Fuzzy *k*-NN Lung Cancer Identification by an Electronic Nose

Rossella Blatt[1], Andrea Bonarini[1], Elisa Calabró[2], Matteo Della Torre[3], Matteo Matteucci[1], and Ugo Pastorino[2]

[1] Politecnico di Milano, Department of Electronics and Information, Milan, Italy
blatt@elet.polimi.it, bonarini@elet.polimi.it, matteucci@elet.polimi.it
[2] Istituto Nazionale Tumori of Milan, Toracic Surgery Department, Milan, Italy
ugo.pastorino@istitutotumori.mi.it, elisa.calabro@istitutotumori.mi.it
[3] Sacmi Imola S.C., Automation & Inspection Systems, Imola (BO), Italy
matteo.della.torre@sacmi.it

Abstract. We present a method to recognize the presence of lung cancer in individuals by classifying the olfactory signal acquired through an electronic nose based on an array of MOS sensors. We analyzed the breath of 101 persons, of which 58 as control and 43 suffering from different types of lung cancer (primary and not) at different stages. In order to find the components able to discriminate between the two classes 'healthy' and 'sick' as best as possible and to reduce the dimensionality of the problem, we extracted the most significative features and projected them into a lower dimensional space, using Nonparametric Linear Discriminant Analysis. Finally, we used these features as input to a pattern classification algorithm, based on Fuzzy *k*-Nearest Neighbors (Fuzzy *k*-NN). The observed results, all validated using cross-validation, have been satisfactory achieving an accuracy of 92.6%, a sensitivity of 95.3% and a specificity of 90.5%. These results put the electronic nose as a valid implementation of lung cancer diagnostic technique, being able to obtain excellent results with a non invasive, small, low cost and very fast instrument.

Keywords: Electronic Nose, E-Nose, Olfactory Signal, Pattern Classification, Fuzzy *k*-NN, MOS Sensor Array, Lung Cancer.

1 Motivation and Methodology

It has been demonstrated that the presence of lung cancer alters the percentage of some volatile organic compounds (VOCs) present in the human breath [7], which may be considered as markers of this disease. This substances can be detected by an electronic nose, that is an instrument that allows to acquire the olfactory signal. The electronic nose includes an array of electronic chemical sensors with partial specificity and an appropriate pattern recognition system able to recognize simple or complex odors [1].

The main objective of this paper is to demonstrate that it is possible to recognize individuals affected by lung cancer, analyzing the olfactory signal of

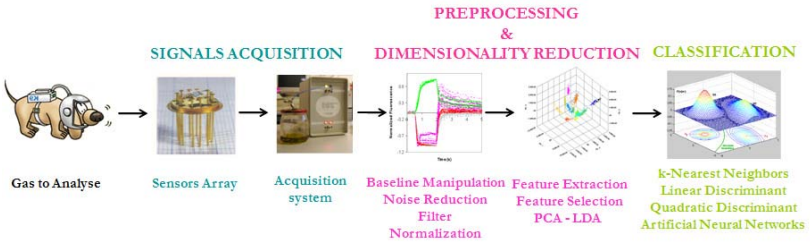

Fig. 1. Block scheme of an electronic nose.

their breath, by the use of an electronic nose and an appropriate classification algorithm.

The experiment has been developed within the Italian MILD (Multicentric Italian Lung Detection) project, promoted by the Istituto Nazionale Tumori of Milan, Italy. The study has been approved from the Ethical Committee of the Institute and we asked all volunteers to sign an agreement for the participation to the study. We analyzed the breath of 101 volunteers, of which 58 healthy and 43 suffering from different types of lung cancer. In particular 23 of them have a primary lung cancer, while 20 of them have different kinds of pulmonary metastasis. Control people do not have any pulmonary disease and have negative chest CT scan. The breath acquisition has been made by inviting all volunteers to blow into a nalophan bag of approximately $400cm^3$. Considering that the breath exhaled directly from lung is contained only in the last part of exhalation, we decided to consider only this portion of the breath. We used a spirometer to evaluate each volunteer exhalation capacity and, at the end of the exhalation, we diverted the flow into the bag. Finally, the air contained in the bag has been input to the electronic nose and analyzed. From each bag we took two measures, obtaining a total of 202 measurements, of which 116 correspond to the breath of healthy people and 86 to diseased ones.

