Forensic Chemistry 27 (2022) 100379

Contents lists available at ScienceDirect

Forensic Chemistry

journal homepage: www.sciencedirect.com/journal/forensic-chemistry

Machine learning model for detecting fentanyl analogs from mass spectra

Phillip Koshute [*], Nathan Hagan, N. Jordan Jameson

Johns Hopkins University Applied Physics Laboratory, 11100 Johns Hopkins Road, Laurel MD 20723, USA

A R T I C L E I N F O

Keywords:
Artificial intelligence
Library matching
Nested cross-validation
Neural network
Random forest
Spectral patterns

A B S T R A C T

In recent years, fentanyl and its analogs have been increasingly abused, leading to tragic outcomes. One way to tackle this problem is to rapidly detect fentanyl analog compounds and prevent them from reaching people who could be harmed by them. However, the emergence of novel fentanyl analogs that evade detection by classic mass spectral library matching has exacerbated the problem. We propose supervised machine learning classification models as a complementary approach to library matching for detecting fentanyl analogs from mass spectra. To develop and apply such models, we extract two dozen peak-based and similarity-based input features from each spectrum of interest. Using techniques such as random forests, neural networks, and logistic regression, we identify patterns within these features' values, resulting in strong detection performance. Within a cross-validation framework, we achieve 99% probability of detection and 1% probability of false alarm on a representative set of several thousand mass spectra. These results suggest that machine learning models may offer a robust complement to library matching. Practitioners from diverse fields, including border security, law enforcement, and military may benefit from this capability to detect drugs of abuse.

1. Introduction

In 1960, researchers developed a synthetic opioid known as fentanyl for use as an intraoperative analgesic that is 100 times more potent than morphine [1]. In recent years, fentanyl and its analogs have been increasingly abused outside of conventional medical settings [2]. Such misuse has tragic effects, frequently leading to deaths due to overdosing [3]. An important aspect to halting this problem is improving existing capabilities to detect and dispose fentanyl analogs [4]. However, detecting fentanyl analogs is increasingly challenging.

Typically, investigators collect spectra of unknown substances via sensing technologies [5] such as mass spectrometry [6], Raman spectroscopy [7], and infrared spectroscopy [8]. They seek to identify those substances through spectral library matching [9]. This approach is effective at identifying substances with spectra already in the library, but it is much less effective for variants of these substances for which library entries do not already exist [10].

While most libraries contain fentanyl and its most common analogs, malicious actors are developing new analogs at an accelerating pace [11]. Motivated by the opportunity to exploit regulatory loopholes [12], malicious actors develop, sell, and otherwise distribute novel fentanyl analogs not represented in available spectral libraries and therein such compounds can evade state-of-the-art sensors.

Recent efforts have offered potential enhancements to classic library matching. Moorthy et al. [13] developed a Hybrid Similarity Search algorithm that considers both observed fragment ions and inferred neutral losses to match a given spectrum with a "nearest-neighbor" library spectrum that differs by a chemical substitution. This approach has been employed for detection of and mapping similarities between novel fentanyl-related compounds [14]. However, it is limited because it requires knowing or estimating each compound's nominal mass, which may not be possible using the electron ionization (EI) mass spectra of fentanyl analogs. In another development, Nan et al. [15] systematically studied the common fragmentation pathways that fentanyl-related compounds undergo during EI mass spectrometry. These results could be utilized in future models, e.g., with rules based on peak locations; to our knowledge, though, these results have not yet been implemented in any formal fentanyl detection platform.

To address the ongoing gap in fentanyl analog detection capability, we developed and evaluated supervised machine learning models that extract patterns from mass spectra of known fentanyl analogs and enable more robust detection of previously unknown fentanyl analogs. Broadly, supervised machine learning is the process of identifying patterns within known samples to aid the prediction or classification of unknown samples [16]. Within this work, we used machine learning techniques to find patterns within the characteristics of spectra collected from known

Abbreviations: PD, Probability of detection; PFA, Probability of false alarm; AUROC, Area under the receiver operator characteristic curve.

[*] Corresponding author.

E-mail address: Phillip.Koshute@jhuapl.edu (P. Koshute).

https://doi.org/10.1016/j.forc.2021.100379
Received 2 September 2021; Received in revised form 3 November 2021; Accepted 9 November 2021
Available online 14 November 2021

fentanyl analogs or known non-fentanyl substances. The resulting models classify the spectra from unknown substances.

Mass spectrometry combined with gas chromatography yields spectra of individual compounds within a mixture [17]. Moreover, the importance of mass spectrometry within laboratory analysis has been increasing [18]. For these reasons, we focus our work on models for detecting fentanyl analogs within mass spectra. Given the advantages of Raman and infrared technologies for portable sensing, we anticipate extending this approach to these technologies in future work.

Machine learning has been applied to mass spectrometry for various applications. Lin et al. [19] used random forests to discriminate between mass spectra from blood with and without metabolic syndrome. Jang et al. [20] developed a neural network to classify mass spectra into three classes of erectile dysfunction (ED) drugs and one class of non-ED drugs. Davidson et al. [21] used linear discriminant analysis (LDA) and random forests to classify animal fecal pellets by species, gender, age, and species strain. Other applications include metabolomics [22], cancer diagnosis [23], and forensics [24]. Wei et al. [25] have used neural networks to predict EI mass spectra from known molecular structures.

Machine learning and other well-established statistical techniques have also been applied to fentanyl analog detection using various sensing technologies. For instance, Xu et al. [26] used principal component analysis (PCA) and logistic regression to classify infrared spectra according to whether they corresponded to fentanyl-related functional groups. Wang et al. [27] similarly used PCA and partial least squares discriminant analysis to distinguish fentanyl from morphine analogs from surface-enhanced Raman spectra. However, only Bonetti [28] has attempted to apply machine learning to fentanyl detection with mass spectra. This research involved PCA and LDA for distinguishing between isomers of two specific fentanyl analogs. Our models, which distinguish between fentanyl analogs and non-fentanyl substances, are considerably more general. We believe that our approach is the first to use supervised machine learning models to undertake this general fentanyl analog detection problem for mass spectra.

