IL NUOVO CIMENTO **47 C** (2024) 316
DOI 10.1393/ncc/i2024-24316-4

COMMUNICATIONS: SIF Congress 2023

Artificial intelligence algorithms for prostate cancer prediction by breath analysis

A. LO SASSO(1)(2)(3)(*), L. BELLANTUONO(3)(4), R. MANETTA(5), A. GAZZERRO(6), R. DI FILIPPO(6), F. PORCELLI(2), L. FACCHINI(2), P. FRASCELLA(2) and R. BELLOTTI(1)(3)

(1) *Dipartimento Interateneo di Fisica, Università degli Studi di Bari - Bari, Italy*
(2) *Predict S.r.l. - Fiera del Levante, Bari, Italy*
(3) *Istituto Nazionale di Fisica Nucleare, Sezione di Bari - Bari, Italy*
(4) *Dipartimento di Biomedicina Traslazionale e Neuroscienze (DiBraiN), Università degli Studi di Bari - Bari, Italy*
(5) *Dipartimento di Emergenza Urgenza e Accettazione, Ospedale S. Salvatore L'Aquila, Italy*
(6) *Scuola di Specializzazione in Radiologia, Ospedale S. Salvatore - L'Aquila, Italy*

received 20 February 2024

Summary. — Breath analysis is emerging as a promising screening technique. Analysis of volatile organic compounds by artificial intelligence can lead to an early warning bell by noninvasive screening. In this work, we analyze volatile organic compounds from patients who had undergone a prostate cancer exam. We expose a computational methodology to discriminate suspected and full-blown patients. We adopt a sample selection and use oversampling to train a neural network based on a Multi-Layer Perceptron in order to predict real data and simulate performances on further patients undergoing prostate cancer screening.

1. – Introduction

Breath analysis-based metabolomics, also called breathomics, is focused on the study of patterns of volatile organic compounds (VOCs) found in the exhaled breath. Breathomics is emerging as a possible new monitoring technique. VOCs are the product of highly dynamic metabolic processes that take place throughout the entire organism. VOCs reach the lungs via the blood and undergo diffusive processes at the level of the alveolar capillary membrane. Recent studies have shown it is already possible to detect cancer patients by sniffing their exhaled [1]. The biological nose is not very reliable

(*) E-mail: andrea.losasso@uniba.it

because it is susceptible to sensitization and influenced by the body's indispositions or external factors. Nowadays we know that the exhaled breath is a complex matrix containing about 250 VOCs. In this work, we outline a methodology based on artificial intelligence for the cancer prostate screening starting from VOCs in gaseous breath samples. First, we describe the data collection ad sampling selection steps and discuss the limitation in adopting classical statistics methodologies. Then, we show the results of the classification framework based on artificial intelligence and explore further perspectives of this research work.

2. – Data collection and sample selection

Sampling is performed by means of Mistral, the breath sampler (medical device) developed and commercialized by Predict S.r.l. [2]. Breath sampling follows a natural flow of exhalation thanks to the mouthpieces constructed to mitigate the resistance due to air pressure. During breath sampling, the system also collects ambient air, in order to better distinguish between exogenous and endogenous molecules [3]. The VOCs collected are then thermally desorbed (Unity-xr, Markes) and analysed by gas chromatography-mass spectrometry (Agilent 8860-MSD5977). Each VOC found in the chromatogram is identified by matching the mass spectrum with the spectral library of National Institute of Standards and Technology, and confirmed with VOCs standard matrix. After the sampling and identification of VOCs abundance in the cartridges, we perform data analysis [4]. VOCs abundances are set in a matrix having a number of rows equal to the number of patients sampled and number of columns equal to the VOCs recorded. The data for each cell (i,j) represents the abundance of VOC j-th for patient i-th.

