Universidade Federal Rural de Pernambuco
Pró-reitoria de Pesquisa e Pós-graduação
Pós-Graduação em Biometria e Estatística Aplicada

Eva Susana Albarracín Estrada

Approach for Drift Representation and Extraction in Gas Sensors Signals by Sample Entropy

Recife - PE

Dezembro, 2020

Eva Susana Albarracín Estrada

Approach for Drift Representation and Extraction in Gas Sensors Signals by Sample Entropy

Tese apresentada ao Programa de Pós-Graduação em Biometria e Estatística Aplicada do Departamento de Estatística e Informática, em cumprimento das exigências legais para obtenção do título de doutor.

Orientador: Prof. PhD. Tiago Alessandro Espínola Ferreira
Universidade Federal Rural de Pernambuco – UFRPE – Brasil

Coorientador: Prof. PhD. Edilson Delgado-Trejos
Instituto Tecnológico Metropolitano – ITM – Colombia

Recife - PE

Dezembro, 2020

Dados Internacionais de Catalogação na Publicação
Universidade Federal Rural de Pernambuco
Sistema Integrado de Bibliotecas
Gerada automaticamente, mediante os dados fornecidos pelo(a) autor(a)

A327a Albarracín Estrada, Eva Susana

Approach for Drift Representation and Extraction in Gas Sensors Signals by Sample Entropy / Eva Susana Albarracín Estrada. - 2020.

127 f. : il.

Orientador: Tiago Alessandro Espinola Ferreira.
Coorientador: Edilson Delgado Delgado-Trejos.
Inclui referências.

Tese (Doutorado) - Universidade Federal Rural de Pernambuco, Programa de Pós-Graduação em Biometria e Estatística Aplicada, Recife, 2021.

1. nariz eletrônico. 2. deriva do sensor de gás. 3. detecção rápida. 4. entropia de amostra. 5. processamento de sinal. I. Ferreira, Tiago Alessandro Espinola, orient. II. Delgado-Trejos, Edilson Delgado, coorient. III. Título

CDD 519.5

Universidade Federal Rural de Pernambuco
Pró-reitoria de Pesquisa e Pós-graduação
Pós-Graduação em Biometria e Estatística Aplicada

Approach for Drift Representation and Extraction in Gas Sensors Signals by Sample Entropy

Eva Susana Albarracín Estrada

Tese julgada adequada para obtenção do título de Doutora em Biometria e Estatística Aplicada, defendida e aprovada por unanimidade e com distinção em 21/12/2020 pela Comissão Examinadora.

Orientador:

____
Prof. PhD. Tiago Alessandro Espínola Ferreira
DEINFO/UFRPE

Coorientador:

____
Prof. PhD. Edilson Delgado-Trejos
Instituto Tecnológico Metropolitano ITM – Colombia

Banca Examinadora:

____
Prof^{a}. PhD. Tatijana Stosic
DEINFO/UFRPE

____
Prof. PhD. Cristian Manuel Durán Acevedo
Universidad de Pamplona – Colombia

____
Prof^{a}. PhD. Adriana Xiomara Reyes Gamboa
Politécnico Colombiano Jaime Isaza Cadavid – Colombia

I dedicate this thesis to my husband for his unconditional and true support. To Sophie, because she is the life that has germinated of my life. I also want to dedicate this work to my mother and my father, who will always live in my heart and memories, although he is not physically present.

Acknowledgments

To Professor Tiago Alessandro Espínola Ferreira for his valuable academic and professional contribution as advisor of this doctoral thesis and his constant contribution to this work's development and completion. I thank him for his quality as a human being and his willingness to illuminate my path.

To Professor Edilson Delgado Trejos, who, with his experience and knowledge, has guided this research and has also been a manager and participant in the whole process carried out for the completion of this doctoral study.

To Professor Tatijana Stosic for her contribution and teachings; to Professors Glauco Estacio Gonçalves and Victor Wanderley Costa de Medeiros for always being available to give feedback and be collaborators in the development of this research; to Professor Luciana Leite de Andrade de Lima for her participation and contribution from her area to this research; to Professors Paulo José Duarte Neto and Moacyr Cunha Filho who gave me their support and guidance in their work as coordinators of the postgraduate program in Biometrics and Applied Statistics. In addition, I am grateful in general to the team of professors of the Post-Graduate Program in Biometrics and Applied Statistics of the Federal Rural University of Pernambuco.

To Marco Santos, because thanks to his kindness and excellent management of the program's secretariat, I always found support in the academic and administrative procedures.

To all those workers from the Federal Rural University of Pernambuco, they showed us the kindness and quality of human beings that exist in this region of Brazil.

To the Instituto Tecnológico Metropolitano - (ITM) for offering all the guarantees, spaces, and equipment of the Research Group on Quality, Metrology, and Production for the project's materialization described in this book.

To my friends and colleagues on the road.

Abstract

Humans and animals perceive the surrounding environment using the physiological mechanisms of perception, commonly called senses (GUERRINI *et al.*, 2017). Bio-inspired by the biological olfactory system, the development of artificial devices that combine chemical sensors array with pattern recognition techniques, commonly termed as "electronic nose" (E-Nose), have been used for recognition of Volatile Organic Compounds (VOCs). The gas sensors' response may contain some disturbances (noise and drift) composed of multiple frequencies, affecting signal processing tasks' performance. The present thesis focused on analyzing the drift behavior in signals from gas sensors used in artificial olfactory devices. For this purpose, one extensive database was used, reported in the literature as a real database with severe drift issues. An exploratory analysis was performed over that database using discrete Wavelet transform, observing the presence of drift, noise perturbance, and the existence of outliers, making it more challenging to treat that database. Additionally, it was estimated the influence of drifts based on Sample Entropy to establish the dynamics caused in the signals of E-Nose. Finally, it was generated several work scenarios using synthetic measurements generator. I was sought to explore the effect of drifts on different portions of signals from electronic nose systems, analyzing the performance of the rapid detection method for electronic nose systems using artificial data.

Keywords: electronic nose, gas sensor drift, rapid detection, signal processing, sample entropy, wavelet

Resumo

Humanos e animais percebem o ambiente circundante usando os mecanismos fisiológicos de percepção, comumente chamados de sentidos (GUERRINI et al., 2017). Bioinspirados pelo sistema olfativo biológico, o desenvolvimento de dispositivos artificiais que combinam uma matriz de sensores químicos com técnicas de reconhecimento de padrões, comumente denominados como "nariz eletrônico" (E - Nose), têm sido usados para o reconhecimento de Compostos Orgânicos Voláteis (COVs). A resposta dos sensores de gás pode conter algumas perturbações (ruído e deriva) compostas por múltiplas frequências, afetando o desempenho das tarefas de processamento de sinal. A presente tese teve como objetivo analisar o comportamento da deriva em sinais de sensores de gás usados em dispositivos olfativos artificiais. Para tanto, foi utilizado um extenso banco de dados, relatado na literatura como um banco de dados real com graves problemas de deriva. Uma análise exploratória foi realizada sobre esse banco de dados usando transformada Wavelet discreta, observando a presença de deriva, perturbação de ruído e a existência de outliers, tornando mais difícil o tratamento desse banco de dados. Adicionalmente, estimou-se a influência dos desvios com base na Entropia da Amostra para estabelecer a dinâmica causada nos sinais do E-Nose. Por fim, foram gerados diversos cenários de trabalho utilizando gerador de medidas sintéticas. Fui procurado para explorar o efeito dos desvios em diferentes partes dos sinais de sistemas de nariz eletrônico, analisando o desempenho do método de detecção rápida para sistemas de nariz eletrônico usando dados artificiais.

Palavras-chave: nariz eletrônico, deriva do sensor de gás, detecção rápida, processamento de sinal, entropia de amostra, wavelet

Contents

List of figures

List of tables

1 INTRODUCTION

During the 1980s, research on artificial olfactory systems leads to a generally accepted definition of an electronic nose as an instrument that comprises an array of heterogeneous electrochemical gas sensors with partial specificity and a pattern recognition system. However, in more recent years, the term electronic nose has been used in a broader sense to refer to gas sensors that measure the ambient gas atmosphere based on the general principle that changes in the gaseous atmosphere characteristically alter the sensor properties. It exists different sensor types for olfactory systems, and the more commonly used are fabricated with the following materials: metal oxides, conducting polymers composites, and intrinsically conducting polymers. Apart from conductive sensors, gas detection has also been done using optical sensors, surface acoustic wave sensors, gas-sensitive field-effect transistors, and quartz microbalance (QMB) sensors (LOUTFI *et al.*, 2015).

Electronic noses use various chemical sensors (arranged as sensors array) with overlapping sensitivities that detect different aromas, meaning that they are not selective to a given chemical compound. However, they are slightly more sensitive to individual chemical families such as organic solvents, fatty acids, sulfurous gases, among others. In this way, the sensors' response consists of characteristic signals for each chemical mixture, being sensitive to a wide variety of products. Once the data from the individual sensors from the array is collected, the E-Nose devices require a suitable signal processing stage to analyze and classify the aroma data. However, so far, the signal pre-processing of sensor array responses represents an essential part of most artificial olfactory applications, affecting the final response (forecasting) (LOUTFI *et al.*, 2015).

Electronic nose systems are useful in multiple applications, such as detecting toxic gas escapes, pollution factors, bombs, narcotics, diagnosis of diseases, and food quality control. They have significant advantages in the agri-food sector, among which the following stand out:

- Allow Non-destructive analysis of the product.
- Obtaining real-time results (in seconds or minutes).
- Generally, they have good portability, robustness, and low price.
- Allow easy adaptation to different quantities and varieties of products.
- Ease of use by unqualified personnel.

1.1 Object of study

A substantial restriction in the technology of the gas sensors, in addition to the limitations in selectivity and sensitivity, is presented from the sensor drifts(ZIYATDINOV, A. *et al.*, 2010). The drift can degrade the response of the system. This effect is even more considerable when the analysis is performed through multiple measurements made over long periods. The sensors' responses vary by aging in the sensing layer because of more cycles of use, even considering measures under controlled conditions.

The drift effects over the responses affect pattern recognition tasks, causing lower accuracy rates (wrong odor discrimination). The drift effects are more evident when trying to classify volatile compounds that have been introduced to the sensor system in a broad time regarding the training dataset. In this way, it is well known that the drifts are a dynamic process caused by chemical changes in the sensors, which give an unstable signal over time. Besides, samples and the operator through contamination of the instrument can also introduce drifts (ARTURSSON *et al.*, 2000). For this reason, this work focuses on analyzing the problem of gas sensors drifts and their effect on odor recognition systems from experiments based on sample entropy and the use of the rapid detection approach proposed by (RODRIGUEZ GAMBOA, Juan C.; ALBARRACIN E., Eva Susana; SILVA, Adenilton J. Da; *et al.*, 2019; RODRIGUEZ GAMBOA, Juan C *et al.*, 2021).

1.2 Objectives

1.2.1 General Objective

Analyze the drift behavior in signals from electronic nose systems based on sample entropy and the rapid detection approach.

1.2.2 Specific objectives

- Estimate the effect of drifts based on the dynamic principles of entropy to establish the dynamics caused in the signals of electronic nose systems.
- Explore the effect of drifts on different portions of signals from electronic nose systems, using synthetically generated data.
- Analyze the performance of the rapid detection method for electronic nose systems using artificial data.

2 THEORETICAL FRAMEWORK AND STATE OF THE ART

This chapter's subject begins by describing the operation and the parts that constitute this class of systems to present a general perspective of the detection and recognition systems of volatile compounds. Below, the content focuses on gas sensors, their workflow, and how they detect volatiles to sequentially generate the electrical signals necessary for odor recognition through the proper use of a data acquisition system, the software elements, and the pattern recognition system used for this specific task. From there, it delves into the topic to be dealt with in this thesis, which corresponds to drifts in chemical sensors and the effects they cause on sensor signals. Therefore, in this chapter, the existing state of the art in terms of the different ways of approaching the problem of drifts is presented, these being the construction of new sensors, the development of robust classifiers, and the treatment of drifts in the processing of the signals.

2.1 Electronic noses: general aspects and applications

"Organisms sense their surroundings in search of food, to secure themselves from predator and prey, to design territory, and choosing interesting mates via emission and detection of volatile chemical compounds" (JHA *et al.*, 2019). Bio-inspired by the olfactory system, the development of artificial devices that combine arrays of chemical sensors with pattern recognition techniques, commonly termed “electronic nose” (E-nose), have been explored for recognition and sensing of volatile organic compounds (VOCs). Its use as inexpensive chemical detectors is an emerging research area that plays a critical function by mimicking the olfactory organ. This mimic can recognize different smells that correlate with a range of fields, including environmental monitoring, disease diagnosis, public security affairs, agricultural production, food industry, and biometric applications, among others (HU *et al.*, 2018; JHA *et al.*, 2019). Intending to contextualize artificial olfaction, this chapter introduces VOCs recognition sensing by chemical sensors using E-Nose systems and its significance for VOCs recognition and sensing in different applications.

2.2 A brief history of the E-Nose systems

The term electronic nose as "instrument, which comprises an array of electronic chemical sensors with partial specificity and an appropriate pattern-recognition system, capable of recognizing simple or complex odors" was presented by (GARDNER; BARTLETT, 1994). However, as informed in that same article, the earliest work on the development of an instrument specifically to detect odors probably dates back to (MONCRIEFF, 1961), and the first electronic noses were reported by (BUCK, T.M. AND ALLEN, F.G. AND DALTON, 1965; DRAVNIEKS; TROTTER, 1965; WILKENS; HARTMAN, 1964). Afterward, another important work representing a landmark of the E-Nose was presented by (PERSAUD, Krishna; DODD, 1982). They proposed an E-nose using semiconductor transducers and reported that this device could reproducibly discriminate between a wide variety of odors without highly specific receptors. It meant an electronic nose concept as an intelligent system that comprises an array of chemical sensors for odor classification. In that way, the investigations in chemical sensing that comprise two main categories: bulk detection and trace vapor detection using chemical sensors, have been growing in recent years. From the number of published reports based on chemical sensing between 1950 and 2017 (216.521 papers), the maximum number of these reports were published between 2011 and 2017 (a total of 74,145 documents). It shows the interest and growth of chemical sensing applications in the last decade (JHA et al., 2019).

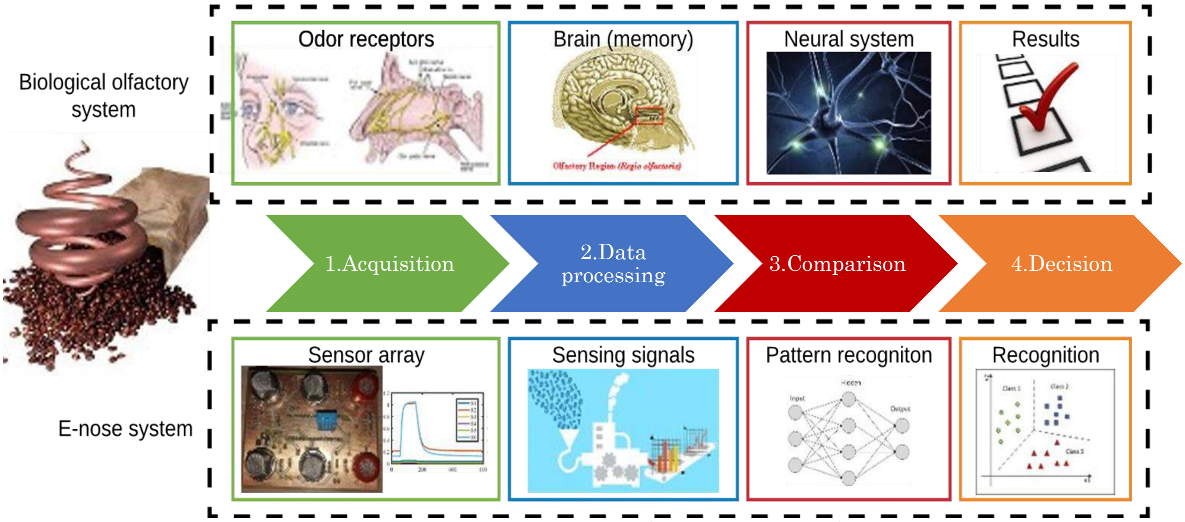
The relatively quick assessment of headspace (volume above a liquid or solid in a closed container), a quantitative representation or finger-print of gases, and cheap sensors easily integrated with the current production processes are some of the important features that have done of these systems a relevant topic for research in the chemical detection field. However, despite these features, there are still relatively few applications of E-Nose adopted in the industry. It could be attributed to difficulties in robustness, selectivity, and reproducibility of the sensors and the need for pattern recognition algorithms to cope with the complex signal analysis. Nonetheless, the use of electronic noses is rapidly expanding, and there have been notable achievements relevant to the food industry, particularly in the past few years. Furthermore, this progress coincides with an increased understanding of the biological mechanisms behind the human olfactory system. Specifically, we now have a greater understanding of the genetics behind the olfactory receptors and the relationships between an odorant's molecular property and the quality of an odor (LOUTFI *et al.*, 2015).

2.3 E-Nose system architecture

The E-Nose system tries to mimic the working mechanism of the biological olfaction system for sensing chemical compounds in different applications. Therefore, an E-Nose consists of a set of partially selective chemical sensors (sensor array), signal conditioning electronics, a pattern recognition unit equivalent to olfactory receptor neurons, olfactory bulb, and olfactory cortex of the biological olfaction system, respectively (JHA *et al.*, 2019). How is depicted in Figure 1, E-nose systems use interactive sensor-arrays (acquisition) that react to analytes on the sensitive materials' surface, accompanying the adsorption, desorption and/or reversible reaction. Meanwhile, the specific responses between the analytes and the sensors array are recorded and transformed into readable digital values (data processing), which can be computed based on statistical models (comparison) to achieve the recognition (decision) of different odors (HU *et al.*, 2018).

An E-Nose comprises different modules that work together to recognize odors. This type of instrument has at least three parts, each one with specific functions detailed below: adequacy of the gas mixture (sampling system), the gas sensor array, and the processing system. Figure 2 shows the working sequence of an artificial olfaction system, whose blocks are described in the subsequent sections.

Figure 1 – Schematic diagram of the working principle of human and artificial olfaction



Source: Adapted from (HU *et al.*, 2018)

2.3.1 Adequacy of the gas mixture (sampling system)

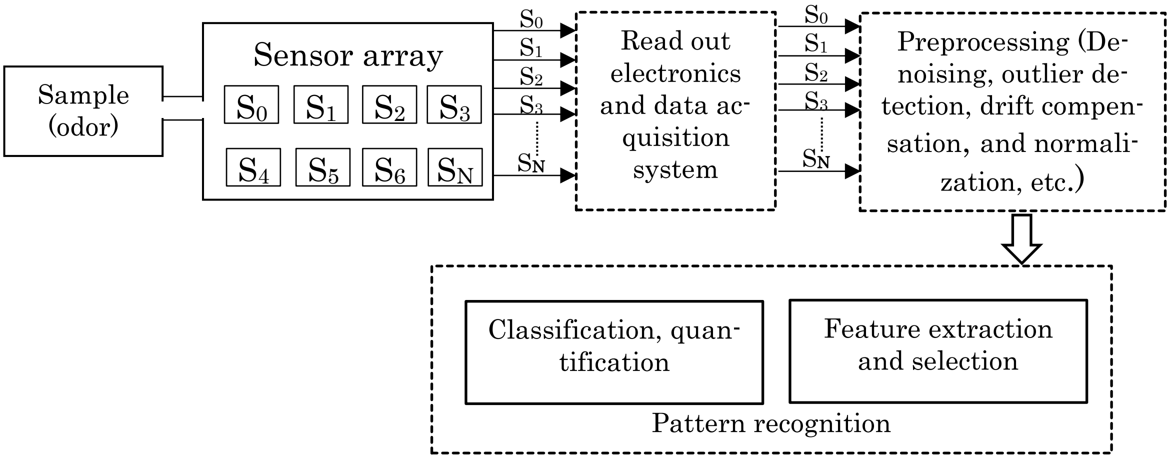
Initially, the sample is conditioned by volatile extraction methods that allow the gas to be analyzed to pass to the sensor array. The sampling system is mainly composed of a place where the sample is preserved (such as a concentration chamber), a control system, and a flow transport system (such as an air pump, mass flow controllers, etc.).

The sample used depends on the applications of the E-Nose. Some of these applications summarized by (HU *et al.*, 2018) and (ZHANG, L.; ZHANG, D., 2018) are listed below.

- **Medical care**: diagnosis of health conditions via the detection and classification of VOCs into one or a combination of body fluids. In general, exhaled breath, skin/sweat, feces, urine, saliva, breast milk, and intestinal gas are one or a combination of the secretion pathways of VOCs emission.

- **Food industry:** The VOCs concentrations in ppm emitted in the food aging, the spoilage process makes many commercial sensors usable and a powerful tool in the food industry. For instance, E-Nose systems have been used to monitor the shelf life of tomato by sensing the aromatic VOCs due to post-harvesting, respiration, fermentation, and phenolic oxidation; fungi contamination in peaches are detected through the analysis of VOCs; and other application areas for several specific foodstuffs: milk, fish and meat, wine, beverage, tea and coffee (RODRÍGUEZ; DURÁN; REYES, 2010). Moreover, the brand and/or the place of origin can be recognized by E-Nose, e.g., the tobacco types and cigarette brands can be identified by a composite polymer-based sensor.

Figure 2 – Block diagram of the E-Nose system. S_0, S_1, ..., S_N correspond to each gas sensor


Sample (odor)
Sensor array
S0 S1 S2 S3
S4 S5 S6 SN
S0
S1
S2
S3
SN
Read out electronics and data acquisition system
S0
S1
S2
S3
SN
Preprocessing (Denoising, outlier detection, drift compensation, and normalization, etc.)
Classification, quantification
Feature extraction and selection
Pattern recognition

Source: Adapted from (JHA et al., 2019)

- **Environment monitoring:** Continuously on-line/in situ detection of gas is desired for most environmental monitoring applications, e.g., the identification of toxic wastes including carbon monoxide (CO), sulfur dioxide (SO_2), ammonia, hydrogen sulfide (H_2S), ozone (O_3); air quality testing is mainly nitric oxide (NO), nitrous oxide (NO_2), carbon monoxide (CO), carbon dioxide (CO_2), sulfur dioxide (SO_2). Soil/water pollution mainly includes methane (CH_4), ammonia, NO, SO_2; the factory emission detection is mostly about toluene, H_2S, SO_2; vehicle exhaust control of CO, CO_2, NO, NO_2; indoor volatile organic compounds are mainly formaldehyde (HCHO), benzene and acetone, etc.

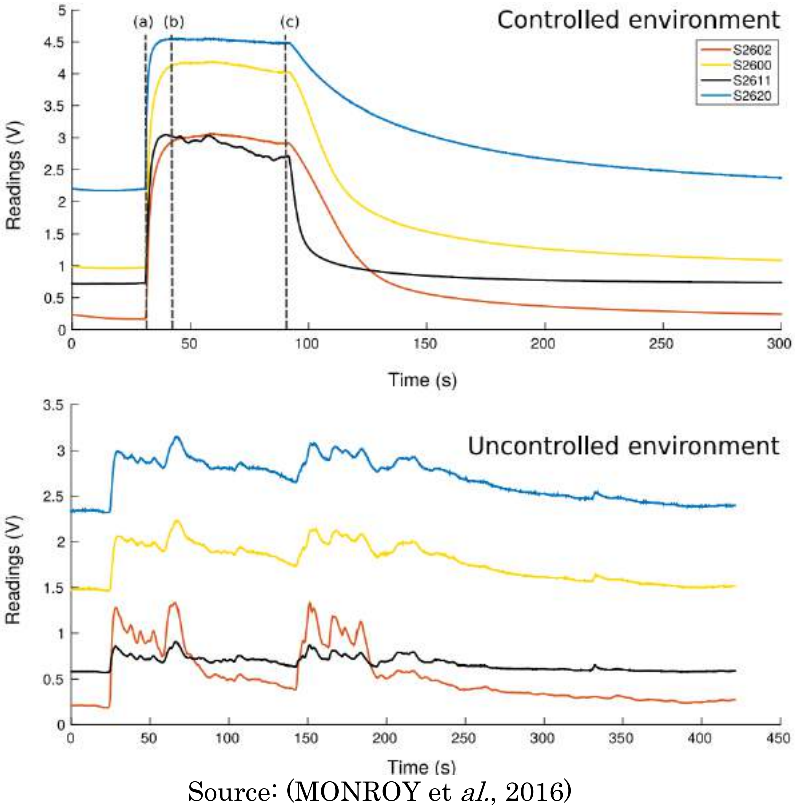
- **Public Security:** The public security, especially anti-terrorism, becomes extremely urgent. The applications in this field mainly focus on detecting explosives and/or nerve agents. Normally, the vapor pressure of explosives, e.g., 2-methyl-1,3,5-trinitrobenzene (TNT), 1,3,5-trinitroperhydro-1,3,5-triazine (RDX), octahydro-1,3,5,7-tetranitro-1,3,5,7-tetrazocine (HMX) are 9 ppb, 4.9 ppt, and 0.25 ppt, respectively.

- **Agricultural Production:** The E-Nose system in agriculture is used for monitoring production processes, disease detection, identify insect infestations in their first stage, and soil/water pollution.

Traditionally, the E-Nose applications have been performed in highly tight-controlled sensing test chambers that isolate the chemical analyte (odor) from its natural, predominantly complex environmental condition. Such isolation enables the chemical sensory system to exhibit chemical signatures that are, to a very large extent, specific to both, the kind of sensory elements used and the chemical analyte being monitored. Controlled sensing test chambers ensure a strict handle over some critical sensing conditions, including environmental temperature, pressure, and ambient flow. (VERGARA *et al.*, 2013). On the other hand, applications such as the detection of toxic chemicals in human environments, or the localization of gas sources by robots, demand a continuous classification of volatile substances which cannot be addressed with the traditional pulse-like excitation. In contrast, when no control is performed over the environment (including the gas-emitting source), signals look much more random and chaotic, being difficult to identify distinctive behaviors or patterns (MONROY *et al.*, 2016).

Figure 3 shows the readings of an E-Nose composed of an array of four metal oxide- semiconductor (MOX) gas sensors when exposed to an ethanol gas source under controlled and uncontrolled environmental conditions. As can be seen, when controlling the environment, reproducible patterns are obtained, and key-points can be easily identified in the response: (a) start of the volatile exposition, (b) end of the transient response and start of the steady phase, and (c) end of the volatile exposition and start of the recovery period. In contrast, when no control is performed over the environment (including the gas-emitting source), signals look much more random and chaotic, being difficult to identify distinctive behaviors or patterns (MONROY *et al.*, 2016). To facilitate understanding, we will continue with the description of an E-Nose system with controlled environmental conditions.

Figure 3 – Readings of an E-Nose exposed to an ethanol gas source under (top) well-controlled environment and measurement, and (bottom) when no control is performed during the measurements



Source: (MONROY et *al.*, 2016)

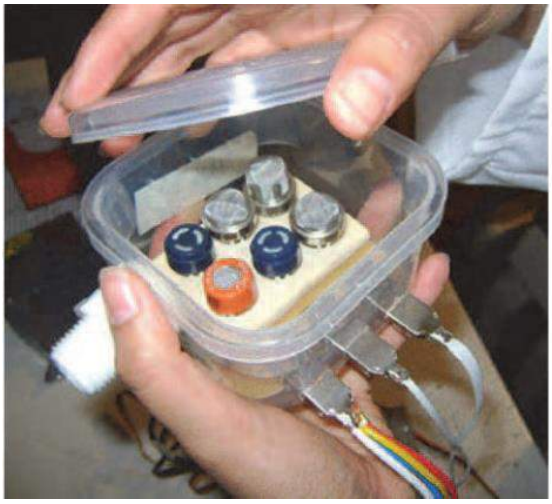
2.3.2 Matrix or array of gas sensors

An odor recognition system has as its main element a matrix of gas sensors responsible for transducing the concentration of volatiles in changes in their resistance. Conveniently, said matrix is in a special chamber or compartment in which the specific conditions for its correct operation are guaranteed. Principally, adequate insulation must be ensured to prevent contaminants from being introduced, and the appropriate pressure and temperature must also be maintained.

Another advantage of using a sensor camera is facilitating the measurement process because the volatiles will be in greater concentration and have more contact with the active element of the sensors, which allows a better and faster response of the same. It has also been found experimentally that the more airtight the sensor chamber, the better the mentioned advantages are exploited. Figure 4 shows the photograph of a sensor chamber and different kinds of chemical sensors. It is important to mention that gas sensor arrays usually use sensors of the same type but different references (Example: TGS822, TGS821, TGS813, etc.) to obtain a more significant overlap between the signals and to facilitate the classification tasks and odor detection.

Additional to this module is the volatile transport system, which conditions the operation and allows the measurement and purging processes of the sensors to be carried out. It is a system that is responsible for transporting to the sensor chamber the volatiles released by the sample or element to be analyzed. Sometimes the odor sample is injected into the sensor chamber manually, with the consequent problems of error and slowness that this implies. At other times, an automatic system is responsible for transporting the odorous or volatile molecules, extracting them from the area where the sample is located through the injection of some type of gas or air until they are taken to the sensor chamber. Also, the electronic smell systems mostly have some cleaning mechanism of the sensor chamber. Successive measurements are made starting from the same initial conditions, and the repeatability of the results is guaranteed.

Figure 4 – Sensing system: concentration chamber



Source: (DURÁN; VELÁSQUEZ; GUALDRON, 2012)

A chemical sensor generates a transient response to a target chemical vapor, which is affected by several parameters, including target and interfering chemicals, the concentration of target chemical, odor flow rate, temperature, pressure, humidity, types of chemical sensor and chemical interface, etc. In chemical vapor sensing applications, steady S_{ij} (response of j-th sensor in the array for i-th chemical analyte) and transient response $S_{ij}(t)$ are useful in chemical identity information extraction. The response vector of a sensor array (set of n sensors) to the i-th chemical compound can be represented as $\mathbf{S}_\mathrm{i} = (S_{i1}, S_{i2}, \ldots\ldots, S_{ij}, \ldots .S_{in})$. Likewise, sensor array response to m different chemical compounds can be represented by a response matrix **S1** as in the equation (2-1)(JHA *et al.*, 2019).

$$\mathbf{S1} = \begin{bmatrix} S_{11} & S_{12} \ldots & \ldots S_{1n} \\ S_{21} & S_{22} \ldots & \ldots S_{2n} \\ \vdots & \vdots & \vdots \\ S_{m1} & S_{m2} \ldots & \ldots S_{mn} \end{bmatrix} \tag{2-1}$$

2.3.3 Processing System

In most cases, the processing system consists of a computer with the appropriate software to track the data obtained from the sensors. Pre-processing techniques are applied to those data to extract the static parameters of the measures and to reduce the amount of information to be analyzed. Multivariate analysis techniques are employed, such as Principal Components Analysis (PCA) and pattern recognition as Networks Artificial neurons (RNA), Support Vectors Machines (SVM), among others, to perform tasks such as classification, discrimination, prediction, quantification of samples according to their organoleptic characteristics.

Table 1 – The most used supervised and unsupervised classification methods

Supervised methods	Unsupervised methods
Back-Propagation Neural Networks - BPNN	Self-Organizing Map (SOM)
SVM	Hierarchical Cluster Analysis (HCA)
K-nearest neighbor - KNN	K-means clustering
Naïve Bayes - NB	Fuzzy clustering
Linear Discriminant Analysis - LDA	
Adaptive Resonance Theory Map - ARTMAP	

Source: Own

Besides visual discrimination of chemical compounds in feature space, selected features are used to input classification methods for their class identification (qualitative recognition) and the input of quantification methods for concentration estimation. The classification methods are categorized into supervised, unsupervised, and reinforcement strategies. The most used supervised classification methods and unsupervised methods are in Table 1. Unsupervised methods are used to group the chemical compounds using the sensor array response or their extracted features. The third category of methods, based on reinforcement learning, doesn't require the obvious sensor response to a chemical compound and their class information but explores sensor response space in some tunable way, such as greedy search. Better class recognition efficiency is achieved with the SVM classifier compared to the rest three classification methods (RODRIGUEZ GAMBOA, Juan C *et al.*, 2021). SVM method is extensively used for classification, feature extraction, clustering, outlier removal, and regression tasks in various disciplines. It is used primarily in class recognition of chemical compounds by the sensor array response processing using PCA, KPCA extracted features as input (JHA *et al.*, 2019).

2.4 Gas sensors

There are different types of gas sensors for use in odor recognition systems. The most used are Chemoresistive gas sensor based on n-type metal oxide nanostructures MOX, FET-Based gas sensor, solid state electrochemical gas sensor (SSES), Quartz Crystal Microbalance (QCM) based gas sensor, and others devices based on surface ionization, optical (photoluminescence), magnetic (magneto-optical Kerr effect), and transduction mechanisms (HU *et al.*, 2018; PONZONI *et al.*, 2017). A brief comparison to the first four types of sensors has been given according to aspects of sensitivity, selectivity, speed, cost, size (Table 2).

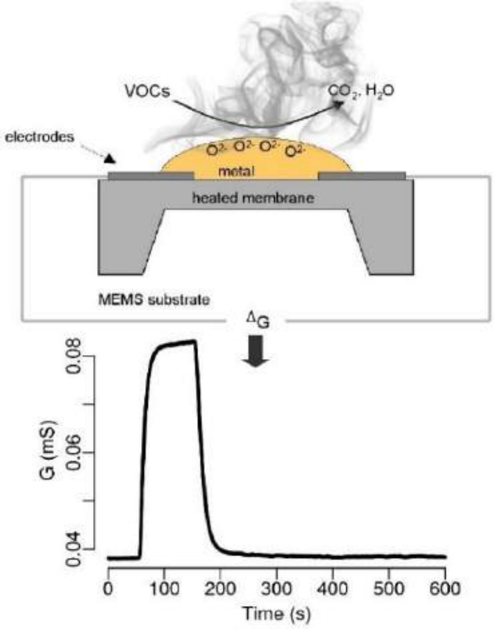
Table 2 – Comparisons of different gas sensing technologies

Gas sensor types	Sensitivity	Selectivity	Speed	Cost	Size
Chemoresistor	High	Medium	Fast	Low	Small
FET	High	Medium	Fast	Medium	Small
SSES	High	Good	Fast	Low	Large
QCM	High	Poor	Medium	High	Medium

Source: (HU *et al.*, 2018)

This thesis focuses on MOX gas sensors. These sensors are devices that consist of two main parts, the first is an active element which changes its physical or chemical properties in the presence of that which it detects, and the second part is a transducer, which converts the changes in the properties of the active element into an electrical signal. These sensors typically have a selective membrane, preventing the passage of particles or unwanted material, acting as a first noise filter. Figure 5 shows a simplified diagram of a device of this type, in which the main parts of a gas sensor and the nature of the inputs and outputs can be seen (RODRÍGUEZ-GAMBOA; ALBARRACÍN-ESTRADA; DELGADO-TREJOS, 2011).

Figure 5 – Top: Simplified scheme of a chemical sensor for the detection of VOCs. Bottom: Signal of a one-single sensor in conductance (G) units expressed in milliSiemens (mS)

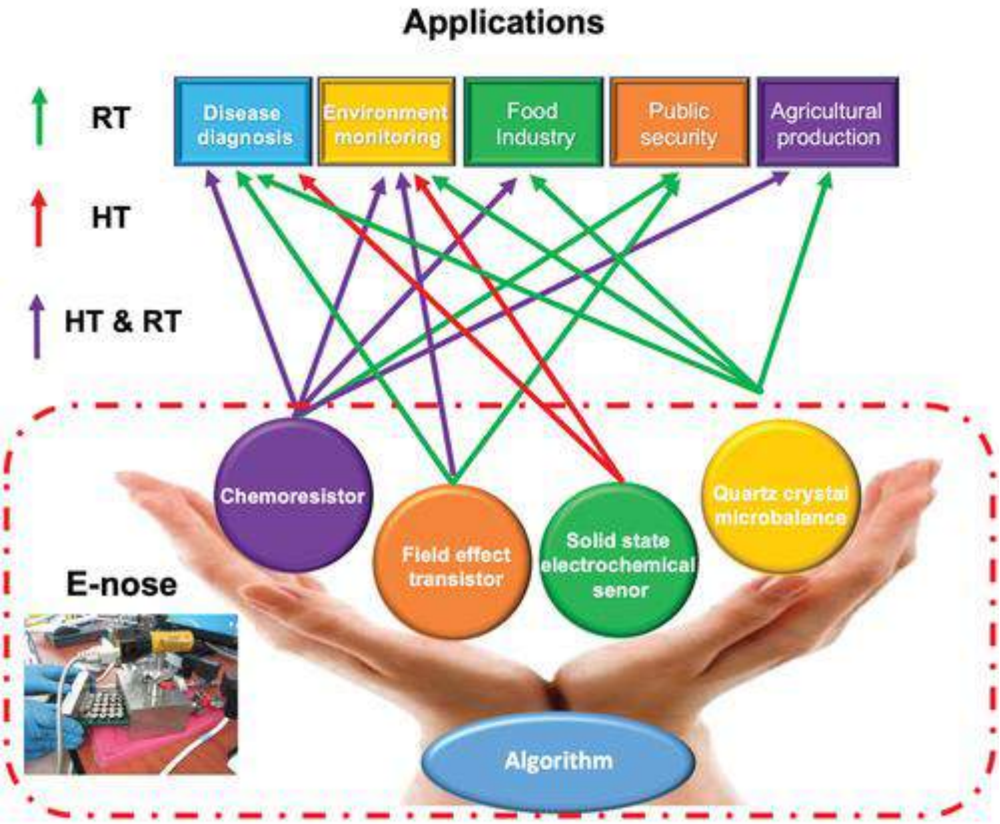


Source: Own

Figure 6 depicts how these sensors work in special applications at different temperature ranges (green arrow: room temperature, red arrow: high temperature, purple arrow: both room and high temperature). For example, chemoresistors and FET can work at both room and high temperature. Solid-state electrochemical gas sensor can work at high temperatures only with the limitation of the solid electrolyte. QCM always works at room temperature. From the viewpoint of applications, all the fields prefer to work at room temperature; however, disease diagnosis, environmental monitoring, and agricultural production also accept working at high temperatures (HU *et al.*, 2018). Given the characteristics of chemoresistors (MOX sensors), these are the most widely used and will focus on this work. The follows paragraphs are centered on this kind of sensor.

