ANNALS OF “DUNAREA DE JOS” UNIVERSITY OF GALATI
MATHEMATICS, PHYSICS, THEORETICAL MECHANICS
FASCICLE II, YEAR X (XLI) 2018, No. 1

Article DOI: https://doi.org/10.35219/ann-ugal-math-phys-mec.2018.1.11

COMPARATIVE ASSESSMENT OF THE MODELING AND DISCRIMINATION POWER OF TWO PATTERN RECOGNITION METHODS APPLIED TO DETECT DESIGNER DRUGS

Mirela Praisler[1], Stefanut Ciochina[2]

[1]"Dunarea de Jos" University of Galati, Department of Chemistry, Physics and Environment
[2]"Dunarea de Jos" University of Galati, Department of Mathematics and Computer Science
Mirela.Praisler@ugal.ro, stefanut27@yahoo.com

Abstract
We are presenting a comparative assessment of the modeling and discrimination power of two pattern recognition methods, i.e. Hierarchical Cluster Analysis (HCA) and the Naive Bayes Classifier (NBC), from the point of view of their efficiency in detecting illicit amphetamines, based on their GC-IRAS laser spectra recorded between 1405 and 1150 cm^{-1}. A special attention was also given to the detection of their main precursors, the ephedrines. The spectra were first preprocessed with a discriminating feature weight w_{TE}. The performances of two automatic detection applications, based on HCA and on NBC, are compared from the point of view of their capacity to correctly recognize illicit amphetamines and ephedrines and distinguish among them according to the Schedules of the *United Nations Convention on Psychotropic Substances*.

Keywords: Amphetamines, ephedrines, pattern recognition.

1. INTRODUCTION

Amphetamines are the stimulants of the central nervous system that are the most frequently abused, ususally for recreational purposes. Amphetamine (α-methylphenethylamine) and some of its analogues may be found in legal preparations treating affections such as narcolepsy and obesity. However, they may be used only under strict medical supervision, as they do posses (moderate) psychological dependence liability and adiction liability. Hence, this class of compounds is listed under Schedule II of the *United Nations Convention on Psychotropic Substances* [1]. On the other hand, amphetamines such as the 3,4-methylenedioxyamphetamines and its analogues have no medical use and are listed under Schedule I of the same convention on controlled substances, as they may cause serious heart disease, dependence liability, as well as high rates of suicidal behaviors [2,3].

As amphetamines are synthetic drugs, the most recent trends in the fight against narcoterorism include the development of analytical instruments capable to perform *in-situ* detections not only of the end-products, but also of their main precursors. In the case of amphetamines, ephedrines are precursors the most frequently used by clandestine laboratories [4,5].

The portable instruments that are currently available may detect a given list of amphetamines [6]. However, in their attempts to avoid legal repercussions, drug dealers constantly introduce new illegal amphetamine analogues on the black market. Hence, portable instruments able to detect *in-situ* not only the known illegal drugs, but also any compound having a similar molecular structure, are highly needed.

Tthis paper presents a combination of artificial intelligence methods which leads to a forensic application operating a portable laser infrared spectrometer designed to detect amphetamine analogues, as well as ephedrines. The results presented in this paper have been obtained with the spectra recorded by using the UT8 quantum cascade laser (QCL), which emits in the 1405 - 1150 cm^{-1} range [7].

2. EXPERIMENTAL PART

The initial database consisted of the infrared spectra of 36 compounds encompassing the structure – activity correlations specific to the targeted drugs of abuse [8]. The spectra have been measured between 1405 and 1150 cm^{-1} with a resolution of 5 cm^{-1}. They belong to 7 illicit stimulant amphetamines (class code M), 6 analogues of ephedrine (class code E), 6 hallucinogenic amphetamines (class code T) and 17 non-amphetamines (class code N) of forensic interest. Previous studies have shown that selecting the most relevant variables before using clustering and / or classification methods is a very useful approach [9, 10]. Therefore, a feature weight emphasizing the most discriminating absorptions has been determined by using the Fisher function [4]:

$$w_{TE} = \frac{\sum \frac{A_I^2}{N_I} + \sum \frac{A_{II}^2}{N_{II}} - 2\sum\sum \frac{A_I A_{II}}{N_I N_{II}}}{\sum \frac{\left(A_I - \overline{A_I}\right)^2}{N_I} + \sum \frac{\left(A_{II} - \overline{A_{II}}\right)^2}{N_{II}}} \tag{1}$$

For this purpose, the spectra of the hallucinogenic amphetamines (T) and ephedrines (E) included in the database have been included in class I and the rest of the spectra in class II.

