Proceedings of SPIE , the International Society for Optical Engineering, Vol. 5826, 2005.

Classification of a target analyte in solid mixtures using principal component analysis, support vector machines and Raman spectroscopy.

Marie-Louise O'Connell,[a] Tom Howley,[b] Alan G. Ryder,[a*] Marc N. Leger,[a] Michael G. Madden.[b]

[a] Department of Chemistry/National Centre for Biomedical Engineering Sciences, National University of Ireland, Galway.
[b] Department of Information Technology, National University of Ireland, Galway.

ABSTRACT

The quantitative analysis of illicit materials using Raman spectroscopy is of widespread interest for law enforcement and healthcare applications. One of the difficulties faced when analysing illicit mixtures is the fact that the narcotic can be mixed with many different cutting agents. This obviously complicates the development of quantitative analytical methods. In this work we demonstrate some preliminary efforts to try and account for the wide variety of potential cutting agents, by discrimination between the target substance and a wide range of excipients. Near-infrared Raman spectroscopy (785 nm excitation) was employed to analyse 217 samples, a number of them consisting of a target analyte (acetaminophen) mixed with excipients of different concentrations by weight. The excipients used were sugars (maltose, glucose, lactose, sorbitol), inorganic materials (talcum powder, sodium bicarbonate, magnesium sulphate), and food products (caffeine, flour). The spectral data collected was subjected to a number of pre-treatment statistical methods including first derivative and normalisation transformations, to make the data more suitable for analysis. Various methods were then used to discriminate the target analytes, these included Principal Component Analysis (PCA), Principal Component Regression (PCR) and Support Vector Machines.

Keywords: Raman, Spectroscopy, Forensic, Classification, Chemometrics, Support Vector Machines.

1. INTRODUCTION

Many spectroscopic tools have been incorporated into the field of forensic analyses.[1,2] Since the early 1990's, technological advances in electronics; as well as materials have catapulted analytical methods, which were previously laboratory based into the field of in-situ investigations, one such method is Raman spectroscopy.[3,4] Based on the Raman effect, this technique provides a wealth of information about narcotic samples. Inelastically (or Raman) scattered light from a sample is analysed using a spectrometer and displayed in the form of a spectrum from which valuable information on the molecular structure and functional groups present may be obtained.[5,6] A change in the polarisability of a molecule is the only condition to be met for a substance to be Raman active. This requirement is unique to Raman spectroscopy and complimentary to IR spectroscopy, which requires a change in dipole moment for the substance to be IR active.[7,8] Although previously seen as simply a complimentary tool to IR spectroscopy, which was difficult to implement experimentally, technological advances such as diode lasers, sensitive CCD detectors, and inexpensive computing, have helped Raman spectroscopy to evolve into a rapid, non-contact, easily implemented technique. However, even with all these advances, the process of illicit narcotic analysis using Raman spectroscopy still faces several difficulties. Some of the materials used to bulk and/or bind narcotic active ingredients can be fluorescent thus obscuring the weaker Raman signal.[9] Using longer wavelength excitation minimises this effect but it still remains a problem in positive narcotic identification and quantification. In addition illicit narcotics can be mixed with several different excipients, all of which will have different Raman scattering efficiencies which can cause some (or all) of the Raman bands of the illicit narcotic to be obscured, making the process of identification or quantification more difficult.

* alan.ryder@nuigalway.ie; phone 353-91-492943; fax 353-91-494596

The goal of this work is to determine the utility of Raman spectroscopy for identifying and quantifying the presence of a target analyte in solid mixtures. Specifically we aim to study the conditions that pertain to illicit narcotics where the mixtures can comprise of a wide range of different materials with a single analyte (e.g. cocaine). For Raman spectroscopy to be an effective tool in quantitative forensics analysis, it is necessary to demonstrate that it can deal with spectral differences due to sample variability; the principal variances encountered are baseline changes due to fluorescence/scatter, and overlapping Raman bands. To make use of Raman spectroscopy effectively, the use of multivariate analysis techniques is required to extract the maximum amount of practical information from the spectra, while taking into account the various interferences that are encountered in Raman spectroscopy.

In this work we investigate the efficacy of Raman spectroscopy and chemometrics for the identification and quantification of a target analyte in solid mixtures. We compare the efficiency of chemometric methods with machine learning (ML) techniques such as Support Vector Machine (SVM).

2. EXPERIMENTAL

Excipient	Abbreviation
Baby Powder	BP
Caffeine	CAF
Cellulose	CEL
Flour	FLO
Glucose	GLU
Lactose	LAC
Magnesium Sulphate	Mg
Maltose	MAL
Sorbitol	SOR
Talc	TAL

Table 1: Abbreviations used for the excipients used in the study.