After being acquired and stored, the sensors' signals are treated by methods of extraction of parameters and pre-processing of data. The technique of extracting parameters is fundamental, especially when using MOX sensors (RODRÍGUEZ-MÉNDEZ *et al.*, 2016). These base their operation on the change of conductivity experienced by the material or active layer of the sensor in the presence of reducing gases and/or oxidants. The conductivity change experiences transients that lead the active layer of the sensor from a resting situation to a conductance that depends on the volatile and its concentration (PONZONI *et al.*, 2017).

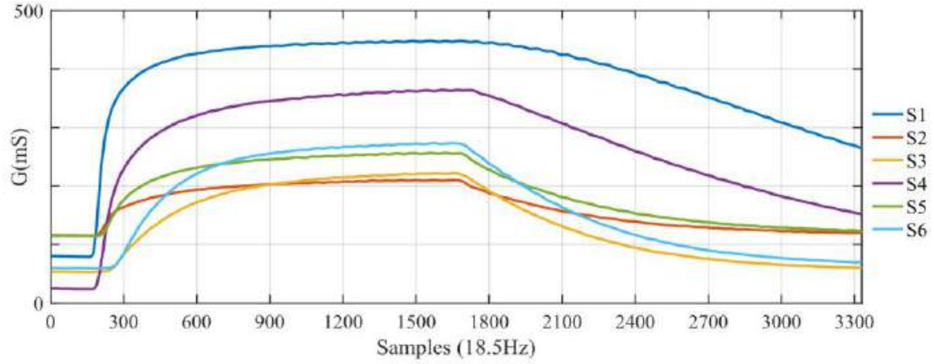
Figure 6 – Network diagram of E-Nose indicates the E-Nose technologies, working conditions, and applications



Source: (HU *et al.*, 2018)

The type of signals obtained from a 6-sensor matrix is shown in Figure 7, where the response of the sensors to a wine measurement is depicted.

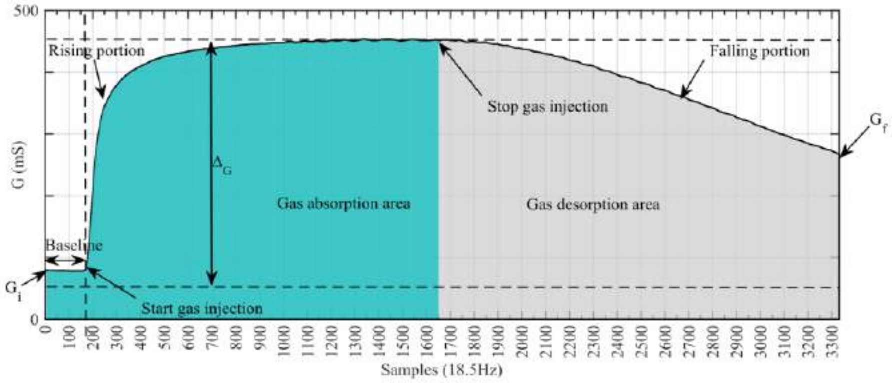
Figure 7 – Wine measurement acquired with O-NOSE; S1, S2..., S6: gas sensor outputs in conductance units G



Source: (RODRIGUEZ GAMBOA, Juan C.; ALBARRACIN E., Eva Susana; SILVA, Adenilton J. Da; *et al.*, 2019)

Figure 8 shows the stages in the measurement (baseline, gas absorption, and gas desorption). Wine measurement acquired with O-NOSE; S1, S2, ..., S6: gas sensor outputs in conductance units G.

Figure 8 – Output of a gas sensor. G_i: initial conductance value, G_f: final conductance value, Δ_G: maximal conductance change concerning the baseline



Source: (RODRIGUEZ GAMBOA, Juan C.; ALBARRACIN E., Eva Susana; SILVA, Adenilton J. Da; *et al.*, 2019)

2.5 Drift in gas sensors

Drift in devices with chemical sensor arrays has a rather complex and inevitable effect. Different sources can generate a drift. The aging of the sensor and the device's poisoning stand out, which is directly reflected by the change in the chemical layer for the volatile detection (reorganization of sensor material and contamination). Also implicit in the experimental operation that includes the thermal and memory effects of the sensors, the changes in the environment, and the appearance of other signals caused by the noise of the system (ZIYATDINOV, A. *et al.*, 2010)

2.6 The concept of drift

The response of a gas sensor contains not only its true signal but also some disturbances. These, in turn, are composed of multiple frequencies, and all of them affect the signal. The part that corresponds to high frequency is called noise, and the part that is made up of low frequency is often known as drift, which can be seen as a gradual change over time in sensor response under constant conditions. Drift is a dynamic process caused by chemical changes in the sensors, specifically in the active layer of the sensors (ARTURSSON *et al.*, 2000).

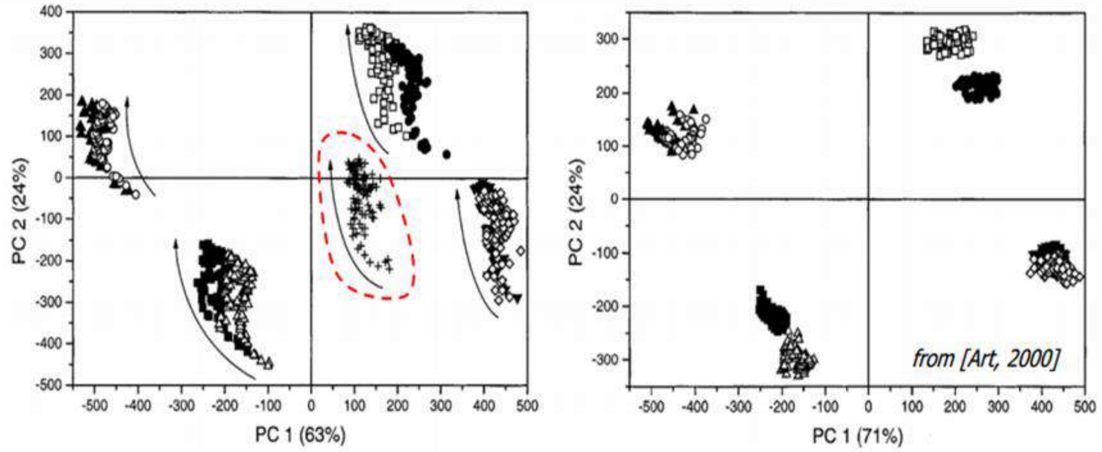
2.6.1 Drift Problems

A limitation of current E-nose built with chemical sensors is the drifts inherent in the signals, causing a slow random variation of the sensor response over time when exposed to some gases under controlled conditions.

The drift effect can affect the sensor baseline part when it is additive and the sensitivity when it is multiplicative. A consequence of the drift influence is that the previous learning of the patterns of the signals delivered by the sensors becomes obsolete over time. Consequently, systems lose the ability to identify already recognized odors. The most effective means of drift compensation is periodic recalibration with a reference gas that is chemically stable and highly correlated with the target analytes in terms of sensor performance. In this way, the response of the sensor matrix for the calibration gas can be directly subtracted from the analytes' response. Thus, a temporal drift model is deduced for each sensor or the sensor matrix (GUTIERREZ-OSUNA, 2002).

To understand the effects of drifts, refer to Figure 9, where the main components of the PCA analysis of the data are plotted. It is observed in the left image how the drift causes the resulting data from the sensor array to present changes visualized as data with greater dispersion and less separation between classes. The image on the right-hand side represents the main components of the data to which drift correction has been applied, using the component correction technique.

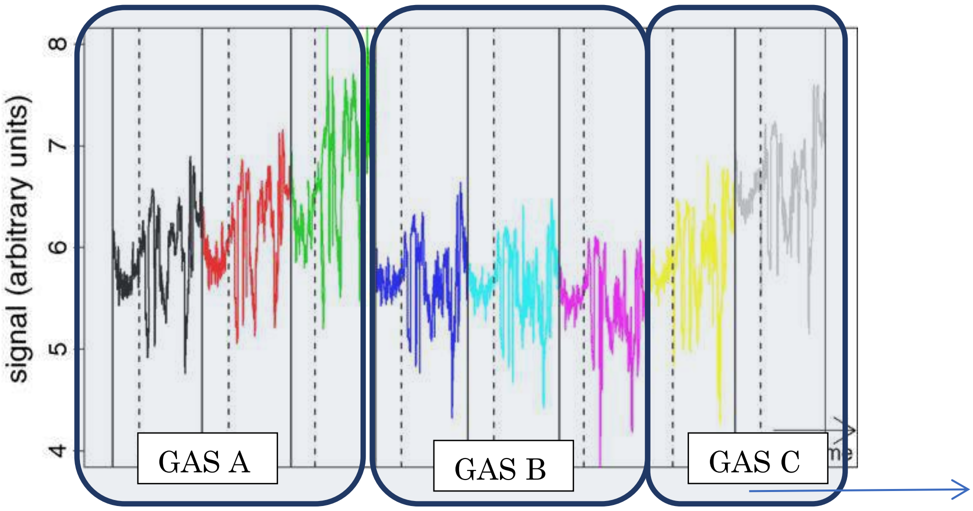
Figure 9 – Projection of the first two Principal Components of the response of a set of sensors to the presence of different gas mixtures. Left before and right after offsetting drift



Source: (GUTIERREZ-OSUNA, 2002)

This same effect can be observed if the response signal of the sensors is analyzed on the time axis, as seen in Figure 10. This Figure was adapted from the work presented by (ZIYATDINOV, A. *et al.*, 2010), to describe the strong influence of drifts on the signals delivered by the sensors. The figure shows the steady-state signals of a sensor in 7 months when the sensor is subjected to three kinds of gases with different concentrations.

Figure 10 – Analysis of the behavior over time of the response signals of a sensor subjected to the presence of three gases with different concentrations, influenced by drifts



Source: (ZIYATDINOV, A. *et al.*, 2010)

The peaks of the signals indicate short-term drifts caused by some temporary changes, such as the heating of the sensors. On the other hand, long-term drifts can be observed in changes in the baseline of similar signals for all classes.

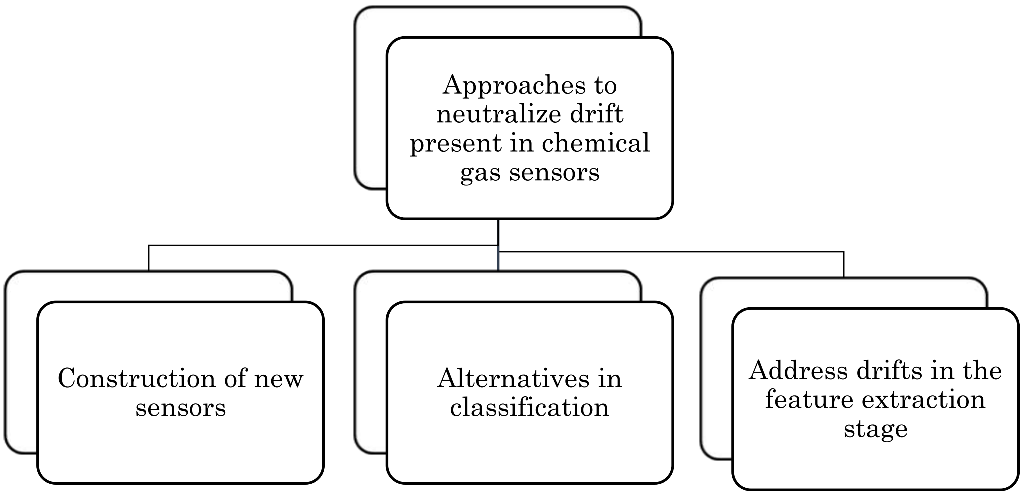
Another important observation of this graph corresponds to the response signals of the sensor in the presence of propanoic acid. These have a more stable behavior over time and are less prone to the effect of drifts. However, it is observable how the sensor response changes over time. In this case, the period analyzed was only seven months, enough time to require recalibration of the sensors.

As reflected in the previous analysis, the behavior of the data causes the classification system resulting from pattern recognition to become obsolete after a certain time.

2.6.2 Ways to address the drift problem

In the literature, there are mainly three different approaches to mitigate the effects of drift (see Figure 11). The first of them seeks to improve the physical part, designing and building new sensors. The second approach aims to correct drift from the processing stage, applying multivariate statistical analysis techniques to achieve an effective representation of drifts and thus be able to eliminate them from the system. Finally, the third points to the development of the classification approach.

Figure 11 – Main approaches to mitigate the effects of drift


Approaches to neutralize drift present in chemical gas sensors
Construction of new sensors
Alternatives in classification
Address drifts in the feature extraction stage

Source: Own

This doctoral thesis restricts the analysis of the drift problem to the approach that focuses on the classification stage, considering that currently working with the raw data of the signals of artificial smell systems is not a limitation. For this reason, the rapid detection method is proposed to analyze the response of the classifier to different drift scenarios in the data. Drift analysis and its effect from sample entropy are also addressed to explain the conditions that can trigger greater drift in data from artificial smell systems.

3 MATERIALS

Two databases were used. The first one corresponds to a database with experimental measurements from the University of California. The second is a set of synthetic data in different work scenarios generated with the *chemosensors* package in R. The following sections detail the characteristics of the databases used.

3.1 Database: Gas Sensor Array Drift Dataset at Different Concentrations

We use the Gas Sensor Array Drift Dataset at Different Concentrations (FONOLLOSA; RODRÍGUEZ-LUJÁN; HUERTA, Ramón, 2015) released in the UCI Machine Learning Repository. This dataset provides 13910 measurements from 16 chemical sensors exposed to six gases at different concentration levels collected over three years, being suitable to tackle a variety of challenges in chemical sensing such as sensor drift, sensor failure, or system calibration, as well discriminatory and quantifying issues. The artificial olfactory system used to generate this dataset had a chemical detection platform with four types of sensors tagged as TGS2600, TGS2602, TGS2610, TGS2620 (four of each type), commercially available metal-oxide gas sensors manufactured and commercialized by Figaro Inc. The odor identity and concentration values in parts-per-million by volume (ppmv) are listed in Table 3. The data distribution over 36 months is shown in Table 4, where it is detailed the dataset organization into ten batches with the number of measurements per class (VERGARA *et al.*, 2012)

Table 5 complements the above with the total of examples per month and batch for each gas. The dotted line in this table highlights the months in which do not make measurements.

Table 3 – Analytes and concentrations in the dataset

Analytes	Concentrations in ppmv
Ammonia	50, 60, 70, 75, 80, 90, 100, 110, 120, 125, 130, 140, 150, 160, 170, 175, 180, 190, 200, 210, 220, 225, 230, 240, 250, 260, 270, 275, 280, 290, 300, 350, 400, 450, 500, 600, 700, 750, 800, 900, 950, 1000
Acetaldehyde	5, 10, 13, 20, 25, 30, 35, 40, 45, 50, 60, 70, 75, 80, 90, 100, 120, 125, 130, 140, 150, 160, 170, 175, 180, 190, 200, 210, 220, 225, 230, 240, 250, 275, 300, 500
Acetone	12, 25, 38, 50, 60, 62, 70, 75, 80, 88, 90, 100, 110, 120, 125, 130, 140, 150, 170, 175, 180, 190, 200, 210, 220, 225, 230, 240, 250, 260, 270, 275, 280, 290, 300, 350, 400, 450, 500, 1000
Ethylene	10, 20, 25, 30, 35, 40, 50, 60, 70, 75, 90, 100, 110, 120, 125, 130, 140, 150, 160, 170, 175, 180, 190, 200, 210, 220, 225, 230, 240, 250, 275, 300
Ethanol	10, 20, 25, 30, 40, 50, 60, 70, 75, 80, 90, 100, 110, 120, 125, 130, 140, 150, 160, 170, 175, 180, 190, 200, 210, 220, 225, 230, 240, 250, 275, 500, 600
Toluene	10, 15, 20, 25, 30, 35, 40, 45, 50, 55, 60, 65,70, 75, 80, 85, 90, 95, 100

Source: (FONOLLOSA; RODRÍGUEZ-LUJÁN; HUERTA, Ramón, 2015)

Table 4 – Data distribution over 36 months

Months	Batch	Number of samples					
		Ethanol	Ethylene	Ammonia	Acetaldehyde	Acetone	Toluene
		Gas 1	Gas 2	Gas 3	Gas 4	Gas 5	Gas 6
1, 2	1	90	98	83	30	70	74
3, 4, 8, 9, 10	2	164	334	100	109	532	5
11, 12, 13	3	365	490	216	240	275	0
14, 15	4	64	43	12	30	12	0
16	5	28	40	20	46	63	0
17, 18, 19, 20	6	514	574	110	29	606	467
21	7	649	662	360	744	630	568
22, 23	8	30	30	40	33	143	18
24, 30	9	61	55	100	75	78	101
36	10	600	600	600	600	600	600
	Total	2565	2926	1641	1936	3009	1833

Source: (FONOLLOSA; RODRÍGUEZ-LUJÁN; HUERTA, Ramón, 2015)

The authors of dataset related in (FONOLLOSA; RODRÍGUEZ-LUJÁN; HUERTA, Ramón, 2015) that the data was collected ensuring a sufficient number of experiments in each batch, as uniformly distributed as possible. In addition, they explain that a few measurements, mainly in batch 7, appear at lower concentration levels than detailed in Table 3. This concentration mismatch is due to some experimental error, but they decided to include those samples in the dataset for the sake of completeness.

Table 5 – Dataset details

Batch	Months	Number of samples							
		Gas 1	Gas 2	Gas 3	Gas 4	Gas 5	Gas 6	per month	per batch
1	1	84	88	76	0	0	0	248	
	2	6	10	7	30	70	74	197	445
2	3	70	140	0	0	7	0	217	
	4	82	170	0	4	0	5	261	
	8	0	20	0	0	0	0	20	
	9	11	4	0	0	0	0	15	
	10	1	0	100	105	525	0	731	1244
3	11	360	146	0	0	0	0	506	
	12	0	334	0	192	0	0	526	
	13	5	10	216	48	275	0	554	1586
4	14	52	43	0	18	0	0	113	
	15	12	0	12	12	12	0	48	161
5	16	28	40	20	46	63	0	197	197
6	17	0	20	0	0	0	0	20	
	18	0	3	0	0	0	0	3	
	19	264	100	110	29	140	9	652	
	20	250	451	0	0	466	458	1625	2300
7	21	649	662	360	744	630	568	3613	3613
8	22	0	0	25	15	123	0	163	
	23	30	30	15	18	20	18	131	294
9	24	0	0	0	25	28	1	54	
	30	61	55	100	50	50	100	416	470
10	36	600	600	600	600	600	600	3600	3600

Source: (FONOLLOSA; RODRÍGUEZ-LUJÁN; HUERTA, Ramón, 2015)

The measurement procedure to generate the dataset related in (FONOLLOSA; RODRÍGUEZ-LUJÁN; HUERTA, Ramón, 2015; VERGARA *et al.*, 2012) consisting of three steps. First, it was circulated synthetic dry air (10% R.H.) through the sensing chamber for 50 s to stabilize the sensors and measure the baseline of the sensor response. Second, it was randomly added one of the analytes of interest to the carrier gas and made it circulate through the sensor chamber during 100 s. Finally, it was re-circulated clean dry air for the subsequent 200 s to acquire the sensors' recovery and have the system ready for a new measurement. The dynamic response of each sensor was recorded at a sample rate of 100 Hz. Hence, each measurement produced a 16-channel time series sequence. The channels were paired with the sensors to acquire sensors' responses. The order of the sensors in the dataset is as follows (CH0-CH15): TGS2602; TGS2602; TGS2600; TGS2600; TGS2610; TGS2610; TGS2620; TGS2620; TGS2602; TGS2602; TGS2600; TGS2600; TGS2610; TGS2610; TGS2620; TGS2620.

Each 16-channel time series acquired in each measurement were represented in the database by an aggregate of features reflecting the dynamic processes occurring at the sensor surface in reaction to the chemical substance being evaluated. In particular, two distinct types of features is referred to in (FONOLLOSA; RODRÍGUEZ-LUJÁN; HUERTA, Ramón, 2015) for the creation of this dataset: (i) the so-called steady-state feature $\Delta R = max_k r[k] - min_k r[k]$, defined as the maximal resistance change with respect to the baseline and its ΔR normalized version ($\|\Delta R\| = (max_k r[k] - min_k r[k])/min_k r[k]$) expressed by the ratio of the maximal resistance and the baseline values, where $r[k]$ is the time profile of sensor resistance, k is the discrete time indexing the recording interval $[0, T]$ when the chemical vapor is present in the test chamber. And (ii), an aggregate of features reflecting the sensor dynamics of the increasing/decaying transient portion of the sensor response during the entire measurement that converts the transient portion of the sensor response into a real scalar by estimating the maximum/minimum value y[k] for the rising/decaying portion of the exponential moving average of the sensor response:

$$y[k] = (1-\alpha)y[k-1] + \alpha(x[k] - x[k-1]) \quad (3\text{-}1)$$

where $[k = 1,2,\ldots,T]$, $y[0]$ its initial condition, set to zero (y$[0] = 0$, and the scalar $\alpha(\alpha \in \{0,1\})$ being a smoothing parameter of the operator such as was defined in (VERGARA *et al.*, 2012). The corresponding authors of the dataset set three different values for $\alpha(\alpha = 0.1,\ \alpha = 0.01, and\ \alpha = 0.001)$ to obtain three different features. They start from the pre-recorded rising portion of the sensor response and three additional features with the same α values for the decaying portion of the sensor response, covering the entire sensor response using the exponential moving average (ema$_\alpha$). Consequently, each 16-channel time series acquired in each measurement was represented in the database as a transformation mapping the sensor response to a lower dimension space preserving the most meaningful portion of the information contained in the original sensor signal. It represented each sensor signal during each measurement by 8-features, and given that the detection platform had 16-channel, measurement results in a 128-dimensional feature vector. The steady-state features and transient that represent the sensors' time series data are summarized in Table 6.

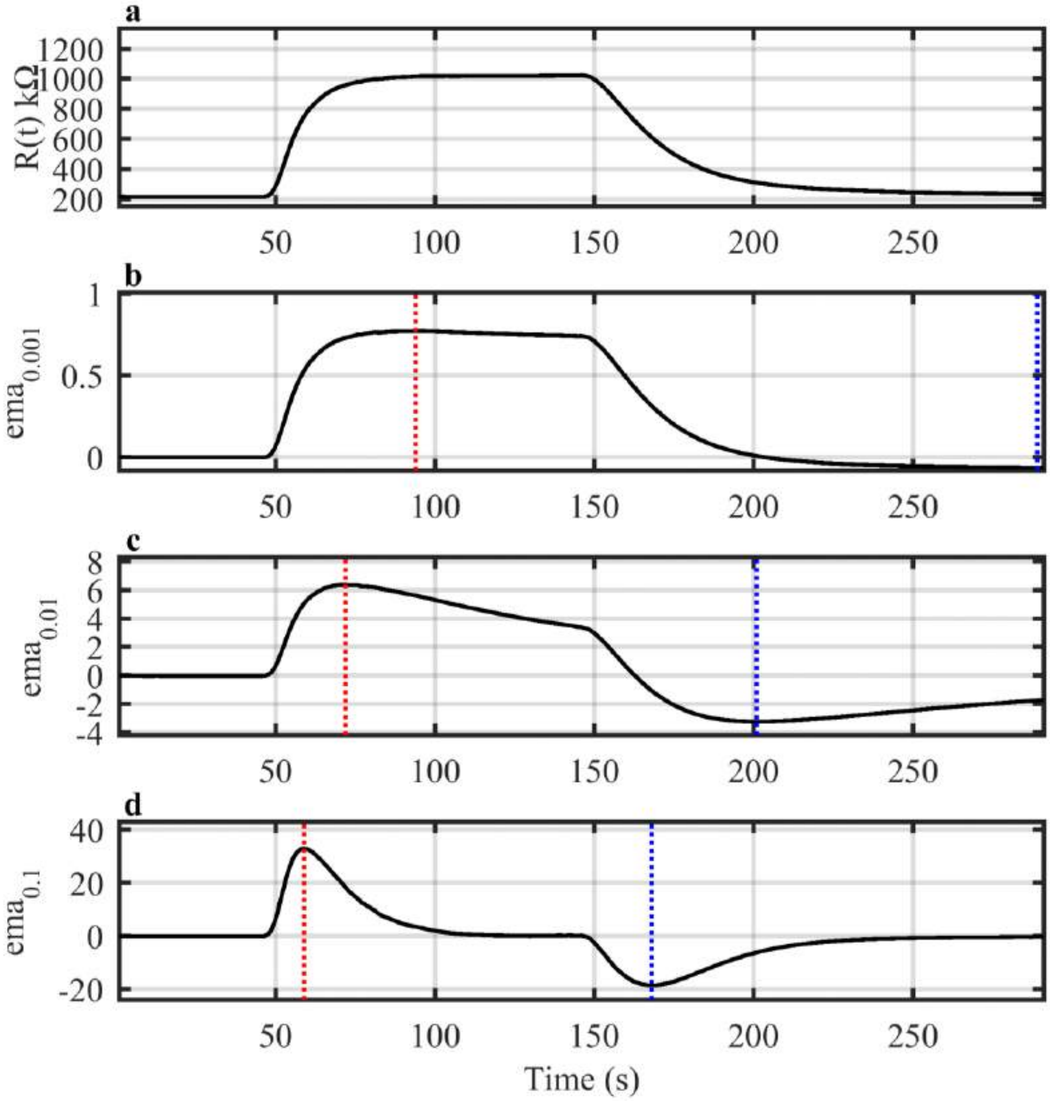
Table 6 – Features extracted from the time series data

Steady-state features	Transient features	
	Rising portion	Decaying portion
ΔR	$max_k\ ema_{\alpha=0.001}(r[k])$	$min_k\ ema_{\alpha=0.001}(r[k])$
$\|\Delta R\|$	$max_k\ ema_{\alpha=0.01}(r[k])$	$min_k\ ema_{\alpha=0.01}(r[k])$
	$max_k\ ema_{\alpha=0.1}(r[k])$	$min_k\ ema_{\alpha=0.1}(r[k])$

Source: (VERGARA *et al.*, 2012)

We plot, in Figure 12 (b)-(d), an example of the exponential moving average of the sensor response, calculated by the equation (3-1), with α=0.1, α=0.01, and α=0.001, for a gas sensor signal. It is possible to identify the six transient features (two per each ema_{α}) extracted from a sensor signal response and stored in the database. The ΔR steady-state feature and the concerned $\|\Delta R\|$ were calculated from the curve response values of the sensor, panel (a).

Figure 12 – Typical response of a chemical gas sensor and the corresponding exponential moving average signals. Panel (a), sensor curve response that shows the three phases of a measurement: baseline measurement (made with pure air), test gas measurement (when the chemical analyte is injected, in gas form, to the test chamber), and the recovery phase (during which the sensor again is exposed to pure air). Panels (b)-(d), exponential moving average of the sensor response for α = 0.1, α = 0.01, and α = 0.001. The dotted red lines signalize the maximum values of the curve ($max_k r[k]$), and the dotted blue lines signalize the minimum values of the curve ($min_k r[k]$)



Source: Own

The database authors related in (FONOLLOSA; RODRÍGUEZ-LUJÁN; HUERTA, Ramón, 2015) some experimental error in a few measurements, mainly in batch 7, in which appear lower concentration levels than detailed in Table 3, but in terms of guaranteeing the data completeness, they decided to include those samples in the dataset.

3.2 Synthetic data

It consists of an array of virtual sensors generated by a package in **R** language, called ***chemosensors***, which is freely accessible. The workflow to generate the synthetic data from ***chemosensors*** consists of creating the work scenario, establishing the parameters of the array of sensors and finally generating the characteristics matrix and the vector of classes or labels. This synthetic data generator package takes as reference the experimental measurements of UNIMAN belonging to the University of Manchester in the United Kingdom, which contains 3925 samples taken in 10 months with 17 sensors and using three analytes, ammonia, propanoic acid and n-butanol at different levels of concentration (ZIYATDINOV, Andrey; PERERA-LLUNA, 2014).

The synthetic data generator package is useful for comparing statistical pattern recognition in artificial smell. It belongs to the *Neurochem* project and contains virtual chemical sensors. The synthetic data is used because of the importance of performing drift analysis, knowing previously the drift components added to the data to perform the exploratory analysis of the impact caused by the drift in the odor recognition systems.

Table 7 presents the list of parameters used in *chemosensors*. This work highlights the use of the *dsd* and *ndcomp* parameters to analyze the effect of drift in the data.

Table 7 – Description of the basic parameters of *SensorsArray* class necessary to generate a virtual sensor array

Parameter	Default value	Range of values	Short description
A	1:2	1,2,... 17	Sensor type
nsensors	2	1,2,... ..	Number of sensors
gnames	3	1,2,3	Number of gases
concUnits	Percentage		Concentration units
alpha	2,25	> 0	Non-linearity of the sensor
beta	2	≥0	Sensor diversity
csd	0,1	≥0	Noise concentration
ssd	0,1	≥0	Sensor noise
dsd	0,1	≥0	Drift noise
ndcomp	1	1,2,3	Number of drift components
ndvar	0,86	(0,1]	Importance of drift components
tunit	1	1,2.....	Gas pulse length

Source: (ZIYATDINOV, Andrey; PERERA-LLUNA, 2014)

3.3 Electronic Nose Dataset for Detection of Wine Spoilage thresholds

This database was collected during the doctorate, and we use it to test and propose the rapid detection approach for E-Nose. The recorded database corresponds to time series obtained for an application of wine quality detection focused on spoilage thresholds, containing 235 recorded measurements of wines divided into three groups and labeled as high quality (HQ), average quality (AQ), and low quality (LQ), in addition to 65 ethanol measurements, which was collected using an electronic nose based on Metal Oxide Semiconductor (MOS) gas sensors, self-developed at the Universidade Federal Rural de Pernambuco (Brazil). A data paper with the details was published in Data in Brief journal (RODRIGUEZ GAMBOA, Juan C.; ALBARRACIN E., Eva Susana; SILVA, Adenilton J. DA; E. FERREIRA, Tiago A., 2019), and also can be accessed publicly at the repository: (RODRIGUEZ GAMBOA, J.C. *et al.*, 2019).

We used 22 bottles of commercial wines of different varieties and vintages, elaborated in four wineries of the *São Francisco* valley (Pernambuco-Brazil). The spoiled samples obtained from 13 of the 22 bottles were randomly selected and left open for six months before starting the measurements (low- quality LQ wines).

Besides, four bottles were opened two weeks before beginning the data collection (average-quality AQ wines). The remaining five bottles opened at the starting time of each measurement (high-quality HQ wines). Also, we measured isolated ethanol in concentrations (v/v): 2, 5, 10, 20, 30, and 40ml of ethanol diluted in distilled water to make solutions of 200 ml. These concentrations allow guaranteeing a range that covers the different possible values in wines with and without spoilage. To ensure the repeatability of the experiments using O-NOSE, we collected between 10 and 11 samples of 1mL of each wine bottle and around 10 and 12 of the ethanol samples at their different concentrations. Therefore, the database contains 235 wines measurements divided into three groups: high quality (HQ), average quality (AQ), and low quality (LQ), with 51, 43, and 141 measurements, respectively, and 65 ethanol measurements (RODRIGUEZ GAMBOA, Juan C.; ALBARRACIN E., Eva Susana; SILVA, Adenilton J. DA; FERREIRA, 2019).

4 METHODS

4.1 Sample Entropy

Sample entropy (SampEn) was introduced by Richman and Moorman (RICHMAN *et al.*, 2000) as a modification of the approximate entropy (ApEn) (PINCUS, 1991). Both methods serve to analyze the dynamics of time-series by evaluating their regularity and complexity level. A greater regularity (lower complexity) produces lower values of SampEn, whereas, for a series with higher complexity, the value of SampEn statistic is higher. The applications of sample entropy include physiology (LAKE *et al.*, 2002; WENG *et al.*, 2017), geophysics (BALASIS *et al.*, 2009), climatology (SHUANGCHENG *et al.*, 2006), hydrology (CHOU, 2014), and engineering (ZHAO; YANG, 2012). The essential advantages of SampEn in comparison with ApEn are the independence of data amount and relatively simple implementation.

SampEn (m, r, N) is defined as the negative natural logarithm of the conditional probability that two sequences of length N, that are similar (within a tolerance level r) for m points, remain similar for m +1 points, where self-matches are not included in calculating the probability. An algorithm for calculating sample entropy can be described as follows [19]. Given a time series of size N, $X = x_1, x_2, \dots, x_N$, first N-m+1 vectors $\mathbf{x}_m(i)$ of size m are constructed where $\mathbf{x}_m(i) = x_i$, x_{i+1}, ..., x_{i+m-1}, and i=1, ..., N-m+1. The distance $d_{i,j}$ between the vectors $\mathbf{x}_m(i)$ and $\mathbf{x}_m(j)$ is calculated as $d_{i,j}[\mathbf{x}_m(i), \mathbf{x}_m(j)] = \max\{|x_{i+k} - x_{j+k}|: k = 0, \dots, m-1\}$, for each $i = 1, \dots, N-m$ and $j = 2, \dots, N-m+1$, where $i \neq j$ and $j > i$ to exclude self-matches. Subsequently, quantities $B_i^m(r) = \frac{B_i}{N-m-1}$ and $A_i^m(r) = \frac{A_i}{N-m-1}$ are calculated, where B_i is the number of vectors $\mathbf{x}_m(j)$ of size m that are similar to vectors $\mathbf{x}_m(i)$ within a tolerance r estimated from $d_{i,j}[\mathbf{x}_m(i), \mathbf{x}_m(j)] \leq r$, and A_i is the number of vectors $\mathbf{x}_{m+1}(j)$ that are similar to vectors $\mathbf{x}_{m+1}(i)$. From the individual $B_i^m(r)$ and $A_i^m(r)$ values, the corresponding mean values $B^m(r) = \frac{1}{N-m}(\sum_{i=1}^{N-m} B_i^m(r))$ and $A^m(r) = \frac{1}{N-m}(\sum_{i=1}^{N-m} A_i^m(r))$ are now calculated, and finally the statistic called sample entropy expressed in (1) is obtained as

$$SampEn(m,r,N) = -ln\left(\frac{A^m(r)}{B^m(r)}\right). \quad (4\text{-}1)$$

4.1.1 Time-dependent sample entropy

While formal statistical analysis of time series assumes their stationarity, this condition is not always met. In these cases, it is necessary to utilize procedures that adequate for nonstationary data series analysis. The time-dependent sample entropy is one such method, corresponding to the quantification of irregularity in the series at different scales, as a function of time, based on the sliding window protocol. This method examines the entropy values from a temporal evolution perspective, allowing the application of this technique in nonstationary conditions since the series are analyzed by segments (MARTINA *et al.*, 2011; STOSIC, Darko *et al.*, 2016). The method to calculate the time-dependent entropy is as follows. Given a series of data $X = x_1, x_2, \dots, x_N$, the sliding window protocol is defined as $X_t = x_{1+t\Delta}, \dots, x_{w+t\Delta}$, $t = 0,1,\dots,\left[\frac{N-w}{\Delta}\right]$ where $w \leq N$ is the window size, $\Delta \leq w$ is the sliding step, and the operator [.] denotes taking integer part of the argument. The time series values in each window X_t are used to compute the $SampEn_{t,\tau}(m,r,w)$ at a given time t and scale τ.

4.2 Rapid Detection Method for E-Nose

Although some E-Nose devices claim to perform real-time monitoring, the realistic approach to process the sensor array outcomes is achieved offline because the measurements need to be completed before the system makes a forecast. The previous issue is significant, taking into count that the measurement process generally takes some minutes. This issue is a limitation when the idea is the massification of this technology and obtaining responses quickly.

The conventional approach for data processing in E-Nose implies to use the complete response curves of the gas sensors, including the rising state, steady-state, recovery phases. Besides, it includes steps such as signal pre-processing and feature generation/extraction, which entails the selection of a suitable method for each stage, increasing the necessary time to find a suitable classifier and forecast models (LIU; MENG; ZHANG, X.-N., 2018; QI; MENG; ZENG, 2017).

Some researchers have focused on reducing the steps and the necessary know-how for model generation in recent articles. For instance, in (LIU; ZENG; MENG, 2019),the authors proposed a bio-inspired data processing method based on a neural network to mimic the mammalian olfactory system with excellent results but using the entire measurement curves. In another work (LÄNGKVIST *et al.*, 2013), the authors proposed a rapid detection system for meat spoilage using an unsupervised technique that considers only the transient response (stacked restricted Boltzmann machines and auto-encoders.) Although the obtained models offer advantages because the features are learned from data instead of being hand-designed, it may produce low suitable and inaccurate models due to the unsupervised method.

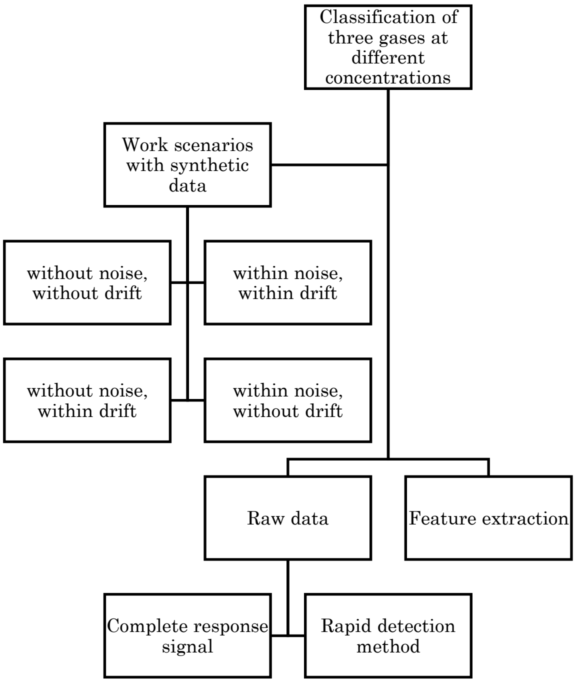
Further, in (PENG, P. *et al.*, 2018; WEI *et al.*, 2019), the authors explored an approach based on raw data treatment. Although this method reduces the steps and the development time, they only tested with the whole response curves, then must wait until the measurement procedure finalization.

Consequently, the mentioned issue motivated the research about a rapid detection approach for the electronic nose systems. We focused on processing an early portion of the signals to reduce the time for making forecasts, testing the proposed method in the collected database with wine samples.

