DNT + 69% toluene.

4.2.2 Training on synthetic data

As discussed in Section 3.3, the additive nature of FTIR spectra allowed us to create synthetic substances by combining the reference spectra for various substances. In this way, we were able to create as many synthetic spectra as we wanted to train the SVM. Since it is hardest for the SVM to correctly classify substances containing very low concentrations of DNT, we created proportionately more synthetic spectra for composite substances containing only small amounts of DNT.

Training the SVM on proportionately more spectra corresponding to synthetic substances containing small amounts of DNT does, indeed, improve the performance of the algorithm. This can be seen in Fig. 4 by its classification confidence for the substance containing 2% DNT + 98% toluene. The training on synthetic spectra, where a proportionately greater number corresponded to low-DNT concentration substances, improved the algorithm's confidence in its classification. For the synthetic training, the SVM was trained using a gamma of 0.00125 and $C = 125$. We were initially able to produce a nearly three-fold improvement in confidence as seen in Fig. 4, by varying the number and distribution of low-DNT concentration training spectra. In addition, Fig. 5 shows how training on synthetic data significantly increases the SVM's confidence in its classifications.

4.3 Gaseous mixtures

We also tested the PCC on gaseous mixtures containing NO, NO_2, and NH_3. Unlike the spectrum of DNT, the spectra of NO, NO_2 and NH_3 contain a great many, closely spaced peaks. For this reason, it was necessary to modify the peak-interval width parameter, as well as the peak-separation and peak-height parameters of the peak-finding routine for these reference gases.

4.3.1 Gaseous mixtures containing NO and NO_2

We first tested the PCC algorithm on gaseous mixtures of NO and air, as well as NO_2 and air. The PCC was able to distinguish between pure air and NO + air at concentrations of 20 and 25 ppm and between pure air and NO_2 + air at concentrations of 10, 30, and 400 ppm as shown in Fig. 6.

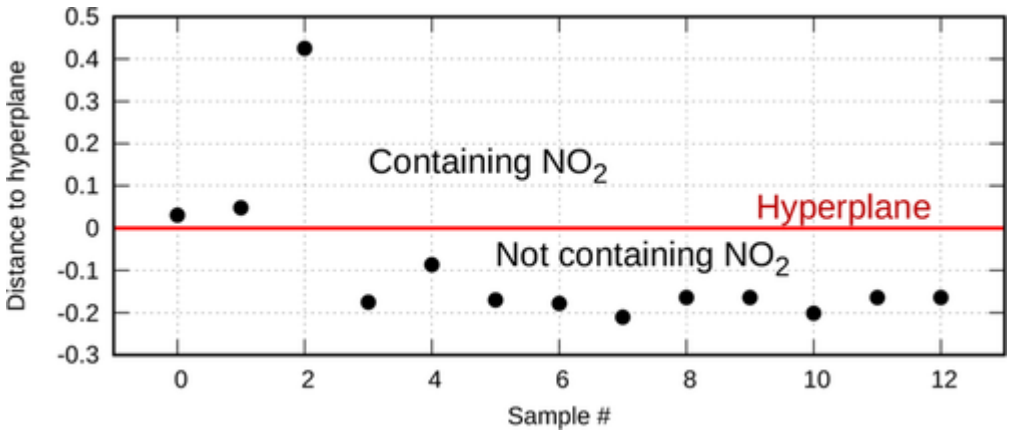

Fig. 6 *Plot of the distance to the hyperplane of the classification of NO_2 samples. Under the hyperplane are the samples not containing NO_2 and above are substances containing NO_2*

As with the tests involving DNT, we began with reference spectra for both NO and NO_2 and proceeded as we did for the powder mixtures. We used peak-interval correlation vectors as input to an SVM with an RBF kernel with a sigma of 10. The algorithm clearly separated the ten samples of air from the two samples of NO, as shown in Fig. 7.

