Direct Identification of Bacteria in Blood Culture Samples using an Electronic Nose

Marco Trincavelli, *Member, IEEE,* Silvia Coradeschi, Amy Loutfi, Bo Söderquist, and Per Thunberg

***Abstract*—In this work we introduce a method for identification of bacteria in human blood culture samples using an electronic nose. The method uses features which capture the static (steady-state) and dynamic (transient) properties of the signal from the gas sensor array and proposes a means to ensemble results from consecutive samples. The underlying mechanism for ensembling is based on an estimation of posterior probability which is extracted from a support vector machine classifier. A large data set representing 10 different bacteria cultures has been used to validate the presented methods. The results detail the performance of the proposed algorithm and show that through ensembling decisions on consecutive samples significant reliability in classification accuracy can be achieved.**

***Index Terms*—Electronic nose, bacteria identification, sepsis.**

I. Introduction

SEPSIS also known as blood poisoning of septicaemia is caused by the presence of micro-organisms, predominantly bacteria in ciruculating blood. Rapid administration of efficient antibiotic treatment is crucial as sepsis can result in septic shock, multiple organ dysfunction and even death. The current standard procedure for diagnosis involves routine microbiological blood cultures. Such procedures can take at least 36 hours to several days before diagnosis can be made. Typically, automated blood culture monitoring systems are first used to incubate a blood culture and monitor the production or reduction of gases (this is normally done via non intrusive methods for detection of CO_2). Once a sample indicates a change in gas tension, a lab technician will culture the sample on plates for further identification. This secondary sub-culturing may require up to 36 hours before a final identification can be made. Thus finding diagnostic methods that provide fast identification of the presence and the type of bacteria thereby allowing proper antibiotic treatment can provide an increased benefit to the patient as well as help to reduce cost to the health care system.

In the past decade, compact gas sensors have been integrated into an instrument called an electronic nose which can provide fast (order of minutes) and non-intrusive measurement of gaseous agents. Within the medical field, electronic noses have been applied to the identification of bacterial agents in different contexts ranging from upper respiratory diseases [1], urine samples [2], and ENT bacteria [3]. The challenge of developing an electronic noses for bacteria identification in blood cultures depends on a number of factors. Firstly, the way in which headspace sampling system introduces the samples to the sensor array have been shown in comparative studies to have an effect on identification rates, where samplers that control temperature and humidity provide better performance results [4]. Secondly, the blood in which the bacteria is sampled may differ from person to person and thus identification should be independent from the medium itself. Thirdly, the post-processing of the signals should be suited for the characteristics of the signals and application. In applications with a larger number of sensors, it is important to extract the relevant features from the signals that can be used for further processing. Although, trials have been made with a single sensor type for identification of bacteria in blood cultures [5], today's off-the-shelf electronic nose devices contain anywhere from 4 to 32 sensors in an array. This allows for a wider range of odours to be detected.

In this work, a new approach for classifying bacteria with an electronic nose is presented. The method evaluates the suitability of a given sample for classification by representing the output from a support vector machine (SVM) with a posterior probability estimation. This estimation is ensembled across ten consecutive reponses of the same sample in order to make the classification more reliable. An electronic nose containing 22 gas sensors with partial and overlapping selectivities and an automatic headspace sampler is used to regulate the samples. The data processing methods presented consist of extracting features that reside on the static (steady-state) and dynamic (transient) properties of the signal. These features are fed into a support vector machine and to the ensembling algorithm. The mechanism of ensembling is based on treating the posterior probabilities as a random sample and estimating the 95% confidence interval for the mean of the posterior of each class. A mean with significant superior confidence interval for a class is disjoint and above all the others then classification is performed (assigning the sequence of samples to that class); otherwise a rejection is declared. Identifications are made among 10 typical bacteria cultures leading to sepsis by directly sampling after the first stage incubation indicates a positive culture growth.

II. Experimental Method

A. Sample Preparation

Bacteria isolates obtained from clinical samples were used. The bacteria were subcultured on blood agar plates and a bacterial suspension solution was adjusted at turbidity of standard MacFarland 1.0 ($3x10^8$ cfu/mL) in NaCl Solution.