2.1. Apparatus and materials

The target analyte for this model system was acetaminophen (4-acetamidophenol 98% – Acros Organics). The bulk of the excipients used were of reagent grade and procured from the Inorganic Chemistry Dept., NUIG and used as received. The sample set (see Table 2) consisted of the target analyte present in various concentrations, mixed with common narcotic excipients.[9] The dataset also contained a subset of pure inorganic materials. The target analyte and excipients were carefully weighed and mixed using an agate mortar and pestle, ensuring homogeneity throughout the sample. Samples were transferred to a sample holder, which comprised of 3 mm depth cylindrical holes drilled into 5 mm deep microscope sized stainless steel plate.

The data was collected using a Horiba Jobin Yvon Labram Infinity at 785nm excitation, within the range of 350-2000 cm^{-1}. The near infrared laser was used to minimise the effects of fluorescent materials on the spectra. The data was viewed and collected through a x10 power objective. The exposure times of the mixtures were tailored to produce the optimal spectrum in a reasonable time; hence these intervals differed for different samples and concentrations. These times ranged from 3 to 30 seconds per scan depending on the sample being examined e.g. a strong Raman scatterer such as crystallised 4-nitroaniline had one of the lowest exposure times of 3 seconds as this was sufficient to obtain a spectrum with strong, well resolved peaks. A pure spectrum of naphthalene and a luminous intensity standard (SRM 2241) were taken for every three target-excipient samples examined. To reduce sampling effects (i.e. local inhomogeneity), each sample was analysed at sixty-four different places on the sample.[10,11] An eight by eight grid was used to take Raman spectra over a large area at each sample. Each set of sixty-four spectra were co-added, averaged and corrected for instrument response. For pure samples a sixteen-point grid (four by four) was used.

% ACETAMINOPHEN IN:									
BP	CAF	CEL	FLO	GLU	LAC	MG	MAL	SOR	TAL
50	50	50	50	50	50	50	50	50	49
20	20	20	20	20	20	20	19	20	20
15	15	15	15	15	15	15	14	15	15
10	10	10	-	-	10	10	10	10	10
9	9	9	9	9	9	9	9	9	9
8	8	8	8	8	8	8	8	-	-
7	7	7	7	7	7	7	7	7	7
6	6	6	6	6	6	6	6	6	6
5	5	5	5	5	5	5	5	5	5
0	0	0	0	0	0	0	0	0	0

Table 2: Percentage of target (Acetaminophen) in each of the most common narcotic excipients.

2.2. Software and hardware

Chemometric analysis was performed using Unscrambler (Version 8.0 1986-2003 Camo Process AS, Oslo, Norway) multivariate analysis software package. All analyses were carried out on desktop PC's. A MATLAB (Version 7.0.0.19920 (R14), The Mathworks Inc.) code was specifically written for this data which both reduced (averaged) each set of sixty-four spectra to one whilst simultaneously correcting it for instrument response.[7,12,13] The full dataset comprised of 217 samples, which were then randomly shuffled before embarking on any data analysis.

2.3. Data Analysis

Within the randomly shuffled dataset 22-fold, "leave 10 out" cross-validation is done on the data, using different classification algorithms. In cross-validation, the dataset is divided into 22 equal segments or folds, a model is built using all the data from folds 2 to 22, and this model is then used to predict the data from the first fold. Then, the second fold is left out of the training (which is now based on folds 1 and 3-22) and used as the test dataset – this process is repeated until all folds have been held out as a test set. In this way, the classification method is tested on all 217 samples of the dataset.

2.4. Chemometrics

Principal component analysis (PCA) is a mathematical method that involves transforming a number of possibly correlated variables into a smaller number of uncorrelated variables or principal components. The first principal component accounts for the largest amount of variation in the data, the second component accounts for the second largest amount of variation, and so forth. Using the information provided by PCA analysis allows the building of an improved model, by identifying unwanted contributory effects such as background fluorescence and residual laser scatter. The dataset was subjected to four different types of pre-processing, as outlined in Table 3. In an attempt to remove the domination of the intense Raman scatterers in the sample set, a maximum normalisation was employed. If all samples are positive when the normalisation is applied, the maximum Raman band intensity is set at +1. This method divides each sample row by its maximum absolute value. Performing this step eliminated the large intensity difference between the spectra. A Savitzky-Golay first derivative, seven-point averaging algorithm was applied to the raw data set. This algorithm fits a polynomial to each successive curve segment, thus replacing the original values with more regular variations. The scores identified in PCA were used as predictors in a multiple linear regression.

Samples	Pre-Processing	Code
All	Raw – No pre-processing	RD-1
All	First Derivative	FD-1
All	Maximum Normalisation	ND-1
All	First derivative + Maximum Normalisation	FDN-1

Table 3: Table of pre-processing methods applied to dataset and corresponding code.

