Screening Unknown Novel Psychoactive Substances Using GC-MS Based Machine Learning

Swee Liang Wong[1*], Li Teng Ng[2], Justin Tan[3], and Jonathan Pan[1]

[1]*Disruptive Technologies Office, Home Team Science and Technology Agency, Singapore*

[2]*Chemical, Biological, Radiological, Nuclear, and Explosives Centre of Expertise, Home Team Science and Technology Agency, Singapore*

[3]*Forensics Centre of Expertise, Home Team Science and Technology Agency, Singapore*

*Corresponding author: wong_swee_liang@htx.gov.sg

Keywords - Artificial intelligence, Chemometrics, Neural network, Machine learning, Random Forest, GC-MS, Spectrometry

Abstract

In recent years, there is a large increase in structural diversity of novel psychoactive substances (NPS), exacerbating drug abuse issues as these variants evade classical detection methods such as spectral library matching. Gas chromatography mass spectrometry (GC-MS) is commonly used to identify these NPS. To tackle this issue, machine learning models are developed to address the analytical challenge of identifying unknown NPS, using only GC-MS data. 891 GC-MS spectra are used to train and evaluate multiple supervised machine learning classifiers, namely artificial neural network (ANN), convolutional neural network (CNN) and balanced random forest (BRF). 7 classes, comprising 6 NPS classes (cathinone, cannabinoids, phenethylamine, piperazine, tryptamines and fentanyl) and other unrelated compounds can be effectively classified with a macro-F1 score of 0.9, averaged across all cross-validation folds. These results indicate that machine learning models are a promising complement as an effective NPS detection tool.

1. Introduction

Novel psychoactive substances have been on the rise recently, with yearly increase in numbers entering the illegal market. These synthetic drugs are controlled as they induce significant social and human health issues [1–3]. Examples of controlled drug substances include fentanyl, cathinone, amphetamine, and many others of which synthetic NPS seek to emulate [4]. These substances are thus regulated under law and their presence is heavily monitored. These synthetic drugs have diverse molecular structures by design, with the intent to exploit regulation loopholes and evade detection efforts [5,6]. Extensive efforts have been made to develop robust chemometric methods for different measurement techniques to better detect them.

There exist various analytical methods that can detect such compounds such as infrared spectroscopy [7,8], Raman spectroscopy [9–11], high-resolution liquid chromatography with tandem mass spectrometry (LC-MS-MS) [12,13] and many others [14–17]. However, GC-MS [18] remains one of the most widely used techniques to identify chemicals due to its ease of operation and ability to distinguish compound mixtures. GC-MS acquires both chromatographic retention time and fragmentation patterns during the electron ionization process. Conventionally, a compound is identified by matching the retention time and mass spectrum obtained against known records in a library through similarity matching algorithms [19]. This requires an extensive GC-MS database for effective identification, which for the case of NPS, is a challenging task as malicious actors create NPSs with ever increasing structural diversity to evade detection. Thus, current GC-MS database are unable to exhaustively cover the spectrum of all possible NPSs. Careful analysis and elucidation of the unknown NPS identity through possible combinations of molecular fragments is possible but it a non-trivial effort and is heavily dependent on the expert user analysing the spectrum. This presents an analytical challenge to identify unknown NPSs.

There exist modified library matching methods for mass spectra that involve a Hybrid Similarity Search algorithm which considers the difference between the nominal molecular mass of an unknown chemical compound and a potential target and shifts the original peaks accordingly [20]. Thereafter, the similarities and hence class of the unknown compound is inferred from likely candidates in an existing library. However, this approach requires knowledge of the nominal mass which may not be possible using only electron ionization (EI) based GC-MS measurements. Thus, it is necessary to study alternative methods that can attribute unknown NPSs to well-defined classes to reduce latency from drug seizure to identification.

Non-library matching methods generally involve two approaches. The first is a generative approach where synthetic GC-MS data is derived from theoretically predicted NPS molecular structures which then serves as references in a library for comparison [21]. Ji et al. also reports a related method where molecular fingerprints are directly predicted from machine learning models and a molecular identity is predicted from the fingerprint combination [22]. However, this approach would require both molecular structure and GC-MS predictions to be acceptably accurate for effective NPS identification. Furthermore, this method would not be able to exhaustively cover all possible NPSs due to resource limitations. Fragmentation pathways for substances of interest can also be predicted, as in the case of fentanyl analogues and synthetic opioids but such a work has not been extended to other NPS classes [23].

Another approach classifies unknown compounds based on statistical consideration or machine learning models trained on related compounds in the library. Recently, principal component analysis has been used in classification of fentanyl analogues based on infrared spectroscopy [7] as well as surface-enhanced Raman spectra [10]. Similarly, Esseiva et al has developed GC-MS based linear discriminant analysis and support vector machine models for distinguishing fiber from cannabis drug-type seedlings [24]. Recent work by Koshute et al have trained various machine learning models for the purpose of fentanyl analogue detection from mass spectra [15]. These chemometric methods have limited coverage, targeting specific subset of NPSs. For classification of a wider range of newly synthesized drugs, Jang et al. demonstrated use of artificial neural network to screen unknown erectile dysfunction drugs based on their LC-MS/MS spectra [12]. Combining it with a hybrid search algorithm, Lee et al. extended this neural network approach for NPS, also using high resolution LC-MS/MS spectra as inputs [13]. Thus, there does not exist a complete solution for classifying unknown NPS from GC-MS spectra.

