analytical chemistry

pubs.acs.org/ac Article

Tile-Based Random Forest Analysis for Analyte Discovery in Balanced and Unbalanced GC × GC-TOFMS Data Sets

Meriem Gaida,* Caitlin N. Cain, Robert E. Synovec, Jean-François Focant, and Pierre-Hugues Stefanuto

Cite This: *Anal. Chem.* 2023, 95, 13519−13527 Read Online

ACCESS | Metrics & More | Article Recommendations | Supporting Information

ABSTRACT: In this study, we introduce a new nontargeted tile-based supervised analysis method that combines the four-grid tiling scheme previously established for the Fisher ratio (F-ratio) analysis (FRA) with the estimation of tile hit importance using the machine learning (ML) algorithm Random Forest (RF). This approach is termed tile-based RF analysis. As opposed to the standard tile-based F-ratio analysis, the RF approach can be extended to the analysis of unbalanced data sets, i.e., different numbers of samples per class. Tile-based RF computes out-of-bag (oob) tile hit importance estimates for every summed chromatographic signal within each tile on a per-mass channel basis (m/z). These estimates are then used to rank tile hits in a descending order of importance. In the present investigation, the RF approach was applied for a two-class comparison of stool samples collected from omnivore (O) subjects and stored using two different storage conditions: liquid (Liq) and lyophilized (Lyo). Two final hit lists were generated using balanced (8 vs Eight comparison) and unbalanced (8 vs Nine comparison) data sets and compared to the hit list generated by the standard F-ratio analysis. Similar class-distinguishing analytes ($p < 0.01$) were discovered by both methods. However, while the FRA discovered a more comprehensive hit list (65 hits), the RF approach strictly discovered hits (31 hits for the balanced data set comparison and 29 hits for the unbalanced data set comparison) with concentration ratios, [OLiq]/[OLyo], greater than 2 (or less than 0.5). This difference is attributed to the more stringent feature selection process used by the RF algorithm. Moreover, our findings suggest that the RF approach is a promising method for identifying class-distinguishing analytes in settings characterized by both high between-class variance and high within-class variance, making it an advantageous method in the study of complex biological matrices.

Nontargeted analysis (NTA) of comprehensive two-dimensional gas chromatography coupled with time-of-flight mass spectrometry (GC × GC-TOFMS) data relates to identifying unknown analytes without prior knowledge of their identity. In a NTA, if the analyte's class affiliation is used as input, the analysis is referred to as supervised; if not, the analysis is simply labeled as unsupervised.[1] Although the increased peak capacity in GC × GC offers a greater insight into the sample's chemical composition, it substantially increases the complexity of the produced data, resulting in a rather daunting data interpretation task.[2] Over the years, multiple chemometric tools have been deployed to provide a comprehensive overview of the GC × GC-TOFMS data, some of which have become essential to the data processing workflow.[3] Among the diverse range of available chemometric tools, one can enumerate data reduction techniques such as principal component analysis[4] (PCA), regression-based methods such as partial least-squares[5] (PLS), or even discriminant analysis approaches such as partial least-squares-discriminant analysis[6] (PLS-DA), to name a few. These methods are harnessed to gain valuable insights into complex data sets, often serving different purposes, such as unveiling

Received: May 1, 2023
Accepted: August 17, 2023
Published: August 30, 2023

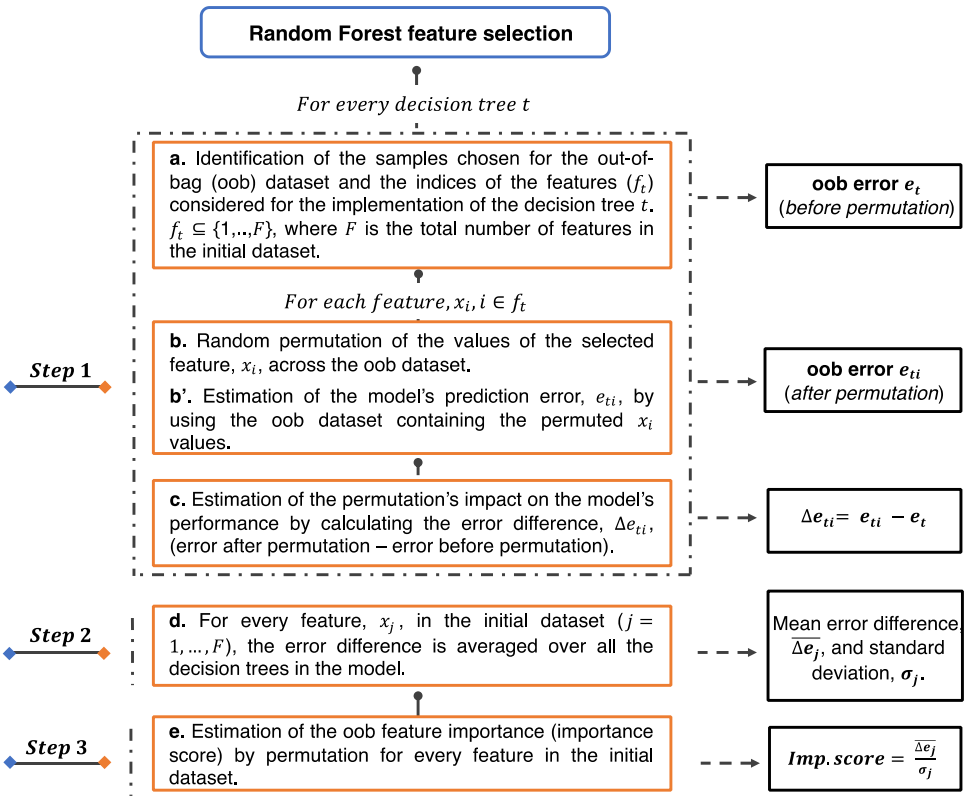

Figure 1. Systematic description of permutation-based feature importance estimation (feature selection) using the machine learning (ML) algorithm random forest (RF). (a) Calculation of the out-of-bag (oob) error before permutation. (b, b′) Calculation of the oob error after permutation. (c) Calculation of the oob error difference. (d) Calculation of the mean oob error difference and the standard deviation. (e) Calculation of the importance score.

clustering patterns within a given data set, correlating chemical features to response factors, or performing classification tasks.[1,3]