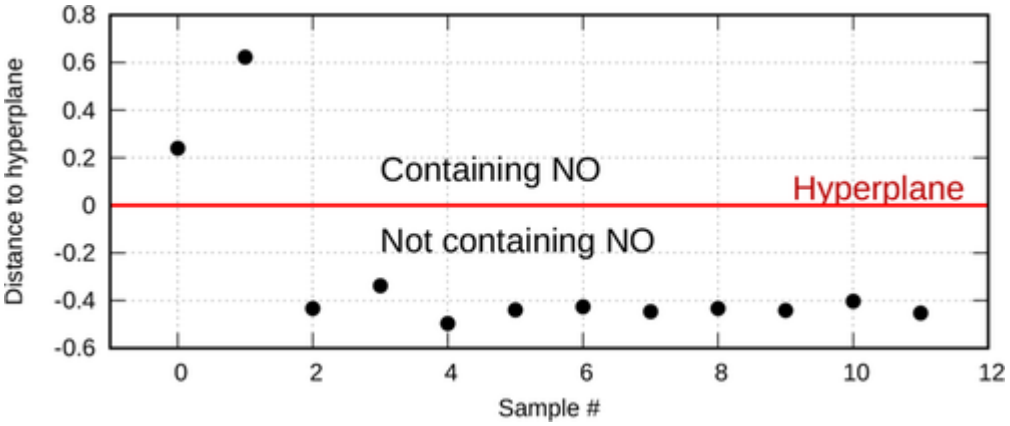

Fig. 7 *Distance to the hyperplane for the classification of NO samples. Under the hyperplane are the samples not containing NO and above are substances containing NO*

4.3.2 Gaseous mixtures with various percentages of NH_3

We then conducted an analysis of gaseous mixtures containing NH_3. Since the peaks of the spectrum of NH_3, such as those of NO and NO_2, are numerous and very close together, we had to modify the peak-separation and peak-height parameters of our peak-finding routine, reducing the peak-interval width to 5. We also smoothed the spectra over a small span of 25 values to eliminate spurious mini-peaks around the main peaks. The peak-finder identified 19 peaks in the pure NH_3 spectrum, as illustrated in red in Fig. 8.

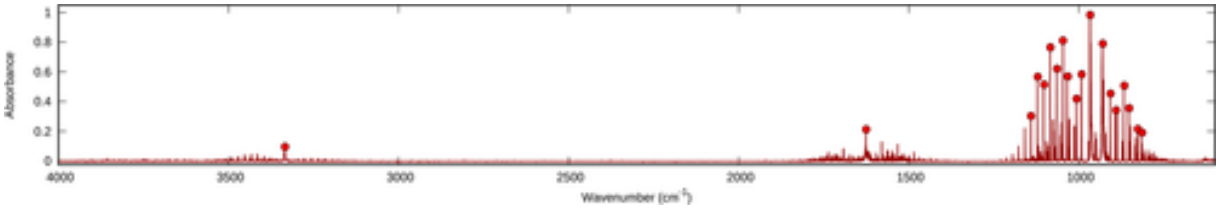

Fig. 8 *Nineteen peaks identified in the pure NH_3 spectrum*

The LCPP provided us with ten samples on which to test the PCC algorithm. Two of these samples consisted of air containing no NH_3. The other samples successively contained half as much NH_3 as the previous sample. Thus, dilution 2 (D2) contained a mixture of 125 ppm of NH_3 and air. D3 was a 62.5 ppm NH_3/air mixture, and so on. The final NH_3-containing sample, D9, contained slightly <1 ppm NH_3.

Fig. 9 shows the spectrum for D4, i.e. a sample containing 62.5 ppm NH_3. We used the full 19-value peak-interval correlation vector as the representation of each of the gaseous mixtures tested. These vectors were given as input to an SVM with an RBF kernel using a gamma of 10 (smaller than that used for the DNT samples because of the proximity of the peaks). The results are shown in Fig. 10. All ten samples were correctly classified and the distances from the SVM separating hyperplane indicate that the algorithm is very confident in its classifications.

