The 2nd International Electronic Conference on Chemical Sensors and Analytical Chemistry

Classification of teas using different feature extraction methods from signals of a lab-made electronic nose

I. Jiménez-López; J. Molina-Quiroga; J.M. Gutiérrez-Salgado

Bioelectronics Section, Department of Electric Engineering, CINVESTAV-IPN Mexico City, Mexico.

*email: irari.jimenezl@cinvestav.mx; jeniffer.molinaq@cinvestav.mx; mgutierrez@cinvestav.mx

1.- INTRODUCTION

Tea and herbal infusions are the world's most consumed aromatic, non-alcoholic beverage after water. They possess multiple human health functions like antioxidation, anti-inflammation, and immune regulation, among others. Also, teas have volatile organic compounds (VOCs) which are responsible for the color, taste and aroma [1]. Analytical methods, like GC-MS and FT-IR spectrometry [2,3], are employed to classify products according to their chemical composition; however, a new method emerge calling electronic nose (e-nose). which detect the "fingerprint" of a chemical component. This work implements a processing strategy using PCA and PARAFAC techniques to extract principal information and compares their relevance using ANN and k-NN. It demonstrates that conventional PCA is better than complex methods as PARAFAC.

2.- MATERIALS AND METHODS

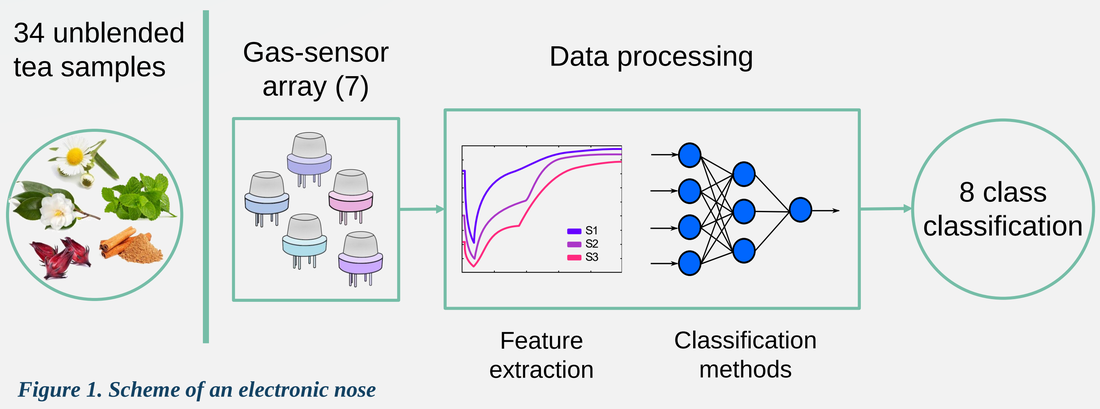

Figure 1. Scheme of an electronic nose

Methodology.

1. **E-nose:** The odor stimuli is controlled of an olfactometer that inject the VOCs into the gas-sensor array which contains seven metal oxide sensors from MQ-series. This sensors detect various gases, including carbon monoxide, liquefied petroleum gas, natural gas, alcohol, benzene, methane, and hydrogen. [4]

2. **Feature extraction:** The data were analyzed using PCA and PARAFAC methods to reduce dimensionality and extract relevant features. PCA finds the linear correlation between the original data variables to produce new uncorrelated linear combinations of these variables using an orthogonal transformation [5], PARAFAC is a multi-way data decomposition method that assumes the existence of a triple path of the data and finds a unique solution [6].

3. **Data Processing:** Two different classification models were used to identify patterns in the data. The first was ANN, which uses a standard trial-and-error process, where several parameters are fine-tuned to find the best configuration to achieve the performance. The second was k-NN that finds a group of k objects in the training set that are near to the test object. k-NN orders the information by computing distances between feature values [14].

