Analytical and Bioanalytical Chemistry

Electronic Supplementary Material

A comparison of different chemometrics approaches for the robust classification of electronic nose data

Piotr S. Gromski, Elon Correa, Andrew A. Vaughan, David C. Wedge, Michael L. Turner, Royston Goodacre
Statistics on model constructions

For each classifier several measurements are reported, including:

**Time:** the amount of time (in seconds) taken for the central processing unit (CPU) to execute the classifiers within R.

**Confusion matrix:** This arrangement permits visualization and estimation of the performance of an algorithm by reporting actual and predicted classifications along with their accuracies. Each column of the matrix represents the occurrences in an actual class, while each row represents the occurrences in a predicted class (Table S1). The resulting tables in our study show the confusion matrix for four classes’ classifiers; in that case the results are considered comparing each factor level to the remaining levels in other words “one versus all” method.

<table>
<thead>
<tr>
<th>Reference model</th>
<th>Event</th>
<th>No Event</th>
</tr>
</thead>
<tbody>
<tr>
<td>Predicted model</td>
<td>Event</td>
<td>TP</td>
</tr>
<tr>
<td></td>
<td>No Event</td>
<td>FN</td>
</tr>
</tbody>
</table>

$sensitivity = \frac{TP}{(TP + FN)}$  \hspace{1cm} $specificity = \frac{TN}{(TN + FP)}$

Where: TP, true positive; TN, true negative; FP, false positive; FN, false negative

**Accuracy:** the proportion of the total number of correct predictions (sum of diagonal elements of a table) to the total number of elements in the table:

$$Accuracy = \frac{(TP + TN)}{(TP + FP + FN + TN)}$$

**95% CI:** 95% confidence intervals.

**No information rate (NIR):** criterion which indicates the highest value among prevalence.

**p-Value [Acc > NIR]:** $p$-value for Acc>NIR.

**Kappa:** Cohen's kappa statistic, which is a criterion of quantity of arrangement after chance arrangement. Kappa measures the percentage of data values in the main diagonal of the confusion matrix and then adjusts these values for the volume of agreement that could be
expected related to chance alone. Kappa ranges between 0 and 1, where 1 is perfect agreement and 0 represents no agreement. Values between 0.8 - 1 are considered as a very good agreement:

\[
\kappa = \frac{\text{observed probability} - \text{chance probability}}{1 - \text{chance probability}}
\]

The numerator represents the divergence amongst the observed probability of success and the probability of success under the hypothesis of an extremely bad case.

**McNemar's Test p-Value:** p-value for the non-parametric technique used for categorical variables. Can be used with two dichotomous measures on the same subjects (repeated measurements). The null hypothesis shows that the two probabilities for each outcome are the same:

\[
p_{TP} + p_{FP} = p_{TP} + p_{FN}
\]

and

\[
p_{FN} + p_{TN} = p_{FP} + p_{TN}
\]

Hence, McNemar’s test is explained by equation:

\[
\chi^2 = \frac{(FP - FN)^2}{FP + FN}
\]

**Prevalence:** the proportion of all positives in the total number of observations:

\[
\text{Prevalence} = \frac{(TP + FN)}{(TP + FP + FN + TN)}
\]

**Precision (Positive Predictive Value):** the proportion of true positives given a positive prediction:
\[ \text{Precision} = \frac{TP}{(TP + FP)} \]

**Negative Predictive Value:** the proportion of true negatives given a negative prediction:

\[ \text{Negative Predictive Value} = \frac{TN}{(TN + FN)} \]

**Detection rate:** the proportion of true positives in the total number of observations:

\[ \text{Detection rate} = \frac{TP}{(TP + FP + FN + TN)} \]

shows distribution of the predicted classes in the confusion matrix.

\[ \text{Detection prevalence} = \frac{(TP + FP)}{(TP + FP + FN + TN)} \]

**SVM parameters selection**

Cost parameter (C), gamma, degree and coef0 these were optimised as described in manuscripts: cost regularizes the number of support vectors, gamma is used to avoid single samples causing too much influence on the model; degree is a measure of the plane angle; coef0 is responsible for support vectors that maximise group separation [1-4].