3. – Prostate cancer case study

Prostate cancer is one of the most common cancers among men, and the risk is directly related to age: between 50 and 60 years old up to 1 in 4 men may have cancerous cells in the prostate [5]. The most advanced technique for prostate tumor diagnosis is multiparametric MRI. The dataset used in this study consists of 147 subjects: 130 labelled as *Suspected* and 17 labelled as *Full-Blown*. Suspected patient label indicates a subject whose PSA value (Prostate-Specific-Antigen test) is equal to 2 or 3, Full-Blown patient label indicates a patient whose PSA value is equals to 4 or 5. For each subject, 257 VOCs have been collected.

3˙1. *Statistical analysis*. – The first approach to data is by means of statistical methodologies. Figure 1 shows the normalized distribution of VOCs abundance for patients labeled as Suspects (dotted line) and patients labeled as Full-Blown (marked dotted line). VOCs abundances are perfectly overlapped, thus it is difficult for statistics to distinguish the sick patients by the distribution of values.

We explore whether the distributions of VOCs abundances of suspected and full-blown patients are statistically distinguishable. To answer this question, we opt for the Wilcoxon rank-sum test, which tests the null hypothesis that two sets of measurements are drawn from the same distribution. However, only 6 VOCs out of 257 report statistically significant differences between the two cohorts: chloromethane (Wilcoxon test value 1.99, p-value 4.63e-02), Acetonitrile (Wilcoxon test value 2.05, p-value 4e-02), Furan, 3-methyl-13.6 (Wilcoxon test value 2.10, p-value 3.56e-02), ion 71 tr 13.88 (Wilcoxon test value 2.97, p-value 2.94e-03), ion 43 tr 18.3 (Wilcoxon test value 2.62, p-value 8.72e-03)

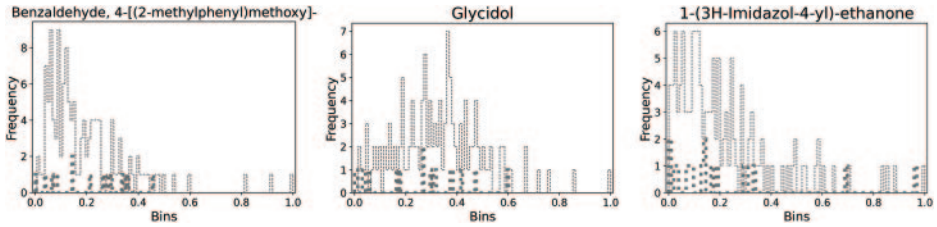

Fig. 1. – Histogram of normalized distributions of Benzaldehyde, 4-[(2-methylphenyl)methoxy]-, Glycidol and 1-(3H-Imidazol-4-yl)-ethanone. VOCs abundances of exhaled breaths from the sick patients (marked dotted line) are perfectly overlapping with VOCs abundances from suspected patients (dotted line). These VOCs are selected from the dataset as a generic example.

and ion 115 tr 29.9 (Wilcoxon test value 2.10, p-value 3.56e-02). Since standard statistical methods does not allow to efficiently discriminate the abundances referred to Suspected and Full-Blown subjects, we move on to approaches based on artificial intelligence.

