Discriminating Artificial Cancer Breath Using an Electronic Nose : K-Nearest Neighbors Versus Long-Short Term Memory Network

Justin MARTIN
Department of Environmental Sciences
University of Liège
Arlon
jdm.martin@uliege.be

Claudia FALZONE
Department of Environmental Sciences
University of Liège
Arlon
cfalzone@uliege.be

Anne-Claude ROMAIN
Department of Environmental Sciences
University of Liège
Arlon
acromain@uliege.be

***Abstract*—This paper presents the use of k-NN for the classification of healthy human breath with or without the addition of lung cancer biomarkers. 236 breath samples collected from 17 persons over four months were analyzed by a custom electronic nose using commercial metal oxide sensors. About 90% of the samples were correctly classed by the model. Long-Short Term Memory Neural Network could show promising results in this task as well and are under investigation.**

Keywords—Breath analysis, e-nose, metal oxide, cancer, kNN, LSTM

I. INTRODUCTION

The early detection of cancer through non-invasive means is a constantly evolving area of medical research, promising significant benefits in patient prognosis and treatment efficiency. Electronic noses (e-noses) have been studied to fill this role for several years [1]. Their fundamental operating principle is based on sensor non-specificity, which means the system provides a general profile of the entire gas mixture, with the ability to classify it without being able to identify its constituents. This makes it highly effective for identifying and monitoring odours. Its applications have been widespread, notably in environmental studies and the food industry [2]. In recent years, the medical field has shown growing interest in these devices.

This paper presents the detection of cancer markers in breath samples using a home made electronic nose (e-nose), coupled with a k-Nearest Neighbors (k-NN) model. The e-nose, equipped with an array of commercial sensors, is designed to identify the general signature of the various Volatile Organic Compounds (VOCs) associated with cancerous conditions.

A LSTM network, a form of neural network, will also be employed to analyze the response (baseline subtracted conductivity) of the sensors, learning to recognize patterns indicative of the healthy/artificial cancer quality of the breath samples. However, results for the LSTM model are not available at this moment and will not be presented in this paper. The two methods will be compared later on.

The e-nose system's ability to discriminate between cancerous and non-cancerous breath samples was assessed under various concentrations levels. This experiment not only reinforces the potential viability of using e-noses for early cancer detection but also showcases the effectiveness of k-NN models in processing multivariate data. The combination of these technologies shows great results which indicate the viability of the approach for a cancer screening device.

II. MATERIALS AND METHODS

A. On breath sampling

To validate the system, sampled breath from healthy volunteers was sampled. The recruitment for healthy breath sample collection commenced with an invitation to volunteers at our laboratory. Participation was voluntary and unpaid. The recruitment strategy aimed to attract a diverse group of volunteers, acknowledging that the number of smokers and ex-smokers might be low due to their minority status on the campus. Achieving equal-sized subgroups based on characteristics like gender, smoking status, or exercise habits was not considered feasible.

Eligibility criteria required volunteers to be at least 18 years old and free from respiratory problems or lung diseases. To facilitate greater participation, volunteers were allowed to choose the most convenient day for them to participate. They were asked to confirm their availability a week in advance and needed to participate at least four times during the study period.

In total, 17 healthy individuals joined the campaign. Each participant was required to complete a brief questionnaire detailing their biometrics (such as sex, age, height, and weight) and lifestyle habits (including smoking history and frequency, and exercise frequency). They were also asked to report any changes in these details throughout the course of the study.

Participants in the breath collection study were required to fast overnight and adhere to a strict policy of "nothing by mouth" on the morning of their sampling. This meant avoiding activities such as smoking, brushing teeth, chewing gum, or eating, although drinking water was permitted. On arrival for sampling, they were provided with water for rinsing their mouths. Subsequently, they filled a 10-liter Fluorinated Ethylene Propylene (FEP) bag with their breath. This was done through a saliva filter used in spirometry, housed in a Polytetrafluoroethylene (PTFE) holder.

Over the course of three months, from January 2023 to March 2023, the breath of the 17 healthy participants was collected on 17 separate days. This resulted in a total of 170 samples (cancer and healthy) generated from 85 unique breath samples. On average, about 5 samples were collected per day. Each participant contributed approximately 6 samples throughout the study, with the median number of samples provided being 3.5.

This work was financed by Interreg FWV for the Pathacov project.

B. *On breath processing*

The collected breath samples undergo a process where half of each sample is infused with VOC (Volatile Organic Compound) biomarkers. To achieve parts per million (ppm) concentration levels, microinjections are employed. Using a micro-syringe, quantities ranging from 0.1 to 2µL of pure compounds are added to an 8L FEP sampling bag (HedeTech®, Dalian, China) already containing the breath. Following the compound injection, the bag is heated for 30 minutes at 60°C. To obtain sub-ppm concentrations, the contents of the injected bag are diluted with human breath from volunteer samples, controlled by mass flow controllers.