M. Trincavelli, S. Coradeschi and A. Loutfi are with the Center for Applied Autonomous Sensor Systems, School of Science and Technology, Örebro University, Örebro, SE - 70182, Sweden e-mail: name.surname@oru.se. B. Söderquist is with Örebro University Hospital and P. Thunberg is with the department of Medical Physics at Örebro University Hospital, Örebro, Sweden email:name.surname@orebroll.se

TABLE I
BACTERIA SPECIES CODE AND TAXONOMY

Species Code	Taxonomy
ECOLI	E.coli
PSAER	Pseudomonas aeruginosa
STA	Staphylococcus aureus
KLOXY	Klebsiella oxytoca
PRMIR	Proteus mirabilis
SRFCL	Enterococcus faecalis
STLUG	Staphylococcus lugdunensis
PASMU	Pasteurella multocida
HSA	Steptococcus pyogenes
HINFL	Hemophilus influenzae

After which, 2 ml of the bacteria suspension were added to an aerobic blood culture bottle (BD BACTEC^{TM} Plus + Aerobic/F) together with 5 ml of pooled human blood. Blood cultured bottles were incubated at 35°C for 24 hours. The bacterial isolates used along with the respective species codes are given in Table I. It is important to note that the viable count when the bottle is positive has not been established and therefore is not known at the time of sampling. In this work only one bacterial strain and one bacteria species is considered at the time as polyclonal or polymicrobial bacteremia is rare and multi-bacteria strains arising from e.g. wounds is beyond the scope of this study.

B. Sampling Procedure

The purpose of the sampling system is to transfer the headspace from the sample to the sensors without altering its composition and properties. The sampling system of the NST 3220 Emission Analyzer, Applied Sensors, Linköping, Sweden uses two adjacent needles, one ’in’ needle and one ’out’ needle. The sample gas drawn through the ’in’ needle is passively replaced with air through the ’out’ needle by the negative pressure created in the sample bottle as the headspace is removed. This ensure a minimal dilution factor, and therefore a minimal alteration of the properties, in particular when compared to systems that use a carrier gas.

The sensor array of the NST Emission analyzer is composed of 10 MOS and 12 MOSFET sensors, for a total of 22 sensors. These two technologies complement each other enabling the array to discriminate between a much wider range of gases than by using just one sensor technology [6].

The sampling cycle, as in most e-nose based systems, is composed by three phases: baseline acquisition, odour sampling and recovery to initial state. In the baseline acquisition phase the sensor array is exposed to a reference gas (air in this case) for 10 seconds and the value of the sensors is recorded. During the odour sampling phases the headspace in the analysis bottle is injected into the sensor chamber for 30 seconds. After this, the sensors are exposed again to the reference gas for 260 seconds in order for the sensors to recover the value they had during the baseline acquisition phase. The total length of the sampling cycle is five minutes.

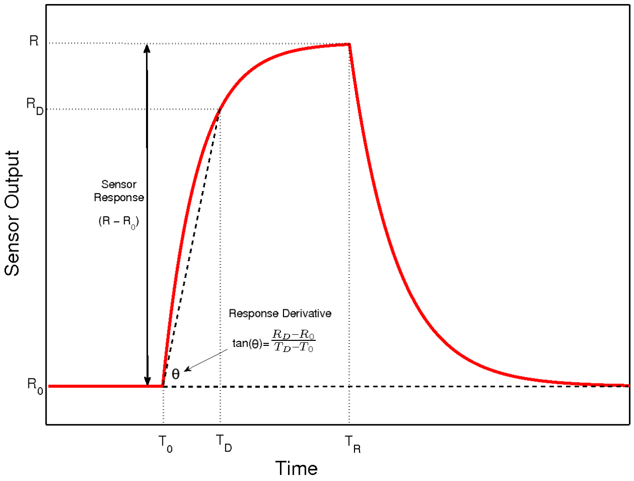

Fig. 1. Graphical interpretation of the two feature extraction methods used in this work.

The sampling cycle is repeated ten times in a row and we refer to a series of ten consecutive sampling cycles as a measurement. A measurement sequence is composed by one measurement for every type of bacteria. The whole data set is composed by 12 measurement sequences, 6 done with a first batch of bacteria cultures and six done with a second batch one week later. Blood samples within a batch came from the same source and different sources were used between batches.