The procedure we have adopted for the optimisation of these SVM is as objective as we could make it and the approach we used compares favourably with previous studies that have also been used in generalised searches for parameter optimisation [1,5-9,3].

**Table S2** Parameters selected for each kernel in SVM

<table>
<thead>
<tr>
<th>Kernel</th>
<th>C</th>
<th>Gamma</th>
<th>Degree</th>
<th>coef0</th>
</tr>
</thead>
<tbody>
<tr>
<td>Linear</td>
<td>1</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>Polynomial</td>
<td>0.18</td>
<td>2</td>
<td>2.5</td>
<td>-</td>
</tr>
<tr>
<td>Radial</td>
<td>0.2</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>Sigmoid</td>
<td>0.05</td>
<td>-</td>
<td>0</td>
<td>-</td>
</tr>
</tbody>
</table>

*C = cost parameter*
Random forests

As shown in Fig. S1 random subsets (1 to \(n\)) of trees are used where each tree starts off with a different subset of input variables (in open boxes) which are randomly selected (typically the number that are selected is \(\sqrt{\text{number of variables}}\)). Red circles correspond to internal nodes within each tree into which particular samples are assign based on judgment points (these are the internal edges or branches of the tree), finally black triangles correspond to leaf nodes where the final classifications are made. As many classifications are made then these are ensemble to give an average classification from the forest [10].

![Cartoon of random forests algorithm](image)

**Fig. S1** Cartoon of random forests algorithm

**Variable importance plot** – reflects variable importance measured by random forests (Fig. S2). These give a direct indication of which inputs are most useful for prediction. The VIPs are reported in two different ways: (A) mean decrease in accuracy and (B) mean decrease in Gini. Where, Gini index is a criterion that measures statistical dispersion. A decrease in Gini demonstrates that a particular predictor variable shows a greater influence in separating the data into the defined classes [10,11].
Validation procedures used in model construction

As discussed in the paper it is very important that chemometric models that employ supervised learning are adequately validated. There are several ways in which this can be achieved and in our work we have used cross-validation and replacement methods like bootstrapping. These are explained below.

**Cross-validation:** this is used to estimate the performance of a predictive model where the data are separated into two sets, called the training set and the test set [12,13]. To improve results for the evaluation of the data, we used $k$-fold cross-validation where $k=10$ such that the data are divided into nearly equal sets where $9/10$ of the data were used for training and the remaining $1/10$ are the independent data test sets (Fig. S3). Each model is repeated 10 times where each $10^{th}$ are used as the test set. This method allows for the estimation of how correct a predictive model will be when implemented. Moreover, this method also permits one to choose how large the final test set should be and how many trials are needed to accomplish this. To approximate the true distribution of the $k$-fold cross-validation the method is repeated multiple times (in our case 100) using different randomly selected folds from the data set [12,13].
**Fig. S3** Cartoon of 10-fold cross-validation. In this example data set of \( n \) samples and \( n \) features (X) with Y as class descriptor is divided into 10 equal parts

**Bootstrapping:** this is a general approach to statistical inference based on building a sampling distribution model for a statistic by resampling from the original data [14]. In non-parametric bootstrapping, samples are repeatedly selected from the original data set (each with an equal probability of selection) and then replaced (Fig. S4 shows this pictorially). Performing this resampling many times allows for the better estimation of prediction accuracy from the population by calculating multiple alternative versions of the single prediction accuracy that would ordinarily be calculated from one selection of the training set. Randomly simulated samples obtained from the original data should approximate the true source population, resulting in a new ‘pseudo-population’. Repeating this process thousands of times allows the calculation of confidence intervals for the statistics of interest. In general, the technique will generate a defined number of independent training and test data sets. In comparison to other cross-validation approaches, this method allows several samples to be selected numerous times in each split [14].
Chemical interpretation of models

In order to interpret the various models we have constructed the vectors that are used in model construction need to be inspected. These vectors give information on which variables are important and these are introduced and depicted below.