2.5. Machine Learning

Machine Learning, in its broadest sense, refers to the study of computer algorithms that improve through experience. In spectral classification, the goal of machine learning is to build a model of the how the Raman spectrum of a sample relates to the presence or absence of a target substance in that sample. The work presented here focuses specifically on the use of machine learning for the detection of acetaminophen. The following sections provide a brief description of the machine learning methods that were used.

2.5.1. Support Vector Machine

The SVM[14] is a powerful machine learning tool that is capable of representing non-linear relationships and producing models that generalise well to unseen data. For binary classification, a linear SVM (the simplest form of SVM) finds an optimal linear separator between the two classes of data, in this case acetaminophen and non-acetaminophen samples. This optimal separator is the one that results in the widest margin of separation between the two classes, as a wide margin implies that the classifier is better able to classify unseen spectra. Typically, it is not possible to find a classifier that completely separates the data. In fact, efforts to produce a classifier that completely separates the data could result in an over fitted model that performs poorly on test data. In order to regulate this overfitting, SVMs have a complexity parameter, C, which determines the trade-off between choosing a large-margin classifier and the amount by which misclassified samples are tolerated. A higher value of C means that more importance is attached to minimising the amount of misclassification than to finding a wide margin model.

To handle non-linear data, kernels (e.g. Radial Basis Function (RBF), Polynomial or Sigmoid) are introduced to map the original data to a new feature space in which a linear separator can be found. In addition to the C parameter, each kernel may have a number of parameters associated with it. For the experiments reported here, two kernels were used: the RBF kernel, in which the kernel width (σ) can be changed, and the Polynomial kernel, which also has one parameter – degree (d). Note that the linear kernel is equivalent to a Polynomial kernel of degree one and corresponds to the original input space. The SVM is considered useful for handling high dimensional data and should therefore be suited to the spectral domain. See Cristianini and Shawe-Taylor for more details on SVM classification.

2.5.2. k-Nearest Neighbours

k-Nearest Neighbours (kNN)[15] is a learning algorithm which classifies a test sample by firstly obtaining the class of the k samples that are the closest to the test sample. The majority class of these nearest samples (or nearest single sample when k = 1) is returned as the prediction for that test sample. Various measures may be used to determine the distance between a pair of samples. In these experiments, the Euclidean distance measure was used. In practical terms, each Raman spectrum is compared to every other spectrum in the dataset. At each spectral data point (wavenumber axis) the difference in intensity between the two spectra is measured (distance). The sum of the squared distances for all the data points (full spectrum) gives a numerical measure of how close the spectra are i.e. nearest neighbours.

2.5.3. C4.5

The C4.5 decision tree[16] algorithm generates a series of "if-then" rules that are represented as a tree structure. Each node in the tree corresponds to a test of the intensity at a particular wavenumber. The result of a test at one node determines which node in the tree is checked next until finally, a leaf node is reached. Each leaf specifies the class to be returned if that leaf is reached.

2.5.4. Ripper

RIPPER[17] (Repeated Incremental Pruning to Produce Error Reduction) is an inductive rule-based learner that builds a set of prepositional rules that identify classes while minimising the amount of error. The number of training examples misclassified by the rules defines the error. RIPPER was developed with the goal of handling large noisy datasets efficiently whilst also achieving good generalisation performance.

2.5.5. Naïve Bayes

Naïve Bayes[18] is a classification technique, which for each class calculates the probability that the test sample belongs to that class, given the intensities at each point in the spectrum. It then returns its prediction as the class with the highest probability. One potential failing of this approach is that it is based on the assumption that each attribute (i.e. wavenumber) is independent, which is not true for spectral data. However, it has been shown in other application domains that Naïve Bayes classifiers can work well, even when this independence assumption does not hold.[19]

3. RESULTS & DISCUSSION

We first examined the data set using traditional chemometrics and followed a step-by-step procedure to investigate the factors that will influence the methods for identifying and quantifying acetaminophen in mixtures.