Inspired by these recent works, we seek to develop a machine learning based model that can classify an unknown NPS to its drug class based solely on its GC-MS, with no knowledge of its nominal molecular mass. In this work, we trained and evaluated several supervised machine learning models based on mass spectra of known NPS. We also present the necessary pre-processing steps to prevent inflated prediction accuracies.

2. Materials and Methods

2.1 Dataset

NPS GC-MS data was obtained from the Scientific Working Group for Seized Drug [25] and Cayman Chemical Spectral Library [26] while GC-MS data of other non-related compounds with nominal mass < 600 was selected from the database distributed with the demonstration software version of NIST MS Search v2.3. The non-related compound class represents a diverse selection, with each chemical compound having only one GC-MS spectrum, to evaluate the model's ability in identifying compounds that do not belong to the NPS classes of interest. The maximum nominal mass allowed is set at 600 as the nominal mass for all NPS compounds in the dataset is < 600.

Distribution of the classes in the dataset is described in Table 1. 689 open-source GC-MS spectra from a selection of NPS classes and GC-MS of 202 other compounds were used to train and evaluate the machine learning models. A typical GC-MS mass spectrum is shown (Figure 1). Each spectrum of the dataset has been normalised against a maximum abundance of 100 for consistency between spectra. All GC-MS data comprises of relative abundance peaks plotted against the mass-to-charge ratio (m/z). All relative abundance of molecular fragments at mass-to-charge ratios from 1 to 600 are used as model inputs. Mass-to-charge ratios are rounded to the nearest integer for a total of 600 features with no further feature engineering performed. Mass-charge ratios with no abundance peaks are assigned a value of zero.

Classification is performed for 7 different classes and their analogues, namely, cathinones [27], cannabinoids [28], phenethylamines [29], piperazines [30], tryptamines [31], fentanyls [32] and a general mixture of non-related compounds. The dataset is manually split into the various classes according to their molecular structure. Choice of classes is to ensure sufficient population in each class for effective model training and hence classification. Chemical identities and their associated classes are provided in the dataset made available in the Supplementary Data. Data augmentation is performed on the dataset prior to training the model, similar to previous work by Skarysz et al. [33]. Augmentation is done by randomly scaling the original abundance peaks by 10% of their values and renormalizing them back to values relative to the largest peak intensity (given a value of 100), to mimic variations in GC-MS measurements. All training data is augmented tenfold.

2.2 Machine Learning Models and Evaluation Metrics

In this work, we constructed supervised machine learning models to classify unknown spectra according to 7 classes (cathinone analogues, cannabinoid analogues, phenethylamine analogues, piperazine analogues, tryptamine analogues, fentanyl analogues and other non-related compounds). Different machine learning models are trained and evaluated, namely artificial neural network (ANN) [34], convolutional neural network (CNN) [35] and balanced random forests models [36]. 4-fold cross validation is carried out for model evaluation, in which 75% of the total data is used for training the model while the remaining 25% is used for evaluation.

Model performance is evaluated based on recall, precision, and F1-scores. The performance metrics are calculated using the following equations:

$$\text{Recall} = \frac{TP}{TP + FN} \tag{1}$$

$$\text{Precision} = \frac{TP}{TP + FP} \tag{2}$$

$$\text{F1 score} = 2\frac{\text{Precision x Recall}}{\text{Precision} + \text{Recall}} \tag{3}$$

where:

True positive (TP): number of correct model predictions on the positive class,

False positive (FP): number of incorrect model predictions on the positive class,

True negative (TN): number of correct model predictions on the negative class,

False negative (FN): number of incorrect model predictions on the negative class.

2.3 Baseline comparison with classical library matching

We baselined the model performance against a library matching method using similarity scores obtained through simple match factor between two spectra, as employed in NIST Mass Spectral Search Program (MS Search). Hybrid match factor scores are not used as we assumed no knowledge of the molecular mass. The dataset and data splits from which the scores are calculated are identical to those used to train and evaluate the machine learning models.

The equation for the simple match factor [14], *sMF* is as follows:

$$sMF(x_1, x_2) = C_1 . \frac{(\sum_i a_{x1}[i] . a_{x2}[i])^2}{(\sum_i a_{x1}[i]^2) . (\sum_i a_{x2}[i]^2)} \tag{4}$$

where C_1 is defined as 999 according to conventions while x_1 and x_2 represents the two compared spectra and *a* is the peak intensity value at mass-charge ratio i. Similarity scores for the spectrum of the unknown compound in validation set is calculated against all spectra present in the training set. The unknown is then classified according to the majority class label present amongst the top ten compounds with the highest simple match factor scores. Only spectra with *sMF* > 700 are considered to exclude compounds that has low match scores with the query molecule. If there is no majority class label, the compound is labelled as unknown and is considered as an incorrect prediction.