In recent years, tile-based approaches have gained popularity as a means to overcome the challenge of data misalignment.[1,7,8] Briefly, the tiling scheme divides the chromatograms along the first (^{1}D) and second (^{2}D) dimensions into small sections, known as "tiles", through a process termed chromatogram binning. At this stage, each chromatogram is binned by four different grids, shifted with respect to each other to ensure that every chromatographic peak is optimally seized by at least one of the gridding schemes. Due to the multigrid tiling process, each peak is sampled multiple times, resulting in hit redundancy. Therefore, the binning step is followed by a redundant hit removal step, involving the use of a "pinning and clustering" algorithm that ensures a single and unique representation for every analyte in the final tile hit list.[8] When first introduced, the tiling approach was used in conjunction with the one-way analysis of variance (ANOVA) statistical model, wherein the Fisher ratio (F-ratio) of the tiles was calculated and the tiles were ranked in descending F-ratio order. The efficiency of the tile-based F-ratio approach (FRA) was thoroughly investigated across various fields of application, and significant efforts were dedicated to increasing its robustness and enhancing its capabilities.[9−14] Additionally, there now exist commercial software that was developed based on the tile-based FRA methodology, making the method more accessible to a broader audience.[15−18] Nonetheless, several research studies have proposed some alterations to the aforementioned method to expand its scope to better suit specific practical settings. In this respect, multiple F-ratio amended calculations were suggested, such as the control-normalized F-ratio[19] (^{CN}FRA) and the minimum variance optimized F-ratio[20] (^{MVO}FRA). Furthermore, multiple studies investigated combining the tiling approach with other metrics, such as relative variance (RSD^2) in the tile-based variance rank initiated-unsupervised sample indexing (VRI-USI) method[21] and the use of a rank metric (RM) for pairwise analysis workflows.[22] Despite the extensive efforts devoted to optimizing and extending the tile-based FRA to accommodate various settings, the scenario of unbalanced data sets, i.e., different numbers of samples per class, remains untreated and calls for further attention. Since the F-ratio is a statistical measure calculated in the context of the ANOVA model, unequal sample sizes could compromise the robustness of the model due to the equal variance assumption violation.[23] Note that when the unequal sample size effect is not substantial, the error might be negligible but may not be when the class sizes are severely unbalanced. Therefore, to eliminate this potential source of error altogether, we introduce in this report a new nontargeted supervised method that combines the previously established tiling approach with one of the most famous machine learning (ML) algorithms Random Forest (RF).

The application of ML algorithms in the interpretation of GC × GC data has lately been gaining significant attraction, primarily due to the advantages they offer over traditional statistical methods.[15,24−27] One of their key advantages is their inherent ability to handle large data sets, which can encompass either a limited number of samples with a high number of features or a larger number of samples with a restricted number

https://doi.org/10.1021/acs.analchem.3c01872
Anal. Chem. 2023, 95, 13519−13527

of features, with relative ease. Additionally, ML algorithms are highly regarded for their ability to reduce the data dimensionality. By combining sampling strategies and feature selection techniques, ML algorithms can considerably alleviate the complexity of the data interpretation task while jointly enhancing model performance.[28,29] Among the wide variety of ML algorithms, RF stands out as a leading example. RF is a supervised method mainly used for classification and regression purposes.[30] It operates by implementing a set of decision trees built upon a random sampling of the investigated data set. Each tree is built on a different subset of the data and operates independently of the other trees in the forest. In classification schemes, each decision tree accounts for a specific class prediction, and the final outcome of the RF model is determined by the class selected by the majority of the trees. RF models are commonly known for their robustness, as they are less sensitive to noise and the presence of outliers in the data. They are particularly suited for the study of large data sets where the number of features outweighs the number of observations and are less prone to overfitting. In contrast to other statistical methods, such as the ANOVA model, RF models do not require any assumptions about the data distribution, making them more flexible and applicable to a diverse range of data sets.[30−32] This key advantage is utilized in the new data processing method proposed in this study and referred to as tile-based RF analysis. Unlike the F-ratio analysis, RF models do not require the use of the same number of samples per class and can be efficiently extended to the analysis of unbalanced data sets. This asset is assigned to the way the RF algorithm operates and will be comprehensively discussed in this paper. Herein, we evaluate the efficiency of the tile-based RF approach within a two-class comparative analysis of stool samples collected from omnivore individuals and stored under two different storage conditions, namely, liquid and lyophilized. Furthermore, we compare its performance with that of the standard tile-based F-ratio analysis.

PRINCIPLES AND THEORY

In this report, we introduce a new nontargeted tile-based supervised analysis method that combines the four-grid tiling scheme previously established for the F-ratio analysis,[7,8] with the estimation of tile hit importance using the ML algorithm RF. This approach is referred to as tile-based RF analysis.

Permutation-based feature selection in RF relies on a rather rudimentary concept. If a feature (variable) is influential in a prediction model, then the permutation of its values across all observations (samples) ought to affect the model's accuracy. If not, then permuting its values should have little to no effect on the model's predictive power. In this scheme, RF quantifies the influence of every feature by the calculation of a measure called out-of-bag (oob) feature importance estimate.[33] The higher this estimate, the more important is the feature. As its name suggests, RF does not implement a single decision tree but rather operates within a *forest* of decision trees. Therefore, feature importance estimates are computed for every decision tree, t, in the model through an iterative procedure (Figure 1).

Due to its affiliation with bagging algorithms, RF randomly partitions the original data set into multiple subsets with an equal number of samples as in the initial data set. These subsets are commonly referred to as bootstrapped data sets. Each bootstrapped data set contains roughly two-thirds of the samples in the original set, and the remaining samples are simply replicated samples. In other words, one sample can be selected more than once. For every decision tree, the samples that are not selected in the bootstrapped data set, i.e., one-third of the samples in the original data set, are simply stored in another set called the out-of-bag data set. This set will act as an external validation set for every decision tree t. Additionally, the RF randomness is boosted by the use of subsets of features, f, for the training of the decision trees.[34] Therefore, when estimating feature importance, RF starts by identifying the oob data set and the features that are used to grow every decision tree t. Since feature importance in RF is assessed by permutation, the ML algorithm first starts by evaluating the model's performance by using the oob data set as a validation set and computes the oob error before permutation, e_t. Typically, this error can be interpreted as a misclassification error and can serve as a good indicator of the accuracy of the model (Figure 1a).

For every decision tree t and for each feature x_i considered to grow the tree, RF proceeds to a random permutation of the feature's values across the oob data set (Figure 1b) and computes an after permutation oob error, e_{ti} (Figure 1b′). Subsequently, an error difference of Δe_{ti} is reported. Since only a subset of features is considered in the growing and splitting of every decision tree, the features that are not selected during the process will be assigned an error difference of 0 ($\Delta e_{ti} = 0$) (Figure 1c). This error difference is then averaged over all of the decision trees in the model, and a standard deviation is calculated (Figure 1d). Finally, the importance of every feature is computed by simply dividing the mean error difference ($\overline{\Delta e_j}$) by the standard deviation, σ_j (Figure 1e). It is important to specify that, if after permutation, multiple features contribute equally to the decrease in the model's accuracy, then it is not uncommon for RF to attribute the same importance estimate to those features. In this report, the oob feature importance estimate for each hit is reported as a value scaled between 0 and 1 and will be referred to as the importance score.