III. PATTERN RECOGNITION ALGORITHM

The pattern recognition algorithm is articulated into five phases, namely feature extraction, dimensionality reduction, classification, posterior probability estimation and ensembling decisions. The next subsections describe each of the phases of the algorithm.

A. Feature Extraction

In several works dealing with feature extraction from gas sensor response [7], [8], it has been demonstrated that features based on the dynamic of the signal can add information for the discrimination task. Therefore two different feature extraction methods are considered in this work:

- **Static Response**: The static response of a sensor is defined as the difference between the value that the sensor has at the end of the sampling phase minus the baseline value for that sensor.
- **Response Derivative**: The response derivative is calculated as the average of the derivative of the sensor during the first 3 seconds of exposition of the array to the headspace.

Figure 1 gives a graphical interpretation of these two features. The extraction of these features generates a 22 dimensions feature vector in case only the static response is considered or 44 dimensions vector if both the static response and the response derivative are considered.

B. Dimensionality Reduction

When the dimensionality of the considered feature space is high and a finite number of samples is available, a good estimation of the discriminant function is difficult to obtain. However, in many practical applications multivariate data in $\Re^n$ usually have an intrinsic dimensionality much lower than n. This is especially true for gas sensing applications, where the gas sensors in an array have a highly correlated response. As a result all the samples can be enclosed in a hyperspace of dimensionality much lower than the one of the original feature space. There are fundamentally two approaches for reducing the dimensionality of a feature space: selecting a subset of the original feature set or projecting the original feature space into a lower dimensional one. In this work we considered a method belonging to the second family, the Linear Discriminant Analysis (LDA). The LDA finds a projection of the data on a lower dimensional space that maximizes a class separability criterium based on the concepts of inter-class and intra-class scatter. In this work we have 10 classes and therefore the original feature space is projected into a 9 dimensional space.

An alternative approach is to consider feature selection in order to be able to optimize the sensor array. This will be considered in future works where the development of a tailored made instrument will be addressed.

C. Classification

The classification algorithm that has been considered in this work is the Support Vector Machine (SVM) [9]. The SVM is a popular kernel based algorithm that projects the data into a high dimensional space in which the problem is solved using a maximum margin linear classifier. The linear decision boundaries in the high dimensional feature space are in general non linear decision boundaries in the original feature space. One of the most important properties of support vector machines is that the estimation of the model parameters is a convex optimization problem and therefore any local solution is also a global optimum. Many variations of the original model of SVM have been proposed, both for classification and regression problems. The model used in this work is the soft margin SVM with Gaussian kernel. This particular model has two hyperparameters that need to be set in advance: γ that is the size of the Gaussian kernel and the regularization parameter C that determines how much samples that fall on the wrong side of the decision boundary have to be penalized.

The SVM is by definition a binary classifier, though it is possible to extend it to the multiclass case using different approaches. The most popular are the *one-versus-the-rest* and the *one-versus-one* approach [10]. Given the number of classes K, in the *one-versus-the-rest* approach one SVM for every class C_k is trained in which samples that belong to C_k are considered the positive class while samples belonging to the other $K-1$ classes are considered the negative class. In the *one-versus-one* approach $K(K-1)/2$ binary SVMs are trained for all the possible pairs of classes. Following the indications presented in [11] we chose to use the *one-versus-one* approach.

D. Posterior Probability Estimation

The SVM is a decision machine and so it does not provide any estimation of the posterior probability $P(C|x)$ of sample x belonging to class C. Though, many methods have been proposed in literature for obtaining such estimation [12], the one considered in this work consists in fitting a sigmoid to every pairwise decision in order to get an estimate of the pairwise coupled posterior probability and then ensembling the pairwise coupled posteriors in order to get a multiclass posterior probability. Given a 2-class dataset where the two classes are encoded by the values $y = \pm 1$ and f is the (unthresholded) output of the SVM, a parametric form of a sigmoid that estimated the posterior probability for class 1 is:

$$r_{+1-1} = P(y = +1|f) = \frac{1}{1 + e^{Af+B}} \tag{1}$$

The parameters A and B are calculated by minimizing the negative log likelihood of the data:

$$\underset{A,B}{\operatorname{argmin}}(-\sum_i t_i log(P_i) + (1 - t_i) log(1 - P_i)) \tag{2}$$

$$\text{where} \qquad t_i = \frac{y_i + 1}{2} \tag{3}$$

More details about how the optimization problem is solved can be found in [13]. It is important to notice that, as Platt pointed out in [14], the SVM decision values f are usually clustered around ± 1 and therefore the probability estimation given by Equation (1) might be inaccurate due to overfitting. In order to limit this overfitting the parameters A and B are estimated performing a further 5-fold cross validation on the data set.