**Scores plot:** for LDA and PLS-DA the projection of the scaled latent variable scores in two dimensions. The scores plot helps to visualise the relationship between observations (samples).

**Loadings plot:** for LDA and PLS-DA this depicts the relationship (correlation) between input variables. The loadings can be regarded as the weights (vector) for each original variable when computing the latent variables. Loadings plots help to interpret patterns seen in the scores plot and to identify variables that are responsible for discrimination. Loadings close to zero have a rather low impact on the identification process, whereas high values indicate a large contribution.
Fig. S5 (a) PLS-DA scores plot of the first two latent variables, and (b) PLS-DA loadings plot representing variables associated with these first and second components

GP results

Table S3 Genetic programming results

<table>
<thead>
<tr>
<th></th>
<th>Acetone</th>
<th>DMMP</th>
<th>Methanol</th>
<th>Propanol</th>
</tr>
</thead>
<tbody>
<tr>
<td>Sensitivity</td>
<td>0.91</td>
<td>0.86</td>
<td>0.79</td>
<td>0.79</td>
</tr>
<tr>
<td>Specificity</td>
<td>0.96</td>
<td>0.88</td>
<td>0.87</td>
<td>0.83</td>
</tr>
</tbody>
</table>
Confusion Matrix and Statistics

class=test acetone DMN F methanol propanol
acetone 3066 0 21 113
methanol 0 0 2759 346
propanol 0 60 120 3970

Overall Statistics

Accuracy : 0.9116
95% CI : (0.9066, 0.9204)
No Information Rate : 0.205
P-Value [H0 > H1] : < 2.2e-16

Model : 0.8874
Nominal’s Test P-Value : NA

Statistics by Class:

Class: acetone Class: DMNF Class: methanol Class: propanol
Sensitivity 0.9761 0.3097 0.9697 0.9786
Specificity 0.9811 0.5713 0.8854 0.5437
Pos Pred Value 1.0000 0.9027 0.9059 0.9201
Neg Pred Value 0.2418 0.2249 0.2293 0.2060
Detection Rate 0.2418 0.2195 0.2169 0.2270
Detection Prevalence 0.2520 0.2820 0.2441 0.2520

Fig. S6 Confusion matrix for the LDA classification where the calculations were based on 100 random runs of ten-fold cross-validation

Confusion Matrix and Statistics

class=pred acetone DMN F methanol propanol
acetone 3114 0 0 16
methanol 143 180 54 8
propanol 427 2164 91 0

Overall Statistics

Accuracy : 0.6620
95% CI : (0.6566, 0.6687)
No Information Rate : 0.7977
P-Value [H0 > H1] : < 2.2e-16

Model : 0.817
Nominal’s Test P-Value : NA

Statistics by Class:

Class: acetone Class: DMNF Class: methanol Class: propanol
Sensitivity 0.8510 0.2001 0.3818 0.6960
Specificity 0.9203 0.5650 0.9317 0.9317
Pos Pred Value 0.9203 0.5650 0.9317 0.9317
Neg Pred Value 0.2001 0.3818 0.6960 0.9203
Prevalence 0.2766 0.2766 0.2766 0.2766
Detection Rate 0.2766 0.2766 0.2766 0.2766
Detection Prevalence 0.2461 0.2461 0.2461 0.2461

Fig. S7 Confusion matrix for the PLS-DA classification where the calculations were based on 100 random runs of ten-fold cross-validation
Confusion Matrix and Statistics

**Fig. S8** Confusion matrix for the RF classification where the calculations were based on 100 random runs of ten-fold cross-validation

Confusion Matrix and Statistics

**Fig. S9** Confusion matrix for the SVM classification based on linear kernel where the calculations were based on 100 random runs of ten-fold cross-validation
Confusion Matrix and Statistics

**SVM-pred**
class.test acetone DMMP methanol propanol
acetone 3000 0 191 9
DMMP 3 258 135 450
methanol 117 0 2062 297
propanol 3 170 149 2877