In the context of this study, the RF computes the importance score for every summed chromatographic signal within each tile on a per-mass channel (m/z) basis. These scores are then used to rank the tile hits in descending order of importance. However, since a four-grid tiling scheme is used, one RF model is built for every single grid, accounting for four separate RF models. An importance score is then computed for every tile hit in every grid on a per-mass channel basis. Similar to the FRA, the redundant hits arising from the multigrid sampling are removed using the "pinning and clustering" algorithm.[1,8]

EXPERIMENTAL SECTION

Sample Collection. The tile-based RF approach was performed on stool samples acquired from the National Institute of Standards and Technology (NIST). These samples were collected from omnivore (O) subjects and stored under two different storage conditions, liquid (Liq) and lyophilized (Lyo), creating a two-class comparison scheme (OLiq vs OLyo). Three biological replicates were provided for each class. Upon collection, the OLiq class samples were already in the liquid state and were kept as such. As for the OLyo class samples, they were dried by using a lyophilizer. All samples were stored at −80 °C temperature.

HS-SPME-GC × GC-TOFMS Conditions. Chromatographic data collection was carried out on a Pegasus BT 4D GC × GC-TOFMS instrument equipped with a cryogenic

https://doi.org/10.1021/acs.analchem.3c01872
Anal. Chem. 2023, 95, 13519−13527

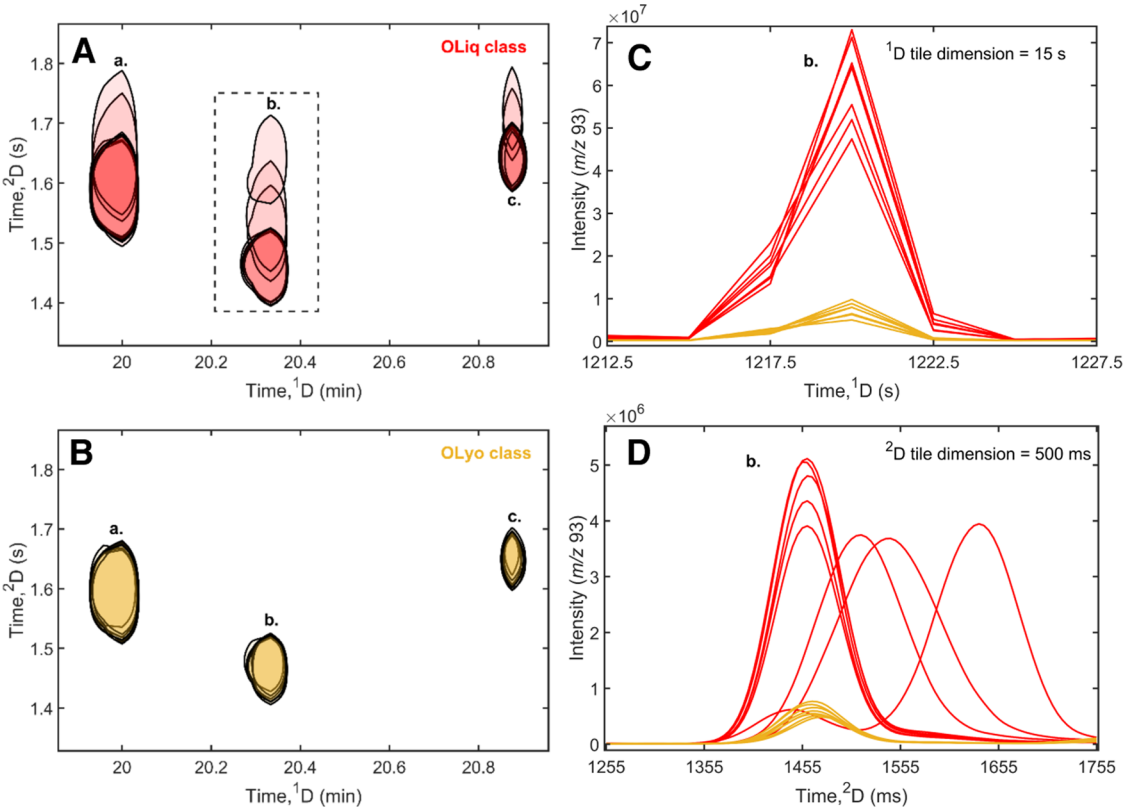

Figure 2. Illustration of the significant second-dimension (^{2}D) retention time shift in the omnivore liquid (OLiq) class chromatograms using the contour plots (A, B) of three example analytes: (a) caryophyllene, (b) trans-α-bergamotene, and (c) humulene. (C) Summed ^{1}D and (D) ^{2}D peaks of trans-α-bergamotene at m/z 93.

modulator (LECO Co., St. Joseph, MI). The headspace (HS) of each stool sample was extracted using a multisolid phase microextraction (SPME) setup. This setup allows for multiple analyses of a single sample by simultaneously using three SPME fibers, hence resulting in the measurement of nine chromatograms per class, i.e., three experimental measurements for each biological replicate. However, due to an adverse experimental incident, one measurement was not obtained in the OLiq class, thus resulting in an unbalanced number of chromatograms per class, namely, 8 chromatograms in the OLiq class and 9 chromatograms in the OLyo class. An example chromatogram for each class is provided in Figure 3. Nitinol-core (NIT) SPME fibers were used for each extraction. The volatile organic compounds (VOCs) were extracted at 40 °C for a total of 20 min and then desorbed at 250 °C in the GC injection port. Prior to the HS-SPME extractions, the OLyo class samples were diluted in 1 mL of Milli-Q water. All samples were analyzed on a normal column set: a nonpolar first dimension (^{1}D) (Rxi-5Sil MS, 30 m × 0.25 mm i_d × 0.25 μm d_f) and a polar second dimension (^{2}D) (Rxi-17Sil MS, 1.5 m × 0.25 mm i_d × 0.25 μm d_f). The GC oven was initially set at 40 °C for 1 min and then first ramped to 200 °C at a heating rate of 5 °C/min. A second temperature ramp of 15 °C/min was set to reach a final temperature of 270 °C (held for 2 min), accounting for a total runtime of 39.67 min. A modulation period of 2.5 s was used for all separations. Mass spectra were collected at a 200 Hz acquisition rate between mass channels 30 and 450 m/z.