4.3 Sceneries of Synthetic Database

This section explains the four work scenarios generated artificially with the chemosensors package to address the deviations present in artificial odor systems. Figure 13 describes the workflow in the design of these scenarios that have been built to analyze the performance of the classifiers using the traditional approach versus the rapid detection approach using data with and without drift. This research leads to a total of ten databases generated from the four scenarios, in which the noise and drift concentration present in the generated data are parameterized. Prior knowledge of these values allows analysis of results using raw data instead of the traditional method for artificial odor systems, which employs feature selection. These 10 synthetic databases are made available to the public as part of the results of this research at Mendeley Data: https://data.mendeley.com/drafts/s7c74xw673.

Figure 13 – Workflow used for validation of rapid detection method with the synthetic data


Classification of three gases at different concentrations
Work scenarios with synthetic data
without noise, without drift
within noise, within drift
without noise, within drift
within noise, without drift
Raw data
Feature extraction
Complete response signal
Rapid detection method

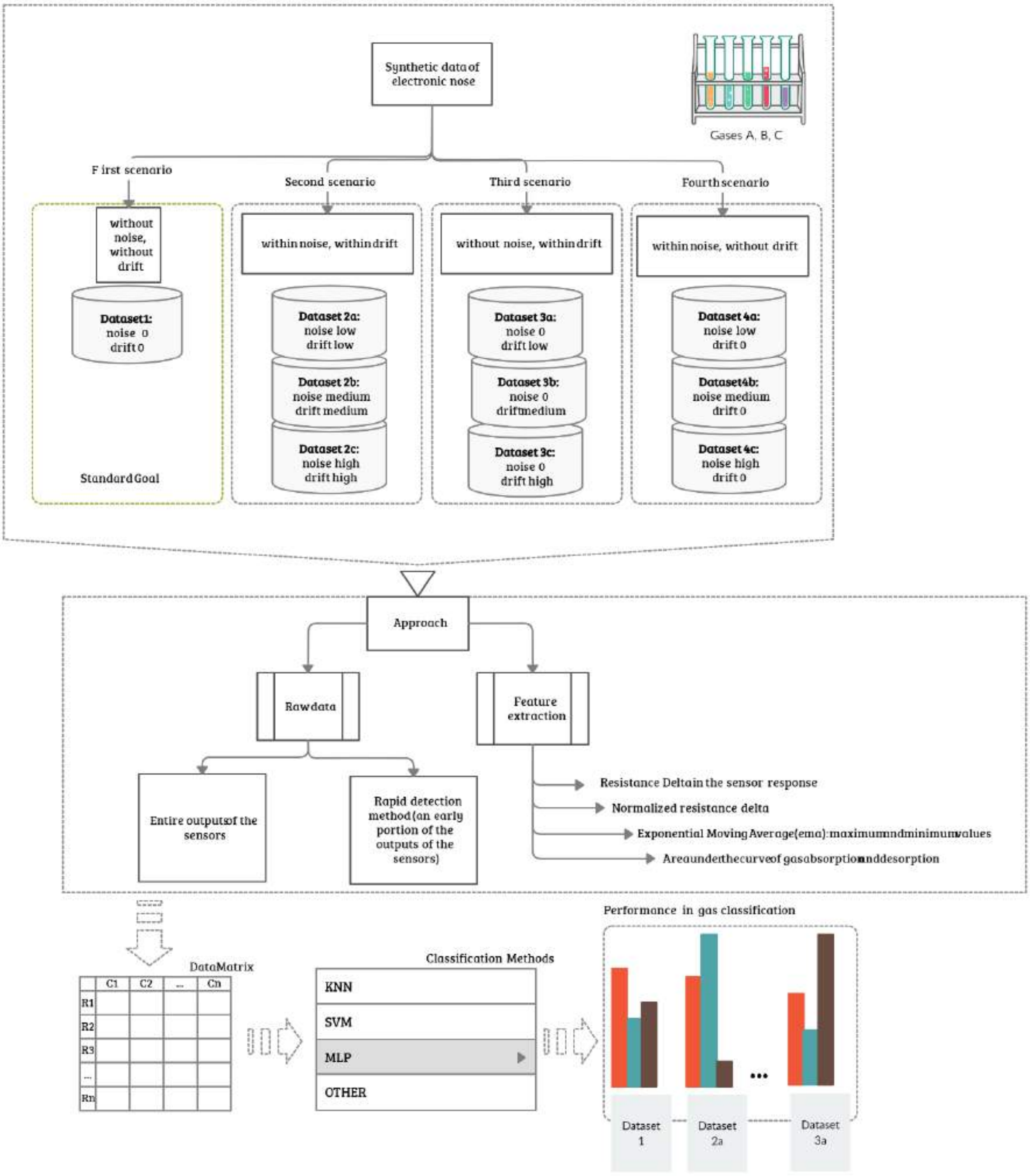
Source: Own

It is represented in Figure 14 the noise and drift parameters established when the synthetic databases were generated with the *chemosensors* package. Note that the first scenario represents the standard goal of the data from electronic nose systems, which ideally contain neither drift nor embedded noise in the response signals of the sensors. The second, third, and fourth scenarios each include three databases in which various combinations of the noise levels and drift included in the data are parameterized.

In total, among the four scenarios, ten synthetic databases were generated. These ten databases are used to compare the traditional method used in artificial odor systems, which consists of selected characteristics in the preprocessing stage, versus raw data from the signals, without a selection of features. In the latter, it is validated whether the rapid detection method manages to find an early portion of the signal that is not affected by drift in terms of the classifier's success rates. Several experiments were performed with different classifiers to test and validate this hypothesis.

Below is a description of the work scenarios generated with three gases A, B, and C at different concentrations. Such scenarios contain measurements that are presented in chronological order to build the corresponding training and validation sets. Table 8 shows the values established for the variables csd, dsd, and ssd. After testing other quantities, these were chosen, finding similar results, so it was decided to analyze these data sets. Additional parameters associated with the generation of the data were taken at their default values.

Figure 14 – Work scenarios with artificial data. Ten synthetic databases were generated to analyze the rapid detection method against the selection of characteristics


Synthetic data of electronic nose
Gases A, B, C
F irst scenario
Second scenario
Third scenario
Fourth scenario
without noise, without drift
Dataset1: noise 0 drift 0
Standard Goal
within noise, within drift
Dataset 2a: noise low drift low
Dataset 2b: noise medium drift medium
Dataset 2c: noise high drift high
without noise, within drift
Dataset 3a: noise 0 drift low
Dataset 3b: noise 0 driftmedium
Dataset 3c: noise 0 drift high
within noise, without drift
Dataset 4a: noise low drift 0
Dataset4b: noise medium drift 0
Dataset 4c: noise high drift 0
Approach
Rawdata
Feature extraction
Entire outputsof the sensors
Rapid detection method(an early portion of the outputs of the sensors)
Resistance Deltain the sensor response
Normalized resistance delta
Exponential Moving Average(ema):maximumndminimumvalues
Areaunderthecurveof gasabsorptionnddesorption
DataMatrix
C1
C2
...
Cn
R1
R2
R3
...
Rn
Classification Methods
KNN
SVM
MLP
OTHER
Performance in gas classification
...
Dataset 1
Dataset 2a
Dataset 3a

Source: Own

Table 8 – Parameters of the databases generated with *chemosensors* package.

	Data Base	csd	ssd	dsd	Noise quantifier	Drift quantifier
First scenario	1	0	0	0	0	0
Second scenario	2A	1	1	1	Low	Low
	2B	1.5	1.5	3	Medium	Medium
	2C	2	2	5	High	High
Third scenario	3A	0	0	1	0	Low
	3B	0	0	3	0	Medium
	3C	0	0	5	0	High
Fourth scenario	4A	1	1	0	Low	0
	4B	2	2	0	High	0
	4C	1.5	1.5	0	Medium	0

Source: Own

Gases A, B, and C were chosen in different concentrations as specified in Table 9.

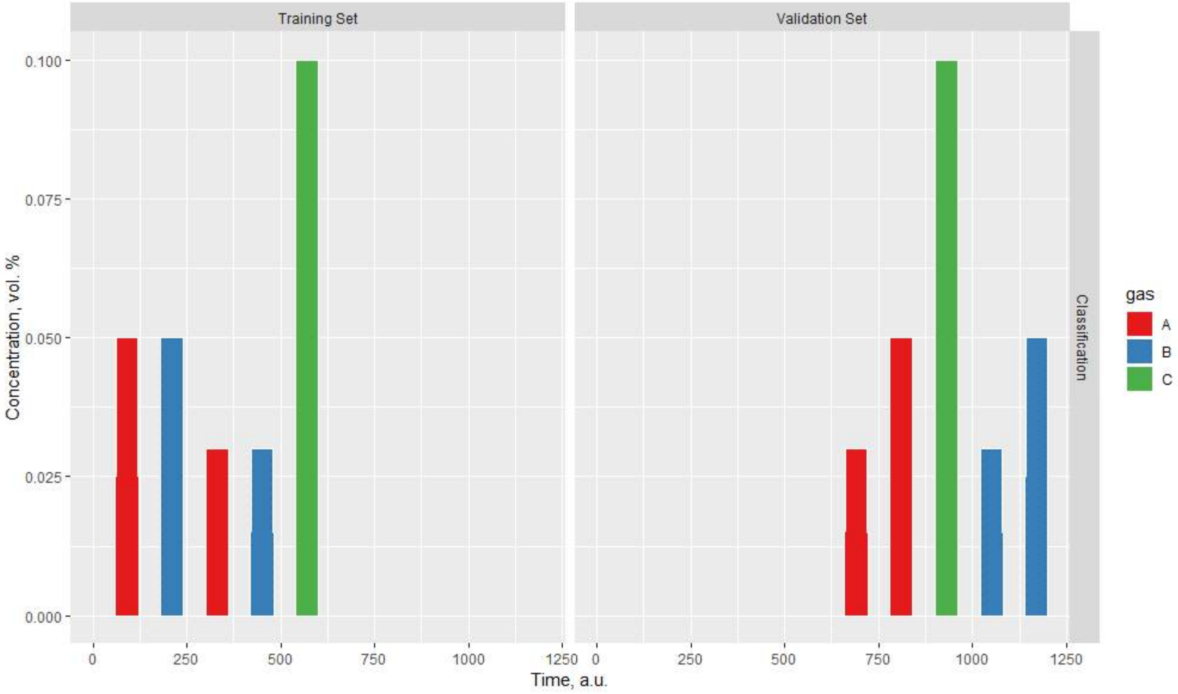
It is seen from Figure 15 an example of the scenario defined as a classification on analyte A, B, and C with both training and validation sets consisting of five pulses of concentrations of A 0.05, A 0.03, B 0.03, B 0.05, C 0.1 vol.%.

Table 9 – Gases concentration in the synthetics databases

Gas or analyte	Concentration (a.u.)
A	0.05
	0.03
B	0.03
	0.05
C	0.1
	0.5

Source: Own

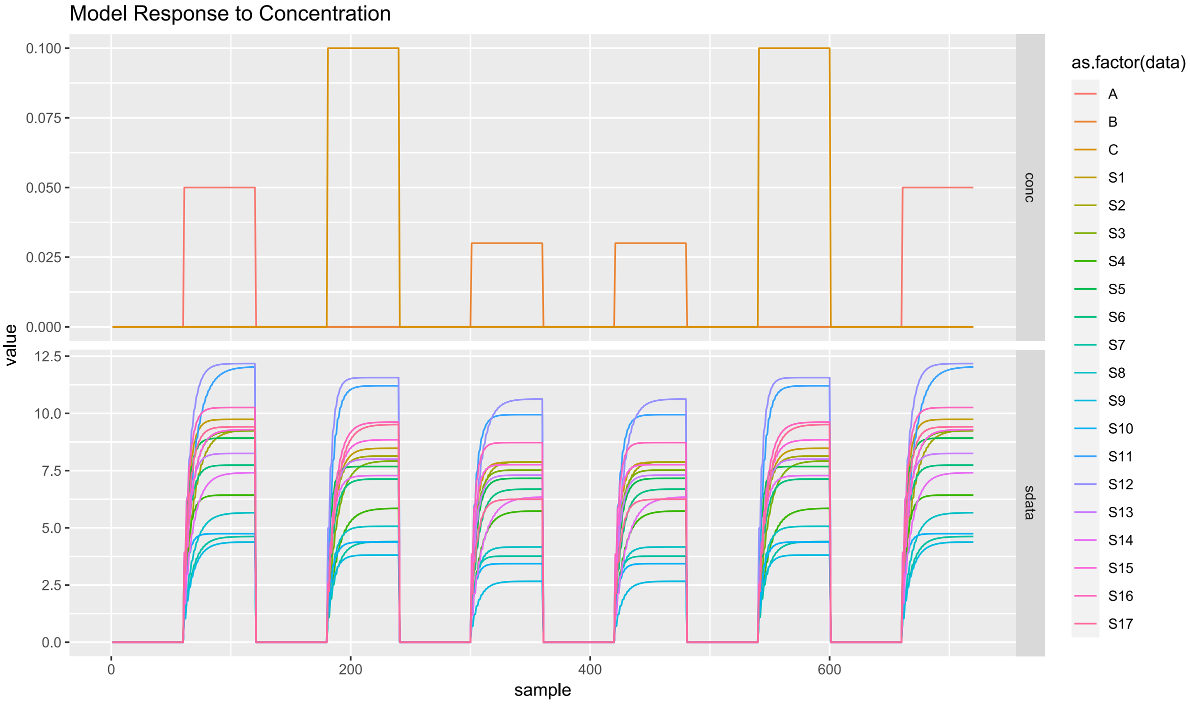
Figure 15 – Plot showing the training and validation set in a classification work scenario with different concentrations of gases A, B, C



Source: Own

Also, each dataset was constructed using arrays of 17 gas sensors considering that the artificial data generator package uses the profiles of 17 sensors used in the UNIMAN database. Therefore, there are arrays of 17 sensors that are used to generate artificial data in each database cited in Table 8. Each database file has 86,400 lines that correspond to 360 sequential measurements of analytes A, B, and C at the concentrations specified in Table 9. Figure 16 illustrates a set of six sequential measurements in which the gas absorption and desorption phases are observed. In this case, the signals' drift and noise are set to zero in the data generator parameters. The upper part of the figure shows the concentration and the gas type injected to each corresponding response of the 17-sensors array represented by S1 to S17. Consequently, Figure 16 represents the behavior of the available signals in the first scenario of synthetic data referenced in Table 8.

Figure 16 – Plot showing the signal responses of the sensor array when analytes at different concentrations of gases A, B, C are injected

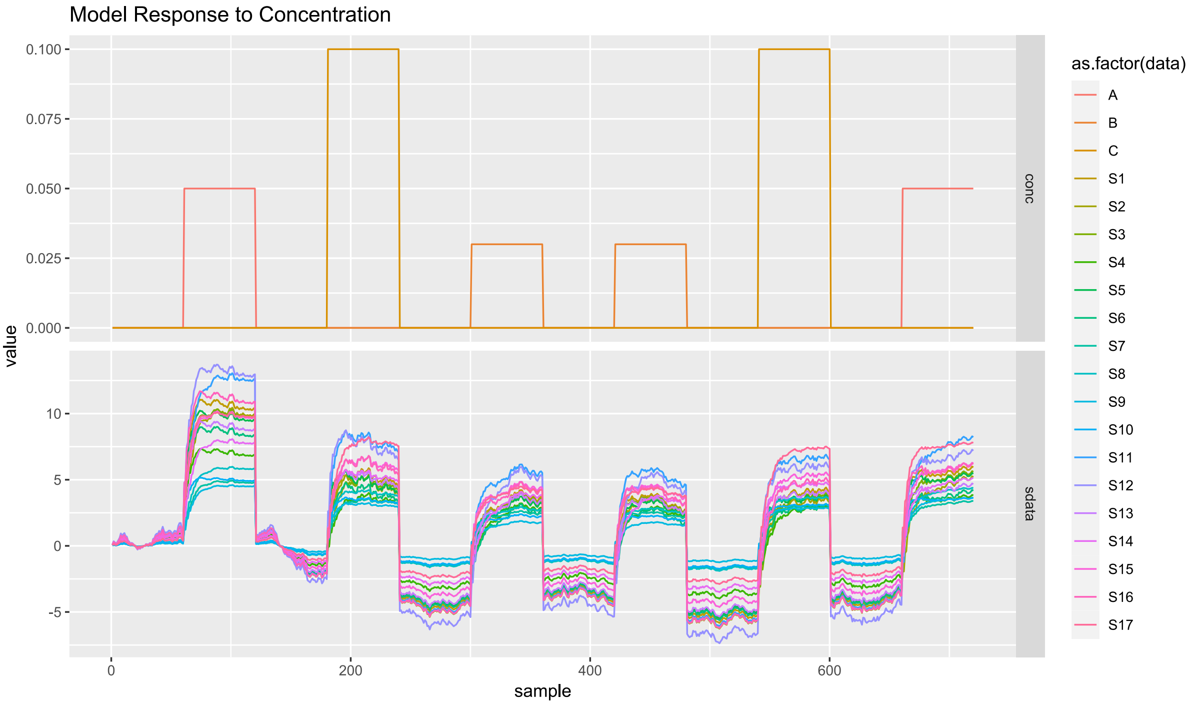


Source: Own

In contrast, a set of six measurements is shown in Figure 17, but with a fraction of drift added in the data generation. However, the noise and noise concentration parameters of the sensors are set to zero. On the other hand, when there is no drift presence in the signals, but there is noise, the signals' behavior is as represented in Figure 19. Finally, a set like the previous one is drawn in Figure 18, but with noise and drifts immersed in the data generation.

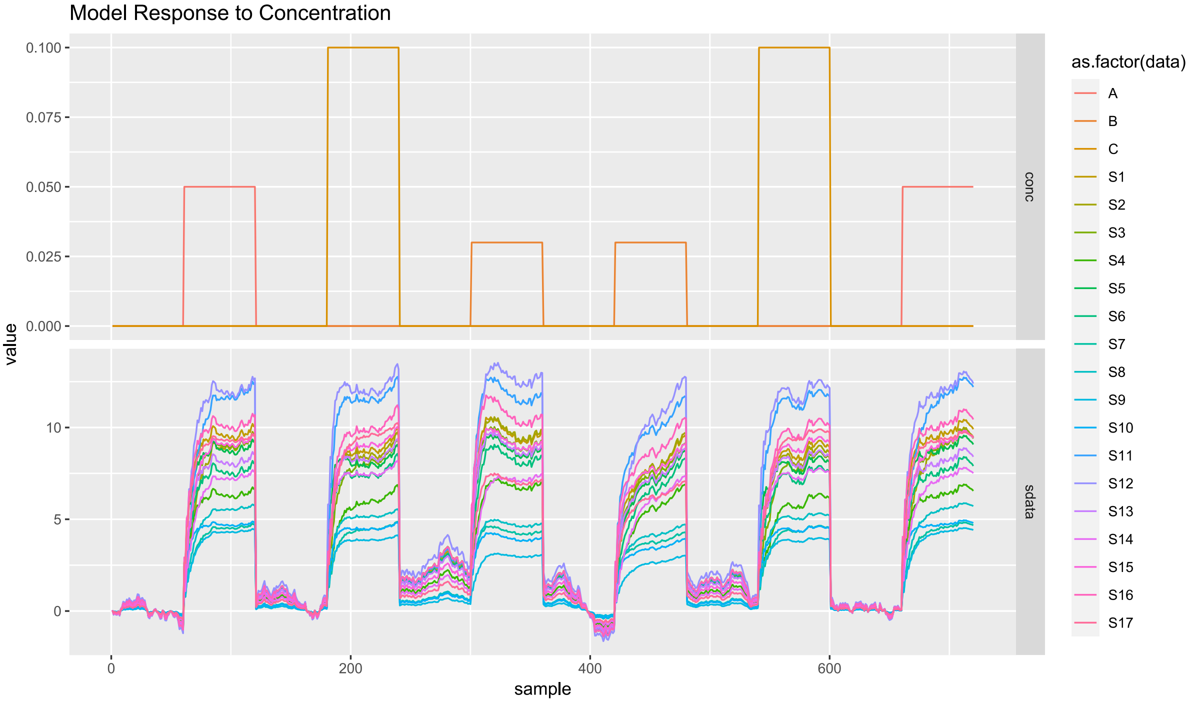
To widen the differences between the four work scenarios, some examples are presented in Figure 20, Figure 21, Figure 22 and Figure 23 from the principal components analysis. These graphs allow us to identify how the presence of drifts in the signals increases the complexity in the system when the groups of analytes represented in the PCA space overlap each other and the random and progressive change in the response signals of the matrix of sensors (see Figure 21 and Figure 22), even when air is being detected.

Figure 17 – Plot representing the second synthetic data scenario. Analytes A, B, and C are injected at different concentrations and there is added noise and drift in the response signals from the sensor array



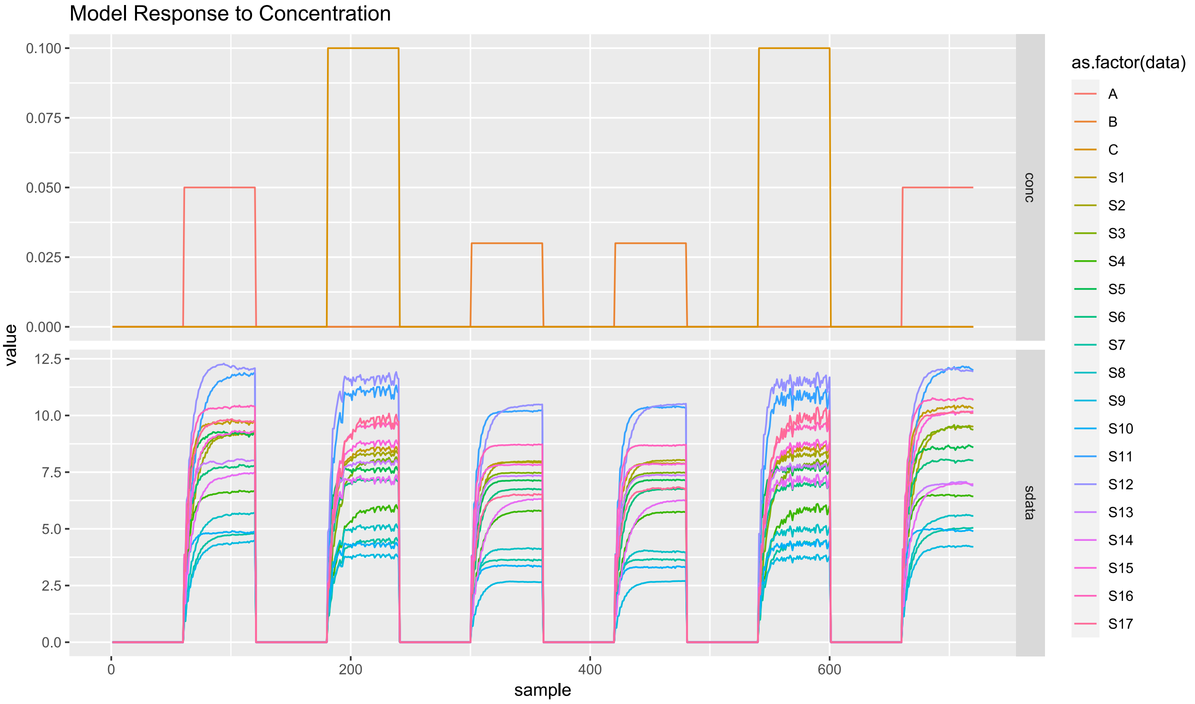
Source: Own

Figure 18 – Plot representing the third synthetic data scenario. Analytes A, B, and C are injected at different concentrations. There is added drift, but noise is set to zero in the response signals from the sensor array



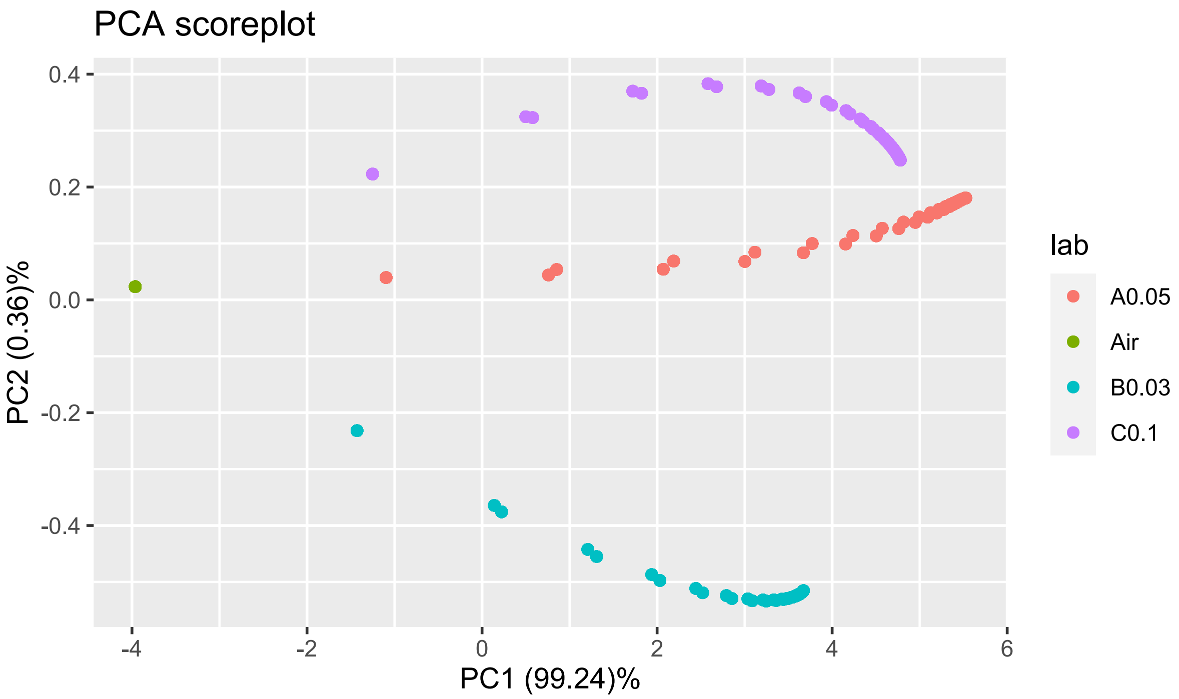
Source: Own

Figure 19 – Plot representing the fourth synthetic data scenario. Analytes A, B, and C are injected at different concentrations, and only noise is added to the responses of the sensor array



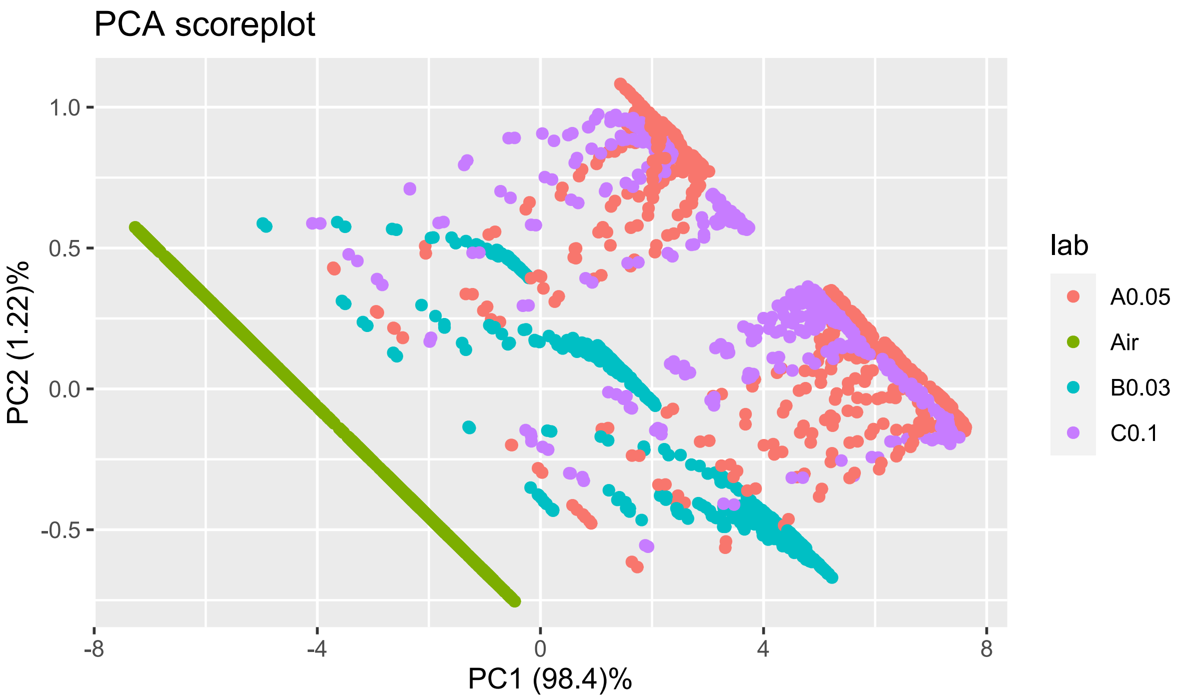
Source: Own

Figure 20 – Score plot corresponding to the Principal Component Analysis of the sensor array to the first work scenario, without noise and drift



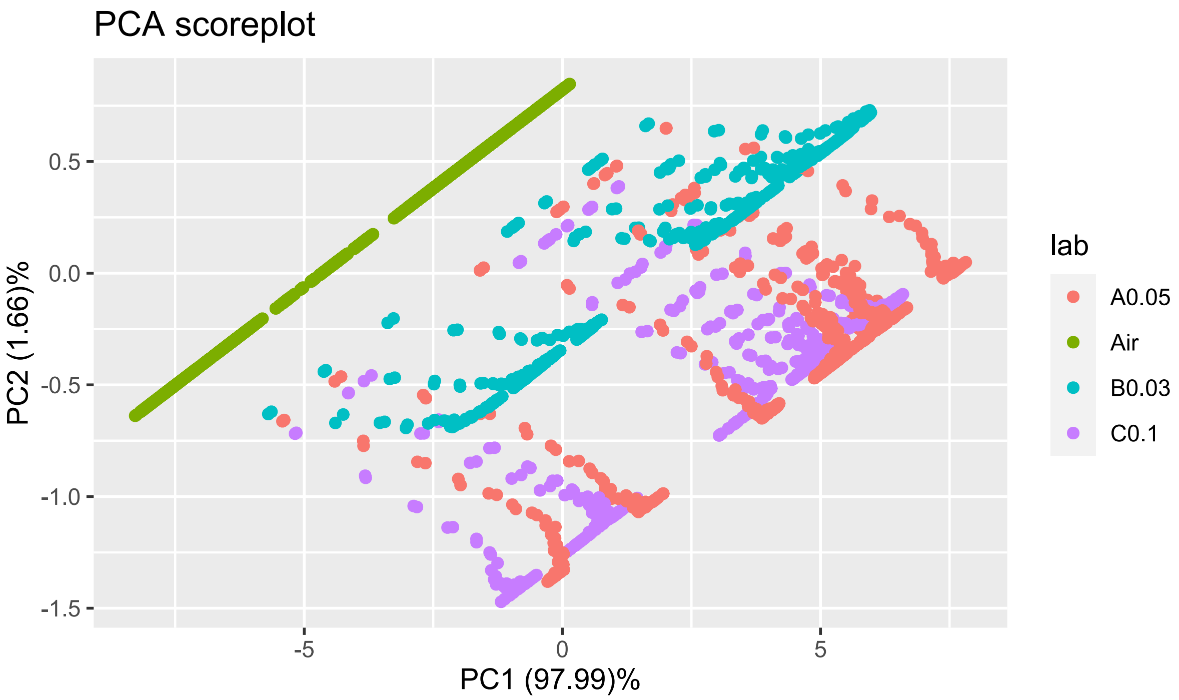
Source: Own

Figure 21 – Score plot corresponding to the Principal Component Analysis of the sensor array to the second work scenario. With noise and with drift



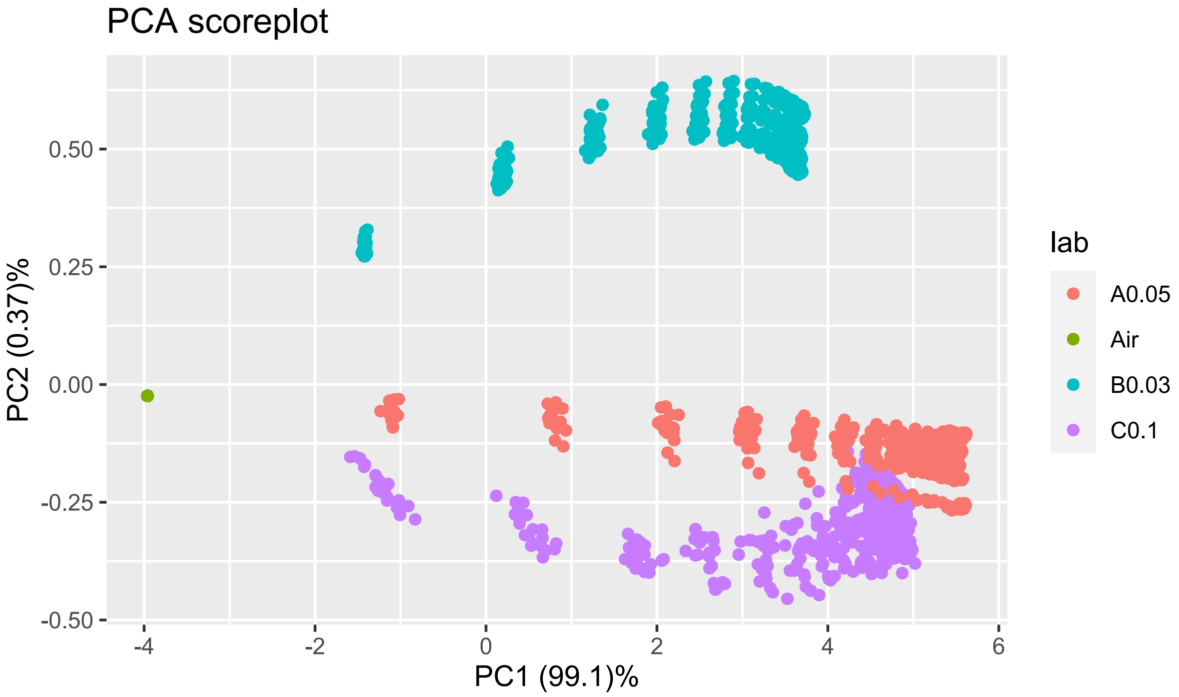
Source: Own

Figure 22 – Score plot corresponding to the Principal Component Analysis of the sensor array to the third work scenario. Without noise and with drift



Source: Own

Figure 23 – Score plot corresponding to the Principal Component Analysis of the sensor array to the fourth work scenario. With noise and without drift.



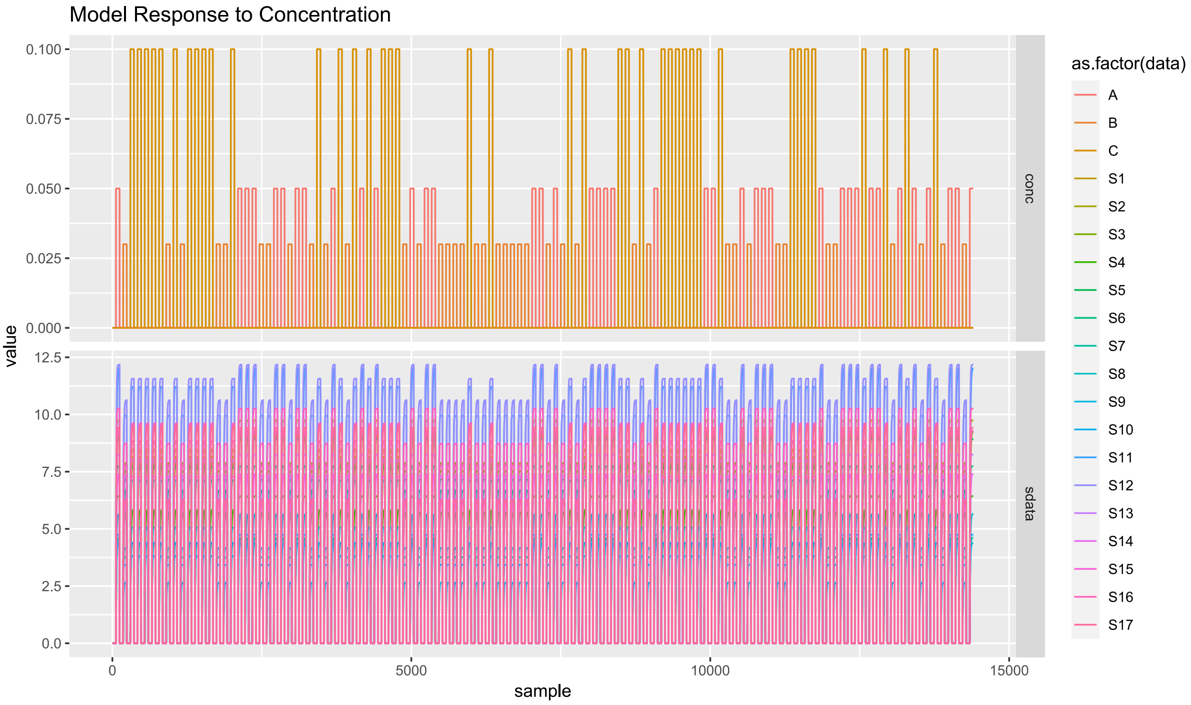
Source: Own

In each workspace, 360 measurements were generated in chronological order. Each measure in time is composed of a matrix of 240 consecutive samples per unit of time (lines of the matrix) for each of the 17 sensors that make up the array. Consequently, each database generated contains 86,400 lines per 17 columns, in addition to the three columns that indicate the kind of gas injected and its respective concentration.

The methodology proposed in this thesis seeks to take advantage of this synthetic data generator by obtaining raw data from response signals from an E-Nose that will be used to present them to several classifiers. The raw data of the signals is necessary to validate the rapid detection methodology that takes advantage of the benefits of classifiers to process large volumes of information. The foregoing is a strong point of this research. In the literature, there are no databases available with measurements in which drift control is had and containing the raw data of the signals. It is clarified that the database delivered to the academic community by (FONOLLOSA; RODRÍGUEZ-LUJÁN; HUERTA, Ramón, 2015) contains the characteristics presented in the table for each measurement but does not provide the raw data.

With the rapid detection approach, the analysis is presented using the complete signal delivered by the sensor matrix or by identifying an early portion of the signal that allows for a correct prediction, mitigating the effect caused by drift. This early potion is determined by applying a sliding window, seeking to find the most effective portion to mitigate the effect of drifts.

Figure 24 – Plot showing the signal responses of the sensor array when analytes at different concentrations of gases A, B, C are injected. The chronological sequence of the 360 measurements is represented in this graph. In the upper part, the injected gases and their concentrations are represented.