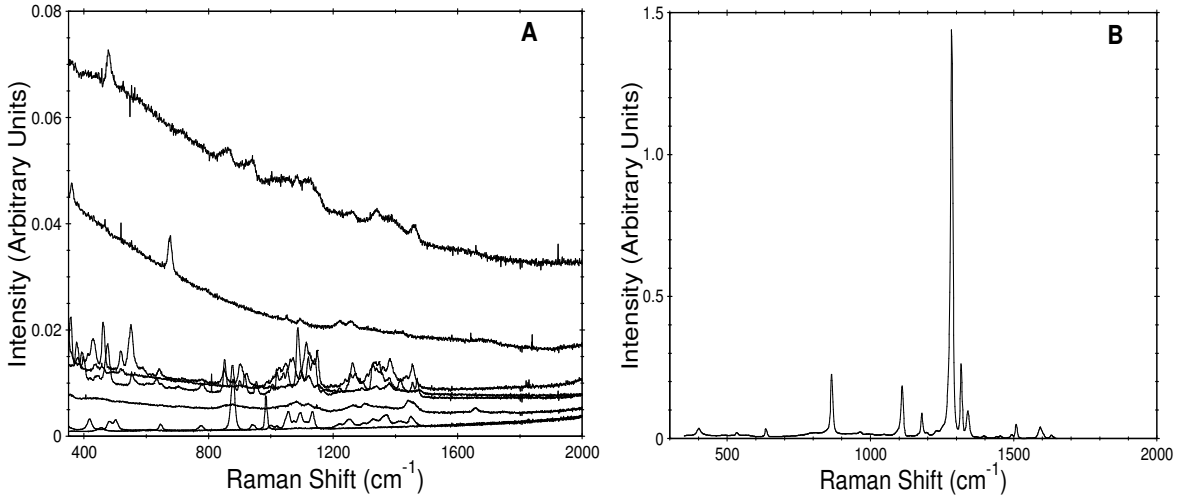

Figure 1: A) Raw dataset showing the varying baseline offsets of some samples. B) 4-Nitroaniline – raw – intense Raman scatterer.

The collection of spectra above (Figure 1) represents some of the raw data collected. No pre-processing has been attempted and no baseline corrections have been made. There is an obviously large variation in Raman intensity between samples (Figure 1 A). The 4-nitroaniline sample (Figure 1 B) is a very strong Raman scatterer with an intensity of more than ten times that of any other sample investigated. In addition, the samples in figure 1 A were recorded using much longer exposure times (20-30 seconds) than the 4-Nitroaniline (3 seconds). This highlights one of the difficulties associated with using Raman spectroscopy, which is the wide variance in spectral intensity, which must be taken into account when developing analytical methods. The second noticeable effect in the raw Raman spectra is the variable baseline, which dominates some samples. This can be attributed to either fluorescence or residual laser scatter, and it increases proportionally with the exposure time, as does the Raman signal. Longer exposure times are needed for samples with these interferences in order to obtain defined Raman peaks, however this also increases the impact of sloping background and baseline offset in multivariate analysis. In order to identify the variation in the Raman dataset due to baseline effects and intensity effects PCA was carried out on the uncorrected spectra (Figure 2).