Once all the set of pairwise posteriors have been obtained they are ensembled according to the second approach proposed in [12]. This method is formulated as an optimization problem as follows:

$$\begin{aligned} \underset{\mathbf{p}}{\text{minimize}} \qquad & \sum_{i=1}^{k} \sum_{j:j \neq i} (r_{ji} p_i - r_{ij} p_j)^2 && (4) \\ \text{subject to} \qquad & \sum_{i=1}^{k} p_i = 1, \forall i \\ & p_i \geq 0, \forall i \\ & && (5) \end{aligned}$$

where r_{ab} is the pairwise posterior probability calculated with (1) for the two-class case, p_a is the multiclass posterior probability for class a and $\mathbf{p}$ is the vector of all the p_a. It is shown in [12] how this problem has a unique solution and can be solved by a simple linear system.

E. Ensembling Decisions

As described in Section II-B during a measurement the sampling cycle is repeated 10 times. The pattern recognition algorithm is applied independently to the 10 samples and therefore 10 estimates $P(C|x_i)$ are obtained, where i is the number of the sample. A step in order to decrease the uncertainty in the identification is to try to ensemble the information coming from the 10 sampling cycles instead of considering them separately. A way to do this is to consider

TABLE II
CONFUSION MATRIX OBTAINED FOR THE WHOLE DATA SET BY THE ALGORITHM THAT RELIES ONLY ON THE RESPONSE OF THE SENSORS (ROWS ARE THE TRUE CLASS, COLUMNS ARE ESTIMATED CLASS).

	1	2	3	4	5	6	7	8	9	10
1	113	5	0	0	0	2	0	0	0	0
2	7	102	2	0	0	0	0	0	0	9
3	0	2	118	0	0	0	0	0	0	0
4	0	0	1	107	1	11	0	0	0	0
5	0	0	0	5	106	9	0	0	0	0
6	0	0	0	16	0	102	0	1	1	0
7	0	0	0	0	0	0	120	0	0	0
8	0	0	0	0	0	0	0	120	0	0
9	6	0	0	0	0	0	3	0	111	0
10	2	0	15	0	0	0	0	0	0	103

TABLE III
CONFUSION MATRIX OBTAINED FOR THE WHOLE DATA SET BY THE ALGORITHM THAT RELIES BOTH ON THE RESPONSE AND ON THE DERIVATIVE OF THE SENSORS (ROWS ARE THE TRUE CLASS, COLUMNS ARE ESTIMATED CLASS).

	1	2	3	4	5	6	7	8	9	10
1	104	5	0	0	0	1	0	1	9	0
2	9	111	0	0	0	0	0	0	0	0
3	0	2	117	0	0	0	0	0	1	0
4	0	0	0	108	0	12	0	0	0	0
5	0	0	1	8	102	8	0	1	0	0
6	0	0	0	5	0	111	0	1	3	0
7	0	0	0	0	0	0	120	0	0	0
8	0	0	0	0	2	0	0	118	0	0
9	2	0	0	0	0	0	0	0	118	0
10	0	2	5	0	0	0	0	0	0	113

the 10 estimates of $P(C|x_i)$ as a random sample and perform a multiple comparison among the means of $P(C_k|\mathbf{x})$ for all the classes k. In this work the multiple comparison is performed using the Tukey's Honestly Significant Differences Test (HSD) [15]. It is important to notice that a multiple comparison procedure is not equivalent to a series of pairwise comparison since it is designed to provide an upper bound (the chosen confidence level α) on the probability that any comparison will be incorrectly found significant. Instead performing multiple pairwise test the confidence level α would apply to each single comparison, so the chance of incorrectly finding a significant difference would increase with the number of comparisons.