A new database, created by preprocessing the spectra of the above mentioned compounds with the w_{TE} feature weight, was further used as input for an exploratory analysis performed by Principal Component Analysis (PCA) [11], by using the MATLAB software.

The number of principal components (PCs) that are necessary in order to eliminate from the system the spectral information that is not relevant for modeling and discrimination has been evaluated. Then the number of clusters that may be clearly established [12] has been identified by using the Silhouette index [13].

The PCA scores have been then subjected to Hierarchical Cluster Analysis (HCA) [14, 15]. A clustering tree has been built by using the agglomerative clustering method. The accuracy with which the clustering tree assigns the class identity of the analyzed compounds has been assessed according to structure – activity correlations, and has been compared with the correct classification rate obtained in the case of the Naïve Bayes classifier [16].

3. RESULTS AND DISCUSSION

The number of PCs necessary for building the PCA models has been evaluated by analyzing the cumulative explained variance. The results indicate that the first five PCs are cumulating an explained variance of 96.56% (see Table 1). On the other hand, the last three PCs, especially PC4 and PC5, are characterized by much smaller explained variances than the first two PCs. Practically, most of the relevant information is described by the first two PCs, which cumulate an explained variance of 84.10%.

Table 1. Explained variance of the first principal components obtained for the w_{TE} preprocessed spectra recorded in the 1405 - 1150 cm^{-1} spectral domain.

Principal component	Explained variance (%)	Cumulated explained variance (%)
PC1	63.02	63.02
PC2	21.08	84.10
PC3	5.80	89.90
PC4	4.12	94.02
PC5	2.54	**96.56**

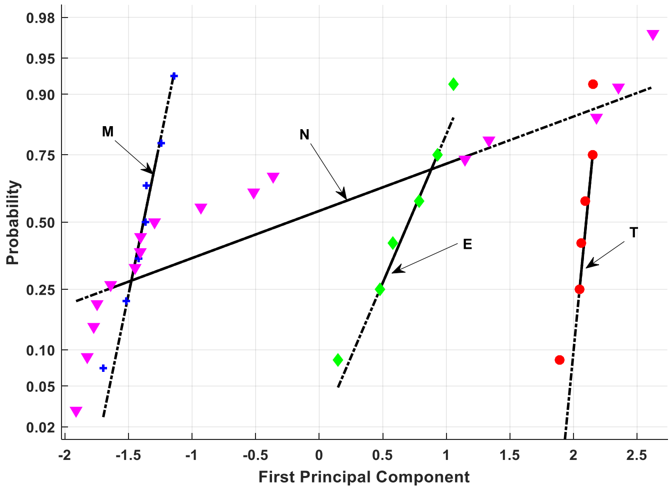
The probability of cluster membership has been assessed for each class of compounds included in the database based on the normal probability plots determined based on the PCA scores. The results obtained for the first three PCs are presented in Fig. 1. They indicate that the distributions of the scores associated to the samples forming the clusters of positive compounds (M, T and E) are relatively close to the normal distribution. As expected, the normal distribution does not apply to the negative samples. This behavior is due to the fact that this class is formed by chemicals having very different molecular structures and thus very dissimilar spectra.

Secondly, the normal probability plots indicate that the scores of the stimulant amphetamines are very similar to those of the negatives for all PCs. Hence, a larger number of misclassifications are to be expected between these two classes of compounds. This is true for all modeled classes of compounds in the case of PC3 and subsequent PCs. This aspect, corroborated with the results of the analysis of the explained variance, has indicated that further analysis should be performed with the first two PCs.

