Microchemical Journal 172 (2022) 106893

Contents lists available at ScienceDirect

Microchemical Journal

journal homepage: www.elsevier.com/locate/microc

Comparison of different processing approaches by SVM and RF on HS-MS eNose and NIR Spectrometry data for the discrimination of gasoline samples

Marta Barea-Sepúlveda [a], Marta Ferreiro-González [a,*], José Luis P. Calle [a], Gerardo F. Barbero [a], Jesús Ayuso [b], Miguel Palma [a]

[a] Department of Analytical Chemistry, Faculty of Sciences, University of Cadiz, Agrifood Campus of International Excellence (ceiA3), IVAGRO, Puerto Real, Cadiz 11510, Spain
[b] Department of Chemical-Physics, Faculty of Sciences, University of Cadiz, INBIO, Puerto Real, Cadiz 11510, Spain

ARTICLE INFO

Keywords:
Gasoline
Research Octane Number
HS-MS eNose
NIRS
Chemometrics
Support Vector Machine
Random Forest
Low-level data fusion

ABSTRACT

In the quality control of flammable and combustible liquids, such as gasoline, both rapid analysis and automated data processing are of great importance from an economical viewpoint for the petroleum industry. The present work aims to evaluate the chemometric tools to be applied on the Headspace Mass Spectrometry (HS-MS eNose) and Near-Infrared Spectroscopy (NIRS) results to discriminate gasoline according to their Research Octane Number (RON). For this purpose, data from a total of 50 gasoline samples of two types of RON-95 and 98-analyzed by the two above-mentioned techniques were studied. The HS-MS eNose and NIRS data were combined with non-supervised exploratory techniques, such as Hierarchical Cluster Analysis (HCA), as well as other supervised classification techniques, namely Support Vector Machine (SVM) and Random Forest (RF). For supervised classification, the low-level data fusion was additionally applied to evaluate if the combined use of the data increases the scope of relevant information. The HCA results showed a clear clustering trend of the gasoline samples according to their RON with HS-MS eNose data. SVM in combination with 5-Fold Cross-Validation successfully classified 100% of the samples with the HS-MS eNose data set. The RF algorithm in combination with 5-Fold Cross-Validation achieved the best accuracy rate for the test set with the low-level data fusion system. Furthermore, it allowed us to identify the most important features that could define the differences between RON 95 and RON 98 gasoline. On the other hand, using the HS-MS eNose and NIRS low-level data fusion reached better results than those obtained using NIRS data individually, with accuracy rates of 100% in both SVM and RF performances with the test set. In general, the performance of the SVM and RF algorithms was found to be similar.

1. Introduction

Gasoline is a product of petroleum refining that contains practically all kinds of volatile hydrocarbons in the C_4 to C_{12} range and aromatics compounds [1,2]. However, its composition, and therefore the quality of this petroleum product, varies according to refineries, fuel properties, and type based on its Research Octane Number (RON) [3]. Therefore, the characterization of the different sorts of gasoline (e.g.: RON 95 and RON 98) could be of high relevance for numerous reasons concerning the quality control of this combustible. There are several methods to determine the RON value, from the reference method established by the American Society for Testing and Materials, the ASTM 2699 [4], to several spectroscopic methods using multivariate regression [5] and some protocols ASTM E1655 and ASTM D6122 [6,7]. However, there is also interest in the on-site determination of RON values using portable devices [8]. For example, both the Headspace Mass Spectrometry (HS-MS eNose) and Near-Infrared Spectroscopy (NIRS) allow for the on-site application and provide interesting information from both volatile and non-volatile compounds in fuels. Literature indicates that HS-MS eNose methodologies have been successfully applied in combination with chemometrics for discrimination purposes of gasoline types according to their RON [9–13]. On the other hand, the application of spectroscopic

* Corresponding author.
E-mail address: marta.ferreiro@uca.es (M. Ferreiro-González).

https://doi.org/10.1016/j.microc.2021.106893
Received 10 September 2021; Received in revised form 30 September 2021; Accepted 1 October 2021
Available online 6 October 2021

techniques complemented by multivariate analysis, such as Near-Infrared Spectroscopy (NIRS), has also been described as an accurate, non-destructive, and rapid technique in this field [5,8,10,14–17]. Nonetheless, the HS-MS eNose and NIR Spectrometry techniques provide a large amount of information in a limited period. To manage this volume of data, it has become essential to use chemometric techniques that enable data transformation into interpretable information. Thus, combining these analytical techniques with the appropriate chemometric tools can enhance their capabilities [18]. Hierarchical Cluster Analysis (HCA) and Principal Components Analysis (PCA) are unsupervised pattern recognition techniques generally applied for handling the multivariate data without previous knowledge about the samples' classification tendency [19,20]. Notwithstanding, these unsupervised algorithms do not allow for future predictions, so it is required to recur to the application of learning/supervised algorithms to generate predictive classification or regression models. The classification and regression supervised algorithms work differently despite the similarity in the overall objective (assign inputs to outputs based on input-output assignments). In classification problems, the algorithm learns a function to map inputs to outputs in which the output value is a discrete class label, whereas regression problems attempt to map inputs to outputs where the output is a continuous value [21]. Nevertheless, despite the potential of regression algorithms, classification algorithms offer certain advantages when the main goal is to facilitate data processing for on-site portable operations [16].

Numerous methods are available for obtaining supervised models for sample classification, but Linear and Quadratic Discriminant Analysis (LDA and QDA) have long been the most employed in petroleum-based products research [22–24]. Nevertheless, in cases where strong similarities or high within-group variability are observed, an efficient separation of non-linear regions is often difficult to achieve with a linear method such as LDA. In these situations, a multivariate technique with higher performance is needed. Within this framework, Support Vector Machine (SVM) and Random Forest (RF) are a new generation of non-parametric learning algorithms that have been brought into chemometrics for classification and regression tasks [25,26], achieving successful results with HS-MS eNose and NIRS data matrices [27–30]. SVM is a set of supervised learning algorithms that shows robust generalization performance and can model non-linear boundaries by using the kernel functions, such as the radial basis function kernel (RBF kernel) [27]. RF is a bootstrapping algorithm that generates several decision trees for prediction or class assignment [31]. One advantage of this algorithm is the extremely rapid decision tree construction, and therefore the training speed of hundreds of them is much faster than training an artificial neural network [29]. Whereas, when SVM and RF are applied to solve real problems, the model parameter selection is a fundamental consideration, as it may influence the model's accuracy and performance [32]. At the same time, the application of different analytical techniques to describe a particular phenomenon has increased the amount of available information. Consequently, combining the results from different types of measurements into a single data matrix, also known as data fusion, is a widely applied methodology in many fields such as analytical chemistry, biology, and computer science not only for obtaining a better understanding of the studied phenomena but also for increasing the accuracy rate when a supervised algorithm is applied [33–37]. To mention some examples, Qui et al. (2015) applied the data fusion methodology on the E-tongue and E-nose data in combination with chemometrics techniques, such as SVM and RF, to trace the quality status of mandarin [29]. On the other hand, Li et al. (2019) used the data fusion methods on the Raman and Near-Infrared Spectroscopies data sets for a rapid analysis of methanol in gasoline [38].