Data Analysis. The raw chromatograms were imported in MATLAB (ver. R2021a, MathWorks, MA) and preprocessed by first applying the rolling ball minimum technique for baseline correction[35] and second by normalizing each chromatogram to the sum of the total ion current (TIC) chromatogram. The chromatograms were binned using the four-grid tiling scheme previously established for the F-ratio analysis.[7,8] A ^{1}D × ^{2}D tile size of 6 modulations (15 s) × 500 ms was used for the generation of all hit lists. Note that due to the prominent ^{2}D retention time shift noticed in 3 of the OLiq class chromatograms (Figure 2), the ^{2}D tile size was purposely optimized in order to embrace this retention time shift while maintaining a good enough selectivity for the individual chromatographic peaks. A cluster window size of 4 modulations (10 s) × 300 ms was used for the redundant hit removal step.[1,8] For both data processing methods, i.e., tile-based RF and tile-based F-ratio, a signal-to-noise (S/N) threshold of 10 was set to exclude tiles with poor chromatographic signals. Additionally, only tiles with at least 3 m/z above the S/N threshold are accounted for in the analyses. Analyte discovery using the standard F-ratio approach was performed using the in-house tile-based F-ratio software developed by the Synovec group at the University of Washington, Seattle, USA, and previously described in the literature.[1,7,8] However, this in-house software was altered to accommodate the specificity of the RF approach.

For each reported hit list, the hits were identified by using the 2011 NIST library, and a tentative analyte identification was assigned only for analytes with match values (MV) ≥ 800 (Tables S1−S3). Also, as an attempt to further enhance analyte identification, multivariate curve resolution-alternating least-squares (MCR-ALS) decomposition[36] was performed using the PLS Toolbox 9.1 (eigenvector Research, Inc., Wenatchee, WA) to provide a better resolution of the pure chromato-

https://doi.org/10.1021/acs.analchem.3c01872
Anal. Chem. 2023, 95, 13519−13527

graphic signals. Analytes with MV < 800 were reported as unknowns (Unk).[37]

For each hit, the concentration ratio between the two classes, [OLiq]/[OLyo], and the p-values were calculated based on a quantification of the total peak area in each studied sample at a specific m/z. Therefore, assessing m/z purity, which relates to m/z free from interfering signals, is of the utmost importance for accurate concentration ratio and p-value calculations. In this regard, peak purity was assessed by setting a mass channel m/z purity criterion that combines a lack of fit (LOF) threshold (LOF $\leq$ 10%) to a p-value threshold (p-value <0.001)[10] (Figure S1). Concentration ratio values were reported only for the m/z that satisfied the purity criterion. The hits where none of the discovered m/z complied with the purity criterion resulted in the reporting of the m/z that corresponded to the highest F-ratio value or the highest importance score in the final hit list tables along with its corresponding p-value (Tables S1−S3). Analytes with p-values <0.01 were labeled "true positives" and analytes with p-values ≥0.01 were labeled "false positives". p-values were calculated at the 99% confidence interval.

RESULTS AND DISCUSSION

Optimization of Random Forest Hyperparameters. Every RF model is governed by internal parameters, called hyperparameters. These parameters have a meaningful influence on the learning process of the model and can significantly influence its performance.[30] Key hyperparameters include the number of decision trees in the forest, the number of candidate features to sample at each node, and the minimum leaf size. Optimal combinations of values for these hyperparameters are of the utmost importance to achieving the best model performance. Nevertheless, determining the appropriate values can be challenging due to the lack of clear available guidance.[38]

In this study, we decided to optimize the RF model by fine-tuning some of its hyperparameters. The first parameter is the number of decision trees. Although, it cannot per se be tuned, as there is no one-size-fits-all answer to what the optimal number should be, it is generally recommended to select a sufficiently high number to achieve a good model performance without falling into the overfitting trap.[39−42] Commonly, in a RF model, the number of trees falls within the range of 100 to 1000.[43] Based on the large number of tiles produced by the binning of the chromatograms, we decided to set the number of trees to 300 to balance the need for accuracy with the need for computational efficiency. Furthermore, we optimized the number of candidate features to sample and the minimum leaf size by performing Bayesian Optimization (BO).[44] Briefly, BO is a technique often used to identify the optimal set of hyperparameters for a given ML model. Its goal is to minimize a so-called objective function, which is typically a measure of the model's performance. The first step in BO requires specifying a range of possible values for each hyperparameter to be optimized, usually based on prior knowledge and domain expertise. This step helps tone down the search space, thus, enabling the algorithm to focus on the most promising regions of the hyperparameter space. In this regard, we selected a range between 1 and 20 for the optimization of the minimum leaf size.[43] The minimum leaf size in the RF defines the depth of the trees. If the trees are too deep, they may overfit the model, and if they are too shallow, they may underfit the model.[38] As for the number of candidate features to sample by every decision tree, the default value is usually set to the square root of the total number of features in the original data set.[38,43] However, we wanted to better tailor this parameter to our specific practical setting by further optimization. Therefore, we selected a search range between 1 and the square root of the total number of features as an attempt to find a more model-dependent value. The rationale behind these selected values is that smaller values lead to more distinct and less correlated decision trees. Nonetheless, lower values may also decrease the model's accuracy in cases where the decision trees are implemented with nonimportant features. On the other hand, larger values may result in more correlated trees and substantially decrease the model's performance.[38] The optimized values of all the discussed parameters are reported in Tables S4−S5 in Supporting Information.

Evaluation of the Tile-Based Random Forest Approach. In the present work, the tile-based RF approach was applied for a two-class comparison (OLiq vs OLyo) of stool samples collected from omnivore individuals and stored using two different conditions: liquid and lyophilized. As previously stated, the two classes have a different sample size, making it an 8 vs Nine comparison, with the OLyo class being the outnumbered class. While the F-ratio approach has demonstrated exceptional data analysis capabilities when it comes to identifying class-distinguishing analytes, it encounters practical challenges when exposed to designs with unequal sample sizes, similar to the data set at our disposal. In fact, the equal variance assumption of the ANOVA model could be breached in an unbalanced data set, which can affect the robustness of the results.[23] Unlike the F-ratio analysis, RF is unaffected by the uneven sample size challenge as it can use stratified sampling when assessing feature importance by permutation.[33] This technique guarantees that each class is equally represented in the samples selected to construct both the bootstrapped data sets and the out-of-bag data sets. Stratified sampling is pivotal in assessing feature importance as it ensures that each class has an equal chance of being represented, thus avoiding any bias toward the majority class. In this respect, tile-based RF can be considered as an alternative method when the ANOVA assumptions are not met. In order to compare the performance of this newly introduced approach to that of the standard F-ratio analysis, one OLyo sample was randomly disregarded to reduce the two-class comparison to an even 8 vs Eight comparison. In this regard, three hit lists were generated, one for the unbalanced Random Forest (URF) analysis (8 vs 9), one for the balanced Random Forest (BRF) analysis (8 vs 8), and one for the F-ratio analysis (8 vs 8). All hit lists are provided in Supporting Information (Tables S1−S3).