Overall Statistics

- **Accuracy**: 0.875
- 95% CI: (0.8691, 0.8807)
- No Information Rate: 0.2841
- P-Value [Acc > NIR]: < 2.2e-16
- Kappa: 0.8333
- Nonparam's Test P-Value: < 2.2e-16

Statistics by Class:

Class: acetone Class: DMMP Class: methanol Class: propanol
Sensitivity 0.9413 0.9875 0.9497 0.7520
Specificity 0.9833 0.9347 0.9566 0.9645
Pos Pred Value 0.9303 0.9821 0.9505 0.9204
Neg Pred Value 0.9833 0.9875 0.9666 0.9991
Prevalence 0.2841 0.2146 0.2489 0.2613
Detection Rate 0.2562 0.2056 0.2115 0.2266
Detection Prevalence 0.2520 0.2520 0.2441 0.2520

Fig. S10 Confusion matrix for the **SVM classification** based on **polynomial kernel** where the calculations were based on 100 random runs of ten-fold cross-validation.

Confusion Matrix and Statistics

**SVM-pred**
class.test acetone DMMP methanol propanol
acetone 3117 1 82 0
DMMP 10 2020 113 252
methanol 99 0 2886 205
propanol 6 200 96 2045

Overall Statistics

- **Accuracy**: 0.9166
- 95% CI: (0.9117, 0.9216)
- No Information Rate: 0.2602
- P-Value [Acc > NIR]: < 2.2e-16
- Kappa: 0.8898
- Nonparam's Test P-Value: < 2.2e-16

Statistics by Class:

Class: acetone Class: DMMP Class: methanol Class: propanol
Sensitivity 0.9811 0.9183 0.9078 0.8617
Specificity 0.9913 0.9605 0.9745 0.9625
Pos Pred Value 0.9741 0.8832 0.9223 0.9500
Neg Pred Value 0.9237 0.9736 0.9697 0.9519
Prevalence 0.2502 0.2418 0.2478 0.2602
Detection Rate 0.2416 0.2220 0.2249 0.2243
Detection Prevalence 0.2520 0.2520 0.2441 0.2520

Fig. S11 Confusion matrix for the **SVM classification** based on **radial kernel** where the calculations were based on 100 random runs of ten-fold cross-validation.
Confusion Matrix and Statistics

svm.pred

class: test actone DMF methanol propanol
actone 2707 3 0 463
DMF 4267 200 389
methanol 20 2544 394
propanol 0 342 335 2523

Overall Statistics

Accuracy : 0.8174
95% CI : (0.8126, 0.8241)
No Information Rate : 0.3094
P-Value [Acc > NIR] : < 2.2e-16

Kappa : 0.7568
Mcnemar's Test P-Value : < 2.2e-16

Statistics by Class:

<table>
<thead>
<tr>
<th>Class: actone</th>
<th>Class: DMF</th>
<th>Class: methanol</th>
<th>Class: propanol</th>
</tr>
</thead>
<tbody>
<tr>
<td>Sensitivity</td>
<td>0.9978</td>
<td>0.6751</td>
<td>0.8262</td>
</tr>
<tr>
<td>Specificity</td>
<td>0.9298</td>
<td>0.2390</td>
<td>0.9222</td>
</tr>
<tr>
<td>Pos Pred Value</td>
<td>0.9459</td>
<td>0.0147</td>
<td>0.9206</td>
</tr>
<tr>
<td>Neg Pred Value</td>
<td>0.9996</td>
<td>0.9608</td>
<td>0.9643</td>
</tr>
<tr>
<td>Prevalence</td>
<td>0.2136</td>
<td>0.2346</td>
<td>0.2424</td>
</tr>
<tr>
<td>Detection Rate</td>
<td>0.2131</td>
<td>0.2053</td>
<td>0.2003</td>
</tr>
<tr>
<td>Detection Prevalence</td>
<td>0.2520</td>
<td>0.2520</td>
<td>0.2441</td>
</tr>
</tbody>
</table>

Fig. S12 Confusion matrix for the SVM classification based on sigmoid kernel where the calculations were based on 100 random runs of ten-fold cross-validation