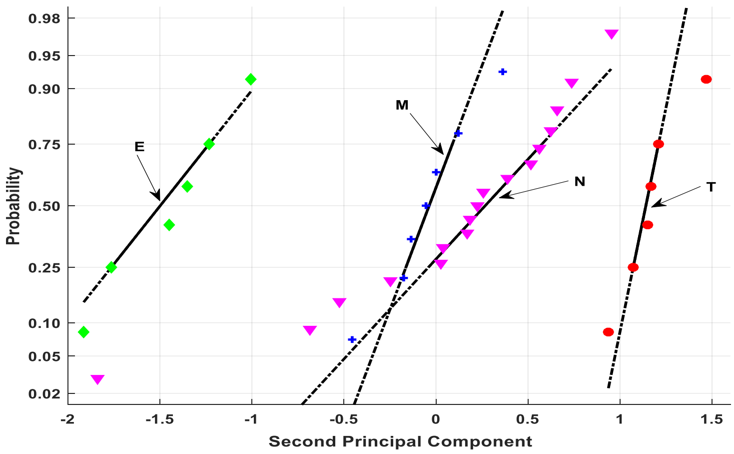
In order to assess if these two clusters may be distinguished well enough in order to obtain an acceptable correct classification rate, the PCA scores have been analyzed based on the Silhouette index. The results indicate that only three clusters may be well distinguished (see Fig. 2).

The nature of these clusters was identified by applying HCA. For this purpose, the PC1 and PC2 scores have been subjected to the agglomerative clustering algorithm. The resulting clustering tree (see Fig. 3) is characterized by a cophenetic correlation coefficient c = 0.8948. It indicates that the three distinguishable clusters are formed by the following classes of compounds: hallucinogenic amphetamines (T); ephedrines (E); stimulant amphetamines and negatives (M and N).

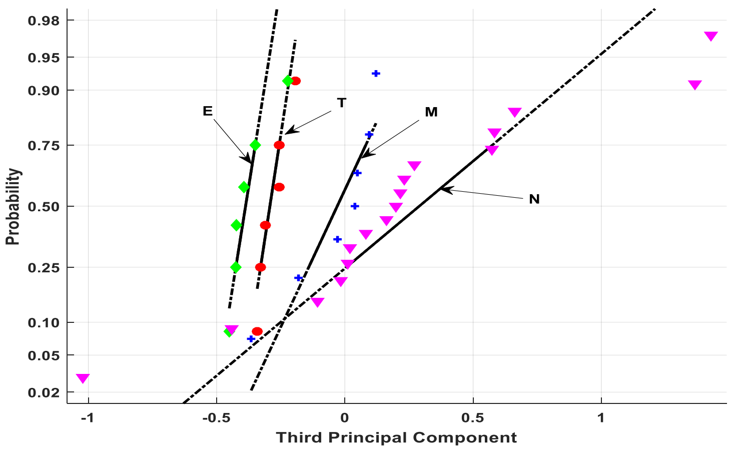
The clustering tree indicates that, from the point of view of assigning the T class identity, the system is remarkably sensitive. No hallucinogenic amphetamine is misclassified. The system is less selective, as a few negatives (i.e. N28, N23, N54 and N55) are classified as false hallucinogens. This may be explained by the fact that, although their full infrared spectra (4000-600 cm^{-1}) are very different, in the narrow spectral window of the UT8 QCL (1405 - 1150 cm^{-1}) these spectra are relatively similar to those of the T compounds (see Fig. 4) [17, 18]. However, as this is a forensic application, the sensitivity of the system is much more important than its selectivity [2].



a)



b)



c)

Figure 1. Normal probability plots determined for the modeled classes of compounds, i.e. hallucinogenic amphetamines (T), ephedrines (E), stimulant amphetamines (M) and negatives (N): a) PC1; b) PC2; c) PC3.

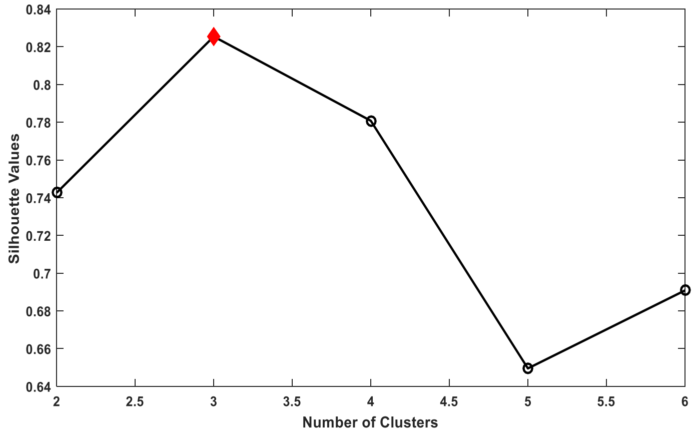


Figure 2. Silhouette index determined based on the PCA scores.