The total ion current (TIC) chromatograms resulting from the GC × GC separation of both class samples exhibit a comparatively simple elution profile with a rather limited number of discernible peaks (Figure 3). The chromatograms display two distinct separation regions centered roughly around 10 and 21 min, respectively. These two regions display higher peak intensities and a greater number of eluting compounds.

Upon thorough analysis of all of the discovered analytes, a dominant presence of terpene-class compounds was noticed across all hit lists (Tables S1−S3), with a prevalent presence of monoterpenes in the first elution region and sesquiterpenes in the second elution region. Additionally, the concentration ratio chromatograms of the analytes discovered by either BRF or URF (Figure 4) mostly exhibit higher concentrations in the

https://doi.org/10.1021/acs.analchem.3c01872
Anal. Chem. 2023, 95, 13519−13527

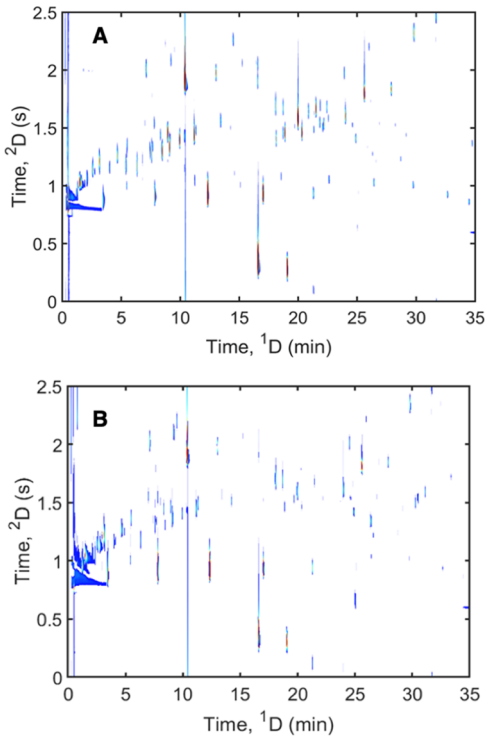

Figure 3. Total ion current (TIC) chromatograms of the GC × GC separations of (A) an omnivore liquid (OLiq) class sample and (B) an omnivore lyophilized (OLyo) class sample.

liquid samples compared to the lyophilized samples. The same observation is applicable to the F-ratio discovered analytes (Figure S2). These results may suggest that the lyophilization process can potentially affect the concentration of certain compounds, as they may undergo chemical or physical changes during the lyophilization process. However, the potential effect of lyophilization on the analyte concentrations cannot be fully assessed based on the available data and falls beyond the scope of this investigation.

A comparison of the hit lists generated by the BRF (Table S2) and URF (Table S3) analyses reveals not only a significant overlap in the number of discovered analytes (31 analytes for BRF and 29 analytes for URF) but also a substantial similarity in the discovered compounds, with a total of 23 common analytes. These observations corroborate the assumptions made about the RF performance not being compromised by the presence of an uneven sample size. Furthermore, the majority of the RF discovered analytes are considered statistically significant (p-value <0.01) and therefore class-distinguishing, except for 2 and 3 false positives (p-value ≥0.01) in the BRF and URF hit lists, respectively (Tables S2–S3). After the removal of any remaining redundant hits, hits identified as siloxane derivatives, and spectra with excessive background noise, the final F-ratio hit list contained 62 analytes. Their F-ratios ranged between 1191.01 and 1.82. Among these analytes, 49 compounds were identified as true positives (p-value <0.01), while the remaining 13 were labeled false positives (p-value ≥0.01). Upon reviewing the F-ratio hit list, it was observed that multiple hits were shared with both RF-generated hit lists, despite minor variations in their rankings. Specifically, the BRF and URF approaches identified 23 and 24 common analytes, respectively.

By preserving all the m/z associated with the tile hits, the tiling approach makes it possible to calculate the concentration ratios and the p-values for all of the discovered m/z of all of the discovered analytes. In this regard, we plotted the statistical significance of every m/z ($-\log_{10}(p_{value})$) vs its corresponding concentration ratio ($\log_2([OLiq]/[OLyo])$) as an attempt to identify the most meaningful changes between the F-ratio and the RF approach (Figure 5). This data representation method is often termed volcano plots. An initial look at the volcano plots suggests that a larger number of m/z (846 in total) were detected by the F-ratio approach (Figure 5A) compared to the RF method that only discovered 169 m/z in both the BRF and URF analyses (Figure 5B,C). However, only 50% of the F-ratio discovered m/z were deemed statistically significant, as only 429 m/z had a p-value of less than 0.01, whereas most of the RF discovered m/z fall below the 0.01 p-value threshold and therefore exhibited a high statistical significance. Furthermore, with regard to the concentration ratio, the volcano plots highlight two distinct regions of interest. The first region portrayed by the blue dots encompasses all m/z values with concentration ratios greater than 2 or less than 0.5. The second region represented by the gray dots corresponds to m/z with concentration ratios falling between 0.5 and 2. In this regard, two main observations can be made. First, it is evident that within the [0.5, 2] range, the F-ratio method portrays a higher incidence of false positives (p-value ≥0.01) as opposed to the RF approach. Second, most of the m/z discovered by RF is associated with concentration ratios exceeding 2 (higher presence in the OLiq class) or falling below 0.5 (higher presence in the OLyo class). In light of these empirical

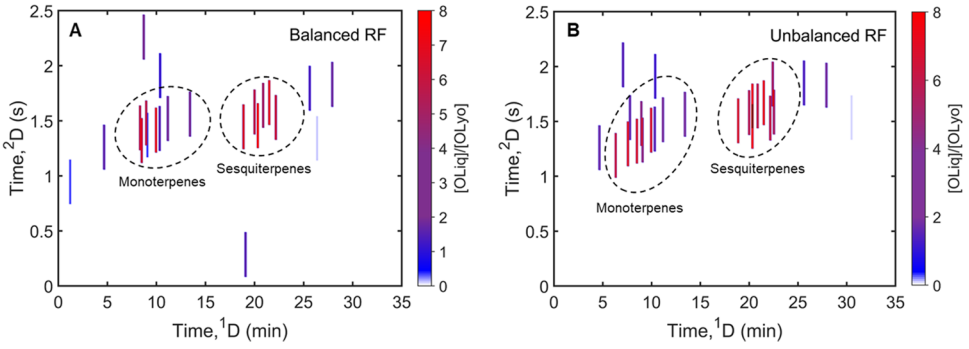

Figure 4. Concentration ratio chromatograms ([OLiq]/[OLyo]) displaying the chromatographic position of the hits discovered by (A) balanced (BRF) and (B) unbalanced (URF) RF analyses. The concentration ratios are only reported for the analytes that possess a top purest m/z satisfying both p-value <0.001 and LOF ≤10%.