2. Material and methods

We followed best practices of machine learning and statistics throughout our work (Fig. 1). We used already collected spectra to develop and assess our detection models. From each spectrum, we extracted features that we used as inputs for the models. We separated the resulting data points into disjoint subsets or “folds,” iteratively using some of these folds for model training and holding out others to evaluate model performance. Within such a cross-validation framework, we were able to characterize our models’ likely performance against spectra that are unknown or otherwise not included in model training. Furthermore, we implemented a classic library matching approach and applied it to the same data sets in order to assess our model’s performance in comparison to widely employed library matching techniques. These steps are further described in the following subsections.

2.1. Data

We obtained 3718 EI mass spectra from the National Institute of Standards and Technology (using the collection of libraries distributed with the NIST05 MS Search demonstration software) and the Scientific Working Group for Seized Drugs [29]. This set includes 195 spectra of fentanyl analogs such as fentanyl, carfentanil, and sufentanil. We categorized the remaining 3523 spectra as non-fentanyl substances. These other substances included other drugs (including other opioids), environmental contaminants, toxic compounds, and common organic small molecules with a variety of functional groups. This set of substances was not compiled specifically for this application, but it represents a diverse sample of substances that might be encountered during relevant investigations. Each spectrum corresponds to a unique substance.

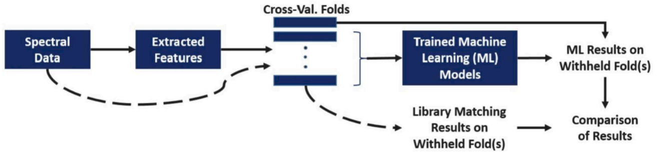
Cross-Val. Folds
Spectral Data
Extracted Features
Trained Machine Learning (ML) Models
ML Results on Withheld Fold(s)
Library Matching Results on Withheld Fold(s)
Comparison of Results

Fig. 1. Process Overview for Model Development and Assessment.

In some cases, our spectra include both a given substance and one of its deuterated analogs (in a deuterated analog, some number of hydrogen atoms have been replaced by deuterium atoms). While the spectra of such pairs of substances are similar, they are not easily matched to one another with simple library matching due to the offset between their peaks. Therefore, we opted to regard such pairs as unique substances and retained them for training and evaluating our methods.

These mass spectra that we obtained are comprised of peaks expressed in terms of two values: mass-to-charge ratio (m/z) and intensity. Therefore, we did not implement any smoothing or peak-finding steps for these spectra, performed usually as first steps when processing raw data directly from the sensor. The number of peaks in a given spectrum within this data set varies from 4 to 394. For each spectrum, we normalized the intensity values such that the maximum intensity for an individual peak in each spectrum was 1000. For consistency with common peak file formats (e.g., as with those from SWGDRUG), we reported intensity values for other peaks as integers.

Fig. 2 compares the mass spectra of fentanyl with acetyl fentanyl (a fentanyl analog) and with acetaminophen (a non-fentanyl substance). Visually, the mass spectrum of acetyl fentanyl is much more similar to the spectrum of fentanyl (e.g., with coinciding peaks at m/z 150 and the slightly shifted peaks near 250) than the spectrum of acetaminophen. Our work aims to automate the process by which spectral similarities (which may be visually evident as in this example or may be subtler) are used to link previously unseen spectra with fentanyl.

2.2. Feature extraction

Machine learning models identify patterns within characteristics of each spectrum; these characteristics are commonly called “features.” One option is to use the intensities from each m/z value as a separate features. We instead extracted twenty-four high-level features as a more efficient and potentially more insightful way of representing each mass spectrum (Table 1 and Table 2). Broadly, our features describe a spectrum’s peaks or its similarity to spectra of known substances.

Examples of peak-based features include the highest mass-to-charge ratio, the mass-to-charge ratio of the peak with the highest intensity (i.e., the base peak), and several measures of the variability across peaks.

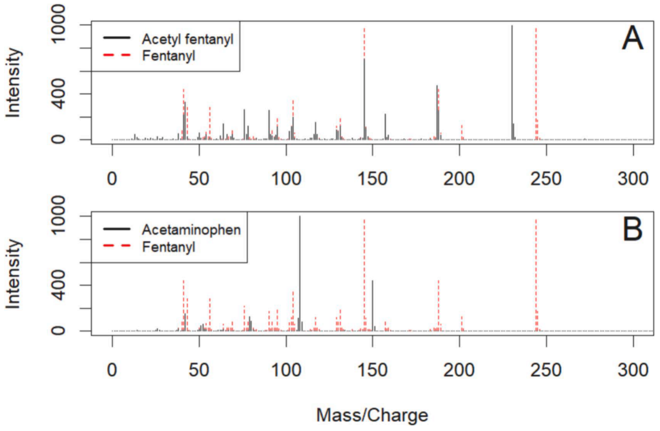

Fig. 2. Comparison of Spectra for (A) Fentanyl with a Fentanyl Analog (acetyl fentanyl) and (B) Fentanyl with a Non-Fentanyl Substance (acetaminophen).

Table 1
Peak-Related Machine Learning Features for Detection Models.

Feature	Description
Base Peak Mass	Mass-to-charge ratio (*m/z*) of the peak with the greatest intensity (i.e., "base peak")
Base Peak Mass Proximity	*m/z* difference between base peak and the peak with the nearest *m/z* value
Maximum Mass	*m/z* of the peak with the greatest mass
Maximum Mass Proximity	*m/z* difference between maximum-mass peak and the peak with the nearest *m/z* value
Number of Peaks	Total number of reported peaks in the spectrum
Intensity Mean	Mean of all peaks' intensity values
Intensity Standard Deviation	Standard deviation of all peaks' intensity values
Intensity Density	Maximum intensity divided by number of peaks
Mass Mean	Mean of all peaks' *m/z*
Mass Standard Deviation	Standard deviation of all peaks' *m/z*
Mass Density	Maximum mass divided by number of peaks
Most Frequent Pair Peakwise Mass Difference (PPMD)	Most frequent *m/z* difference among all pairs of peaks
Mean PPMD	Mean of *m/z* differences among all pairs of peaks

Table 2
Similarity-Related Machine Learning Features for Detection Models.