2 Processing and Classification of the Olfactory Signal

An electronic nose is an instrument able to detect and recognize odors, namely the volatile organic compounds present in the analyzed substance. It consists in three principal components (Figure 1): a *Gas Acquisition System*, a *Preprocessing and Dimensionality Reduction phase* and a *Classification Algorithm*. In particular the acquisition of the olfactory signal is done through a sensor array that converts a physical or chemical information into an electrical signal. MOS sensors are characterized by high sensitivity (in the order of ppb), low cost, high speed response and a relatively simple electronics. Considering that most of the VOCs markers of lung cancer are present in the diseased people's breath in very small quantities, varying from parts per million to parts per billion, we chose to use this kind of sensors rather than others. In particular, we used an array composed of six MOS sensors (developed by SACMI s.c.), that react to gases

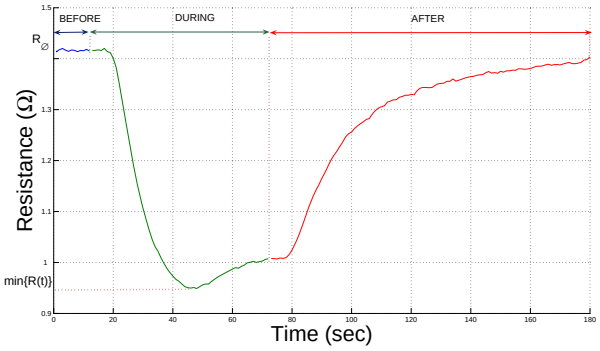

Fig. 2. Example of a typical sensor response.

with a variation of resistance. The VOCs interact with a doped semiconducting material deposited between two metal contacts over a resistive heating element, which operates from 200 °C to 400 °C. As a VOC passes over the doped oxide material, the resistance between the two metal contacts changes in proportion to the concentration of the VOC. The registered signal corresponds to the change of resistance through time produced by the gas flow [3]. In Figure 2 it is possible to see a typical response of a MOS sensor. In particular, each measure consists of three main phases:

1. **Before**: during this time the instrument inhales the reference air, showing in its graph a relatively constant curve;
2. **During**: it is the period in which the electronic nose inhales the analyzed gas, producing a change of the sensors' resistance. It is the most important part of the measurement because it contains informations about how sensors react to the particular substance;
3. **After**: during this phase the instrument returns to the reference line.

After the electronic nose has acquired the olfactory signal, the pre-processing phase begins; its purpose is to reduce the effect of humidity, to normalize the obtained signal and to manipulate the baseline. The latter transforms the sensor response w.r.t. its baseline (e.g., response to a reference analyte) for the purposes of contrast enhancement and drift compensation [2].

After pre-processing, we performed dimensionality reduction to extract the most relevant information from the signal. We reached this objective through *Features Extraction*, *Features Selection* and *Features Projection* in a lower dimensional space. The first operation extracts those descriptors from the sensors' responses able to represent data characteristics in the most efficient way. Feature selection finds, among all possible features, those ones that maximize the informative components and, thus, the accuracy of classification. In particular, we applied the non-parametric test of Mann-Whitney-Wilcoxon [11] with a significance level equal to $\alpha = 0.0001$ to select only discriminant descriptors. In order

to evaluate the discriminative ability of the combination of more features, we performed an Analysis of Variance (ANOVA) [11] and several scatter plots. Let define $R(t)$ the curve representing the resistance variation during the measurement and R_0 the value of the resistance at the beginning of the measurement (as indicated in Figure 2), we found as the most discriminative features between the two classes 'healthy' and 'sick':

- **Single Point**. It is the minimum value of resistance reached during the measurement: $S = \min(R(t))$;
- **Delta**. It corresponds to the resistance change of sensors during the measurement: $\delta = R_0 - \min(R(t))$;
- **Classic**. It is the ratio between the reference line and the minimum value of resistance reached during the measurement: $C = R_0 / \min(R(t))$;
- **Relative Integral**. It is calculated as: $I = \int R(t)/(t \cdot R_0)$;
- **Phase Integral**. It represents the closed area determined by the plot of the state graph of the measurement [8]: $x = R, \quad y = dR/dt$.