https://doi.org/10.1021/acs.analchem.3c01872
Anal. Chem. 2023, 95, 13519−13527

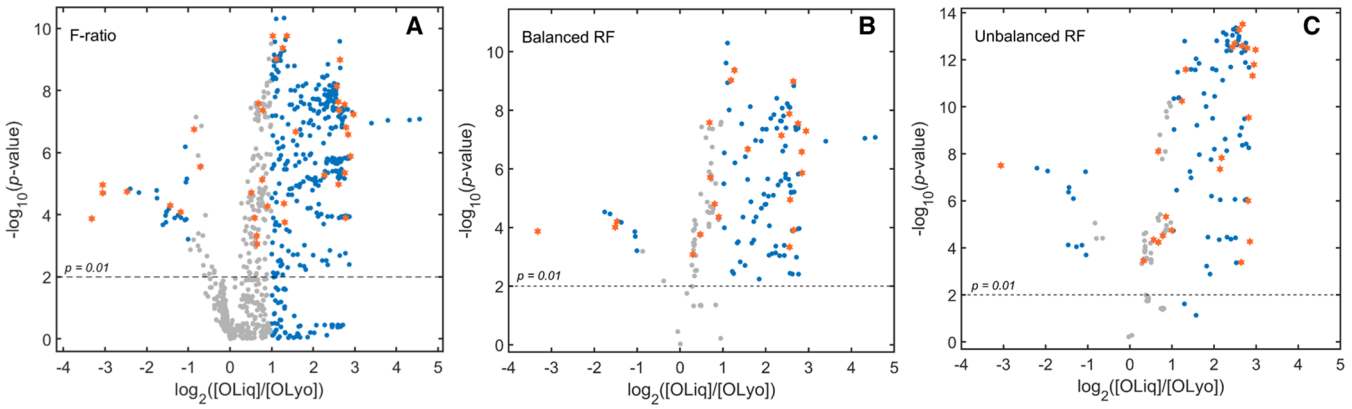

Figure 5. Volcano plot representations of all the m/z discovered through the three different tile-based analyses, associated with (A) the 62 analytes identified by the F-ratio analysis, (B) the 31 analytes discovered by the balanced random forest (BRF) analysis, and (C) the 29 analytes found through the unbalanced random forest (URF) analysis. The dashed line corresponds to a p-value = 0.01. m/z with concentration ratios $0.5 < [\text{OLiq}]/[\text{OLyo}] < 2$ ($-1 < \log_2([\text{OLiq}]/[\text{OLyo}]) < 1$) are highlighted in gray. For each discovered hit and if applicable, the top purest m/z (p-value <0.001 and LOF $\leq$ 10%) are highlighted with orange hexagrams.

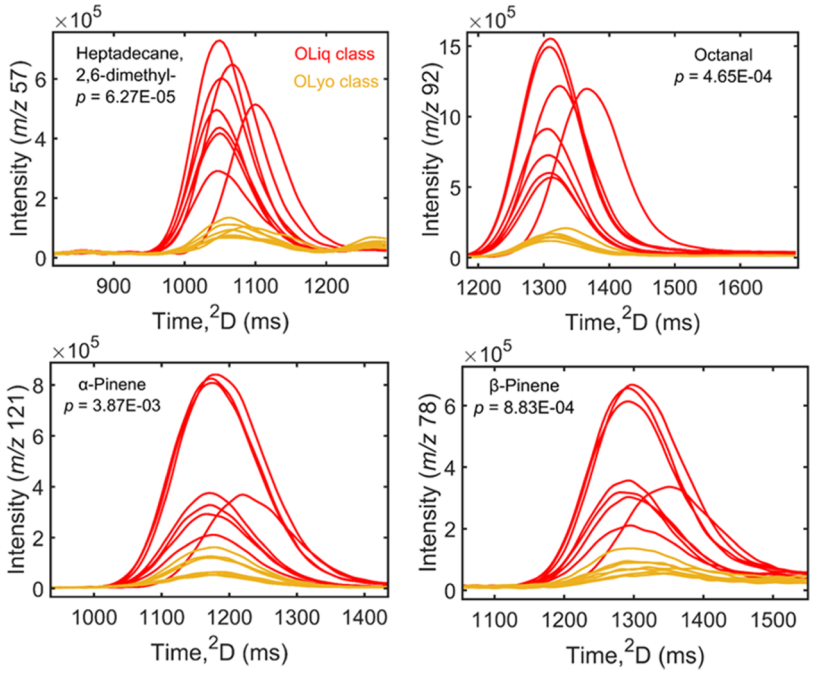

Figure 6. Illustration of the significant within-class variance in the omnivore liquid (OLiq) class samples. The plots correspond to the summed ^{2}D peaks of four analytes exclusively discovered by the tile-based Random Forest approach. Their p-values were calculated within the context of a balanced class comparison. α-Pinene and β-pinene were also discovered by the unbalanced random forest approach. The ^{2}D retention time windows portrayed in all of the plots correspond to the actual ^{2}D tile size.

observations, it can be inferred that the feature selection process in RF is governed by a rather more stringent methodology compared to the standard F-ratio approach, mainly because of how the permutation-based feature selection process operates. In the present two-class comparison context, m/z values with concentration ratios close to 1 are generally indicative that their corresponding features are approximately equally present in both classes. Permuting the values of such features is unlikely to disrupt model accuracy. Thus, RF assigns low to 0 importance scores to such nonclass distinguishing features. Moreover, the tile-based RF algorithm works in such a way that features with null importance scores are excluded from the final hit list. Conversely, being a statistical measure, the F-ratio does not inherently discriminate between features based on their concentration ratios. A nonzero F-ratio value is attributed to every discovered m/z and, subsequently, to its every associated feature as long as there are variations in the distribution of its values between the two classes, regardless of the magnitude and the origin of the variation.