In this work we use the HSD with a confidence level $\alpha = 0.05$ and we perform a classification only if, according to the HSD result, the mean of the posterior probability for one class k is significantly superior of the means of all the other classes. If this does not happen we declare a rejection for that measurement.

IV. Results

A 12-fold cross validation on the collected data set has been performed to evaluate the proposed pattern recognition algorithm. In every fold, one sequence of measurements have been left out and used for testing the algorithm trained with the remaining eleven sequences. The hyperparameters of the SVM have been estimated with an exhaustive grid search in the interval $[-10, 10]$ with step 1 in the base 2 logarithmic scale for both C and γ. An 8-fold cross validation, where the folds have been extracted randomly from the training set, has been performed at every point in the grid. The confusion matrices obtained using only the features based on the response of the sensor and using the features based both on the response and the derivative are displayed respectively in Table II and III. The classification accuracy obtained in the two cases is $91.8\% \pm 11.5\%$ and $94\% \pm 12.7\%$ respectively. Before ensembling the posterior probabilities calculated for the sampling cycles, an analysis of reliability of the estimate of the posterior probabilities as a confidence measure for the classification algorithm is made as well as an analysis of the distribution of errors with respect to the measuring cycles and measuring sequences. In order to check the validity of the posteriors as a confidence measure a threshold is introduced so that, if not exceeded by the maximum of the posteriors, a rejection is declared. Figure 2 shows the results of varying this threshold in the range (0,1) for both the considered feature sets (response and response + derivative). The fact that the error rate decreases when the rejection threshold increases indicates that the estimation of the posterior is a reliable confidence measure for the classification algorithm. Moreover it can also be observed how the addition of the derivative based features diminish the error of roughly the 25% over all the rejection threshold spectrum without increasing the rejection rate. This confirms that the dynamic characteristics of the signal contains useful information for the discrimination of odours.

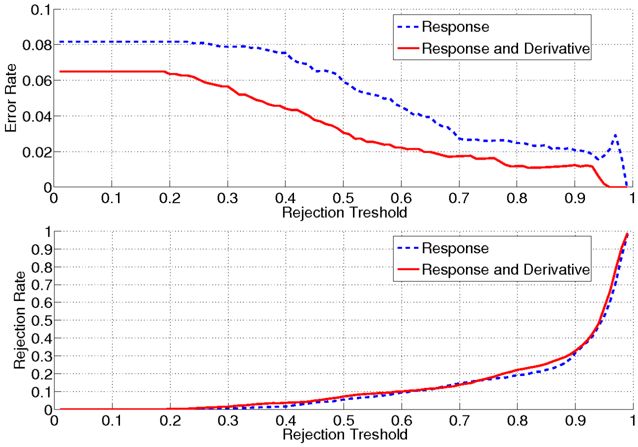

Fig. 2. Performance of the classification algorithm with a varying rejection threshold. The upper figure shows the error rate and the lower figure the rejection rate. The dashed lines represents the performance obtained by the algorithm that uses only the response based features, while the solid lines represent the performance obtained by the algorithm that uses both the response and derivative based features.

A second aspect to analyze is how errors are spread across measurement sessions and sampling cycles. Table IV shows the performances obtained in the twelve measurement sessions. It is evident how measurement sessions 1 and 7 obtain a performance much worse than the other sessions.

This can be explained by the fact that these two sessions are the ones recorded in the beginning of the two experiment batches. Therefore, we can suppose that this degradation of performance can be due to interference in the measuring system, like humidity deposited on the sensors surface, the sensors were not fully warmed or stagnant air was present in the sampling system. An analysis of how the errors are spread across measurement cycles is given in Figure 4. It can be observed how the number of errors made during the first measuring cycle is larger than the errors in the other cycles. This can be due to the fact that the purging procedure of the nose at the end of a measuring cycle is not perfect and therefore some leftover from the previous sampling cycle is still there. This effect is particularly evident in the first measuring cycle since the bacteria that was smelled in the cycle before was different. Indeed it is possible to use a sample which will better "condition" the sensors to the bacteria infected blood, which is a common practice in the electronic nose community. The effect of using a conditionner may result in a better classification performance for the first sample and this will be investigated in future works.