3. Results and Discussion

The dataset includes spectra of NPS which are isomers of each other. Structural isomers, especially ring positional isomers, can give nearly identical results when measured under single energy EI GC-MS, if no additional methods such as low energy ionization [37] are used. Such isomer pairs (Figure A1) can easily be matched to each other through direct similarity measures. Thus, proper care should be taken when splitting GC-MS data into training/validation sets. We compare two different splitting methods here to demonstrate this effect. One method is isomer agnostic, where all samples in the dataset are treated equally when distributed between training and validation sets. A different method takes into consideration the presence of position isomers, ensuring that these isomers are all found in the same training or validation split. Baseline performance using our simple match factor strategy is compared (Table 2) between the two methods. Using recall averaged across all cross-validation folds as a performance metric, we observed that the performance of the simple match factor strategy for isomer agnostic data splits is much higher across all classes. Such a difference shows that the high structural similarity and hence GC-MS spectra between position isomers cannot be ignored as this results in an over estimation of the model's performance. To counter that, isomeric consideration is taken during data splitting for subsequent experiments.

Machine learning results shown here are based on three top performing models, Artificial Neural Network (ANN), Convolutional Neural Network (CNN) and Balanced Random Forests (BRF) are presented. BRF is used as there exists class imbalances in the dataset. Table 3 describes the F1 scores of the respective models. BRF model performs best, scoring highest across almost all classes. The receiver operating characteristic (ROC) [38] curves for the BRF model are described for all NPS classes (Figure 2). In the ROC graph, model performance across all classes is presented, with the vertical axis showing TP rate and the horizontal axis showing FP rate. The area under the ROC curve (AUC) values are also labelled. The closer the AUC to 1, the better the model performs in identifying the correct class while minimising false positives.

Compared to a simple match factor classification strategy, all machine learning models perform better for all classes. This demonstrates the effectiveness of a supervised machine learning approach in classifying unknown NPSs over a classical library matching method. BRF model, depending on the class, presents a 20% to 700% increase in recall scores compared to a simple match factor strategy (Figure 3). Compared to previous work which uses both LC-MS/MS and hybrid match factors [13], our model which is solely based on GC-MS mass spectra, albeit classifying for a smaller range of NPS classes, have greater model performance while having no knowledge of the molecular mass. Our classifier can also predict a wider range of NPS classes compared to a recent work which focuses only on fentanyl analogues [15]. Comparatively, our deep learning models have lower performance, likely due to the lower diversity in the minority classes such as piperazines.

To further investigate the model's ability to discern chemicals of interest from a larger population, we evaluated the trained BRF model against the remaining mass spectra (nominal weight < 600) present in the database distributed with NIST MS Search v2.3 (total of 2167 spectrum). The model has a high detection rate of 97 ± 1%, despite being trained on a much smaller dataset set size of 202 mass spectrum in the non-related compounds class. The results are slightly lower than that of Koshute et al [15], likely due to the larger number of classes being predicted for and the smaller class size being trained. This reflects the generalisability of the model in detecting compounds belonging to the NPS classes of interest.

4. Conclusion

In this work, we have shown that machine learning approaches enables effective classification of unknown NPS with BRF being the best performing model. The model's accuracy outperforms classical library matching. We have also shown that the model is able to discern the NPS types of interest from the larger chemical population, indicating that the model constructed can be used to detect NPS from unknown chemical mixtures. Additionally, our approach requires no feature engineering and takes in the complete distribution of relative abundance at the various mass-charge ratios of the GC-MS mass spectrum as input. Isomeric considerations are also taken into account to minimise inflated model performance scores. This facilitates easy integration with existing systems with minimal pre-processing performed on the measurement data.

As the model is trained on a single type of measurement, determining the exact molecular structure of an unknown substance is not straightforward due to limited information. A further improvement would be to predict the molecular structure from GC-MS data, but that would likely require inputs from complementary spectroscopic techniques such as Raman and infrared. These data inputs can be combined in the future to increase the predictive capabilities of the models. In addition, spectra obtained from samples with mixed composition of chemical compounds can be used to train the model, enhancing its robustness towards mixed samples. Furthermore, the retention time could potentially be included in the model as an input feature.

In all, our current work suggests a promising complement to existing methods for identifying unknown NPS not present in the library. The methodology developed is not limited to NPS but can be extended to other chemicals of interest.

Declaration of Competing Interest

The authors declare that there is no conflict of interests that can affect or influence the results and discussion in this work.

Acknowledgements

The authors acknowledge the Home Team Science and Technology Agency for financial support.

Supplementary Data

Supplementary data to the article can be found online. The python file describes the code and architecture of the various models used. The dataset used in this work, which contains mass spectrum from both SWGDRUG and MS Search and its description, is also provided.

References

[1] M.C. van Hout, A. Benschop, M. Bujalski, K. Dąbrowska, Z. Demetrovics, K. Felvinczi, E. Hearne, S. Henriques, Z. Kaló, G. Kamphausen, D. Korf, J.P. Silva, Ł. Wieczorek, B. Werse, Health and social problems associated with recent novel psychoactive substance (NPS) use amongst marginalised, nightlife and online users in six European countries, Int. J. Ment. Health Addict. 16 (2018) 480–495. https://doi.org/10.1007/s11469-017-9824-1.

[2] A. Peacock, R. Bruno, N. Gisev, L. Degenhardt, W. Hall, R. Sedefov, J. White, K. v Thomas, M. Farrell, P. Griffiths, New psychoactive substances: challenges for drug surveillance, control, and public health responses, The Lancet 394 (2019) 1668–1684. https://doi.org/10.1016/S0140-6736(19)32231-7.