Additionally, upon closer examination of the F-ratio and RF hit lists, an important observation can be made. Some analytes with high statistical significance (p-value <0.01) and concentration ratios greater than 2 were solely discovered by the F-ratio method. This observation is also evident in the F-ratio volcano plot that shows a greater number of m/z with p-value <0.01 and $[\text{OLiq}]/[\text{OLyo}] \geq 2$ compared to the plots associated with the BRF and URF analyses. As mentioned in a previous section, if a feature is not sampled by any of the

https://doi.org/10.1021/acs.analchem.3c01872
Anal. Chem. 2023, 95, 13519−13527

decision trees in the model, its values will not be permuted when assessing feature importance. It will not affect the model's performance and thus will be attributed an importance score of 0. In light of this, it is important to keep in mind that in some cases, an importance score of 0 does not necessarily mean that the feature is irrelevant to the classification task. It can also mean that the feature was simply not selected for the training of the decision trees. This could be a plausible explanation to why, in some cases, RF does not depict some statistically significant features, with high concentration ratios. To address this issue and reduce the risk of overlooking important features, increasing the number of decision trees used to build the RF model could be a potential solution. In fact, as the number of trees increases, the likelihood that an important feature will be selected by at least one decision tree increases.[43] However, one should keep in mind that increasing the number of trees in the model can lead to overfitting and, hence, decreasing the model's accuracy. It can also lead to a longer computational time and higher memory requirements, which cannot be feasible in certain settings. Therefore, in this study, we deemed it fit to continue working with the initial set of 300 decision trees.

Further scrutiny of the F-ratio and RF hit lists revealed that certain analytes with high statistical significance were exclusively discovered by the RF approach. Specifically, 8 compounds in the BRF hit list and 5 in the URF hit list were overlooked by the F-ratio method. This initially seemed intriguing, as the FRA is intended to provide a comprehensive final hit list. However, closer examination of the summed ^{2}D peaks of these analytes reveals a common pattern: a significant within-class variance in the omnivore liquid class when compared to the omnivore lyophilized class (Figures 6, S3, and S4). In fact, it was shown that when the within-class variance of one of the classes is substantial, the standard F-ratio approach may fail to identify the most important features. In this regard, an amended version of the standard F-ratio calculation was proposed by Prebihalo et al. that can tackle this issue, known as the control-normalized F-ratio method.[19] For the analytes in question, the OLiq samples exhibit not only a prominent ^{2}D retention time shift but also differences in signal intensities due to biological differences (Figure 6). This most likely severely impacted the standard F-ratio's ability to discover them. These results also suggest that the RF method is less sensitive to high within-class variances and has the potential to unravel class-distinguishing analytes even in challenging settings.

CONCLUSIONS

In this study, we introduced a novel nontargeted tile-based supervised analysis method called tile-based RF analysis. This new method combines the ML algorithm RF to chromatogram binning to facilitate the identification of class-distinguishing analytes in unbalanced class settings, presenting a significant advantage over the conventional tile-based F-ratio analysis. The efficiency of this method was assessed through a two-class comparative study involving stool samples stored under two different storage conditions: liquid and lyophilized. The method performed similarly in both balanced and unbalanced data set comparisons, generating hit lists with comparable class-distinguishing compounds. The performance of the RF method was further evaluated in comparison to that of the standard F-ratio analysis. To enable this comparison, the unbalanced data set was adjusted to a balanced setting by randomly disregarding one sample from the outnumbered class to meet the requirements of the F-ratio analysis. Despite both methods identifying similar analytes, the RF approach yielded a more stringent hit list mainly consisting of statistically significant analytes ($p < 0.01$) with concentration ratios greater than 2 or less than 0.5. In contrast, the F-ratio analysis produced a more comprehensive final hit list but contained a substantial number of false positives ($p \geq 0.01$). Moreover, the RF approach showed quite promising results in detecting statistically significant features in situations where high between-class variance was accompanied by a high within-class variance. This makes it a highly favorable method for analyzing complex biological samples that exhibit such tendency.

ASSOCIATED CONTENT

Supporting Information

The Supporting Information is available free of charge at https://pubs.acs.org/doi/10.1021/acs.analchem.3c01872.

Hit lists for the F-ratio, the balanced, and the unbalanced random forest analyses; optimized random forest hyperparameters; mass channel m/z purity criterion; F-ratio concentration ratio chromatogram; illustrations of the significant within-class variance in the OLiq class samples (PDF)

AUTHOR INFORMATION

Corresponding Author

Meriem Gaida − *Organic and Biological Analytical Chemistry Group, Molecular Systems Research Unit, University of Liège, 4000 Liège, Belgium;* orcid.org/0000-0003-4932-472X; Email: Meriem.gaida@uliege.be

Authors

Caitlin N. Cain − *Department of Chemistry, University of Washington, Seattle, Washington 98195-1700, United States;* orcid.org/0000-0001-8367-5799

Robert E. Synovec − *Department of Chemistry, University of Washington, Seattle, Washington 98195-1700, United States;* orcid.org/0000-0001-7032-3911

Jean-François Focant − *Organic and Biological Analytical Chemistry Group, Molecular Systems Research Unit, University of Liège, 4000 Liège, Belgium;* orcid.org/0000-0001-8075-2920

Pierre-Hugues Stefanuto − *Organic and Biological Analytical Chemistry Group, Molecular Systems Research Unit, University of Liège, 4000 Liège, Belgium;* orcid.org/0000-0002-1224-8869

Complete contact information is available at: https://pubs.acs.org/10.1021/acs.analchem.3c01872

Funding

This research was funded by the FWO/FNRS Belgium EOS Grant 30897864 "Chemical Information Mining in a Complex World", the University of Liège, F.R.S.-F.N.R.S, and Léon Fredericq Foundation scientific grants.

Notes

The authors declare no competing financial interest.

ACKNOWLEDGMENTS

The authors would like to thank LECO Corporation for its instrumental support as well as Merck for providing the SPME

NIT device. Additionally, the authors would like to acknowledge Thibaut Dejong for his contribution to the experimental measurements.