Feature	Description
Similarity to Alfentanil	Cosine similarity with alfentanil spectrum
Similarity to Carfentanil	Cosine similarity with carfentanil spectrum
Similarity to Fentanyl	Cosine similarity with fentanyl spectrum
Similarity to Remifentanil	Cosine similarity with remifentanil spectrum
Similarity to Sufentanil	Cosine similarity with sufentanil spectrum
Peakwise Similarity to Alfentanil	Cosine similarity with alfentanil's distribution of peak pairwise mass differences (PPMD)
Peakwise Similarity to Carfentanil	Cosine similarity with carfentanil's distribution of PPMD
Peakwise Similarity to Fentanyl	Cosine similarity with fentanyl's distribution of PPMD
Peakwise Similarity to Remifentanil	Cosine similarity with remifentanil's distribution of PPMD
Peakwise Similarity to Sufentanil	Cosine similarity with sufentanil's distribution of PPMD
Library Matching Score	Maximum score from library matching (maximum modified cosine similarity with the spectrum of any fentanyl analog in the library)

We also tracked the difference in mass between all pairs of peaks of a given spectrum, referring to them as "peak pairwise mass differences" (PPMD). Conceptually, these differences may correspond to the masses of neutral fragments dissociated from the intact molecular ion during mass spectral acquisition.

Similarity-based features include comparisons of raw spectra and of PPMD distributions. We compared each spectrum with the spectra of alfentanil, carfentanil, fentanyl, remifentanil, and sufentanil. We selected these five compounds because they are widely known and likely to be in most reference libraries. To quantify the similarity between raw spectra or PPMD distributions, we used cosine similarity, which is the cosine of the angle between the vectors representing each spectrum. This metric is also commonly used in library matching approaches. Numerically, the cosine similarity (CS) is calculated as

$$CS = (\sum_i u_i v_i)/(\sqrt{\sum_i {u_i}^2}\sqrt{\sum_i {v_i}^2}) \tag{1}$$

Cosine similarity can only be calculated on vectors of the same length (i.e., with the same number of elements). Thus, to calculate cosine similarity, we "padded" each spectrum's vector with zeros at non-peak masses so that each spectrum had intensity values for every mass-to-charge value and thus all vectors had the same length.

As an additional similarity-based feature, we also used the output score from the library matching method. This value reflects the similarity between a given spectrum and any known fentanyl analog spectrum using a slightly different similarity measure known as the simple match factor or modified cosine similarity (MCS). The MCS is obtained from the CS by taking the square root of each intensity and squaring the resulting quotient. Taking the square root reduces the effect of variation in intensity measurements between sensors. This metric has also been used in other recent fentanyl detection research [13]. The MCS is calculated as

$$MCS = \left(\sum_i \sqrt{u_i v_i}\right)^2/(\sum_i u_i \sum_i v_i \tag{2}$$

We selected these features based on both structural and statistical considerations. For instance, the maximum mass feature of a spectrum often corresponds to the nominal mass of the underlying fentanyl analog and mass differences between spectral peaks may correspond to the masses of neutral fragments. Similarly, the proximity features may provide insights into characteristics fragmentation behavior. We included various statistical measures such as mean and standard deviation as alternative ways to characterize spectra and potentially identify patterns between them.

Table 1 and Table 2 show the full set of peak-related and similarity-related features, respectively, that we used for our machine learning models. We used this full set of features as input to our machine learning models. In order to focus on whether machine learning might generally provide an alternative detection capability, we opted not to undertake any feature selection analysis.

2.3. *Model construction and evaluation*

Supervised machine learning involves identifying patterns in available data for which the outcome is known and applying those patterns to make predictions about new data for which the outcome is not known. (In contrast, unsupervised machine learning involves identifying patterns without knowing any outcomes.) Machine learning models may be used to predict continuous-valued responses or discrete classes, prompting the use of regression or classification models respectively. For this study, we sought to classify unknown spectra either as fentanyl analogs or non-fentanyl substances. Thus, we constructed and evaluated binary classification models.

We evaluated each model according to a 10-fold cross-validation scheme. In this scheme, all known spectra are partitioned into ten disjoint subsets or "folds," each with the same distribution of fentanyl analogs and non-fentanyl substances present in the full data set. Nine of the ten folds are included in a training data set that is used to identify patterns, i.e., "train" the model. The remaining fold is called the "test data set" or "test fold." The resulting model is evaluated on the "test fold," yielding a model output score for each spectrum in the test fold. If the model's score for a spectrum exceeds the given model's detection threshold, the model "predicts" or "classifies" that spectrum as a fentanyl analog. This process is repeated for all folds, such that each fold is withheld from the training data and used to evaluate the resulting model exactly once. Since each spectrum has been assigned to exactly one fold, this also means that each spectrum is withheld from the training data and used to evaluate the resulting model exactly once. Aside from stratification by fentanyl status, we selected the folds randomly.

By comparing each spectrum's predicted class with its actual class, we calculated several model performance metrics. Probability of detection (PD) is the proportion of fentanyl analogs that a model correctly classifies as fentanyl analogs. Probability of false alarm (PFA) is the proportion of non-fentanyl substances that a model incorrectly classifies as fentanyl analogs. Because PD only involves fentanyl analogs and PFA only involves non-fentanyl substances, they are not affected by imbalanced data sets such as ours. A model's PD and PFA are closely linked to its detection threshold. Lower thresholds will lead more spectra to be classified as fentanyl analogs, which generally prompts increases in both PD and PFA. To assess the model's capability independent of a specific detection threshold, we also constructed a receiver operator characteristic (ROC) curve by iteratively varying the detection threshold and plotting the resulting PD and PFA on a coordinate plane's y- and x-axes respectively. The area on the plane between $x = 0$, $x = 1$, $y = 0$, and the ROC curve is known as the area under the ROC curve (AUROC). The best AUROC values are generally close to 1 [30].

Since the spectra from a given test fold were not included in training the model upon which they were evaluated, the models' performance against these spectra provides an estimate of likely AUROC values against unseen spectra for a model trained with this approach. Additionally, we are able to estimate the "calibrated PD" that is achievable while maintaining no more than a fixed PFA, using model scores from test set spectra to determine the necessary detection threshold. For this, we set the PFA limit at 1%. This approach, involving a "test-set guided threshold," is akin to "constant false alarm detectors" used within radar surveillance [31].