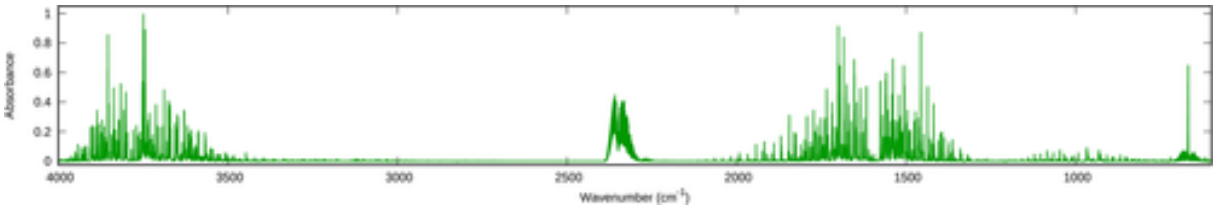

Fig. 9 *Spectrum for a 62.5 ppm NH_3 + air mixture*

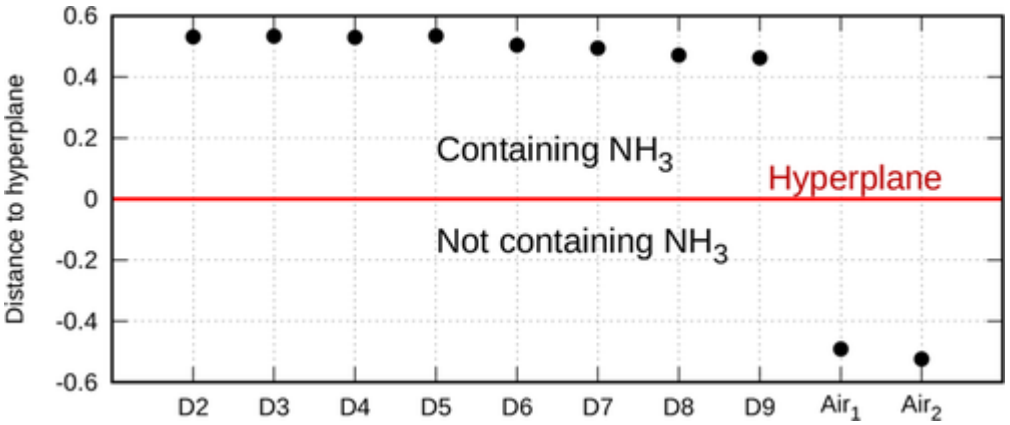

Fig. 10 *Separation of the samples containing NH_3 and pure air by an SVM using the full correlation vector representation of each sample. The figure shows the distance of the representations from the SVM hyperplane. Under the hyperplane are the gaseous mixtures not containing NH_3; above are the mixtures containing various dilutions of NH_3*

4.3.3 Estimating the percentage of NH_3 in the sample

One question left unanswered by the PCC algorithm is what percentage of the mixture consists of the reference substance, in this case, NH_3. Once the PCC algorithm has determined that NH_3 is present in a given mixture, it can predict approximately how much NH_3 is present in the sample. Consider the two overlapping spectra of NH_3 and air shown in Fig. 3. The spectrum for the mixture of air is shown in blue. The spectrum for pure NH_3 is shown in red. We only consider the parts of these spectra that are contained in the union of the intervals about the defining peaks of NH_3. We refer to this union of peak intervals as the NH_3 critical zone. The Beer–Lambert spectra-additivity assumption for mixtures of non-interacting substances allows us to assume that in the NH_3 critical zone for a mixture of NH_3 and air, the amplitude of the spectrum for NH_3 + air will be higher than that of NH_3 alone. Therefore, for each mixture of NH_3 + air, we calculate the difference between the values of its spectrum and the spectrum for pure NH_3 over the NH_3 critical zone, for each of the mixtures, starting with the 0.5/0.5 NH_3/air mixture (D2) and going down to D9. If we plot these differences with respect to the pure NH_3, it can be seen that the curve is almost perfectly logarithmic, which is in agreement with the Beer–Lambert law of optical absorption [30]. Since we know that the concentration of each successive NH_3 dilution was half the previous dilution, we know the NH_3 concentration for each dilution. As a result, we can plot (Fig. 11) the logarithm of the spectrum differences with respect to the pure NH_3 spectrum against real NH_3 concentration levels and we obtain what is, for all intents and purposes, a linear relationship ($R^2 = 0.99$). Interpolating between each of these values, one can determine the concentration of NH_3 in any sample known to contain NH_3.