Therefore, due to all the above-mentioned, this study aimed to evaluate the application of the SVM and RF algorithms on the HS-MS eNose and NIRS data for the classification of RON 95 and RON 98 gasoline samples provided by different Spanish refineries. As explained before, there are some regression methods allowing for good prediction of the RON value based on spectroscopic data; however, for both the producers and the consumers, a reliable discrimination method between the two commercially available gasoline products is much interesting. Additionally, a low-level data fusion method was applied for creating the HS-MS eNose and NIRS fusion system to assess whether the simultaneous use of both data sets increases the scope of useful information and lead to achieve better accuracy rates in the supervised models. Furthermore, hyperparameter tuning was also conducted to achieve the most accurate SVM and RF models for each data set.

2. Materials and methods

2.1. Samples

The two commonly available consumer gasoline types in Spain, RON 95 and 98, were selected for this study. Concretely, a total of 50 samples belonging to 25 RON 95 and 25 RON 98 gasoline from different Spanish refineries were analyzed.

2.2. HS-MS eNose spectrum acquisition and HS-MS eNose data set

All the gasoline samples were analyzed using an HS-MS Alpha Moss system (Toulouse, France) based on an HS 100 static headspace autosampler and an α Kronos quadrupole mass spectrometer (MS). The samples (80 μL) were stored in 10 mL sealed vials (Agilent Crosslab), and these were placed into the autosampler oven to be heated and agitated to generate the headspace. Headspace was taken from the vial employing a gas syringe and then injected into the mass spectrometer detector without any chromatographic separation. The experimental HS and MS conditions used for the analysis were previously optimized and described by Ferreiro et al. (2014) [11]: incubation temperature 145 °C, incubation time 10 min, 500 rpm of agitation speed, fill speed 100 μL/s, injection volume 4.5 mL, syringe type 5 mL, syringe temperature 150 °C, injection speed 75 μL/s, and flushing time 120 s. Furthermore, 2 μL of perfluorotributylamine (TBPFA) were added to all samples as an internal standard. Instrumental control was achieved using the Residual Gas Analysis software package and Alpha Soft 7.01 software (Alpha Moss, Toulouse, France). Total Ion Spectra (TIS) from gasoline samples were obtained and arranged into a two-dimension data matrix (D*nxm*) to form the HS-MS eNose data set, where *n* is the number of gasoline samples ($n = 50$), and *m* is the *m/z* intensities ($m = 156$). All the TIS were normalized at the *m/z* of 131, which is the significant *m/z* of the internal standard of TBPFA [12].

2.3. NIRS spectrum acquisition and NIRS data set

Each NIR spectra was recorded at room temperature using an AvaSpec-NIR 256–1.7 (Avantes, Louisville, CO, USA) equipped with a tungsten halogen lamp and a transmittance probe with a path length of 10 mm. All the gasoline samples were analyzed in the range of 891–1812 nm and with a spectral resolution of 3.4 nm. The NIR spectra for each gasoline sample was placed in a two-dimension data matrix (D*nxp*) to form the NIRS data set, where *p* is the number of absorbance values ($p = 256$) without any pretreatment and *n* the number of gasoline samples ($n = 50$).

2.4. HS-MS eNose/NIRS fusion data set

Data fusion is the integration of multiple sources of information to generate a more specific and complete data set. In this study, the low-level data fusion method was used to generate the HS-MS eNose and NIRS fusion matrix. This fusion method consists of concatenating signals from different analytical instruments to form a single matrix where the rows are equal to the number of samples analyzed and the columns are formed by signals (variables). For our specific case, the HS-MS eNose and NIRS fusion matrix (D*nxz*) consisted of a data set where *n* is the

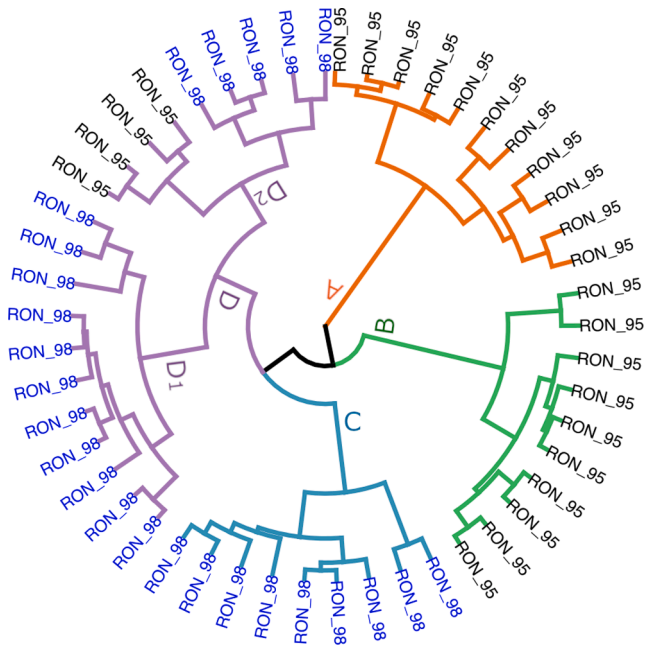

Fig. 1. Circular dendrogram resulting from the Hierarchical Cluster Analysis (HCA) based on the HS-MS eNose data set (D_{50x156}). The gasoline samples were colored according to their RON: black for RON 95 and blue for RON 98.