We carried out this cross-validation and PD calibration on three types of machine learning models (logistic regression, shallow neural networks, and random forests), reporting AUROC and calibrated PD for each model type. Logistic regression involves generalized linear models with binomial (i.e., two-class) responses [32]. Neural networks are nonlinear weighted sums of input feature values, represented by layers of nodes connected by edges [33]. Unlike deep neural networks that may include tens or hundreds of layers, we considered shallow two-layer neural networks, which have less demanding computational requirements and are less prone to overfitting. Random forests are ensembles of decision trees, each constructed with a statistically strategic random subset of the full set of available training data [34]. We selected these three techniques because they provide a representative sample of the most prominent types of supervised machine learning classification models. We describe these machine learning techniques in further detail in Appendix A.

To account for the random components of our model training and cross-validation, we conducted multiple rounds of cross-validation. Randomness occurs during model training for neural networks (in selecting an initial set of edge weights) and for random forests (in selecting subsets of training data for each decision tree). Additionally, we randomly separated all available data into folds as part of cross-validation. In this work, we conducted 50 rounds of 10-fold cross-validation; for all metrics, we reported the mean and the empirical 95% confidence interval as reported by the *quantile* function in the *stats* package in R [35]. We opted for 50 rounds as a balance between exploring many instances of our model training process's random components and reasonably limiting computational demands.

Notably, even with multiple rounds of cross-validation, this PD calibration process does not yield an unbiased estimate of the PD and PFA achievable in a real-world deployment setting; we incur bias by reporting results on the same data that we use to determine the detection threshold. To eliminate this bias, we also carried out nested cross-validation according to the following process. By nesting a second cross-validation loop inside of a typical cross-validation loop, we select a best model type and a corresponding detection threshold without incorporating the held-out test fold at all. In this way, our nested cross-validation process mimics a deployment scenario in which a best model and detection threshold would necessarily be selected prior to assessing unknown spectra collected in the field. We describe this nested cross-validation process in further detail in Appendix B. To account for randomness in how spectra are separated into folds and models are constructed, we conducted 50 rounds of this nested cross-validation process.

2.4. *Feature importance analysis*

To gain some insight into which features drive model performance, we performed brief feature analysis. We constructed random forest models using all available mass spectra (withholding only five prominent fentanyl analogs) using both the full set of features and an abbreviated set without the five similarity and five peakwise similarity features. We tracked each model's calibrated PD values from "out-of-bag" (OOB) predictions. For a given spectrum, OOB predictions are only made on trees for which that spectrum was not part of the tree construction. To account for the random components in how the random forest models are constructed, we repeated this step 50 times and calculated the mean calibrated PD.

Additionally, we calculated the mean permutation importance for each feature across these 50 rounds. Permutation importance is the amount by which OOB accuracy decreases when a given feature's values are permuted. We obtained these values directly from the *randomForest* output.

2.5. *Comparison with classic library matching*

Traditionally, fentanyl and other substances of interest have been detected via library matching. Accordingly, we assessed the performance of our machine learning models by comparing their results with results achieved with simple library matching.

Simple library matching involves comparing an unknown spectrum with each spectrum within a reference library. If the unknown spectrum has a "close enough" match with the spectrum reference library, the unknown substance has been "detected."

Typically, the degree of similarity between the unknown spectra and a given reference spectrum is quantified with a similarity metric such as the CS or MCS. Consistent with others' work, we have used the MCS for library matching [13], which also may be called the simple match factor. (In contrast, we used the CS to compare each spectrum with prominent fentanyl analogs as part of our feature extraction, as described above.) As with cosine similarity, MCS can only be calculated on vectors of the same length; therefore, we "padded" each spectrum's vector with zeros at non-peak m/z values prior to calculating MCS. The MCS between two vectors u and v is given by Equation (2) above.

In practice, a substance of interest is regarded as having been detected if the MCS between an unknown spectrum and that substance's reference spectrum exceeds some score threshold (e.g., 0.75). Within this work, we classify an unknown spectrum as a fentanyl analog if its similarity score with the spectrum of any known fentanyl analog within the reference library exceeds the detection threshold. This library matching classification is not influenced by the similarity scores between the unknown spectrum and spectra of known non-fentanyl substances.

As shown in Table 3, we assessed four different versions of simple library matching. For general surveillance, an investigator might use a "standard" reference library with only a small number of fentanyl analogs. We assumed that the fentanyl analogs included in such a library would be alfentanil, carfentanil, fentanyl, remifentanil, and sufentanil. Rather than always holding the detection threshold constant at 0.75, we also allowed the threshold to vary in order to achieve the maximum PD, while still limiting the PFA to no more than 1%. When paired with the standard reference library of five prominent fentanyl analogs, we refer to this approach as "adaptive library matching." We also implemented the library matching technique with an expanded or "targeted" reference library including all available fentanyl analogs (up to 176 within a 10-fold cross-validation scheme), both with and without variable thresholds; we refer to these approaches as "targeted library matching" and "enhanced library matching" respectively. These four library matching approaches are summarized in Table 3. For even comparison, we used the same fentanyl analogs for training machine learning models and as reference for library matching.

In the following section, we compare these library matching approaches with each other and with our machine learning models.

Table 3
Library Matching Approaches.

Approach	Number of Fentanyl Analogs in Reference Library	Threshold
Standard	5	Fixed at 0.75
Adaptive	5	Varied to maximize PD, while limiting PFA to 1% or less
Targeted	176	Fixed at 0.75
Enhanced	176	Varied to maximize PD, while limiting PFA to 1% or less

3. Results

As described in the previous section, we used 50 rounds of 10-fold cross-validation to evaluate our machine learning models and library matching methods. According to our PD calibration process, we allowed the threshold to vary such that the PD was maximized while limiting the PFA to no more than 1%.

Within this framework, we first compared different versions of simple library matching, showing results in Table 4. In all rounds, the standard and targeted approaches yielded 0% PFA, while the adaptive and enhanced approaches yielded 1%. The standard and targeted approaches use the same library of five fentanyl analogs for all cross-validation folds and thus have no variation in their PD.

Among the library matching approaches under consideration, it is clear that enhanced library matching yields the best performance (with mean values of 96.3% PD and 1.00% PFA). We obtained these results using optimized score thresholds that varied between 0.46 and 0.48. Historically, chemists have used a fixed decision threshold with library matching (as we used in “standard library matching” and “targeted library matching”). Therefore, the good detection performance of enhanced library matching, which allows the threshold to vary and considers values that are smaller than traditional threshold values, is somewhat surprising and warrants deeper analysis in further studies.

