Analyst

Accepted Manuscript

ROYAL SOCIETY OF CHEMISTRY

www.rsc.org/analyst

Assessment of the statistical significance of classifications in infrared spectroscopy based diagnostic models

David Pérez-Guaita[1], Julia Kuligowski[2], Salvador Garrigues[3], Guillermo Quintás[4] and Bayden.R. Wood[1*]

[1]Centre for Biospectroscopy, School of Chemistry, Monash University, Clayton, 3800 VIC, Australia.

[2] Neonatal Research Group, Health Research Institute La Fe, 46026 Valencia, Spain

[3]Analytical Chemistry Department, University of Valencia, Edifici Jeroni Muñoz, 46100 Burjassot, Spain[4] Leitat Technological Center, Bio In Vitro Division, 46026 Valencia, Spain

*To whom correspondence should be addressed

Fourier transform infrared (IR) spectroscopy in combination with multivariate data analysis is a versatile tool that can be applied for the diagnosis. However, a rigorous validation of the obtained models is necessary in order to obtain robust results. This work evaluates the advantages of the use of permutation testing for determining the statistical significance of the misclassification errors obtained from IR based diagnostic models through cross validation (CV). The model performance, estimated by CV, is compared to a distribution of CV-performance values obtained using randomly permuted class labels. The distribution of 'random CV-values' is considered as a null distribution and used to establish the significance of the model estimators obtained using real class labels. ATR-FTIR spectra of serum samples were classified using random forest (RF) classifiers according to two criteria, the tag number (a randomly assigned pseudo class membership) and the level of urea (real class). CV errors obtained were compared to the null distribution of CV errors from a permutation test and an independent validation set. The procedure was evaluated testing typical conditions leading to overoptimistic estimations provided by the CV like e.g. the size of subsamples used during CV, variable selection and the use of replicates. Results show that for the tag number (pseudo class), CV indicated classification errors between 23 and 33 % depending on the subsample size employed. Those values were even lower when variable selection or replicates were used. However, permutation testing indicated that those CV errors were non-significant. In contrast, for sample classification according to their levels of urea, all cross validation errors were found to be significant. Although the proposed method is computationally intensive, it provides a simple way of calculating an empirical p-value of the CV-estimator, thus establishing the statistical significance and providing a feasibility indicator especially useful for studies where the number of samples is limited.

1. Introduction

In the last decade, experimental studies revealed the potentials of vibrational spectroscopy for the diagnosis of disease in a wide variety of biological fluids and tissues[1]. Specifically, the combination of multivariate analysis and Raman or infrared (IR) spectroscopy has been repeatedly proposed for medical diagnosis based on the high amount of information encrypted in IR and Raman spectra, which provides a bimolecular fingerprint of the metabolic state of the patient[2]. A literature review between 2012 to 2014 reports its application for the diagnosis of several diseases such as asthma[3] or diverse types of cancer including glioma[4], breast[5], ovarian[6,7], oral[8] or bladder cancer[9]. In addition, in the case of blood and related samples such as serum and plasma, attenuated total reflectance (ATR) IR measurements offer a minimum-invasive, fast and cost-effective diagnostic tool[10]. The access to well characterized samples is fundamental for exploiting the potentials of vibrational spectroscopy. Selection of the sample size required to derive a statistically significant estimation is critical to produce robust results and it depends on a number of factors including the size of the effect and intrinsic biological and instrumental variation. However, difficulties in obtaining well characterized samples (e.g. blood, tissues) may hinder the development of this type of technology and, frequently, reported studies include a limited number of samples. In these situations, external independent test sets for the evaluation of the prediction capabilities of the models are not always available and, alternatively, a number of cross validation (CV) procedures are employed. In spite of its usefulness, CV-based approaches might lead to an underestimation of the classification error[11] leading to overly optimistic descriptions of the model discriminant performance. In the fields of metabolomics and proteomics it is not unusual to generate mega-variate datasets with a high variable to sample ratio[12]. Because of the need to establish reliable indicators of the performance of the classifications when biomarker identification is involved, the aforementioned problems related with CV have been already addressed. One approach is based on non-parametric permutation tests, which imply the random reallocation of class labels in order to establish the statistical significance

of a figure of merit obtained using CV[13,14]. In spite of being a quite common approach in other fields such as proteomics[13] or metabolomics[12], this strategy is not yet common practice in IR biospectroscopy and besides a conference proceedings of Lloyd et al[15], it is difficult to find any application of this procedure in the biospectroscopy field .

The aim of this paper is to introduce permutation testing as a feasible indicator of the statistical significance of CV-figures of merit of IR based diagnostic models. For this purpose, ATR-Fourier transform infrared (FTIR) spectra of serum samples were recorded and discriminant models by means of random forest (RF) classifiers were developed. Reproducing preliminary studies, a limited number of samples were used for calculating and cross validating the RF classifiers. Results of a permutation test performed over the CV error were compared with the actual prediction capability of those models, established by validating them with a large and representative set of samples. Serum samples were classified in two groups according to: i) the urea concentrations in sample, and ii) to an arbitrary sample number un-related to the sera composition. The importance of the access to statistically reliable performance parameters was further underlined in common situations of IR based diagnostic models, which are known to enhance overoptimistic CV errors, i.e. the use of measurement replicates and variable selection.

2. Experimental

2.1 Samples and Spectral Acquisition

ATR-FTIR Spectra of 122 serum samples from different patients were acquired as described elsewehere[16]. Briefly, spectra containing 235 data points recorded in the range 1800-850 cm^{-1} with a spectral resolution of 4 cm^{-1} were recorded using a Tensor 27 spectrophotometer from Bruker (Bremen, Germany), equipped with temperature-stabilized deuterated lanthanum tryglycine sulfate (DLATGS) detector and a nine reflection diamond/ZnSe DuraDisk from Smiths Detection Inc. (Warrington, UK). 150 microliters of serum were deposited on the ATR cell and the spectrum was

measured using the empty clean cell as a background. Both, spectrum and background were obtained averaging 100 scans, and the water contribution to the spectrum was eliminated by subtracting a spectrum of water measured in the same conditions.

Measurements for each sample were recorded in triplicate and the average of the replicates was used in sections 3.1 and 3.2 and the actual individual replicates were used in section 3.3.

2.2 Samples Sets and Class Assignment

As we have stressed in the introduction, in most of the cases the number of samples are limited on the studies for the diagnosis of illnesses based on IR spectroscopy. Since our investigation aimed to simulate those preliminary studies, only 30 samples were employed for calibration. Besides, reproducing the conditions of a representative blind test, 92 samples were used only as an independent test set in order to establish the actual prediction capability of the models against a real-world external set of samples not included in any step of both the calibration and selection of RF parameters. Sample classes were assigned according to two criteria:

1) First, samples were classified based on the concentration of urea measured by employing an enzymatic reference method used at the hospital for the conventional routine analysis. 61 samples (15 for calibration and 46 for validation) contained urea concentration values above 60 mg/dL and the other 61 (15 for calibration and 46 for validation) below 40 mg/dL. Urea concentration in serum can be measured using FTIR spectroscopy[16] and so, a discriminative model should be a priori to provide good predictive performance.

2) Samples were distributed based on an arbitrary hospital tag number: 56 samples (15 for calibration and 41 for validation) with a tag finished by 1, 2, 3 and 4 were assigned to ‘class A’ and 66 samples (15 for calibration and 51 for validation) with a tag finished by 0, 5, 6, 7, 8 and 9 were assigned to ‘class B’. This

classification was therefore arbitrary and thus, any classification obtained considering this pseudo-class could be attributed to chance, model overfitting or to an optimistic estimation of the CV-figures of merit.

2.3 Random Forest Classification

Random forest classification was performed in Matlab 2012b (The Matwhorks, Natick, USA) using the routine available at https://code.google.com/p/randomforest-matlab/ and further calculations were carried out using in-house written scripts.

In order to increase the number of variables, an augmented data matrix was obtained by concatenating the raw and the first derivate spectra calculated using a Savitzky-Golay routine, resulting in a calibration dataset with 30 samples and 986 variables. The tuning parameters of the RF classifiers were set following the recommendations of Ollesch et al[9]. The number of trees employed for each random classifier was three times the number of variables and the number of wavenumber-features used in each tree was a third of the number of variables). All other parameters of the RF classifiers were set to the default values established in the routines.

2.4 Cross validation, external validation and permutation test

For performing the Montecarlo CV only the calibration samples were used. A set of M/n samples, where "M" is the number of samples in the calibration set and "n" the number of data splits used during cross validation, were randomly selected and predicted by the RF classifiers obtained using the remaining samples. This step was repeated 60 times, changing the calibration and internal validation subsets selected and registering the CV-predicted classes in order to obtain a representative value of the CV from the all possible subsets. Based on CV-predicted classes, the samples were classified using a majority rule: each sample was classified in the group of the majority of CV-predicted values. The error rate was then obtained by dividing the number of the bad classified samples by the numbers of samples of the calibration set. Then, the external test set

of 92 samples, was used for establishing the actual classification performance of the models. In this case the whole calibration dataset was employed for computing the RF classifiers and the same parameters employed for the CV in terms of tuning parameters of the RF classifiers. In the case of the external validation after variable selection, RF classifiers were performed including only the variables selected in the routine.

Permutation testing included a random class permutation followed by an ensemble RF using the dataset with the permuted class labels. CV was carried out as described above. Those two steps were repeated 200 times and the CV error was obtained. The distribution of the so-called 'random CV-values' was used as a null distribution for establishing the statistical significance of the model CV estimators obtained using real class labels. This statistical significance was calculated as the number of permuted values lower than the CV error obtained using the real classes divided by the number of permuted values.

2.5 Variable selection

The variable selection (selection of the wavenumbers which provide the best classification) procedure was based on a Montecarlo procedure as described in a previous reference[9], setting the tunable parameters to less restrictive values to facilitate the selection of correlated but uninformative variables. Montecarlo feature selection utilizes the repetition of a cyclic routine randomly selecting 23 samples in each cycle. In the cyclic procedure, an ensemble RF classifier was created and validated using the Montecarlo CV as described in the section 2.3, recording the mean Gini values for each variable and the CV error. Then, 20% of variables with the lowest Gini values were eliminated and the CV was repeated. This cycle of CV and elimination of variables was repeated 12 times until only 45 values were retained. From the variables selected from the 10 RF models, those selected by the model with the best classification error rate were finally chosen as variables selected for the cyclic routine. This cyclic routine was repeated 10 times and the final variables chosen were those that were selected in 3 or more out of the 10 cyclic routines.

3. Results

3.1 Statistical significance of cross validation errors

The selection of the employed n-folds (i.e. number of sub-samples used during CV) is critical when CV errors are used as the selection of a high n-fold could lead to an overoptimistic evaluation of the model. RF models for the classification by sample tag number and urea levels were performed and validated using an independent validation set (external validation) and CV with different n-folds (15, 10 and 4). Results (see Table 1) evidenced a strong association between the CV error and the n-fold. In all cases, the CV error was lower than the one obtained using the independent validation set and increasing the n-fold, the CV error decreased. More importantly, in the case of the classification according to the sample tag number, classification errors between 26 and 44% were obtained. Those values are significantly lower than the one obtained by the independent validation, which was 52±1% and hence corresponded to the expected value for a non-statistically significant class difference. This demonstrates the capability of RF for recognizing apparent discriminant patterns even in random data, which only was evidenced after a proper external validation with an independent sample set and might be not detected using CV-figures of merit.

Table 1: Error parameters obtained from the validation of the RF classifiers for different n-fold values.

Class	Parameter	n-fold 15	n-fold 10	n-fold 4
TAG NUMBER	Independent validation error (%)	51.8±0.5%	51.8±0.5%	55.4%±1.0%
	CV Error (%)	26±4%	38±4%	44%±3%
	p-value	**0.36**	**0.33**	**0.35**
UREA LEVEL	Independent validation error (%)	32±2%	32±3%	32±2%
	CV Error (%)	8.2±0.9%	13.3±0.7%	15.7±0.7%
	p-value	**<0.005**	**0.005**	**0.005**

Permutation testing was applied for calculating the statistical significance of the aforementioned CV errors. The values of the CV errors obtained for the classification of the permuted classes are represented in Figure 1. It is remarkable that in the case of the 15-fold CV (leave-2-out), the distribution of CV errors obtained for 200 permutations are always lower than the theoretically expected 50% error value. This shows the ability of RF to provide low CV errors from IR based classification/diagnostic methods no matter which class/illness is under study. However, comparing the CV errors to the distribution of the CV errors obtained from permuted classes enables the evaluation of the statistical significance of the model under study.

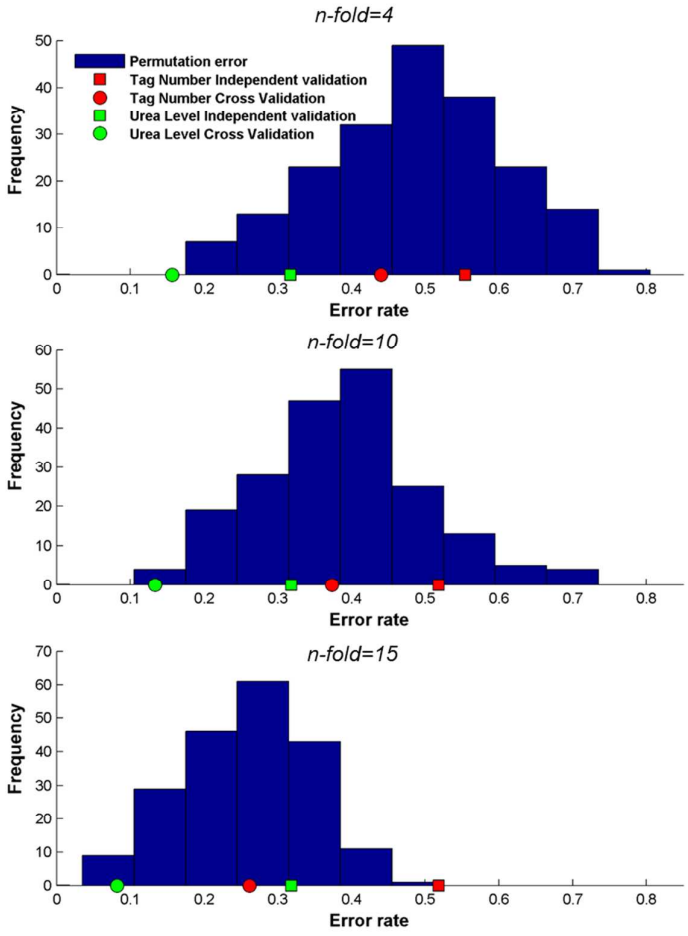

Figure 1: Distribution of errors obtained from the independent and cross validation of the models using RF classifiers.

The mean of the Gaussian distribution of the prediction moves to lower values as the n-fold increases. Therefore, although the CV error also decreases with increasing n-fold, their statistical significance does not show substantial changes throughout the different n-folds under study. In the case of the arbitrary classes based on the sample tag number, the obtained CV was in the range of the mean CV error obtained from permuted classes as shown in Figure 1 and the calculated statistical significance was $p>0.05$ as shown in Table 1. Thus, the ability of permutation testing for detecting overoptimistic estimations of the CV error at different n-folds is demonstrated. In contrast, for the classification between samples with low and high levels of urea, the CV errors obtained were significantly lower than those obtained by the permuted classes ($p\leq0.005$). It has to be also considered that the prediction error obtained using the independent validation dataset was two times higher than the one obtained using CV, which indicates, as previously shown[12], that permutation testing focuses on assessing the significance of the classification without taking into account the real predictability of the population of samples to predict. An n-fold value of 4 (leave-7-out) was selected for the further calculations.

3.2 Statistical significance of CV errors after feature selection

The development of diagnostic tools based on the modelling of IR spectra faces a critical challenge regarding the sample to variable ratio, especially considering early studies where the number of samples is limited. Therefore, feature selection is normally used for "producing a smaller number of variables ("feature = input variable") that are more informative than the original whole set of wavenumber-variables"[17]. The main aim of this procedure is the improvement of the diagnostic model performance and the detection of spectral biomarkers that can help to improve the understanding of the biological mechanisms of the illness, also providing a scientific justification of the IR based classification. However, the introduction of this feature selection could lead to overoptimistic results and the apparent improvement in the CV error obtained after variable selection can also be originated from the susceptibility of variable selection towards chance

correlations[18]. This effect is a well-known source of complications in the metabolomics field and the likelihood of obtaining a model based on chance correlation depends on the sample to variable ratio and the correlation structure of the data[19]. In addition, the presence of chance correlation on regression models for the prediction of glucose concentration has been widely studied [20,21].

The aim of this section was to assess the use of the permutation testing for evaluating the statistical significance of IR models after variable selection. Variable selection was carried out for the classification of samples according to the sample tag number and low/high urea levels followed by cross validation and permutation testing. From results shown in Figure 2 it can be appreciated that, as compared to the 4-fold distribution before variable selection (see Figure 1), the variable selection procedure improved the CV error of the classification according to the sample tag number from 44% to 27%. However, the CV distribution of errors obtained from the permuted values also moves to lower errors and consequently no significant difference to the CV error was observed (p=0.46). In contrast, for the classification according to the urea levels, the CV error obtained after variable selection was still significantly lower than errors obtained from permutation testing (p=0.015).

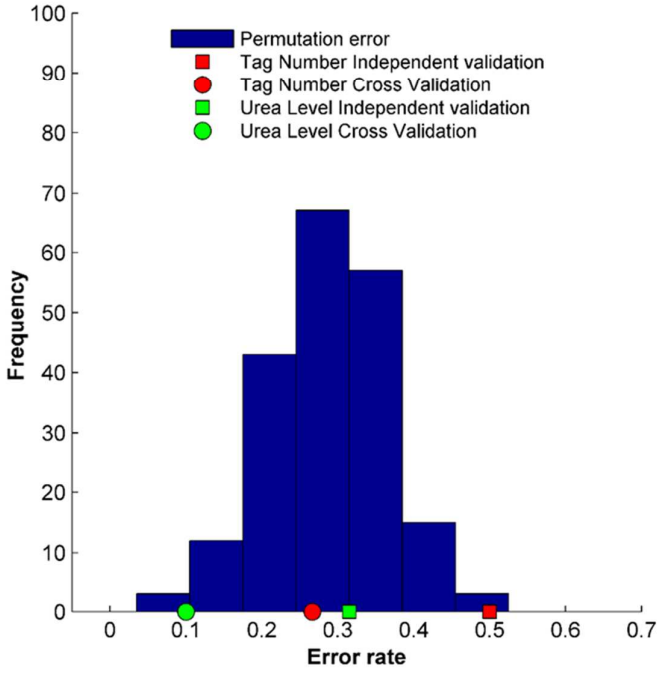

Figure 2: Distribution of errors obtained from the independent and cross validation of the models performed for classification of the samples after variable selection using the RF classifiers.

Figure 3 shows the frequency of the selected variables in each of the 10 cycles performed during the variable selection process, where red dots correspond to the variables that were finally selected. The distribution of the variables selected differs from the classification considered. In the case of the arbitrary determination of the sample tag number (see Figure 3b), several wavenumber values arbitrarily distributed across the IR bands were selected. It should be noted that the use of these variables improved the CV error in comparison to the use of the whole spectrum, and if the statistical significance was not considered, they could have been mistakenly assigned as "biomarkers" for the sample tag number. These results demonstrate that the variable selection procedure could find correlations on trivial wavenumber values which apparently improved the prediction capability of the model, but which did not have any relationship with the studied parameter. Nevertheless, the use of the permutation test and a thorough investigation of the wavenumber values "feature selected" demonstrates the irrelevance of these markers and models, which are likely associated with chance correlations without statistically significant results.

In the case of the classification according to urea levels, the variables selected correspond to characteristic urea bands, such as a shoulder of the band at 1040 (ρNH_2)[22] cm^{-1} in the untreated and the first derivative spectrum and the bands at 1552 (νC-N)[22] and at 1155 (ρNH_2)[22] cm^{-1} in the first derivative spectrum. It has to be noted that the bands selected are not the most intense bands of the urea spectrum. However, this might be due to strong interference in the 1600-1400 cm^{-1} region caused by amide I and II bands. The selected wavenumber values are in good agreement with the selectivity ratio[23] obtained from a PLS regression performed on the data set for the prediction of the concentration of urea (see black spectra) [16].

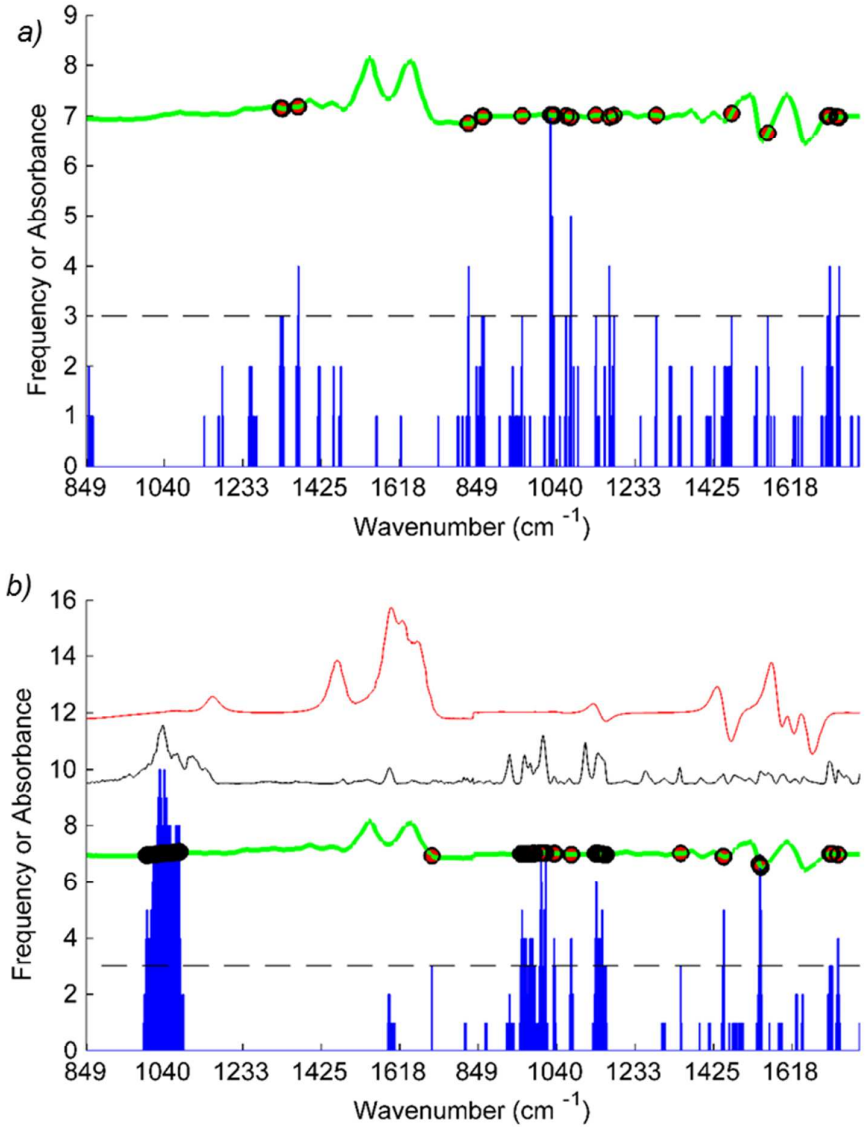

Figure 3: Wavenumber values chosen by the feature selection procedure for the classification according to the sample tag number (a) and urea levels (b). Blue bars indicate the variable selection frequency in the 10 cyclic routines performed. Variables with a frequency higher than the threshold (dashed line) were finally retained and are indicated as red points with black circles in the mean sample spectrum (green line). Panel (b) also shows the ATR spectrum of a urea standard solution at 1000 mg/dL (red line) and the selectivity ratio values for a PLS regression model (black

line).The Y axis label "Frequency" corresponds to blue bars and the Y axis label "Absorbance" corresponds to the spectra. Selectivity ratio is unitless.

3.3 The effect of replicates on the statistical significance on RF models

The high variable to sample ratio can be also compensated by including measurement replicates of the sample in the model. The use of measurement replicates is usually justified by the need of taking into account the variability of different spots of the samples. The 'sample replicate trap' occurs when replicates of the same physical samples are introduced in both the calibration and the validation subsets employed during model CV and external validation[24]. This effect can be corrected by using a continuous block cross validation employing sub models of m-blocks being m the number of replicates. However this process can be troublesome if the number of replicates is not the same for all samples.

The effect of the use of replicates during cross validation of RF models was evaluated employing permutation testing. Figure 4 represents the CV and independent validation errors obtained for the two class labels considered as well as the distribution of the CV errors for 200 permutations when two and three replicates are used. Once again, the distribution of the errors obtained using the permutated classes was dramatically shifted to lower errors as the number of replicates increased. According to the CV error, the RF was able to correctly classify 82 and 93% of the samples according to the sample tag number using 2 and 3 replicates, respectively. As expected, the permutation testing indicated that those classifications were not statistically significant ($p=0.6$ and 0.5). Regarding the classification of the samples according to urea levels, the prediction errors were found to be significantly different from those obtained from randomly permuted classes in all cases ($p<0.05$). It has to be remarked that, although the use of replicates improved the CV errors, it

decreased the real prediction capability of the model, as evidenced by the increase of the independent prediction errors obtained using 1, 2 and 3 replicates.

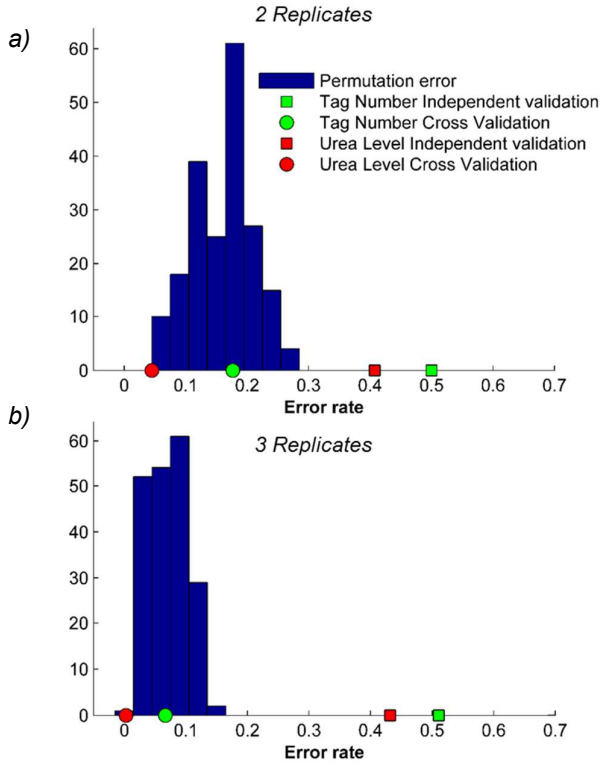

Figure 4: Distribution of the classification errors of the RF classifiers obtained using 2 (a) and 3 replicates (b).

4. Conclusions

Results obtained show that the application of permutation testing is suitable for calculating the statistical significance of IR based diagnostic models. In the case of the sample tag number, the manifested CV errors obtained were not statistically significant after permutation testing. In the case of the classification according to urea levels, CV errors were found to be significant. It has to be stipulated that the method did not provide an estimation of the predictive capacity of the model, which can only be achieved by an external validation set. However, it provides a significance value of the classification, which contains information about the significance of the differences between the spectra of different classes. Therefore this method can be used as an early estimator of the

validity of a hypothesis or for making a decision about the utility of incorporating further samples to the system. Although there are several repetitions of the procedures involved in the modelling such as the CV and variable selection, which can become computationally intensive, it is nonetheless a versatile tool that can be used in systems with a high variable to sample ratio or with an unbalanced number of samples in each class and hence we recommend the use of permutation testing especially in the case of preliminary studies.

Acknowledgements

JK acknowledges the Sara Borrell CD12/00667 grant (*Instituto Carlos III,* Ministry of Economy and Competitiveness, Spain). B.R.W. is supported by an Australian Research Council (ARC) Future Fellowship (FT120100926) and the project is supported by an ARC Discovery grant DP120100431.

1. C. Kendall, M. Isabelle, F. Bazant-Hegemark, J. Hutchings, L. Orr, J. Babrah, R. Baker, and N. Stone, *Analyst*, 2009, **134**, 1029–1045.
2. D. I. Ellis and R. Goodacre, *The Analyst*, 2006, **131**, 875–885.
3. A. Sahu, K. Dalal, S. Naglot, P. Aggarwal, and C. Murali Krishna, *PLoS ONE*, 2013, **8**, e78921.
4. J. R. Hands, P. Abel, K. Ashton, T. Dawson, C. Davis, R. W. Lea, A. J. S. McIntosh, and M. J. Baker, *Anal. Bioanal. Chem.*, 2013, **405**, 7347–7355.
5. M. Khanmohammadi and A. B. Garmarudi , *TrAC Trends Anal. Chem.*, 2011, **30**, 864–874.
6. K. Gajjar, J. Trevisan, G. Owens, P. J. Keating, N. J. Wood, H. F. Stringfellow, P. L. Martin-Hirsch, and F. L. Martin, *The Analyst*, 2013, **138**, 3917–3926.
7. G. L. Owens, K. Gajjar, J. Trevisan, S. W. Fogarty, S. E. Taylor, B. Da Gama-Rose, P. L. Martin-Hirsch, and F. L. Martin, *J. Biophotonics*, 2014, **7**, 200–209.
8. A. Sahu, S. Sawant, H. Mamgain, and C. M. Krishna, *Analyst*, 2013, **138**, 4161–4174.
9. J. Ollesch, S. L. Drees, H. M. Heise, T. Behrens, T. Brüning, and K. Gerwert, *Analyst*, 2013, **138**, 4092–4102.
10. G. Bellisola and C. Sorio, *Am. J. Cancer Res.*, 2011, **2**, 1–21.
11. K. H. Esbensen and P. Geladi, *J. Chemom.*, 2010, **24**, 168–187.

12. C. M. Rubingh, S. Bijlsma, E. P. P. A. Derks, I. Bobeldijk, E. R. Verheij, S. Kochhar, and A. K. Smilde, *Metabolomics*, 2006, **2**, 53–61.
13. S. Smit, M. J. van Breemen, H. C. J. Hoefsloot, A. K. Smilde, J. M. F. G. Aerts, and C. G. de Koster, *Anal. Chim. Acta*, 2007, **592**, 210–217.
14. J. A. Westerhuis, H. C. J. Hoefsloot, S. Smit, D. J. Vis, A. K. Smilde, E. J. J. van Velzen, J. P. M. van Duijnhoven, and F. A. van Dorsten, *Metabolomics*, 2008, **4**, 81–89.
15. G. R. Lloyd, J. Hutchings, L. M. Almond, H. Barr, C. Kendall, and N. Stone, 2012, vol. 8219, p. 82190C–82190C–6.
16. D. Perez-Guaita, J. Ventura-Gayete, C. Pérez-Rambla, M. Sancho-Andreu, S. Garrigues, and M. de la Guardia, *Microchem. J.*, 2013, **106**, 202–211.
17. J. Trevisan, P. P. Angelov, P. L. Carmichael, A. D. Scott, and F. L. Martin, *Analyst*, 2012, **137**, 3202–3215.
18. J. Kuligowski, D. Pérez-Guaita, J. Escobar, M. de la Guardia, M. Vento, A. Ferrer, and G. Quintás, *Talanta*, 2013, **116**, 835–840.
19. K. Baumann, *QSAR Comb. Sci.*, 2005, **24**, 1033–1046.
20. B. Zhao, Y. Cao, R. Liu, and K. Xu, *Guang Pu Xue Yu Guang Pu Fen Xi Guang Pu*, 2012, **32**, 934–938.
21. R. Liu, W. Chen, X. Gu, R. K. Wang, and K. Xu, *J. Phys. Appl. Phys.*, 2005, **38**, 2675–2681.
22. J. Grdadolnik and Y. Maréchal, *J. Mol. Struct.*, 2002, **615**, 177–189.
23. T. Rajalahti, R. Arneberg, F. S. Berven, K.-M. Myhr, R. J. Ulvik, and O. M. Kvalheim, *Chemom. Intell. Lab. Syst.*, 2009, **95**, 35–48.
24. K. A. Bakeev, *Process Analytical Technology: Spectroscopic Tools and Implementation Strategies for the Chemical and Pharmaceutical Industries*, John Wiley & Sons, 2008.