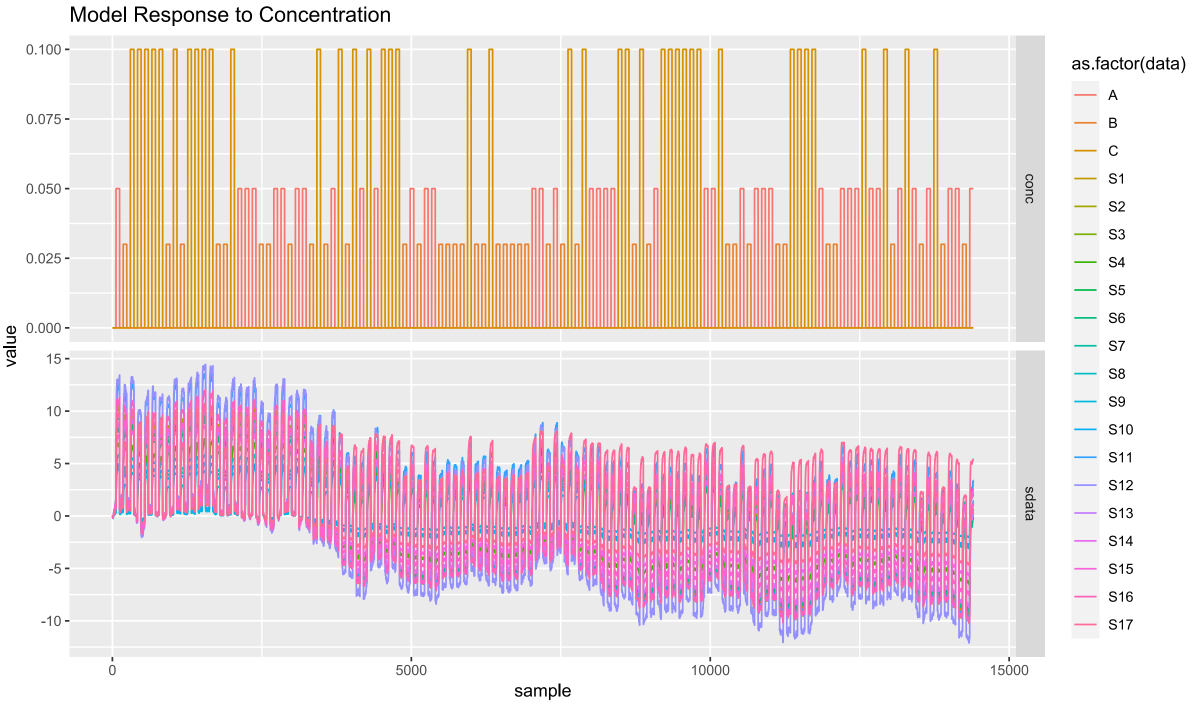


Source: Own

Below are the plots with the sensor response data. The 360 consecutive measurements are plotted in each graph in time for the different work scenarios proposed. In the lower part of the graphs, the response signals of the sensors are represented, and in the upper part of the graph, the input gases and their concentrations are represented. Figure 24 contains the response signals of the first working scenario in which data is established without noise and drift.

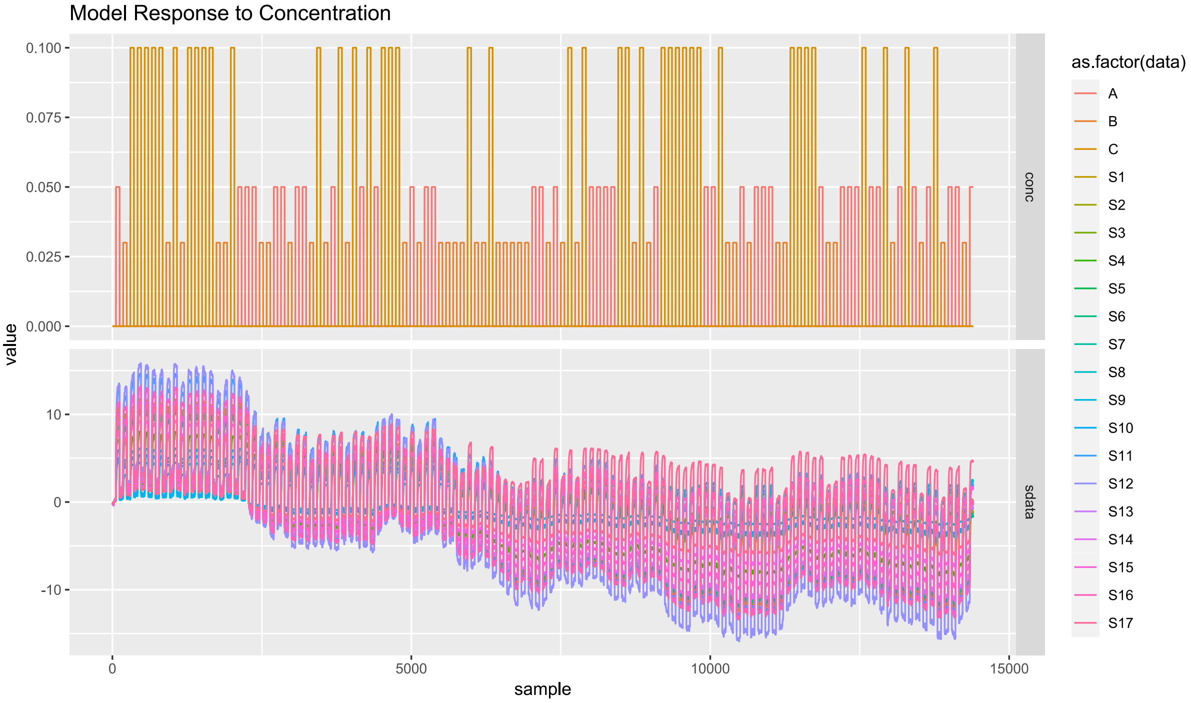
Note that Figure 25 and Figure 26 clearly show the effect of drifts on the gas sensor signals. This effect generates such degradation in the signals that the set of data presented to the classifier in chronological order makes the patterns learned in training obscure with time and the artificial smell system. On the contrary, in Figure 24 and Figure 27 that correspond to the first and fourth work scenarios, respectively, the drift and degradation of the sensor responses are not noticeable thanks to the drift in these scenarios being set to zero.

Figure 25 – Plot showing the signal responses of the sensor array when analytes at different concentrations of gases A, B, C are injected. The chronological sequence of the 360 measurements is represented in this graph. These correspond to the second work scenario (with noise and drifts). In the upper part, the injected gases and their concentrations are represented



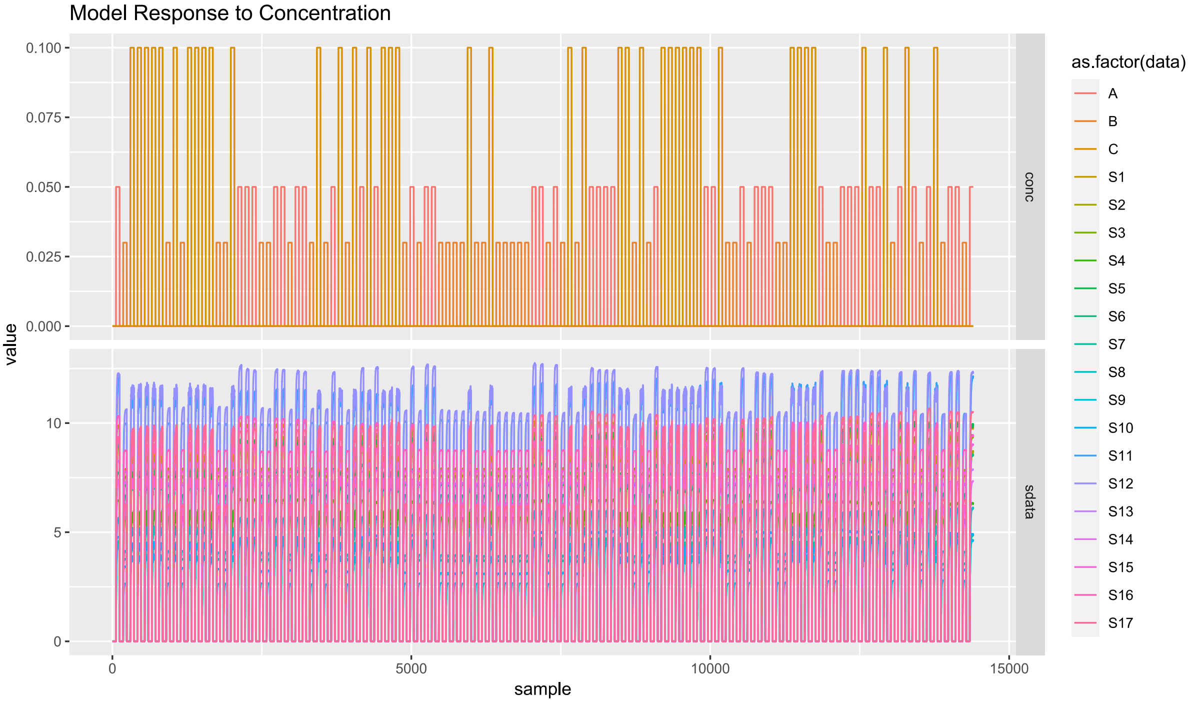
Source: Own

Figure 26 – Plot showing the signal responses of the sensor array when analytes at different concentrations of gases A, B, C are injected. The chronological sequence of the 360 measurements is represented in this graph. These correspond to the third work scenario (without noise and with drifts). In the upper part, the injected gases and their concentrations are represented



Source: Own

Figure 27 – Plot showing the signal responses of the sensor array when analytes at different concentrations of gases A, B, C are injected. The chronological sequence of the 360 measurements is represented in this graph. These correspond to the fourth work scenario (with noise and without drifts). In the upper part, the injected gases and their concentrations are represented



Source: Own

4.3.1 Work Scenarios using Feature extraction

The most common groups of characteristics extracted from the gas sensors signals are the steady and transient state features (YAN *et al.*, 2015). We used eight features to capture each gas sensor's dynamic and static behavior, based on (RODRIGUEZ GAMBOA, Juan C *et al.*, 2021). We obtained a 136 columns characteristics matrix (synthetic database using 17 sensors), where each row represents the fingerprint of one measurement and three additional columns indicating the gas type and the corresponding concentration. It was chosen the following steady-state characteristics: $\Delta G = max_k g[k] - min_k g[k]$ defined as the maximal conductance change concerning the baseline, and its normalized version$\|\Delta G\| = (max_k g[k] - min_k g[k])/min_k g[k]$, as well, : $\Delta G = |g[N] - g[1]|$ defined as the conductance change between the final point to the initial point, and its normalized version $\|\Delta G\| = (|\, g[N] - g[1]|)/g[1]$. Besides, the area under the curve in the absorption $absorption\ area = \int_{g[1]}^{max_k g[k]} g[k]$ and desorption portions of the gas $desorption\ area = \int_{max_k g[k]}^{g[N]} g[k]$. Additionally, we had an aggregate of features reflecting the dynamics of the rising/falling transient portion of the sensor response using an exponential moving average filter (emaα) that converts the transient portion into a real scalar by estimating the maximum/minimum value y[k]=(1-α)y[k-1]+α(x[k]-x[k-1]), where [k=1,2,...,T], y[0] its initial condition, set to zero (y[0]=0), and the scalar α (α∈{0,1}) being a smoothing parameter of the operator such as was defined in (MUEZZINOGLU *et al.*, 2009; VERGARA *et al.*, 2012). We tested three different values for α=0.1, α=0.01, and α=0.001; but it was chosen the max emaα with α=0.01 as an informative transient feature and his relative position.

4.4 Classification methods

Different models were generated using the Python programming language. An SVM algorithm available in the scikit-learn library was used, with the following parameters to optimize the model: A Radial Bayes Function (RBF) as the kernel, the regularization parameter C as 10, and the other settings as the default value. The second architecture corresponds to a simple Deep MLP model with only fully connected layers based on the model used in (RODRIGUEZ GAMBOA, Juan C.; ALBARRACIN E., Eva Susana; SILVA, Adenilton J. Da; *et al.*, 2019). The configuration of the MLP model consists of eight layers with *Tanh* as the activation function except for the output layer, in which we used *softmax*. The input layer has 100 neurons, and all the hidden layers have 30 neurons. Other three DL architecture implementation types were used for the classification tasks based on (RODRIGUEZ GAMBOA, Juan C *et al.*, 2021), called Sniff ConvNet, Sniff ResNet, and Sniff Multinose.

4.4.1 Training Configurations

The three sets of DL models were trained to reach 20 epochs using the Stochastic Gradient Descent (SGD) algorithm for optimization with a learning rate of 0.001 and a momentum of 0.9. Besides, we used the categorical cross-entropy loss function.

Regarding the training process in all tested classification methods, all datasets were split as follows, the training group including 20% of measurements and the validation group with 80%.

4.5 Discrete Wavelet Transform

The wavelet analysis has been applied to problems that including noise removal in signals. Due to its better time-frequency resolution, it overcomes other classical methods, such as short time Fourier Transform, for instance. One of the advantages when using wavelets is the computational efficiency of Mallat's pyramidal algorithm. This algorithm is indeed a two-channel filter bank that splits the input signal into low and high frequencies using quadrature mirror filters

(OLIVEIRA, DE *et al.*, 2018). Wavelet analysis proves effectively analyzing the distorted signals in time-frequency domain, and it can be described mathematically as follows as defined in (GAROUSI; SHAKARAMI; NAMDARI, 2016).

$$f(x) = \sum_{ij} a_{i,j}\psi_{i,j}(x). \tag{4-2}$$

where i and j represent the integer values and $\psi_{i,j}(x)$ stands for wavelet expansion functions. a_{ij} Stands for the two coefficients of discrete wavelet transform (DWT) of $f(x)$. These coefficients have the formula:

$$a_{i,j} = \int_{-\infty}^{+\infty} f(x)\psi_{i,j}(x) \tag{4-3}$$

where $\psi_{\mathrm{i,j}}(\mathrm{x})$ represent the mother wavelet and can gain its parameters through:

$$\psi_{i,j}(x) = 2^{-i/2}\psi(2^{-j}x - j) \tag{4-4}$$

where i represent the scaling parameter in wavelet and j for the translation one. For multiresolution satisfaction, the difference of two scale equation is given as:

$$\phi(\mathrm{x}) = \sqrt{2}\sum_{\mathrm{k}} \mathrm{h(k)}\phi(2\mathrm{x} - \mathrm{k}) \tag{4-5}$$

where $h(k)$ gives the wavelet function a unique value by satisfying wavelet conditions and $\phi(x)$ is scaling function, which has a relation with the mother of wavelet as follows:

$$\psi(x) = \sqrt{2}\sum_{k} g(k)\phi(2x - k) \tag{4-6}$$

where h in (4-5) and g in (4-6) can be considered filters of wavelet of low-pass filter and high pass filter. From all the above equation, the j wavelet value can be determined as:

$$f_0(x) = \sum_k a_{0,k} \phi_{0,k}(k) = \sum_k a_{J+1,k} \phi_{J+1,k}(x) + \sum_{j=0}^{J} d_{j+1,k} \psi_{j+1,k}(x) \tag{4-7}$$

where $a_{0,k}$, $a_{J+1,k}$, $d_{j+1,k}$ are the coefficients at scale j+1 and can be determined under the condition of the availability of scale j as follow:

$$a_{j+1,n} = \sum_k a_{j,k} h(k-2n) \tag{4-8}$$

$$d_{j+1,n} = \sum_k a_{j,k} g(k-2n) \tag{4-9}$$

where $a_{j+1,n}$is the approximation coefficient and $d_{j+1,n}$ is the detailed one at scale j+1 defined.

5 RESULTS

5.1 Exploratory analysis of the data

It is plotted in the figures the measurements by gas existing in the database to pre-explore the University of California database, mentioned previously in this document in section 3.1.

The figure shows the behavior of the response signals when Gas 1 is sensed. These data comprise the set of 8 characteristics listed in Table 6. In turn, the graph in the upper part includes the 2-D plot, and in the lower part, the 3-D plot of the characteristics presented in the database.

In order to explore the database of the University of California, mentioned earlier in this document in section 3.1, the existing measurements by gas in this database are graphed in Figure 28 to Figure 33. Each measurement is expressed in a data matrix of 16 sensors by eight characteristics, which results in a multidimensional array of 128 columns by the number of measures (rows) reported for each gas class.

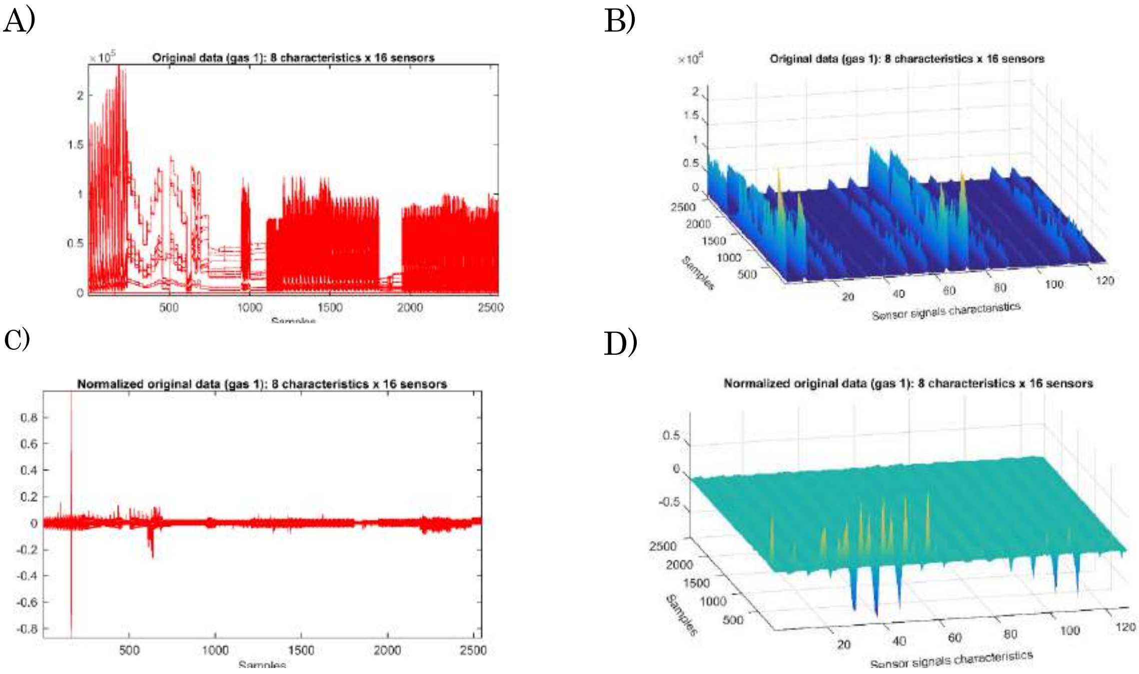
The exploratory analysis begins with the Gas 1 measurements. A total of 2565 measurements were reported for this gas. Therefore, the data matrix for this gas is an array that contains 128 columns (characteristics) by 2565 rows (measurements) with the values provided by the authors of the database. Panel A of Figure 28 shows the resulting 2-D plot for this kind of gas, and the 3-D plot is presented in Panel B of this same figure. This 3-D plot allows appreciating the magnitude of the reported values more clearly. It is identified that, in the first measurements, there is substantial contamination in the sensors caused by the high values of the first feature. This corresponds to the first of the features of the stable state of each sensor's response signal (resistance delta), which in the graph are observed as the peaks in yellow. It is identified that, for this gas, the sensors that suffer the most remarkable saturation caused by these peaks in response to gas 1 are sensors 1, 2, 9, and 10. Given the above, it is inferred that Gas 1 saturates the aforementioned sensors to a greater extent. It is also observed that the sensors were contaminated by the operator, affecting the measurement process.

Given the above, it is inferred that Gas 1 saturates the sensors mentioned above to a greater extent. It is also observed that the sensors were contaminated by the operator, affecting the measurement process. Additionally, in Panels C and D, the normalized data are presented, identifying outliers and measurements with an instrumental error. These factors affect the difficulty of achieving the classifier's prediction. Therefore, it is decided to remove them.

Continuing the analysis, the response of the system when gas 2 is sensed is reported in Figure 29. The authors of the database reported 2926 measurements made with this gas. Similarly to that expressed for gas 1, the matrix of characteristics of this gas is made up of 128 columns with 2926 rows. It is observed in the 3D plot (panel B) that the sensors respond in a similar way to gas, with a difference of sensors 5, 6, 13, and 14 in which an almost null response is observed. There are also findings of outliers and deviations that affect the data identified in the yellow peaks in the 3D plots. Such deviations for this gas occur at the beginning and the end of the data collection.

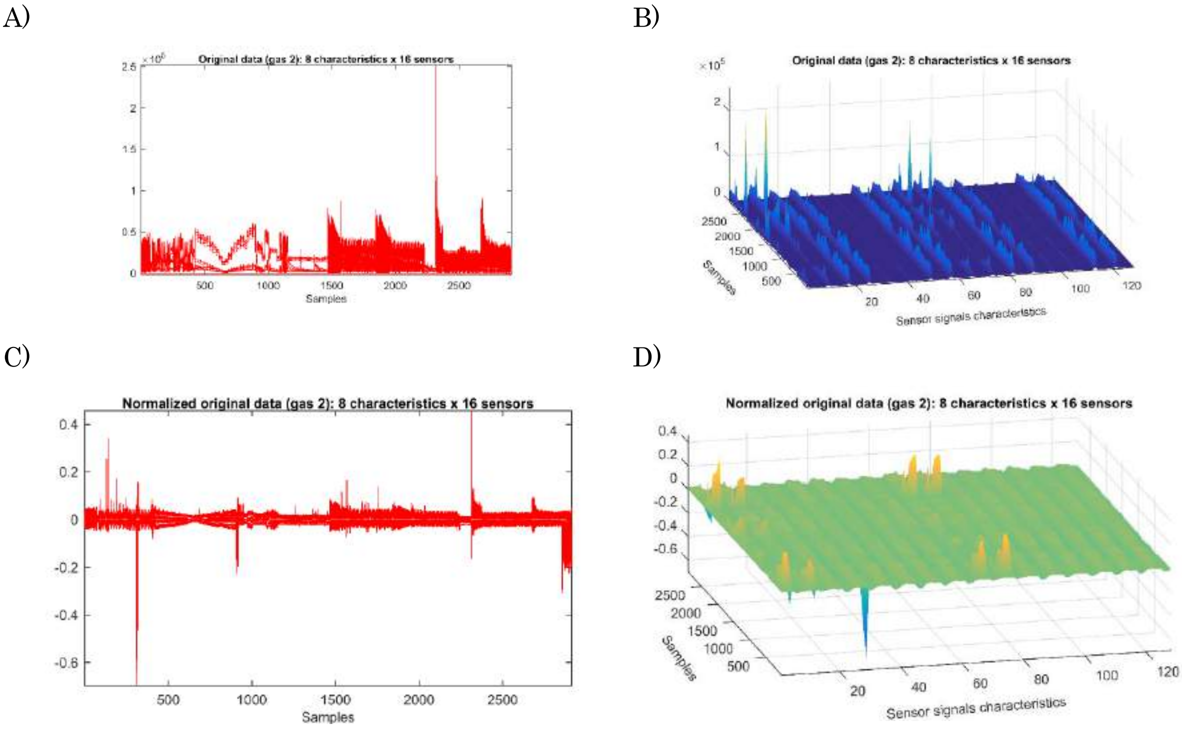
The gases sensed and represented in Figure 28 and Figure 29 correspond to Ammonia and Acetaldehyde in the concentrations reported in Table 2.

Figure 28 – Plot showing the measurements reported for gas 1. In panel A, the 2-D characteristics matrix of this class of gas is drawn. Panel B shows the feature matrix in 3D. Panel C shows a matrix of normalized 2D features, and Panel D presents a Matrix of normalized 3-D



Source: Own

Figure 29 – Plot showing the measurements reported for gas 2. In panel A, the 2-D characteristics matrix of this class of gas is drawn. Panel B shows the feature matrix in 3D. Panel C shows a matrix of normalized 2D features, and Panel D presents a Matrix of normalized 3-D

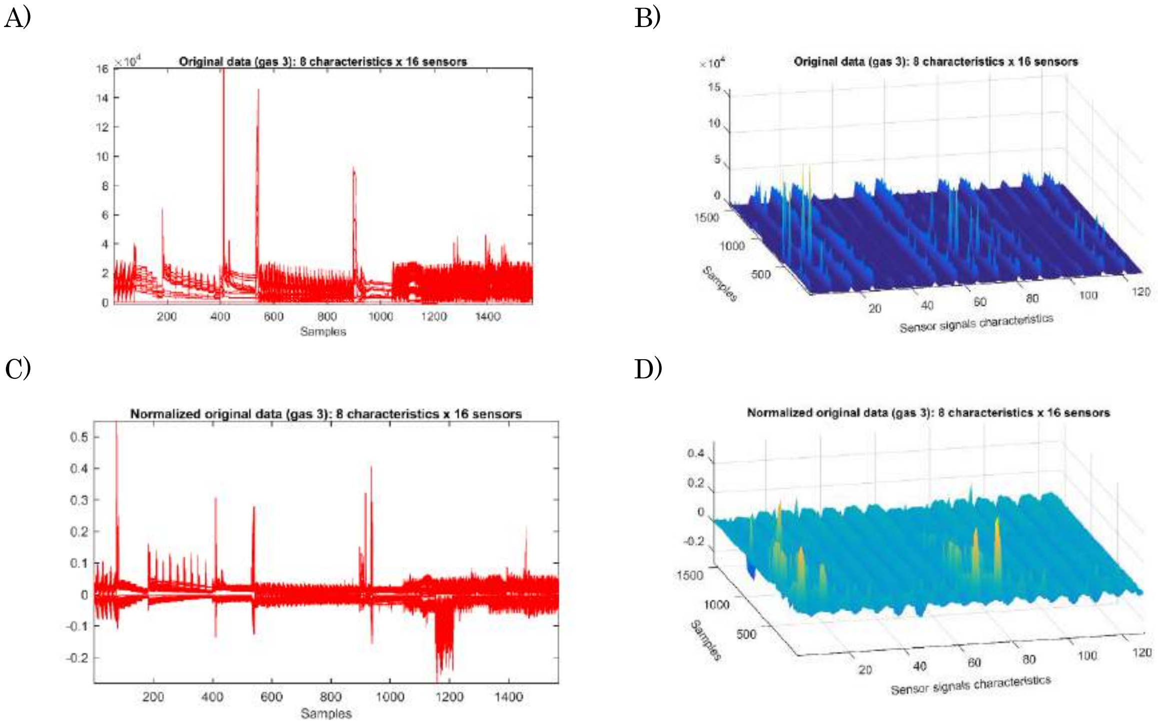


Source: Own

Figure 30 represents the behavior of the response signals of the sensors when gas 3 is injected. In this case, the behavior of the sensors is similar to that reported for gas 2. However, there is a slight improvement in the response of sensors 5, 6, 13, and 14, unlike what is observed in the previous plot. Additionally, it can be seen in Figure 31 that for gas 4, the sensors that best respond to this gas are sensors 1, 2, 9, and 10. For their part, sensors 3, 4, 7, 8, 11, 12, 15, and 16 provide an acceptable response, while sensors 5, 6, 13, and 14 show almost zero behavior with respect to the other sensors. Again, the peaks reflected in the 3D plots are identified for the two figures cited in this paragraph.

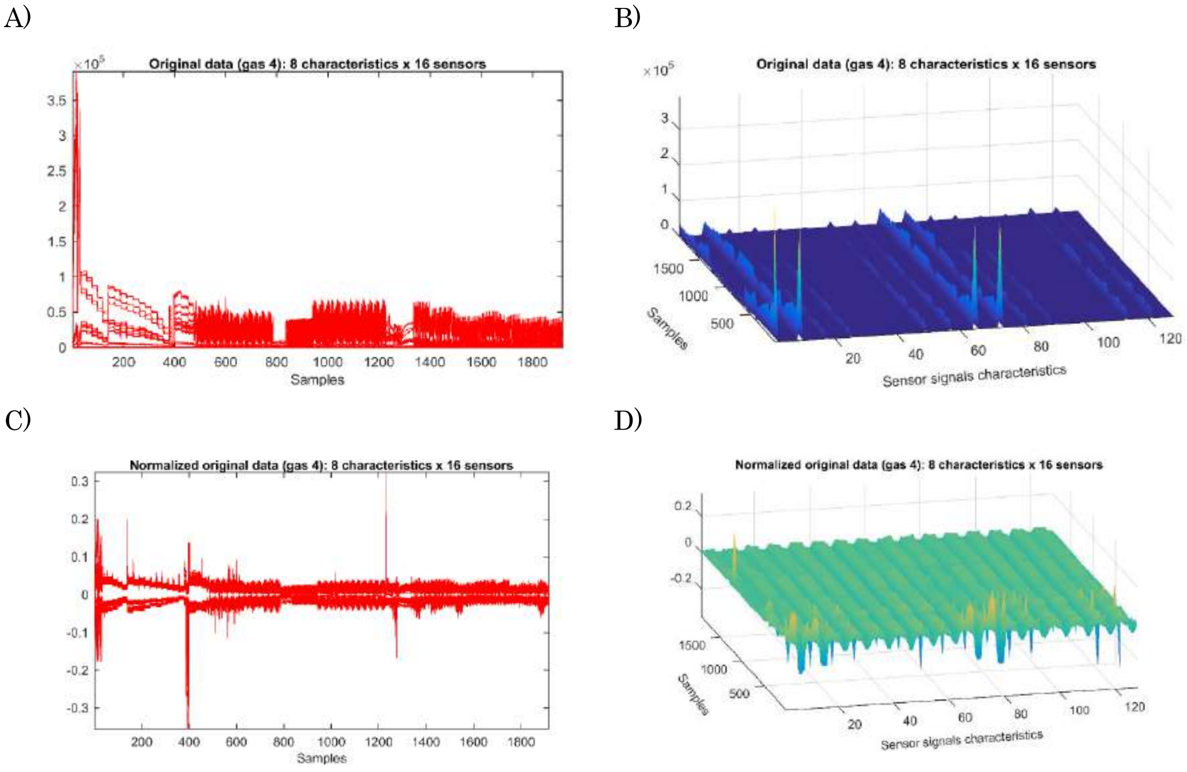
In this case, the gases reported in Figure 30 and Figure 31 corresponds to Acetone and Ethylene in the concentrations expressed in Table 3.

Figure 30 – Plot showing the measurements reported for gas **3**. In panel A, the 2-D characteristics matrix of this class of gas is drawn. Panel B shows the feature matrix in 3D. Panel C shows a matrix of normalized 2D features, and Panel D presents a Matrix of normalized 3-D



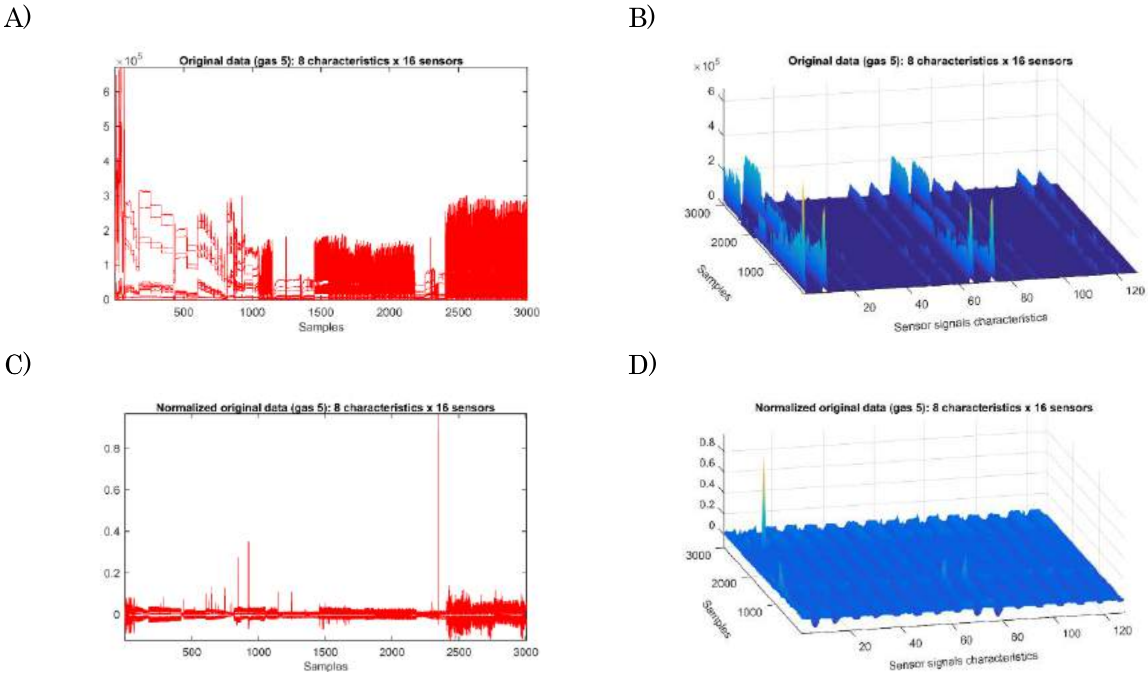
Source: Own

Figure 31 – Plot showing the measurements reported for gas **4**. In panel A, the 2-D characteristics matrix of this class of gas is drawn. Panel B shows the feature matrix in 3D. Panel C shows a matrix of normalized 2D features, and Panel D presents a Matrix of normalized 3-D



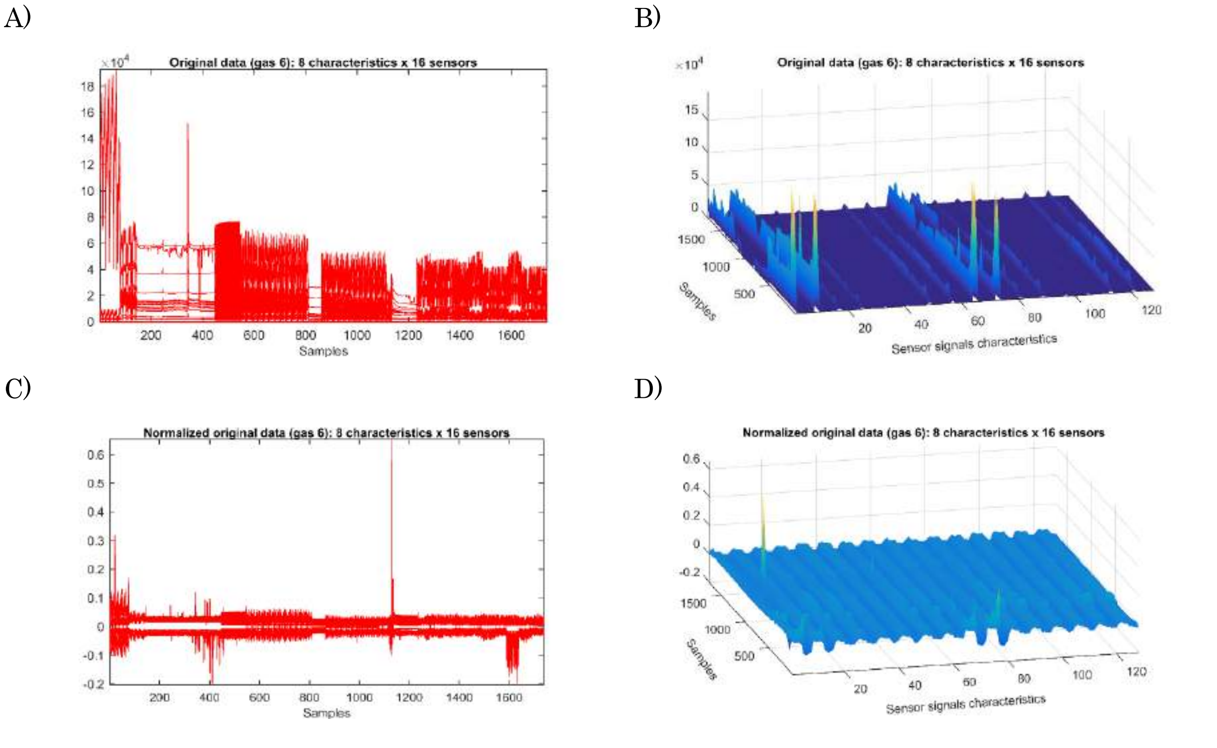
Source: Own

Figure 32 – Plot showing the measurements reported for gas **5**. In panel A, the 2-D characteristics matrix of this class of gas is drawn. Panel B shows the feature matrix in 3D. Panel C shows a matrix of normalized 2D features, and Panel D presents a Matrix of normalized 3-D



Source: Own

Figure 33 – Plot showing the measurements reported for gas 6. In panel A, the 2-D characteristics matrix of this class of gas is drawn. Panel B shows the feature matrix in 3D. Panel C shows a matrix of normalized 2D features, and Panel D presents a Matrix of normalized 3-D



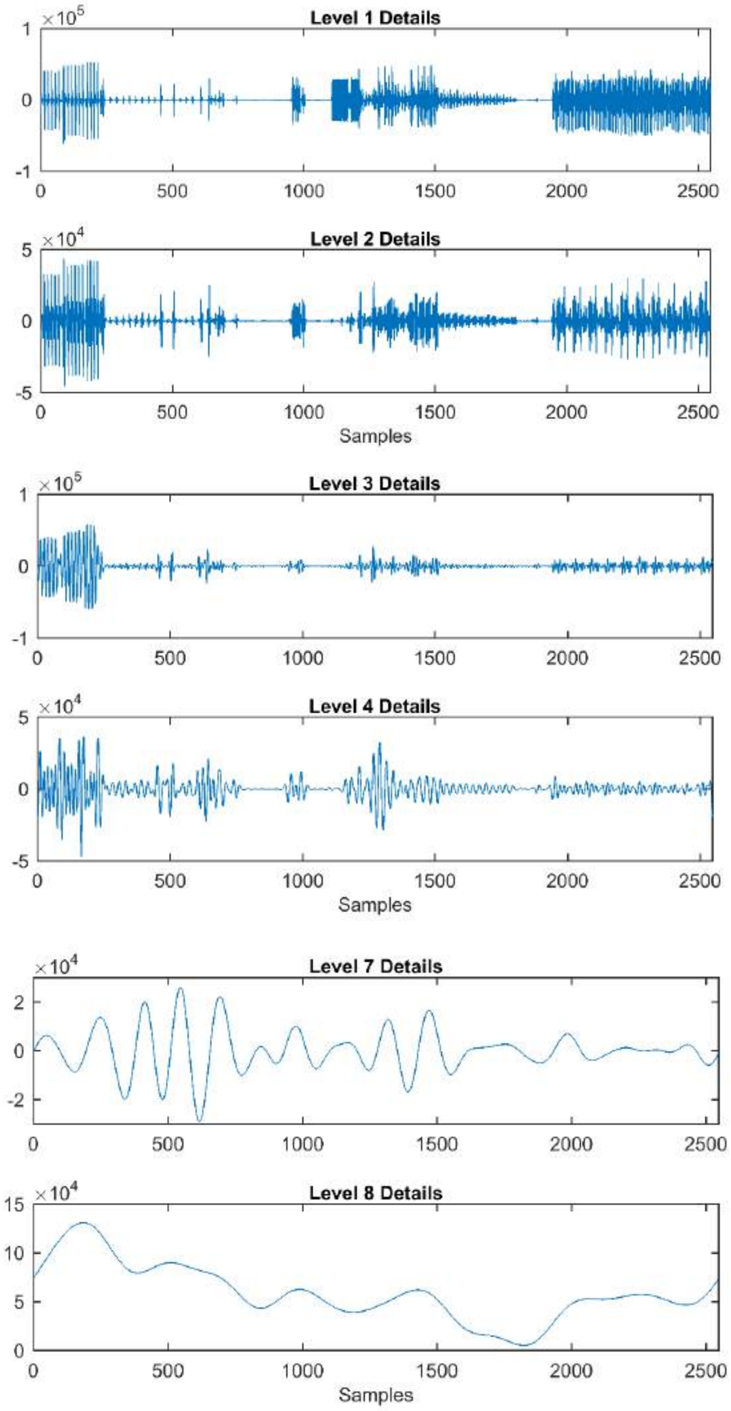
Source: Own

Finally, Figure 32 and Figure 33 show the measurements corresponding to gases 5 and 6, which are equivalent to Ethanol and Toluene in the concentrations referenced in Table 3. In this case, the sensors that best respond to Ethanol are sensors 1, 2, 9, and 10 of the array sensors. For its part, an acceptable response is identified from sensors 3, 4, 7, 8, 11, 12, 15, and 16. Regarding Toluene, a behavior of the sensors similar to that reported for Ethanol is identified.