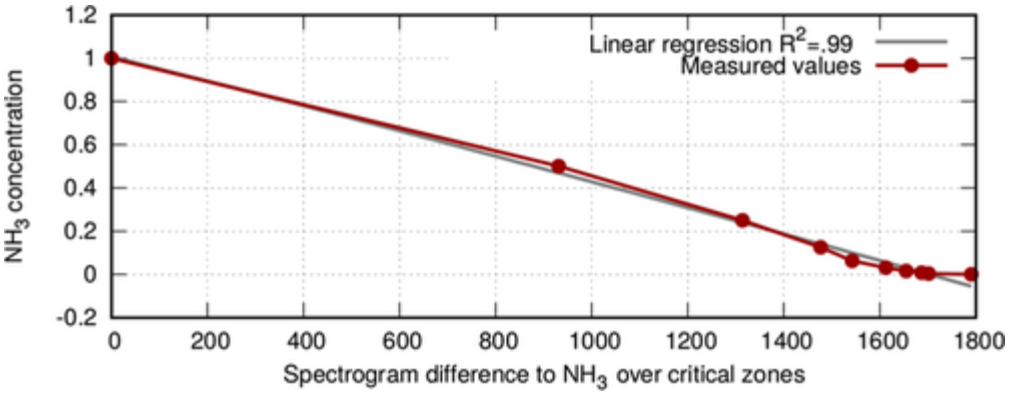

Fig. 11 *Plot of the computed relative concentration of NH_3 in the sample being examined*

4.3.4 Testing the limits of the sensitivity of PCC for gaseous samples

A reasonable criticism of the samples compared, i.e. air and NH_3 + air, is that there is a region of the two spectra between 0 and 1300 cm^{-1} where there is a considerable spectral contribution from NH_3 and very little contribution from the air spectrum. Thus, the PCC might only be able to detect NH_3 in an NH_3 + air mixture because in this critical region the two spectra (i.e. air and NH_3 + air), are very different.

We, therefore, decided to make the NH_3-detection and concentration-estimation task significantly harder. This allowed us to test the sensitivity of the PCC algorithm for pairs of spectra with considerable overlap and numerous peaks with little separation between them. We, therefore, truncated the NH_3 and the NH_3 + air spectra from 0 to 1300 cm^{-1}, thereby removing the segments of their respective spectra where they differ most noticeably.

Thus, for a 0.5/0.5 mixture of pure air and NH_3 we obtain the spectra shown in Fig. 3, with the distinguishing values of both spectra beyond 1300 cm^{-1}. We compared the truncated spectra of a total of ten different samples of pure air with the spectra of the eight D2–D9 samples previously tested. When the peak-correlation vectors are given to an SVM classifier, even with these severely truncated spectra, Fig. 12 shows that the PCC algorithm is still able to distinguish pure air from air containing NH_3. As before, we computed the approximate percentage of NH_3 in the samples. To do this, we found the same logarithmic curve as for the untruncated samples. Also, even though the fit is, unsurprisingly, not as good as for the original samples with their full spectra, we observe that the logarithmic fit to data still has an $\boldsymbol{R^2}$ of 0.87. When this logarithmic curve is transformed into a linear curve, as was done in Fig. 11, the fit to linearity produces an $\boldsymbol{R^2}$ of 0.96.