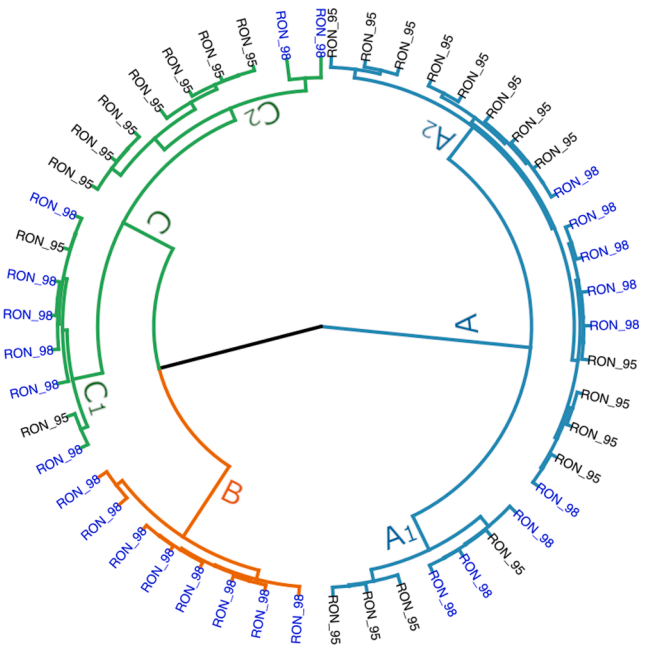

Fig. 2. Circular dendrogram resulting from the Hierarchical Cluster Analysis (HCA) based on the NIRS data set (D_{50x256}). The gasoline samples were colored according to their RON: black for RON 95 and blue for RON 98.

number of gasoline samples ($n = 50$), and z is the number of absorbance and m/z values obtained from the measurements of the two analytical techniques ($z = 412$). Min-Max normalization was applied to the low-level data fusion matrix as a pretreatment.

2.5. Multivariate analysis and software

All data analyses were performed with RStudio (R version 4.0.5, Boston, MA, USA). Non-supervised analysis, namely Hierarchical Cluster Analysis (HCA), was performed using the *hclust* function from the *stats* package. The choice of the Linkage method for the HCA was established by calculating and comparing the agglomerative coefficient obtained from different Linkage's methods (Average, Single, Complete, and Ward) using the *agnes* function of the *clust* package. The supervised analysis, including Support Vector Machine (SVM) and Random Forest (RF), were carried out by using the *caret* package in the R Project for Statistical Computing.

3. Results and discussion

3.1. Exploratory study

Firstly, the tendency of the gasoline samples to cluster according to their RON was tested. For this purpose, each sample's ($n = 50$) HS-MS eNose spectrum normalized at the m/z of 131 ($m = 156$), and each sample's ($n = 50$) NIRS spectrum without any pretreatment ($m = 256$) was subjected to a Hierarchical Cluster Analysis (HCA). This class of clustering method produces a hierarchical classification of data based on their similarity. For this analysis, Euclidean distance was chosen for inter-individual similarity matrix calculation, and Ward's method was selected as the inter-group measure. The choice of Ward's method in the present study was established by calculating and comparing the agglomerative coefficient obtained from different Linkage methods (Average, Simple, Full, and Ward). This coefficient allows finding the Linkage method that can identify stronger clustering structures, with values closer to 1 suggesting a strong clustering structure. In this case, with the Ward's method the highest agglomerative coefficients were obtained with values of 0.94 and 0.98 for the HS-MS eNose and NIRS data, respectively, thus identifying the strongest clustering structure of the four methods evaluated. The results of the HCA have been graphically displayed in the circular dendrograms in Fig. 1 and Fig. 2.

The resulting dendrogram for HS-MS eNose (Fig. 1) shows a notable trend for the samples to be classified according to their RON, with all gasoline samples grouped into four main clusters (A, B, C, and D). Clusters A and B exclusively contain samples of a single octane rating, namely RON 95 (black). In addition, Cluster C includes only gasoline samples of RON 98 (blue). Meanwhile, Cluster D is divided into two subclusters, D_1 and D_2. It can be observed that the RON 98 gasoline samples have a greater tendency to be classified within both clusters. Nonetheless, in contrast to subcluster D_1, the D_2 subcluster also contains samples of RON 95. Thus, the results obtained through this analysis seem to indicate that the volatile organic compounds (VOCs) present in both types of gasoline allow them to be discriminated against according to their RON. Nevertheless, this clustering was not completely consistent since there is not a perfect separation. The HCA results for NIRS (Fig. 2) revealed that the gasoline samples tended to group into 3 principal clusters (A, B, and C). Unlike the HS-MS eNose dendrogram, the tendency to cluster according to RON is somewhat less well defined with the NIRS data set. Here, only Cluster B contains samples of one octane rating type, specifically RON 98 (blue). For their part, Clusters A and C are subdivided into two subclusters and contain samples of both RON types. Subcluster A_2 contains 66% of RON 95 (black) and 34% of RON 98 gasoline samples, while subcluster A_1 contains 57% of RON 95 and 43% of RON 98 gasoline. Regarding Cluster C, subcluster C_1 is mainly formed by RON 98 gasoline samples and subcluster C_2 by RON 95 gasoline samples.

Given the results obtained through HCA, the application of multivariate techniques that include supervised pattern recognition algorithms is required to enable an accurate classification and to guarantee the generation of a mathematical model allowing for future predictions. For this purpose, two supervised classification methods, namely SVM and RF, were applied and compared to predict the octane number of gasoline samples from the HS-MS eNose and NIRS data. The evaluation of the generated SVM and RF models was carried out using accuracy as a metric, which is the ratio between the number of correct predictions and the total number of input samples. In addition, these two algorithms