3˙2. *Artificial intelligence analysis*. – The algorithm we adopt is a neural network based on a Multi-Layer Perceptron (MLP), implemented the activation function set to sigmoid. We assign label 0 to Suspected patients and label 1 to Full-Blown patients. Due the limited size of dataset (130 suspected patients and 17 Full-Blown patients), we decide to over-sample the dataset. This approach allows to train MLP with more data. We use the SMOTE algorithm to perform the over-sampling. Following leave-one-out validation, we remove one patient from the real dataset and construct a dataset of 10000 synthetic patients (5000 per class) from the features of the remaining real patients. We used 75% of this dataset as training set and 25% as test set. After identifying the optimal number of epochs in the algorithm using the validation loss and training curves, we evaluate the performance by using leave-one-out validation. Validation test is performed predicting the label of the patient removed before over-sampling that MLP never analyzed either in training or in test. We repeat this workflow 147 times, one for each real patient. MLP correctly identifies samples labeled as Suspected in 83% of cases (specificity), but samples labeled as Full-Blown are well predicted only in 12% of cases (sensitivity). This disproportion in performance highlights a lack of generalization. Furthermore, the fact that MLP classifies 88% of the data labeled as Full-Blown in Suspected is an indication that there are patterns in the Suspected samples that lead back to Full-Blown and confound MLP in the discrimination. To solve this issue, we performed an unsupervised classification with Self-Organized Map (SOM). SOM is a clustering algorithm based on neural networks. SOM discriminates the set of Suspected patients into two subgroups, containing 87 and 43 samples, respectively. The process of data splitting, labelling, analysis, and validation of the data is shown discussed in detail [6]. The cluster consisting of 87 patients is selected as the unambiguous one, which does not contain confounding patterns, while the cluster containing 43 samples is discarded. Hence, we assume that the 87 samples are the *superpure* Suspected patients, *i.e.*, a group without confounding patterns. We run MLP on the selected dataset, consisting of 87 *superpure* Suspected patients and 17 Full-Blown patients. To follow leave-one-out validation, we remove one patient from the real dataset and make the over-sampling to 10000 synthetic patients, 5000 cases per class as presented before. We split this dataset using 75% of data as training set and 25% as testing set, stratifying samples of two classes among training and test set. Validation is then performed by predicting the

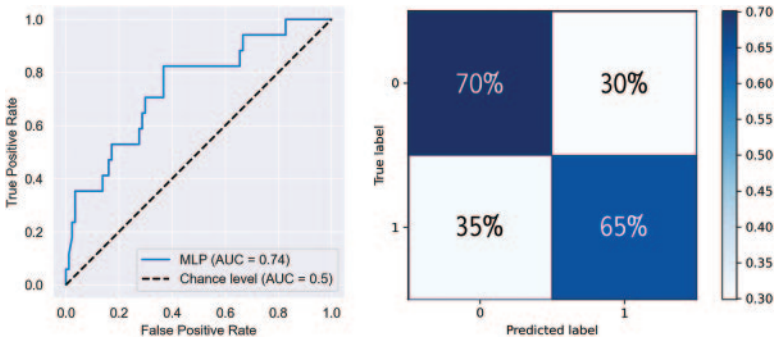

Fig. 2. – MLP performance after the SOM clustering selection. On the left, the ROC curve indicates that MLP reaches an Area Under Curve (AUC) equal to 0.74. On the right, the confusion matrix reports the percentages of patients which are well predicted or not.

patient removed before over-sampling that MLP never analyzed either in training or in test. We repeat this workflow 104 times, one for each real patient. The results of the predictions are shown in fig. 2. The performance increases sharply with this set-up, reaching 70% in accuracy, 70% in specificity and 64% in sensitivity.

4. – Conclusion

Prostate cancer affects up to 1 in 4 men between 50 and 60 years old. The number of people contracting this disease will grow in the coming years according to the latest estimates. So, it is important to find an effective screening technique to intervene early. Combining breath analysis and artificial intelligence is emerging as a promising screening technique. In addition, breath analysis has the worth of being completely noninvasive, since it allows to repeat several times the screening without any damage to the patient. The easy reproducibility of sampling allows to perform a patient monitoring during the different stages of disease. The results shown in this study provide an effective discrimination between suspected and full-blown label. On the other hand, this work can be improved by increasing the sampling and allowing MLP to avoid selection bias. In addition, it will be important to understand why some data are confusing, such as the ones removed by SOM.

REFERENCES

[1] Sonoda H. *et al.*, *Gut*, **60** (2011) 814.
[2] Predict S.r.l., *Breath Technologies*, `https://www.predictcare.it/mistral`.
[3] Di Gilio A. *et al.*, *Molecules*, **25** (2020) 5823.
[4] Di Gilio A. *et al.*, *Cancers*, **12** (2020) 1262.
[5] Rawla P., *World J. Oncol.*, **10** (2019) 63.
[6] Bellotti R. *et al.*, *Proc. SPIE*, **2492** (1995) 1153.