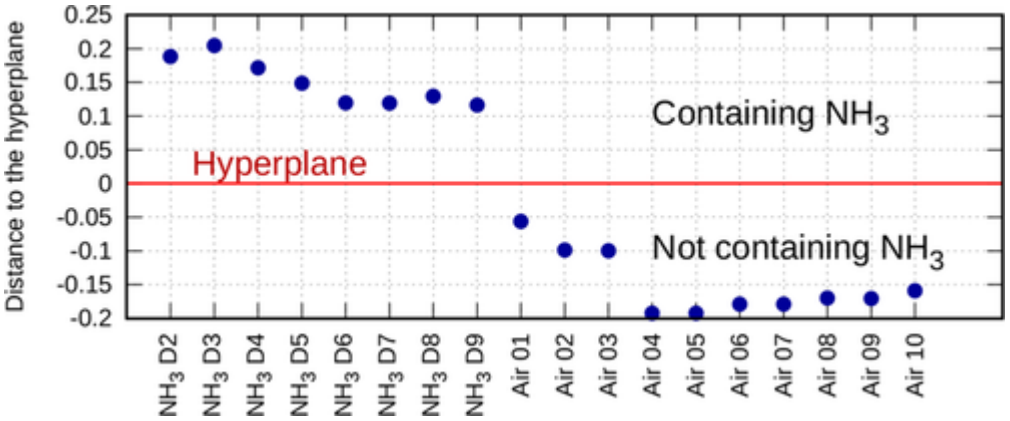

Fig. 12 *PCC discriminates pure air (above the SVM hyperplane) and mixtures of air + NH_3 (below the hyperplane) even when severely truncated spectra are used*

4.4 Blind classification

The final test of the PCC algorithm was done with eight substances that had been prepared by one of the authors, but the presence or absence of DNT in each substance was not revealed to the other two authors. The PCC algorithm was trained on 12 mixtures known to contain DNT and nine mixtures known not to contain DNT, they are indicated in Fig. 5. We then tested the PCC on the unknown samples. Some of these samples contained DNT, sometimes in very small quantities, others had no DNT, and still, others contained substances whose chemical structure very closely resembled that of DNT.

4.5 Comparison with other classifiers

We also compared different classifiers with the SVM. We based our comparison on the classifiers used for chemiometry and described in [23]: feedforward, backpropagation networks [18] with three different hidden-layer sizes the feedforward and the LDA algorithm [19].

For the PCC algorithm, the SVM generally performs better, in terms of both classification accuracy and robustness [27], than the other classifiers.

For the substances tested in the blind-classification test, SVM is considerably more robust than the other classifiers tested. The robust nature of SVM classification has also been demonstrated elsewhere [27] and we also observed this in the present context.

We considered a wide range of peak-interval widths (from 2 to 100 cm^{-1}) and reference-peak separations (from 2.5 to 75 cm^{-1}). It turns out that a peak-interval width of either 8 or 10 cm^{-1} gives the best classification performance, shown in Fig. 2 (the area in yellow), which indicates the perfect classification of the USs.

While it is true that the SVM's best performance over all peak-separation distances and peak-interval widths is between 1 and 7% better than the best performance of any of the other classifiers tested, this difference, in itself, is not large enough to justify the use of an SVM classifier. The crucial difference, as has been observed in other contexts, is the robustness of this algorithm. For a given reference substance, it is not always obvious what values to use for peak-separation and peak-interval widths. However, this choice is far less important for an SVM classifier, compared to the three different backpropagation networks and LDA. So, for example, consider the number of correct classifications for all peak-intervals averaged over all peak-separation distances. The variability of these values is between 2.7 and 7.7 times higher for the other classifiers compared to SVM. Similarly, if we look at the number of times each algorithm correctly classified all of the USs over all peak-separation values and all peak-interval widths (140 values), we find that the SVM has between 1.4 and 2.7 times as many correct classifications as the other algorithms.