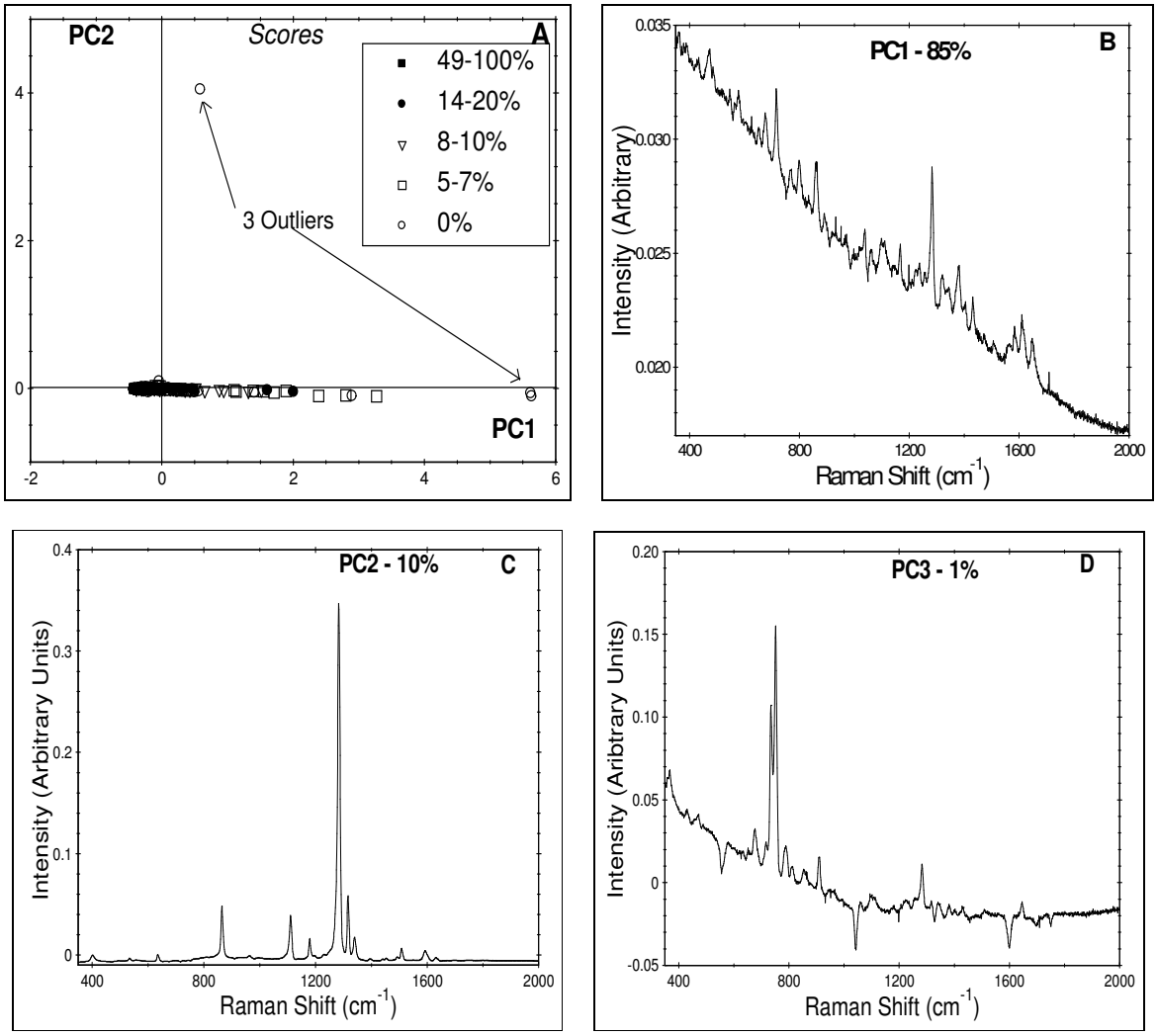

Figure 2: A) PCA Scores loading B) PCA x-loadings spectrum PC1 C) PCA x-loadings spectrum PC2 D) PCA x-loadings spectrum PC3 for RD-1 dataset.

The scores plot between the 2 major components (Fig. 2A) shows that there is very little separation of samples according to the target analyte concentration. PC1 (Fig. 2B) essentially describes underlying baseline effects and accounts for a lot of variation (85% explained variance). Separation along PC1 is due to the baseline effect.
PC2, 10% explained variance, (Figure 2 C) essentially represents 4-Nitroaniline, the most intense Raman scatterer sampled (Figure 1B). The third PC, 1% explained variance, (Fig 2 D) is due to pure potassium iodate (another intense Raman scatterer) with another baseline contribution. These 3 PC's represent 96% of the spectral variation in the raw spectral data, which hinders the identification of the target analyte. It is therefore obvious from PCA that the data needs to be corrected for both baseline and Raman intensity effects before it can be used for analyte classification.

The only automated and reproducible method for background correction, which is available to us at the moment, is the use of derivatisation. While this has drawbacks in making the data more difficult to understand, it does enable the reproducible removal of baseline effects. Therefore, we performed a first derivative to remove the baseline effects and the PCA model run once more under identical conditions. PC1 (not shown) in this case now accounts for 68% of the spectral variation, and when examined is seen to be due to the 4-nitroaniline sample (strong Raman scatterer). The large

disparity due to the background effects was eliminated. However, the dominance of PC1 due to Raman intensity effects still needs to be accounted for, otherwise it will have an overly large influence on the chemometric process.

The next step was to remove the absolute Raman intensity effects by normalizing each Raman spectrum to the most intense peak. This should remove the large weightings being assigned to individual Raman spectra like that of 4-nitroanailine. PCA was repeated on this dataset and the scores plot for the first 2 PC's is shown in Figure 3.

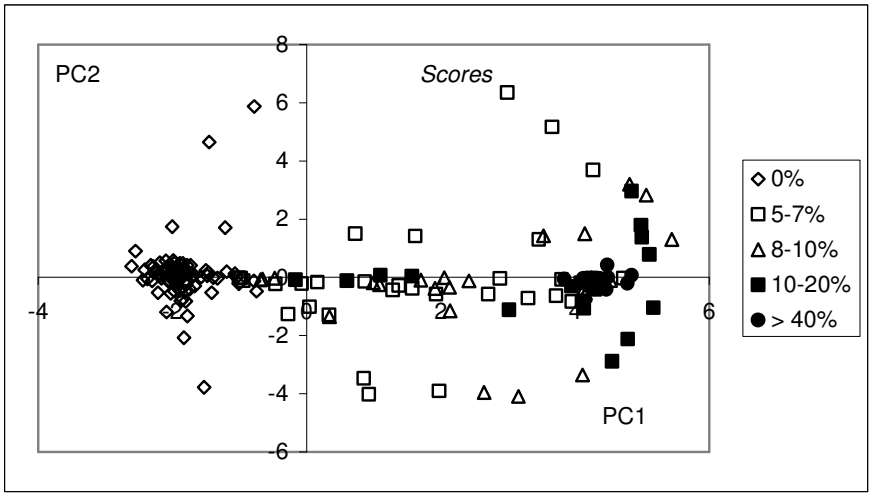

Figure 3: Scores plot of PC1 vs. PC2, on 1st derivative and normalised dataset showing separation of samples based on acetaminophen concentration.