Given these results, we compared enhanced library matching with three types of machine learning models. These results are shown in Table 5. As with adaptive and enhanced library matching, we allowed the decision threshold to vary for each of the machine learning models such that PD was maximized while limiting PFA to no more than 1%. AUROC provides a complementary metric that does not require setting a decision threshold.

Our machine learning results are from models that include the full set of input features listed in Tables 1 and 2. Because we have set the decision threshold based on model scores from test set spectra, these results do not provide an unbiased estimate of PD and PFA in deployment settings. However, they do enable clear and fair comparison between machine learning models and library matching.

Across the 50 rounds of 10-fold cross-validation, the random forest models achieve both the highest mean PD (99.7%) and mean AUROC (0.9998). As is evident from these metrics’ empirical 95% confidence intervals (shown by the length of the error bars in Fig. 3), the random forest models have limited variability, indicating that the effects of the random components of the model and cross-validation folds are small. While enhanced library matching also achieves very good performance (96.3%), it is not as good as the random forest models. Our methods outperform classic library matching, even when a large library of fentanyl analogs are available. Fig. 3 further illustrates these results. In this figure, the marker indicates the mean value and the bars indicate empirical 95% confidence intervals.

Table 4
Comparison of Library Matching Approaches.

Method	Mean Probability of Detection*(empirical 95% confidence interval)*
Standard library matching	6.3% *(no variation)*
Adaptive library matching	44.7% *(no variation)*
Targeted library matching	67.3% *(64.5%, 68.9%)*
Enhanced library matching	96.3% *(95.3%, 97.4%)*

Table 5
Comparison of Enhanced Library Matching with Machine Learning Models.

Method	Mean Calibrated Probability of Detection*(empirical 95% confidence interval)*	Mean Area Under the ROC Curve*(empirical 95% confidence interval)*
Enhanced library matching	96.3% *(95.3%, 97.4%)*	0.9977 *(0.9965, 0.9981)*
Logistic regression	94.3% *(91.9%, 97.9%)*	0.9854 *(0.9545, 0.9957)*
Neural network	96.8% *(94.9%, 98.3%)*	0.9968 *(0.9965, 0.9981)*
Random forest	99.7% *(99.5%, 100.0%)*	0.9998 *(0.9997, 0.9999)*

To eliminate the bias accompanying calibrated PD (Table 5), we also separately conducted 50 rounds of nested cross-validation, showing the results in Table 6. Since a best model is selected for each of the 10 outer-loop cross-validation folds, there are 500 total instances in which a best model is selected. As in Table 5, the random forest model consistently outperforms other model types.

These results are very encouraging, suggesting that the bias stemming from a test set-guided threshold has minimal effect. Even after eliminating this bias, our machine learning approach maintains very high PD (99.8%) and reasonably low PFA (0.89%). The models’ high detection performance includes consistent success against fentanyl analogs with low similarity to fentanyl, e.g., lofentanil (CS with fentanyl = 0.091), which are often missed by enhanced library matching. These results provide the best estimate of our models’ likely performance against unseen spectra.

Regarding feature importance, when we exclude the five similarity and five peakwise similarity features, OOB calibrated PD decreases from 99.7% to 98.9%. Additionally, the feature with the greatest permutation importance is the library matching score, highlighting how our methods complement and build upon existing library matching techniques. Further detail on feature importance is provided in Appendix C.

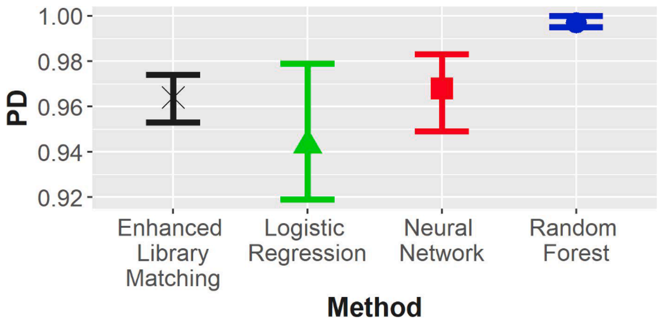

Fig. 3. Comparison of PD Across Methods with PFA Calibrated to 1% or Less.

Table 6
Nested Cross-Validation Results.

Frequency That Each Method is Selected as the Best Model			
Enhanced Library Matching (LM)	Logistic Regression (LR)	Neural Network (NN)	Random Forest (RF)
0	0	0	500

Mean Probability of Detection from Best Model *(empirical 95% confidence interval)*	Mean Probability of False Alarm from Best Model *(empirical 95% confidence interval)*
99.8% *(99.5%, 100.0%)*	0.89% *(0.77%, 0.99%)*

4. Conclusions and next steps

While enhanced library matching offers substantial improvement over standard library matching, random forest models provide even better performance. Our results suggest that random forest models are capable of achieving 99% PD and 1% PFA against unseen spectra. Such detection performance offers a promising capability for fentanyl detection within mass spectra. Accordingly, we recommend further research on random forests and other machine learning models as a complement to library matching.

There are several avenues by which our efforts might be advanced. We have trained and evaluated our models on isolated “pure” mass spectra. Often, mass spectrometry is combined with online separation and thus yields a time series of mass spectra of chromatographically separated analytes (i.e., gas chromatography mass spectrometry). Further effort is needed to adapt our approach to such time series.

Moreover, we have considered a very coarse set of popular machine learning techniques: logistic regression, neural networks, and random forests, generally using default parameter options. Our models might be further improved by exploring additional machine learning techniques, tuning model parameters, or adding a separate feature selection step. The various model types might also be combined into an ensemble model.

While mass spectrometry is prominent in laboratory analysis of collected samples, other sensor types such as infrared or Raman are more commonly used in portable settings. Further work is needed to extend our approach to spectra collected with these sensor types. Indeed, our initial attempt to adapt our fentanyl detection methods to Raman spectra (not described in this document) has yielded promising results.

We have used cross-validation to mimic the scenario in which our models would be exposed to previously unseen fentanyl analogs. Our models could be further validated with predictions of the structure [36] and spectra of fentanyl analogs that have not yet been observed but could potentially be synthesized. Such predictions would enable a valuable capability for assessing model performance against possible future threats, but further work is needed to reliably predict mass spectra from chemical structures.

In practice, detection models are trained on the spectra of analogs known at the time that they are constructed with the ambition of detecting novel analogs that emerge later. If particular fentanyl analogs could be assigned a “date of emergence,” our models could be further validated by separating spectra according to such dates and evaluating the performance of earlier models upon later analogs.