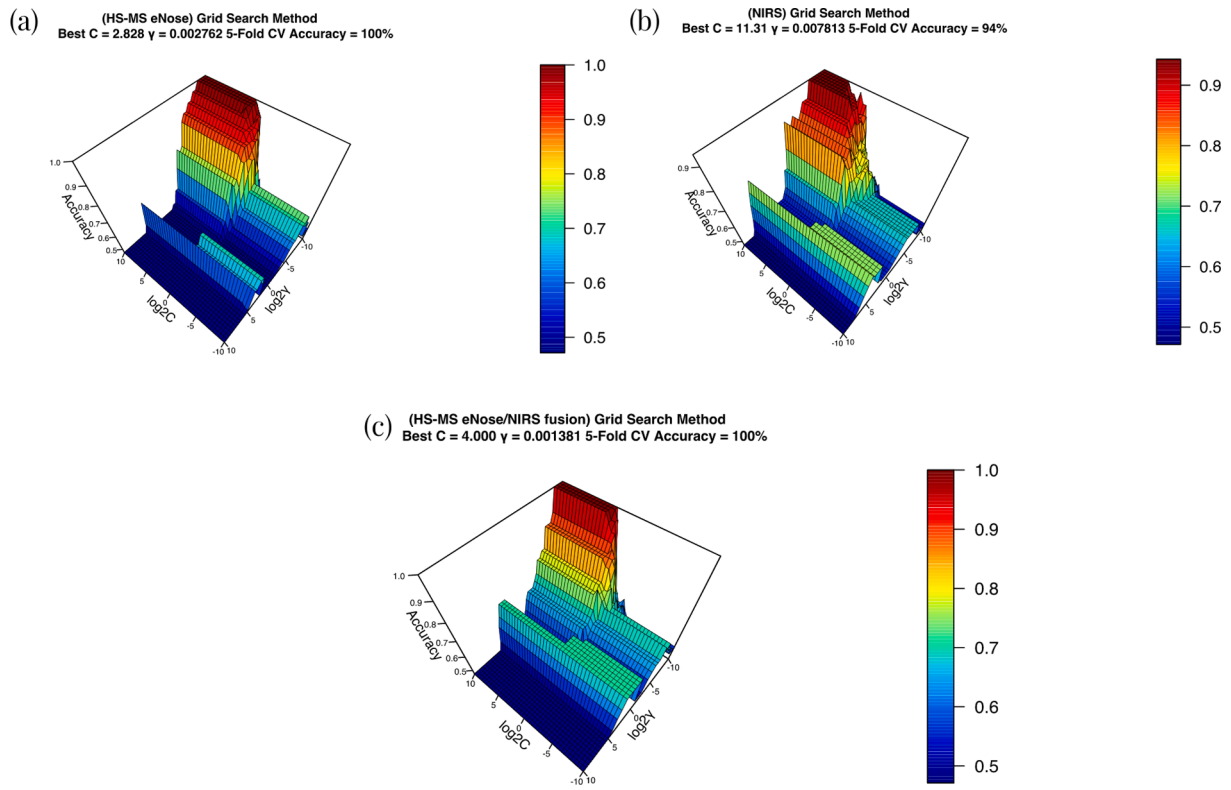

Fig. 3. The Grid Search Method results for the searching of the best C and γ according to the 5-Fold Cross-Validation accuracy rate: (a) Based on the HS-MS eNose output; (b) Based on the NIRS output; (c) Based on the HS-MS eNose/NIRS fusion output.

were also applied on the low-level data fusion matrix of the two systems to evaluate whether it is possible to achieve better accuracy rates.

3.2. Classification based on Support Vector Machine (SVM)

Support Vector Machine (SVM) is a non-parametric supervised algorithm commonly used to classify data into different classes. Compared to LDA, where the classification of samples is based on Fisher's linear discriminant function, the main idea of the SVM method is to find the optimal hyperplane (boundary) that maximizes the margin between the support vectors (data points closest to the hyperplane), which results in the segregation of the classes with a lower classification error. As in LDA, the SVM algorithm is generally applied to a data set where the response variables are linearly separable. Nevertheless, unlike LDA, when the problems that we encounter are not linearly separable the SVM algorithm can be used for classifying by using the kernel trick, which means transforming data into another dimension with a clear dividing margin between classes. For this purpose, each data set was randomly divided into the training set (split = 0.7), resulting in the training set containing 35 samples, and the test set, containing the remaining 15 samples. Radial basis function (RBF) was chosen as the core function. In the SVM RBF Kernel, there are only two parameters that need to be tuned: the penalty factor (C) and the kernel parameter (γ). The former controls the number of support vectors and the balance between bias and variance, while the latter controls the behavior of the Gaussian kernel.

To optimize these two hyperparameters, the Grid Search method with the exponential growth of the C and γ was selected. Here, $\log_2$C and $\log_2\gamma$ were in the range from –10 to 10 at 0.5 intervals. Each combination of parameter choices was checked by using 5-Fold Cross-Validation, and the parameters with the best cross-validation accuracy were selected. Note that in 5-Fold Cross-Validation the training set was divided into five subsets of equal size. Sequentially, one subset was tested using the classifier trained on the remaining four subsets. This process was repeated for each one of the subsets. Therefore, 8405 models were generated, i.e., 41x41 (C and γ combinations) ×5 (subsets). The searching of the best C and γ parameters with the three data sets is presented in Fig. 3, representing the $\log_2\gamma$ (y-axis), $\log_2$C (x-axis), and the accuracy rate obtained (z-axis) on a surface plot. On the one hand, it can be observed that the accuracy rate increases with higher values of C for all three systems and, since this hyperparameter affects the bias-variance trade-off, this suggests the hyperplane would allow less misclassified observations and, therefore, there would be fewer support vectors, resulting in a less biased model but with a higher variance. On the other hand, it is possible to see that the best results were obtained

Table 1
Comparison of classification results of HS-MS eNose, NIRS, and HS-MS eNose/NIRS fusion systems based on the Support Vector Machine algorithm.

Dataset	Parameters			Accuracy Rate for 5-Fold Cross Validation (%)	Accuracy Rate for Training set (%)	Accuracy Rate for Test set (%)
	Penalty Factor (C)	Kernel Parameter (γ)	nSVs[a]			
HS-MS eNose system	2.828	0.002762	26	100	100	100
NIRS system	11.31	0.007813	23	94	100	93
HS-MS eNose/NIRS fusion system	4.000	0.001381	28	100	100	100

[a] nSVs: Number of Support Vectors (SVs).