After feature selection we performed data projection: we considered Principal Component Analysis (PCA) [10] and Nonparametric Linear Discriminant Analysis (NPLDA) [12], that is based on nonparametric extensions of commonly used Fisher's linear discriminant analysis [10]. PCA transforms data in a linear way projecting features into the directions with maximum variance. It is important to notice that PCA does not consider category labels; this means that the discarded directions could be exactly the most suitable for the classification purpose. This limit can be overcome by NPLDA, which looks for the projection able to maximize differences between different classes and minimize those intra-class. In particular, NPLDA removes the unimodal gaussian assumption by computing the between scatter-matrix S_b using local information and the k nearest neighbors rule; as a result of this, the matrix S_b is full-rank, allowing to extract more that c-1 features (where c is equal to the number of considered classes) and the projections are able to preserve the structure of the data more closely [12]. As evident from Figure 3, NPLDA is able to separate the projected features more clearly than PCA, which plot shows a more evident overlap of samples. This means that NPLDA is more suitable, for the problem considered, in terms of classification performance. Moreover, the plot and the obtained eigenvalues clearly indicated that only one principal component is needed.

Once the most representative characteristics are found, it is possible to perform the analysis of the data, that, in this case, consists in a *pattern recognition* algorithm. In particular, we considered Fuzzy k-Nearest Neighbors (Fuzzy k-NN) classifier, a variation of the classic k-NN, based on a fuzzy logic approach [13]. The basic idea of k-NN is to assign a sample to the class of the k closest samples in the training set. This method is able to do a non linear classification, starting from a small number of samples. The algorithm is based on a measure of the distance (in this case, the Euclidean one) between the normalized features and it has been demonstrated [10], that the k-NN is formally a non parametric approximation of the Maximum A Posteriori MAP criterion. The asymptotic

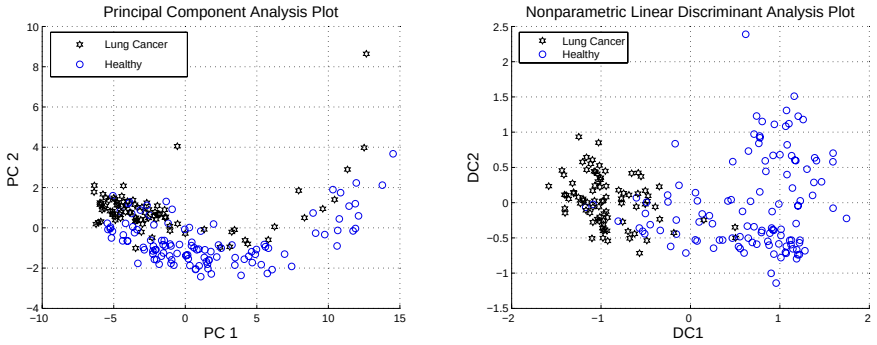

Fig. 3. The result of dimensionality reduction through PCA on the left and NPLDA on the right.

performance of this simple and powerful algorithm, is almost optimum: with an infinite number of samples and setting k=1, the minimum error is never higher than the double of the Bayesian error (that is the theoretical lower bound reachable) [10]. One of the most critical aspects of this method regards the choice of parameter k with a limited number of samples: if k is too large, then the problem is too much simplified and the local information loses its relevance. On the other hand, a too small k leads to a density estimation too sensitive to outliers. For this reason, we decided to consider the Fuzzy k-NN, a variation of the classic k-NN that assigns a fuzzy class membership to each sample and provides an output in a fuzzy form. In particular, the membership value of unlabeled sample x to i^{th} class is influenced by the inverse of the distances from neighbors and their class memberships:

$$\mu_i(x) = \frac{\sum_{j=1}^{k} \mu_{ij} (\|x - x_j\|)^{\frac{-2}{m-1}}}{\sum_{j=1}^{k} (\|x - x_j\|)^{\frac{-2}{m-1}}} \tag{1}$$

where μ_{ij} represents the membership of labeled sample x_j to the i^{th} class. This value can be crisp or it can be calculated according to a particular fuzzy rule: in this work we defined a fuzzy triangular membership function with maximum value at the average of the class and null outside the minimum and maximum values of it. In this way, the closer the sample j is to the average point of class i, the closer its membership value μ_{ij} will be to 1 and vice versa. The parameter m determines how heavily the distance is weighted when calculating each neighbor's contribution to the membership value [14]; we chose $m = 2$, but almost the same error rates have been obtained on these data over a wide range of values of m.