If the data collected during measurement sessions 1 and 7 are removed from the set we obtain $96.4\%\pm4.3\%$ classification accuracy with the feature based on the static response and $98.9\% \pm 1.4\%$ with the features that include the dynamic of the sensors. Figure 3 shows the effect of the introduction of a threshold on the posterior probability on this reduced data set. The observations made for the full data set are confirmed and in this case the improvement obtained with the introduction of the features based on the derivative of the signal is even more significant. Figure 5 shows that most of the errors are done in the first measuring cycle also when measurement session 1 and 7 are left out.

Results from ensembling the decisions for the full data set and for the data set without sequence 1 and 7 are shown in Figure 6 and 7 respectively. It is important to notice how neglecting the first cycle improves the performance of the ensemble. This confirms that the first cycle contains additional noise with respect to the subsequent cycles. In particular in Figure 7, the performance of the ensemble for the data set without measurement session 1 and 7 is displayed. After only 4 sampling cycles perfect discrimination is obtained as both the error and rejection rate are zero.

V. Conclusion

This paper advocates the use of electronic noses to discriminate among various bacteria regularly found in the blood cultures. This is an important application of electronic olfaction that could significantly improve the current methodologies and be successfully used in clinical settings. The results presented show that the bacteria can be accurately discriminated using the method. Further the proposed methods have been tested on a large data set, (an order of magnitude larger than previous studies). Our next step will be the starting of clinical trials to test the robustness of the method and its applicability in a clinical setting. In particular we will examine the effect of the geneology of the bacteria (i.e. different strains of the same species on discrimination performance).

TABLE IV
Classification accuracy for the twelve measurement sessions.

# Session	Response Features	Response and Derivative Features
1	73%	69%
2	91%	99%
3	93%	98%
4	100%	100%
5	100%	100%
6	100%	100%
7	65%	64%
8	88%	97%
9	97%	100%
10	100%	100%
11	98%	99%
12	97%	96%

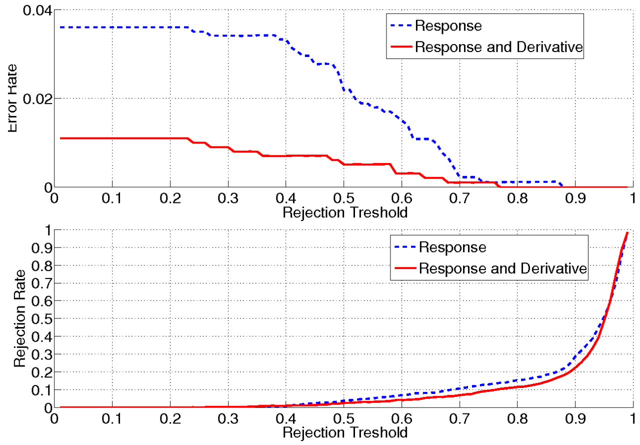

Fig. 3. Performance of the classification algorithm with a varying rejection threshold for the data set where sequence 1 and 7 have been removed. The upper figure shows the error rate and the lower figure the rejection rate. The dashed lines represents the performance obtained by the algorithm that uses only the response based features, while the solid lines represent the performance obtained by the algorithm that uses both the response and derivative based features.

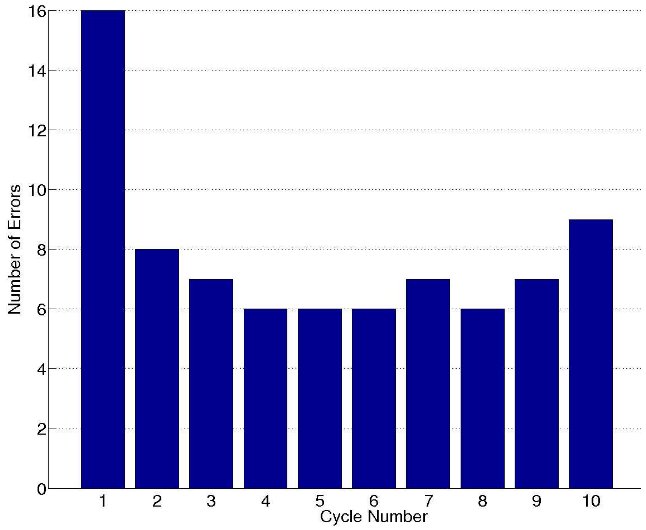

Fig. 4. Number of errors committed by the classification algorithm in the different measuring cycles. It is evident how the first measurement cycle is more subject to erroneous decisions.