The specific compounds used for injection are selected based on a citation frequency analysis from literature on GCMS studies of lung cancer biomarkers in breath, which has been presented in a previous publication. For more information, the readers are invited to look in the reference [3].

Following vaporization, the bag is placed in a pressure chamber and pressurized to +1 bar. The flow of these concentrated biomarkers is then adjusted to achieve the desired final concentrations for artificial cancer breath. Standard flow rate for literature concentrations is set at 3.93mL/min, but prior trials have shown that such low concentrations give poor classification results. To better grasp the limitations of the technology, it was chosen to raise the concentration. Multiples of the "realistic" flow rate are used: 15.72mL/min (4x), and 30.0mL/min (8x), all diluted against a baseline of 1300mL/min of breath.

Each sample, whether healthy or artificial cancer breath, was analyzed using the e-nose for a duration of five minutes. To ensure accuracy, reference air was drawn in for a minimum of five minutes before and after each sample's analysis. The flow rate during the process was consistently maintained at 200mL/min, regulated by a downstream pump and flowmeter. The reference air, consisting of pure analytical-grade air provided by Air Liquide® (Paris, France) was humidified at 20°C and stored in a 25-liter FEP bag. A manual 3-way valve was employed for alternating between sample and reference air.

C. *About the electronic nose*

The composition and specifics of the electronic nose have been discussed in previous publications [3,4].

In the data analysis process for each sample, the highest conductance value recorded by each of the six sensors is selected. This maximum value is then adjusted by subtracting the baseline conductance, which is the stable conductance level of the sensor in reference air. Following this, drift correction is applied to the data for accuracy (detailed methodology can be found in the supplementary materials of [4]). The final step in the data processing is normalization, as described in equation 1.

Normalization is crucial because it cancels the influence of concentration on the data. This adjustment is significant as electronic noses are designed to differentiate gas mixtures based on composition rather than concentration. The aim is to detect relative concentrations of compounds within a mixture rather than their absolute amounts. In the normalization equation, x_{ij} represents the original (un-normalized) conductance of the sensor, while x'_{ij} is the normalized conductance. Here, 'j' denotes the sample number, and 'i' refers to the specific sensor.

Equation 1: $$x'_{ij} = \frac{x_{ij}}{\sqrt{\sum_{i=1}^{n} x_{ij}^2}}$$

Evaluating the discrimination capacity of the electronic nose using normalized data is essential. This is because variations in concentration levels in breath samples can occur due to factors unrelated to the cancer presence. Not accounting for these variations could result in incorrect interpretations, such as false positives or, more critically, false negatives.

D. *About kNN and the LSTM neural network*

K-Nearest Neighbors (kNN) is well-known classification tool, which has been used in similar works in the past [5]. kNN operates by first locating the "k" closest data points (or "neighbors") in the feature space to the new data point that needs to be categorized. This proximity is typically measured using a distance metric (e.g. Euclidean distance). The algorithm then determines the output for the new data point based on the outputs of these nearest neighbors: the most common class label among the neighbors is assigned to the new point. The kNN analysis of the dataset was done on the JMP® software. The dataset was divided into three parts: learning (60%), validation (30%), and test (10%). The selected samples are randomised with a restriction to have equal number of healthy/cancer samples per group, and the samples for the test part were taken to be from a different sampling day altogether. The first 30 values of k were explored. Due to the bigger amount of datapoints in the LSTM dataset (due to every sample being represented by a time serie), the values of k explored were raised to 100. For each model, a 5-fold cross validation was done.

Long short-term memory (LSTM) neural networks (NN) are notably good at predicting future changes in a sequence based on previous observations [6]. Usually, machine learning with e-nose data is done after variable selection (e.g. isolating the maximum response of each sensor while exposed to a breath sample). LSTM is not fit to work with only a few data points per sample, but rather data such as the conductance in relation to time (called a "time series") for every processed sample. In this case, the data is several hundred points per sample.

The LSTM neural network analysis is currently under work in python and R language. A comparison of the LSTM performance compared to a more classical kNN will be presented later.

III. Results and Discussion

kNN is as performing as a previous experiment of cancer detection on real patients seen in literature [5]. For this dataset, the classification errors during training were, on average across cross validation, satisfactory (10%) and lowered to 9% during validation and for the test set as well. This indicates that the model, on this dataset, is able to correctly classify about 90% of the samples. Best performing k was between 1 and 16 during cross validation.

To better visualize, a Principal Component Analysis (PCA) was realised on this dataset (Fig. 1). The third component seems to contribute the most to the separation of

the data. The sensors contributing to this component are G2530 (62%), MP901 (23%), G3530 (11%) and T2603 (4%) (see [3,4] for more information on sensors).

Of course, this does not indicate that the electronic nose would perform in a similar way for a medical application, a clinical trial is still needed to compare this approach with the real performance of the device. A much larger sample pool is also required to validate the device for medical use. As shown on a previous publication [4], realistic levels of concentration have much poorer performance however and without technological improvements the quality of the results obtained here are unlikely to be obtained in clinical trials.