3 Results and Conclusion

The performance of the classifier has been evaluated through the obtained confusion matrix and performance indexes. Being 'TruePositive' (TP) a sick sample

classified as sick, 'TrueNegative' (TN) a healthy sample classified as healthy, 'FalsePositive' (FP) a healthy sample classified as sick and 'FalseNegative' (FN) a sick sample classified as healthy, performance indexes are defined as:

- Accuracy (Non Error Rate NER)=$(TP+TN)/(TP+FP+TN+FN)$;
- Sensitivity (True Positive Rate TPR)=$(TP)/(TP+FN)$;
- Specificity (True Negative Rate TNR)=$(TN)/(TN+FP)$;
- Precision w.r.t. diseased people ($PREC_{POS}$)=$(TP)/(TP+FP)$;
- Precision w.r.t. healthy people ($PREC_{NEG}$)=$(TN)/(TN+FN)$.

To obtain indexes able to describe in a reliable way the performances of the algorithm, it is necessary to evaluate these parameters on new and unknown data, validating the obtained results. Considering the not so big dimension of population and that for every person we had two samples, we opted for a modified Leave-One-Out approach: each test set is composed by the pair of measurements corresponding to the same person, instead of a single measure as would be in the normal Leave-One-Out method. Doing this way, we avoided that one of these two measures could belong to the training set, while using the other in the test set. In order to deeply understand the relevance of the obtained performance indexes, we calculated the corresponding confidence intervals, which lower and upper bounds are defined as:

$$\bar{X} - t_{\frac{\alpha}{2}} \frac{\sigma}{\sqrt{n}} \leq \mu_x \leq \bar{X} + t_{\frac{\alpha}{2}} \frac{\sigma}{\sqrt{n}} \tag{2}$$

where $\bar{X}$ is the registered index value, n is the number of the degrees of freedom, σ is the standard deviation and $t_{\frac{\alpha}{2}}$ is the quantile of the t-student distribution corresponding to the degrees of freedom-1 of the problem.

Results obtained by Fuzzy k-NN are very satisfactory, leading to an accuracy of 92.6%. The confusion matrix obtained by this algorithm is shown in Table 1(a), where elements along the principal diagonal represent respectively the TruePositive and the TrueNegative values, while those off-diagonal are respectively the FalseNegative and the FalsePositive values. Performance indexes and their corresponding confidence intervals (set CI=95%), are reported in Table 1(b). A relevant consideration regards the robustness of Fuzzy k-NN to k changes: we considered different values of k, but the algorithm demonstrated to be very robust to these changes, keeping its results invariant.

In order to prove the effectiveness of Fuzzy k-NN for the considered problem, we evaluated also other families of classifiers: in particular we considered performance achieved by the classic k-NN, by a feedforward artificial neural network (ANN) and by two classifiers based, respectively, on linear and quadratic discriminant functions. All obtained results were comparable or worst than those achieved by Fuzzy k-NN in terms of average accuracy. Considering the single indexes we noticed that sensitivity and precision w.r.t healthy people were higher using Fuzzy k-NN classifier. This consideration is very important because in diagnosis sensitivity is more relevant than specificity because it is more important to recognize correctly a sick person instead of a healthy one; in the same way,

Table 1. Confusion matrix (a) and performance indexes (b) obtained from Fuzzy k-NN algorithm (k=1,3,5,9,101).

(a)

CONFUSION MATRIX		ESTIMATED LABELS	
		Positive	Negative
TRUE LABELS	Positive	82	4
	Negative	11	105

(b)

Indexes	Average Index	Confidence Interval (CI = 95%)
Accuracy	92.6%	[88.5-96.7]
Sensitivity	95.3%	[91.8-98.9]
Specificity	90.5%	[86.0-95.0]
PREC_{POS}	88.2%	[82.3-94.1]
PREC_{NEG}	96.3%	[93.2-99.4]

Table 2. Comparison of lung cancer diagnosis performance and corresponding confidence intervals (set CI=95%) reached by the electronic nose presented in this work and current diagnostic techniques. Data from [9]. Note that results regarding CAT and PET have been obtained from a different dataset than the one analyzed by the E-Nose.

	Accuracy	Sensitivity	Specificity	PREC_{POS}	PREC_{NEG}
CAT	Nd	75%	66%	Nd	Nd
Confidence Interval		[60-90]	[55-77]		
PET	Nd	91%	86%	Nd	Nd
Confidence Interval		[81-100]	[78- 94]		
E-Nose	92.6%	95.3%	90.5%	88.2%	96.3%
Confidence Interval	[88.5-96.7]	[91.8-98.9]]	[86.0-95.0]	[82.3-94.1]	[93.2-99.4]

precision w.r.t. negative samples is more relevant than precision w.r.t. positive ones, because it is worse to classify a person as healthy when he or she is actually sick, than the opposite. Moreover the robustness showed by the Fuzzy k-NN's to k changes is not verified in the classic k-NN, that lead to different results according to different values of k. However, performing a Student's t-test between all pair of classifiers, no relevant differences emerged; this means that implemented classifiers' results are comparable for the problem considered.