Finally, we have trained our models with over 3000 mass spectra, including nearly 200 spectra from fentanyl analogs. These models have shown great success and promise. In general, we would expect that our model’s performance would decrease if fewer spectra were available for model training. Learning curve analysis (cf., [37;38]) would provide insight into such a tradeoff. This analysis would be particularly important for estimating data needs for applying our approach to other classes of substances of interest as described above.

Within this work, we present an approach for developing a model to detect a certain class of substances (fentanyl analogs). Our approach might also be applied to develop detection models for other classes of substances of interest (e.g., certain types of chemical weapon agents). Following our approach, an analyst would extract features from spectra from the class of interest, identify patterns within those features via machine learning, and apply the resulting models to detect substances from the class of interest within unseen spectra. If successful, such extensions would add valuable detection capability for a wide variety of applications.

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgements

The authors thank Dr. P. Demirev, Johns Hopkins University Applied Physics Laboratory (JHUAPL), for thoughtful feedback and editorial suggestions on the manuscript. The authors would also thank Dr. D. Portnoy and Dr. B. Jennison, both at JHUAPL, for numerous fruitful technical exchanges regarding library matching. Additionally, the authors thank D. George, also from JHUAPL, for writing code to format spectral peak data. This work was funded by the JHUAPL internal research and development program and other internal JHUAPL funding.

Appendix

A. Machine learning techniques

Logistic regression for binary classification involves calculating a probability of the reference class (fentanyl analog for our study) according to the logistic equation $P(y|x) = 1/(1 + e^{-\beta x})$, where vector β contains the learned coefficients for each feature. A classification is made by comparing the model output with a specified detection threshold T_{LR}; we selected T_{LR} to be as low as possible while still limiting PFA to no more than 1%. To construct our logistic regression models, we used the *glm* function in the R *stats* package [35].

Neural networks involve layers of nodes, in which each layer’s values are functions of the previous layer’s nodes. For binary classification (as with our fentanyl analog detection models), there is also a single node in the outer layer and a class is predicted according to whether that node’s value exceeds the specified detection threshold T_{NN}; as with logistic regression, we selected T_{NN} to be as low as possible while still limiting PFA to no more than 1%.

To construct our neural network models, we used the *neuralnet* function in the R *neuralnet* package [39]. Our neural network models included two hidden layers of 15 and 10 nodes respectively, with each node using the logistic activation function. We also scaled each feature’s value by subtracting that feature’s mean from all values of that feature and dividing by that feature’s standard deviation. Such scaling increases the likelihood that the *neuralnet* function will converge to a solution. We set the convergence threshold to 0.1, used one repetition, and set the error function as cross-entropy. Higher converge thresholds increase the likelihood that a solution will be found, but also generally reduce the performance of the resulting model. We sought a convergence threshold that was as low as possible but still reliably converged.

Random forests are ensembles of decision trees that separate samples into small sets according to logical rules based on feature values, e.g., *BasePeakMass* greater than 300. The decision trees are constructed from various random subsets of the original training set, selected in a way that minimizes correlation between the trees, thereby reducing variance of the overall model. For binary classification, the positive class (i.e., fentanyl analog) is predicted if the proportion of votes received for that class exceeds the specified detection threshold T_{RF}; as with the other model types, we selected T_{RF} to be as low as possible while still limiting PFA to no more than 1%. To construct random forest models, we used the *randomForest* function in the R *randomForest* package [40], including 500 trees in each model.

Table A1 summarizes the methods that we implemented.

Table A1
Implemented Supervised Learning Classification Methods.

Method	R Function	Notes
Logistic Regression	*glm*	T_{LR} set to maximize PD while limiting PFA to 1% or less

Neural Networks	*neuralnet*	# hidden layers = 2;
		# nodes in each hidden layer = 15, 10;
		convergence threshold = 0.1;
		activation function = logistic;
		error function = cross-entropy;
		T_{NN} set to maximize PD while limiting PFA to 1% or less

Random Forests	*randomForest*	# trees = 500; T_{RF} set to maximize PD while limiting PFA to 1% or less

Appendix B. . Unbiased evaluation via nested cross-validation

If our models were deployed in an operational setting, we would need to select a detection threshold based solely on training data. In other words, since spectra of unknown samples would not be available beforehand, we could not use them to determine any part of our detection scheme (e.g., the specific model type or accompany detection threshold). Above, we have used the test-set spectra (which are assumed to be "unknown" within cross-validation) to determine the detection threshold such that test set PD is maximized while limiting test set PFA to less than 1%. This incurs a bias.

To eliminate this bias, we also carried out a nested cross-validation according to the following process. By nesting a second cross-validation loop inside of a typical cross-validation loop, we select a best model type and a corresponding detection threshold without incorporating the held-out test fold at all. Herein, our nested cross-validation process mimics a deployment scenario in which a best model and detection threshold would necessarily be selected prior to assessing unknown spectra collected in the field.

We conducted an inner cross-validation on nine folds, holding out an outer test fold. We trained models on eight folds and iteratively evaluated these models on a ninth inner test fold, repeating this process for all nine folds. Using the results from the nine inner test folds, we selected a detection threshold and a best model. As illustration, each model's results for each of the ten outer test folds from one of the rounds of nested cross-validation are shown in Table B1. In this round, the random forest model yielded the best calibrated PD for all ten inner cross-validation loops (as is highlighted in boldface); therefore, we selected random forest as the model type and used its corresponding threshold for each of the ten outer test folds.

Following this model selection step, we trained a new model of the selected type (in this case, random forests) solely using the spectra from the nine folds; unlike with the preliminary results shown in Table 5, we did not incorporate test fold results in our selection of a detection threshold and a best model type. We applied this single model, along with the selected detection threshold to the held-out test fold, predicting whether each test fold spectrum came from a fentanyl analog by comparing the model score with the detection threshold. We repeated this process with each of the ten outer-loop folds as the held-out test fold. Based on these results, we calculated PD and PFA. Since this process does not use the test spectra to determine any aspect of the detection scheme, it enables an unbiased estimate of our approach's likely performance against unseen spectra. To account for randomness in how spectra are separated into folds and models are constructed, we conducted 50 rounds of this nested cross-validation process. As an example, Table B1 shows results from a single nested cross-validation round (Table B1).