5.2 Obtain a Multiresolution Analysis (MRA) of the data using the SYM4 Wavelet

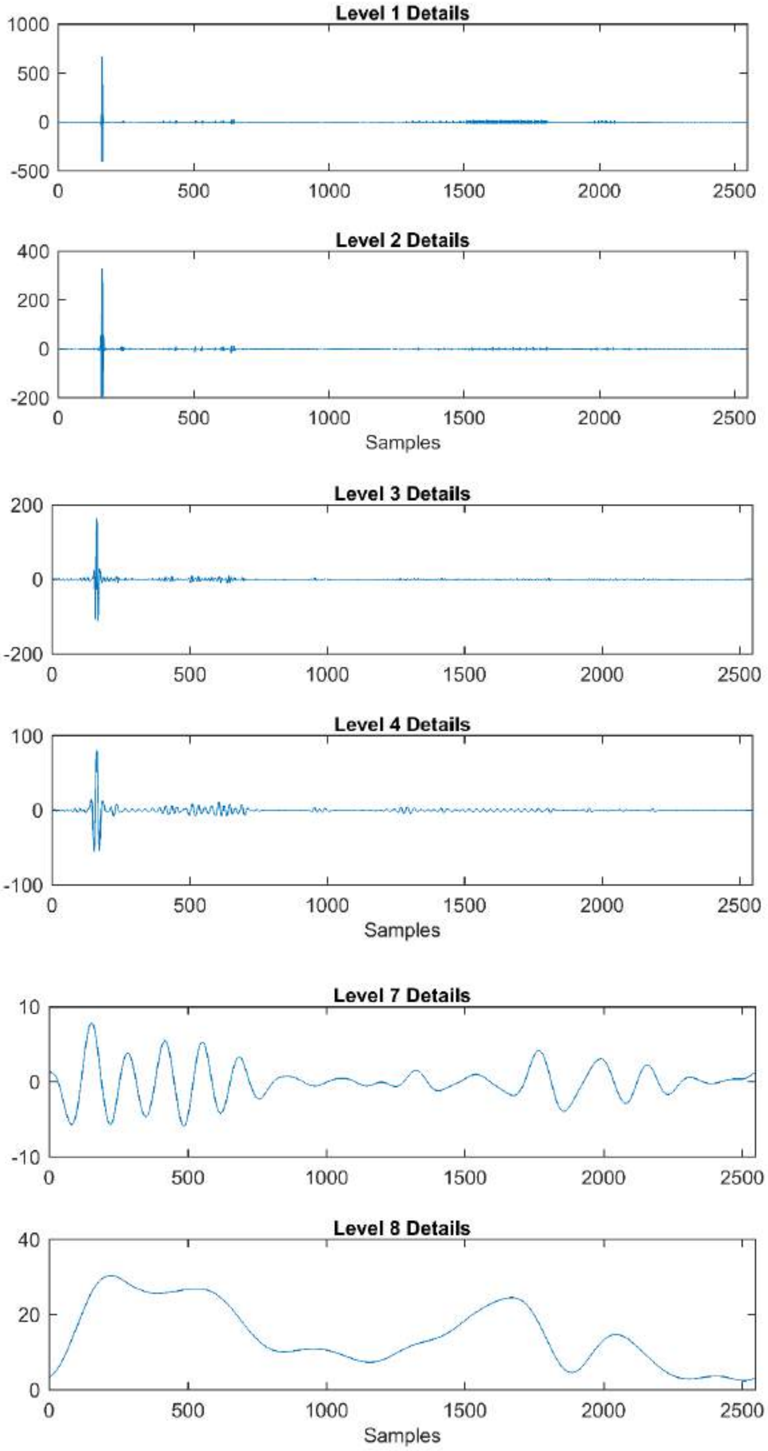
It performed a wavelet decomposition at level 8 using the ' sym4 ' wavelet. For this, it has chosen a wavelet and a level of decomposition 8, and then it computed the wavelet decompositions of the signals at level 8. Plot details Gas 1-Sensor 1 (ch 1). The presence of high and low frequencies in the decomposition is evidenced, indicating noise and drift, respectively. Figure 34 and Figure 35 depict the MRA details of the response signals to Gas 1. For the analysis, the first characteristic of Sensor 1 is chosen because it corresponds to the resistance delta of the sensor.

Figure 34 – MRA decomposition of the first feature in the gas 1 response signals in the first sensor.



Source: Own

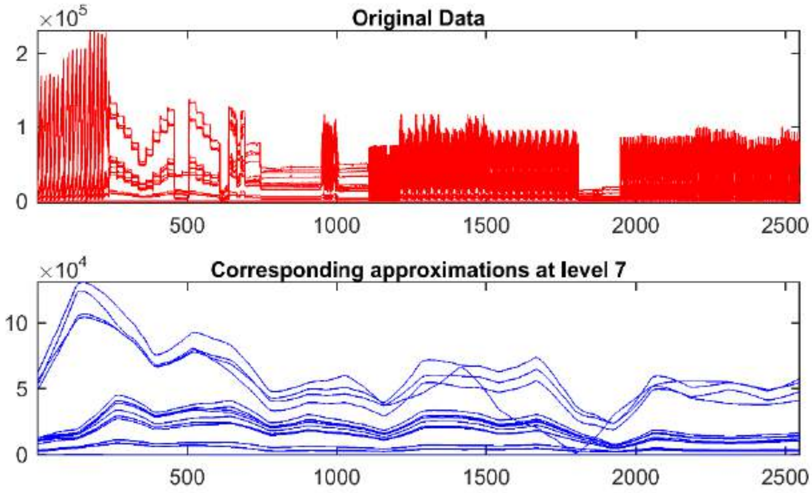
Figure 35 – MRA decomposition of the second feature in the gas 1 response signals in the first sensor



Source: Own

To complete the analysis, Figure 36 shows the MRA of the signals corresponding to Gas 1. The presence of low frequencies in the signal is observed, which implies the presence of drifts introduced in the data

Figure 36 – MRA decomposition of the full set of characteristics in the response signals of gas 1 at the first sensor



Source: Own

Finally, to carry out a more subtle simplification, the multisignal denoising is carried out. The denoising procedure carried out is summarized below:

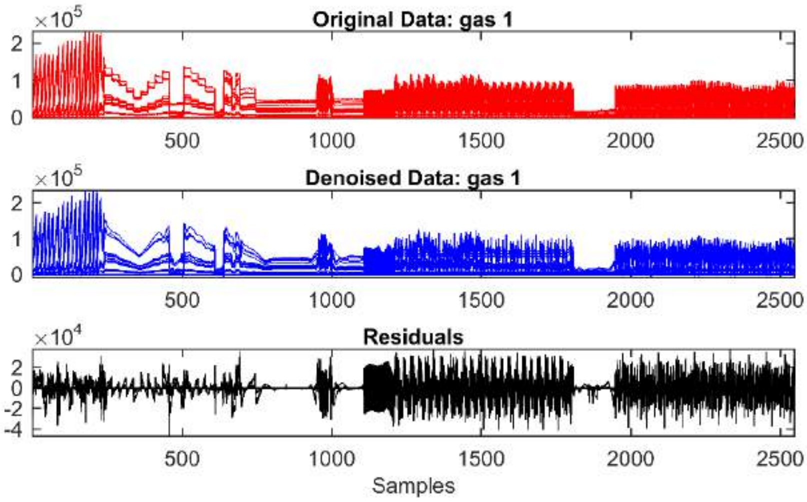
1) Decomposition: a wavelet is chosen, and a level of decomposition N, and then is computed the wavelet decompositions of the signals at level N.

2) Thresholding: For each level from 1 to N, and each signal, a threshold is selected, and thresholding is applied to the detail coefficients.

3) Reconstruction: wavelet reconstructions are computed using the original approximation coefficients of level N and the modified detail coefficients of 1 to N levels.

Let us now choose the level of decomposition N = 5 instead of N = 7 used previously.

Figure 37 – Multisignal denoising of Gas 1



Source: Own

Observing the resulting residual noise in the signals allows inferring large amounts of noise in the data, as shown in Figure 37. Consequently, the presence of these large amounts of noise and associated drifts and the existence of outliers make this database a difficult to treat and forecast data set. The prediction does not depend solely on the drift problem, but there is also the implicit instrumental and operational failure of the system and an exaggerated number of samples and gas concentrations in short periods of time. Given the above, it is difficult to predict drifts' behavior in this database when there are multiple factors associated with the deviation of the sensor response.

5.3 Time-dependent sample entropy

This section presents the analysis performed on the University of California database using time-dependent entropy, given the data's non-stationarity conditions. Time-dependent entropy is a useful method for analyzing non-stationary series, which corresponds to the quantification of the irregularity in the series at different time scales, according to the sliding window protocol. This method examines the entropy values from a time evolution perspective, allowing the application of this technique in non-stationary conditions since the series are analyzed by segments.

It is employed time-dependent SampEn statistics in overlapping sliding windows to analyze the temporal evolution of regularity of sensors responses series.

Figure 38 to Figure 43 show the results obtained for the standard values window size w = 150 and τ=2, m=2 and r=0.2 applied to the sensor response series of the University of California data. The vertical lines in each of the figures indicate the divisions by batch according to the organization of the data reported in Table 5.

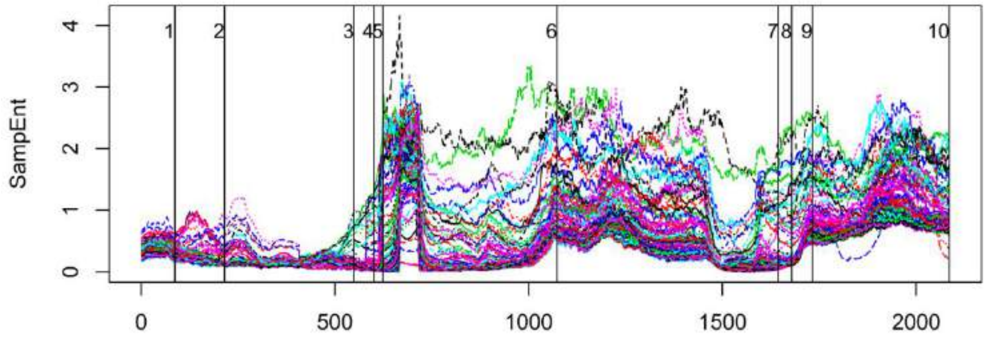
It is seen from Figure 38, Figure 40, and Figure 41 that the entropy values increase after the third batch, indicating changes in sensor response. These substantial changes in the sensor response signals generated by strong contamination of the sensors, indicating the increase in the degree of difficulty for classification given that the behavior of the series is significantly different in the initial portion of the data (Batch 1, 2, and 3) that are used for training.

Figure 39 and Figure 42 show an incremental entropy behavior from batches 6 and 7. Finally, in Figure 43, it is observed that the highest values of entropy occur from batch 6 onwards. Note that for gas F, no measurements were made in batch 3, 4, and 5.

In this analysis of the time-dependent entropy statistics values, it is found that the entropy increased in the batches in which a greater number of measurements of each of the sensed gases in this database (see Table 3). Therefore, the batches with the highest number of measurements show the highest values of entropy proportionally. This distribution of the number of measurements carried out per batch causes strong contamination of the sensors, directly associated with drifts. Therefore, it is inferred that entropy is an appropriate indicator to establish drift levels in the data.

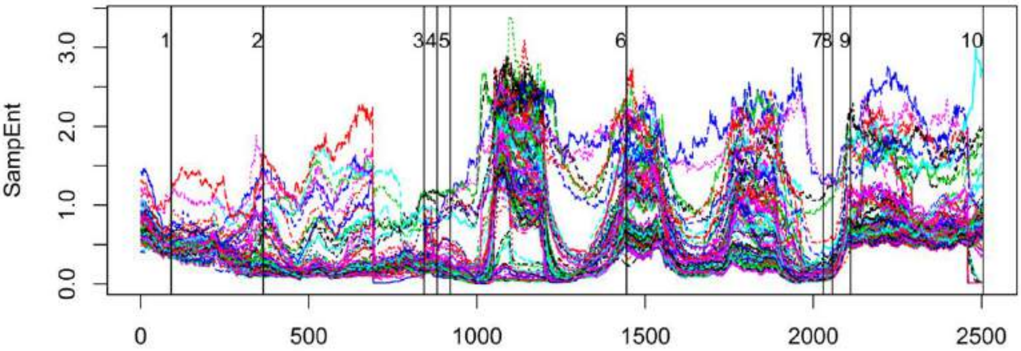
Finally, in Figure 44 the experiments carried out with windows of different sizes (w=120, w=150, and w=200) used to compare the performance of entropy at different window sizes are plotted, allowing to choose the window of 150 parameters used in Figure 38 to Figure 43.

Figure 38 – Time-dependent SampEn statistics Gas 1 (w=150). The 16-time series are presented for each of the sensors. Each color represents the time-dependent entropy values of each of the time series of the responses of each of the 16 sensors



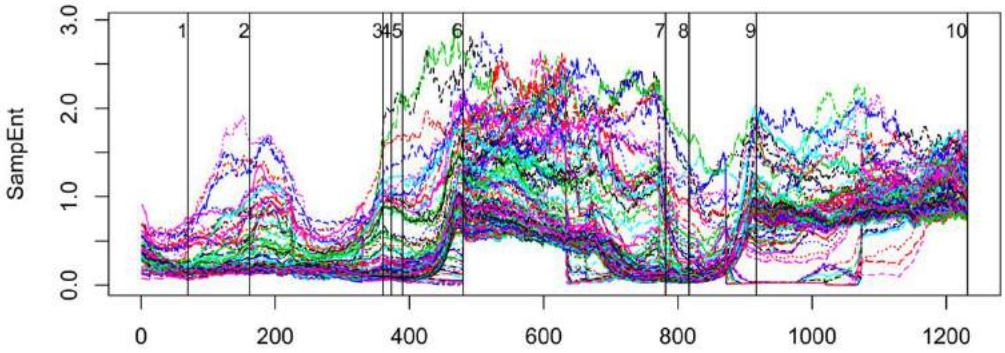
Source: Own

Figure 39 – Time-dependent SampEn statistics Gas 2 (w=150). The 16-time series are presented for each of the sensors. Each color represents the time-dependent entropy values of each of the time series of the responses of each of the 16 sensors



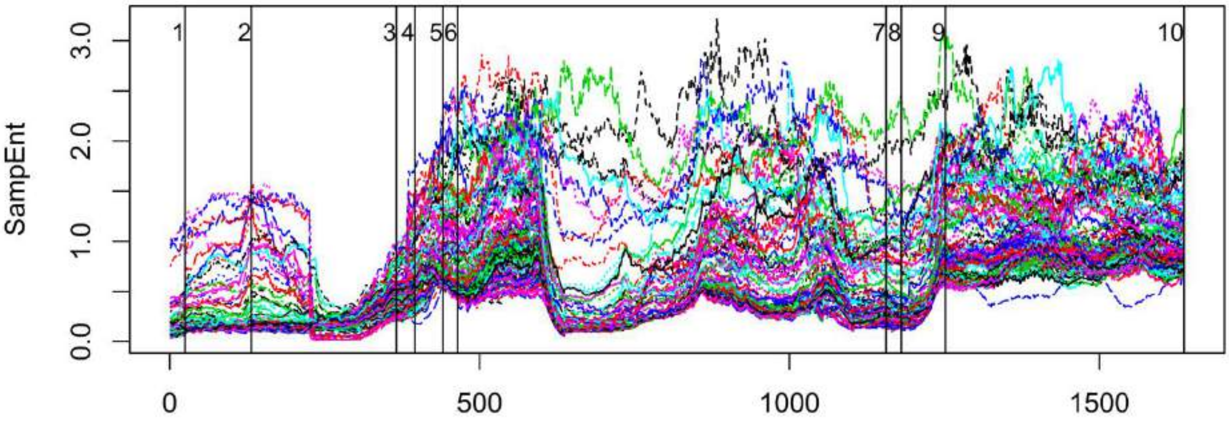
Source: Own

Figure 40 – Time-dependent SampEn statistics Gas 3 (w=150). The 16-time series are presented for each of the sensors. Each color represents the time-dependent entropy values of each of the time series of the responses of each of the 16 sensors



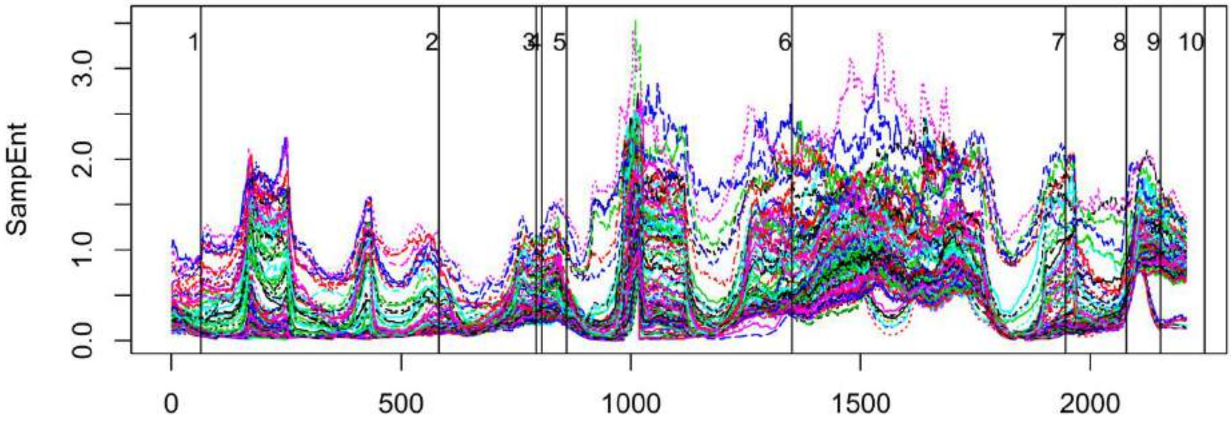
Source: Own

Figure 41 – Time-dependent SampEn statistics Gas 4 (w=150). The 16-time series are presented for each of the sensors. Each color represents the time-dependent entropy values of each of the time series of the responses of each of the 16 sensors



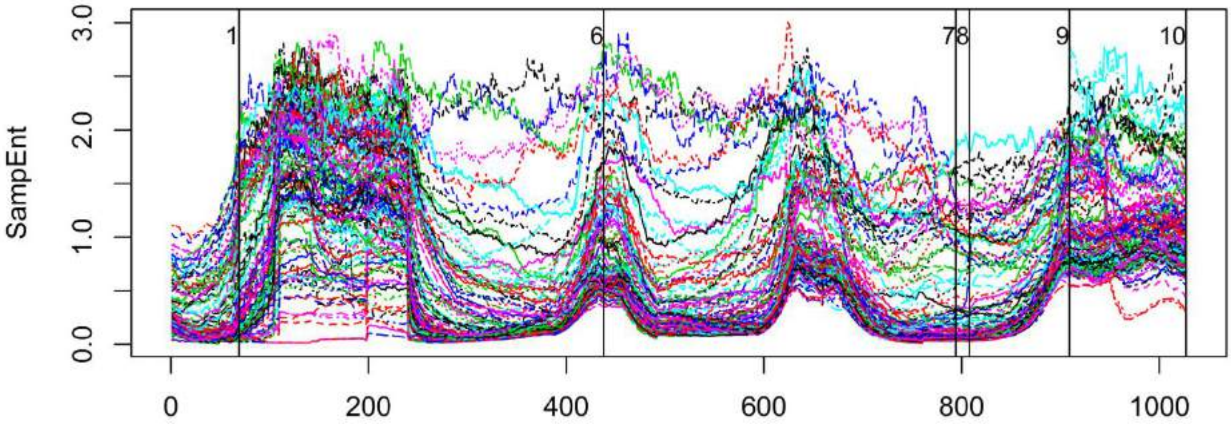
Source: Own

Figure 42 – Time-dependent SampEn statistics Gas 5 (w=150). The 16-time series are presented for each of the sensors. Each color represents the time-dependent entropy values of each of the time series of the responses of each of the 16 sensors.



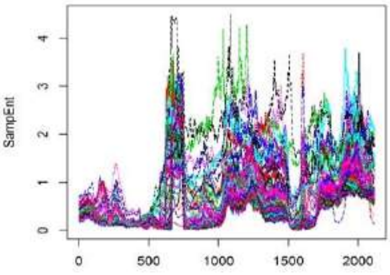
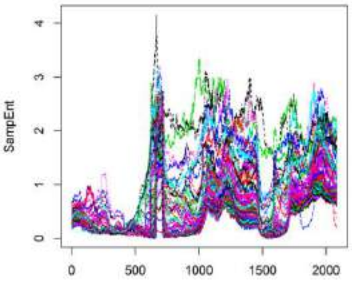
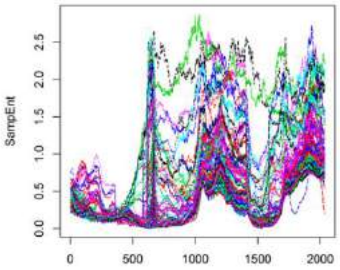
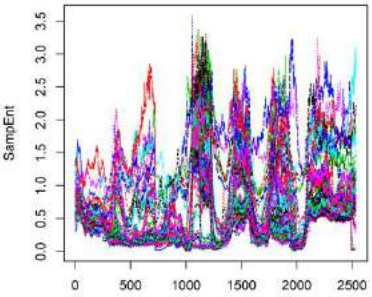
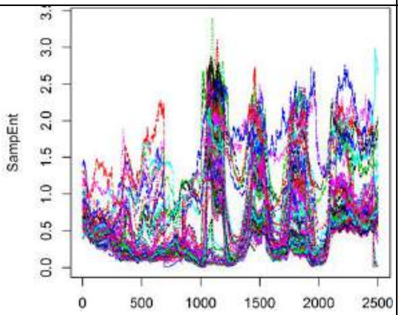
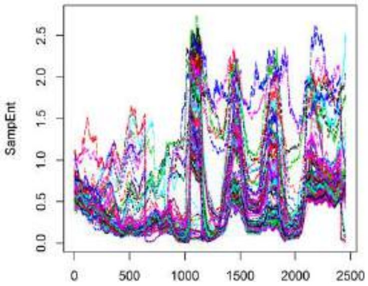
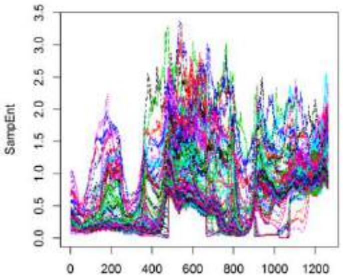
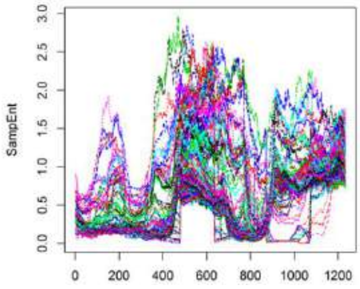
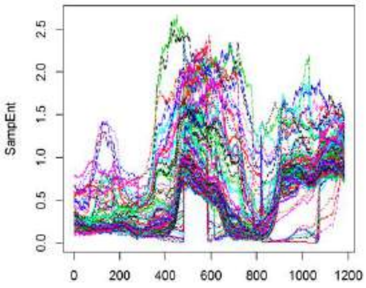
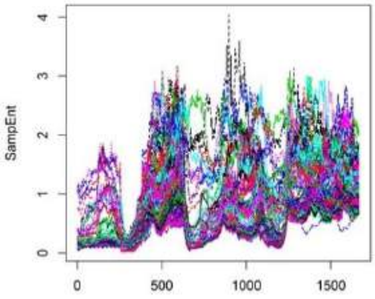
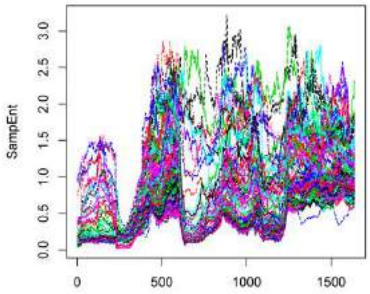
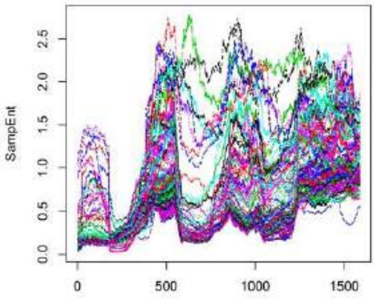
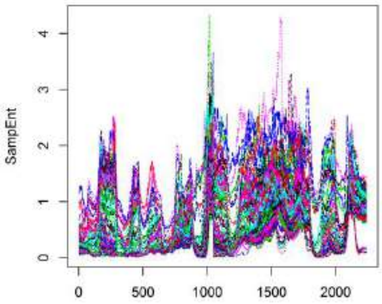
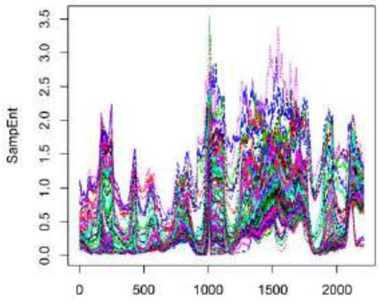
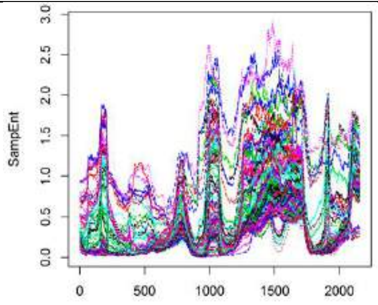
Source: Own

Figure 43 – Time-dependent SampEn statistics Gas 6 (w=150). The 16-time series are presented for each of the sensors. Each color represents the time-dependent entropy values of each of the time series of the responses of each of the 16 sensors

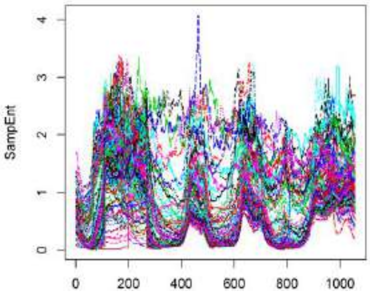
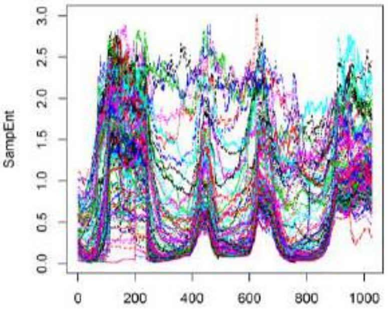
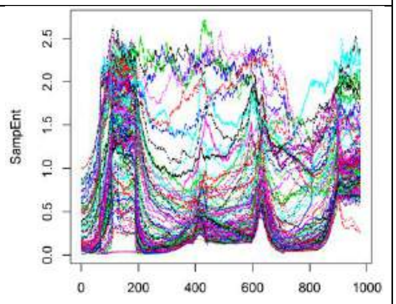


Source: Own

Figure 44 – Sliding windows time-dependent sample entropy (w=120, 150, and 200), gases one to six. Each color represents the time-dependent entropy values of each of the time series of the responses of each of the 16 sensors

	W=120	W=150	W=200
GAS 1			
GAS 2			
GAS 3			
GAS 4			
GAS 5			

GAS 6

Source: Own

To extend the results obtained so far, we also calculated the mean values and the standard deviation of the time-dependent sample entropy at w = 150 for each of the eight characteristics of the response signals from the 16 sensors. In this case, there is no discrimination by the gas class. The results presented in Table 10 and Table 11 allow us to conclude that the characteristics that reflect greater entropy are 5 and 8, which correspond to the characteristics $max_k\ ema_{\alpha=0.1}(r[k])$ and $min_k\ ema_{\alpha=0.1}(r[k])$ of the transient portion stated in Table 6.

Table 10 – Mean values of time-dependent entropy with w = 150. The lines indicate the sensor and the columns correspond to each of the 8 characteristics reported in the database.

	V1	V2	V3	V4	V5	V6	V7	V8
1	0.703142	0.533181	0.601553	0.647526	0.934113	0.57885	0.725241	1.064504
2	0.450273	0.409921	0.461482	0.499588	0.741055	0.447424	0.580043	0.963464
3	0.412897	0.402405	0.414641	0.449188	0.747394	0.418729	0.494259	0.886342
4	0.410866	0.4147	0.408238	0.45382	0.728227	0.414047	0.486605	0.895956
5	0.375759	0.395468	0.417259	0.550298	1.332696	0.409296	0.612693	1.453507
6	0.374427	0.423794	0.409276	0.557587	1.342078	0.410209	0.679574	1.500408
7	0.418979	0.394787	0.410029	0.465122	0.638835	0.409716	0.516399	0.858794
8	0.439982	0.412457	0.417168	0.457003	0.676756	0.428982	0.534808	0.905792
9	0.472879	0.389012	0.410695	0.461023	0.748409	0.412936	0.538867	0.940602
10	0.469484	0.391389	0.410923	0.443956	0.78856	0.417075	0.515061	0.960708
11	0.419265	0.386878	0.409807	0.450152	0.73429	0.413794	0.482391	0.888017
12	0.414706	0.398863	0.40165	0.455779	0.753308	0.414641	0.465935	0.885615
13	0.382689	0.411019	0.417532	0.497987	1.028584	0.420601	0.541611	1.328539
14	0.388184	0.411694	0.417933	0.498233	1.056212	0.42421	0.535794	1.319615
15	0.4138	0.401586	0.435339	0.453908	0.631544	0.432649	0.516006	0.815572
16	0.403214	0.389952	0.43126	0.464553	0.612331	0.416717	0.488135	0.803905

Source: Own

Table 11 – Standard deviation values of time-dependent entropy with w = 150. The lines indicate the sensor and the columns correspond to each of the 8 characteristics reported in the database

	V1	V2	V3	V4	V5	V6	V7	V8
1	0.613936	0.444787	0.532043	0.551232	0.573806	0.508866	0.548903	0.529923
2	0.372422	0.296939	0.381723	0.398833	0.431983	0.366787	0.426379	0.474078
3	0.331177	0.305335	0.335114	0.316677	0.455466	0.353924	0.413188	0.516662
4	0.323324	0.310676	0.336978	0.331388	0.415183	0.363607	0.397063	0.519527
5	0.309485	0.281802	0.383457	0.431214	0.695145	0.393976	0.432806	0.59271
6	0.305453	0.305317	0.317352	0.41407	0.718907	0.342549	0.469897	0.59689
7	0.334876	0.28612	0.327939	0.351257	0.399485	0.368091	0.434231	0.506667
8	0.374827	0.302314	0.319172	0.346504	0.44796	0.391255	0.422069	0.514509
9	0.422121	0.290138	0.357744	0.368707	0.430851	0.339165	0.403341	0.451341
10	0.411187	0.292518	0.356917	0.374436	0.455695	0.360406	0.387542	0.478234
11	0.326112	0.286676	0.344678	0.344237	0.442025	0.34293	0.388702	0.471782
12	0.322691	0.292132	0.350797	0.358167	0.462083	0.36235	0.381991	0.484809
13	0.31556	0.292599	0.342813	0.407702	0.601938	0.353632	0.421199	0.616405
14	0.296277	0.296396	0.353372	0.40225	0.642161	0.376216	0.393705	0.630479
15	0.323979	0.27901	0.343234	0.371028	0.408027	0.364651	0.426676	0.44122
16	0.317279	0.275668	0.353437	0.375903	0.388029	0.383684	0.397775	0.446237

Source: Own

5.4 Rapid Detection Method for E-Nose using Synthetic Data

As explained in section 4.3, the work scenarios generated with the *chemosensors* package were tested to analyze the classifiers' performance applying the traditional approach versus the rapid detection approach over data with and without drift.

The traditional approach uses the entire response curves, implementing preprocessing techniques to extract the features and later data processing. Moreover, the rapid detection approach is based on processing an early portion of raw signals and a rising window protocol. The following tables summarized the experiments performed to achieve results.

Based on the experimental results, there are some exciting things to remark:

1. As expected, all classifiers made good forecasts in the first scenario (without noise and drift).

2. All classifiers worked well in the fourth scenario, datasets 4A, 4B, and 4C (with noise and without drift), meaning that the noise can be treated without problem.

3. Although the second scenario seems to be more complex because it includes noise and drift, the results suggest that the third scenario (without noise and with drift) turned out to be more challenging.

4. SVM outperformed the performance compared against the other classifiers, followed by the Sniff Resnet model, which dealt well with the more challenging scenario.

5. The results let us infer that the models generated using the rapid detection approach dealt very well under the work scenarios, with the advantage of lets to achieve faster results using only an early portion of the signals.