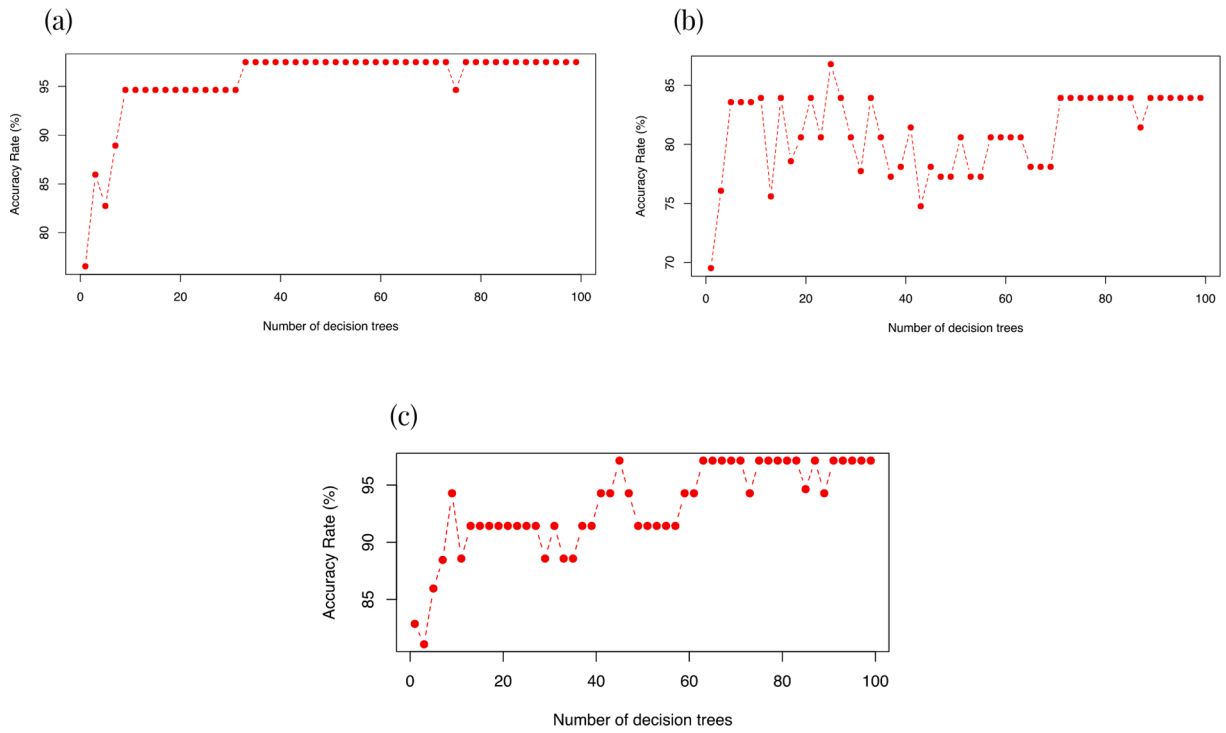

Fig. 4. The RF performance according to the number of decision trees: (a) Based on the HS-MS eNose output; (b) Based on the NIRS output; (c) Based on the HS-MS eNose/NIRS fusion output.

with the lowest γ values, suggesting that the limit is almost linear. For this reason, the most accurate values of C and γ in the 5-Fold Cross-Validation were selected to avoid overfitting and to obtain a good accuracy rate. Firstly, for the HS-MS eNose system (Fig. 3a) the best value for C was 2.828 ($\log_2 C = 1.5$) and for γ was 0.002762 ($\log_2 \gamma = -8.5$). Secondly, for the NIRS system (Fig. 3b) the best C value was 11.31 ($\log_2 C = 3.5$) and the best γ value was 0.007813 ($\log_2 \gamma = -7$). Finally, for the fusion system (Fig. 3c) the best γ was 0.001381 ($\log_2 \gamma = -9.5$) and the C was 4.000 ($\log_2 C = 2$).

The evaluation of the SVM models' performance was carried out through the 5-Fold Cross-Validation and the training and test set accuracy rates. As seen in Table 1, the performance of HS-MS eNose and fusion systems were 100% accurate in the 5-fold Cross-Validation set, training set and test set. On the other hand, the SVM model was satisfied with an accuracy rate of 100 % in the training set. Nonetheless, the 5-Fold Cross-Validation and test sets for the NIRS system were 94 and 93% accurate, respectively. Results obtained through the SVM models confirmed the applicability of these analytical techniques for the discrimination of gasoline according to their RON. Especially with the HS-MS eNose system, whose results are highly promising for this purpose given the excellent performance of the model, thus indicating that volatile organic compounds (VOCs) are more suitable in terms of gasoline discrimination according to the octane number. In addition, the application of this technique in this field can be an alternative to conventional analytical techniques such as GC–MS, offering multiple advantages like faster analysis, lower costs, easy to handle in routine analysis, and absence of residues [10,16]. On the other hand, the accuracy rates obtained for the HS-MS eNose indicate that the application of low-level data fusion to achieve the perfect discrimination of gasoline samples would not be necessary when applying the SVM algorithm due to the excellent performance of this technique for this purpose.

The SVM model has shown excellent performance. However, the nature of the algorithm itself does not allow the selection of the most relevant features for the construction of the model. Therefore, another non-parametric technique known as Random Forest (RF) was used to pursue this goal to examine the variables that could define the differences between RON 95 and RON 98 gasoline samples for classification purposes.

3.3. Classification based on Random Forest (RF)

The Random Forest (RF) is a widely used non-parametric supervised algorithm for classification and regression tasks. It consists of several independent decision trees running as an ensemble. Each tree grows on a bootstrap sample taken with replacement from the original data, which means that 2/3 of the original data, known as ''inside the bag'' data, is used for training, and 1/3 of the original data, known as ''out of bag'' (OOB) data, is used for testing. As an ensemble model, each tree in the random forest votes for a category prediction, and the top-voted one is then used to make the final prediction. In RF there are a few hyperparameters that need to be tuned. Nevertheless, the *mtry* and *ntree* parameters are, perhaps, the most likely to have the highest significant effects on the final accuracy. The former is the number of variables to be randomly selected in each partition for each tree in the forest, while the latter is the number of trees to grow.

Prior to the hyperparameter optimization, each data set was randomly divided into the training set (split = 0.7), containing 35 samples, and the test set (split = 0.3), containing the remaining 15 samples. For classification purposes, the square root of the total number of predictors is generally used as the optimum value of *mtry*. Therefore, based on the variable dimension of the three feature sets (HS-MS eNose, NIRS, and HS-MS eNose/NIRS fusion), the *mtry* values were held constant at 12.49, 16.03, and 20.30, respectively. Besides, it is necessary to establish a specific number of trees to be used in the RF. In this sense, a large number of trees would not imply a risk of overfitting, although it could have repercussions in terms of longer computational times. To determine the number of trees to be used, the *ntree* values in this study were set from 2 to 100 at 2 trees interval and the 5-Fold Cross-Validation accuracy was considered as the evaluation criteria. The results for each system are graphically displayed in Fig. 4. As can be seen, the accuracy rate in the HS-MS eNose (Fig. 4a) tends to stabilize from 40 decision trees and is maintained up to 75 decision trees. Then, a decrease in

Table 2
Comparison of classification results of HS-MS eNose, NIRS, and HS-MS eNose/NIRS fusion systems based on the Random Forest algorithm.

Dataset	Parameters		Accuracy Rate for 5-Fold Cross Validation (%)	Accuracy Rate for Training Set (%)	Accuracy Rate for Test Set (%)
	mtry	*ntree*			
HS-MS eNose system	12.49	100	98	100	93
NIRS system	16.03	100	84	100	93
HS-MS eNose/NIRS fusion system	20.30	100	97	100	100

accuracy is observed at 76 decision trees. Nevertheless, after 77 trees, the accuracy rate stabilizes again and is maintained up to 100 trees. In this case, the number of decision trees for the HS-MS eNose system should be from 77 to 100. On the other hand, the accuracy rate in the NIRS system (Fig. 4b) tends to stabilize at 78 decision trees, however, a decrease can be observed at 88 trees. Then, at 89 trees the accuracy rate increases and remains stable up to 100. According to these results, the number of decision trees for the NIRS system should be in the range of 89 – 100. Finally, it is observed that the accuracy rate in the fusion system tends to stabilize at 67 trees, although, decreases for the accuracy rate are observed up to 91 trees, where a complete stabilization of accuracy is observed up to 100 trees. For this reason, the number of decision trees for the fusion system should be in the range of 91 – 100. Therefore, based on the results obtained and in order to find a compromise between computation time and stabilization of the accuracy rate, 100 trees were chosen for the three systems.