Confusion Matrix and Statistics

LDA classification

class: test actone DMF methanol propanol
actone 11076 6 160 428
DMF 23 8698 170 1628
methanol 1 4 9611 1714
propanol 2 460 707 10614

Overall Statistics

Accuracy : 0.866
95% CI : (0.8621, 0.8696)
No Information Rate : 0.3091
P-Value [Acc > NIR] : < 2.2e-16

Kappa : 0.848
Mcnemar's Test P-Value : < 2.2e-16

Statistics by Class:

<table>
<thead>
<tr>
<th>Class: actone</th>
<th>Class: DMF</th>
<th>Class: methanol</th>
<th>Class: propanol</th>
</tr>
</thead>
<tbody>
<tr>
<td>Sensitivity</td>
<td>0.9977</td>
<td>0.9647</td>
<td>0.9029</td>
</tr>
<tr>
<td>Specificity</td>
<td>0.9832</td>
<td>0.9496</td>
<td>0.9821</td>
</tr>
<tr>
<td>Pos Pred Value</td>
<td>0.9491</td>
<td>0.8446</td>
<td>0.8487</td>
</tr>
<tr>
<td>Neg Pred Value</td>
<td>0.9993</td>
<td>0.9668</td>
<td>0.9702</td>
</tr>
<tr>
<td>Prevalence</td>
<td>0.2386</td>
<td>0.2228</td>
<td>0.2295</td>
</tr>
<tr>
<td>Detection Rate</td>
<td>0.2380</td>
<td>0.2127</td>
<td>0.2072</td>
</tr>
<tr>
<td>Detection Prevalence</td>
<td>0.2508</td>
<td>0.2518</td>
<td>0.2441</td>
</tr>
</tbody>
</table>

Fig. S13 Confusion matrix for the LDA classification where the calculations were based on 1000 bootstraps
Confusion Matrix and Statistics

**pLDA.pReduction**

Class: test acetone DMNF methanol propanol

<table>
<thead>
<tr>
<th></th>
<th>acetone</th>
<th>DMNF</th>
<th>methanol</th>
<th>propanol</th>
</tr>
</thead>
<tbody>
<tr>
<td>acetone</td>
<td>11852</td>
<td>179</td>
<td>75</td>
<td>151</td>
</tr>
<tr>
<td>DMNF</td>
<td>97</td>
<td>9655</td>
<td>936</td>
<td>1371</td>
</tr>
<tr>
<td>methanol</td>
<td>200</td>
<td>577</td>
<td>9594</td>
<td>1470</td>
</tr>
<tr>
<td>propanol</td>
<td>81</td>
<td>1897</td>
<td>1642</td>
<td>816</td>
</tr>
</tbody>
</table>

**Overall Statistics**

- **Accuracy**: 0.8404
- **95% CI**: (0.8327, 0.8477)
- **No Information Rate**: 0.2561
- **P-Value [Acc > NIR]**: < 2.2e-16
- **Kappa**: 0.7072
- **Nonemar's Test P-Value**: < 2.2e-16

**Statistics by Class**:

<table>
<thead>
<tr>
<th>Class: acetone</th>
<th>Class: DMNF</th>
<th>Class: methanol</th>
<th>Class: propanol</th>
</tr>
</thead>
<tbody>
<tr>
<td>Sensitivity</td>
<td>0.9960</td>
<td>0.9354</td>
<td>0.8233</td>
</tr>
<tr>
<td>Specificity</td>
<td>0.9920</td>
<td>0.9600</td>
<td>0.9494</td>
</tr>
<tr>
<td>Pos Pred Value</td>
<td>0.9762</td>
<td>0.9495</td>
<td>0.9445</td>
</tr>
<tr>
<td>Neg Pred Value</td>
<td>0.8902</td>
<td>0.9437</td>
<td>0.9418</td>
</tr>
<tr>
<td>Prevalence</td>
<td>0.2529</td>
<td>0.2561</td>
<td>0.2804</td>
</tr>
<tr>
<td>Detection Rate</td>
<td>0.2445</td>
<td>0.2139</td>
<td>0.2062</td>
</tr>
<tr>
<td>Detection Prevalence</td>
<td>0.2500</td>
<td>0.2515</td>
<td>0.2441</td>
</tr>
</tbody>
</table>