Table 12 – Classification accuracy rates using SVM to compare the conventional and the rapid detection approach over the synthetic database. FS: Feature Selection pre-processing. W1: Window1, W2: Window2, …, W10: Window10. Scenarios as described in Section 4.3

Scenarios	Task	FS	W1	W2	W3	W4	W5	W6	W7	W8	W9	W10
1	Train	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
	Valid	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
2A	Train	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
	Valid	0.96	0.96	0.96	0.96	0.97	0.95	0.95	0.95	0.93	0.92	0.91
2B	Train	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
	Valid	0.95	0.95	0.95	0.95	0.96	0.95	0.93	0.92	0.90	0.85	0.81
2C	Train	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	0.99	0.99
	Valid	0.96	0.93	0.94	0.92	0.91	0.89	0.87	0.84	0.76	0.69	0.67
3A	Train	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
	Valid	0.96	0.91	0.84	0.80	0.77	0.74	0.76	0.80	0.80	0.79	0.78
3B	Train	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
	Valid	0.94	0.93	0.78	0.74	0.74	0.72	0.72	0.70	0.68	0.68	0.68
3C	Train	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
	Valid	0.94	0.90	0.76	0.72	0.69	0.68	0.66	0.64	0.65	0.67	0.66
4A	Train	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
	Valid	1.00	1.00	1.00	0.99	0.95	0.94	0.93	0.93	0.93	0.93	0.93
4B	Train	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
	Valid	0.97	0.97	0.98	0.94	0.94	0.93	0.92	0.92	0.92	0.92	0.92

4C	Train	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
	Valid	0.98	0.99	0.97	0.94	0.91	0.96	0.94	0.94	0.94	0.94	0.94

Source: Own

Table 13 – Classification accuracy rates using MLP to compare the conventional and the rapid detection approach over the synthetic database. FS: Feature Selection pre-processing. W1: Window1, W2: Window2, …, W10: Window10. Scenarios as described in Section 4.3

Scenarios	Task	FS	W1	W2	W3	W4	W5	W6	W7	W8	W9	W10
1	Train	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
	Valid	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
2A	Train	0.83	0.76	0.76	0.85	0.80	0.96	0.76	0.80	0.76	0.99	0.87
	Valid	0.73	0.60	0.63	0.72	0.53	0.80	0.67	0.60	0.63	0.67	0.66
2B	Train	0.86	0.85	0.70	0.68	0.75	0.70	0.80	0.76	0.87	0.70	0.83
	Valid	0.70	0.60	0.47	0.57	0.53	0.47	0.60	0.54	0.58	0.53	0.59
2C	Train	0.87	0.70	0.61	0.70	0.68	0.72	0.69	0.94	0.76	0.70	0.68
	Valid	0.62	0.55	0.53	0.53	0.43	0.44	0.56	0.69	0.59	0.43	0.54
3A	Train	0.85	0.97	0.99	0.82	0.80	0.90	0.96	1.00	0.86	0.90	0.93
	Valid	0.77	0.60	0.52	0.53	0.51	0.46	0.59	0.74	0.53	0.59	0.60
3B	Train	0.65	0.63	0.85	0.75	0.77	0.80	0.92	0.75	0.83	0.79	0.92
	Valid	0.48	0.46	0.48	0.57	0.53	0.51	0.62	0.58	0.68	0.59	0.63
3C	Train	0.70	0.69	0.66	0.79	0.58	0.90	0.83	0.76	0.77	0.76	0.69
	Valid	0.57	0.41	0.38	0.51	0.32	0.63	0.50	0.58	0.60	0.63	0.43
4A	Train	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
	Valid	0.98	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	0.99
4B	Train	0.70	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
	Valid	0.64	0.92	0.99	0.99	0.96	0.97	0.98	0.97	1.00	0.92	0.93
4C	Train	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
	Valid	1.00	0.97	0.98	1.00	0.86	1.00	0.93	1.00	1.00	1.00	1.00

Source: Own

Table 14 – Classification accuracy rates using Sniff ConvNet to compare the conventional and the rapid detection approach over the synthetic database. FS: Feature Selection pre-processing. W1: Window1, W2: Window2, …, W10: Window10. Scenarios as described in Section 4.3

Scenarios	Task	FS	W1	W2	W3	W4	W5	W6	W7	W8	W9	W10
1	Train	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	0.70	1.00
	Valid	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	0.66	1.00
2A	Train	1.00	0.96	0.99	0.89	0.92	0.90	0.92	0.54	0.99	0.89	0.86
	Valid	0.96	0.63	0.95	0.64	0.58	0.68	0.82	0.56	0.71	0.69	0.59
2B	Train	0.89	0.70	0.86	0.93	0.80	0.66	0.87	0.76	0.69	0.90	0.72
	Valid	0.82	0.48	0.56	0.72	0.62	0.47	0.70	0.53	0.56	0.66	0.55

2C	Train	0.90	0.93	0.75	0.82	0.80	0.89	0.80	0.87	0.90	0.77	0.83
	Valid	0.76	0.66	0.50	0.64	0.51	0.69	0.59	0.68	0.60	0.62	0.68
3A	Train	0.99	0.87	0.87	1.00	0.90	0.75	0.94	0.92	0.87	0.97	0.94
	Valid	0.90	0.55	0.48	0.60	0.50	0.69	0.63	0.46	0.52	0.60	0.60
3B	Train	0.96	0.83	0.83	1.00	0.70	0.86	0.77	0.70	0.82	0.90	0.70
	Valid	0.92	0.45	0.49	0.59	0.38	0.42	0.45	0.35	0.44	0.50	0.39
3C	Train	0.90	0.85	0.72	0.97	0.87	0.96	0.85	0.87	0.87	0.80	0.92
	Valid	0.87	0.53	0.43	0.52	0.38	0.67	0.52	0.43	0.54	0.41	0.38
4A	Train	1.00	1.00	1.00	0.73	0.82	1.00	0.83	1.00	1.00	1.00	1.00
	Valid	0.90	1.00	1.00	0.99	0.85	1.00	0.99	1.00	1.00	1.00	1.00
4B	Train	1.00	1.00	1.00	1.00	1.00	1.00	0.93	1.00	1.00	0.94	1.00
	Valid	0.99	0.99	1.00	1.00	0.98	0.99	0.88	0.99	0.95	0.96	1.00
4C	Train	1.00	1.00	1.00	0.97	1.00	1.00	1.00	1.00	1.00	1.00	0.62
	Valid	0.96	0.98	1.00	0.95	1.00	1.00	1.00	1.00	1.00	1.00	0.68

Source: Own

Table 15 – Classification accuracy rates using Sniff Resnet to compare the conventional and the rapid detection approach over the synthetic database. FS: Feature Selection pre-processing. W1: Window1, W2: Window2, …, W10: Window10. Scenarios as described in Section 4.3

Scenarios	Task	FS	W1	W2	W3	W4	W5	W6	W7	W8	W9	W10
1	Train	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
	Valid	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
2A	Train	0.97	0.99	0.94	1.00	0.79	1.00	1.00	1.00	0.62	0.86	1.00
	Valid	0.83	0.84	0.67	0.97	0.61	0.83	0.92	0.88	0.49	0.58	0.78
2B	Train	0.92	1.00	0.86	1.00	0.94	0.99	0.87	0.89	0.89	1.00	0.96
	Valid	0.81	0.80	0.74	0.78	0.86	0.79	0.75	0.72	0.65	0.93	0.76
2C	Train	1.00	0.97	1.00	0.76	0.96	0.94	0.97	0.82	0.82	0.75	0.69
	Valid	0.84	0.88	0.81	0.59	0.65	0.92	0.67	0.59	0.59	0.49	0.52
3A	Train	0.99	1.00	0.97	1.00	0.99	0.97	0.99	0.94	0.99	0.96	0.89
	Valid	0.95	0.65	0.65	0.65	0.74	0.53	0.54	0.60	0.70	0.62	0.52
3B	Train	0.94	1.00	1.00	0.97	1.00	1.00	0.89	0.58	0.85	0.89	0.94
	Valid	0.81	0.81	0.62	0.64	0.61	0.80	0.62	0.41	0.52	0.56	0.69
3C	Train	0.79	0.97	0.99	1.00	0.99	0.96	1.00	0.96	0.87	0.90	0.97
	Valid	0.57	0.47	0.73	0.82	0.53	0.59	0.59	0.61	0.59	0.67	0.54
4A	Train	1.00	1.00	1.00	1.00	0.90	1.00	1.00	1.00	1.00	0.76	1.00
	Valid	0.99	1.00	1.00	1.00	0.83	1.00	0.98	0.92	1.00	0.79	0.98
4B	Train	0.94	1.00	0.99	1.00	1.00	1.00	1.00	1.00	1.00	0.96	0.99
	Valid	0.91	1.00	0.95	0.99	0.96	0.99	0.92	0.85	0.81	0.71	0.85
4C	Train	1.00	1.00	1.00	0.99	1.00	1.00	1.00	1.00	0.79	0.99	0.87
	Valid	0.91	0.99	0.99	0.96	0.75	0.99	0.96	0.97	0.57	0.79	0.86

Source: Own

Table 16 – Classification accuracy rates using Sniff Multinose to compare the conventional and the rapid detection approach over the synthetic database. FS: Feature Selection pre-processing. W1: Window1, W2: Window2, ..., W10: Window10. Scenarios as described in Section 4.3

Scenarios	Task	FS	W1	W2	W3	W4	W5	W6	W7	W8	W9	W10
1	Train	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
	Valid	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
2A	Train	1.00	0.94	0.86	0.99	0.99	1.00	1.00	1.00	0.99	1.00	1.00
	Valid	0.98	0.71	0.58	0.66	0.69	0.66	0.82	0.67	0.68	0.73	0.76
2B	Train	0.90	0.66	0.83	0.87	0.93	0.96	0.94	0.99	0.90	0.97	0.96
	Valid	0.85	0.39	0.50	0.46	0.48	0.54	0.66	0.66	0.65	0.65	0.61
2C	Train	0.93	0.65	0.77	0.72	0.82	0.85	0.93	0.93	0.90	0.90	0.89
	Valid	0.74	0.35	0.41	0.38	0.48	0.45	0.57	0.55	0.61	0.57	0.54
3A	Train	0.99	0.97	1.00	0.92	0.99	1.00	0.99	0.99	1.00	1.00	1.00
	Valid	0.97	0.63	0.60	0.42	0.48	0.61	0.58	0.66	0.60	0.66	0.66
3B	Train	0.83	0.80	0.85	0.92	0.96	0.93	0.96	0.99	0.94	0.96	0.99
	Valid	0.67	0.54	0.50	0.60	0.46	0.43	0.57	0.63	0.62	0.60	0.58
3C	Train	0.96	0.76	0.79	0.94	0.92	0.94	0.93	0.93	0.97	0.89	0.89
	Valid	0.88	0.36	0.40	0.44	0.41	0.40	0.49	0.59	0.47	0.63	0.58
4A	Train	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
	Valid	0.98	1.00	1.00	1.00	1.00	1.00	1.00	1.00	0.99	1.00	1.00
4B	Train	0.96	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
	Valid	0.91	0.97	0.99	0.97	0.98	0.94	0.97	0.92	1.00	1.00	0.98
4C	Train	0.99	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
	Valid	0.84	1.00	1.00	1.00	0.97	1.00	1.00	1.00	1.00	1.00	1.00

Source: Own

6 BIBLIOGRAPHICAL PRODUCTION

This section lists the papers published during the doctoral studies at the *Universidade Federal Rural de Pernambuco* (UFRPE) in the post-graduate program of Biometria e Estatística Aplicada.

Complexity analysis of Brazilian agriculture and energy market. **Physica A: Statistical Mechanics and its Applications** (ALBARRACÍN E. *et al.*, 2019)

Wine quality rapid detection using a compact electronic nose system: application focused on spoilage thresholds by acetic acid. **LWT Food Science and Technology** (RODRIGUEZ GAMBOA, Juan C.; ALBARRACIN E., Eva Susana; SILVA, Adenilton J. Da; *et al.*, 2019).

Electronic nose dataset for detection of wine spoilage thresholds. **Data in Brief** (RODRIGUEZ GAMBOA, Juan C.; ALBARRACIN E., Eva Susana; SILVA, Adenilton J. DA; E. FERREIRA, Tiago A., 2019)

Validation of the rapid detection approach for enhancing the electronic nose systems performance, using different deep learning models and support vector machines. **Sensors and Actuators, B: Chemical** (RODRIGUEZ GAMBOA, Juan C *et al.*, 2021)

7 CONCLUSION

During the doctoral studies, it was developed an E-Nose system. That device was used to obtain a database to detect wine spoilage thresholds, using an array of six metal-oxide-semiconductor (MOS) gas sensors. Using this database, we proposed a novel approach for the electronic nose systems, treating an early portion of the raw signals (while the measurement process is still running and without applying preprocessing techniques.) The proposed methodology focused on reducing the necessary time for making the forecast, accelerating the response time. The mentioned approach achieved excellent results against the traditional methodology. The database is available[1].

The time dependent SampEn statistic in overlapping sliding windows, used to analyze the temporal evolution of the regularity of the series of responses from the sensors, is a consistent statistic to analyze drift conditions in databases of E-Nose systems. The coincidence was identified between the increase in the entropy value and the increase in the sensors' contamination caused by their saturation in carrying out many consecutive measurements in short periods. This generates the sensors' poisoning, and entropy becomes an indicator of this poisoning that may not be noticeable to the operator. The quantification of this parameter makes it possible to generate a protocol for the distribution in time of the measurements made with an artificial smell system. Likewise, it was concluded that the forecast fails does not depend solely on the drift problem, but there is also the implicit instrumental and operational failure of the system and an exaggerated number of samples and gas concentrations in short periods of time. Given the above, it is difficult to predict drifts' behavior in the University of California database when there are multiple factors associated with the deviation of the sensor response.

[1] https://data.mendeley.com/datasets/vpc887d53s/3

The results presented allow us to conclude that the characteristics that reflect greater entropy are the characteristics $max_k\, ema_{\alpha=0.1}(r[k])$ and $min_k\, ema_{\alpha=0.1}(r[k])$ of the transient portion of the signal.

Subsequently, we focused on validating if the rapid detection approach proposed for the electronic nose is suitable to be applied in diverse E-Nose scenarios with and without noise and drift. Consequently, the rapid detection approach could deal very well under drift and noise conditions based on the experimental results with the advantage of lets to achieve faster results using only an early portion of the signals against the traditional approach that needs the whole measurement information. Moreover, an interesting finding suggests that a drift scenario and without noise could be more complex against a scenario with drift and noise. Maybe parts of information are contained in the noise. Likewise, it was evidenced that the classifiers dealt well under noise conditions. SVM classifier outperformed the performance compared against the other classifiers, followed by the Sniff Resnet model that is competitive in the more challenging scenario (with drift and without noise).

It is recommended as future work to propose a method to quantify drifts' effect based on the dynamic principles of entropy to establish a soft metrological scheme that quantifies the contamination of an E-Nose database, intending to determine rest times of the sensor instrument and/or sensors replacement by aging.

REFERENCES

ALBARRACÍN E., E. S. *et al.* Complexity analysis of Brazilian agriculture and energy market. **Physica A: Statistical Mechanics and its Applications**, jun. 2019. v. 523, p. 933–941.

ARTURSSON, T. *et al.* Drift correction for gas sensors using multivariate methods. **Journal of Chemometrics**, 2000. v. 14, n. 5–6, p. 711–723.

BALASIS, G. *et al.* Investigating dynamical complexity in the magnetosphere using various entropy measures. **Journal of Geophysical Research: Space Physics**, set. 2009. v. 114, n. A9, p. n/a-n/a.

BUCK, T.M. AND ALLEN, F.G. AND DALTON, J. V. **Detection of Chemical Species by Surface Effects on Metals and Semiconductors**. Volume 496 ed. [S.l.]: Bell Telephone Laboratories, 1965.

CHOU, C.-M. Complexity analysis of rainfall and runoff time series based on sample entropy in different temporal scales. **Stochastic environmental research and risk assessment**, 2014. v. 28, n. 6, p. 1401–1408.

DRAVNIEKS, A.; TROTTER, P. J. Polar vapour detector based on thermal modulation of contact potential. **Journal of Scientific Instruments**, ago. 1965. v. 42, n. 8, p. 624–627.

DURÁN, C.; VELÁSQUEZ, A.; GUALDRON, O. Implementación de una nariz electrónica para detectar pacientes con EPOC desde el aliento exhalado. **Ingeniería y Desarollo**, 2012. v. 30, n. 2, p. 143–159.

FONOLLOSA, J.; RODRÍGUEZ-LUJÁN, I.; HUERTA, Ramón. Chemical gas sensor array dataset. **Data in Brief**, 2015. v. 3, p. 85–89.

GARDNER, J. W.; BARTLETT, P. N. A brief history of electronic noses. **Sensors and Actuators B: Chemical**, 1 mar. 1994. v. 18, n. 1–3, p. 210–211.

GAROUSI, M. R.; SHAKARAMI, M. R.; NAMDARI, F. Detection and classification of power quality disturbances using parallel neural networks based on discrete wavelet transform. **Journal of Electrical Systems**, 2016. v. 12, n. 1, p. 158–173.

GUERRINI, L. *et al.* Smelling, Seeing, Tasting - Old Senses for New Sensing. **ACS Nano**, 2017. v. 11, n. 6, p. 5217–5222.

GUTIERREZ-OSUNA, R. Pattern analysis for machine olfaction: A review. **IEEE Sensors Journal**, 2002. v. 2, n. 3, p. 189–202.

HU, W. *et al.* Electronic Noses: From Advanced Materials to Sensors Aided with Data Processing. **Advanced Materials Technologies**, 14 dez. 2018. v. 4, n. 2, p. 1800488.

JHA, S. K. *et al.* Recognition and sensing of organic compounds using analytical methods, chemical sensors, and pattern recognition approaches. **Chemometrics and Intelligent Laboratory Systems**, 15 fev. 2019. v. 185, p. 18–31.

LAKE, D. E. *et al.* Sample entropy analysis of neonatal heart rate variability. **American Journal of Physiology-Regulatory, Integrative and Comparative Physiology**, 2002. v. 283, n. 3, p. R789--R797.

LÄNGKVIST, M. *et al.* Fast Classification of Meat Spoilage Markers Using Nanostructured ZnO Thin Films and Unsupervised Feature Learning. **Sensors**, jan. 2013. v. 13, n. 2, p. 1578–1592.

LIU, Y.-J.; MENG, Q.-H.; ZHANG, X.-N. Data Processing for Multiple Electronic Noses Using Sensor Response Visualization. **IEEE Sensors Journal**, nov. 2018. v. 18, n. 22, p. 9360–9369.

YING-JIE, L.; ZENG, M.; MENG, Q.-H. Electronic nose using a bio-inspired neural network modeled on mammalian olfactory system for Chinese liquor classification. **Review of Scientific Instruments**, fev. 2019. v. 90, n. 2, p. 025001.

LOUTFI, A. *et al.* Electronic noses for food quality: A review. **Journal of Food Engineering**, 1 jan. 2015. v. 144, p. 103–111.

MONCRIEFF, R. W. An instrument for measuring and classifying odors. **Journal of applied physiology**, jul. 1961. v. 16, p. 742–9.

MARTINA, E. *et al.* Multiscale entropy analysis of crude oil price dynamics.

Energy Economics, 2011. v. 33, n. 5, p. 936–947.

MONCRIEFF, R. W. An instrument for measuring and classifying odors. **Journal of applied physiology**, jul. 1961. v. 16, p. 742–9.

MONROY, J. G. *et al.* Chemometrics and Intelligent Laboratory Systems Continuous chemical classification in uncontrolled environments with sliding windows. **Chemometrics and Intelligent Laboratory Systems**, 2016. v. 158, p. 117–129.

MUEZZINOGLU, M. K. *et al.* Acceleration of chemo-sensory information processing using transient features. **Sensors and Actuators, B: Chemical**, 2009. v. 137, n. 2, p. 507–512.

OLIVEIRA, B. R. DE *et al.* A wavelet-based method for power-line interference removal in ECG signals. **Research on Biomedical Engineering**, 2018. v. 34, n. 1, p. 73–86.

PENG, P. *et al.* Gas Classification Using Deep Convolutional Neural Networks. **Sensors**, jan. 2018. v. 18, n. 2, p. 157.

PERSAUD, Krishna; DODD, G. Analysis of discrimination mechanisms in the mammalian olfactory system using a model nose. **Nature**, set. 1982. v. 299, n. 5881, p. 352–355.

PINCUS, S. M. Approximate entropy as a measure of system complexity. **Proceedings of the National Academy of Sciences of the United States of America**, mar. 1991. v. 88, n. 6, p. 2297–301.

PONZONI, A. *et al.* Metal oxide gas sensors, a survey of selectivity issues addressed at the SENSOR lab, Brescia (Italy). **Sensors (Switzerland)**, 2017. v. 17, n. 4.

QI, P.-F.; MENG, Q.-H.; ZENG, M. A CNN-based simplified data processing method for electronic noses. [S.l.]: IEEE, 2017. p. 1–3.

RICHMAN, J. S. *et al.* Physiological time-series analysis using approximate entropy and sample entropy. **Am J Physiol Heart Circ Physiol**, 2000. v. 278, p. 2039–2049.

RODRÍGUEZ-GAMBOA, J. C.; ALBARRACÍN-ESTRADA, E. S.; DELGADO-TREJOS, E. Quality Control Through Electronic Nose System. **Modern Approaches To Quality Control**, 2011. p. 505–522.

RODRÍGUEZ-MÉNDEZ, M. L. *et al.* Electronic Noses and Tongues in Wine Industry. **Frontiers in Bioengineering and Biotechnology**, 25 out. 2016. v. 4, p. 81.

RODRIGUEZ GAMBOA, J.C. *et al.* Mendeley Data - Electronic nose dataset for detection of wine spoilage thresholds. **Mendeley Data**, 2019.

RODRIGUEZ GAMBOA, Juan C.; ALBARRACIN E., Eva Susana; SILVA, Adenilton J. Da; *et al.* Wine quality rapid detection using a compact electronic nose system: application focused on spoilage thresholds by acetic acid. **LWT - Food Science and Technology**, 2019. v. 108, p. 377–384.

RODRIGUEZ GAMBOA, Juan C *et al.* Validation of the rapid detection approach for enhancing the electronic nose systems performance, using different deep learning models and support vector machines. **Sensors and Actuators, B: Chemical**, jan. 2021. v. 327, p. 128921.

RODRÍGUEZ, J.; DURÁN, C.; REYES, A. Electronic nose for quality control of Colombian coffee through the detection of defects in "Cup Tests". **Sensors (Basel, Switzerland)**, jan. 2010. v. 10, n. 1, p. 36–46.

SHUANGCHENG, L. *et al.* Measurement of climate complexity using sample entropy. **International journal of climatology**, 2006. v. 26, n. 15, p. 2131–2139.

STOSIC, Darko *et al.* Correlations of multiscale entropy in the FX market. **Physica A: Statistical Mechanics and its Applications**, 2016. v. 457, p. 52–61.

VERGARA, A. *et al.* Chemical gas sensor drift compensation using classifier ensembles. **Sensors and Actuators B: Chemical**, 2012. v. 166, p. 320–329.

VERGARA, A. *et al.* On the performance of gas sensor arrays in open sampling systems using Inhibitory Support Vector Machines. **Sensors and Actuators, B: Chemical**, 2013. v. 185, p. 462–477.

WEI, G. *et al.* Development of a LeNet-5 Gas Identification CNN Structure for Electronic Noses. **Sensors (Basel, Switzerland)**, 8 jan. 2019. v. 19, n. 1.

WENG, W. C. *et al.* Altered resting-state EEG complexity in children with tourette syndrome: A preliminary study. **Neuropsychology**, 1 maio. 2017. v. 31, n. 4, p. 395–402.

WILKENS, W. F.; HARTMAN, J. D. An electronic analog for the olfactory processes. **Annals of the New York Academy of Sciences**, 1 jul. 1964. v. 116, n. 2 Recent Advanc, p. 608–612.

YAN, J. *et al.* Electronic Nose Feature Extraction Methods: A Review. **Sensors**, 2015. v. 15, n. 11, p. 27804–27831.

ZHANG, L.; ZHANG, D. Efficient Solutions for Discreteness , Drift , and Disturbance (3D) in Electronic Olfaction. **IEEE Transactions on Systems, Man, and Cybernetics: Systems**, 2018. v. 48, n. 2, p. 242–254.

ZHAO, Z.; YANG, S. Sample entropy-based roller bearing fault diagnosis method. **Journal of Vibration and Shock**, 2012.

ZIYATDINOV, A. *et al.* Drift compensation of gas sensor array data by common principal component analysis. **Sensors and Actuators, B: Chemical**, 2010. v. 146, n. 2, p. 460–465.

ZIYATDINOV, Andrey; PERERA-LLUNA, A. Data simulation in machine olfaction with the R package chemosensors. **PLoS ONE**, 2014. v. 9, n. 2.

PUBLISHED PAPER: COMPLEXITY ANALYSIS OF BRAZILIAN AGRICULTURE AND ENERGY MARKET

Physica A 523 (2019) 933–941

Contents lists available at ScienceDirect

Physica A

journal homepage: www.elsevier.com/locate/physa

Complexity analysis of Brazilian agriculture and energy market

Eva Susana Albarracín E., Juan C. Rodríguez Gamboa, Elaine C.M. Marques, Tatijana Stosic [*]

Departamento de Estatística e Informática, Universidade Federal Rural de Pernambuco, Rua Dom Manoel de Medeiros s/n, Dois Irmãos, 52171-900 Recife, PE, Brazil

HIGHLIGHTS

- We study complexity of ethanol and sugar prices before and after the crisis.
- Market efficiency of both commodities increases after global and European crisis.
- Energy market is less regular (more efficient) than the agricultural market.

ARTICLE INFO

Article history:
Received 12 November 2018
Available online 9 April 2019

Keywords:
Multiscale entropy
Biofuel
Agriculture market

ABSTRACT

We investigate regularity and asynchrony in Brazilian energy (ethanol) and agriculture (sugar) market with focus on 2008 global economic crisis, using multiscale entropy method. We applied this method on sugar and ethanol return series for different temporal scales and in sliding windows to analyze temporal evolution of regularity of price dynamics. The results show that for both ethanol and sugar return series the entropy values increase after 2008 and 2012, indicating the increase of market efficiency in post-crisis periods. During the crisis periods sugar and ethanol return series present some deviations from the expected decreasing behavior for higher timescales, which is more evident for the ethanol. Overall, higher entropy values are found for ethanol series indicating less regularity and higher market efficiency in energy market.

1. Introduction

Sugarcane is the dominant raw material used in Brazil in the manufacture of sugar and ethanol, where plant residues (bagasse, tops and leafs) are used for energy generation (steam and electricity for the production itself, as well as surplus electricity) [1,2]. This agricultural product has proved valuable for the country's economy because its derivatives, crystal sugar and hydrated ethanol, represent two of the leading commodities traded on the stock market. Brazil has been recognized as the largest producer of sugarcane in the world for several years [3], however, the country faces various socioeconomic and environmental issues related to sugarcane expansion: displacement of extensive livestock and soybeans production to remote areas (mostly to the border of Amazon region) which leads to additional deforestation, increased land-market dynamics — merging of small properties into larger units (more feasible for large-scale sugarcane production) which is accompanied by use of heavy agricultural machinery, and management practices that lead to soil erosion and degradation of soil physical properties [4,5]. One consequence of sugarcane burning is the increased

[*] Corresponding author.
E-mail address: tastosic@gmail.com (T. Stosic).

https://doi.org/10.1016/j.physa.2019.04.134
0378-4371/© 2019 Published by Elsevier B.V.

concentration of aerosol particles which can cause serious health problems for local population. Finally, as most of sugarcane in Brazil today is still harvested manually, improvement of working conditions for sugarcane cutters is a big challenge for sugarcane employers and Brazilian Ministry of Labor [4]. All these concerns are directly related to the increase in ethanol production and its competition with sugar production, making the two markets highly interconnected, being both influenced by global factors (crude oil prices) and specific local features of Brazilian economic development, including government policies and technological advances such as flex plants which can adjust production of these commodities depending on their relative prices. The relation between Brazilian ethanol and sugar market was studied using econometric methods [6–11], however the emerging methods from complex system science can reveal some new aspects of the nature of the stochastic processes that generate financial temporal series [12–17]. The objective of this work is to investigate the temporal variability of ethanol and sugar prices, employing entropy, a classical concept of information theory widely used to quantify disorder and uncertainty of dynamic systems [18]. Thus, we use multiscale sample entropy, time-dependent sample entropy, and cross-sample entropy [19,20] for analyzing the influence of the economic recession on these time series, particularly the subprime crisis (2008–2010) which was triggered since 2006 in the United States with the bankruptcy of mortgage loans and strongly influenced the whole world stock markets [21]. The results show how this approach can be used for analysis of financial time series, revealing their complexity or similarity, with the potential of detecting economic recession.

The rest of this paper is organized as follows: Section 2 describes the data and the methods employed for the analysis of entropy; Section 3 provides a comprehensive description of the empirical results, and Section 4 presents the discussions and conclusions.

2. Materials and methods

2.1. Data

We analyze two time series with 875 observations of weekly prices of crystal sugar (in USD/50 kg) and hydrated ethanol fuel (in USD/liter) [22]. These values correspond to the closing prices on Fridays of each week during the period from July 7, 2000 to May 19, 2017.

The temporal series of sugar and ethanol prices are shown in Fig. 1(a) and (b), respectively, and their corresponding series of returns calculated as $R_t = \ln(P_t/P_{t-1})$ are shown in Fig. 1(c) and (d), where P_S and P_E correspond to the weekly prices of crystal sugar and hydrated ethanol respectively, and R_S and R_E represent the corresponding return series.

The shaded area in Fig. 1 represents the period from mid-September 2008 to December 2010, when the US economy fell into a severe recession influencing the behavior of these commodities (increased prices, higher returns and volatilities). The increase of prices in the mid 2000s, was partly driven by a surge in US demand, followed with decrease due to collapse in crude oil prices in the second half of 2008. Ethanol prices recovered in 2009 as a result of increased global demand for sugar resulting in strong increases in sugar prices, which were passed on to ethanol prices, until the break of 2011, after which the prices were back to the pre-crisis level [10,23]. We apply the entropy analysis to measure the diversity of patterns contained in these time series to explore the emergence of substantial changes, particularly those associated with significant events in market dynamics, in particular, before, during and after the economic crisis. In addition, we use cross-sample entropy to analyze the behavior of the sugar versus ethanol markets.

2.2. Sample entropy

Sample entropy (SampEn) was introduced by Richman and Moorman [19] as a modification of the approximate entropy (ApEn) [24], both methods were designed to analyze the dynamics of time series by evaluating its regularity and level of complexity. A greater regularity (lower complexity) produces lower values of SampEn, whereas for a series with higher complexity the value of SampEn statistic is higher. The applications of sample entropy include physiology [25,26], geophysics [27] climatology [28], hydrology [29] and engineering [30]. The key advantages of SampEn in comparison with ApEn are the independence of the amount of data and a relatively simple implementation, reasons for which we selected it for the proposed analysis in this work, since the available dataset is not very large (weekly data).

SampEn (m, r, N) is defined as the negative natural logarithm of the conditional probability that two sequences of length N, that are similar (within a tolerance level r) for m points, remain similar for m +1 points, where self-matches are not included in calculating the probability. An algorithm for calculating sample entropy can be described as follows [19]. Given a time series of size N, $X = x_1, x_2, \ldots, x_N$, first N-m+1 vectors $\mathbf{x}_m(i)$ of size m are constructed where $\mathbf{x}_m(i)$ = x_i, x_{i+1}, …, x_{i+m-1}, and $i = 1$, …, N-m+1. The distance $d_{i,j}$ between the vectors $\mathbf{x}_m(i)$ and $\mathbf{x}_m(j)$ is calculated as $d_{i,j}\left[\mathbf{x}_m(i), \mathbf{x}_m(j)\right] = \max\left\{\left|x_{i+k} - x_{j+k}\right| : k = 0, \ldots, m-1\right\}$, for each $i = 1, \ldots, N-m$ and $j = 2, \ldots, N-m+1$, where $i \neq j$ and $j > i$ to exclude self-matches. Subsequently, quantities $B_i^m(r) = \frac{B_i}{N-m-1}$ and $A_i^m(r) = \frac{A_i}{N-m-1}$ are calculated, where B_i is the number of vectors $\mathbf{x}_m(j)$ of size m that are similar to vectors $\mathbf{x}_m(i)$ within a tolerance r estimated from $d_{i,j}\left[\mathbf{x}_m(i), \mathbf{x}_m(j)\right] \leq r$, and A_i is the number of vectors $\mathbf{x}_{m+1}(j)$ that are similar to vectors $\mathbf{x}_{m+1}(i)$. From the individual $B_i^m(r)$ and $A_i^m(r)$ values, the corresponding mean values $B^m(r) = \frac{1}{N-m}\left(\sum_{i=1}^{N-m} B_i^m(r)\right)$ and $A^m(r) = \frac{1}{N-m}\left(\sum_{i=1}^{N-m} A_i^m(r)\right)$ are now calculated, and finally the statistic called sample entropy expressed in (1) is obtained as

$$SampEn\,(m, r, N) = -ln\left(\frac{A^m\,(r)}{B^m\,(r)}\right). \tag{1}$$

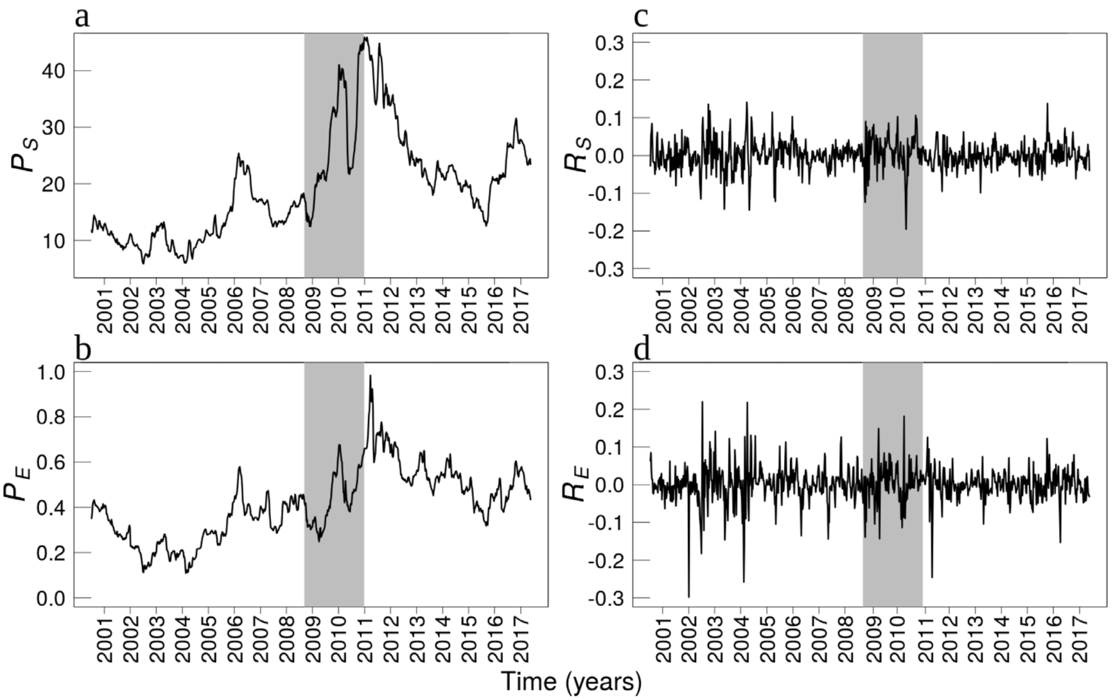


Fig. 1. Time series of weekly prices for (a) sugar, (b) ethanol, and returns for (c) sugar, (d) ethanol.

2.3. Cross sample entropy

To determine the similarity of sugar and ethanol series, we use the statistic known as cross-sample entropy (Cross-SampEn), which is based on SampEn [19]. While SampEn determines how many vectors in a time series of data occur within a statistically significant range that can be defined as similarity, analysis of similarity of parallel sequences in the two nonstationary time series is achieved using the Cross-SampEn cross-correlation method, which proceeds as follows [19]. Given two simultaneous time series $u = u_1, u_2, \ldots, u_N$ and $v = v_1, v_2, \ldots, v_N$ of size N, first it is necessary to configure the parameters m and r, the length of the vector to be compared and the tolerance to accept the matches, respectively. Second the vectors $\mathbf{x}_m(i) = [u_i, u_{i+1}, \ldots, u_{i+m-1}]$: $1 \leq i \leq N - m + 1$ and $\mathbf{y}_m(j) = [u_j, u_{j+1}, \ldots, u_{j+m-1}]$: $1 \leq j \leq N-m+1\}$, are constructed for u and v. For each $i \leq N-m+1$, we define $B_i^m(r)(v||u) = \frac{B_i}{N-m}$, where B_i is the number of vectors $\mathbf{y}_m(j)$ within tolerance r from $\mathbf{x}_m(i) : d_{i,j}[\mathbf{x}_m(i), \mathbf{y}_m(j)] \leq r$, with $d_{i,j}[x_m(i), y_m(j)] = \max\left\{\left|u_{i+k} - v_{j+k}\right| : k = 0, \ldots, m-1\right\}$ is the maximum difference in their respective scalar components. Then we define $B^m(r)(v||u) = \frac{\sum_{i=1}^{N-m} B_i^m(r)(v||u)}{N-m}$, where $B^m(r)$ is the probability that two vectors (from two simultaneous series u and v) will match for $m$points. We repeat this calculation for vectors of length $m+1$ and define $A_i^m(r)(v||u) = \frac{A_i}{N-m}$ and $A^m(r)(v||u) = \frac{\sum_{i=1}^{N-m} A_i^m(r)(v||u)}{N-m}$ where $A^m(r)$ is the probability that two vectors (from two different series) will match for $m+1$points. Finally, Cross-SampEn is estimated as

$$Cross - SampEn\,(m, r, N) = -ln\left(\frac{A^m(r)(v||u)}{B^m(r)(v||u)}\right). \tag{2}$$

Cross-sample entropy represents a conditional probability that sequences (from two different series) that are similar (within certain tolerance level) over m consecutive data points will remain similar after addition of one consecutive data point. Higher values of Cross-SampEn indicate less synchronization between analyzed temporal series [19]. Cross-SampEn was used in analyzing physiological [31,32], geophysical [33], and financial data [34,35].

In practice, SampEn and Cross-SampEn are usually applied on standardized time series such as: $u^* = \frac{u_i - \overline{u}}{\sigma_u}$, where $\overline{u}$ and σ_u represent the mean and standard deviation of the time series, with values of r varying between 0.1–0.25, and $m = 1, 2, 3$. For time series with length N between 100–5000 samples, $m = 2$ and $r = 0.2$ are the most common parameters [19,36].

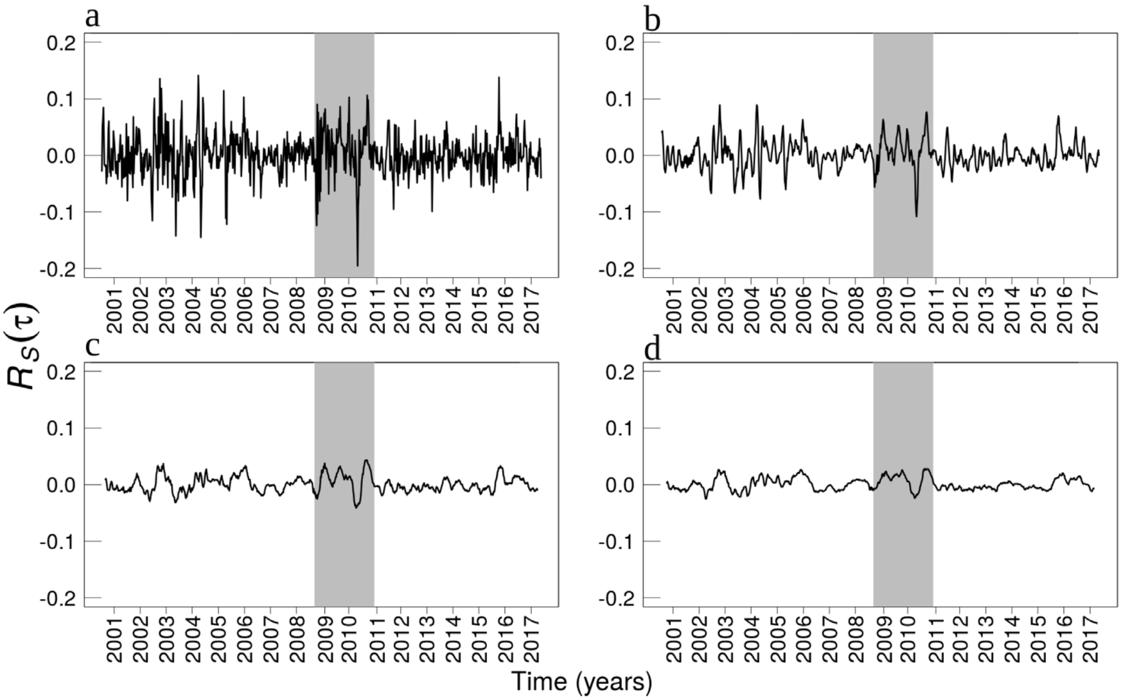


Fig. 2. Moving average filter applied on the series of sugar returns (a) $\tau = 1$, (b) $\tau = 5$, (c) $\tau = 15$ and (d) $\tau = 25$. The shaded area represents the period of crisis.