Voltage response

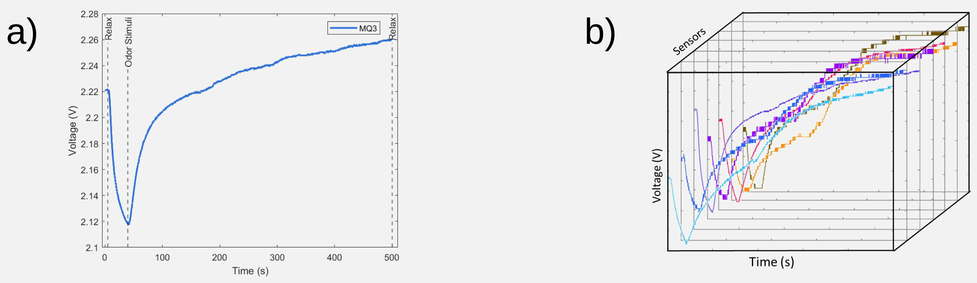

Figure 2. a) The response a white tea detects of sensor MQ3. b) The 3d-representation of one sample

The set of samples was formed by 34 unblended tea samples from commercial brands. Each tea sample was places in the e-nose platform and recorded 10 times to analyze the experiment's repeatability.
Finally, the database was shaped as a tridimensional matrix of 2499 samples, 340 records (34 teas x 10 repetitions), and 7 sensors.

5. ACKNOWLEDGMENTS

Authors would like to express their gratitude to the Mexican National Council of Science and Technology (CONAHCyT) for the financial support and master degree.

6. REFERENCES

[1] G. Y. Tang et al. Int. J. Mol. Sci. vol. 20, pp. 6196, 2019.
[2] W. Chen et al. Food Control., vol. 140, pp. 109103, 2022.
[3] F. Yousefbeyk et al. Fharm. Sci., vol. 29, pp. 100, 2023.
[4] L. Valdez et al. Sensors., vol. 16, pp. 1745, 2016.
[5] I. Jolliffe et al. Principal Components., pp. 1-6, 2016.
[6] E. Acar et al. IEEE Trans. Knowl Data Eng, vol. 21, pp. 6, 2009.
[7] S. Kaushal et al. Agriculture, vol. 12, pp. 1359, 2022

3.- RESULTS AND DISCUSSION

PCA and PARAFAC results

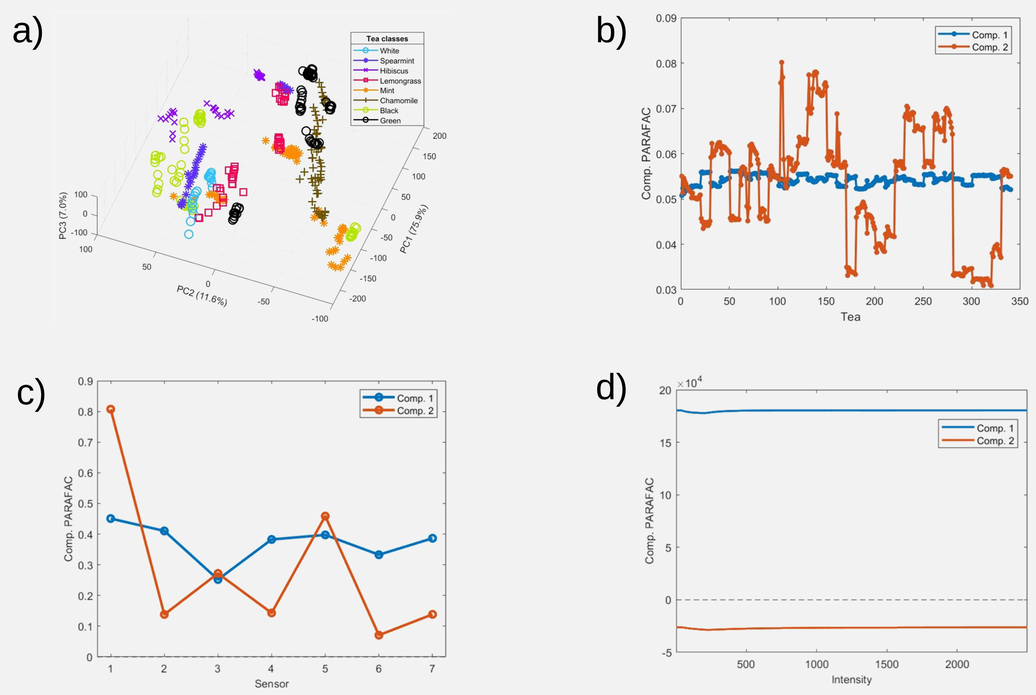

Figure 3. a) PCA score plot of the three first components from eight tea classes. PARAFAC results loadings for b) tea, c) intensities, and c) sensor.