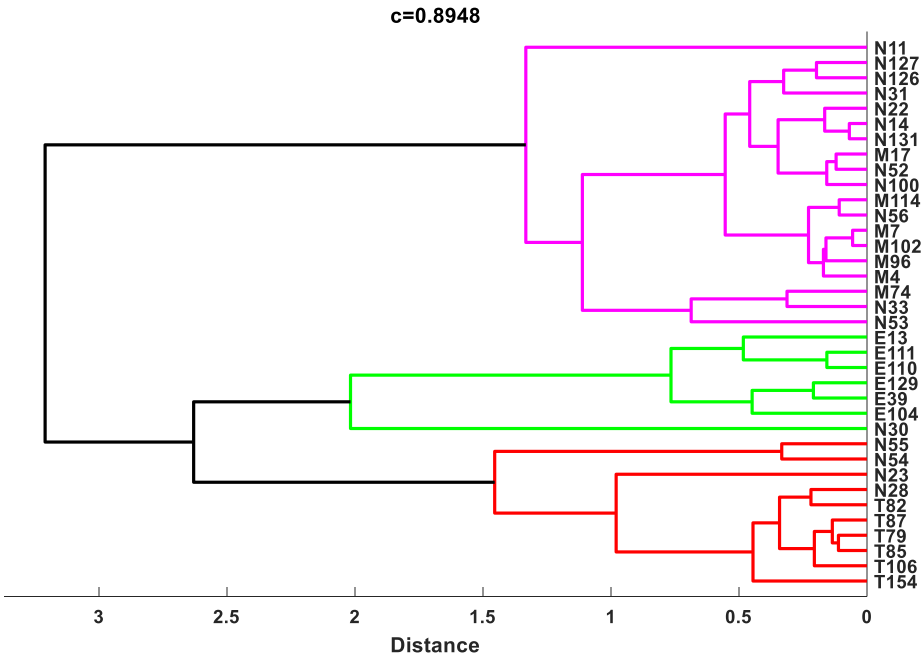


Figure 3. Clustering tree determined based on the PC1 and PC2 scores.

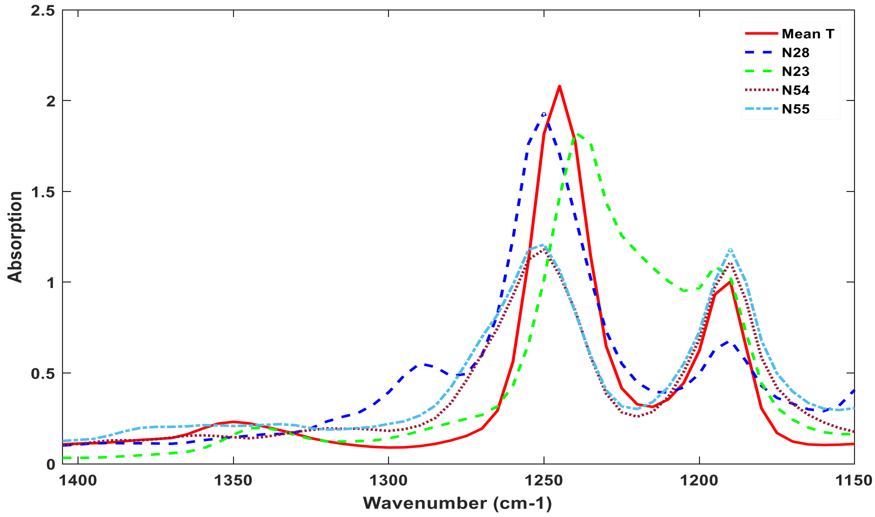


Fig. 4. Mean spectrum of the modeled hallucinogenic amphetamines and of the negatives classified as (false) hallucinogens

The system is best performing in the case of the ephedrines. No ephedrine is misclassified and only one negative (N30) is classified as a false ephedrine.