2.4. Multiscale sample entropy MSE

As entropy is scale dependent, a pattern may look less regular depending on the choice of timescale. In financial temporal series this scale dependency exhibits short-term and long-term trends; fluctuations are more complex for smaller timescales, while higher timescales are characterized by more regular fluctuations [37]. Therefore multiscale entropy is frequently used to represent variability over a broad range of scales in time series [37–41]. The multiscale entropy procedure consists in the application of the moving average filter, which removes high-frequency components, generating a new series with different timescales τ [37,39,41], for which entropy is calculated using (1) and (2). The following procedure is used to obtain the new filtered series. Given the time series $X = x_1, x_2, \ldots, x_N$, the moving-average filter is applied to each timescale τ, obtaining a series $Y_\tau = \frac{1}{\tau}\sum_{j=0}^{\tau-1} x_{i+j}$, where $i = 1, N\text{-}\ \tau+1$. Thus, Y_τ holds the values of X for each timescale greater than τ, or frequencies smaller than $f = \frac{1}{\tau}$, removing short-term fluctuations at higher values of τ, which effectively reduces the complexity of the time series. We applied the MSE method on return series of sugar and ethanol for $\tau = 1, 2, 3, \ldots, 26$. Fig. 2 shows how the moving average filter, using $\tau = 1, 5, 15, 25$, removes the short-term fluctuations in the sugar returns series.

2.5. Time-dependent entropy

While standard statistical analysis of time series assumes their stationarity, this condition is not always met, in which cases it is necessary to utilize procedures that adequate for nonstationary data series analysis. The time-dependent entropy is one such method, corresponding to quantification of irregularity in the series at different scales, as a function of time, based on the sliding window protocol. This method examines the entropy values from a perspective of temporal evolution, allowing the application of this technique in nonstationary conditions since the series are analyzed by segments [39,41]. The method to calculate the time-dependent entropy is as follows. Given a series of data $X = x_1, x_2, \ldots, x_N$, the sliding window protocol is defined as $X_t = x_{1+t\Delta}, \ldots, x_{w+t\Delta}$, $t = 0, 1, \ldots, \left[\frac{N-w}{\Delta}\right]$ where $w \leq N$ is the window size, $\Delta \leq w$ is the sliding step, and the operator [.] denotes taking integer part of the argument. The time series values in each window X_t are used to compute the $SampEn_{t,\tau}(m, r, w)$ and $Cross - SampEn_{t,\tau}(m, r, w)$ at a given time t and scale τ.

3. Results and discussion

As described in Section 2.1, P_S and P_E represent data series containing the weekly prices of crystal sugar and hydrated ethanol and R_S and R_E are the corresponding series of logarithmic returns. Considering 52 data per year, we set on the

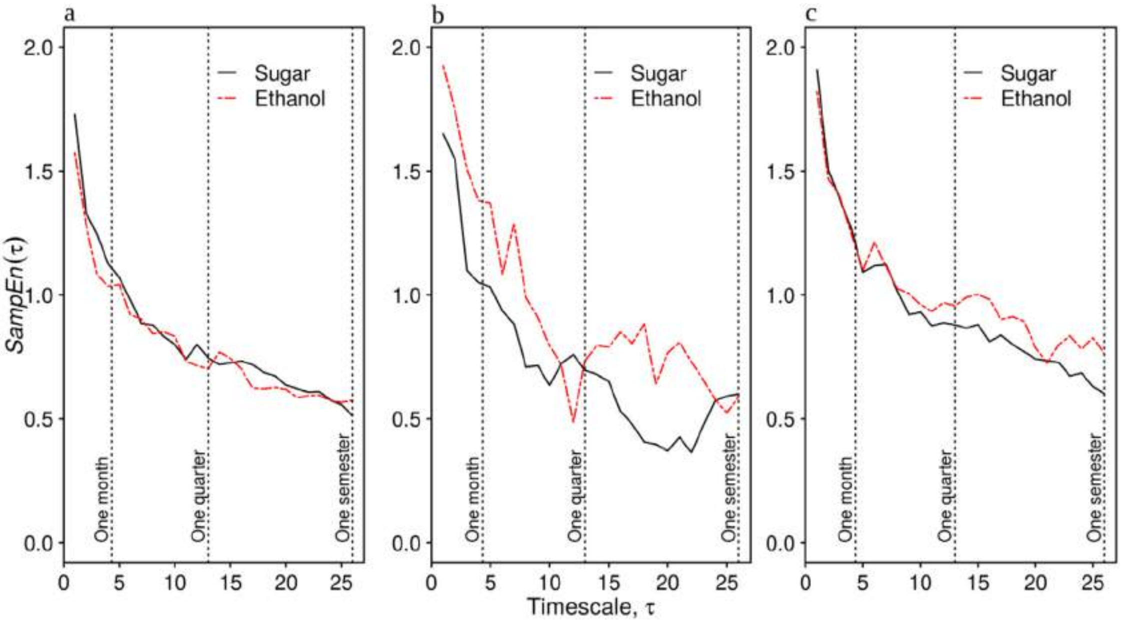


Fig. 3. Multiscale SampEn statistics for the series of returns of sugar and ethanol. (a) Before (b) during, and (c) after the crisis.

timescales $\tau = 1, 2, 3, \ldots, 26$, where $\tau = 4$, $\tau = 13$ and $\tau = 26$ represent the intervals of one month, one quarter and one semester, respectively. In addition, we configured the parameters $r = 0.2$ and $m = 2$ in the SampEn and Cross-SampEn methods, as explained at the end of Section 2.3. To analyze the entropy dependence over time, we used overlapping sliding windows with $\Delta = 1$ for window size $w = 200$ (about four years) and we attributed the resulting value to the midpoint of the chronological time.

3.1. Multiscale sample entropy

To analyze the behavior of sugar and ethanol time series, these were divided into three periods: before, during and after the global economic downturn, taking as the recession start the date the company Lehman Brothers bankruptcy was officially established, in September 2008 [42]. Thus, we adopted September 12, 2008 as the date of the beginning of the crisis, and as the date of the culmination of the downturn two years later, December 30, 2010, dividing the data into three new subseries for prices P_S and P_E with size $N = 426$, $N = 121$ and $N = 328$, respectively. Fig. 3 shows the multiscale sample entropy $SampEn(\tau)$ for the series of returns R_s and R_E for each of the analyzed periods. It is seen from Fig. 3 that: (i) The behavior of $SampEn(\tau)$ before the crisis is as may be generally expected, i.e., the entropy value decreases as the timescale increases, in the long-run the time series looses pattern diversity and fluctuations become more regular (Fig. 3(a)); (ii) During the crisis the $SampEn(\tau)$ values of ethanol show an anomalous behavior for timescales higher than one quarter since the entropy values increase indicating that in the long-run, the ethanol time series becomes less regular and more complex and uncertain (Fig. 3(b)). Similar results were obtained by Martina et al. [41] for crude oil prices during the period of Gulf war 1990. Finally, (iii) the $SampEn(\tau)$ values after the recession exhibit a higher entropy (indicating higher market efficiency) then before and during the crisis, and show a decreasing behavior, but non-monotonically, Fig. 3(c).

3.2. Time dependent multiscale entropy

We also employed multiscale SampEn and Cross-SampEn statistics in overlapping sliding windows to analyze temporal evolution of regularity and asynchrony of sugar and ethanol series. Fig. 4 shows the results obtained for window size $w = 200$ and $\tau = 1, 5, 10$. It is seen from Fig. 4(a) and (b) that the entropy values increase after 2008 and 2012 indicating the increase of market efficiency after global and European crisis, respectively. This effect is more pronounced for sugar series, having as a consequence the increase in cross-sample entropy values (increased asynchrony between two series) in the same periods (Fig. 4(c)).

The heat map layout in Fig. 5 displays the outcomes obtained for returns with window size $w = 200$ and $\tau = 1, 2, \ldots, 26$. We observed at higher timescales, in both series higher entropy values occurring after the crisis, indicating less regularity in return series and more efficient market. This effect is more evident in the period between 2012 and 2013, which is related to the economic crisis in Europe in 2012 [43]. Again this behavior is more pronounced for sugar

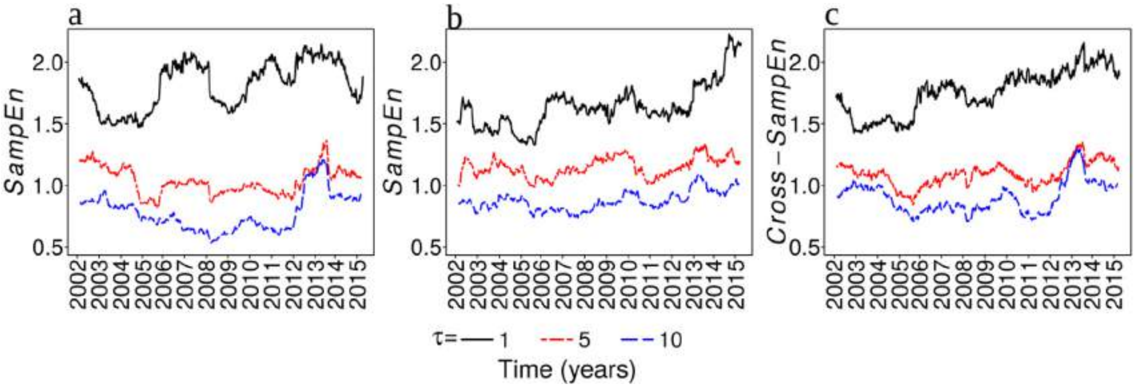


Fig. 4. Time-dependent entropy of (a) sugar, (b) ethanol, and (c) cross-sample entropy between sugar and ethanol, for timescale $\tau = 1$, 5,10 and window size $w = 200$.

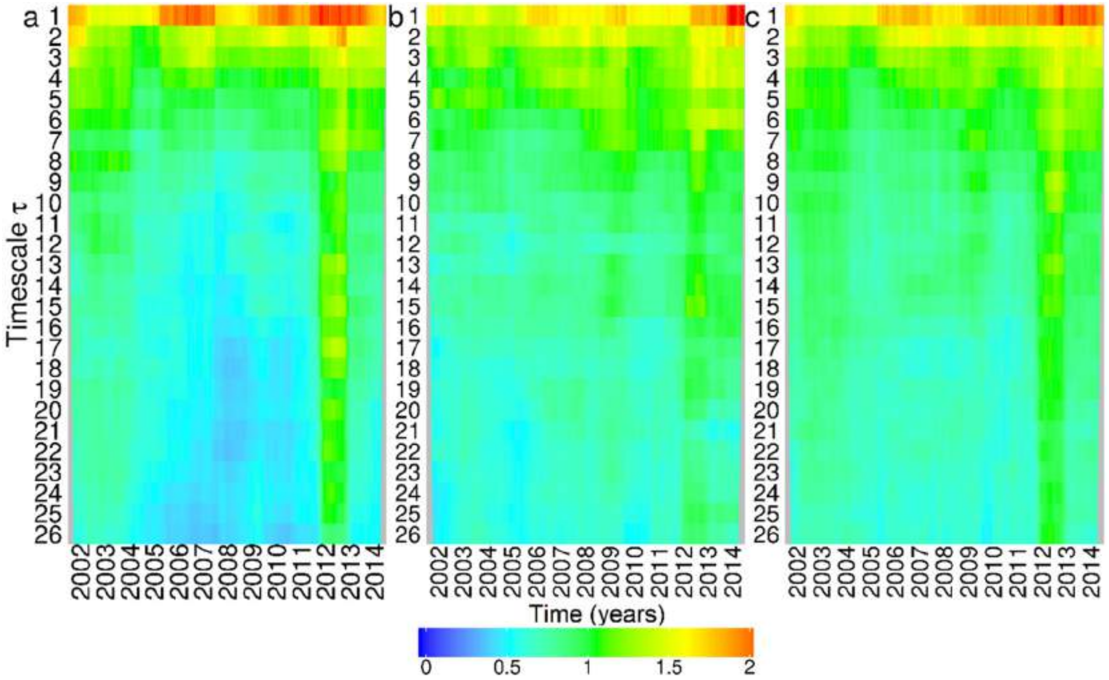


Fig. 5. Time-depended multiscale entropy for (a) sugar, (b) ethanol and (c) cross sample entropy between sugar and ethanol.

than for ethanol which also (for some timescales) shows the increase in entropy before and during the crises, indicating that overall, ethanol series are less regular but also less responsive to effects of crisis.

It is well-known that for financial time series multiscale patterns tend to decrease monotonously with the timescale, indicating that the long-term trend of the time series loses patterns diversity (i.e., becomes more regular) regarding the short-term tendency as seen from heat maps of time depended multiscale entropy (Fig. 5). Following Martina et al. [41], we display in Fig. 6 the multiscale patterns for the most affected years by economic depression (vertical lines in heat maps in Fig. 5). In these years, the application of the SampEn method suggests that sugar and ethanol return series present some deviations from the expected decreasing behavior, being more evident for the ethanol. That suggests that in the long-run, the price fluctuations of this commodity do not gain regularity, become more uncertain since the diversity of patterns (entropy) increases with the timescale. The entropy values for 2012 are higher than in other years, revealing the post-crisis effect as well as the influence of economic crisis in Europe [43].

To support the results obtained so far, we also calculated the mean values and the standard deviation of time-dependent multiscale sample entropy for each timescale, as shown on Fig. 7. Then, assuming independence between both commodities and with 5% significance level, we performed the statistical comparison tests as shown below. (i) To examine the normality conditions, we ran the *Shapiro* test, finding that the means did not fulfill the condition while the

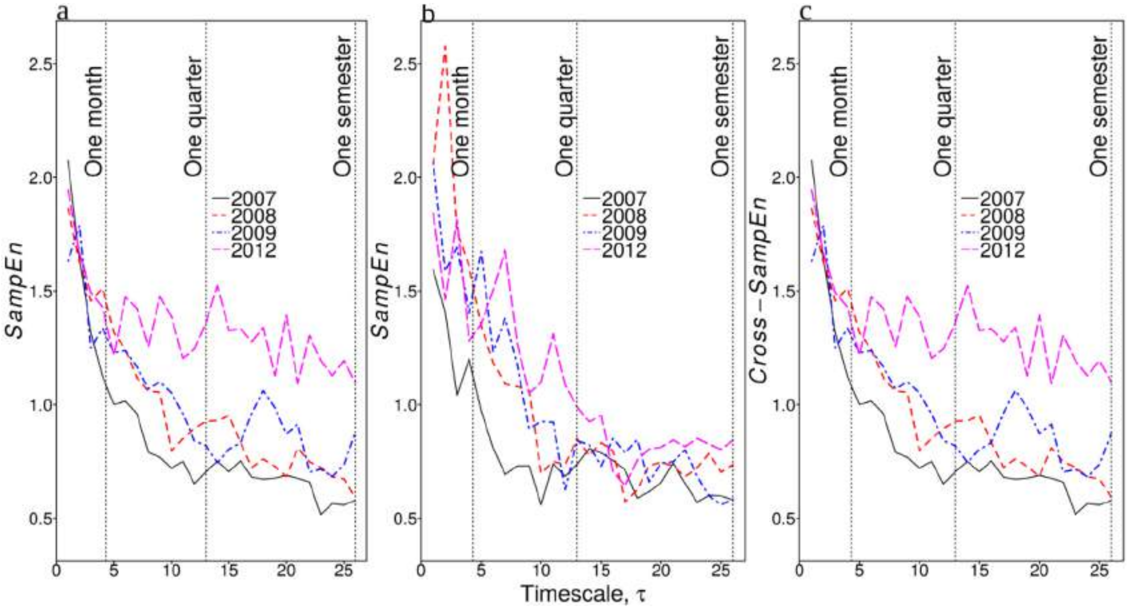


Fig. 6. Multiscale sample entropy of returns series: (a) sugar, (b) ethanol, and (c) cross-sample entropy between sugar and ethanol for one-year periods mostly affected by economic crisis.

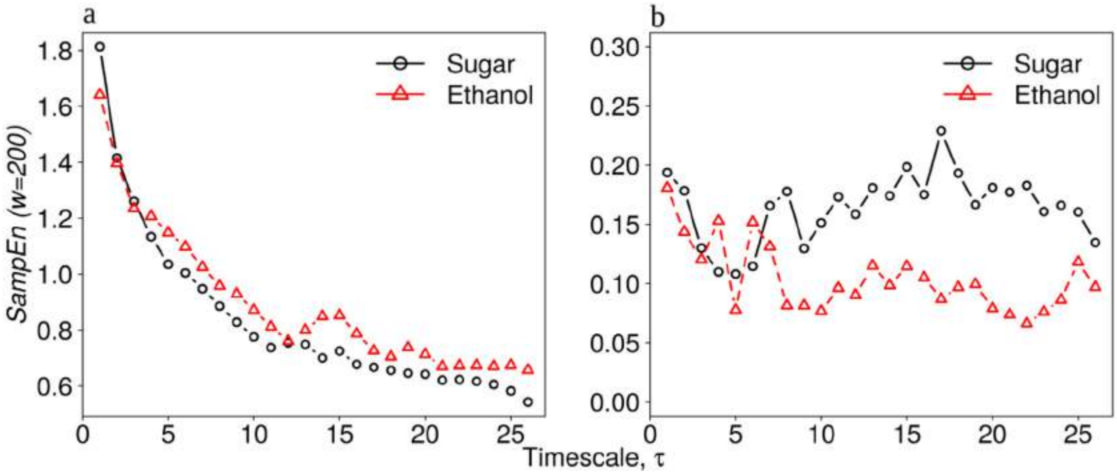


Fig. 7. Time dependent multiscale sample entropy: (a) mean values, (b) standard deviation values.

standard deviations satisfied the test. (ii) We conducted the *Mann–Whitney–Wilcoxon* test that revealed that there is no enough evidence to say the mean values are different. (iii) For the standard deviation, we executed the T-test and the results exhibited that their values are statistically different, being higher for sugar, indicating that for the sugar series in different timescales there are periods of high entropy and periods of low entropy. It is seen from Fig. 7(a) that at most timescales ethanol series exhibits higher entropy than sugar series which together with previous results (Figs. 3–6) indicates that the ethanol series is more complex when compared to the sugar series, meaning less regularity in price fluctuations of this commodity.

4. Conclusions

In this work we analyze complexity of temporal series of prices of ethanol and sugar which represent Brazilian energy and agriculture market, with focus on influence of global economic crisis in 2008. These commodities are produced from sugarcane and their production changes depending on their relative prices which can be influenced by various internal (i.e governmental policies and demand/supply ratio) and external (i.e. oil prices and exchange rate) factors. We use well established methods sample entropy and cross-sample entropy which were designed to evaluate regularity of temporal series and asynchrony between two temporal series. We applied these methods on sugar and ethanol return series for different temporal scales to investigate short term and long term behavior, and in overlapping sliding windows to analyze temporal evolution of regularity of price dynamics, and to detect the influence of economic crisis. The results show that for both ethanol and sugar return series the entropy values increase after 2008 and 2012 indicating the increase of market efficiency after global and European crisis, respectively. At higher timescales, in both series higher entropy values (less regularity and higher market efficiency) also occurred after the crisis, with more evidence in the period between 2012 and 2013, which is related to the economic crisis in Europe in 2012. This behavior is more pronounced for sugar than for ethanol. Analyzing entropy values for the years most affected by economic crisis we found that both sugar and ethanol return series present some deviations from the expected decreasing behavior for higher timescales, which is more evident for ethanol, indicating that in the long-run the price fluctuations of this commodity become more uncertain. By applying appropriate statistical tests on time dependent multiscale entropy we compared mean values and standard deviations of entropies of sugar and ethanol returns for different temporal scales using the values calculated in each sliding window. While there is no statistical difference between mean values, although they are slightly higher for ethanol, standard deviation values are statistically different, being higher for sugar. Overall, higher entropy values are found for ethanol series indicating that the ethanol series is more complex when compared to the sugar series, meaning less regularity and higher market efficiency in energy market.

Our work complements some recent results [37,41,44–48] that indicate that entropy is a promising measure to monitor the evolution of financial variables (in this case ethanol and sugar prices) for different timescales, and could be useful to identify different market phases, particularly those related to macroeconomic events such as financial crisis.

Acknowledgments

We acknowledge financial support of Brazilian agencies CAPES, Brazil and CNPq, Brazil.

References

[1] M.O. de Souza Dias, R. Maciel Filho, P.E. Mantelatto, O. Cavalett, C.E.V. Rossell, A. Bonomi, M.R.L.V. Leal, Sugarcane processing for ethanol and sugar in Brazil, Environ. Dev. 15 (2015) 35–51.
[2] J.R. Moreira, V. Romeiro, S. Fuss, F. Kraxner, S.A. Pacca, BECCS potential in Brazil: Achieving negative emissions in ethanol and electricity production based on sugar cane bagasse and other residues, Appl. Energy 179 (2016) 55–63.
[3] D. Khatiwada, S. Leduc, S. Silveira, I. McCallum, Optimizing ethanol and bioelectricity production in sugarcane biorefineries in Brazil, Renew. Energy 85 (2016) 371–386.
[4] L.A. Martinelli, S. Filoso, Expansion of sugarcane ethanol production in Brazil: environmental and social challenges, Ecol. Appl. 18 (4) (2008) 885–898.
[5] G. Sparovek, G. Berndes, A. Egeskog, F.L.M. de Freitas, S. Gustafsson, J. Hansson, Sugarcane ethanol production in Brazil: an expansion model sensitive to socioeconomic and environmental concerns, Biofuels, Bioprod. Biorefining: Innov. Sustain. Econom. 1 (4) (2007) 270–282.
[6] T. Serra, Volatility spillovers between food and energy markets: a semiparametric approach, Energy Econ. 33 (6) (2011) 1155–1164.
[7] A. Dutta, Cointegration and nonlinear causality among ethanol-related prices: evidence from Brazil, GCB Bioenergy 10 (5) (2018) 335–342.
[8] T. Serra, D. Zilberman, J. Gil, Price volatility in ethanol markets, Eur. Rev. Agric. Econom. 38 (2) (2011) 259–280.
[9] K. Balcombe, G. Rapsomanikis, Bayesian estimation and selection of nonlinear vector error correction models: the case of the sugar-ethanol-oil nexus in Brazil, Am. J. Agric. Econom. 90 (3) (2008) 658–668.
[10] L. Kristoufek, K. Janda, D. Zilberman, Comovements of ethanol-related prices: evidence from Brazil and the USA, GCB Bioenergy 8 (2) (2016) 346–356.
[11] D. Bentivoglio, A. Finco, M.R.P. Bacchi, Interdependencies between biofuel, fuel and food prices: The case of the Brazilian ethanol market, Energies 9 (6) (2016) 464.
[12] V. Plerou, P. Gopikrishnan, B. Rosenow, L.A.N. Amaral, T. Guhr, H.E. Stanley, Random matrix approach to cross-correlations in financial data, Phys. Rev. E 65 (2002) 066126.
[13] K. Yamasaki, L. Muchnik, S. Havlin, A. Bunde, H.E. Stanley, Scaling and memory in volatility return intervals in stock and currency markets, Proc. Natl. Acad. Sci. USA 102 (2005) 9424–9248.
[14] D. Stošić, D. Stošić, T. Stošić, H.E. Stanley, Multifractal Properties of Price Change and Volume Change of Stock Market Indices.
[15] L. Zunino, B.M. Tabak, F. Serinaldi, M. Zanin, D.G. Pérez, O.A. Rosso, Commodity predictability analysis with a permutation information theory approach, Physica A 390 (5) (2011) 876–890.
[16] B.M. Tabak, T.R. Serra, D.O. Cajueiro, Topological properties of commodities networks, Eur. Phys. J. B 74 (2) (2010) 243–249.
[17] L. Kristoufek, K. Janda, D. Zilberman, Correlations between biofuels and related commodities before and during the food crisis: A taxonomy perspective, Energy Econ. 34 (5) (2012) 1380–1391.
[18] H. Kantz, T. Schreiber, Nonlinear Time Series Analysis, Vol. 7, Cambridge university press, 2004.
[19] J.S. Richman, et al., Physiological time-series analysis using approximate entropy and sample entropy, Am. J. Physiol. Hear. Circ. Physiol. 278 (2000) 2039–2049.
[20] M. Costa, A.L. Goldberger, C.K. Peng, Multiscale entropy analysis of complex physiologic time series, Phys. Rev. Lett. 89 (6) (2002) 068102.
[21] M. Dooley, M. Hutchison, Transmission of the US subprime crisis to emerging markets: Evidence on the decoupling–recoupling hypothesis, J. Int. Money Financ. 28 (8) (2009).
[22] Center for Advanced Studies in Applied Economics / Luiz de Queiroz College of Agriculture / University of São Paulo (Centro de Estudos Avançados em Economia Aplicada/Escola Superior de Agricultura Luiz de Queiroz /Universidade de São Paulo) - CEPEA / Esalq / USP; https://www.cepea.esalq.usp.br/br (Accessed: 22-Jul-2018).
[23] T. Serra, Volatility spillovers between food and energy markets: a semiparametric approach, Energy Econ. 33 (6) (2011) 1155–1164.
[24] S.M. Pincus, Approximate entropy as a measure of system complexity, Proc. Natl. Acad. Sci. 88 (6) (1991) 2297–2301.
[25] D.E. Lake, J.S. Richman, M.P. Griffin, J.R. Moorman, Sample entropy analysis of neonatal heart rate variability, Am. J. Physiol.-Regul., Integr. Comp. Physiol. 283 (3) (2002) R789–R797.

[26] W.C. Weng, C.F. Chang, L.C. Wong, J.H. Lin, W.T. Lee, J.S. Shieh, Altered resting-state EEG complexity in children with tourette syndrome: A preliminary study, Neuropsychology 31 (4) (2017) 395.

[27] G. Balasis, I.A. Daglis, C. Papadimitriou, M. Kalimeri, A. Anastasiadis, K. Eftaxias, Investigating dynamical complexity in the magnetosphere using various entropy measures, J. Geophys. Res. Space Phys. 114 (A9) (2009).

[28] L. Shuangcheng, Z. Qiaofu, W. Shaohong, D. Erfu, Measurement of climate complexity using sample entropy, Int. J. Climatol.: J.R. Meteorol. Soc. 26 (15) (2006) 2131–2139.

[29] C.M. Chou, Complexity analysis of rainfall and runoff time series based on sample entropy in different temporal scales, Stoch. Environ. Res. Risk Assess. 28 (6) (2014) 1401–1408.

[30] Z. Zhao, S. Yang, Sample entropy-based roller bearing fault diagnosis method, J. Vib. Shock 31 (6) (2012) 136–140.

[31] T. Zhang, Z. Yang, J.H. Coote, Cross-sample entropy statistics as a measure of complexity and regularity of renal sympathetic nerve activity in the rat, Exp. Physiol. 92 (2007) 659–669.

[32] J.S. Chang, S.D. Lee, G. Ju, J.-W. Kim, K. Ha, I.-Y. Yoon, Enhanced cardiorespiratory coupling in patients with obstructive sleep apnea following continuous positive airway pressure treatment, Sleep Med. 14 (2013) 1132–1138.

[33] R. Hernández-Pérez, L. Guzmán-Vargas, A. Ramírez-Rojas, F. Angulo-Brown, Pattern synchrony in electrical signals related to earthquake activity, Physica A 389 (2010) 1239–1252.

[34] L.-Z. Liu, X.-Y. Qian, H.-Y. Lu, Cross-sample entropy of foreign exchange time series, Physica A 389 (2010) 4785–4792.

[35] W. Shi, P. Shang, Cross-sample entropy statistic as a measure of synchronism and cross-correlation of stock market, Nonlinear Dynam. 71 (2013) 539–554.

[36] J.M. Yentes, N. Hunt, K.K. Schmid, J.P. Kaipust, D. McGrath, N. Stergiou, The appropriate use of approximate entropy and sample entropy with short data sets, Ann. Biomed. Eng. 41 (2) (2013) 349–365.

[37] J. Alvarez-Ramirez, E. Rodriguez, J. Alvarez, A multiscale entropy approach for market efficiency, Int. Rev. Financ. Anal. 21 (2012) 64–69.

[38] M. Costa, C.K. Peng, A.L. Goldberger, J.M. Hausdorff, Multiscale entropy analysis of human gait dynamics, Physica A 330 (1–2) (2003) 53–60.

[39] D. Stosic, D. Stosic, T. Ludermir, T. Stosic, Correlations of multiscale entropy in the FX market, Physica A 457 (2016) 52–61.

[40] Y. Li, M. Xu, Y. Wei, W. Huang, A new rolling bearing fault diagnosis method based on multiscale permutation entropy and improved support vector machine based binary tree, Measurement 77 (2016) 80–94.

[41] E. Martina, E. Rodriguez, R. Escarela-Perez, J. Alvarez-Ramirez, Multiscale entropy analysis of crude oil price dynamics, Energy Econ. 33 (5) (2011) 936–947.

[42] A. Nobi, S.E. Maeng, G.G. Ha, J.W. Lee, Effects of global financial crisis on network structure in a local stock market, Phys. A Stat. Mech. Appl. 407 (2014) 135–143.

[43] M. Karanikolos, et al., Financial crisis, austerity, and health in Europe, Lancet 381 (9874) (2013) 1323–1331.

[44] R. Gençay, N. Gradojevic, The tale of two financial crises: An entropic perspective, Entropy 19 (6) (2017) 244.

[45] W.A. Risso, The informational efficiency and the financial crashes, Res. Int. Bus. Finance 22 (3) (2008) 396–408.

[46] D. Stosic, D. Stosic, T. Ludermir, W. de Oliveira, T. Stosic, Foreign exchange rate entropy evolution during financial crises, Physica A 449 (2016) 233–239.

[47] L. Zunino, A.F. Bariviera, M.B. Guercio, L.B. Martinez, O.A. Rosso, Monitoring the informational efficiency of european corporate bond markets with dynamical permutation min-entropy, Physica A 456 (2016) 1–9.

[48] L. Zunino, B.M. Tabak, F. Serinaldi, M. Zanin, D.G. Pérez, O.A. Rosso, Commodity predictability analysis with a permutation information theory approach, Physica A 390 (5) (2011) 876–890.

PUBLISHED PAPER: WINE QUALITY RAPID DETECTION USING A COMPACT ELECTRONIC NOSE SYSTEM: APPLICATION FOCUSED ON SPOILAGE THRESHOLDS BY ACETIC ACID

Contents lists available at ScienceDirect

LWT - Food Science and Technology

journal homepage: www.elsevier.com/locate/lwt

Wine quality rapid detection using a compact electronic nose system: Application focused on spoilage thresholds by acetic acid

Juan C. Rodriguez Gamboa[a,1,*], Eva Susana Albarracin E[a,1], Adenilton J. da Silva[b], Luciana L. de Andrade Lima[c], Tiago A. E. Ferreira[a]

[a] Departamento de Estatística e Informática, Universidade Federal Rural de Pernambuco - UFRPE, Recife, PE, Brazil
[b] Centro de Informática, Universidade Federal de Pernambuco - UFPE, Recife, PE, Brazil
[c] Departamento de Tecnologia Rural, Universidade Federal Rural de Pernambuco, Recife, PE, Brazil

ARTICLE INFO

Keywords:
Beverage quality control
Wine spoilage
Online early detection

ABSTRACT

It is crucial for the wine industry to have methods like electronic nose systems (E-Noses) for real-time monitoring thresholds of acetic acid in wines, preventing its spoilage or determining its quality. In this paper, we prove that the portable and compact self-developed E-Nose, based on thin film semiconductor (SnO_2) sensors and trained with an approach that uses deep Multilayer Perceptron (MLP) neural network, can perform early detection of wine spoilage thresholds in routine tasks of wine quality control. To obtain rapid and online detection, we propose a method of rising-window focused on raw data processing to find an early portion of the sensor signals with the best recognition performance. Our approach was compared with the conventional approach employed in E-Noses for gas recognition that involves feature extraction and selection techniques for preprocessing data, succeeded by a Support Vector Machine (SVM) classifier. The results evidence that is possible to classify three wine spoilage levels in 2.7 s after the gas injection point, implying in a methodology 63 times faster than the results obtained with the conventional approach in our experimental setup.

1. Introduction

Wine flavor depends on 20 or more compounds, besides water and ethanol, that with subtle alterations in concentration determine its quality (Jackson, 2008). The most important technique used to determine wine quality is directly related to the organoleptic characteristics evaluation by trained experts (Aleixandre, Cabellos, Arroyo, & Horrillo, 2018; Cretin, Dubourdieu, & Marchal, 2018; Sáenz-Navajas et al., 2015). Since the analytical panels are expensive, time-consuming, and they are not always available, the wine is also characterized using gas and liquid chromatography or spectrophotometry, that require on reagents and experienced personnel (Martins et al., 2018; Perestrelo, Rodriguez, & Câmara, 2017; Stupak, Kocourek, Kolouchova, & Hajslova, 2017; Vazallo-Valleumbrocio, Medel-Marabolí, Peña-Neira, López-Solís, & Obreque-Slier, 2017). Besides, E-Noses are used as an alternative to traditional methods for wines discrimination regarding the organoleptic characteristics. Their purpose is to analyze aroma profiles by registering signals produced by the mixture of gases (as the human nose does) and then comparing the pattern of responses generated by different samples (Lozano, Santos, & Horrillo, 2016; Peris & Escuder-Gilabert, 2016; Rodríguez-Méndez et al., 2016; Zhao et al., 2017). However, most E-Noses are designed for general purpose, and sometimes they are not portable to use on-site.

Volatile acidity (VA) measurements, generally interpreted as acetic acid content ($g{\cdot}l^{-1}$), are used routinely as an indicator of wine spoilage (Zoecklein, Fugelsang, Gump, & Nury, 1995). Thereby, it is crucial for the wine industry and consumers to have methods for real-time monitoring of VA thresholds. There are previous works which the wine spoilage was characterized using E-Noses developed with special sensors or combined with other technologies and methods. Some common characteristics of those systems are the instrumentation complexity, most of them involve the use of additional equipment that requires experienced personnel, and they do not realize online detection. For instance, a metalloporphyrin based optoelectronic nose was developed in Amamcharla & Panigrahi (2010) for the simultaneous prediction of Volatile Organic Compounds (VOCs) concentrations in binary mixtures (acetic acid and ethanol) using partial least square regression (PLSR) and multilayer perceptron neural network (MLP-NN). Besides, in Gil-

[*] Corresponding author. Departamento de Estatística e Informática, Universidade Federal Rural de Pernambuco - UFRPE, Rua Dom Manoel de Medeiros s/n, Dois Irmãos, 52171-900, Recife, PE, Brazil.

E-mail address: juan.gamboa@ufrpe.br (J.C. Rodriguez Gamboa).

[1] Joint first authors.

https://doi.org/10.1016/j.lwt.2019.03.074
Received 27 November 2018; Received in revised form 6 March 2019; Accepted 25 March 2019
Available online 01 April 2019

Sánchez et al. (2011), it is reported the wine spoilage analysis when in contact with air using a combined system of a potentiometric electronic tongue and a humid E-Nose.

The acetic acid detection was studied by Macías et al. (2012) using a commercial E-Nose for general purpose, in combination with a neural network classifier (MLP). They detected only the excessive concentrations of acetic acid, equal to or greater than 2 g·l^{-1} in synthetic wine samples (aqueous ethanol solution at 10% v/v). However, levels higher than 1.2 g l^{-1} of VA cause that the wine takes on vinegar aromas (unpleasant), reducing its quality; hence the governments forbid their commercialization (Normative instruction N° 14, 2018; Zoecklein et al., 1995). Thus, our work was aimed to detect lower levels and with a quick identification in real wine samples with several spoilage thresholds using the self-developed E-Nose, without using any reagent to reduce the environmental impact, as well with a smooth and safe operation interface (no occupational risk for the operator and with minimal training).

This study presents the self-developed E-Nose based on commercially available gas sensors for early detection of spoilage thresholds by VA in routine tasks of wine quality control. We recorded electrical signals corresponding to odorant profiles of wines samples with different spoilage levels.[2] Afterward, we compared the conventional data processing approach used in E-Noses against our online data processing approach to accelerate the responses. In the conventional approach was applied the preprocessing and feature extraction before an SVM classifier to obtain the main odorant parameters (which requires that the measurement process had finished before the data processing stage). By contrast, we focused on an online solution, that let to achieve faster results, using an early portion of the signals while the measurement process is still running. Our approach is based on the training of a deep MLP classifier using the raw data.

2. Materials and methods

2.1. Electronic Nose

We used an E-Nose, that we named O-NOSE, comprising principally of an array of six metal-oxide gas sensors (Table 1), used to detect the volatile compounds. Fig. 1 shows O-NOSE on the left side, and the sensors board with two layers for a compact design on the right side.

2.1.1. Experimental setup

In Fig. 2a, we depict the O-NOSE measurement process divided into three stages. (i) Concentration stage: we used 1 ml wine samples to accumulate the volatiles for 30 s inside the concentration chamber. (ii) Data acquisition: 10 s after the initialization of this stage, the VOCs push toward the sensors chamber for 80 s generating change in the sensor resistance (gas absorption). Subsequently, the gas injection stops, and it begins the desorption for 90 s. Therefore, the acquired data corresponds to 180 s with 18.5 Hz sample rate. (iii) Purge: the goal is to clean and remove volatile residues for 600 s. Fig. 2b shows the standard block diagram for the whole experiments, the electrical signals acquired are processed using the pattern recognition techniques after finished the data acquisition stage in the conventional approach or online applying our approach.

2.2. Data

2.2.1. Wine samples

We used 22 bottles of commercial wines, and to obtain spoiled samples, 13 of the 22 bottles were randomly selected, opened and left in an uncontrolled environment six months before starting the measurements. These bottles were labeled as low-quality (LQ) wines. Besides, another four bottles were opened two weeks before beginning the data collection. These four bottles were labeled as average-quality (AQ) wines, and the remaining five bottles were labeled as high-quality (HQ) wines.

Table 1
Gas sensors array setup. The sensors manufactured by Hanwei Sensors[a] are commercially available. They have been chosen because of their high sensitivity to organic, natural, ethanol, methanol, and combustible gases, as well as its simplicity of use and low financial cost.

Number	Sensor	Description	Load resistance
1, 4	MQ-3	High sensitivity to alcohol and small sensitivity to Benzine	22 kΩ
2, 5	MQ-4	High sensitivity to CH4 and natural gas	18 kΩ
3, 6	MQ-6	High sensitivity to LPG, iso-butane, propane	22 kΩ

[a] www.hwsensor.com.

The 22 wine bottles were characterized as follows: (i) the VA quantification was performed in triplicate according to official methods for wine analysis of the International Organization of Vine and Wine (OIV. International Organization of vine and wine, 2014). (ii) Acetic acid was identified by High Performance Liquid Chromatography (HPLC) with UV/Vis absorption detector, following the procedure detailed in (De Andrade Lima et al., 2010), and the ranges obtained are shown in Table 2. It is known that at normal levels in wines (< 0.3 g l^{-1}) the VA can be a desirable flavor, adding to the complexity of taste and odor, as well, a content of less than 0.70 g l^{-1} seldom imparts spoilage character. However, a progressive increment in VA gives to the wines a sour taste and taints its fragrance (Jackson, 2008; Zoecklein et al., 1995). Brazilian Ministry of Agriculture, Livestock and Supply (Instrução Normativa N° 14, 2018) establishes that the maximum level of VA in wine is 1.2 g l^{-1}.