[3] O.H. Drummer, Fatalities caused by novel opioids: a review, Forensic Sci. Res. 4 (2019) 95–110. https://doi.org/10.1080/20961790.2018.1460063.

[4] R. Graddy, M.E. Buresh, D.A. Rastegar, New and emerging illicit psychoactive substances, Med. Clin. North Am. 102 (2018) 697–714. https://doi.org/10.1016/j.mcna.2018.02.010.

[5] E. Underwood, A new drug war, Science (1979). 347 (2015) 469–473. https://doi.org/10.1126/science.347.6221.469.

[6] A. Shafi, A.J. Berry, H. Sumnall, D.M. Wood, D.K. Tracy, New psychoactive substances: a review and updates, Ther. Adv. Psychopharmacol. 10 (2020) 2045125320967197. https://doi.org/10.1177/2045125320967197.

[7] L.S.A. Pereira, F.L.C. Lisboa, J. Coelho Neto, F.N. Valladão, M.M. Sena, Screening method for rapid classification of psychoactive substances in illicit tablets using mid infrared spectroscopy and PLS-DA, Forensic Sci. Int. 288 (2018) 227–235. https://doi.org/10.1016/j.forsciint.2018.05.001.

[8] H.Z. Shirley Lee, H.B. Koh, S. Tan, B.J. Goh, R. Lim, J.L.W. Lim, T.W. Angeline Yap, Identification of closely related new psychoactive substances (NPS) using solid deposition gas-chromatography infra-red detection (GC–IRD) spectroscopy, Forensic Sci. Int. 299 (2019) 21–33. https://doi.org/10.1016/j.forsciint.2019.03.025.

[9] J. Omar, B. Slowikowski, C. Guillou, F. Reniero, M. Holland, A. Boix, Identification of new psychoactive substances (NPS) by Raman spectroscopy, J. Raman Spectrosc. 50 (2019) 41–51. https://doi.org/10.1002/jrs.5496.

[10] H. Muhamadali, A. Watt, Y. Xu, M. Chisanga, A. Subaihi, C. Jones, D.I. Ellis, O.B. Sutcliffe, R. Goodacre, Rapid Detection and Quantification of Novel Psychoactive Substances (NPS) Using Raman Spectroscopy and Surface-Enhanced Raman Scattering, Front. Chem. 7 (2019). https://doi.org/10.3389/fchem.2019.00412.

[11] A. Guirguis, S. Girotto, B. Berti, J.L. Stair, Identification of new psychoactive substances (NPS) using handheld Raman spectroscopy employing both 785 and 1064nm laser sources, Forensic Sci. Int. 273 (2017) 113–123. https://doi.org/10.1016/j.forsciint.2017.01.027.

[12] I. Jang, J.U. Lee, J.M. Lee, B.H. Kim, B. Moon, J. Hong, H. bin Oh, LC-MS/MS Software for screening unknown erectile dysfunction drugs and analogues: artificial neural network classification, peak-count scoring, simple similarity search, and hybrid similarity search

algorithms, Anal. Chem. 91 (2019) 9119–9128. https://doi.org/10.1021/acs.analchem.9b01643.

[13] S.Y. Lee, S.T. Lee, S. Suh, B.J. Ko, H. bin Oh, Revealing unknown controlled substances and new psychoactive substances using high-resolution LC–MS-MS machine learning models and the hybrid similarity search algorithm, J Anal. Toxicol. 46 (2021) 732-742. https://doi.org/10.1093/jat/bkab098.

[14] A.S. Moorthy, A.J. Kearsley, W.G. Mallard, W.E. Wallace, Mass spectral similarity mapping applied to fentanyl analogs, Forensic Chem. 19 (2020) 100379. https://doi.org/10.1016/j.forc.2020.100237.

[15] P. Koshute, N. Hagan, N.J. Jameson, Machine learning model for detecting fentanyl analogs from mass spectra, Forensic Chem. 27 (2022). https://doi.org/10.1016/j.forc.2021.100379.

[16] C.H. Wang, A.C. Terracciano, A.E. Masunov, M. Xu, S.S. Vasu, Accurate prediction of terahertz spectra of molecular crystals of fentanyl and its analogs, Sci. Rep. 11 (2021). https://doi.org/10.1038/s41598-021-83536-y.

[17] J.L. Bonetti, S. Samanipour, A.C. van Asten, Utilization of machine learning for the differentiation of positional NPS Isomers with direct analysis in real time mass spectrometry, Anal. Chem. 94 (2022) 5029–5040. https://doi.org/10.1021/acs.analchem.1c04985.

[18] R.M. Silverstein, G.C. Bassler, Spectrometric identification of organic compounds, J. Chem. Educ. 39 (1962) 546. https://doi.org/10.1021/ed039p546.

[19] A. Samokhin, K. Sotnezova, V. Lashin, I. Revelsky, Evaluation of mass spectral library search algorithms implemented in commercial software, J. Mass Spectrom. 50 (2015) 820–825. https://doi.org/10.1002/jms.3591.