The performance of the Random Forest models for each system is presented in Table 2. According to the results, the accuracy of the models in the training set was excellent, reaching 100% in all three systems. Besides, the 5-fold cross-validation sets were found to be accurate in the range of 84 – 98%, obtaining the highest value for HS-MS eNose and the lowest for NIRS. As for the test set, 93 – 100% accuracy was satisfactory for all three systems. In this case, the simultaneous use of HS-MS eNose and NIRS (fusion data) results provides the highest accuracy rate (100%) in the test set, indicating that the RF algorithm requires the application of low-level data fusion to obtain a perfect classification of gasoline according to its octane rating.

Considering the nature of the RF model, the most relevant features for classification can be selected. In this case, the *varImp* function from the *caret* package in R has been used to estimate the contribution of each variable to the model. For RF this function computes the prediction accuracy in the out-of-bag portion of the data for each tree. Then, the same is then done after permuting each predictor variable. Finally, the difference between the two accuracies is averaged over all trees and normalized by the standard error [39]. The top 20 features and their relative importance in the RF models with the three data sets are shown in Fig. 5. Among the 20 most important features, there are some which present a greater contribution (≥80 of relative importance) in the creation of the RF models. In the case of HS-MS eNose, *m/z* 59 has the highest contribution, followed by *m/z* 60. The wavelengths that provide the most information for the creation of the RF model using the NIRS data are 1614, 1169 and 1199 nm, being 1614 nm the one with the highest relative importance. Moreover, using the low-level data fusion of the HS-MS eNose and NIRS data sets, it was found that the most influential features are *m/z* 59 and *m/z* 87, observing that the contribution of the NIRS variables has a lower relative importance compared to those of the HS-MS eNose, thus indicating that the later technique would provide more relevant information in terms of classification of gasoline

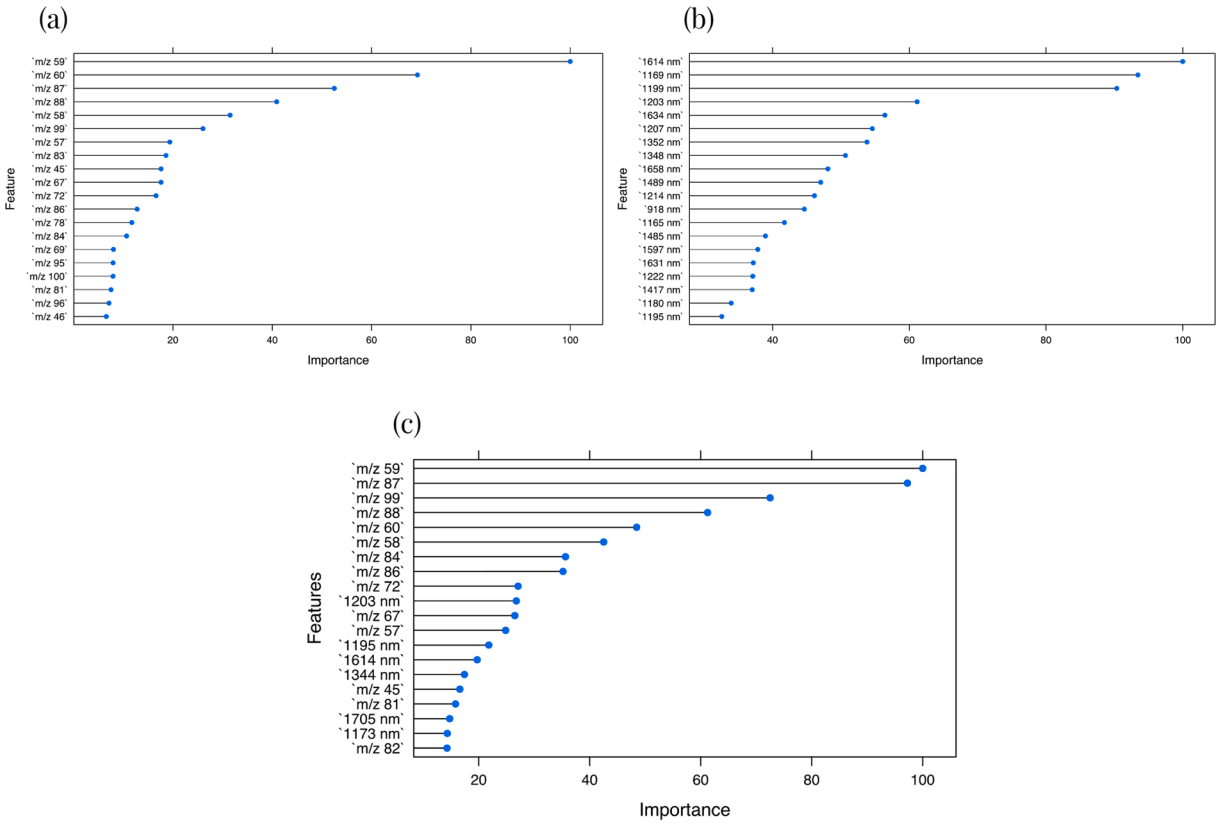

Fig. 5. List of the 20 most important features in the RF model and their relative importance: (a) Based on the HS-MS eNose output; (b) Based on the NIRS output; (c) Based on the HS-MS eNose/NIRS fusion output.

according to its Research Octane Number. However, it is also needed for the NIRS data to reach a 100% accuracy rate for the test set.

4. Conclusions

The combination of HS-MS eNose and NIRS together with chemometric tools have proven to be suitable analytical techniques for the characterization of gasoline according to its RON. HCA was able to group RON 95 and 98 gasoline samples using HS-MS eNose data, but this approach was not as successful when using NIRS data. Nevertheless, the overall information extracted from HS-MS eNose and NIRS allowed the accurate dis-crimination of these two categories (RON 95 and 98) using non-parametric tools, such as SVM and RF. The HS-MS eNose data achieved an excellent performance with an accuracy rate of 100% after searching for the best hyperparameters for SVM. Satisfactory results were also obtained for the NIRS data with an accuracy of 94%, 100%, and 93% for the 5-Fold Cross-Validation, training, and test sets. The RF algorithm also displayed a great performance, reaching the highest accuracy of 98%, 100%, and 93% in the 5-Fold Cross-Validation, training, and test sets with the HS-MS eNose data. Furthermore, the HS-MS eNose and NIRS fusion data achieved 97 – 100% accuracy rates in the SVM and RF performances, reaching better results than those obtained using the NIRS data separately. Nevertheless, for the RF, it was possible to verify that this is since the variables that contribute the most to the model developed with the data fusion are those of the HS-MS eNose. The aforementioned would indicate that this technique gives more relevant information in terms of the classification of gasoline according to its RON; however, NIRS data is also needed to reach 100% accuracy in the test set. Meanwhile, the performance of the SVM and RF algorithms was in general similar for the 95 and 98 RON gasoline samples discrimination.