The use of an electronic nose as lung cancer diagnostic tool is reasonable if it gives some advantage compared to current lung cancer diagnostic techniques, namely Computed Axial Tomography (CAT) and Positron Emission Tomography (PET). Not only this is verified in terms of performance, as illustrated in Table 2, but also because the electronic nose, unlike the classical approaches, is a low cost, robust, small (and thus, eventually portable), very fast and, above all, non invasive instrument.

In literature there are three other main research works regarding lung cancer diagnosis by an electronic nose [4–6]. Accuracy indexes obtained from these works were respectively equal to 90.32%, 88.16% and 80%. Moreover, in [5] and [6], no cross-validation techniques has been applied to obtain such results; this means that results have been obtained from one realization and, therefore, they are not necessarily representative of the real generalization capability of the classifier.

An ambitious research prospective regards the individuation of risk factors connected to lung cancer (as smoke or food). Involving a larger population and partitioning it according to different disease stages, it would be possible to study the possibility of early diagnosis, that is the most important prospective of research that this work should follow.

References

1. Gardner J.W., Bartlett P.N., Electronic noses. Principles and applications, Oxford University Press, USA, 1999
2. Osuna R.G., Nagle H.T., Shiffman S.S., The how and why of electronic nose, IEEE Spectrum, pp. 22-34, September 1998
3. Pardo M., Sberveglieri G., Electronic Olfactory Systems Based on Metal Oxide Semiconductor Sensor Arrays, In Material Research Society Bullettin, Volume 29, No. 10, October 2004
4. Di Natale C., Macagnano A., Martinelli E., Paolesse R., D'Arcangelo G., Roscioni C., Finazzi-Agro A., D'Amico A., Lung cancer identification by the analysis of breath by means of an array of non-selective gas sensors, Biosensors and Bioelectronics, vol. 18, no 10, pp. 1209-1218, 2003
5. Machado R.F., Laskowski D., Deffenderfer O., Burch T., Zheng S., Mazzone P.J., Mekhail T., Jennings C., Stoller J.K., Pyle J., Duncan J., Dweik R.A., Erzurum S., Detection of Lung Cancer by Sensor Array Analyses of Exhaled Breath, American Journal of Respiratory and Critical Care Medicine, vol. 171, pp. 1286-1291, 2005
6. Chen X., Cao M., Li Y., Hu W., Wang P., Ying K., Pan H., A study of an electronic nose for detection of lung cancer based on a virtual SAW gas sensors array and imaging recognition method, Measurement science & technology, vol. 16, no 8, pp. 1535-1546 , 2005
7. Phillips M.D., Cataneo R.N., Cummin A.R.C., Gagliardi A.J., Gleeson K., Greenberg J., Maxfield R.A., Rom W.N., Detection of lung cancer with volatile markers in the breath, Chest, vol. 123, no 6, pp. 2115-2123, 2003
8. Martinelli E., Falconi C., D'Amico A., Di Natale C., Feature Extraction of chemical sensors in phase space, Sensors and Actuators B:Chemical, vol. 95, no 1, pp. 132-139, Elsevier Science, 2003
9. Pieterman R.M., Van Putten J.W.G., Meuzelaar J.J., Mooyaart E.L., Vaalburg W., Koter G.H., Fidler V., Pruim J., Groen H.J.M., Preoperative staging of non-small-cell lung cancer with positron-emission tomography, The New England journal of medicine, vol. 343, no 4, pp. 254-261, 2000
10. Fukunaga K., Introduction to statistical pattern recognition, Academic Press, Second Edition, San Diego, 1990
11. Lyman Ott R., Longnecker M.T., An Introduction to Statistical Methods and Data Analysis, Duxbury Press, 5th Edition, 2001
12. Fukunaga K., Mantock J.M., Nonparametric discriminant analysis, IEEE Transactions on pattern analysis and machine intelligence, vol. PAMI 5, no 6,pp. 671-678, 1983
13. Zadeh L., Fuzzy sets, Information and Control, vol. 8, pp. 338-353, 1965
14. Keller J.M., Gray M.R., Givens J.A., A fuzzy k-Nearest neighbor algorithm, IEEE Transactions on systems, man and cybernetics, vol. 15, no 4, pp. 580-585, July-August 1985