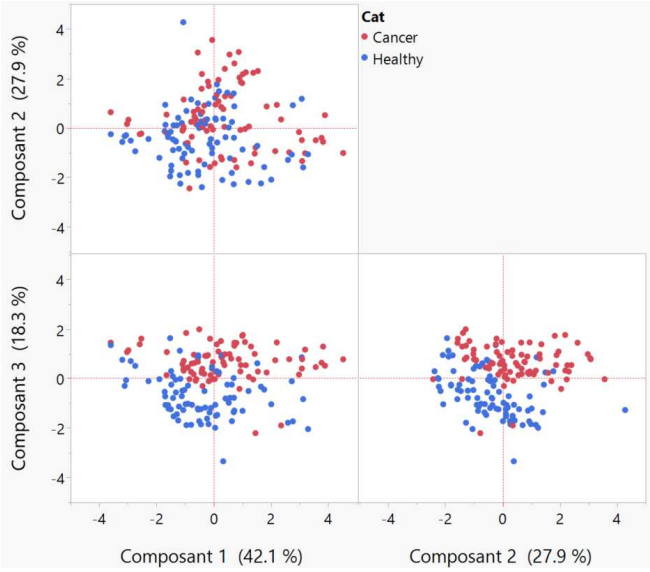

Fig. 1. Principal Component Analysis Score Plot showing the three first components of the dataset used for k-NN model training, validation and testing. The two categories (Cancer in red and healthy in blue) are clearly separated on the 3rd component.

Since the dataset for LSTM has more data points, it was considered interesting to test running the kNN on this data to observe the results and show that presenting the data in this manner had no influence over the classification. The best k was between 25 and 95 during cross validations. The training error rate dropped to 0.3%. Validation however maintains 10%. The test dataset error rate was slightly lowered with 7% misclassification. The obtained classification errors are therefore similar between the two datasets, which was expected.

While it is possible that LSTM picks up information that would not be present in the dataset used for kNN (which is limited to maximum conductance values for each sample), every machine learning (ML) algorithm is limited by the quality of the dataset it is trained on (the famous ML saying "garbage in, garbage out" comes to mind). Provided overfitting is prevented, models can only go so far, and the discrimination power ultimately relies on the quality of the sensing device – the electronic nose.

The betterment of electronic noses is therefore central in their deployment as screening devices in the medical field. A recently published article using the present dataset explores a way to evaluate the power score of an electronic nose device [4].

IV. Conclusion

Using an experimental metal oxide sensor-based electronic nose and artificial cancer breath made from healthy volunteers' breath, we obtained an overall correct classification rate of about 80% using a k-Nearest Neighbour model. However, this result was obtained using concentration up to 8 times the concentrations of biomarkers found in breath.

This is encouraging for the further development of the device and its future use in clinical trials, where it will be tested in real usage conditions and the validity of the findings presented in this paper will be challenged. The need to raise the concentration to obtain good classification rates underlines the need for e-nose technology improvements, with either pre-concentration or more sensitive sensor technology.

Comparison of performance with a LSTM neural network is ongoing. While it is possible that LSTM neural network performs better, special caution will be taken to avoid overfitting of the model.

Acknowledgment

The team would like to thank Simon-Pierre Liégeois for its invaluable contribution to the project, and the numerous contributors that donated their breath and enabled this research.

References

[1] M. Fleischer, E. Simon, E. Rumpel, H. Ulmer, M. Harbeck, M. Wandel, C. Fietzek, U. Weimar, H. Meixner, Detection of volatile compounds correlated to human diseases through breath analysis with chemical sensors, (2002) 5.

[2] J.W. Gardner, P.N. Bartlett, Electronic noses, principles and applications, New York, NY, 1999.

[3] J.D.M. Martin, A.-C. Romain, Building a Sensor Benchmark for E-Nose Based Lung Cancer Detection: Methodological Considerations, Chemosensors 10 (2022) 444. https://doi.org/10.3390/chemosensors10110444.

[4] J. Martin, C. Falzone, A.-C. Romain, How well does your E-nose detect cancer? Application of artificial breath analysis for performance assessment, J. Breath Res. (2024). https://doi.org/10.1088/1752-7163/ad1d64.

[5] R. Blatt, A. Bonarini, E. Calabró, M. Della Torre, M. Matteucci, U. Pastorino, Fuzzy k-NN Lung Cancer Identification by an Electronic Nose, in: F. Masulli, S. Mitra, G. Pasi (Eds.), Applications of Fuzzy Sets Theory, Springer, Berlin, Heidelberg, 2007: pp. 261–268. https://doi.org/10.1007/978-3-540-73400-0_32.

[6] L. Zhao, W. Li, S. Wang, A Concentration Prediction and Gas Classification Model Based on LSTM-Attention Multi-task Learning Framework Network, J. Phys.: Conf. Ser. 2537 (2023) 012020. https://doi.org/10.1088/1742-6596/2537/1/012020.