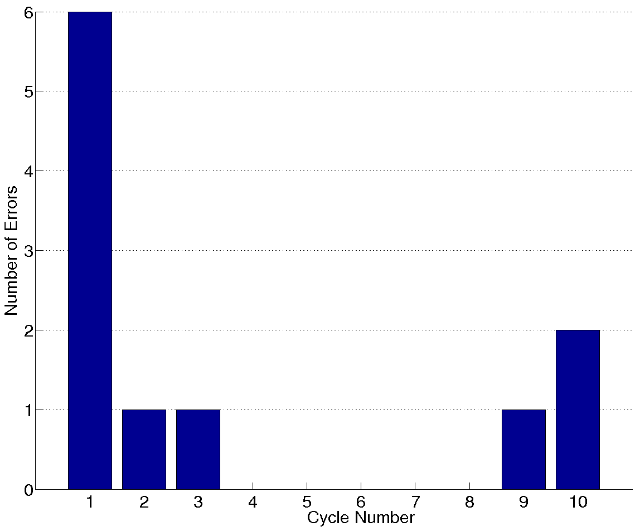

Fig. 5. Number of errors committed by the classification algorithm in the different measuring cycles for the data set where sequence 1 and 7 have been removed. It is evident how the first measurement cycle is more subjected to erroneous decisions.

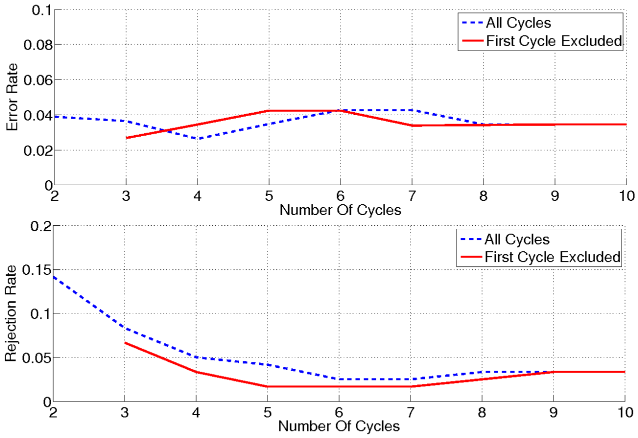

Fig. 6. Performance of the ensembled classification algorithm with a varying number of measuring cycles. The upper figure shows the error rate and the lower figure the rejection rate. The solid lines represents the performance obtained by the algorithm that uses all the measuring cycles, while the dashed lines represent the performance obtained by the algorithm that neglects the first measurement cycle. Notice that the solid line starts from cycle two and the dashed line from cycle three. This is because at least two samples are needed to calculate an uncertainty.

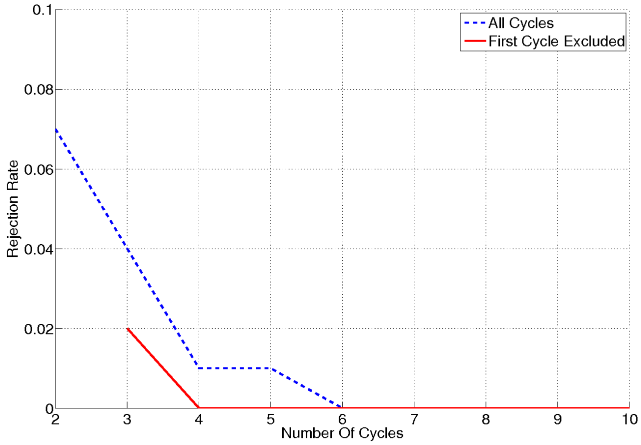

Fig. 7. Performance of the ensembled classification algorithm with a varying number of measuring cycles for the data set where sequence 1 and 7 have been removed. Only the rejection rate is shown since the error is constantly zero. The solid lines represents the performance obtained by the algorithm that uses all the measuring cycles, while the dashed lines represent the performance obtained by the algorithm that neglects the first measurement cycle. Notice that the solid line starts from cycle two and the dashed line from cycle three. This is because at least two samples are needed to calculate an uncertainty.