REFERENCES

(1) Trinklein, T. J.; Cain, C. N.; Ochoa, G. S.; et al. *Anal. Chem.* **2023**, *95*, 264−286.
(2) Hantao, L. W. *LCGC North Am.* **2023**, *41*, 105−111.
(3) Stefanuto, P. H.; Smolinska, A.; Focant, J. F. *TrAC, Trends Anal. Chem.* **2021**, *139*, No. 116251, DOI: 10.1016/j.trac.2021.116251.
(4) Bro, R.; Smilde, A. K. *Anal. Methods* **2014**, *6*, 2812−2831, DOI: 10.1039/c3ay41907j.
(5) Wold, S.; Sjöström, M.; Eriksson, L. *Chemom. Intell. Lab. Syst.* **2001**, *58*, 109−130, DOI: 10.1016/s0169-7439(01)00155-1.
(6) Lee, L. C.; Liong, C. Y.; Jemain, A. A. *Analyst* **2018**, *143*, 3526−3539, DOI: 10.1039/C8AN00599K.
(7) Marney, L. C.; Siegler, W. C.; Parsons, B. A.; et al. *Talanta* **2013**, *115*, 887−895.
(8) Parsons, B. A.; Marney, L. C.; Siegler, W. C.; et al. *Anal. Chem.* **2015**, *87*, 3812−3819.
(9) Trinklein, T. J.; Synovec, R. E. *J. Chromatogr. A* **2022**, *1677*, No. 463321.
(10) Ochoa, G. S.; Prebihalo, S. E.; Reaser, B. C.; Marney, L. C.; Synovec, R. E. *J. Chromatogr. A* **2020**, *1627*, No. 461401.
(11) Trinklein, T. J.; Jiang, J.; Synovec, R. E. *Anal. Chem.* **2022**, *94*, 9407−9414.
(12) Sudol, P. E.; Ochoa, G. S.; Synovec, R. E. *J. Chromatogr. A* **2021**, *1644*, No. 462092.
(13) Ochoa, G. S.; Sudol, P. E.; Trinklein, T. J.; Synovec, R. E. *Talanta* **2022**, *236*, No. 122844.
(14) Ochoa, G. S.; Billingsley, M. C.; Synovec, R. E. *Anal. Bioanal. Chem.* **2023**, *415*, 2411−2423.
(15) Zou, Y.; Gaida, M.; Franchina, F. A.; Stefanuto, P. H.; Focant, J. *Molecules* **2022**, *27*, No. 1806.
(16) Sudol, P. E.; Galletta, M.; Tranchida, P. Q.; et al. *J. Chromatogr. A* **2022**, *1662*, No. 462735, DOI: 10.1016/j.chroma.2021.462735.
(17) Mikaliunaite, L.; Synovec, R. E. *Talanta* **2022**, *244*, No. 123396, DOI: 10.1016/j.talanta.2022.123396.
(18) Spadafora, N. D.; Mascrez, S.; McGregor, L.; Purcaro, G. *Food Chem.* **2022**, *383*, No. 132438, DOI: 10.1016/j.foodchem.2022.132438.
(19) Prebihalo, S. E.; Ochoa, G. S.; Berrier, K. L.; et al. *Anal. Chem.* **2020**, *92*, 15526−15533.
(20) Schöneich, S.; Ochoa, G. S.; Monzón, C. M.; Synovec, R. E. *J. Chromatogr. A* **2022**, *1667*, No. 462868.
(21) Sudol, P. E.; Ochoa, G. S.; Cain, C. N.; Synovec, R. E. *Anal. Chim. Acta* **2022**, *1209*, No. 339847.
(22) Cain, C. N.; Trinklein, T. J.; Ochoa, G. S.; Synovec, R. E. *Anal. Chem.* **2022**, *94*, 5658−5666.
(23) Blanca, M. J.; Alarcón, R.; Arnau, J.; Bono, R.; Bendayan, R. *Behav. Res. Methods* **2018**, *50*, 937−962.
(24) Reichenbach, S. E.; Zini, C. A.; Nicolli, K. P.; et al. *J. Chromatogr. A* **2019**, *1595*, 158−167.
(25) Beccaria, M.; Mellors, T. R.; Petion, J. S.; et al. *J. Chromatogr. B* **2018**, *1074−1075*, 46−50.
(26) Purcaro, G.; Rees, C. A.; Melvin, J. A.; Bomberger, J. M.; Hill, J. E. *J. Breath Res.* **2018**, *12*, No. 046001.
(27) Beccaria, M.; Franchina, F. A.; Nasir, M.; et al. *Molecules* **2021**, *26*, No. 4600, DOI: 10.3390/molecules26154600.
(28) Janiesch, C.; Zschech, P.; Heinrich, K. *Electron. Mark.* **2021**, *31*, 685−695.
(29) Crisci, C.; Ghattas, B.; Perera, G. *Ecol. Modell.* **2012**, *240*, 113−122.
(30) Breiman, L. *Mach. Learn.* **2001**, *45*, 5−32.
(31) Parmar, A.; Katariya, R.; Patel, V.A Review on Random Forest: An Ensemble Classifier. In *International Conference on Intelligent Data Communication Technologies and Internet of Things (ICICI)*; Springer, 2019; Vol. *26*, pp 758−763.
(32) Fawagreh, K.; Gaber, M. M.; Elyan, E. *Syst. Sci. Control Eng.* **2014**, *2*, 602−609, DOI: 10.1080/21642583.2014.956265.
(33) Mathworks *Statistics and Machine Learning Toolbox User's Guide* 2021, 1 10862.
(34) Fratello, M.; Tagliaferri, R.Decision Trees and Random Forests. In *Encyclopedia of Bioinformatics and Computational Biology*; Elsevier, 2019; Vol. *1*−*3*, pp 374−383.
(35) Schmarr, H. G.; Bernhardt, J. *J. Chromatogr. A* **2010**, *1217*, 565−574.
(36) Tauler, R. *Chemom. Intell. Lab. Syst.* **1995**, *30*, 133−146.
(37) Stein, S. *Anal. Chem.* **2012**, *84*, 7274−7282.
(38) Probst, P.; Wright, M. N.; Boulesteix, A. *WIREs Data Min. Knowl. Discovery* **2019**, *9*, No. e1301, DOI: 10.1002/widm.1301.
(39) Probst, P.; Boulesteix, A. L. *J. Mach. Learn. Res.* **2017**, *18*, 6673−6690.
(40) Oshiro, T. M.; Perez, P. S.; Baranauskas, J. A.How Many Trees in a Random Forest?. In *Machine Learning and Data Mining in Pattern Recognition*; Springer, 2012; Vol. *7376*, pp 154−168.
(41) Díaz-Uriarte, R.; de Andrés, S. A. *BMC Bioinf.* **2006**, *7*, No. 3.
(42) Scornet, E. *ESAIM: Proc. Surv.* **2017**, *60*, 144−162.
(43) Hastie, T.; Tibshirani, R.; Friedman, J.Random Forests. In *Elements of Statistical Learning: Data Mining, Inference, and Prediction*; Springer, 2009; pp 587−604.
(44) Wu, J.; Chen, X.-Y.; Zhang, H.; et al. *J. Electron. Sci. Technol.* **2019**, *17*, 26−40.

https://doi.org/10.1021/acs.analchem.3c01872
Anal. Chem. 2023, 95, 13519−13527