PCA

Data were organized in a two-dimensional array, which the rows denoting teas and the columns denoting measurements for each sensor (340 x 17493). Three principal components failed to achieve the recommended 95% of the accumulated explained variance. So, four PCs (ca. 96.8%), were used to feed the classification models.

PARAFAC

The analysis was performed in the formed tridimensional matrix. To choose the appropriate number of components, a CORCONDIA evaluation was done achieving a core consistency value of 99.2%. The tea loadings represent tea variability; the intensity matrix shows changes in voltage values; finally, the sensor loadings describe the responses of each sensor.

Confusion matrix

Table 1. Confusion matrix and classification rate of PCA and PARAFAC using ANN of 10-fold cross validation.

	Classifed as:	White		Spearmint		Hibiscus		Lemongrass		Mint		Chamomile		Black		Green		Class. Rate	
		FE_1*	FE_2**	FE_1	FE_2	FE_1	FE_2	FE_1	FE_2	FE_1	FE_2	FE_1	FE_2	FE_1	FE_2	FE_1	FE_2	FE_1	FE_2
True class	White	**100**	**90**	0	10	0	0	0	0	0	0	0	0	0	0	0	0	1.00	0.99
	Spearmint	0	0	**90**	**86.6**	10	3.3	0	10	0	0	0	0	0	0	0	0	0.99	0.96
	Hibiscus	0	0	3.3	6.6	**86.6**	**60**	6.6	6.6	0	20	0	6.6	3.3	0	0	0	0.97	0.93
	Lemongrass	0	0	0	5	7.5	12.5	**85**	**57.5**	7.5	17.5	0	7.5	0	0	0	0	0.97	0.87
	Mint	0	0	0	0	0	2	2	20	**94**	**58**	4	16	0	4	0	0	0.98	0.85
	Chamomile	0	0	0	0	0	0	0	5	0	8.3	**93.3**	**75**	5	8.3	1.6	3.3	0.98	0.88
	Black	0	0	0	0	0	0	0	2	0	8	0	12	**100**	**74**	0	4	0.98	0.91
	Green	0	0	0	0	0	0	0	0	0	0	0	1.6	5	8.3	**95**	**90**	0.99	0.96
																	Total	0.98	0.92
		Predicted class																	

Table 2. Confusion matrix and classification rate of PCA and PARAFAC using ANN of 10-fold cross validation.

	Classifed as:	White		Spearmint		Hibiscus		Lemongrass		Mint		Chamomile		Black		Green		Class. Rate	
		FE_1*	FE_2**	FE_1	FE_2	FE_1	FE_2	FE_1	FE_2	FE_1	FE_2	FE_1	FE_2	FE_1	FE_2	FE_1	FE_2	FE_1	FE_2
True class	White	**100**	**95**	0	0	0	0	0	5	0	0	0	0	0	0	0	0	1.00	0.99
	Spearmint	0	0	**100**	**90**	0	0	0	10	0	0	0	0	0	0	0	0	0.99	0.98
	Hibiscus	0	0	3.3	0	**90**	**73.3**	0	3.3	0	13.3	0	6.6	6.6	3.3	0	0	0.99	0.97
	Lemongrass	0	2.5	7.5	7.5	0	0	**80**	**77.5**	12.5	10	0	0	0	0	0	2.5	0.97	0.94
	Mint	0	0	2	0	0	2	2	6	**92**	**78**	0	0	4	14	0	0	0.97	0.91
	Chamomile	0	0	0	0	0	0	0	3.3	0	0	**91.6**	**86.6**	0	0	8.3	10	0.98	0.95
	Black	0	0	0	0	0	0	0	0	0	16	0	0	**100**	**84**	0	0	0.99	0.95
	Green	0	0	0	0	0	0	0	0	0	0	0	8.3	0	0	**100**	**91.6**	0.98	0.96
																	Total	0.98	0.96
		Predicted class																	

4. CONCLUSIONS

- PCA has a superior metrics than using PARAFAC. This is corroborated by PCA-ANN combination that achieved the most remarkable accuracy. So, using only the variance as the main feature allows a better data evaluation.
- Both feature extraction techniques focused on discrimination tasks related to qualitative analyses, considering the content of VOCs. E-noses could be sensitive to the mixture of VOCs per tea, allowing their possible quantification from MOXs signals.