[20] A.S. Moorthy, W.E. Wallace, A.J. Kearsley, D. v. Tchekhovskoi, S.E. Stein, Combining fragment-ion and neutral-loss matching during mass spectral library searching: A new general purpose algorithm applicable to illicit drug identification, Anal. Chem. 89 (2017) 13261–13268. https://doi.org/10.1021/acs.analchem.7b03320.

[21] M.A. Skinnider, F. Wang, D. Pasin, R. Greiner, L.J. Foster, P.W. Dalsgaard, D.S. Wishart, A deep generative model enables automated structure elucidation of novel psychoactive substances, Nat. Mach. Intell. 3 (2021) 973–984. https://doi.org/10.1038/s42256-021-00407-x.

[22] H. Ji, H. Deng, H. Lu, Z. Zhang, Predicting a molecular fingerprint from an electron ionization mass spectrum with deep neural networks, Anal. Chem. 92 (2020) 8649–8653. https://doi.org/10.1021/acs.analchem.0c01450.

[23] Q. Nan, W. Hejian, X. Ping, S. Baohua, Z. Junbo, D. Hongxiao, Q. Huosheng, S. Fenyun, S. Yan, Investigation of fragmentation pathways of fentanyl analogues and novel synthetic opioids by electron ionization high-resolution mass spectrometry and electrospray ionization high-resolution tandem mass spectrometry, J. Am. Soc. Mass Spectrom. 31 (2020) 277–291. https://doi.org/10.1021/jasms.9b00112.

[24] J. Broséus, F. Anglada, P. Esseiva, The differentiation of fibre- and drug type Cannabis seedlings by gas chromatography/mass spectrometry and chemometric tools, Forensic Sci. Int. 200 (2010) 87–92. https://doi.org/10.1016/j.forsciint.2010.03.034.

[25] Scientific Working Group for the Analysis of Seized Drugs. SWGDRUG Mass Spectral Library. Version 3.11. URL https://www.swgdrug.org/ms.htm.

[26] Cayman Chemical (2020) Version CaymanSpectralLibrary_v09222020. URL https://www.caymanchem.com/forensics/publications/csl.

[27] C.L. German, A.E. Fleckenstein, G.R. Hanson, Bath salts and synthetic cathinones: An emerging designer drug phenomenon, Life Sci. 97 (2014) 2–8. https://doi.org/10.1016/j.lfs.2013.07.023.

[28] M.S. Castaneto, D.A. Gorelick, N.A. Desrosiers, R.L. Hartman, S. Pirard, M.A. Huestis, Synthetic cannabinoids: Epidemiology, pharmacodynamics, and clinical implications, Drug Alcohol Depend. 144 (2014) 12–41. https://doi.org/10.1016/j.drugalcdep.2014.08.005.

[29] L.A. King, New phenethylamines in Europe, Drug Test Anal. 6 (2014) 808–818. https://doi.org/10.1002/dta.1570.

[30] S. Elliott, Current awareness of piperazines: pharmacology and toxicology, Drug Test Anal. 3 (2011) 430–438. https://doi.org/10.1002/dta.307.

[31] A.M. Araújo, F. Carvalho, M. de L. Bastos, P. Guedes de Pinho, M. Carvalho, The hallucinogenic world of tryptamines: an updated review, Arch. Toxicol. 89 (2015) 1151–1173. https://doi.org/10.1007/s00204-015-1513-x.

[32] P. Armenian, K.T. Vo, J. Barr-Walker, K.L. Lynch, Fentanyl, fentanyl analogs and novel synthetic opioids: A comprehensive review, Neuropharmacology. 134 (2018) 121–132. https://doi.org/10.1016/j.neuropharm.2017.10.016.

[33] A. Skarysz, Y. Alkhalifah, K. Darnley, M. Eddleston, Y. Hu, D.B. McLaren, W.H. Nailon, D. Salman, M. Sykora, C.L.P. Thomas, A. Soltoggio, Convolutional neural networks for automated targeted analysis of raw gas chromatography-mass spectrometry data, 2018 International Joint Conference on Neural Networks (IJCNN) (2018) pp. 1–8. https://doi.org/10.1109/IJCNN.2018.8489539.

[34] O.I. Abiodun, A. Jantan, A.E. Omolara, K.V. Dada, N.A. Mohamed, H. Arshad, State-of-the-art in artificial neural network applications: A survey, Heliyon. 4 (2018) e00938. https://doi.org/10.1016/j.heliyon.2018.e00938.

[35] Z. Li, F. Liu, W. Yang, S. Peng, J. Zhou, A survey of convolutional neural networks: Analysis, applications, and prospects, IEEE Trans. Neural. Netw. Learn Syst. (2021) 1–21. https://doi.org/10.1109/TNNLS.2021.3084827.

[36] G. Lemaître, F. Nogueira, C.K. Aridas, Imbalanced-learn: A python toolbox to tackle the curse of imbalanced datasets in machine learning, J. Mach. Learn. Res. 18 (2017) 559–563.

[37] R.F. Kranenburg, D. Peroni, S. Affourtit, J.A. Westerhuis, A.K. Smilde, A.C. van Asten, Revealing hidden information in GC–MS spectra from isomeric drugs: Chemometrics based identification from 15 eV and 70 eV EI mass spectra, Forensic Chem. 18 (2020) 100225. https://doi.org/10.1016/j.forc.2020.100225.