Appendix C. . Feature importance results

As described in Section 2.4, we tracked the permutation importance for each feature within 50 random forest models trained with all available data. These importance values are shown in Table C1. Permutation importance is the amount by which out-of-bag (OOB) accuracy decreases when a given feature's values are permuted. For a given spectrum, OOB predictions are only made on trees for which that spectrum was not part of the tree construction. We obtained these values directly from the *randomForest* output (Table C1).

Table C1
Peak-Related Machine Learning Features for Detection Models.

Feature	Permutation Importance (Mean Decrease in Accuracy)_
Base Peak Mass	0.31%
Base Peak Mass Proximity	0.04%
Maximum Mass	0.76%
Maximum Mass Proximity	0.06%
Number of Peaks	0.38%
Intensity Mean	0.26%
Intensity Standard Deviation	0.16%
Intensity Density	0.47%
Mass Mean	0.23%
Mass Standard Deviation	0.37%
Mass Density	0.30%
Most Frequent Pair Peakwise Mass Difference (PPMD)	0.00%
Mean PPMD	0.32%
Similarity to Alfentanil	0.11%
Similarity to Carfentanil	0.23%
Similarity to Fentanyl	0.37%
Similarity to Remifentanil	0.07%
Similarity to Sufentanil	0.29%
Peakwise Similarity to Alfentanil	0.94%
Peakwise Similarity to Carfentanil	0.76%
Peakwise Similarity to Fentanyl	0.57%
Peakwise Similarity to Remifentanil	0.84%
Peakwise Similarity to Sufentanil	0.29%
Library Matching Score	4.52%

Table B1
Results from an example round of nested cross-validation.

OuterCross-Validation Fold	Inner-Loop Cross-Validation Calibrated PD Results					Outer-Loop Cross-Validation Results		
	Enhanced Library Matching (LM)	Logistic Regression (LR)	Neural Network (NN)	Random Forest (RF)	Best Model	Best Model Threshold	Best Model PD	Best Model PFA
1	95.9%	95.3%	97.7%	**100.0%**	RF	0.141	100.0%	1.14%
2	94.7%	95.9%	95.3%	**99.5%**	RF	0.150	100.0%	0.57%
3	95.9%	93.0%	95.9%	**100.0%**	RF	0.135	100.0%	0.28%
4	95.3%	96.5%	95.9%	**100.0%**	RF	0.149	100.0%	1.14%
5	96.5%	94.7%	97.0%	**100.0%**	RF	0.151	100.0%	0.85%
6	96.5%	91.8%	96.5%	**99.5%**	RF	0.153	100.0%	0.28%
7	95.9%	96.5%	96.5%	**99.5%**	RF	0.155	100.0%	1.14%
8	95.3%	96.5%	94.7%	**99.5%**	RF	0.153	94.7%	1.13%
9	94.7%	97.7%	95.9%	**100.0%**	RF	0.153	100.0%	1.14%
10	95.9%	97.1%	95.9%	**100.0%**	RF	0.155	100.0%	1.70%
				Mean:		0.150	99.5%	0.94%

Appendix D. . Data summary

The enclosed data set includes the spectra that we used to train and evaluate our models. As mentioned in Section 2.1., we obtained these data from the National Institute of Standards and Technology (using the collection of libraries distributed with the NIST05 MS Search demonstration software) and the Scientific Working Group for Seized Drugs [29].

These spectra are represented as a single matrix. Each row corresponds to a single spectrum and each column (except for the first column) corresponds to the intensity at a given integer *m/z* value. The first column indicates whether that row's spectrum corresponds to a fentanyl analog (indicated by a value of 1) or to a non-fentanyl substance (indicated by a value of 0). The spectra in the final five rows correspond to prominent fentanyl analogs: fentanyl, alfentanil, carfentanill, remifentanil, sufentanil.

For each spectrum, *m/z* values without peaks are assigned an intensity value of 0.

Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.forc.2021.100379.

References

[1] T.H. Stanley, The history and development of the fentanyl series, J. Pain Symptom Manage. 7 (3) (1992) S3–S7, https://doi.org/10.1016/0885-3924(92)90047-L.

[2] Kuczyńska, Katarzyna, et al. "Abuse of fentanyl: An emerging problem to face." Foren. Sci. Int. 289 (2018): 207-214. https://doi.org/10.1016/j.forsciint.2018.05.042.

[3] O'Donnell, Julie, et al. "Notes from the field: opioid-involved overdose deaths with fentanyl or fentanyl analogs detected—28 states and the District of Columbia, July 2016–December 2018." *Morbidity and Mortality Weekly Report* 69.10 (2020): 271.

[4] R.L. Rothberg, K. Stith, Fentanyl: a whole new world? J. Law, Med. Ethics 46 (2) (2018) 314–324, https://doi.org/10.1177/1073110518782937.

[5] de Araujo, R. William, et al. "Portable analytical platforms for forensic chemistry: a review." Anal. Chim. Acta 1034 (2018): 1-21. https://doi.org/10.1016/j.aca.2018.06.014.

[6] W.MA. Niessen, D. Falck, Introduction to mass spectrometry, a tutorial, in: Analyzing Biomolecular Interactions by Mass Spectrometry 1, 2015, pp. 1–54, https://doi.org/10.1002/9783527673391.ch1.

[7] E. Smith, G. Dent, Modern Raman spectroscopy: a practical approach, John Wiley & Sons, 2019. Doi: 10.1002/0470011831.

[8] B.C. Smith, Fundamentals of Fourier transform infrared spectroscopy, CRC Press, 2011. Doi: 10.1201/b10777.

[9] S.E. Stein, D.R. Scott, Optimization and testing of mass spectral library search algorithms for compound identification, J. Am. Soc. Mass Spectrom. 5 (9) (1994) 859–866, https://doi.org/10.1016/1044-0305(94)87009-8.

[10] Reitzel, Lotte A., et al. "Identification of ten new designer drugs by GC-MS, UPLC-QTOF-MS, and NMR as part of a police investigation of a Danish Internet company." Drug Testing Anal. 4.5 (2012): 342-354. https://doi.org/10.1002/dta.358.

[11] 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.

[12] Helander, Anders, et al. "Intoxications involving acrylfentanyl and other novel designer fentanyls–results from the Swedish STRIDA project." *Clinical toxicology* 55.6 (2017): 589-599. https://doi.org/10.1080/15563650.2017.1303141.