To sum up, the results obtained in the present study demonstrate how the HS-MS eNose and NIRS techniques, in combination with suitable chemometric tools such as SVM and RF, can be an alternative to conventional interpretation methods for analysts to evaluate analytical results in a faster and, above all, objective approach.

CRediT authorship contribution statement

Marta Barea-Sepúlveda: Formal analysis, Data curation, Software, Visualization, Writing – original draft, Writing – review & editing. **Marta Ferreiro-González:** Conceptualization, Investigation, Data curation, Resources, Funding acquisition, Project administration, Supervision, Writing – review & editing. **José Luis P. Calle:** Software, Writing – review & editing. **Gerardo F. Barbero:** Investigation, Writing – review & editing. **Jesús Ayuso:** Methodology, Validation. **Miguel Palma:** Conceptualization, Methodology, Resources, Funding acquisition, Project administration, Supervision, Writing – review & editing.

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.

Acknowledgments

Marta Barea-Sepúlveda gratefully thanks the University of Cadiz and the Cátedra Fundación Cepsa for a Ph.D. contract under the program FPI UCA/TDI-4-19. The authors are grateful to the Instituto de Investigación Vitivinícola y Agroalimentario (IVAGRO) for providing the necessary facilities to carry out this research.

Funding

This work has been co-financed by the 2014-2020 ERDF Operational Programme and by the Department of Economy, Knowledge, Business and University of the Regional Government of Andalusia. Project reference: "FEDER-UCA18-107214".

References

[1] E. Stauffer, J.A. Dolan, R. Newman, Flammable and Combustible Liquids, A. Press (Ed.). Fire Debris Anal. Elsevier. 2008. 199–233. https://doi.org/10.1016/b978-012663971-1.50011-7.
[2] M.A. Kamrin, Gasoline, in: Encycl. Toxicol. Third Ed. Elsevier. 2014. 700–701. https://doi.org/10.1016/B978-0-12-386454-3.00391-2.
[3] I. Barnett, M. Zhang, Discrimination of brands of gasoline by using DART-MS and chemometrics, Forensic Chem. 10 (2018) 58–66, https://doi.org/10.1016/j.forc.2018.07.003.
[4] ASTM, D2699–16e1, Standard Test Method for Research Octane Number of Spark-Ignition Engine Fuel, West Conshohocken, PA (2016), https://doi.org/10.1520/D2699-16E01.
[5] A.A. Kardamakis, N. Pasadakis, Autoregressive modeling of near.IR spectra and MLR to predict RON values of gasolines, Fuel. 89 (1) (2010) 158–161, https://doi.org/10.1016/j.fuel.2009.08.029.
[6] ASTM, E1655, Standard practices for infrared multivariate quantitative analysis, West Conshohocken, PA (2005).
[7] ASTM, D6122, Standard practice for validation of the performance of multivariate process infrared spectropho- tometers, West Conshohocken, PA (2009).
[8] M. Voigt, R. Legner, S. Haefner, A. Friesen, A. Wirtz, M. Jaeger, Using fieldable spectrometers and chemometric methods to determine RON of gasoline from petrol stations: A comparison of low-field ^{1}H NMR@80 MHz, handheld RAMAN and benchtop NIR, Fuel. 236 (2019) 829–835, https://doi.org/10.1016/j.fuel.2018.09.006.
[9] M. Ferreiro-González, J. Ayuso, J.A. Álvarez, M. Palma, C. G. Barroso, New headspace-mass spectrometry method for the discrimination of commercial gasoline samples with different research octane numbers, Energy and Fuels. 28 (10) (2014) 6249–6254, https://doi.org/10.1021/ef5013775.
[10] M. Ferreiro-González, J. Ayuso, J.A. Álvarez, M. Palma, C.G. Barroso, Gasoline analysis by headspace mass spectrometry and near infrared spectroscopy, FUEL. 153 (2015) 402–407, https://doi.org/10.1016/j.fuel.2015.03.019.
[11] M. Ferreiro-González, G. Barbero, M. Palma, J. Ayuso, J. Álvarez, C. Barroso, Determination of ignitable liquids in fire debris: Direct analysis by electronic nose, Sensors (Switzerland). 16 (5) (2016) 695, https://doi.org/10.3390/s16050695.
[12] M. Ferreiro-González, G.F. Barbero, J. Ayuso, J.A. Álvarez, M. Palma, C.G. Barroso, Validation of an HS-MS method for direct determination and classification of ignitable liquids, Microchem. J. 132 (2017) 358–364, https://doi.org/10.1016/j.microc.2017.02.022.
[13] B. Falatová, M. Ferreiro-González, J.L.P. Calle, J.Á. Álvarez, M. Palma, Discrimination of ignitable liquid residues in burned petroleum-derived substrates by using HS-MS eNose and chemometrics, Sensors (Switzerland). 21 (2021) 1–12, https://doi.org/10.3390/s21030801.
[14] R.M. Correia, E. Domingos, V.M. Cáo, B.R.F. Araujo, S. Sena, L.U. Pinheiro, A. M. Fontes, L.F.M. Aquino, E.C. Ferreira, P.R. Filgueiras, W. Romão, Portable near infrared spectroscopy applied to fuel quality control, Talanta. 176 (2018) 26–33, https://doi.org/10.1016/j.talanta.2017.07.094.
[15] D. Özdemir, Determination of octane number of gasoline using near infrared spectroscopy and genetic multivariate calibration methods, Pet. Sci. Technol. 23 (9-10) (2005) 1139–1152, https://doi.org/10.1081/LFT-200035547.
[16] R.M. Balabin, R.Z. Safieva, E.I. Lomakina, Gasoline classification using near infrared (NIR) spectroscopy data: Comparison of multivariate techniques, Anal. Chim. Acta. 671 (1-2) (2010) 27–35, https://doi.org/10.1016/j.aca.2010.05.013.
[17] V.K. Yadav, K. Nigam, A. Srivastava, Forensic investigation of arson residue by infrared and Raman spectroscopy: From conventional to non-destructive techniques, Med. Sci. Law. 60 (3) (2020) 206–215, https://doi.org/10.1177/0025802420914807.
[18] K. Héberger, Chemoinformatics-multivariate mathematical-statistical methods for data evaluation. In: Med. Appl. Mass Spectrom. Elsevier, 2008. 141–169. https://doi.org/10.1016/B978-044451980-1.50009-4.
[19] E.E. Waddell, J.L. Frisch-Daiello, M.R. Williams, M.E. Sigman, Hierarchical cluster analysis of ignitable liquids based on the total ion spectrum, J. Forensic Sci. 59 (5) (2014) 1198–1204, https://doi.org/10.1111/jfo.2014.59.issue-510.1111/1556-4029.12517.
[20] B. Falatová, M. Ferreiro-González, C. Martín-Alberca, D. Kačíková, Š. Galla, M. Palma, C.G. Barroso, Effects of fire suppression agents and weathering in the analysis of fire debris by HS-MS enose, Sensors (Switzerland). 18 (2018) 1933, https://doi.org/10.3390/s18061933.
[21] Géron. A, Chapter 5: Support Vector Machine. Hands-On Machine Learning with Scikit-Learn & TensorFlow, Second Ed., O'Reilly. 2019. 153–174.
[22] P.M.L. Sandercock, E. Du Pasquier, Chemical fingerprinting of unevaporated automotive gasoline samples, Forensic Sci. Int. 134 (1) (2003) 1–10, https://doi.org/10.1016/S0379-0738(03)00081-1.
[23] M.E. Sigman, M.R. Williams, Assessing evidentiary value in fire debris analysis by chemometric and likelihood ratio approaches, Forensic Sci. Int. 264 (2016) 113–121, https://doi.org/10.1016/j.forsciint.2016.03.051.
[24] M. Ferreiro-González, G.F. Barbero, M. Palma, J. Ayuso, J.A. Álvarez, C.G. Barroso, Characterization and differentiation of petroleum-derived products by E-nose fingerprints, Sensors (Switzerland). 17 (2017) 2544, https://doi.org/10.3390/s17112544.