4.6 Analysis of the results for 5000 synthetic samples

We tested 5000 mixtures of synthetic spectra containing widely varying concentrations of the reference substance, including mixtures in which the reference substance was absent. The training set consisted of 6400 mixtures that differed from those in the test set. They were

generated by randomly mixing real spectra from 3-nitrophenol, acetone, $NaHCO_3$, cane sugar, wheat flour, musk ketone, olive oil mixed with sucrose. Two-thirds of the mixtures contained various concentrations of the reference substance and at least one other substance, the other third consisted of mixtures that did not contain the reference substance. For both the test and training sets, two classes were generated; (a) one containing the reference substance at various concentrations ranging from 3 to 100% and (b) one containing other substances that did not include the reference substance. The substance to be detected was DNT. Fig. 13 shows the results obtained for 5000 test spectra constructed as above. There were no false positives and only a single false negative, which contained a low concentration of 24-DNT (4.8%) and musk ketone (95.2%). It is shown in Fig. 13 by the red dot (low-DNT concentration, close to the hyperplane, meaning low confidence in its (incorrect) classification). Musk ketone was used for its chemical proximity to the DNT, to test the limits of the PCC algorithm. Unsurprisingly, the distance to the SVM hyperplane varies from one mixture to another, depending on the concentration of DNT in the mixture (Fig. 13). This is clearly visible for the positive class (violet) of spectra containing DNT. All elements containing no DNT (green dots) are correctly classified. For the spectra of substances containing DNT, even in low concentrations, the distances to the hyperplane is generally quite high.

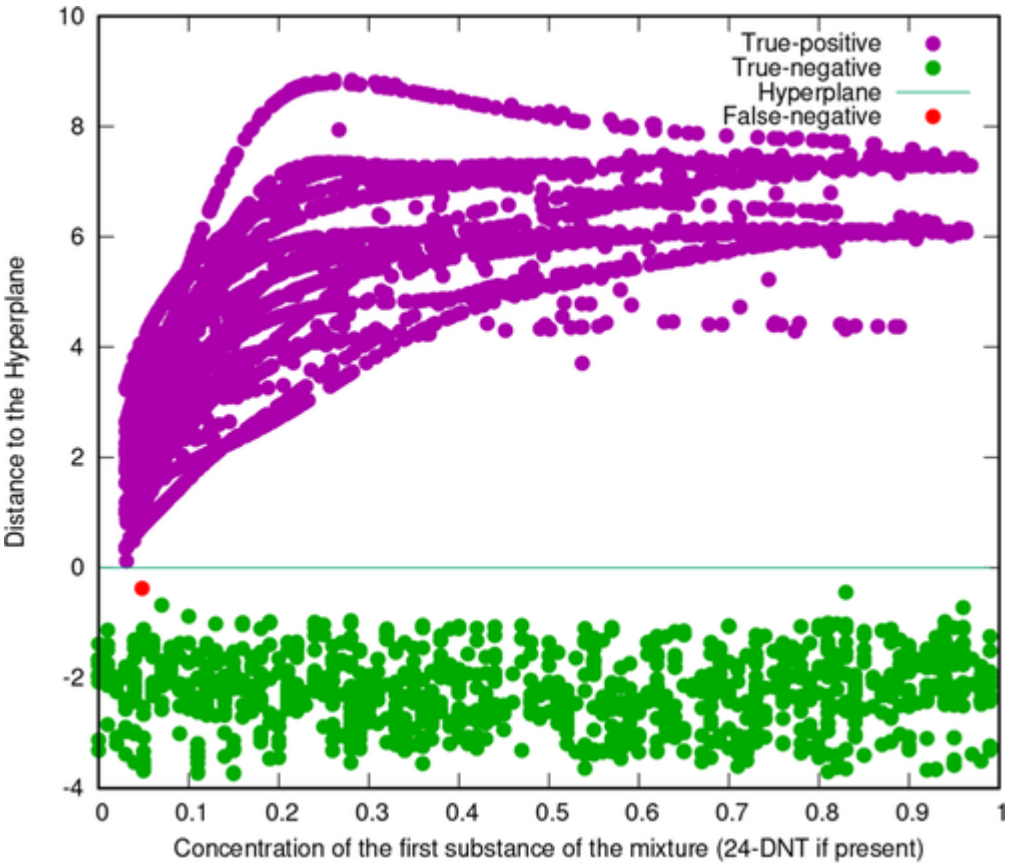

Fig. 13 *Plot of the distances to the hyperplane and the classifications for spectra from substances containing different concentrations of DNT or no DNT. The points below the hyperplane (negative distance) are classified as not containing DNT, the points above the hyperplane are classified as containing DNT*

The average time required to classify a spectrum is 3.20 ms (including file accesses and transfers) on an intel i9 i9-8950HK core, running at 2.9 GHz. The peak memory usage was ~42 MB.