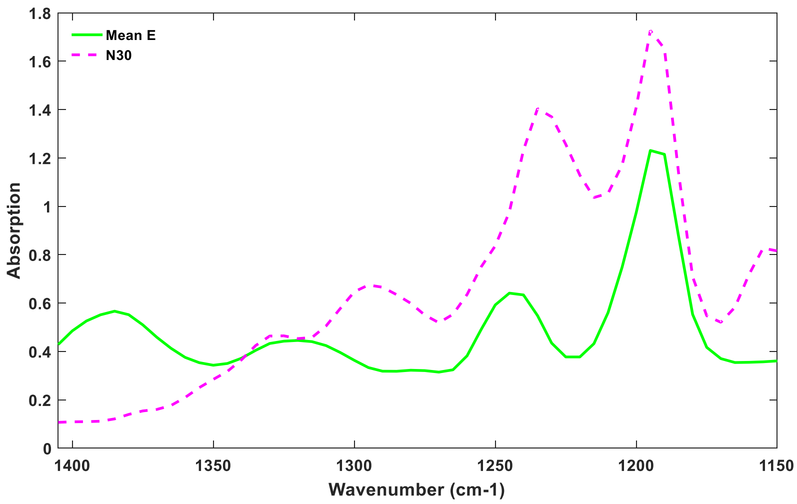


Figure 5. Mean spectrum of the modeled ephedrines and of a misclassified negative.

Much better results are obtained when the spectra are analyzed by using the Naïve Bayes classifier and the PCA scores obtained for the first two PCs (see Fig. 6). Table 2 presents the class identity assignments obtained with the Naïve Bayes classifier in the case of the negatives misclassified by HCA. Only one negative (N28) is still misclassified as a hallucinogen, the rest of the compounds being correctly recognized as negatives. Hence, the system based on the Naïve Bayes classifier is not only very sensitive, but also very selective.

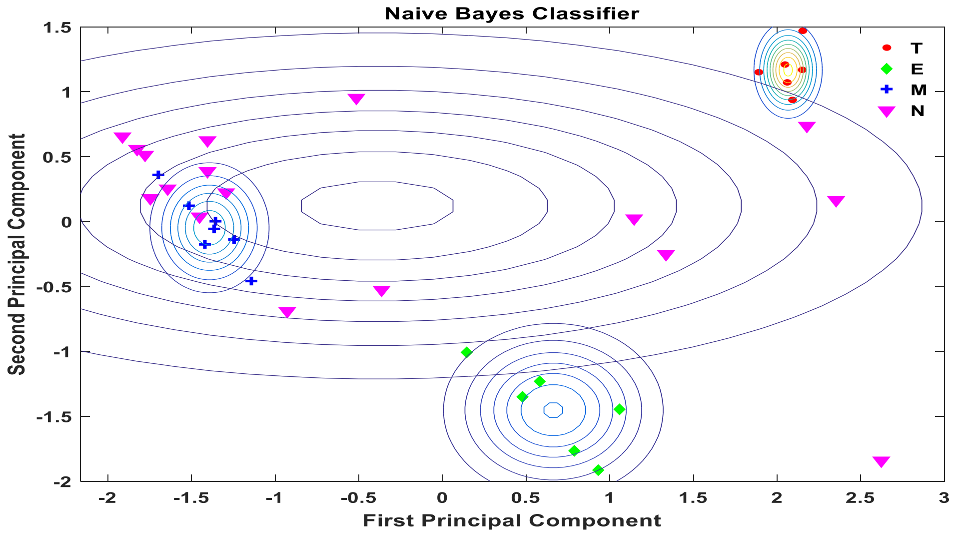


Figure 6. Class identity assignment based on the Naïve Bayes classifier.

Table 2. Class identity assignment based on the Naïve Bayes classifier performed for the negatives misclassified by the system based on Hierarchical Cluster Analysis.

Tested	Answer	Posterior				Cost			
		T	E	M	N	T	E	M	N
N28	T	0.7493	0	0	0.2507	0.2507	1	1	0.7493
N23	N	0	0	0	1	1	1	1	0
N54	N	0	0.0001	0	0.9999	1	0.9999	1	0.0001
N55	N	0	0.0015	0	0.9985	1	0.9985	1	0.0015
N30	N	0	0	0	1	1	1	1	0

4. CONCLUSION

We may conclude that HCA is useful only for distinguishing ephedrines and hallucinogenic amphetamines from other types of compounds. However, the system is designed to perform only *in-situ* screening. The compounds classified as positives are subsequently tested in laboratory conditions, based on their full GC-FTIR spectra. At this stage, the substances that have been misclassified as positives are easily identified.

On the other hand, a much better accuracy may be obtained if the class identity is performed on the basis of the Naïve Bayes classifier. In this case, the system becomes not only very sensitive, but also remarkably selective. Its accuracy recommends it as an efficient forensic tool screening for ephedrines and hallucinogenic amphetamines.

Acknowledgements

The work of Stefanut Ciochina was funded by the Romanian Ministry of European Funds within the POSDRU/107/1.5/S/76822 project. The authors are grateful for the financial support.

References

1. https://www.unodc.org/pdf/convention_1971_en.pdf
2. S. Karch, *Drug of Abuse Handbook,* 2nd ed. Boca Raton: CRC Press, 2007.
3. R. Laing (Ed,), *Hallucinogens. A Forensic Drug Handbook*, London: Academic Press, 2003.
4. M. Praisler, S. Ciochina, M. Coman, Hunting for Illicit Psychoactive Substances and Precursors: a Multivariate Approach, in Kloetzer, M; Ferariu, L. (Eds), *2017 21st International Conference On System Theory, Control And Computing (ICSTCC),* Book Series: International Conference on System Theory Control and Computing, 2017, pp. 248-253. Article number 8107042.
5. M. Praisler, S. Ciochina, C. Negoita, Improved Selectivity in Detecting Controlled Amphetamines and their Main Precursors based on Laser Infrared Spectra, *2017 E-Health and Bioengineering Conference, EHB 2017*, 28 July 2017, pp. 233-236. Article number 7995404.
6. J. M. Chalmers, H. G. M. Edwards, M. D. Hargreaves, *Infrared and Raman Spectroscopy in Forensic science*, Chichester: Wiley, 2012.
7. M. Praisler, S. Ciochina, M. Coman, Screening for Illicit Psychoactive Drugs Based on Pattern Recognition Methods, *5th International Symposium on Electrical and Electronics Engineering, ISEEE 2017*, 20-22 October 2017, Galati, Romania
8. S. Gosav, M. Praisler, D. O. Dorohoi, G. Popa, Structure – Activity Correlations for Illicit Amphetamines Using ANN and Constitutional Descriptors, *Talanta -The International Journal of Pure and Applied Chemistry* 70 (2006) 922-928.
9. S. Gosav, M. Praisler, D. O. Dorohoi, ANN Expert System Screening for Illicit Amphetamines using Molecular Descriptors, *Journal of Molecular Structure* 834-836 (2007) 188-194.
10. S. Gosav, M. Praisler, Artificial Neural Networks Built for the Recognition of Illicit Amphetamines Using a Concatenated Database, *Romanian Reports of Physics* 54-9/10 (2009) 929–935.
11. I. T. Jolliffe, *Principal Component Analysis*, 2nd ed. New York: Springer, 2002.
12. M. Praisler, S. Ciochina, Global clustering quality coefficient assessing the efficiency of PCA class identity assignment, *Journal of Analytical Methods in Chemistry*, volume 2014 (2014), Article ID 342497.
13. A. Starczewski, A. Krzyżak A. *Performance Evaluation of the Silhouette Index*, in L. Rutkowski, M. Korytkowski, R. Scherer, R. Tadeusiewicz, L. Zadeh, J. Zurada J. (Eds.) *Artificial Intelligence and Soft Computing* ICAISC 2015, Lecture Notes in Computer Science, vol. 9120, Springer, Cham
14. D. Mullner, *Modern hierarchical, aglomerative clustering algoritms*, arXiv:1109.2378v1, 2011.
15. L. Kaufman, P.J. Rousseeuw, *Finding Groups in Data. An Introduction to Cluster Analysis*, Willey, 2005.
16. P. Cichosz, *Naïve Bayes classifier*, in P. Cichosz (ed.), *Data Mining Algorithms: Explained Using R*, Chichester: Wiley, 2015.
17. S. Ciochina, M. Praisler, M., M. Coman, Hierarchical Cluster Analysis Applied for the Automated Recognition of Psychoactive Substances and of Their Main Precursors, 2017 5th International Symposium on Electrical and Electronics Engineering (ISEEE), 20-22 October 2017, Galati, Romania.
18. S. Ciochina, M. Praisler, C. Negoita, Cluster Analysis Evaluating the Automated Detection of Drugs of Abuse with a New Hollow Fiber based Quantum Cascade Laser Infrared Spectrometer, *2017 E-Health and Bioengineering Conference, EHB 2017*, 28 July 2017, pp. 237-240. Article number 7995405.