Acknowledgment

This work has been supported by EU structural funds Novamedtech. The author would explicitly like to thank Åsa Öström at Örebro University for providing access to the instrumentation used in this work. The authors would also like to thank Lena Barkman at Örebro University Hospital for her assistance in the sample preparation and her support.

References

[1] S. Y. Lai, O. F. Deffenderfer, W. Hanson, M. P. Phillips, and E. R. Thaler, "Identification of upper respiratory bacterial pathogens with the electronic nose," *Laryngoscope*, vol. 112, pp. 975–979, Jun 2002.

[2] S. Aathithan, J. C. Plant, A. N. Chaudry, and G. L. French, "Diagnosis of bacteriuria by detection of volatile organic compounds in urine using an automated headspace analyzer with multiple conducting polymer sensors," *J. Clin. Microbiol.*, vol. 39, pp. 2590–2593, Jul 2001.

[3] M. Holmberg, F. Gustafsson, E. G. Hörnsten, F. Winquist, L. E. Nilsson, L. Ljung, and I. Lundström, "Feature extraction from sensor data on bacterial growth," *Biotechnology Techniques*, vol. 12, no. 4, pp. 319–324, 2004.

[4] G. C. Green, A. D. Chan, and R. A. Goubran, "An investigation into the suitability of using three electronic nose instruments for the detection and discrimination of bacteria types," *Conf Proc IEEE Eng Med Biol Soc*, vol. 1, pp. 1850–1853, 2006.

[5] M. Bruins, A. Bos, P. L. Petit, K. Eadie, A. Rog, R. Bos, G. H. van Ramshorst, and A. van Belkum, "Device-independent, real-time identification of bacterial pathogens with a metal oxide-based olfactory sensor," *Eur. J. Clin. Microbiol. Infect. Dis.*, vol. 28, pp. 775–780, Jul 2009.

[6] K. Arshak, E. Moore, G. Lyons, J. Harris, and S. Clifford, "A review of gas sensors employed in electronic nose applications," *Sensor Review*, vol. 24, no. 2, pp. 181–198, October 2004.

[7] T. Eklöv, P. Martensson, and I. Lundström, "Enhanced selectivity of mosfet gas sensors by systematical analysis of transient parameters," *Analytica Chimica Acta*, vol. 353, no. 2-3, pp. 291 – 300, 1997.

[8] R. Gutierrez-Osuna, H. T. Nagle, and S. S. Schiffman, "Transient response analysis of an electronic nose using multi-exponential models," *Sensors and Actuators B: Chemical*, vol. 61, no. 1-3, pp. 170 – 182, 1999.

[9] C. J. C. Burges, "A tutorial on support vector machines for pattern recognition," *Data Mining and Knowledge Discovery*, vol. 2, no. 2, pp. 121–167, 1998.

[10] C. M. Bishop, *Pattern Recognition and Machine Learning (Information Science and Statistics)*. Springer, August 2006.

[11] C.-W. Hsu and C.-J. Lin, "A comparison of methods for multi-class support vector machines," *IEEE Transactions on Neural Networks*, vol. 13, no. 2, pp. 415–425, 2002.

[12] T.-F. Wu, C.-J. Lin, and R. C. Weng, "Probability estimates for multi-class classification by pairwise coupling," *Journal of Machine Learning Research*, vol. 5, pp. 975–1005, 2004.

[13] H. T. Lin, C. J. Lin, and R. C. Weng, "A note on platt's probabilistic outputs for support vector machines," *Mach. Learn.*, vol. 68, no. 3, pp. 267–276, October 2007.

[14] J. C. Platt, "Probabilistic outputs for support vector machines and comparisons to regularized likelihood methods," in *Advances in Large Margin Classifiers*, 1999, pp. 61–74.

[15] Y. Hochberg and A. C. Tamhane, *Multiple Comparison Procedures (Wiley Series in Probability and Statistics)*. Wiley, November 1987.