This scores plot demonstrates the effect of background removal and the normalisation of the dataset. There is now a clear distinction between samples with no target analyte present and those with the acetaminophen present. The zero acetaminophen concentration samples are found at negative PC1 scores whilst those most heavily concentrated with the target analyte (50-100% range) are found at positive PC1 scores. The smaller concentrations (5-10%) lie scattered in between the two sets. PC1 which contributes the most to the model (26%) is related to acetaminophen, as can be seen by comparing the loadings plot with the 1st derivative plot of the Raman spectrum of acetaminophen Figure 4. The reason for the separation of samples along PC2 is less clear-cut and examination of the loadings plot (not shown) and Raman spectra did not reveal any significant explanation. However, it does appear that the some of the separation is related to the sugar based excipients, with predominately high concentration glucose mixtures in the upper two quadrants and lactose based mixtures in the lower two.

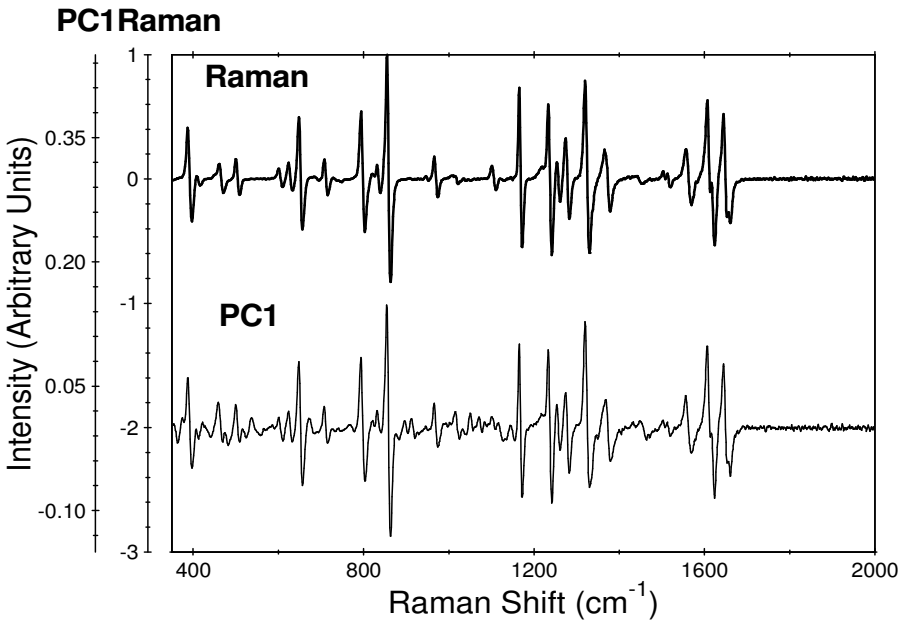

Figure 4: Top - First Derivative of acetaminophen; Bottom – From PCA overview – PC(x-expl): 1(26%) from FND-1 dataset.

The samples in Figure 3 have been separated according to the acetaminophen concentration by applying first derivative and maximum normalisation to correct for baseline offsets and intensity differences respectively. In order to determine numerically the accuracy of classifying samples according to the presence of acetaminophen, we adopted a simple threshold test based on Principal Component Regression (PCR). If the predicted concentration of acetaminophen in a sample is lower than the threshold then it is said not to contain acetaminophen. If the predicted concentration is higher than the threshold then it is said not to contain acetaminophen. The errors described in Table 4 give the frequency at which this binary determination was incorrect.

Dataset	**Threshold**		
	2%	3%	5%
RD-1	56.0%	55.1%	48.6%
FD-1	57.1%	56.1%	49.3%
ND-1	18.4%	16.6%	12.0%
FND-1	**7.4%**	**4.6%**	**5.1%**

Table 4: Percentage % error of PCR classification.

This is a simple binary test for comparing the PCR results with the machine learning results in Table 5. The PCR results reflect the gradual improvement of the dataset model due to the PCA. Surprisingly the first derivative PCR results were all less accurate than the raw dataset. When faced with these results it was decided to repeat calculations but on normalised data only i.e. without the first derivative applied. This treatment produced better classification accuracy than both the raw dataset and first derivative only. From this we can assume that a more efficient method for correcting for the inherent weighting between spectra in this Raman dataset. Applying the first derivative algorithm in combination with a maximum normalisation transformation and setting a threshold value of three percent produced a classification accuracy of 4.6% which was the best model developed.