[13] Moorthy, Arun S., et al. "Combining fragment-ion and neutral-loss matching during mass spectral library searching: A new general purpose algorithm applicable to illicit drug identification." *Analytical chemistry* 89.24 (2017): 13261-13268. https://doi.org/10.1021/acs.analchem.7b03320.

[14] Moorthy, S. Arun, et al. "Mass spectral similarity mapping applied to fentanyl analogs." *Forensic Chemistry* 19 (2020): 100237. https://doi.org/10.1016/j.forc.2020.100237.

[15] Nan, Qin, et al. "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." *Journal of the American Society for Mass Spectrometry* 31.2 (2020): 277-291. https://doi.org/10.1021/jasms.9b00112.

[16] L. Lo Vercio, K. Amador, J.J. Bannister, S. Crites, A. Gutierrez, M.E. MacDonald, J. Moore, P. Mouches, D. Rajashekar, S. Schimert, N. Subbanna, A. Tuladhar, N. Wang, M. Wilms, A. Winder, N.D. Forkert, Supervised machine learning tools: a tutorial for clinicians, J. Neural Eng. 17 (6) (2020) 062001, https://doi.org/10.1088/1741-2552/abbff2.

[17] Sparkman, O. David, Zelda Penton, and Fulton G. Kitson. *Gas chromatography and mass spectrometry: a practical guide*. Academic press, 2011. https://doi.org/10.1016/C2009-0-17039-3.

[18] Brown, Hilary M., et al. "The current role of mass spectrometry in forensics and future prospects." *Analytical Methods* 12.32 (2020): 3974-3997. https://doi.org/10.1039/D0AY01113D.

[19] Lin, Zhang, et al. "Exploring metabolic syndrome serum profiling based on gas chromatography mass spectrometry and random forest models." Anal. Chim. Acta 827 (2014): 22-27. https://doi.org/10.1016/j.aca.2014.04.008.

[20] Jang, Inae, et al. 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. 91.14 (2019): 9119-9128. https://doi.org/10.1021/acs.analchem.9b01643.

[21] Davidson, B. Nicola, et al. "Rapid identification of species, sex and maturity by mass spectrometric analysis of animal faeces." BMC Biol. 17.1 (2019): 1-14. https://doi.org/10.1186/s12915-019-0686-9.

[22] U.W. Liebal, A.N.T. Phan, M. Sudhakar, K. Raman, L.M. Blank, Machine learning applications for mass spectrometry-based metabolomics, Metabolites 10 (6) (2020) 243, https://doi.org/10.3390/metabo10060243.

[23] Huang, Ying-Chen, et al. "Predicting breast cancer by paper spray ion mobility spectrometry mass spectrometry and machine learning." Anal. Chem.istry 92.2 (2019): 1653-1657. https://doi.org/10.1021/acs.analchem.9b03966.

[24] Z. Zhou, R.N. Zare, Personal information from latent fingerprints using desorption electrospray ionization mass spectrometry and machine learning, Anal. Chem. 89 (2) (2017) 1369–1372, https://doi.org/10.1021/acs.analchem.6b0449810.1021/acs.analchem.6b04498.s001.

[25] J.N. Wei, D. Belanger, R.P. Adams, D. Sculley, Rapid prediction of electron–ionization mass spectrometry using neural networks, *ACS Cent. Sci.* 5(4) 5 (4) (2019) 700–708, https://doi.org/10.1021/acscentsci.9b00085.

[26] Xu, Mengyu, et al. "High accuracy machine learning identification of fentanyl-relevant molecular compound classification via constituent functional group analysis." Scient. Rep. 10.1 (2020): 1-10. https://doi.org/10.1038/s41598-020-70471-7.

[27] K. Wang, B. Xu, J. Wu, Y. Zhu, L. Guo, J. Xie, Elucidating fentanyls differentiation from morphines in chemical and biological samples with surface-enhanced Raman spectroscopy, Electrophoresis 40 (16-17) (2019) 2193–2203, https://doi.org/10.1002/elps.v40.16-1710.1002/elps.201900004.

[28] J. Bonetti, Mass spectral differentiation of positional isomers using multivariate statistics, Forensic Chem. 9 (2018) 50–61, https://doi.org/10.1016/j.forc.2018.06.001.

[29] Scientific Working Group for the Analysis of Seized Drugs. SWGDRUG Mass Spectral Library. Version 3.5. URL https://www.swgdrug.org/ms.htm.

[30] V. Bewick, L. Cheek, J. Ball, Statistics review 13: receiver operating characteristic curves, Crit. Care 8 (6) (2004) 1–5, https://doi.org/10.1186/cc3000.

[31] Baldygo, William, et al. "Artificial intelligence applications to constant false alarm rate (CFAR) processing." *The Record of the 1993 IEEE National Radar Conference*. IEEE, 1993. https://doi.org/10.1109/NRC.1993.270451.

[32] A. Agresti, Categorical data analysis, John Wiley & Sons, 2003. Doi: 10.1002/0471249688.

[33] C.M. Bishop, Neural networks for pattern recognition, Oxford University Press, 1995 https://dl.acm.org/doi/10.5555/235248.

[34] L. Breiman, Random forests, Mach. Learn. 45 (1) (2001) 5–32, https://doi.org/10.1023/A:1010933404324.

[35] R Core Team (2019). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. URL https://www.R-project.org.

[36] Walters, W. Patrick, and Regina Barzilay. "Applications of deep learning in molecule generation and molecular property prediction." Account. Chem. Res. 54.2 (2020): 263-270. https://doi.org/10.1021/acs.accounts.0c00699.

[37] Figueroa, L. Rosa, et al. "Predicting sample size required for classification performance." *BMC medical informatics and decision making* 12.1 (2012): 1-10. https://doi.org/10.1186/1472-6947-12-8.

[38] Koshute, Phillip, Jared Zook, Ian McCulloh. "Recommending Training Set Sizes for Classification." *arXiv preprint arXiv:2102.09382* (2021).

[39] Fritsch, Stefan, Frauke Guenther and Marvin N. Wright. Neuralnet: Training of Neural Networks. R package version 1.44.2. 2019. https://CRAN.R-project.org/package=neuralnet.

[40] A. Liaw, M. Wiener, Classification and Regression by randomForest, R News 2 (3) (2002) 18–22.