[25] Y. Xu, S. Zomer, R.G. Brereton, Support vector machines: A recent method for classification in chemometrics, Crit. Rev. Anal. Chem. 36 (3-4) (2006) 177–188, https://doi.org/10.1080/10408340600969486.
[26] M. Belgiu, L. Drăgu, Random forest in remote sensing: A review of applications and future directions, ISPRS J. Photogramm. Remote Sens. 114 (2016) 24–31, https://doi.org/10.1016/j.isprsjprs.2016.01.011.
[27] O. Devos, C. Ruckebusch, A. Durand, L. Duponchel, J.-P. Huvenne, Support vector machines (SVM) in near infrared (NIR) spectroscopy: Focus on parameters optimization and model interpretation, Chemom. Intell. Lab. Syst. 96 (1) (2009) 27–33, https://doi.org/10.1016/j.chemolab.2008.11.005.
[28] S. Lee, H. Choi, K. Cha, H. Chung, Random forest as a potential multivariate method for near-infrared (NIR) spectroscopic analysis of complex mixture samples: Gasoline and naphtha, Microchem. J. 110 (2013) 739–748, https://doi.org/10.1016/j.microc.2013.08.007.
[29] S. Qiu, J. Wang, C. Tang, D. Du, Comparison of ELM, RF, and SVM on E-nose and E-tongue to trace the quality status of mandarin (Citrus unshiu Marc.), J. Food Eng. 166 (2015) 193–203, https://doi.org/10.1016/j.jfoodeng.2015.06.007.
[30] H. Men, S. Fu, J. Yang, M. Cheng, Y. Shi, J. Liu, Comparison of SVM, RF and ELM on an Electronic Nose for the Intelligent Evaluation of Paraffin Samples, Sensors 18 (2018) 285, https://doi.org/10.3390/s18010285.
[31] L. Breiman, Random forests, Mach. Learn. 45 (2001) 5–32, https://doi.org/10.1023/A:1010933404324.
[32] F. Tian, J. Yan, S. Xu, J. Feng, Q. He, Y. Shen, P. Jia, C. Kadri, Classification of Electronic Nose Data on Wound Infection Detection Using Support Vector Machine Combined GA, J. Comput. Inf. Syst. 8 (2012) 3349–3357. https://www.researchgate.net/publication/268437419_Classification_of_Electronic_Nose_Data_on_Wound_Infection_Detection_Using_Support_Vector_Machine_Combined_GA (accessed May 27, 2021).
[33] A. Smolinska, J. Engel, E. Szymanska, L. Buydens, L. Blanchet, General Framing of Low-, Mid-, and High-Level Data Fusion With Examples in the Life Sciences, Data Handl, Sci. Technol. 31 (2019) 51–79, https://doi.org/10.1016/B978-0-444-63984-4.00003-X.
[34] F. Huang, H. Song, L. Guo, P. Guang, X. Yang, L. Li, H. Zhao, M. Yang, Detection of adulteration in Chinese honey using NIR and ATR-FTIR spectral data fusion, Spectrochim. Acta – Part A Mol. Biomol. Spectrosc. 235 (2020) 118297, https://doi.org/10.1016/j.saa.2020.118297.
[35] X.-M. Wu, Q.-Z. Zhang, Y.-Z. Wang, Traceability of wild Paris polyphylla Smith var. yunnanensis based on data fusion strategy of FT-MIR and UV–Vis combined with SVM and random forest, Spectrochim. Acta – Part A Mol. Biomol. Spectrosc. 205 (2018) 479–488, https://doi.org/10.1016/j.saa.2018.07.067.
[36] A.K. Smilde, I. Van Mechelen, A Framework for Low-Level Data Fusion, Data Handl, Sci. Technol. 31 (2019) 27–50, https://doi.org/10.1016/B978-0-444-63984-4.00002-8.
[37] A. Rudnitskaya, D. Kirsanov, A. Legin, K. Beullens, J. Lammertyn, B.M. Nicolaï, J. Irudayaraj, Analysis of apples varieties – comparison of electronic tongue with different analytical techniques, Sensors Actuators B Chem. 116 (1-2) (2006) 23–28, https://doi.org/10.1016/j.snb.2005.11.069.
[38] M. Li, J. Xue, Y. Du, T. Zhang, H. Li, Data Fusion of Raman and Near-Infrared Spectroscopies for the Rapid Quantitative Analysis of Methanol Content in Methanol-Gasoline, Energy & Fuels. 33 (12) (2019) 12286–12294, https://doi.org/10.1021/acs.energyfuels.9b03021.
[39] Max Kuhn et al. Caret package: Classification and Regression Training (R package version 6.0-86). 2019-03-27. URL: https://topepo.github.io/caret/.