**Fig. S14** Confusion matrix for the **PLS-DA classification** where the calculations were based on **1000 bootstraps**

Confusion Matrix and Statistics

**rf.pReduction**

Class: test acetone DMNF methanol propanol

<table>
<thead>
<tr>
<th></th>
<th>acetone</th>
<th>DMNF</th>
<th>methanol</th>
<th>propanol</th>
</tr>
</thead>
<tbody>
<tr>
<td>acetone</td>
<td>9223</td>
<td>507</td>
<td>550</td>
<td>352</td>
</tr>
<tr>
<td>DMNF</td>
<td>75</td>
<td>10902</td>
<td>19</td>
<td>1228</td>
</tr>
<tr>
<td>methanol</td>
<td>162</td>
<td>111</td>
<td>9221</td>
<td>1866</td>
</tr>
<tr>
<td>propanol</td>
<td>47</td>
<td>616</td>
<td>871</td>
<td>10049</td>
</tr>
</tbody>
</table>

**Overall Statistics**

- **Accuracy**: 0.8466
- **95% CI**: (0.8453, 0.8499)
- **No Information Rate**: 0.2922
- **P-Value [Acc > NIR]**: < 2.2e-16
- **Kappa**: 0.7954
- **Nonemar's Test P-Value**: < 2.2e-16

**Statistics by Class**:

<table>
<thead>
<tr>
<th>Class: acetone</th>
<th>Class: DMNF</th>
<th>Class: methanol</th>
<th>Class: propanol</th>
</tr>
</thead>
<tbody>
<tr>
<td>Sensitivity</td>
<td>0.9719</td>
<td>0.8454</td>
<td>0.8663</td>
</tr>
<tr>
<td>Specificity</td>
<td>0.9939</td>
<td>0.9587</td>
<td>0.9606</td>
</tr>
<tr>
<td>Pos Pred Value</td>
<td>0.9417</td>
<td>0.7721</td>
<td>0.8117</td>
</tr>
<tr>
<td>Neg Pred Value</td>
<td>0.9919</td>
<td>0.9459</td>
<td>0.9695</td>
</tr>
<tr>
<td>Prevalence</td>
<td>0.2172</td>
<td>0.2619</td>
<td>0.2287</td>
</tr>
<tr>
<td>Detection Rate</td>
<td>0.2111</td>
<td>0.2214</td>
<td>0.1982</td>
</tr>
<tr>
<td>Detection Prevalence</td>
<td>0.2598</td>
<td>0.2515</td>
<td>0.2441</td>
</tr>
</tbody>
</table>

**Fig. S15** Confusion matrix for the **RF classification** where the calculations were based on **1000 bootstraps**
Fig. S16 Confusion matrix for the SVM classification based on linear kernel where the calculations were based on 1000 bootstraps.

Fig. S17 Confusion matrix for the SVM classification based on polynomial kernel where the calculations were based on 1000 bootstraps.
Confusion Matrix and Statistics

**SVM prediction**

<table>
<thead>
<tr>
<th>Class</th>
<th>acetone</th>
<th>DMF</th>
<th>methanol</th>
<th>propanol</th>
</tr>
</thead>
<tbody>
<tr>
<td>acetone</td>
<td>10523 64 786</td>
<td>299</td>
<td>DMF</td>
<td>313 8861 644 1901</td>
</tr>
<tr>
<td>methanol</td>
<td>620 93 9946 1961</td>
<td></td>
<td>propanol</td>
<td>132 1313 883 9675</td>
</tr>
</tbody>
</table>

**Overall Statistics**

- Accuracy: 0.821
- 95% CI: (0.8175, 0.8245)
- No Information Rate: 0.2802
- P-Value (Acc > NIR): < 2.2e-16
- Kappa: 0.7614

Nonparametric Test P-Value: < 2.2e-16

Statistics by Class:

<table>
<thead>
<tr>
<th>Class</th>
<th>acetone</th>
<th>DMF</th>
<th>methanol</th>
<th>propanol</th>
</tr>
</thead>
<tbody>
<tr>
<td>Sensitivity</td>
<td>0.9081</td>
<td>0.8627</td>
<td>0.8031</td>
<td>0.7268</td>
</tr>
<tr>
<td>Specificity</td>
<td>0.9672</td>
<td>0.9212</td>
<td>0.9423</td>
<td>0.9311</td>
</tr>
<tr>
<td>Pos Pred Value</td>
<td>0.9027</td>
<td>0.7642</td>
<td>0.8227</td>
<td>0.8042</td>
</tr>
<tr>
<td>Neg Pred Value</td>
<td>0.9695</td>
<td>0.9695</td>
<td>0.9349</td>
<td>0.8975</td>
</tr>
<tr>
<td>Prevalence</td>
<td>0.2490</td>
<td>0.2207</td>
<td>0.2501</td>
<td>0.2802</td>
</tr>
<tr>
<td>Detection Rate</td>
<td>0.2261</td>
<td>0.1904</td>
<td>0.2099</td>
<td>0.2036</td>
</tr>
<tr>
<td>Detection Prevalence</td>
<td>0.2508</td>
<td>0.2518</td>
<td>0.2441</td>
<td>0.2532</td>
</tr>
</tbody>
</table>

**Fig. S18** Confusion matrix for the **SVM classification** based on **radial kernel** where the calculations were based on **1000 bootstraps**

Confusion Matrix and Statistics

**SVM prediction**

<table>
<thead>
<tr>
<th>Class</th>
<th>acetone</th>
<th>DMF</th>
<th>methanol</th>
<th>propanol</th>
</tr>
</thead>
<tbody>
<tr>
<td>acetone</td>
<td>9634 108 699</td>
<td>1490</td>
<td>DMF</td>
<td>123 9262 604 1793</td>
</tr>
<tr>
<td>methanol</td>
<td>92 111 975 1882</td>
<td></td>
<td>propanol</td>
<td>56 1908 1422 8997</td>
</tr>
</tbody>
</table>

**Overall Statistics**

- Accuracy: 0.7988
- 95% CI: (0.7951, 0.8024)
- No Information Rate: 0.303
- P-Value (Acc > NIR): < 2.2e-16
- Kappa: 0.7316

Nonparametric Test P-Value: < 2.2e-16

Statistics by Class:

<table>
<thead>
<tr>
<th>Class</th>
<th>acetone</th>
<th>DMF</th>
<th>methanol</th>
<th>propanol</th>
</tr>
</thead>
<tbody>
<tr>
<td>Sensitivity</td>
<td>0.9726</td>
<td>0.8352</td>
<td>0.8104</td>
<td>0.6381</td>
</tr>
<tr>
<td>Specificity</td>
<td>0.9444</td>
<td>0.9207</td>
<td>0.9406</td>
<td>0.8141</td>
</tr>
<tr>
<td>Pos Pred Value</td>
<td>0.8255</td>
<td>0.7905</td>
<td>0.8165</td>
<td>0.7656</td>
</tr>
<tr>
<td>Neg Pred Value</td>
<td>0.9222</td>
<td>0.9475</td>
<td>0.9325</td>
<td>0.9322</td>
</tr>
<tr>
<td>Prevalence</td>
<td>0.2119</td>
<td>0.2383</td>
<td>0.2485</td>
<td>0.3030</td>
</tr>
<tr>
<td>Detection Rate</td>
<td>0.2070</td>
<td>0.1990</td>
<td>0.1993</td>
<td>0.1993</td>
</tr>
<tr>
<td>Detection Prevalence</td>
<td>0.2508</td>
<td>0.2518</td>
<td>0.2441</td>
<td>0.2532</td>
</tr>
</tbody>
</table>

**Fig. S19** Confusion matrix for the **SVM classification** based on **sigmoid kernel** where the calculations were based on **1000 bootstraps**
References

2. Gunn SR (1998) Support Vector Machines for Classification and Regression. University of Southampton,

For more information on the statistical explanations/terms used above we refer the reader to: