The need for external validation in machine olfaction: emphasis on health-related applications

S. Marco[1,2]

[1] Signal and Information Processing for Sensing Systems

Dept. of Biomedical Signals and Instrumentation, Institute for Bioengineering of Catalonia

Baldiri Reixac, 4-8, 08028-Barcelona, Spain.

[2] Intelligent Signal Processing

Dept. of Electronics, Faculty of Physics, Universitat de Barcelona

Martí I Franqués 1, 08028- Barcelona, Spain

e-mail: smarco@ibecbarcelona.eu

Abstract: Over the last two decades electronic nose research has produced thousands of research works. Many of them were describing the ability of the e-nose technology to solve diverse applications in domains ranging from food technology, to safety, security or health. It is in fact, in the biomedical field where e-nose technology is finding a research niche in the last years. While few success stories exist, most described applications never found the road to industrial or clinical exploitation. Most described methodologies were not reliable and were plagued by numerous problems that prevented practical application beyond the lab. This work emphasizes the need of external validation in machine olfaction. I describe some statistical and methodological pitfalls of the e-nose practice and I give some best practice recommendations for researchers in the field.

Keywords: chemical sensor arrays, pattern recognition, chemometrics, electronic noses, robustness, signal and data processing.

I. Introduction

In the last decade the use of electronic noses for health applications has been advocated by many different groups reporting good results for the detection of a number of diseases due to volatiles emitted by a variety of biomedical fluids. Electronic noses have been often used for the analysis of breath[1-13], but also for the analysis of body odours[14], wound infections[14], urine[15], stools[16], etc. While most of the reported results are encouraging, we have to keep a healthy skeptical attitude due to the potential statistical and methodological artifacts arising in small sample studies. In fact, electronic noses have a very good record of successful lab applications that never reached equivalent good results in real-life settings.

On the other hand, sets of partially selective sensors should necessarily rely on signal and data processing to overcome the inherent difficulties of low selectivity[17,18]. Today's chemical instrumentation including electronic noses is based on embedded (Figure 1) or desktop computers that control the operation of the instrument and additionally can be used for on-line data evaluation. In fact, signal processing and data analysis are of increasing importance not only for machine olfaction but for advanced chemical instrumentation (chemometrics) and –omics data analysis (bioinformatics). It is well-known that current instrumentation offers enormous capabilities for signal recording and storage, while the signal/data analysis and interpretation is the bottleneck of the process.

On the other hand, it is true that the biological sense of olfaction has remarkable capacities from a chemical instrumentation point of view relying on sets of receptors with partial and overlapped sensitivities. This astonishing system is still the main inspiration for machine olfaction[19,20]. However, we are still far today from understanding olfaction from a measurement science perspective, and even further away to be able to mimic this important physiological function.

In Machine Olfaction, but also in other instrumental techniques (Gas Chromatography-Mass Spectrometry (CG-MS), Near Infrared Spectroscopy (NIR), Fourier Transform Infrared Spectroscopy (FT-IR), Ion Mobility Spectrometry (IMS), etc), signal and data processing is required for a large variety of needs, namely (list not exhaustive):

i) Quantification ~~y~~ of components in simple chemical mixtures
ii) Identification of ~~y~~ chemical products by a chemical fingerprint
iii) Monitoring chemical / biochemical processes from volatile emissions patterns
iv) Self-diagnosis for instrument mal-functioning.
v) Error correction
vi) Signal compression
vii) Drift counteraction
viii) Rejection of cross-sensitivities to environmental parameters or background changes.
ix) Design of experiments and design of method validation
x) Noise reduction
xi) Baseline removal

Since insight on raw signals is provided by complex and sometimes even obscure algorithms (in the sense that their interpretation is not easy (e.g multilayer perceptrons)), it is important that the researcher behind predictive algorithm development, in addition of being well versed in terms of data analysis theory and the intricacies of the measurement protocol, keeps an healthy skepticism attitude towards the obtained results.

Several authors have been critical towards the intensive use of advanced statistical and machine learning tools particularly for health applications in the domain of –omic data analysis[21,22]. It is a fact that complex algorithms sometimes give researchers a false sense of safety concerning their results. This safety feeling heavily contrasts with the difficulties in replication of results by other groups~~for results replication by other groups~~. In fact, complex techniques reporting overoptimistic results that were never confirmed has lead to a mistrust about complex algorithms by electronic nose practitioners.

While today we have a large set of tools for the analysis and processing of sensor arrays signals[17,18] we have to be aware that most methods require optimization and the optimum parameters are data dependent. Application with fixed parameters may provide sub-optimal results~~infra-optimal results~~. It is not easy, and it is not recommended, to use most of the methods as black-box systems. The user has to understand their principles and limitations. However, it is difficult to demand that a regular chemist or application engineer has the advanced knowledge that state of the art techniques arising from statistics, machine learning and chemometrics require for proper (or optimum use). The use of sophisticated techniques is made easier because of the availability of many routines in the public domain for different programming languages as MATLAB, R or Python. Additionally there are many commercial software products with attractive graphical user interfaces (GUI) that let anyone to play easily with the algorithms and their parameters. Misuse of these tools is common in the literature[23].

In this framework, we have to admit that electronic noses are considered unreliable instruments by practitioners in several sectors. Robustness improvement is a must to enlarge enose market penetration. E-noses suffer from poor time stability (figure 2) and a number of data processing solutions have been proposed to counteract instrument drift[24-30]. Knobloch et al.[31] report that temperature and mass flow variations can lead to serious stability problems in headspace analysis of liquid samples (figure 3). In the same study (see figure 4), Knobloch reports on the lack of reproducible responses across identical instruments (from a constructive point of view). This variability hinders transferring calibration models from a master instrument to slave instruments and leads to individual e-nose calibration and higher added costs. Kuske et al.[33,34], when studying the detection of moulds growing on building materials, found that the substrate had a large effect on the sensor array response (figure 5). In other words, the application of classification models built with some materials sets, leads to degraded responses when applied to other building materials. In other words, when trying to detect volatile metabolites produced by fungi we should be aware that changes in the background other can lead ~~to problems for~~ detection problems. Similarly Adam et al.[34], when monitoring an anaerobic digester with an electronic nose, noticed that the sensor array pattern was clearly dependent on the materials used for feeding the digester. That is, digester normal operation control models have to be dependent on the feeding regime. Changes on the feeding materials provide larger background variability leading to less specific models.

In this sense, it is important to remark the definition of robustness by the World Health Organization[35]: "Robustness (or Ruggedness) is the ability of the procedure to provide analytical results of acceptable accuracy and precision under a variety of conditions: The results from separate samples are influenced by changes in operational or environmental conditions". It is also important to remark that in the context of intelligent instruments the robustness evaluation has to consider the full analytical procedure: from the sampling method to the e-nose response, but also the full set of signal and data processing blocks that use sensor outputs to produce a predictive response either qualitative (label prediction) or quantitative(e.g. concentration prediction). For instance, it is well-known that with exactly the same sensors, some algorithms can make the instrument more resilient to drift, or to the scarcity of calibration samples[17].

For other instrumental techniques, and mostly in health applications, it has been long recognized that prediction models perform better on the data used to build the model than on fresh data. This effect is particularly clear when the number of samples is much smaller than the dimensionality of the data. In this context, we would like to emphasize the importance of strict validation methods for ultimate performance assessment of the proposed instrumental method and subsequent data evaluation for qualitative or quantitative prediction. While in early 90s there were many publications reporting only basic exploratory data analysis (the ever present score plot from principal component analysis) today most journals require some method validation to be reported and discussed. However, in many occasions only internal validation methods are reported.

While the potential of the instruments is not in dispute, it is absolutely clear that research needs to shift towards methodology improvement at the procedure level, but also at the level of instrumental design. This road has been followed by competitive technologies (e.g. Ion Mobility Spectrometry) or other technologies initially suffering from similar problems (e.g. NIR). As an example, e-nose instruments are plagued with memory effects. That is the output pattern when testing a certain sample, is influenced by remnants from previous analysis. While this is a pretty obvious problem, few efforts have been devoted to reduce analyte absorption for instance by temperature control of the sensor chamber and associated tubing, or by the design of proper washing cycles in the sequence of analysis by the instrument.

In this paper we will discuss the importance of external validation and the need to include this requirement for archival journal publication in the area of machine olfaction. This review is organized as follows. Section II will elaborate on the robustness problems in enoses. Section III will cover the basics of validation methods including performance estimation methods. Section IV will discuss the need to develop predictive models and escape from simple exploratory methods that avoid any validation of the findings. Section V will review the main problems found in the development of classification models. Section VI discusses the relationships between the statistical concept of 'bias' and the analytical concept of robustness. Section VII introduces the considerations we should have in mind when designing validation methodologies to avoid major problems previously presented. I end the review with a summary.

II. Robustness

As I have mentioned in the introduction, robustness improvement is a pending objective in electronic nose research. For robustness improvement, many different factors may obviously contribute, namely: sensor technology improvements and better instrument design including sampling methodology. However, for the purpose of this paper we will focus on the importance of proper validation as a sanity check to ensure a good performance of the method in the operation phase (beyond method development).

The reasons for the lack of robustness in electronic noses are many:

1) Due to lack of selectivity, the instrument calibration degrades fast when new chemicals appear in the sample under analysis, that is, components not present in the calibration phase. In a similar setting, the calibration degrades when the background mixture changes to a composition beyond those encountered in the calibration phase[36].
2) Most sensor technologies are sensitive to humidity. Since humidity values can change drastically in many different scenarios (e.g. in open air sampling conditions), rejection of humidity cross-sensitivity is a must[37,38].
3) Sensor technologies are also temperature dependent. In industrial and environmental applications (open air conditions), temperature could depart strongly from the typical indoor lab conditions[38, 39]
4) Chemical sensors are prone to drift. This drift is typically a mixture of a systematic drift and a random component. An example of sensor drift can be seen in figure 2. An array of conducting polymer sensors drift when exposed to reference chemicals. This drift makes calibration models obsolete soon. For instance, in a classification problem, difficult cases due to class proximity or colinearity requiring complex classifiers are also the most sensitive to drift.
5) Identical sensor components typically have large tolerances: not only regarding the base value (response to pure air), but also regarding sensitivities. This fact, makes calibration transfer and sensor replacement particularly challenging[40].
6) Many different research works, propose the use of dynamical sensor signal features when the sensors are switched from a reference gas (typically purified air) to target stream[41-45]. These dynamic features could be sensitive to instabilities in the flow.

In addition to the traditional sources of perturbations considered in the usual definition of robustness, the e-nose practitioner could be interested in learning about the robustness of the technique to other factors. Because of the high cost of calibration samples, the user would like to learn about the robustness of the system when decreasing the size of the training set[25]. In this particular case, it is clear that simple calibration models could be more robust against a shortage of calibration samples. Similarly, it could be that the more accurate models just after calibration are the ones that degrade the faster in time (figure 6).

III. Basics of Validation

III.a. Main concepts

Validation is a basic component of any method development in the areas of chemometrics, pattern recognition and regression. Validation is motivated to address two fundamental issues: complexity control (also known as model selection) and performance estimation. Every class of predictors for the analysis of instrumental data (including electronic nose data) requires some kind of complexity control. For instance, (i) in a Partial Least Squares Discriminant Analysis approach we should decide on the number of latent variables, (ii) in a Multilayer Perceptron we should decide on the number of neurons in the hidden layer, (iii) in a k- Nearest Neighbour classifier the number of neighbors k should be optimized. In regression problems, the complexity control may take the form of a regularization parameter (that is the case of Ridge Regression). A similar case is the choice of the hyper-parameters in a support vector machine. Even in univariate regression with a polynomial basis the maximum order of the polynomial has to be selected. In figure 7, the typical evolution of prediction errors, in the training set and in the internal validation dataset, against the model complexity is illustrated. Note that typically when the ratio number of samples to dimensionality is low, optimum models tend to be of lower complexity with larger errors.

In this sense, the calibration set should be divided in two parts: one of them used for parameter estimation and the other one for complexity control. The dataset for this second task is known as the internal validation dataset. We choose as the best model complexity the one that provides the best performance in the internal validation set. Taking this performance estimation as the one characterizing the model provides usually over-optimistic results, since this data has been used for model building and in consequence the model is biased to this particular dataset. Instead, strict model validation requires the presence of an external validation dataset composed of 'fresh' or 'blind' samples that have not used at all during the model building process. In summary, the recommended practice is to have a dataset for calibration or model building, and another one for external validation. The calibration dataset is however further divided to have this internal validation that provides model complexity control and initial (but overoptimistic) performance estimation. Despite this is common knowledge, internal validation is only reported in many research papers on the basis that data scarcity precludes to have a separate dataset with blind samples for external validation.

Sometimes the terminology training dataset and test or validation dataset is used. While there is nothing wrong with this, sometimes leads to confusion arising from the fact that this terminology does not explicitly differentiate between internal and external validation. In this paper we will refer to calibration dataset (all data used during the model building process) and the external validation dataset. In our view, the data used for internal validation belongs to the calibration dataset. For us, the training set (it could be named also estimation set), is the part of the calibration set that is typically used to fit the model once the complexity has been selected. In some cases, after complexity selection the whole calibration set is used to estimate the optimum parameters.

For internal validation many diverse techniques have been proposed for an effective use of the available data. Among them we should mention leave-one-out (LOO), k-fold cross-validation, random subsampling and bootstrap. Leave-one-out is sometimes advised when the total availability of samples is limited, which unfortunately is the case in many preliminary studies in the health domain. LOO has been accused to provide overoptimistic results. It has been even proved that in the worst-case the estimation of the performance of LOO can as biased as the direct estimation over the training set[46]. Beyond these results, there are settings where the overoptimistic behavior of LOO becomes clear. We may find cases where a 1-NN classifier is used and performance estimation is done by LOO. If the dataset contains several replicas of the same experimental condition, it is clear that by LOO there would be always a replica in the training dataset, geometrically very close in feature space (assuming some reproducibility in the measurement) to the sample to be evaluated. In consequence, under this conditions LOO may suggest that the predictor is performing extremely well, and this is just an artifact.

Among the internal validation techniques bootstrapping is typically recommended (see comparison by Steryerberg et al.[47]), since it has been shown that provides almost unbiased estimates of predictive accuracy with low variance. However, only pure sampling variability is considered in bootstrap and other potential changes in the samples under analysis is not considered. In consequence, even the more strict internal validation methodology is not sufficient and indicative of the model performance in future samples.

For internal validation in machine olfaction data, we advocate the use of what we call leave one block out (LOBO). Most studies in electronic nose research use lab set-ups that permit to control the conditions of the experiment. In such cases, the experiment is parameterized according a number of parameters, namely: analyte concentrations, temperatures, humidity levels, levels of interferent analytes, etc. Typically and in order to have sufficient statistics, several replicas of the same conditions are available, although it is recommended that those replicas are not consecutive in time. Proper randomization of the sequence of conditions is a basic requirement in the design of experiments, despite leading to longer and more expensive experiments. In LOBO internal validation, a full set of conditions is set aside for validation purposes and consequently it is not present in the training set. The underlying intention is to understand if the system will be able to predict accurately the sample label (or concentration in quantitative analysis) in conditions not present in the training set (never seen before). In fact this is the case in the real operation. For instance, figure 8 shows a potential internal validation method when trying to avoid the simultaneous presence of samples from the same measurement day in the training and in the internal validation set.

Typically, the operation conditions in the real case will never be exactly the same as those in the calibration set. It may be argued that if the space of operation parameters is densely sampled there will be always some operation condition in the training set close to the validation set. However, since for practical and economical reasons the calibration set must have a minimal size, typically the sampling of the space of parameters defining the experiment is necessarily sparse. Otherwise, when exactly the same operations are present in the training and the internal validation set, then the validation conditions are inherently weak, since a minimal repeatability is sufficient to ensure that a sample in the training set will be sufficiently close to the validation sample.

III.b Performance estimation

Once the validation methodology has been decided, a final step consists in performance estimation in an external (or internal) validation set. In order to be concise, but illustrative of the main issues in performance estimation I will focus on performance estimation for binary classifiers. Binary classifiers are particularly important due to the prevalence of this problem in scientific research and particularly on biomedical applications. In most biomedical settings the problem can be set in a binary decision scenario: condition vs control, treatment vs placebo, etc. A similar discussion could be done regarding performance estimation for quantitative models but this would make the paper unnecessarily long.

Before going deep into the matter, I would like to remind that in the case of a binary classifier, we need a single discriminant function and a threshold. The discriminant function is G: $R^n \rightarrow R$, where we are assuming that the sample to be analyzed is characterized by *n* features. At this point, we do not enter in the discussion relative to feature extraction and selection. The decision rule is:

*Assign **x** to class 1 if g(**x**)> threshold, otherwise assign **x** to class 0.*

For a probabilistic interpretation it is desired that $0 \leq g(\mathbf{x}) \leq 1$. In many settings class 0 will be considered either normal or control conditions, while class 1 is considered an alarm or a condition detection.

Binary classifiers are typically evaluated using a Confusion Matrix.

	Decided Normal	Decided Alarm
Real Normal	*True-negatives (TN)*	*False positive (FP)*
Real Alarm	*False negative (FN)*	*True-positives (TP)*

Where *TN, FN, FP,* and *TP* are the numbers of true-negatives, false-negatives, false positives and true-positives respectively. A number of associated figures of merit can be computed from this table. A not exhaustive list follows:

Accuracy= *(TP+TN)/(TP+TN+FP+FN)*

Sensitivity (Recall)=*TP/(TP+FN)*

Specificity=*TN/(TN+FP)*

Positive Predictive Power=*TP/(TP+FP)*

Negative Predictive Power=*TN/(TN+FN)*

What I want to remark is that all of them are estimated with a finite sample size. Consequently, we have only an estimator of those quantities. This estimator should be considered as a random variable, and it is a good practice to report not only the estimated value but also the associated confidence limits.

Let us consider a classification scenario, where we want to estimate the overall accuracy of the classifier. We have a blind dataset of *n* samples, where n_g are correctly classified and n-n_g are not. We assume the proportion to follow a binomial distribution. The estimated accuracy and the estimated error rate are:

$$\hat{p} = \frac{n_g}{n}$$

$$\hat{q} = \frac{n - n_g}{n} = 1 - \hat{p}$$

Under the condition that:

$$n\hat{p} = n_g > 5$$

$$n\hat{q} = n - n_g > 5$$

We can assume that the estimator is normally distributed: $N\left(\hat{p}, \hat{\sigma}_{\hat{p}}\right)$

In such a case the standard deviation is estimated as:

$$\hat{\sigma}_p = \sqrt{\frac{\hat{p}\hat{q}}{n}}$$

Let us put some numbers. Imagine that we have a validation set of 120 samples, and our method provides correct classification for 110 of them. In those conditions we can report the estimated classification rate as:

$$\hat{p} = \frac{110}{120} = 0.92 \pm z_{\alpha/2}0.025 = 0.92 \pm 0.05$$

Since proportion's variance is inversely proportional to the square root of the sample count, in small samples conditions the confidence interval of the estimator grows and in many cases encloses the random choice performance (0.5 for two classes with the same prior probability). In other words, under small sample conditions it could be the case that we cannot have sufficient statistical power to claim the classifier is performing significantly better than a random label selection.

I believe that a when reporting results, we must test if the developed method has a predictive power beyond random choice. An interesting way to pose this problem is by the use of permutation tests[48-51]. In those tests the distribution of the null hypothesis is obtained by shuffling the object labels a large number of times and applying the predictive method developed. Then we test if the predictive accuracy obtained could have been obtained from the distribution of the null hypothesis. This leads to a p-value for the obtained classifier accuracy. An advantage of permutation test is that they are exact even if the observations are not normally distributed.

In electronic nose research is common to report on the comparison of several methods, either from a pure data evaluation point of view, or even different sampling conditions or instrument operation conditions. Rigorous claiming that a method A performs better than a method B, necessarily requires a hypothesis test. Hypothesis test for this kind of research should be required by editors and reviewers in the area. Unfortunately this is not the standard practice in published research. Practitioners have to be also aware that because of a statistical version of the *law of diminishing returns* claiming to beat a method whose performance is already close the limit (e.g 97%) could require a number of validation samples that in many cases is not available for practical reasons. It is also important to realize that the binomial distribution losses Gaussian character close to the upper and lower classification limits. In such cases, proper calculations of the confidence limits should be carried out using the binomial distribution.

When testing binary classifiers, confusion matrix is fully dependent on the position of the threshold[52]. There is a trade-off between sensitivity and specificity. By moving the threshold up and down, we can exchange sensitivity by specificity or vice versa. Receiver Operating Curves (ROC) curves display the probability of detection (sensitivity) against the probability of false alarms (1-Specificity) (see figure 9), when scanning the threshold in the discriminant function. In this representation, the trade-off between the probability of detection and the probability of false alarms is clearly depicted for all possible values of the threshold. The closer the curve to the upper left corner (maximum probability of detection and minimum probability of false alarms), the better the classifier is.

The selection of the optimum threshold is typically application dependant. The final user is most suited to take this decision. Miss-classification costs and prior probabilities of each class should be known for a good selection of the best threshold. Unfortunately enough, in many occasions those numbers are subject to large uncertainties, making this approach difficult to apply. In those cases, it is better to rely on Receiver Operating Characteristic Curves.

When using ROC curves, a new figure of merit appears: the Area Under the Curve (AUC) takes a value of 1 for the perfect binary classifier and a value of 0.5 for the random choice in balanced clases. ROC curves can be estimated using parametric probability density models for the condition and control classes or empirically using histogram approximations of the probability density functions. However, in this later case a large number of test cases are necessary in order to avoid excessive granularity in the estimation of the curve. A review on the use of ROC curves can be found in Fawcett et al. [52] and Lasko et al.[53].

IV. Hypothesis test vs. Predictive models: the need for model validation

In the biomedical literature, it is quite usual to characterize the goodness of a certain indicator in terms of its p-value under a certain test of hypothesis. In the last two decades there has been an abuse of the hypothesis testing approach in the quest for statistical significance in medical journals. This approach leads to "binary thinking": the null hypothesis is either accepted or rejected. In fact, it is usually more informative to think about the size of the effect (observed difference between the control and the condition), and its uncertainty due to the sample size[54]. In fact, the abuse of p-values has been pointed by Bayesian statisticians[55]. According to them, the p-value creates an illusion that knowledge can be summarized in a single number, without further information external to the experiment. Goodman points several examples from the biomedical literature where the p-value alone does not make much sense. Only by incorporation of prior knowledge the interpretation of the experiment can be done correctly. Goodman proposes instead the use of Bayes factors[56]. In its simpler form, Bayes factors are a measure of how the models explain the observed data. In this formalism, both hypothesis (null and alternative) are considered as probabilistic models that can generate the observed data. Bayes factors are also called the weight of the evidence and they tell us how the ratio of the probabilities for both hypotheses change after observing the data. While it has a strong foundation on Bayesian statistics, however in the biomedical use of electronic noses this formalism has never been used as far as I know. The theory behind Bayes factors is beyond the scope of this review. An in depth introduction is found in the review paper by Kass[57].

Additionally, it is very important to realize that p-value differs from predictive power. This is clear with a simple example from univariate statistics. Let us consider that we have a human population and we take height as an indicator to predict gender. It is clear that even with a moderate sample size, the null hypothesis (both genders have the same height) will be clearly rejected with a very small p-value. However, if we build a predictive model only on the basis of this single indicator (height), the accuracy of this predictor will be rather poor. In other terms, it is possible to have indicators with small p-values and very moderate predictive powers. Take note of the next example taken from Broadhurst[58] (see figure 10): Both ANOVA and t-test produce p-values on the order of 10^{-4}. In the usual medical practice this would be considered as a clear sign of the statistical significance of the indicator. However, the indicator delivers a moderate Area under the curve AUC=0.68 (AUC-see section III.b). We have to take into account which hypothesis are we testing. For example, when using the t-test, we are evaluating if the mean of two populations differ. If we have enough number of examples the power of the test is such that even with populations generated by probability density functions with a large overlap, the test will see that the means are different. We can also argue that ROC curves, and related figures of merit like the AUC, are more sensitive to the full shape of the involved probability density functions, while some hypothesis test (e.g. ANOVA) are only based in variance estimates.

In a second example, also cited by Broadhurst[58] but originally from Kenny et al.[59] , a binary classifier working with 286 univariate features and 174 samples is tested. Every individual

feature is tested according the Area Under the Curve measurement but also the p-value of the t-test. Figure 11 shows that feature with non-significant p-values at the usual risk levels are able to provide very good AUCs. We have to take into account that p-values can be very sensitive to the sample size and to the adequacy of the underlying hypothesis of the test.

In fact, the larger the sample size, the smaller has to be the difference (relative to the standard deviation) to have enough statistical power. It is well-known that the power of a t-test to observe differences in the mean of both populations increases with the sample size, so larger samples give smaller p-values. However, this does not automatically translate to a better accuracy on a predictive binary classifier. For a 1-D problem with two normal classes, characterized by the same variance and different mean, a larger sample size brings better accuracy in the estimation of the means and the variance, but the asymptotic Bayesian error is fixed, and a larger sample size would not bring any decrease in the classification error rate beyond this point. In this context it is important to recall the statement by Gardner and Altman[54]:" Small differences of no real interest can be statistically significant with large sample sizes, whereas clinically important effects maybe statistically non-significant only because of the number of subjects studied was small".

V. Main errors in the development of predictive classifier models

In this section, I will review the main errors artificial olfaction researchers fall in, when building classification models. While the description is mostly based on binary classifiers the arguments can be easily generalized to a larger number of classes.

1) **Overfitting:** the research work fails to use independent/blind samples which are held back from model optimization and used only to test the robustness of the prediction. In many occasions, only the cross-validation error is given. In other words, model performance is estimated with the same samples that have been used to decide on the optimum model complexity. This leads to overfitting and to report over-optimistic model performance. Despite this being a well-known error, it still appears frequently due to small sample datasets. A related error that appears frequently is the use of sample replicas to increase the sample count for the study. However, since most methods assume that the samples are identically and independently sampled from a larger population, the use of replicas breaks the assumption of independence. In consequence, reported results (using an apparently bigger dataset) may be overoptimistic.
2) **Confusing statistical significance with predictive accuracy:** A number of features can be providing very small p-values and still provide low predictive accuracy. This misinterpretation of the meaning of the p-value mostly appears in the biomedical literature. This has been already discussed in the previous section.
3) **Ignoring the prior probabilities of the classes.** Classifiers are built with balanced classes (to avoid unbalanced datasets that pose problems to many classification algorithms), but they are later applied ignoring the prior probabilities. This error mostly appears in circumstances where the estimation of the prior probabilities is difficult or maybe it is not possible.

4) **Bias.** This is the most important cause of error and it has multiple origins. We refer to bias when a feature is differentially distributed between the classes but it happens to be correlated with another uncontrolled variable that truly underlies the variance of the feature. In the biomedical literature, this confounding variables may originate from the patient (smoking status, gender, diet….) but in many other occasions it can be also of instrumental origin (different time, different location, different operator, different instrument, different environmental variables, sampling differences, etc). I will review this issue later on in section VI.

According to Ransohoff[60] "Bias is the most important threat to validity". In fact, a healthy skeptical view in Machine Olfaction results is a recommended best practice. Among the potential confounding factors leading to bias two different cases should be considered: (i) Potential confounding factors that have been identified, (ii) Confounding factors ignored by the researcher.

In the first case, the researcher is advised to investigate the potential exploratory power of those factors. If the approach is based on test of hypothesis, we recommend using techniques able to deal with multiple factors whenever possible (e.g. Multi-way ANOVA). If a predictive model is built, try to use techniques based on orthogonal projections, such that the confounding factor is approximately rejected.

Among the many sources of confounding factors, I would like to remark instrumental shifts or system drifts as potential confounding factor. It is very important, to block those potential confounding factors during the phase of experimental design. An example of a bad experimental design would be as follows. Imagine that we are using an electronic nose to investigate the potential of breath analysis for lung cancer diagnostics. If we test 10 lung cancer subjects in month 1, and month 3 we test 10 control subjects, unless further evidences are shown it is impossible to attribute the observed differences to different time or to different condition. The obvious correct design involves testing 5 cancer plus 5 controls in each month. While this is an extreme case, presented only for illustrative purposes, it is clear that unbalanced distribution may easily appear: e.g 8 cancer and 2 controls in month 1, and 7 controls and 3 cancer in month 3. Even in this later case, time can be a clear source of the observed variance between cancer and control groups.

A famous case in the area of proteomics is controversial report in ovarian cancer screening using serum proteomics published in 2002 claiming excellent sensitivity and specificiy[61]. Those excellent results were proven later to be caused by bias, and that the method had no real discrimination power. In this study, spurious classification performance is obtained by storage artifacts when samples were stored at -20ºC. A particular mass m/z=6638.41, was apparently a cancer biomarker, however it was correlated with sample storage time. Since healthy samples were collected first and ill-patients later, the storage time was a confounding factor that in this particular case was correctly identified[62]. However, in the general case the possibility of hidden confounding factors requires skepticism and strict validation.

VI. The relationship between Bias and Robustness.

In the introduction the very important concept of Robustness (or Ruggedness) has been already presented. The importance of robustness assessment in method validation has been already pointed out by Vander-Heyden[63] and Zaiter[64,65].

Gas sensor arrays ~~suffer from the consideration of~~are generally considered ~~–~~unreliable instruments. While literature reports a large number of successful applications, probably they are very sensitive to experimental conditions since translation to the industry or to the clinic is a daunting task.

Gas sensor arrays based instruments are particularly sensitive to many factors. For this reason we have to be careful in our analysis. We can encounter two diverse situations, namely:

(i) Some ignored factors are the ultimate reason for the observed differences between the classes and a success is reported. In this case we talk about biased experiments.

(ii) Correct experimental design and extreme care, makes external factors constant in lab conditions and then the observed differences are exactly due to the investigated condition. However, in real settings (industrial or clinical) those factors cannot be conveniently controlled leading to the practical impossibility to apply the proposed method. In this case we talk about lack of robustness. That is, the method fails to properly work beyond the limits of the analysis.

If we want to avoid both problems, the most important recommendation is to identify potential sources of variance beforehand in the experimental design. At this point, we should be fully aware of the expected working conditions of the instrument in the potential application targeted. This is relevant, since it could be the case that some of those external factors could be appropriately controlled, but others maybe not because of practical or economical reasons. Incorporation of the potential sources of variance in our analysis leads to a much richer discussion. Only when including those sources of variance in the experiment we would be able to understand their relevance and the inform the reader of the research report about: (i) the applicability limits of the proposed solution and consequently the need to control important factors during the application of the method, (ii) the need to devise methods to algorithmically counteract them.

VII. Discussion: strategies for model validation and prediction reliability estimation

Assessment of prediction reliability is a main objective for model evaluation. This is specially challenging when the number of available samples is small. In order to make an efficient use of the available samples a number of validation methodologies have been proposed in the literature[66] . A basic concept in model validation is to differentiate between internal validation and external validation. While most papers describe and use classical internal validation procedures such as leave one out (LOO), k-fold cross-validation and bootstrapping, we

emphasize in this paper that external validation by an external test set is a must[67]. External validation refers to use samples that were not used in the model estimation. Those samples are external samples, to differentiate from internal validation samples. It is important to note, that internal validation samples belong to the calibration set and are used basically for algorithmic complexity control and model optimization and in consequence the model is tuned to the samples used for the training set. In many biomedical studies, portioning the available data in training, internal validation and external validation sets leads to a very low cardinality for this last set. In those conditions, performance assessment with external validation suffers from a high variance. To solve this issue the repeated double cross-validation procedure has been proposed[68] . Repeated double cross-validation uses two nested loops. The outer loop is the division between calibration set and external validation set and it is used to estimate the prediction performance. The inner loop uses the calibration set for model selection and parameter optimization, using classical cross-validation procedures (CV), that is dividing this in training and internal validation sets. The strategies for data partitioning in the outer loop are similar to the CV procedures. In principle, this method allows to use all the available data for performance estimation, largely decreasing the variance problems mentioned above. It has to be taken into account that this procedure leads to a multiplicity of optimal models (*N*) corresponding to different calibration sets in each outer loop. The stability of those models in terms of errors but also in terms of architecture and parameters can give further insight on the sensitivity of the model to the training data. If this sensitivity is assessed as too high, extra caution has to be taken concerning the final results.

Care has to be taken with the strict definition of external validation. In many occasions, and in particular in the repeated cross-validation procedure just described, the external validation set is only a random selection of points from the experimental dataset. While it is true that those samples are not used for building the model, we have to consider that: (i) in a random division between the calibration and the external validation, it is possible that there is always one sample in the calibration set that is very close to the sample to be predicted in feature space. In this case, prediction models based on local information can overfit easily. (ii) This external validation procedure is weak in the sense that the calibration set and the external validation set will contain samples taken in the same operational conditions, with the same operator, with the same environmental parameters. In random selection, this will happen even if the original dataset spans a diversity of operational conditions. In consequence, since conditions appear simultaneously in both sets, this validation method is not able to prove that the model will generalize conveniently under a change in those operational conditions.

I believe that it must be a recommended practice to state in the description of the model a definition of the application domain. This application domain is directly related to the operation conditions spanned by the training set. Unfortunately, many research works in machine olfaction are not concerned with these issues when developing applications and the application limits of the models remain unknown.

We should emphasize that we cannot expect that models produce reliable predictions in the entire universe of operation conditions. It is a good practice to project the measurement sample in the input space and check if this sample can be considered an outlier. Methods for outlier detection can range from T2 and Q2 measures derived from principal component

analysis to more sophisticated methods[69]. In any case, extrapolation is always dangerous and prediction should be treated with caution.

In fact, from a multivariate calibration and pattern recognition point of view, predictive models fail because the final application the conditions at real-world application are different from the conditions of predictive model buildingin which we use the instrument differ from the condition in which the predictive models where built. It is our point that external validation should consider all the expected conditions in which the instrument should operate. Only in those cases, the estimated predictive power could be considered as realistic. Here I reconsider restate the main points that are typically ignored in basic validation.

(i) Drift (Instrumental Shifts): Predictive models can be very sensitive to the time elapsed since last calibration. Periodic re-calibration is the ultimate solution, however, since the calibration of electronic noses is costly and time consuming, full recalibration is not a desired option. The issue of re-calibration with a minimum number of experimental conditions has received insufficient attention in the literature.

(ii) Change in the operation conditions: A clear example is the influence of the sampling conditions. Many AO Machine Olfaction systems explore the headspace of a certain material. In those cases, the temperature of the material and the flow sweeping the headspace are clear sources of problems. They have to be under control to avoid undesired results.

(iii) Humidity changes: Most chemical sensors are also sensitive to humidity. Humidity control is not desired in many settings. In those cases, independent measurement of humidity and their compensation is a must.

(iv) Background or Matrix Effect: It could be easily the case, that we are able to have a controlled background in the lab so we are able to see differences in the sensor signals due to the target condition. However, the effect of a change in the background can be much larger than the difference that we have identified, making predictive models unusable. Background changes can strongly shift our pattern response, but not only due to cross-sensitivities due to the background components in an additive manner. Non-linearities (due in some cases to competitive effects among analytes) can also change the sensitivity to the target analyte. In some cases, high concentrations of a particular component can decrease orf fully inhibit the response to key analytes in the target condition that has to be detected.

(v) Sensor replacement: This is a very practical case, but not less important. Sensor components degrade and ultimately fail for diverse reasons. It is also well-known that sensor tolerances, both in baseline values but also in sensitivity are large. Replacement of even a single sensor can make predictive models obsolete.

(vi) Calibration transfer: Independent calibration of each instrument for the targeted application can be burdensome. Extensive calibration can be carried out in a single (or a few) instrument. The performance of the models in equivalent instruments from a constructive point of view needs investigation. The issue of calibration transfer in the e-nose literature has received insufficient attention.

External validation could be used to explore the limits of the model application domain. We have mentioned just before the main reasons that can cause model failure. Now some recommendations are given to assess those limits by means of a more strict external validation policy. Let us revise them briefly.

(i) Drift. The robustness of the method in time needs to be ~~checked~~tested. For this several considerations are in order. (1) First it is a must that there is an external validation test in the future of the calibration data. It is obvious that this is the condition of any real application. The use of the predictive model is always in the future of the calibration. In other words: first the instrument is calibrated (the predictive model is built) and the instrument is applied to new samples. This is not the case in most cross-validation strategies (with the exception of hold-out), where the label of external validation sample is predicted taking into account samples measured before and after this particular sample. ~~calibration set contains samples in the future of the sample in the external validation set.~~ Leave-one out, random subsampling, k-fold methods, and even bootstrap ignore that the instrument could be a time-varying system. My recommendation is that they can be used for complexity control (model selection), but never for final estimation of performance, since those methods ignore the basic fact that in the real case the samples under analysis are in the future of the calibration. Any research work that reports only validation on those conditions should be presumed of being overoptimistic. This is also the case for random division in calibration and external validation, and it is also the case for weak external validation methods as the repeated double cross-validation. In this setting, again there will be calibration samples in the future of validation samples.

The research report would be even richer if the issue of model life time is addressed. The estimation of the life-time of a predictive model is a complex issue since diverse conditions of operation can lead to different lifetimes. Few papers have reported methodologies for the user to detect when the predictive model is becoming obsolete. Blind application of the predictive model could lead to errors. A basic sanity check is that the new measurement, the new obtained pattern is within the limits of the calibration data. If the new measurement could be considered as an outlier falling outside the range of the calibration data, the measurement and the outcome of the predictive model has to be marked for further investigation.

(ii) Change of the operation conditions (including temperature/humidity). It is recommended that operation conditions suspicious to have a big influence in the system output have to be included in the experimental design, particularly if they cannot be easily controlled in the final application. That is the calibration set data should contain also a variation of those factors (could be systematic or random). Other less recommended option is to check for sensitivity to those parameters after building the predictive model. The obvious drawback of this later approach is that if the results prove the parameter to have a large influence on the pattern distribution, we will be forced to repeat the calibration experiments taking into account this additional parameter.

(iii) Background changes: Background variations are a clear problem for applications that should operate in field conditions. At the lab, it is in many occasions difficult to replicate the complexity of the backgrounds that the instrument can encounter in real, field conditions. In this case, the clear recommendation is that the predictive model has to be built with samples obtained in field conditions and taking care that a sufficient coverage of the potential changes of the background is obtained.

(iv) Sensor Replacement and Calibration transfer: In many occasions this could be a limiting factor for the real applicability of the method and the instrument. I invite (challenge) researchers in the area to test if their models are robust enough to keep accuracy when they change an instrument component (e.g. sensor in a sensor array) or when they apply the same prediction model to a new "identical" instrument.

VIII. Conclusions

Electronic Noses are very sensitive instruments and rather unspecific. Research in the e-nose domain has been dominated during decades by sensor technology developers trying to show how good their sensors were to solve numerous applications in diverse fields. However, it is clear today, that some of those studies were at least naive. The weakness of those studies was in most cases obscured by poor validation strategies, and generalization claims beyond the actual experimental evidences brought by the research work. In fact, the main conclusion of the present paper is that the validation methodologies in electronic noses research should be necessarily more strict and rigorous. We propose that any e-nose application study should support their findings by using external validation datasets or the so-called blind samples. This external validation samples have to be in the future of the training (calibration) dataset. Additionally, the design of the external validation has to address potential pitfalls of the analysis the authors could identify. A much restricted validation set, possibly indicates that the method is inherently weak. Authors should be forced to prove the robustness of their proposed method beyond current standards. The time stability of the results, background shifts, environmental parameters and other disturbance factors should be included in the study for maximum credibility and potential of future translation to real applications beyond the lab. External validation should help authors to identify the limits of applicability of the developed predictive models.

Acknowledgements

This work has been funded by the Spanish Ministerio de Economía y Competitividad under the project TEC2011-26143. Santiago Marco is member of the consolidated research group SGR2009-0753 by the Generalitat de Catalunya.

References

1. Dragonieri S, Schot R, Mertens BJ, Le Cessie S, Gauw SA, Spanevello A, Resta O, Willard NO, Vink TJ, Rabe KF, Bel EH, Sterk PJ (2007) An electronic nose in the discrimination of patients with asthma and controls. The Journal of allergy and clinical immunology 120:856–862. doi: 10.1016/j.jaci.2007.05.043
2. Montuschi P, Mores N, Trové A, Mondino C, Barnes PJ (2013) The electronic nose in respiratory medicine. Respiration; international review of thoracic diseases 85:72–84. doi: 10.1159/000340044
3. Greulich T, Hattesohl A, Grabisch A, Koepke J, Schmid S, Noeske S, Nell C, Wencker M, Jörres RA, Vogelmeier CF, Köhler U, Koczulla AR. (2013) Detection of obstructive sleep aponea by an electronic nose. The European respiratory journal : official journal of the European Society for Clinical Respiratory Physiology. Vol. 42, 145-155. doi: 10.1183/09031936.00091712
4. Lazar Z, Fens N, van der Maten J, van der Schee MP, Wagener AH, de Nijs SB, Dijkers E, Sterk PJ. (2010) Electronic nose breathprints are independent of acute changes in airway caliber in asthma. Sensors, 10:9127–9138. doi: 10.3390/s101009127
5. Biller H, Holz O, Windt H, Koch W, Müller M, Jörres RA, Krug N, Hohlfeld JM. (2011) Breath profiles by electronic nose correlate with systemic markers but not ozone response. Respiratory medicine 105:1352–1363. doi: 10.1016/j.rmed.2011.03.002
6. Chapman E, Thomas PS, Stone E, Lewis C, Yates DH, (2012) A breath test for malignant mesothelioma using an electronic nose. The European respiratory journal 40:448–454. doi: 10.1183/09031936.00040911
7. Hattesohl A, Jörres R, Dressel H, Schmid S, Vogelmeier C, Greulich T, Noeske S, Bals R, Koczulla AR. (2011) Discrimination between COPD patients with and without alpha 1-antitrypsin deficiency using an electronic nose. Respirology 16:1258–1264. doi: 10.1111/j.1440-1843.2011.02047.x
8. Valera J, Togores B, Cosio B (2012) Use of the electronic nose for diagnosing respiratory diseases. Archivos de bronconeumología 48:187–188. doi: 10.1016/j.arbres.2011.08.004
9. Fens N, Zwinderman A, van der Schee M, de Nijs SB, Dijkers E, Roldaan AC, Cheung D, Bel EH, Sterk PJ. (2009) Exhaled breath profiling enables discrimination of chronic obstructive pulmonary disease and asthma. American journal of respiratory and critical care medicine 180:1076–82. doi: 10.1164/rccm.200906-0939OC
10. Di Natale C, Macagnano A, Martinelli E, Paolesse R, D'Arcangelo G, Roscioni C, Finazzi-Agrò A, D'Amico A. (2003) Lung cancer identification by the analysis of breath by means of an array of non-selective gas sensors. Biosensors & bioelectronics 18:1209–18
11. Machado R, Laskowski D, Deffenderfer O, Burch T, Zheng S, Mazzone PJ, Mekhail T, Jennings C, Stoller JK, Pyle J, Duncan J, Dweik RA, Erzurum SC.(2005) Detection of lung cancer by sensor array analyses of exhaled breath. American journal of respiratory and critical care medicine 171:1286–1291. doi: 10.1164/rccm.200409-1184OC
12. Fens N, de Nijs S, Peters S, Dekker T, Knobel HH, Vink TJ, Willard NP, Zwinderman AH, Krouwels FH, Janssen HG, Lutter R, Sterk PJ. (2011) Exhaled air molecular profiling in relation to inflammatory subtype and activity in COPD. The European respiratory journal 38:1301–1309. doi: 10.1183/09031936.00032911
13. Pennazza G, Marchetti E, Santonico M, Mantini G, Mummolo S, Marzo G, Paolesse R, D'Amico A, Di Natale C. (2008) Application of a quartz microbalance based gas sensor array for the study of halitosis. Journal of breath research 2:017009. doi: 10.1088/1752-7155/2/1/017009
14. Persaud K (2005) Medical applications of odor-sensing devices. The international journal of lower extremity wounds 4:50–6. doi: 10.1177/153473460527513.

15. Kodogiannis V, Lygouras J, Tarzynski A, Chowdrey H (2008), Artificial odor discrimination system using electronic nose and neural networks for the identification of urinary tract infection, IEEE transactions on information technology in biomedicine, 12, 707-703, doi: 10.1109/TITB.2008.917928.
16. Covington J, Wedlake L, Andreyev J, Ouaret N, Thomas MG, Nwokolo CU, Bardhan KD, Arasaradnam RP. (2012) The detection of patients at risk of gastrointestinal toxicity during pelvic radiotherapy by electronic nose and FAIMS: a pilot study, Sensors, 12, 13002-13018, doi: 10.3390/s121013002
17. Marco S, Gutierrez-Galvez A (2012) Signal and Data Processing for Machine Olfaction and Chemical Sensing: A Review. IEEE Sensors Journal. doi: 10.1109/JSEN.2012.219292
18. Gutierrez-Osuna R (2002) Pattern analysis for machine olfaction: a review. IEEE Sensors Journal. doi: 10.1109/JSEN.2002.800688
19. Gutiérrez, A., Marco, S. (Eds.). (2009). Biologically Inspired Signal Processing for Chemical Sensing (Studies in Computational Intelligence Vol. 188). Springer.
20. Persaud KC, Marco S, Gutierrez-Galvez A (Eds) (20139, Neuromorphic Olfaction, in Frontiers in Neuroengineering, CRC Press.
21. Ioannidis J (2005) Why most published research findings are false. PLoS medicine 2:e124. doi: 10.1371/journal.pmed.002012.
22. Cornfield J (1966) Sequential Trials, Sequential Analysis and the Likelihood Principle. The American Statistician. doi: 10.1080/00031305.1966.10479786.
23. Defernez M, Kemsley EK (1997) The use and misuse of chemometrics for treating classification problems, TrAC Trends in Analytical Chemistry, Vol. 16, pp- 216-221. doi: 10.0116/S0165-9936(97)00015-0.
24. Marco S, Ortega A, Pardo A, Samitier J (1998) Gas identification with tin oxide sensor array and self-organizing maps: adaptive correction of sensor drifts. Instrumentation and Measurement, IEEE Transactions on 47:316–321.
25. Padilla M, Perera A, Montoliu I, Chaudry A, Persaud K, Marco S, (2010) Drift compensation of gas sensor array data by orthogonal signal correction. Vol. 100, pp. 28-35.
26. Ziyatdinov A, Marco S, Chaudry A, Persaud K, Caminal P, Perera A. (2010) Drift compensation of gas sensor array data by common principal component analysis. Sensors and Actuators B: Chemical. Vol. 146, pp. 460-465. doi: 10.1016/j.snb.2009.11.034.
27. Martinelli E, Magna G, Vito SD, Di Fuccio R, Di Francia F, Vergara A, Di Natale C (2013) An adaptive classification model based on the Artificial Inmune System for chemical sensor drift mitigation, Sensors and Actuators B: Chemical, Vol. 177, pp. 1017-1026, doi: 10.1016/j.snb.2012.11.107
28. Di Carlo S, Falasconi M, Sanchez E, Scionti A, Squillero G, Tonda A. (2011) Increasing pattern recognition accuracy for chemical sensing by evolutionary based drift compensation, Pattern Recognition Letters Vol 32. , pp. 1594-1603.
29. Gutierrez-Osuna, R. (2000) Drift reduction for metal-oxide sensor arrays using canonical correlation regression and partial least squares. *Proceedings of the 7th International Symposium On Olfaction & Electronic Nose*. Pp. 147-152, IOP Press, London.
30. Artursson T, Eklov T, Lundstrom I, Martensson, Sjostrom M, Holmberg M, "Drift correction for gas sensors using multivariate methods." Journal of chemometrics 14.5-6 (2000): 711-723.
31. Knobloch H, Turner C, Spooner A, Chambers M (2009) Methodological variation in headspace analysis of liquid samples using electronic nose. Sensors and Actuators B: Chemical 139:353–360.
32. Kuske, M., Rubio, R., Romain, A. C., Nicolas, J., & Marco, S. (2005). Fuzzy *k*-NN applied to moulds detection. Sensors and Actuators B: Chemical,106(1), 52-60.
33. Kuske M, Padilla M, Romain AC, Nicolas J, Rubio R, Marco S. Detection of diverse mould species growing on building materials by gas sensor arrays and pattern recognition. Sensors and Actuators B: Chemical 119.1 (2006): 33-40.

34. Adam G, Lemaigre S, Romain A-C, Nicolas J, Delfosse P. (2013) Evaluation of an electronic nose for the early detection of organic overload of anaerobic digesters. Bioprocess and biosystems engineering 36:23–33. doi: 10.1007/s00449-012-0757-6.
35. QualityAssurance of Pharmaceuticals: A Compendium of Guidelines and Related Materials, 2007 :World Health Organization
36. Gujral P., Amrhein M., Wise B. M., Bonvin, D. (2010). Framework for explicit drift correction in multivariate calibration models. Journal of Chemometrics, 24(7-8), 534-543.
37. Nimsuk, N., Nakamoto, T. (2008). Study on the odor classification in dynamical concentration robust against humidity and temperature changes. Sensors and Actuators B: Chemical, 134(1), 252-257.
38. Kashwan, K. R., Bhuyan M. (2005) "Robust electronic-nose system with temperature and humidity drift compensation for tea and spice flavour discrimination." *IEEE Asian Conference on Sensors and the International Conference on new Techniques in Pharmaceutical and Biomedical Research,* 5-7 Sept. 2005, Kuala Lumpur, Malasya, pp 154-158.
39. Romain, A. C., Nicolas, J., & Andre, P. (1997). In situ measurement of olfactive pollution with inorganic semiconductors: Limitations due to humidity and temperature influence. In Seminars in Food analysis, 2, 283-296.
40. Tomic, O., Eklöv, T., Kvaal, K., & Haugen, J. E. (2004). Recalibration of a gas-sensor array system related to sensor replacement. Analytica chimica acta, 512, 199-206.
41. Marco S, Pardo A, Davide FA, Di Natale C, D'Amico A, Hierlemann A, Mitrovics J, Schwweizer M, Weimar U, Göpel, W. (1996). Different strategies for the identification of gas sensing systems. Sensors and Actuators B: Chemical, 34(1), 213-223.
42. Marco S, Samitier J, Morante JR (1995) A novel time-domain method to analyse multicomponent exponential transients. Measurement Science and Technology, 6(2), 135.
43. Samitier J, Lopez-Villegas J M, Marco S, Camara L, Pardo A, Ruiz O, Morante JR (1994). A new method to analyse signal transients in chemical sensors. Sensors and Actuators B: Chemical, 18(1), 308-312.
44. Gutierrez-Osuna R, Nagle HT, Schiffman SS (1999) Transient response analysis of an electronic nose using multi-exponential models.Sensors and Actuators B: Chemical, 61(1), 170-182.
45. Vilanova X, Llobet E, Alcubilla R, Sueiras JE, Correig X (1996). Analysis of the conductance transient in thick-film tin oxide gas sensors.Sensors and Actuators B: Chemical, 31(3), 175-180.
46. Kearns M, Ron D (1999) Algorithmic stability and sanity-check bounds for leave-one-out cross-validation. Neural computation 11:1427–1453. doi: 10.1162/08997669930001630.
47. Steyerberg E, Harrell F, Borsboom G, Eijkemans MJ, Vergouwe Y, Habbema JD. (2001) Internal validation of predictive models: efficiency of some procedures for logistic regression analysis. Journal of clinical epidemiology 54:774–81
48. Lindgren F., Hansen B, Karcher W, Sjöström M, Eriksson L (1996), Model validation by permutation tests: Applications to variable selection. J. Chemometrics, 10: 521–532.
49. Ojala M., Garriga GC (2010) Permutation tests for studying classifier performance, J. of Machine Learning Research, 11, 1833-1863.
50. Westerhuis, JA, Hoefsloot HC, Smit S, Vis DJ, Smilde AK, van Velzen EJ, . van Duijnhoven JPM, van Dorsten FA(2008). Assessment of PLSDA cross validation. Metabolomics, 4(1), 81-89.
51. Rubingh CM, Bijlsma S, Derks EP, Bobeldijk I, Verheij ER, Kochhar S, Smilde AK (2006), Assessing the performance of statistical validation tools for megavariate metabolomics data, Metabolomics, 2, 53-61.
52. Fawcett T (2006). An introduction to ROC analysis. Pattern recognition letters, 27(8), 861-874.
53. Lasko TA, Bhagwat JG, Zou KH, Ohno-Machado L. (2005): "The use of receiver operating characteristic curves in biomedical informatics." Journal of biomedical informatics 38,404-415.

54. Gardner M, Altman D (1986) Confidence intervals rather than P values: estimation rather than hypothesis testing. Br Med J. 292, 746-750.
55. Goodman S (1999) Toward evidence-based medical statistics. 1: The P value fallacy. Annals of internal medicine 130:995–1004
56. Goodman S (1999) Toward evidence-based medical statistics. 2: The Bayes factor. Annals of internal medicine 130:1005–1013
57. Kass R, Raftery A (1995) Bayes factors. J. American Statistical Association, vol. 90, 773-795. doi: 10.1080/01621459.1995.10476572.
58. Broadhurst DI, Kell, DB (2006). Statistical strategies for avoiding false discoveries in metabolomics and related experiments. Metabolomics, 2, 171-196.
59. Kenny LC, Dunn WB, Ellis DI, Myers J, Baker PN, Kell DB (2005). Novel biomarkers for pre-eclampsia detected using metabolomics and machine learning. Metabolomics, 1, 227-234.
60. Ransohoff DF (2005). Lessons from controversy: ovarian cancer screening and serum proteomics. Journal of the National Cancer Institute, 97, 315-319.
61. Petricoin EF, Ardekani AM, Hitt BA, Levine PJ, Fusaro VA, Steinberg SM, Mills GB, Simone C, Fishman DA, Kohn EC, Liotta LA (2002). Use of proteomic patterns in serum to identify ovarian cancer. The lancet, 359(9306), 572-577.
62. West-Nørager M, Bro R, Marini F, Høgdall EV, Høgdall CK, Nedergaard L, Heegaard NH. al. (2009) Feasibility of serodiagnosis of ovarian cancer by mass spectrometry. Analytical chemistry 81:1907–13. doi: 10.1021/ac802293g
63. Vander Heyden Y, Nijhuis A, Smeyers-Verbeke J, Vandeginste BG, Massart DL.et al. (2001) Guidance for robustness/ruggedness tests in method validation. Journal of pharmaceutical and biomedical analysis 24:723–53.
64. Zeaiter M, Roger J-M, Bellon-Maurel V, Rutledge DN (2004) Robustness of models developed by multivariate calibration. Part I. TrAC Trends in Analytical Chemistry. Vol 23, 157-170, doi: 10.1016/S0165-9936(04)00307-3
65. Zeaiter M, Roger J-M, Bellon-Maurel V (2005) Robustness of models developed by multivariate calibration. Part II: The influence of pre-processing methods. TrAC Trends in Analytical Chemistry. Vol, 24, pp. 437-445. doi: 10.1016/j.trac.2004.11.023.
66. Lavine BK (2009), Validation of Classifiers in *Comprehensive Chemometrics*, Vol. 3, 587-598.
67. Esbensen KH, Geladi P, (2010) Principles of Proper Validation: use and abuse of re-sampling for validation, J. of Chemometrics, 34, 168-187.
68. Filzmoser P, Liebmann B, Varmuza K, (2009) Repeated double cross validation, J. of Chemometrics, vol. 23, 160-171.
69. Rousseeuw J, Debruyne M, Engelen S, Hubert M (2006), Robustness and Outlier Detection in Chemometrics, Critical Reviews in Analytical Chemistry, 36, 221-242.

Figure 1: Block diagram of an electronic nose featuring an embedded pc for on-board signal and data processing.

Figure 2. Left) Response of a conducting polymer sensor to three reference chemicals in a 300 days time span. Right) Principal Components score plot of a 17 Conducting Polymer Sensor Array. Arrows show the time evolution of the patterns in time. (after Padilla et al.[25])

Figure 3. When sampling the headspace of a liquid, the sensor array response becomes heavily dependent on the flow over the headspace (after Knobloch, et al.[31])

Figure 4. Response patterns for identical instruments in the same conditions have been found to largelly differ (after Knocloch et al.[31])

Figure 5. Scoreplot of a MOX sensor array for indoor fungi detection. The sensor patterns are heavily dependent on the material where the fungi could grow (after Kuske et al.[33]).

Figure 6: The evaluation of the robustness of the method should include from sampling, to the instrument operation but also the predictive model. With the same sampling method and instrument, different algorithms could behave differently concerning time stability or regarding the scarcity of data in the calibration set.

Figure 7. Evolution of the errors (in training and in external validation) depending on the model complexity for two case of the ratio between sample number and dimensionality.

Figure 8. Example of the internal validation methodology to explore the effect of the measurement day. The training set does not contain measurements from the day used for internal validation. For a better use of the available data for internal validation, this scheme (leave one day out) is repeated for all measurement days.

Figure 9, Evolution of the Receiver Operating Characteristic plot for two normal populations of the same variance and increasing means difference.

Figure 10: Synthetic example that features very small p-values and only a moderate AUC (from Broadhurst[58])

Figure 11. Binary classifier with 286 features and 174 samples. Univariate tests of individual features vs. Area Under the Curve in an independent test set (from Broadhurst[58]).

Figure 1.

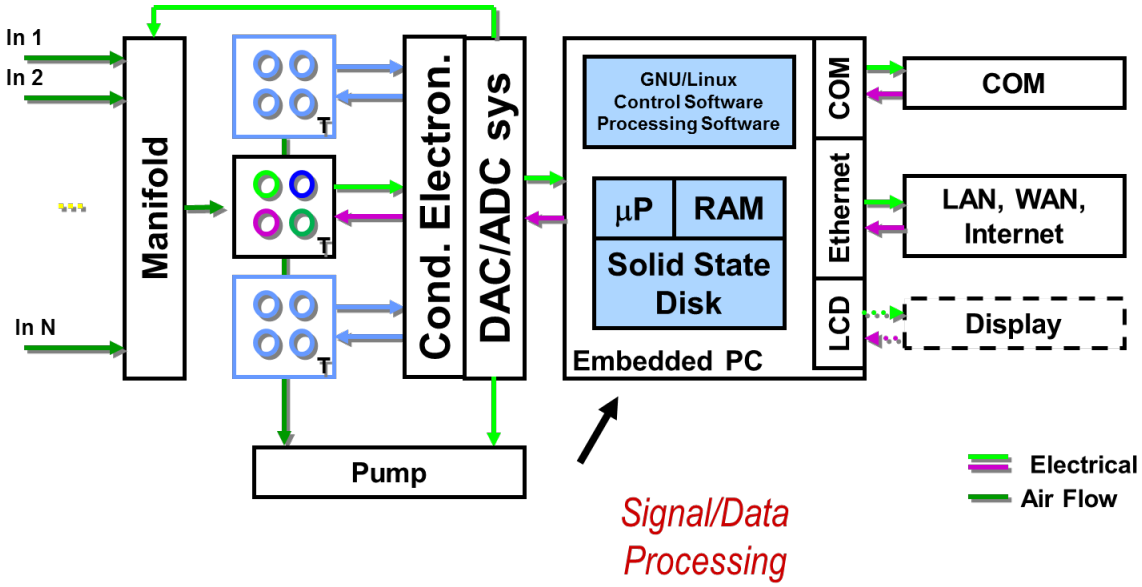
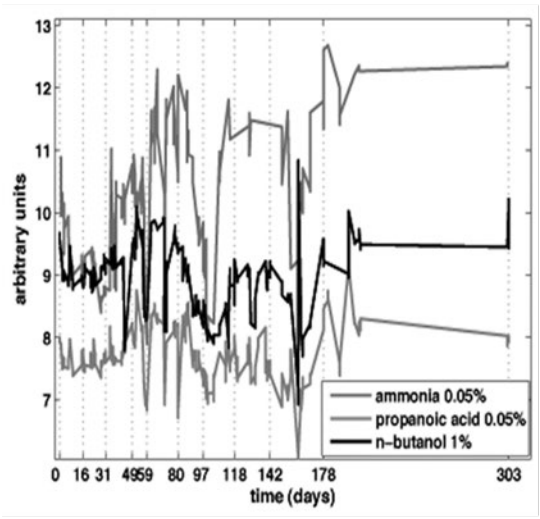

In 1
In 2
In N
Manifold
T
T
T
Cond. Electron.
DAC/ADC sys
GNU/Linux
Control Software
Processing Software
µP
RAM
Solid State
Disk
Embedded PC
COM
Ethernet
LCD
COM
LAN, WAN,
Internet
Display
Pump
Electrical
Air Flow
Signal/Data
Processing

Figure 2


arbitrary units
time (days)
ammonia 0.05%
propanoic acid 0.05%
n-butanol 1%

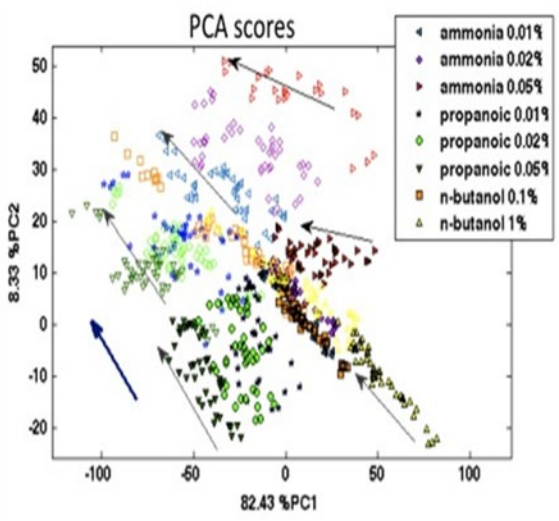

PCA scores
ammonia 0.01%
ammonia 0.02%
ammonia 0.05%
propanoic 0.01%
propanoic 0.02%
propanoic 0.05%
n-butanol 0.1%
n-butanol 1%
8.33 %PC2
82.43 %PC1

Figure 3

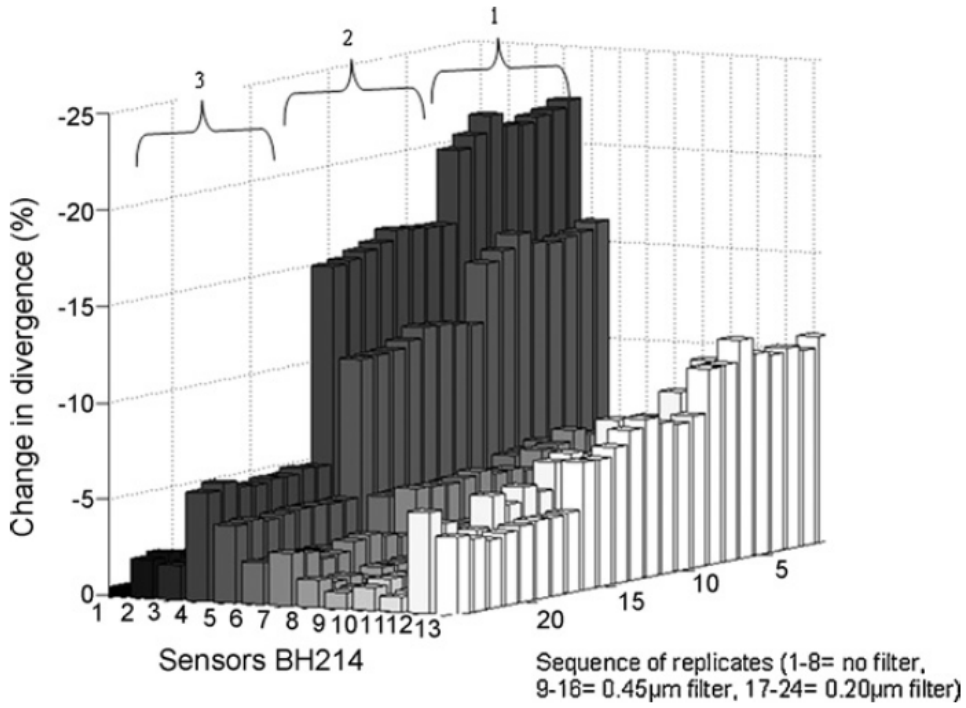

1
2
3
-25
-20
-15
-10
-5
0
Change in divergence (%)
1 2 3 4 5 6 7 8 9 10 11 12 13
Sensors BH214
20
15
10
5
Sequence of replicates (1-8= no filter,
9-16= 0.45µm filter, 17-24= 0.20µm filter)

Figure 4.

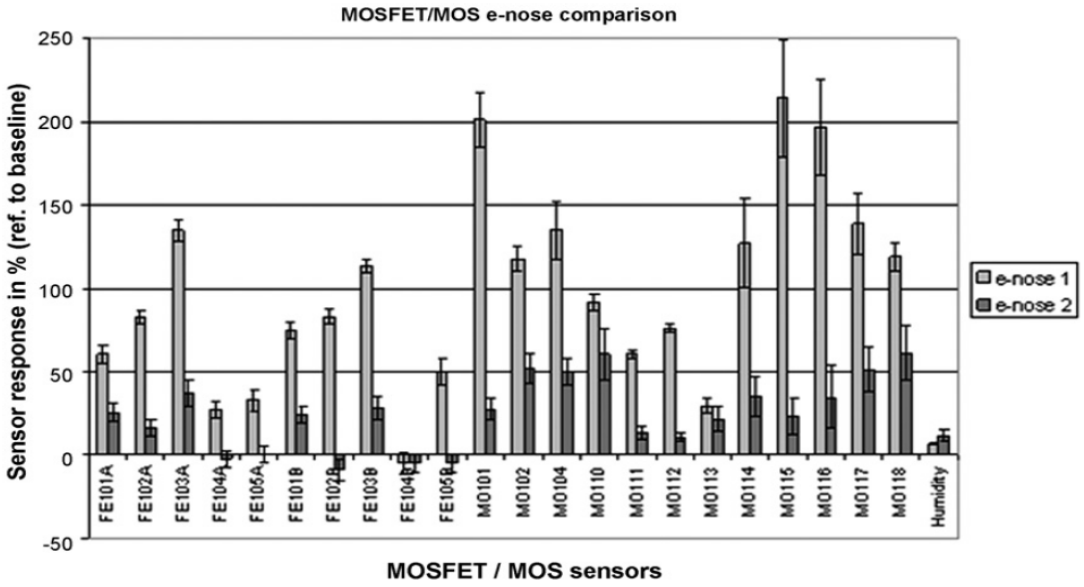

MOSFET/MOS e-nose comparison
Sensor response in % (ref. to baseline)
250
200
150
100
50
0
-50
FE101A
FE102A
FE103A
FE104A
FE105A
FE101B
FE102B
FE103B
FE104B
FE105B
MO101
MO102
MO104
MO110
MO111
MO112
MO113
MO114
MO115
MO116
MO117
MO118
Humidity
MOSFET / MOS sensors
e-nose 1
e-nose 2

Figure 5.

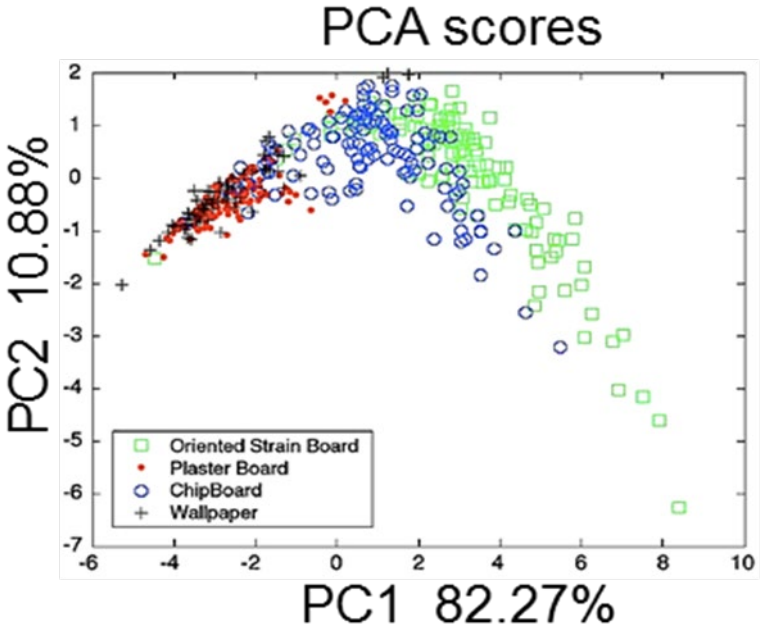

PCA scores
PC2 10.88%
PC1 82.27%
Oriented Strain Board
Plaster Board
ChipBoard
Wallpaper
2
1
0
-1
-2
-3
-4
-5
-6
-7
-6
-4
-2
0
2
4
6
8
10

Figure 6

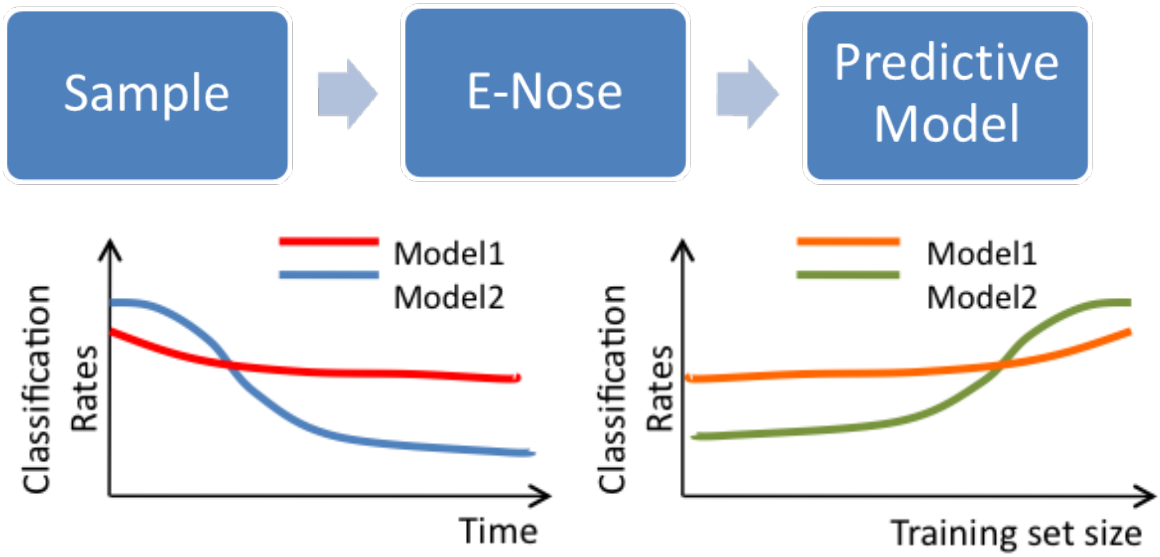

Sample
E-Nose
Predictive Model
Classification Rates
Model1
Model2
Time
Classification Rates
Model1
Model2
Training set size

Figure 7

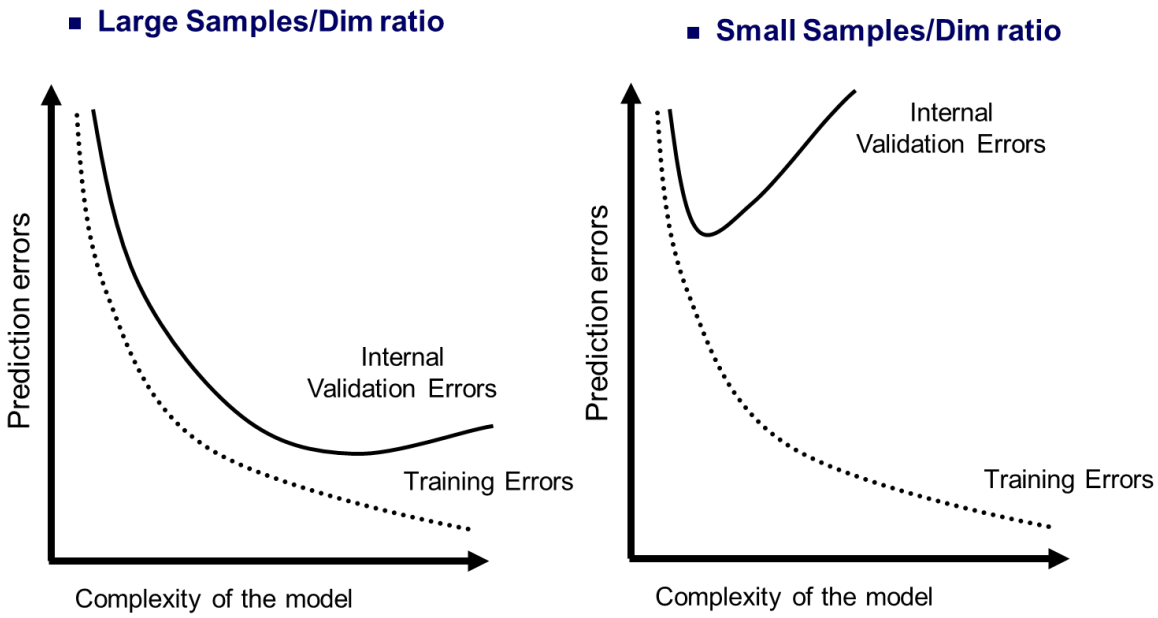

Large Samples/Dim ratio
Prediction errors
Internal
Validation Errors
Training Errors
Complexity of the model
Small Samples/Dim ratio
Prediction errors
Internal
Validation Errors
Training Errors
Complexity of the model

Figure 8

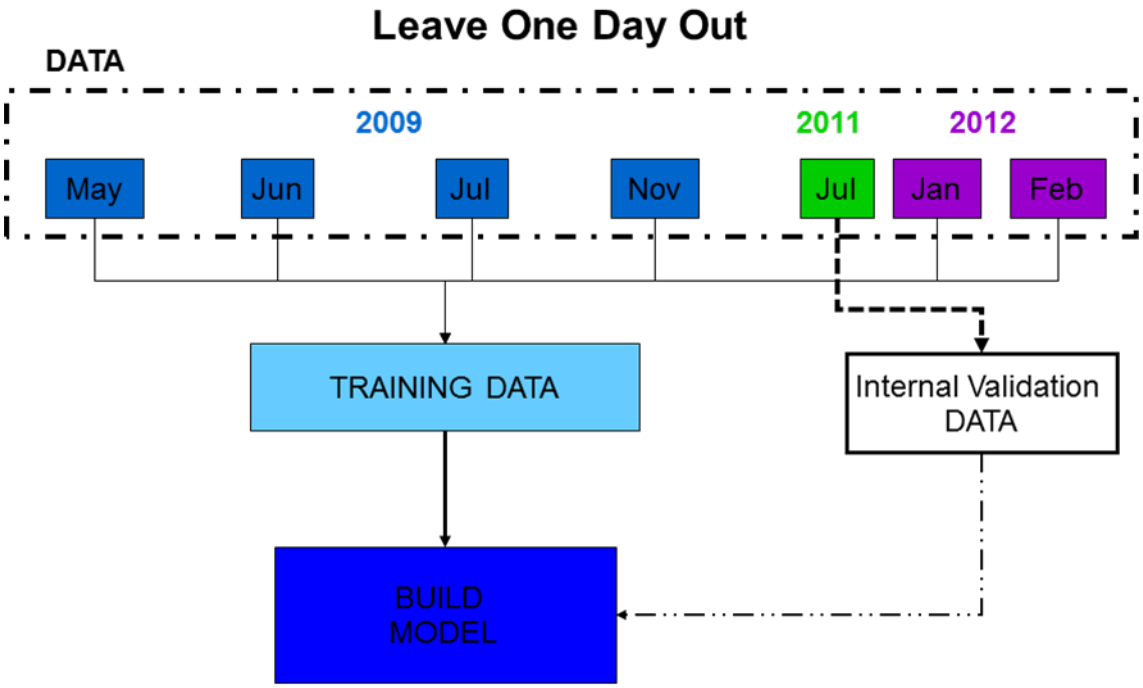

Leave One Day Out
DATA
2009
2011
2012
May
Jun
Jul
Nov
Jul
Jan
Feb
TRAINING DATA
Internal Validation DATA
BUILD MODEL

Figure 9.

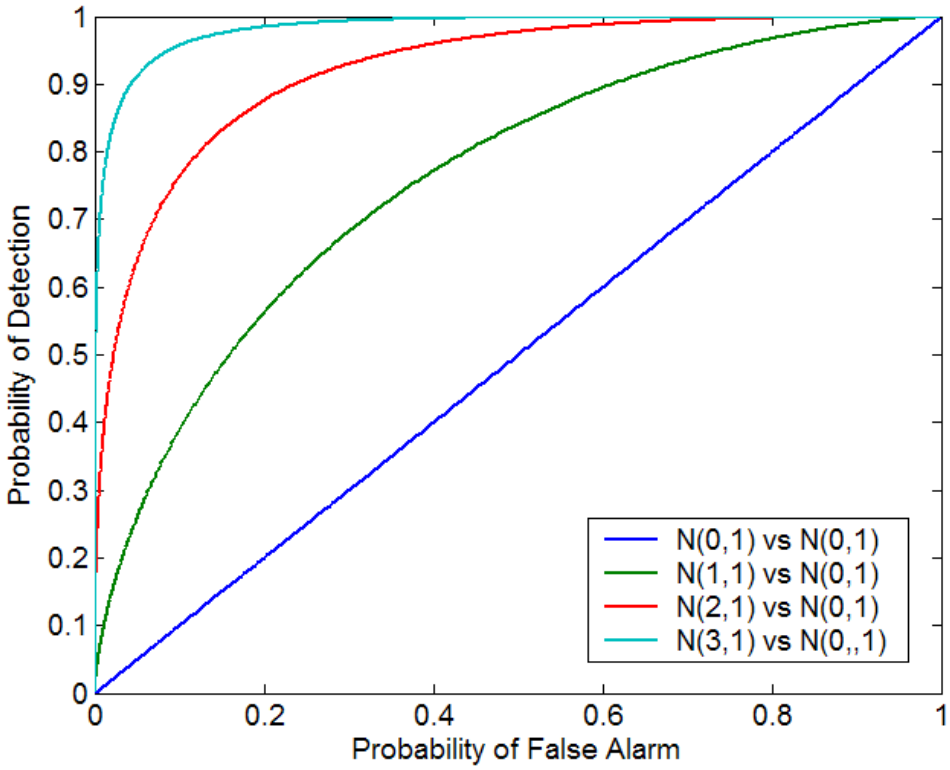

Probability of Detection
Probability of False Alarm
0
0.1
0.2
0.3
0.4
0.5
0.6
0.7
0.8
0.9
1
0
0.2
0.4
0.6
0.8
1
N(0,1) vs N(0,1)
N(1,1) vs N(0,1)
N(2,1) vs N(0,1)
N(3,1) vs N(0,,1)

Figure 10.

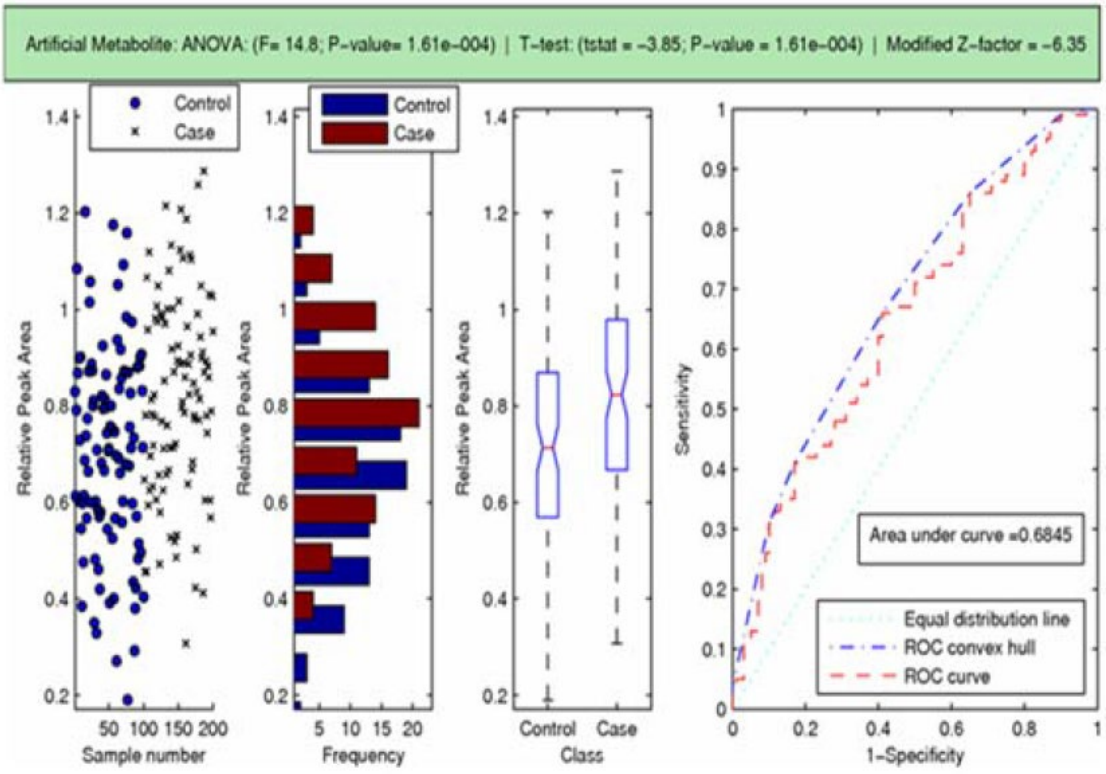
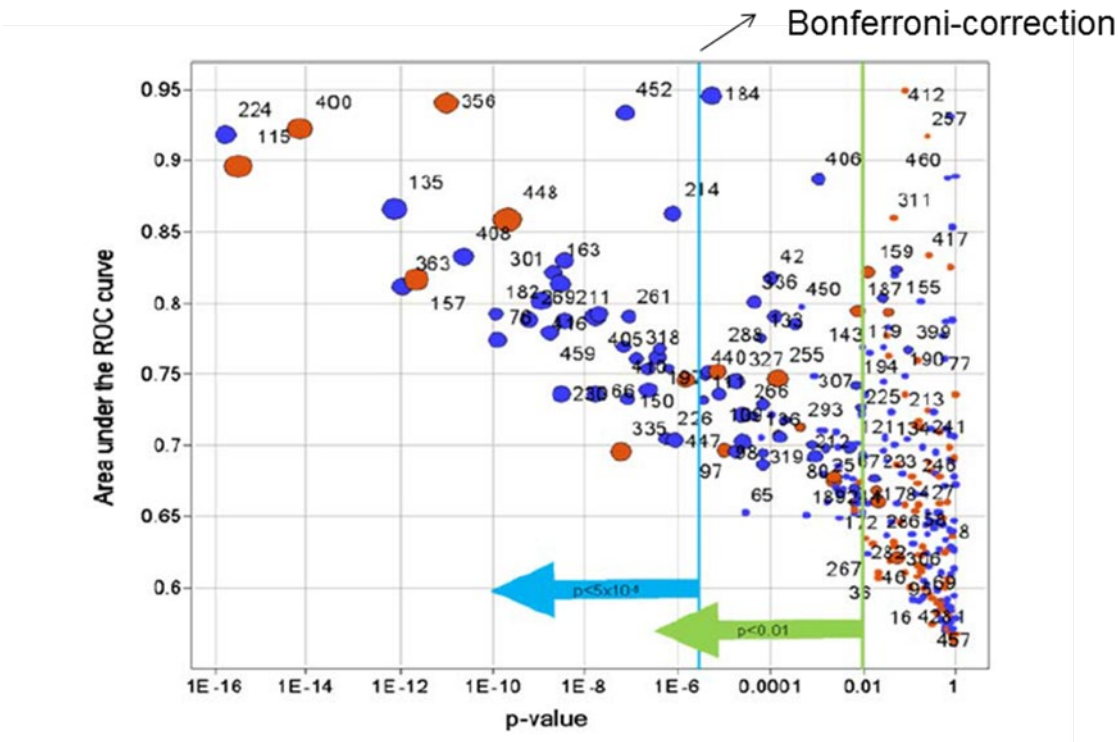

Artificial Metabolite: ANOVA: (F= 14.8; P-value= 1.61e-004) | T-test: (tstat = -3.85; P-value = 1.61e-004) | Modified Z-factor = -6.35
Control
Case
Relative Peak Area
Sample number
Frequency
Class
Sensitivity
1-Specificity
Area under curve =0.6845
Equal distribution line
ROC convex hull
ROC curve

Figure 11


Bonferroni-correction
Area under the ROC curve
p-value
0.95
0.9
0.85
0.8
0.75
0.7
0.65
0.6
1E-16
1E-14
1E-12
1E-10
1E-8
1E-6
0.0001
0.01
1
p<5x10-4
p<0.01