[38] T. Fawcett, An introduction to ROC analysis, Pattern Recognit. Lett. 27 (2006) 861–874. https://doi.org/10.1016/j.patrec.2005.10.010.

Table 1
Dataset distribution.

Class	Population	Species
Cathinone analogues	89	88
Cannabinoid analogues	149	149
Phenethylamine analogues	112	112
Piperazine analogues	50	48
Tryptamine analogues	67	67
Fentanyl analogues	222	222
Other compounds	202	202

.

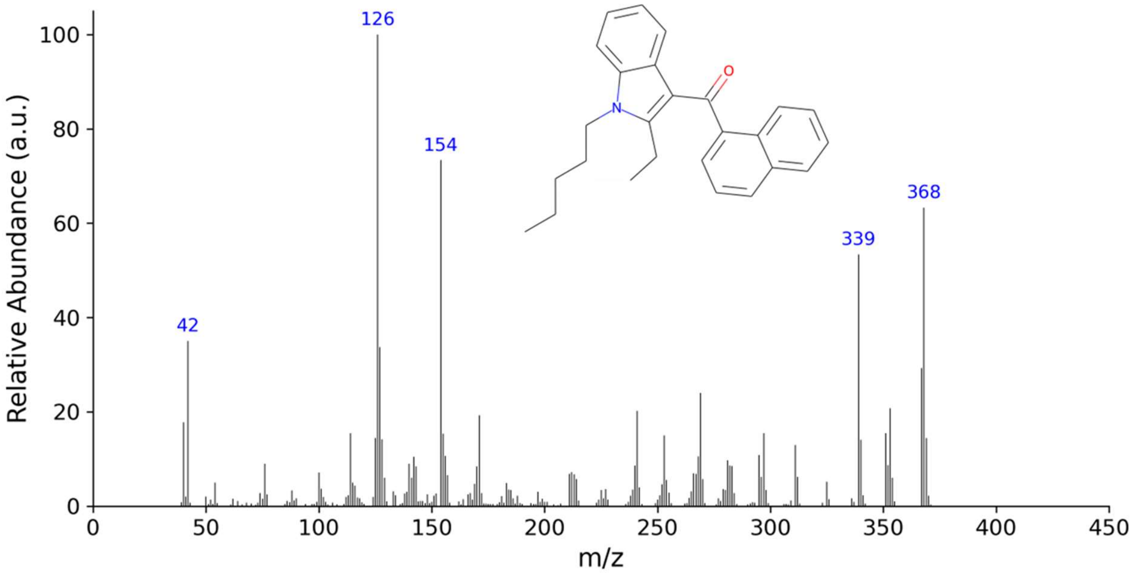


Figure 1. GC-MS spectrum of a cannabinoid-analogue NPS, JWH-116 (structure shown in inset). Blue numbers represent mass of molecular fragments which are more abundant.

Table 2
Recall scores of a simple match factor strategy.

	Mean Recall	
Class	Isomer Agnostic	Isomers in Same Split
Cathinone analogues	0.751	0.650
Cannabinoid analogues	0.375	0.377
Phenethylamine analogues	0.616	0.348
Piperazine analogues	0.599	0.082
Tryptamine analogues	0.627	0.540
Fentanyl analogues	0.738	0.504
Other compounds	0.153	0.149

Table 3
Comparison between different machine learning models.

	F1 Score		
Class	ANN	CNN	BRF
Cathinone analogues	0.786	**0.886**	0.873
Cannabinoid analogues	0.857	0.867	**0.930**
Phenethylamine analogues	0.726	0.843	**0.866**
Piperazine analogues	0.414	0.489	**0.852**
Tryptamine analogues	0.820	0.886	**0.899**
Fentanyl analogues	0.920	0.949	**0.980**
Other compounds	0.754	0.842	**0.901**

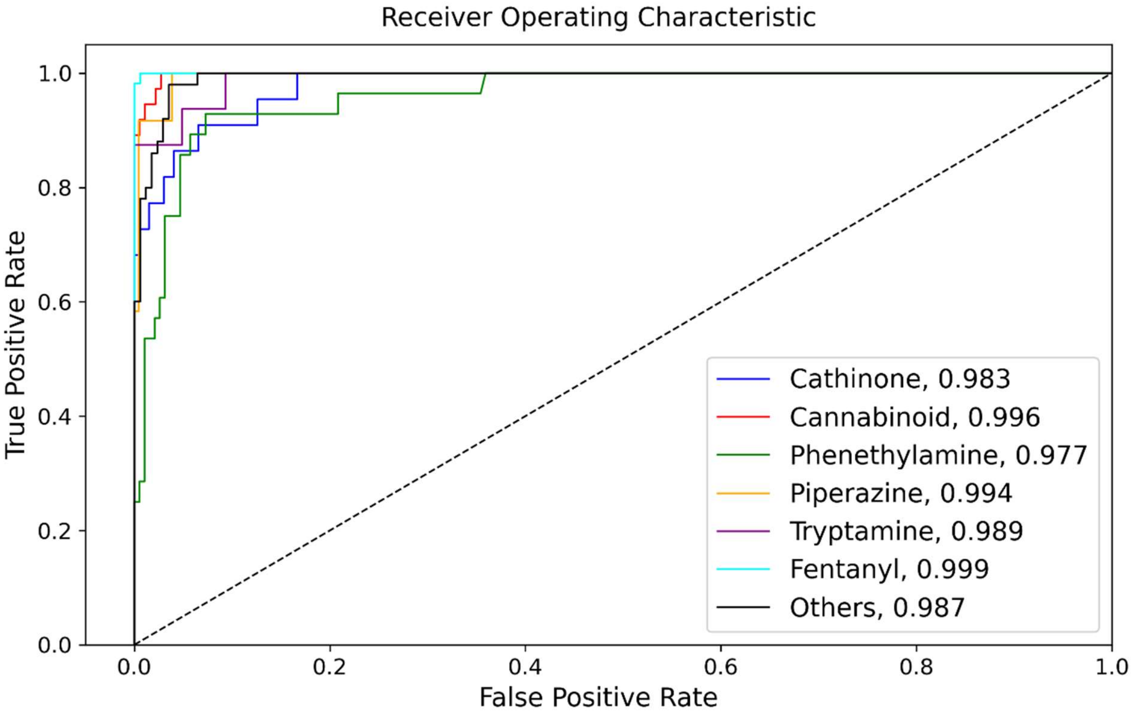


Figure 2. ROC of BRF model for single cross-validation fold. Legend shows average AUC values.

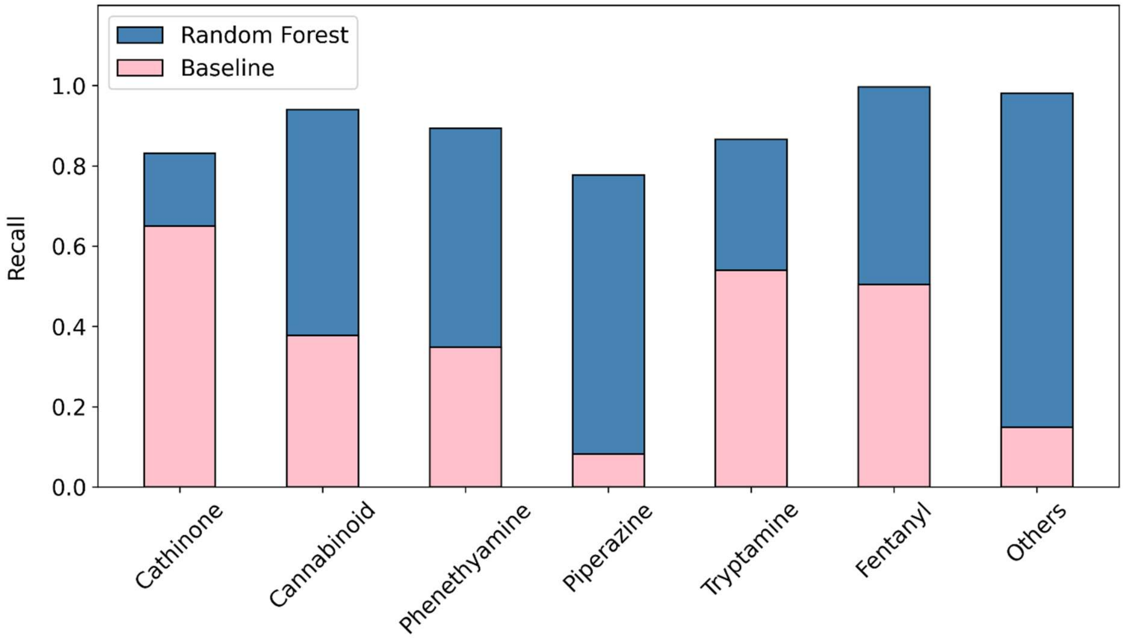


Figure 3. Recall scores of BRF model versus baseline simple match factor strategy.

Appendix

A. GC-MS Mass Spectrum of an Isomer Pair

A representative example of isomers with highly similar mass spectra. Mass spectra belonging to the pair of cathinone analogues shown in Figure A1 are almost similar with a high simple match factor score of 982.9 when calculated using Equation 4.

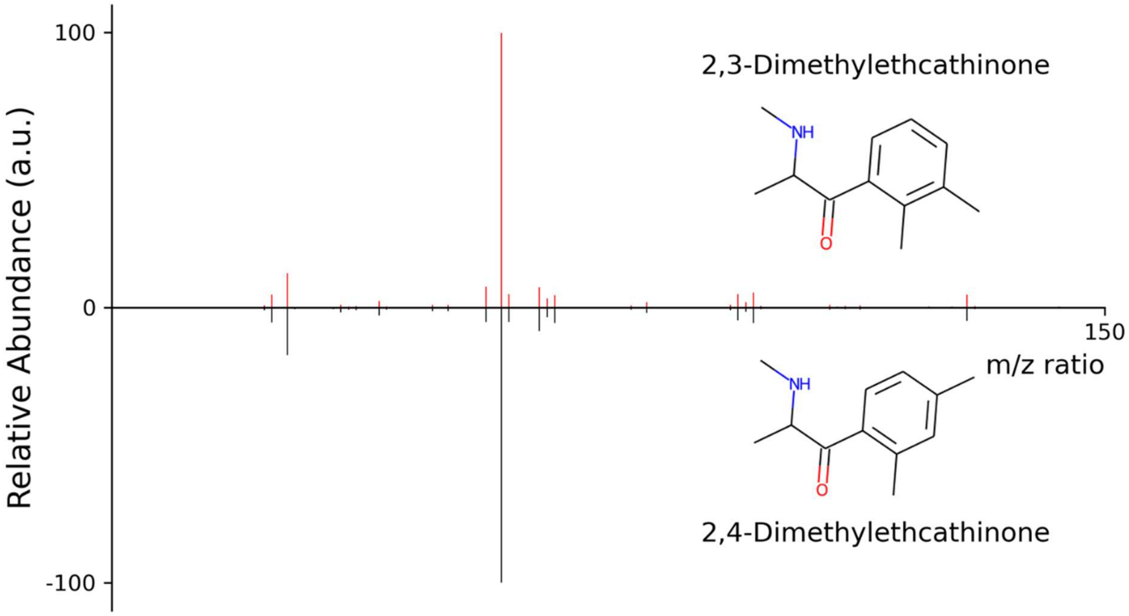


Figure A1. Mass spectrum of two position isomers, showing their similarity.

B. Machine Learning Models

Balanced Random Forests (BRF) model consist of multiple decision trees which votes and classifies the object based on a random choice of feature inputs, in this case the GC-MS mass peak abundance values. BRF model used in this work is constructed using the Python *imbalanced-learn* package [36]. ANN and CNN models are trained using the Python *keras* package.

Table A1
Hyperparameters used for the ML models

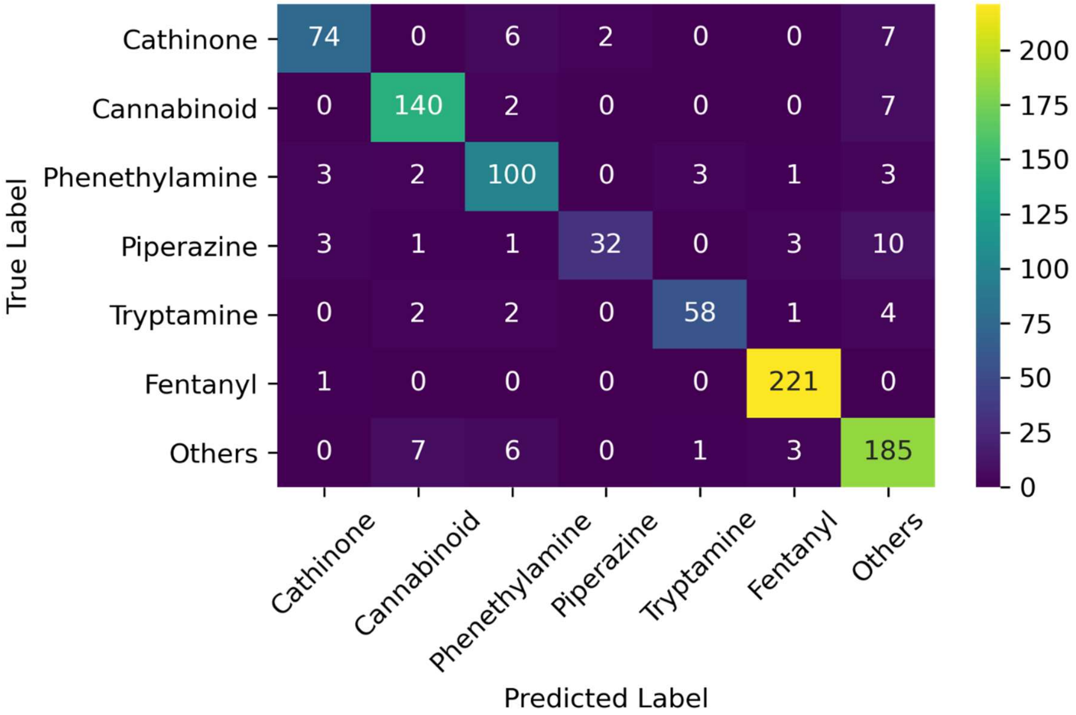
Model	Hyperparameters used

Balanced Random Forest	*maximum depth* = 50;
	maximum features = 0.1;
	no. of estimators = 1000;
	criterion = 'gini'

Artificial Neural Network	*no. of densely connected layers* = 3;
	nodes per layer = [300, 30, 7];
	optimizer = 'Adam';
	loss = categorical cross entropy

Convolutional Neural Network	*no. of convolutional 1D layers* = 3;
	kernel size = 3;
	padding = 'valid';
	channels per convolution layer = [16, 32, 64];
	no. of densely connected layers = 3;
	nodes per layer = [480, 48, 7];
	optimizer = 'Adam';
	loss = categorical cross entropy

C. Predictions of the Balanced Random Forest Model



True Label \ Predicted Label	Cathinone	Cannabinoid	Phenethylamine	Piperazine	Tryptamine	Fentanyl	Others
Cathinone	74	0	6	2	0	0	7
Cannabinoid	0	140	2	0	0	0	7
Phenethylamine	3	2	100	0	3	1	3
Piperazine	3	1	1	32	0	3	10
Tryptamine	0	2	2	0	58	1	4
Fentanyl	1	0	0	0	0	221	0
Others	0	7	6	0	1	3	185

Figure A2. Confusion matrix describing predictions of the Balanced Random Forest model across all cross-validation folds.