As shown in Table 5, the machine learning analysis methods were found to surpass the traditional chemometrics methods in classification of samples.

Method	RD-1	FD-1	FND-1
RBF SVM	**5.1%** *(C=1000, σ =0.01)*	**0.9%** *(C=10, σ=0.01)*	**0.9%** *(C=10, σ=0.01)*
Polynomial SVM	**5.1%** *(C=100, d =1)*	**4.1%** *(C=0.1, d=1)*	**0.9%** *(C=1, d=1)*
K-Nearest Neighbours	12.0% *(k=1)*	**5.1%** *(k=6)*	**3.2%** *(k=1)*
C4.5	11.5%	**1.8%**	**1.4%**
Ripper	14.7%	**1.4%**	**1.4%**
Naïve Bayes	24.0%	25.3%	6.0%

Table 5: Percentage error of Machine Learning methods in the classification of acetaminophen presence in solid mixtures e.g. a result of 4.6% means that 10 samples were misclassified in total.

Table 5 shows the results of five different machine learning classification methods (the results for both RBF and Polynomial kernel SVMs are shown) on three versions of the acetaminophen dataset, i.e. raw, first derivative, and first derivative normalised. For all these methods, the WEKA[15] implementation was used. The default settings were used for C4.5, Ripper, and Naïve Bayes. For SVMs, RBF, and Polynomial kernels with different parameter settings were tested. The parameter settings that achieved the best results are shown in parentheses. Similarly, for kNN, the table shows the value for *k* (number of neighbours) that resulted in the lowest percentage error. The kNN method was tested for values of *k* from 1 to 20.

In Table 5, the lowest error of 0.9% was achieved by the SVM on the FD-1 (using a RBF kernel with C=10 and σ=0.01) and on the FND-1 dataset (using a linear kernel with C=1.0, a RBF kernel also managed this level of accuracy). However, a paired t-test shows that there is no significant difference the results achieved by the SVM on FND-1 and all the other classifiers whose error values are highlighted in bold in the table. The paired t-test is based on a 5% confidence level and, because a cross-validation test was run, uses a corrected variance estimate, as recommended by Nadeau & Bengio.[20] The error of 0.9% equates to the misclassification of two samples out of the 217. The RBF kernel (on FD-1) and the linear kernel (on FDND-1) misclassified two different samples, which implies that the misclassification is an inherent property of the different classifiers used, as opposed to being caused by an outlier in the dataset.

A significant difference was observed between the best PCR model (using a threshold of 3% on the FND-1 dataset) and the linear kernel SVM on the FND-1 dataset (Note that only the RBF SVM and Polynomial SVM significantly improve on the best PCR result). Overall, the SVM appears to exhibit the best results, matching or outperforming all other methods on the raw and pre-processed data. As with the chemometric methods, three of the classifiers (kNN, C4.5 and Ripper) improve their accuracy when based on the pre-processed versions of the dataset. The Naïve Bayes classifier does not improve until a first derivative followed by a maximum normalisation has been applied to the data, whereas the SVM achieves a good performance on all versions of data.

These results have shown that machine learning methods are capable of outperforming the traditional chemometric methods in the classification of the target acetaminophen based on spectral data. They have also shown that the classification accuracy of these machine learning methods can be improved by pre-processing techniques such as first derivative and maximum normalisation.

4. CONCLUSIONS

The Raman spectra of a wide variety of solid mixtures containing the target analyte (acetaminophen) in various concentrations were collected. The resulting spectra were analysed using traditional chemometrics to investigate the principal factors that influence the development of qualitative and quantitative analysis methods. Initial investigations identified these as baselines due to fluorescence and/or residual laser scatter and Raman intensity effects. Pre-processing techniques such as first derivative and maximum normalisation reduced the impact of these unwanted effects and allowed sample separation according to acetaminophen concentration. Performing PCR and SVM gave a numerical accuracy of this classification technique. The SVM classifier based on data, which has had a first derivative followed by maximum normalisation applied to it, has been shown to outperform PCR in the identification of Acetaminophen. The superior performance of the SVM over all other methods used in this research is particularly evident when the raw data is used as the input. The majority of machine learning methods have also been shown to benefit from the pre-processing techniques of first derivative and maximum normalisation.