5 Conclusions

We have presented a simple algorithm, the PCC, to determine whether or not certain mixtures contain a particular target substance. This algorithm is computationally efficient because the training of the SVM can be done off-line. We have shown that this method can be used effectively to detect the presence or absence of target substances in powder mixtures as well as in gaseous mixtures. It can also be used to determine the percentage of the target substance in a given mixture. Finally, because spectral data from non-interacting substances are additive, synthetic spectra are created from reference spectra and used to train the SVM. In this way, the performance of the algorithm on mixtures containing low concentrations of the target substance can be significantly improved. Further work will focus on testing the algorithm on more substances and providing a user-friendly graphical interface.

6 Acknowledgments

This work was funded by the French joint-ministerial CBRN-E Program. The authors would like to thank Dr Thanh-Toan Truong, Dr Nadine Fourrier, and Thomas Pianelli of the 'Laboratoire Central de la Préfecture de Police de Paris' (LCPP) for their support of this project, and in particular, for providing the spectra for gaseous mixtures containing NH_3, NO, and NO_2. The authors are grateful to Paul Malfrait for helping with the analysis of the results and to Nicola Martin for proof-reading the final revision.

7 References

[1] Virkler, K., Lednev, I.K.: 'Analysis of body fluids for forensic purposes: from laboratory testing to non-destructive rapid confirmatory identification at a crime scene', Forensic Sci. Int., 2009, 188, (1), pp. 1–17

[2] Kumar, R., Sharma, V.: 'Chemometrics in forensic science', TRAC Trends Anal. Chem., 2018, 105, pp. 191–201

[3] Vapnik, V.N.: 'Statistical learning theory' (Wiley, New York, USA, 1998)

[4] Luts, J., Ojeda, F., de Plas, R.V., et al.: 'A tutorial on support vector machine-based methods for classification problems in chemometrics', Anal. Chim. Acta, 2010, 665, (2), pp. 129–145

[5] Noble, W.S.: 'What is a support vector machine?', Nat. Biotechnol., 2006, 24, pp. 1–13

[6] Zhang, X., Lu, X., Shi, Q., et al.: 'Recursive SVM feature selection and sample classification for mass-spectrometry and microarray data', BMC Bioinf., 2006, 7, (1), p. 197

[7] de Boves Harrington, P.: 'Support vector machine classification trees based on fuzzy entropy of classification', Anal. Chim. Acta, 2017, 954, pp. 14–21

[8] Wu, J., Ji, Y., Zhao, L., et al.: 'A mass spectrometric analysis method based on PPCA and SVM for early detection of ovarian cancer', Comput. Math. Methods Med., 2016, 2016, pp. 6169249:1–6169249:6

[9] Yi, L., Dong, N., Yun, Y., et al.: 'Chemometric methods in data processing of mass spectrometry-based metabolomics: a review', Anal. Chim. Acta, 2016, 914, pp. 17–34

[10] Parfait, S., Walker, P.M., Créhange, G., et al.: 'Classification of prostate magnetic resonance spectra using support vector machine', Biomed. Signal Proc. Control, 2012, 7, (5), pp. 499–508

[11] Webb-Robertson, B.J.M.: 'Support vector machines for improved peptide identification from tandem mass spectrometry database search', in Lipton, M.S., Paša-Tolic, L. (eds.): Methods in Molecular Biology (Humana Press, Totowa, NJ, 2009), pp. 453–460

[12] Zou, A.M., Ding, J., Shi, J.H., et al.: 'Charge state determination of peptidetandem mass spectra using support vector machine (svm)'. 2008 8th IEEE Int. Conf. on BioInformatics and BioEngineering, Athens, Greece, 2008, pp. 1–6

[13] Zhang, X., Lin, T., Xu, J., et al.: 'Deepspectra: an end-to-end deep learning approach for quantitative spectral analysis', Anal. Chim. Acta, 2019, 1058, pp. 48–57

[14] Pierna, J.A.F., Baeten, V., Renier, A.M., et al.: 'Combination of support vector machines (SVM) and near-infrared (NIR) imaging spectroscopy for the detection of meat and bone meal (MBM) in compound feeds', J. Chemom., 2004, 18, (7–8), pp. 341–349

[15] Eylenbosch, D., Bodson, B., Baeten, V., et al.: 'NIR hyperspectral imaging spectroscopy and chemometrics for the discrimination of roots and crop residues extracted from soil samples', J. Chemometr., 2018, 32, (1), p. e2982

[16] Langeron, Y., Doussot, M., Hewson, D.J., et al.: 'Classifying NIR spectra of textile products with kernel methods', Eng. Appl. Artif. Intell., 2007, 20, (3), pp. 415–427

[17] Allegrini, F., Pierna, J.A.F., Fragoso, W.D., et al.: 'Regression models based on new local strategies for near infrared spectroscopic data', Anal. Chim. Acta, 2016, 933, pp. 50–58

[18] Rumelhart, D.E., McClelland, J.L., PDP Research Group (eds.): 'Parallel distributed processing: explorations in the microstructure of cognition, vol. 1: foundations' (MIT Press, Cambridge, MA, USA, 1986)

[19] Fisher, R.A.: 'The use of multiple measurements in taxonomic problems', Ann. Eugenics, 1936, 7, pp. 179–188

[20] Shen, J., Castan, S.: 'An optimal linear operator for step edge detection', CVGIP, Graph. Models Image Process., 1992, 54, (2), pp. 112–133

[21] Pearson, K.: 'Note on regression and inheritance in the case of two parents', Proc. R. Soc. Lond. Ser. I, 1895, 58, pp. 240–242

[22] Barron, L., Gilchrist, E.: 'Ion chromatography-mass spectrometry: a review of recent technologies and applications in forensic and environmental explosives analysis', Anal. Chim. Acta, 2014, 806, pp. 27–54

[23] Lavine, B.K., Blank, T.R.: 'Feed-forward neural networks', in Brown, S., Tauler, R., Walczak, B. (Eds.): 'Comprehensive chemometrics' (Elsevier, Oxford, 2009, 2nd edn.), pp. 543–554

[24] Vapnik, V.N.: 'The nature of statistical learning theory' (Springer-Verlag, Berlin, Heidelberg, 1995)

[25] Nair, S.S., French, R.M., Laroche, D., et al.: 'The application of machine learning algorithms to the analysis of electromyographic patterns from arthritic patients', IEEE Trans. Neural Syst. Rehabil. Eng., 2010, 18, (2), pp. 174–184

[26] French, R.M., Glady, Y., Thibaut, J.P.: 'An evaluation of scanpath-comparison and machine-learning classification algorithms used to study the dynamics of analogy making', Beh. Res. Meth., 2017, 49, (4), pp. 1291–1302

[27] Raczko, E., Zagajewski, B.: 'Comparison of support vector machine, random forest and neural network classifiers for tree species classification on airborne hyperspectral APEX images', Eur. J. Remote Sens., 2017, 50, (1), pp. 144–154

[28] Wright, A.H.: 'Genetic algorithms for real parameter optimization', in Rawlins, G.J.E. (ed.): 'Foundations of genetic algorithms' vol. 1, (Elsevier, USA, 1991), pp. 205–218

[29] King, D.E.: ‘Dlib-ml: a machine learning toolkit’, J. Mach. Learn. Res., 2009, 10, pp. 1755–1758

[30] McNaught, A.D., Wilkinson, A.: ‘IUPAC compendium of chemical terminology: the gold book’ (WileyBlackwell, UK, 1997, 2nd Revised edn)