To further deal with problems such as misclassification of samples, future work could investigate the use of PCA in combination machine learning methods such as SVM. This might contribute an illustrative identification element to machine learning, which would permit recognition of misclassified samples in the form of Raman spectra. The development of an automated baseline removal programme would also greatly improve pre-processing techniques

5. ACKNOWLEDGEMENTS

This work was funded via a Basic Research grant from Enterprise Ireland: "Quantitative Raman Spectroscopy using Machine Learning Methods: Hazardous and Illicit Substance Analysis" (SC/2003/334/Y). The Raman instrumentation was provided by the National Centre for Biomedical Engineering Science as part of the Irish Higher Education Authority's Programme for Research in Third Level Institutions. The support of Science Foundation Ireland is also acknowledged.

6. REFERENCES

[1] I.R. Lewis, H.G.M. Edwards, *Handbook of Raman Spectroscopy –From the Research Laboratory to the Process Line*, Marcel Dekker, Inc., New York, 2001.

[2] I.P. Hayward, T.E. Kirkbride. D.N. Batchelder, R.J. Lacey, "Use of a Fibre Optic Probe for the Detection and Identification of Explosive Materials by Raman-Spectroscopy", *J. Forensic Sci.*, **40**, 883-884, (1995).

[3] B.J. Kip, T. Berghmans, P. Palmen, A. Van Der Pol, M. Huys, H. Hartwig, M. Scheepers, D. Wienke, "On the use of recent developments in vibrational spectroscopic instrumentation in an industrial environment: quicker, smaller and more robust", *Vib. Spectrosc.*, **24**, 75–92, (2000).

[4] S.D. Harvey, M.E. Vucelick, R.N. Lee, B.W. Wright, "Blind field test evaluation of Raman spectroscopy as a forensic tool", *Forensic Sci. Int.*, **125**, 12–21, (2002).

[5] A.G. Ryder, G.M. O'Connor, T.J. Glynn, "Quantitative analysis of cocaine in solid mixtures using Raman spectroscopy and chemometric methods," *J. Raman Spectros.*, **31**, 221-227, (2000).

[6] A.G. Ryder, "Classification of narcotics in solid mixtures using Principal Component Analysis and Raman spectroscopy", *J. Forensic Sci.*, **47**, 275-284, (2002).

[7] R.L. McCreery, *Handbook of Vibrational Spectroscopy*, 251-289, John Wiley & Sons Ltd., Chichester, 2002.

[8] A.G. Ryder, G.M. O'Connor, T.J. Glynn, "Identifications and quantitative measurements of narcotics in solid mixtures using near-IR Raman spectroscopy and multivariate analysis", *J. Forensic Sci.*, **44,** 1013-1019, (1999).

[9] S.E.J. Bell, L.J. Barrett, D.T. Burns, A.C. Dennis, S.J. Speers, "Tracking the distribution of "ecstasy" tablets by Raman composition profiling: A large scale feasibility study", *The Analyst*, **128**, 1331-1335 (2003).

[10] S.E.J. Bell, D.T. Thorburn. A.C. Dennis, J.S. Speers, "Rapid analysis of ecstasy and related phenethylamines in seized tablets by Raman spectroscopy", *Analyst*, **125,** 541-544, (2000).

[11] P. Niemelä, M. Päällysaho, P. Harjunen, M. Koivisto, VP. Lehto, J. Suhonen, K. Järvinen, "Quantitative analysis of amorphous content of lactose using CCD-Raman spectroscopy", *Journal of Pharmaceutical and Biomedical Analysis*, [Article in Press] (2004).

[12] K.G. Ray, R.L.McCreery, "Simplified calibration of instrument response function for Raman spectrometers based on luminescent intensity standards", *Appl. Spectrosc.*, **51**, 108-116, (1997).

[13] K. J. Frost, R.L. McCreery, "Calibration of Raman spectrometer instrument response function with luminescence standards: an update", *Appl. Spectrosc.*, **52**, 1614-1618, (1998).

[14] N. Cristianini and J. Shawe-Taylor, *An Introduction to Support Vector Machines*, Cambridge University Press, (2000).

[15] I.H. Witten & E. Frank, *Data Mining: Practical Machine Learning Tools and Techniques with Java Implementations.* Morgan Kaufmann Publishers (2000).

[16] J.R. Quinlan, Learning Logical Definitions from Relations, *Machine Learning* **5**, 239-266 (1990).

[17] W. Cohen, *Fast Effective Rule Induction,* In Proceedings of the 12th International Conference on Machine Learning, 115-123, (2002).

[18] P. Langley, W. Iba, and K. Thompson, *An Analysis of Bayesian Classifiers,* In Proceedings of the 10th International Conference on Artificial Intelligence, 223-228, (1992).

[19] S.J. Russell and P. Norvig, *Artificial Intelligence: A Modern Approach*, 2nd Edition, pg 482 (2003).

[20] C. Nadeau and Y. Bengio, *Inference for the generalization error.* Advances in Neural Information Processing Systems 12, MIT Press, (2000).