The database collected using O-NOSE has 235 wines measurements as follow: 51, 43, and 141 measurements of HQ, AQ, and LQ respectively. Besides, we collected 65 ethanol measurements in concentrations (v/v) of 2, 5, 10, 20, 30, and 40 ml of ethanol diluted in distilled water to make solutions of 200 ml.

2.3. Feature extraction and selection

The most common groups of characteristics extracted from the gas sensors signals are the steady and transient state features (J. Yan et al., 2015). We used 23 features to capture the dynamic and static behavior of each gas sensor. So, we obtained a 138 columns characteristics matrix, where each row represents the fingerprint of one measurement. One example of the raw data (Fig. 3a) evidences the sensor sensitivity regarding VOCs analyzed. In Fig. 3b–c, we show the steady and transient features for the response of one sensor during the three intervals of the acquisition procedure explained at the end of Section 2.1.

Afterward, we applied the SVM Recursive Feature Elimination Cross Validation (RFECV) method to reduce the dimensionality, looking to generate parsimonious and robustness models (Lin et al., 2012; K. Yan & Zhang, 2015). Thus, it was chosen the followings steady-state characteristics: $\Delta G = max_k g[k] - min_k g[k]$, defined as the maximal conductance change concerning the baseline, and its normalized version ($\Delta G = (max_k g[k] - min_k g[k]) / min_k g[k]$), as well, the area under the curve in the absorption and desorption portions of the gas, blue and gray areas in Fig. 3b, respectively. Additionally, we had an aggregate of features reflecting the dynamics of the rising/falling transient portion of the sensor response using an exponential moving average filter (ema$_\alpha$) that converts the transient portion into a real scalar by estimating the maximum/minimum value $y[k] = (1 - \alpha) y[k-1] + \alpha(x[k] - x[k-1])$, where $[k = 1,2,...,T]$, $y[0]$ its initial condition, set to zero ($y[0] = 0$, and the scalar $\alpha (\alpha \in \{0,1\})$ being a smoothing parameter of the operator such as was defined in

[2] The generated dataset is publicly available to the research community at https://data.mendeley.com/datasets/vpc887d53s/2.

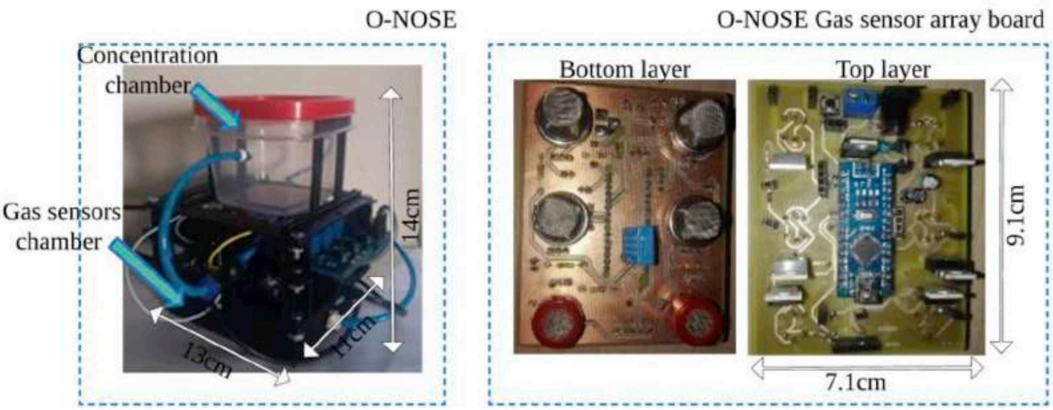


Fig. 1. O-NOSE system. On the left side: system appearance and dimensions. Into 100 ml concentration chamber is placed the wine sample. The sensors array is into the 200 ml chamber. On the right side: the main board with the gas sensors and the microcontroller. The gas is sensed by its effect on the sensitive layer of tin dioxide (SnO_2), resulting from changes in conductivity brought about by chemical reactions on the surface of the tin dioxide particles.

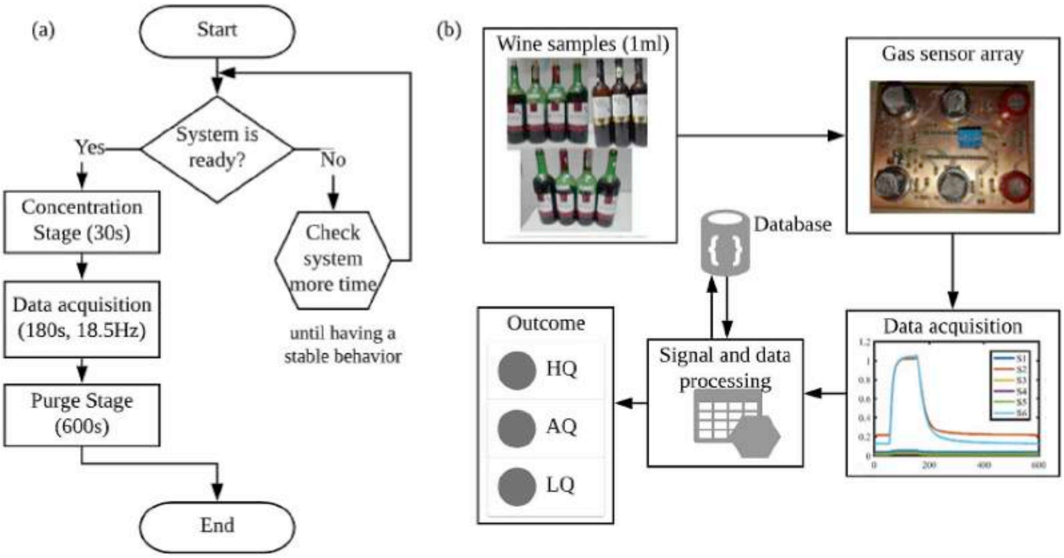


Fig. 2. (a) Flowchart of the measurement setup. (b) Block diagram for wine spoilage detection using 1 ml samples, the outcome is according to the wine quality.

Table 2
Ranges detected of volatile acidity and acetic acid according to the wine spoilage thresholds. The ranges presented correspond to the minimum and maximum values of the analysis.

Wine quality level	Volatile acidity in $g{\cdot}l^{-1}$	Acetic acid in $g{\cdot}l^{-1}$
HQ	[0.15, 0.3]	[ND, 0.23]
AQ	[0.31, 0.41]	[0.24, 0.34]
LQ	[0.8, 3]	[0.74, 2.75]

ND: not detected.

Muezzinoglu et al. (2009); Vergara et al. (2012). We tested three different values for $\alpha = 0.1$, $\alpha = 0.01$, and $\alpha = 0.001$ as shown in Fig. 3c; and by RFECV feature selection, it was chosen the **max** ema_{α} with $\alpha = 0.01$ as an informative transient feature.

2.4. Classification methods

We use two approaches for the classification tasks in this application. The first one consists in applying feature extraction and selection before the classifier. And the second one consists in processing an early portion of the raw data.

2.4.1. Conventional approach to classification using SVM

In this approach, it is necessary to have the whole measurement to obtain the main odorant parameters. We tested various kernels on an SVM classifier and selected a gaussian kernel; then it was trained the model. The block diagram of this approach (depicted in Fig. 4) exhibits the steps performed that includes a feature extraction block generating the $C_{i,j}$ vector, where $i = 1,2...,23$ is the number of characteristics and $j = 1,2...,6$ is the number of sensors. Afterward, the characteristics vector feed the feature selection block, and finally, the chosen variables are carried to the inputs of the SVM classifier.

2.4.2. Rapid and online detection approach using deep MLP

This approach is based on a neural network classifier that is feed with the raw data to perform the discrimination tasks (Peng, Zhao, Pan, & Ye, 2018). Inspired by the mentioned approach and looking to accelerate the response, we propose a rapid detection method in wine quality control, focused on an online solution that lets to achieve faster results using only an early portion of the signals, similar to the presented in (Längkvist, Coradeschi, Loutfi, & Balaguru Rayappan, 2013) for a meat spoilage application, but using a supervised method: deep MLP neural network. The goal with this approach is to offer the possibility to make estimations a few seconds after beginning the measurement process while it is still running. Note that we did not consider the baseline of the sensor since generally in this slice there is no change.

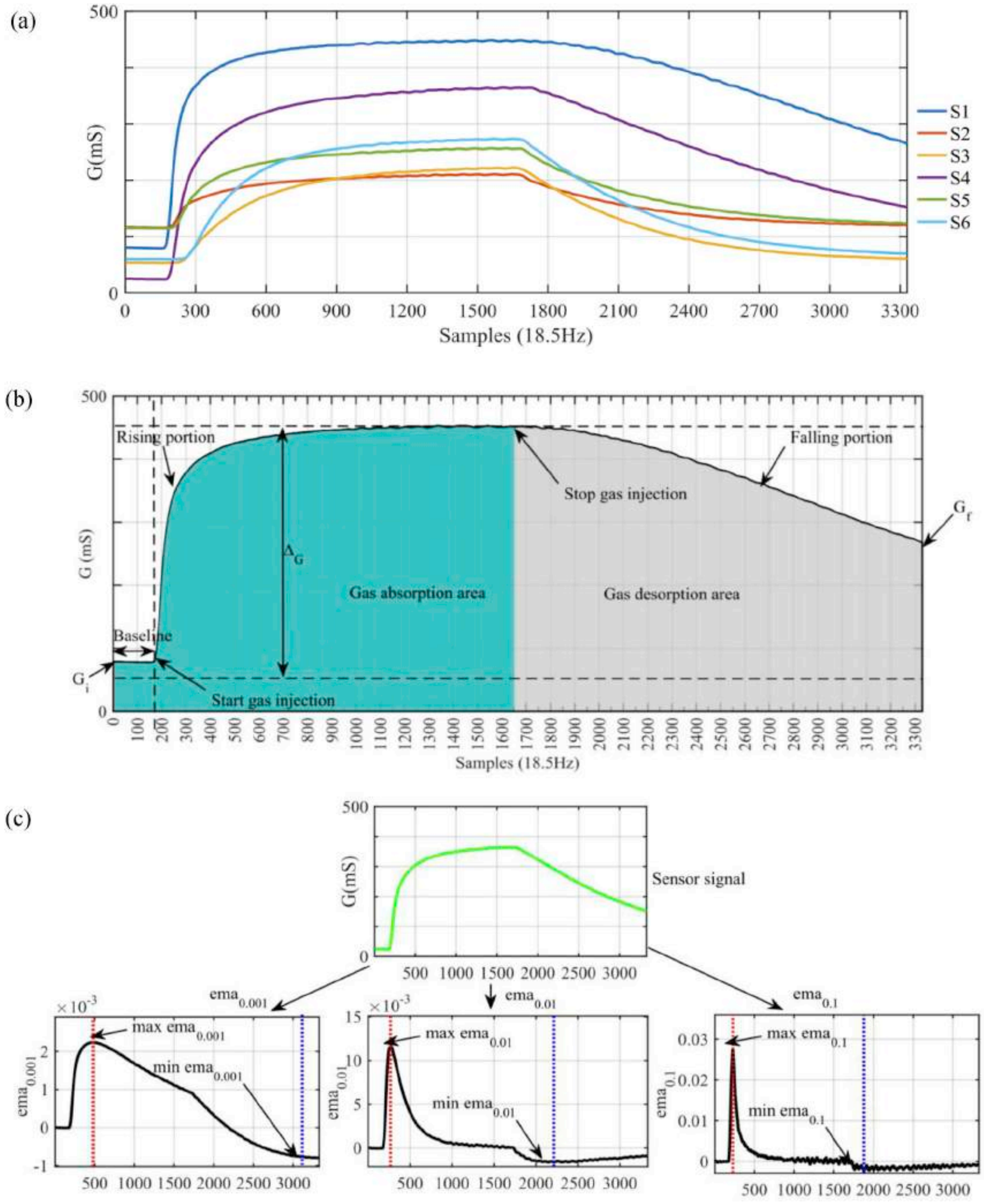


Fig. 3. (a) Wine measurement acquired with O-NOSE; S1, S2, …, S6: gas sensor outputs in conductance units G. (b) Output of a gas sensor; G_i: initial conductance value, G_f: final conductance value, ΔGmaximal conductance change concerning the baseline. (c) Dynamics of the rising/falling transient portion using an exponential moving average filter (ema_α) for α = 0.1, α = 0.01, and α = 0.001.

Consequently, the data processing starts instantly before the gas injection. A rising window method was applied to find the minor portion of information with the best performance of the classifier. This reduces the effort to obtain discrimination models since as complicated preprocessing techniques no need to be applied and it is feasible by the computational acceleration in the last years. Fig. 5 depicts the approach employed.

The method to find the minor portion of information is as follows. Given the data series $X_j = x_1, x_2, ..., x_N$, that represents the gas sensor response $j = 1,2...,6$, their corresponding rising windows are defined as: $X_{j,t} = x_{j,1}, ..., x_{j,t\Delta}$, where $t = 1, ..., \left[\frac{N}{\Delta}\right]$, the step is $\Delta \le N \wedge \ \Delta \, \varepsilon \, \mathbb{N}$, the window size is $t\Delta$, and the operator [.] denotes taking the integer part of the argument. The time series in each window $X_{j,t}$ are used to train the deep MLP classifier. Fig. 5 exhibits the application of the rising windows protocol in our dataset with $\Delta = 50$, hence each $X_{j,1}$ window has 50 points, each $X_{j,2}$ window has 100 points, and so on. The example architecture of the deep MLP, shown in the same figure, corresponds to the neural network used to process the data for the $X_{j,1}$ window. In this case, the input layer size corresponds to the first window ($t = 1$), six sensors, step $\Delta = 50$; then, it has $6\,(1\text{x}50) = 300$ points. The only data preprocessing applied before the deep MLP neural network was a simple data scaling in each window.

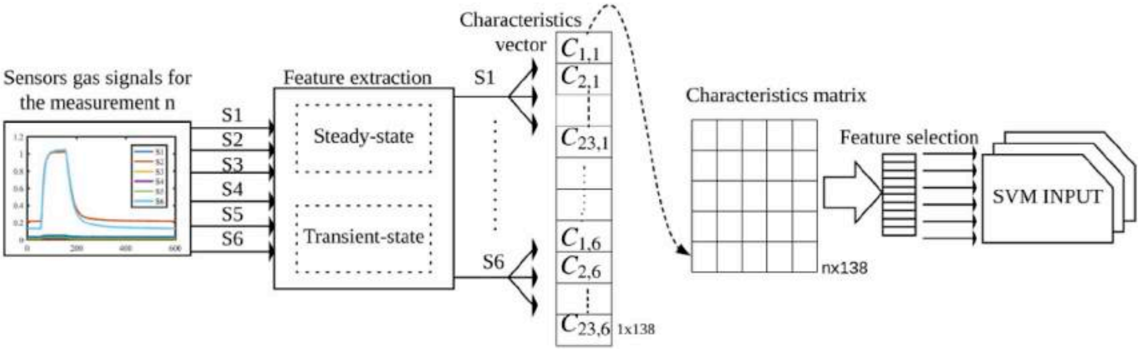


Fig. 4. Block diagram of the conventional approach to classification using SVM. This diagram comprises a Feature Extraction block (FE), a Feature Selection block (FS), and subsequently, the characteristics matrix feeds an SVM classifier.

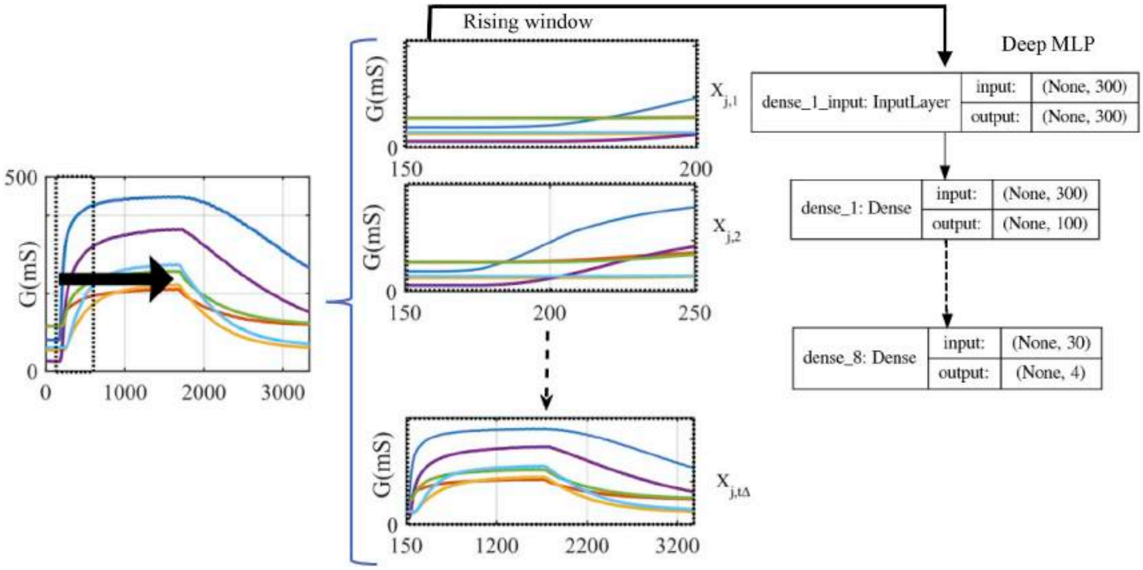


Fig. 5. Rapid and online detection approach. Rising window protocol applied to the raw data searching for the minor portion of data to train the deep MLP classifier with the best performance. On the right side is depicted the neural network architecture for the $X_{j,1}$ window with an input size of 300 points and four outputs (three wine spoilage levels and ethanol). The meaning of "None" is unspecified input because we reshaped the data in a flatted array.

3. Results

3.1. Data exploratory analysis

We performed the database exploratory analysis using the Principal Components Analysis (PCA). The scores for the first components (2D and 3D plots) for the wines are shown in Fig. 6. We also graph the PCA scores of ethanol jointly wines, as shown in Fig. 7.

Based on this exploratory analysis, we performed two experiments with the aim of comparing the performance when the classes are only wines with three spoilage thresholds, and when the ethanol is present as an additional class, which is evidenced as a more complex problem

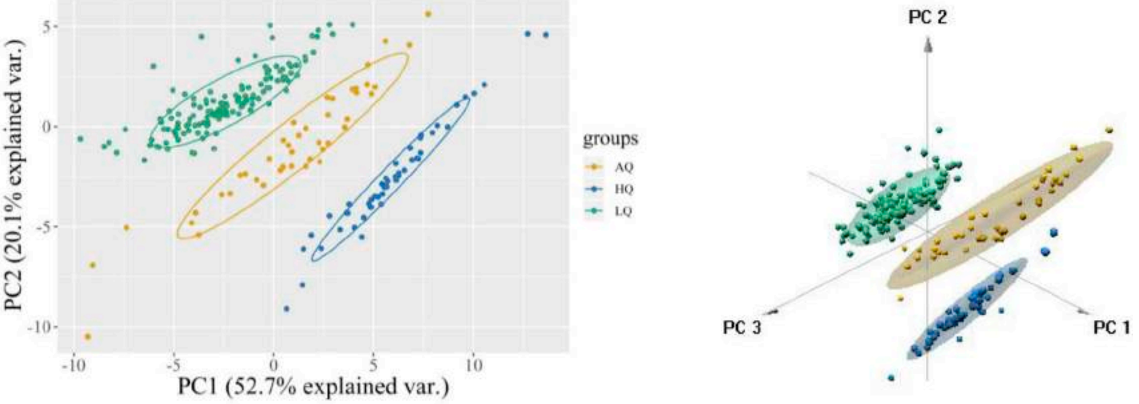


Fig. 6. PCA for the three wine groups HQ, AQ, and LQ. On the left side in 2D and the right side in 3D. It is revealed that O-NOSE detects differences between the three groups according to its quality and spoilage threshold. In this case, the three principal components capture a cumulative variance of 81%.

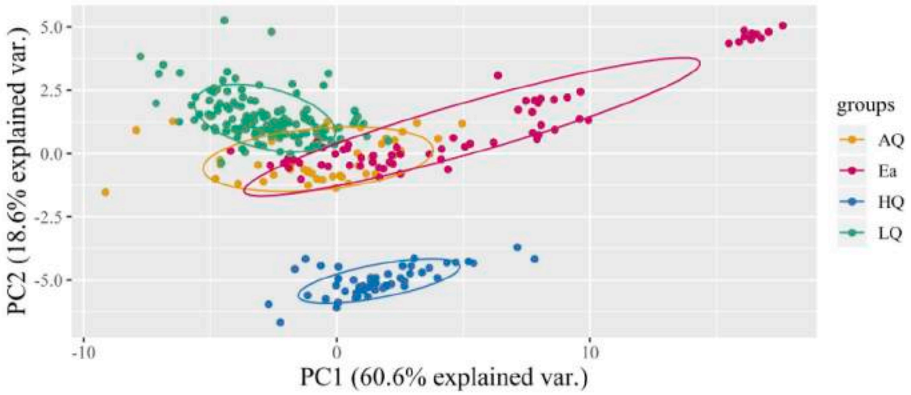


Fig. 7. PCA for the three wine groups (HQ, AQ, LQ) and ethanol (Ea). The close relationship between ethanol and wine is evidenced more strongly for the wines labeled as AQ. The groups labeled as HQ and LQ have greater separation regarding the ethanol. In the case of HQ, the organoleptic characteristics are rich in other elements that characterize the excellent taste. For the LQ, the taste is commonly described as vinegar or metallic taste and low level of ethanol.

because the ethanol is an essential wine component. These two experiments were performed so much for the conventional approach using SVM, as for the rapid and online detection approach using deep MLP neural network, and the results were compared at the end of Section 3.

3.2. Conventional approach to classification using SVM

We used an SVM classifier applying the technique known as Leave One Out (LOO), selecting the measurements of one bottle for the validation group and the remaining for the training group. Since as the dataset contains twenty two bottles, we performed this procedure that quantity of times, and we applied five folds cross-validation technique to prevent the overfitting in the training set. We implemented the scripts for this approach using Matlab R2016a and the Statistics and Machine Learning Toolbox - version 10.2; and, to ensure the integrity of the results, we repeated the procedure 100 times with data shuffling before each training. Then, we averaged the accuracy of each experiment.

In Table 3 are shown the parameters set on the SVM classifier for the two experiments performed: experiment 1 to discriminate among the three wine thresholds (LQ, AQ, and HQ); and experiment 2 to classify among the three wine thresholds and ethanol (LQ, AQ, HQ, and Ea). The recognition accuracy for training and validation, in the first experiment, was 99.78% and 97.34%, and, for the second experiment, 98.31% and 96.23%, respectively.

3.3. Rapid and online detection approach using deep MLP

We did several simulations to find an early portion of the raw data with the best recognition performance in the two experiments. To achieve this, we applied the rising window protocol searching for the minor portion of data to train the deep MLP classifier and averaging the accuracy of each experiment. In this way, we applied the LOO-protocol like the before experiments (Section 3.2), but now training the deep MLP models of eight layers with full neurons connections as detailed in Table 4 (architecture examples of experiment 2).

Table 3
Parameters of the SVM classifiers used for each experiment.ss.

Parameter	Experiment 1	Experiment 2
Kernel function	Gaussian	Gaussian
Kernel parameter scale (gamma)	8.3	19
Box constraint level (C penalty parameter)	10	10
Multiclass method	One-vs-One	One-vs-One
Standardize data	True	True
Feature selection: variables used in the model	69	56
PCA	disabled	disabled

Table 4
Network architecture of three models for the classification using deep MLP, where $\boldsymbol{X_{j,t}}$ is the time series in each window $\boldsymbol{t}$. The trainable parameters are computed as the multiplication between the inputs and the number of neurons in each layer plus the bias number (see the examples for the layers one and eight in the $\boldsymbol{X_{j,1}}$ model).

Layer	Neurons	Trainable parameters		
		$X_{j,1}$ model	$X_{j,12}$ model	$X_{j,63}$ model
1	100	$(300 \times 100)+100 = 30100$	360100	1.8901E+6
2	30	3030	3030	3030
3	30	930	930	930
4	30	930	930	930
5	30	930	930	930
6	30	930	930	930
7	30	930	930	930
8	4	$(30 \times 4)+4 = 124$	124	124

The original raw data have 3330 points, but as was explained in Section 2, the baseline is not considered. Thus, we defined the interval to analyze from the point 150 to the point 3300 (to ensure an integer $\left\lceil \frac{N}{\Delta} \right\rceil$). Since as the step was $\Delta = 50$, we trained 63 models that correspond to each $X_{j,t}$ window using python 3.5.3, repeating the procedure 100 times with data shuffling. In the first experiment with the rapid and online detection approach, the accuracy for the windows with the best performance in the training data was 100%, that occurred 97% of the times in windows with a size less or equal than $X_{j,24}$. This corresponds to the first 64.86 s of the raw data interval. In validation data, the accuracy for the windows with the best performance was 97.68%, that occurred 88% of the times in the first window ($X_{j,1}$). This represents only an early portion of the raw data, that is equivalent to the first 2.7s, indicating a significant reduction in the time for the recognition when compared to the conventional approach using the feature extraction/selection method.

The results for the second experiment with this approach indicated that the best performance occurred in windows with a size less or equal than $X_{j,13}$, corresponding to the first 35.13 s of the raw data interval. The accuracy was 99.99%, and 96.34%; occurring 54% and 61% of the times in training and validation, respectively. Note that, the separability of the data in this experiment is more complex than the experiment 1 that includes only the three wine spoilage levels, causing that the early portion time necessary for the recognition task being greater. However, it is still less than using the conventional approach

Table 5
Comparison between the conventional and the rapid detection approach.

Summary of test results	Conventional approach		Rapid and online detection approach	
	Experiment1	Experiment2	Experiment1	Experiment2
Average accuracy (%)	97.34 ± 0	96.23 ± 0	$97.68 \pm 4.6 \times 10^{-3}$	$96.34 \pm 4.6 \times 10^{-3}$
Time for recognition (s)	171.89	171.89	**2.7**	**35.13**
Data preprocessing	FE + FS	FE + FS	Scaling	Scaling
Online	NA	NA	Yes	Yes
Input size	69	56	300	3900
Time for training (s)	16	27	99	130
Time for validation (s)	«1	«1	«1	«1

Average accuracy is presented as the mean ± standard deviation obtained from 100 repetitions. The Mann-Whitney-Wilcoxon test was conducted with ($P > 0.05$). FE: Feature extraction; FS: Feature selection; NA: Not available.

Table 6
Comparison of the rapid detection approach with other similar works.

	Peng et al. (2018)	Längkvist, Coradeschi, Lotfi, & Balaguru Rayappan (2013)		Proposed work	
		Result1	Result2	Result1	Result2
Model	DCNN	DBN		Deep MLP	
Method	Supervised	Unsupervised		Supervised	
Application or gases	CO, CH_4, H, and C_2H_4	Ethanol and TMA		Wine samples and ethanol	
Gas sensor type	MOS	Nanostructured ZnO		MOS	
Online	Not	Yes		Yes	
Average accuracy (%)	95.2	60 ± 4.5	83.7 ± 4.1	97.68	96.34
Time for recognition (s)	100	5	25	2.7	35.13
Time for training (s)	154	NA	NA	99	130

CO: carbon monoxide; CH_4: methane; H: hydrogen; C_2H_4: ethylene; TMA: trimethylamine; DBN: Deep Belief Network; DCNN: Deep Convolutional Neural Networks; MOS: Metal oxide semiconductor; NA: Not available.

which consumes the whole measurement time, suggesting outperformance for the online detection approach using deep MLP.

4. Discussion

The comparison based on the test results between the two discussed approaches is presented in Table 5. We highlight the gain in timing for recognition wine quality with our approach, and the possibility of using this approach for online detection without preprocessing techniques.

The rapid and online detection approach has the highest computational time in the training. However, the training is performed offline and in most cases is performed just once. Besides, the computational time using the trained model is about a few milliseconds («1s) for the two approaches and experiments. Finally, to support the results obtained and assuming independence between both approach with 5% of significance level, we performed the statistical comparison tests. The results revealed that there is enough evidence to say that in the two experiments the accuracy values for the forecasting with the conventional approach is less than the accuracy values for rapid and online detection approach.

In Table 6, we compared the results of (Peng et al., 2018) and (Längkvist et al., 2013) with our results. We chose these approaches because, unlike the classical feature selection method used in artificial olfactory systems, they also used the raw data to process the gas signals. In that way, in (Peng et al., 2018) was presented an approach based on a Deep Convolutional Neural Network (DCNN) tailored for gas classification but using the entire signal measurement of the gas sensors, resulting in a disadvantage regarding to our approach that lets to achieve faster results using only an early portion of the signals. In (Längkvist et al., 2013), similar to the approach proposed in our work, they considered only the transient response centered on an online solution but using unsupervised learning techniques (stacked restricted Boltzmann machines and auto-encoders), although they also focused on obtaining a rapid response, the accuracy of the system is not high. Therefore, our results are better in terms of the time needed to perform the detection. The comparison suggests that it is possible to obtain better results in accuracy and time, using our method. Therefore, our approach is promising for online analyses in E-Nose with low complexity in hardware using standard gas sensors.

5. Conclusions

In this paper, we prove that it is possible to detect wine quality thresholds in a rapid and online way using a deep MLP classifier processing an early portion of the raw data. We obtained an estimation in 2.7 s after the gas injection started when we classified three wine spoilage thresholds, and 35.13 s when we included ethanol measurements as a class. Therefore, the rapid detection method lets to make predictions 63 times faster for experiment 1, and at least five times faster for experiment 2, when compared with the conventional approach that needs the whole measurement to obtain the main odorant parameters and involves preprocessing techniques.

In this application, we employed Brazilian commercial wines. For future works, it is expected that more researches been conducted including other varieties of wines and more spoilage thresholds. Besides, the rapid detection approach could be extended to other E-Nose applications.

Acknowledgments

This work was supported by the Serrapilheira Institute (grant number Serra-1709-22626), Coordenação de Aperfeiçoamento de Pessoal de Nível Superior – Brazil (CAPES) – Finance Code 001, and Conselho Nacional de Desenvolvimento Científico e Tecnológico – CNPq (grant number 420319/2016-6). We want to thank Centro de Tecnologias Estratégicas do Nordeste (CETENE), Laboratório de Pesquisa em Infraestrutura Computacional (Juá Labs UFRPE), and Laboratório Interdisciplinar de Formação de Educadores (LIFE - UFRPE) for contributing with the development of this project.

References

Aleixandre, M., Cabellos, J. M., Arroyo, T., & Horrillo, M. C. (2018). Quantification of wine mixtures with an electronic nose and a human panel. *Frontiers in Bioengineering and Biotechnology, 6*, 1–7. February https://doi.org/10.3389/fbioe.2018.00014.

Amamcharla, J. K., & Panigrahi, S. (2010). Simultaneous prediction of acetic acidethanol concentrations in their binary mixtures using metalloporphyrin based opto-electronic nose for meat safety applications. *Sensing and Instrumentation for Food Quality and Safety, 4*(2), 51–60. https://doi.org/10.1007/s11694-010-9092-2.

Cretin, B. N., Dubourdieu, D., & Marchal, A. (2018). Influence of ethanol content on sweetness and bitterness perception in dry wines. *Lebensmittel-Wissenschaft und -Technologie- Food Science and Technology, 87*, 61–66. https://doi.org/10.1016/j.lwt.2017.08.075.

De Andrade Lima, L. L., Alexandre, S., Guerra, N. B., Pereira, G. E., De Andrade Lima, T. L., & Rocha, H. (2010). Otimização e validação de método para determinação de ácidos orgânicos em vinhos por cromatografia líquida de alta eficiência. *Quimica Nova, 33*(5), 1186–1189. https://doi.org/10.1590/S0100-40422010000500032.

Gil-Sánchez, L., Soto, J., Martínez-Máñez, R., Garcia-Breijo, E., Ibáñez, J., & Llobet, E. (2011). A novel humid electronic nose combined with an electronic tongue for assessing deterioration of wine. *Sensors and Actuators A: Physical, 171*(2), 152–158.

International, O. I. V. (2014). Organization of vine and wine. *Compendium of Internacional Methods of Analysis of wine and Musts, 1 §*.

Jackson, R. S. (2008). *Wine science: Principles and applications. Igarss 2014*. Elsevierhttps://doi.org/10.1007/s13398-014-0173-7.2.

Längkvist, M., Coradeschi, S., Loutfi, A., & Balaguru Rayappan, J. B. (2013). Fast classification of meat spoilage markers using nanostructured ZnO thin films and unsupervised feature learning. *Sensors (Switzerland), 13*(2), 1578–1592. https://doi.org/10.3390/s130201578.

Lin, X., Yang, F., Zhou, L., Yin, P., Kong, H., Xing, W., et al. (2012). A support vector machine-recursive feature elimination feature selection method based on artificial contrast variables and mutual information. *Journal of Chromatography B: Analytical Technologies in the Biomedical and Life Sciences, 910*, 149–155. https://doi.org/10.1016/j.jchromb.2012.05.020.

Lozano, J., Santos, J. P., & Horrillo, M. C. (2016). *Wine applications with electronic noses. Electronic noses and tongues in food science.* Elsevier Inchttps://doi.org/10.1016/B978-0-12-800243-8.00014-7.

Macías, M., Manso, A., Orellana, C., Velasco, H., Caballero, R., & Chamizo, J. (2012). Acetic acid detection threshold in synthetic wine samples of a portable electronic nose. *Sensors, 13*(12), 208–220. https://doi.org/10.3390/s130100208.

Martins, N., Garcia, R., Mendes, D., Costa Freitas, A. M., da Silva, M. G., & Cabrita, M. J. (2018). An ancient winemaking technology: Exploring the volatile composition of amphora wines. *Lebensmittel-Wissenschaft & Technologie, 96*, 288–295. NoveMber 2017 https://doi.org/10.1016/j.lwt.2018.05.048.

Muezzinoglu, M. K., Vergara, A., Huerta, R., Rulkov, N., Rabinovich, M. I., Selverston, A., et al. (2009). Acceleration of chemo-sensory information processing using transient features. *Sensors and Actuators B: Chemical, 137*(2), 507–512. https://doi.org/10.1016/j.snb.2008.10.065.

Normative instruction N° 14 (2018). *Brazil.* Retrieved from http://www.agricultura.gov.br/noticias/mapa-atualiza-padroes-de-vinho-uva-e-derivados/INMAPA142018PIQVinhoseDerivados.pdf.

Peng, P., Zhao, X., Pan, X., & Ye, W. (2018). Gas classification using deep convolutional neural networks. *Sensors, 18*(1), 157. https://doi.org/10.3390/s18010157.

Perestrelo, R., Rodriguez, E., & Câmara, J. S. (2017). Impact of storage time and temperature on furanic derivatives formation in wines using microextraction by packed sorbent tandem with ultrahigh pressure liquid chromatography. *Lebensmittel-Wissenschaft und -Technologie- Food Science and Technology, 76*, 40–47. https://doi.org/10.1016/j.lwt.2016.10.041.

Peris, M., & Escuder-Gilabert, L. (2016). Electronic noses and tongues to assess food authenticity and adulteration. *Trends in Food Science & Technology, 58*, 40–54. https://doi.org/10.1016/j.tifs.2016.10.014.

Rodríguez-Méndez, M. L., De Saja, J. A., González-Antón, R., García-Hernández, C., Medina-Plaza, C., Garcíia-Cabezón, C., et al. (2016). Electronic noses and tongues in wine industry. *Frontiers in Bioengineering and Biotechnology, 4*, 81. OCT https://doi.org/10.3389/fbioe.2016.00081.

Sáenz-Navajas, M. P., Avizcuri, J. M., Ballester, J., Fernández-Zurbano, P., Ferreira, V., Peyron, D., et al. (2015). Sensory-active compounds influencing wine experts' and consumers' perception of red wine intrinsic quality. *Lebensmittel-Wissenschaft und -Technologie- Food Science and Technology, 60*(1), 400–411. https://doi.org/10.1016/j.lwt.2014.09.026.

Stupak, M., Kocourek, V., Kolouchova, I., & Hajslova, J. (2017). Rapid approach for the determination of alcoholic strength and overall quality check of various spirit drinks and wines using GC–MS. *Food Control, 80*, 307–313. https://doi.org/10.1016/j.foodcont.2017.05.008.

Vazallo-Valleumbrocio, G., Medel-Marabolí, M., Peña-Neira, Á., López-Solís, R., & Obreque-Slier, E. (2017). Commercial enological tannins: Characterization and their relative impact on the phenolic and sensory composition of Carménère wine during bottle aging. *Lebensmittel-Wissenschaft und -Technologie- Food Science and Technology, 83*, 172–183. https://doi.org/10.1016/j.lwt.2017.05.022.

Vergara, A., Vembu, S., Ayhan, T., Ryan, M. A., Homer, M. L., & Huerta, R. (2012). Chemical gas sensor drift compensation using classifier ensembles. *Sensors and Actuators B: Chemical, 166*, 320–329. https://doi.org/10.1016/j.snb.2012.01.074.

Yan, J., Guo, X., Duan, S., Jia, P., Wang, L., Peng, C., et al. (2015). Electronic nose feature extraction methods: A review. *Sensors, 15*(11), 27804–27831. https://doi.org/10.3390/s151127804.

Yan, K., & Zhang, D. (2015). Feature selection and analysis on correlated gas sensor data with recursive feature elimination. *Sensors and Actuators B: Chemical, 212*, 353–363. https://doi.org/10.1016/j.snb.2015.02.025.

Zhao, Z., Yang, X., Zhao, X., Bai, B., Yao, C., Liu, N., et al. (2017). Vortex-assisted dispersive liquid-liquid microextraction for the analysis of major Aspergillus and Penicillium mycotoxins in rice wine by liquid chromatography-tandem mass spectrometry. *Food Control, 73*, 862–868. https://doi.org/10.1016/j.foodcont.2016.09.035.

Zoecklein, B. W., Fugelsang, K. C., Gump, B. H., & Nury, F. S. (1995). Volatile acidity. *Wine analysis and production* (pp. 192–198). Boston, MA: Springer. https://doi.org/10.1007/978-1-4757-6978-4_11.

PUBLISHED PAPER: ELECTRONIC NOSE DATASET FOR DETECTION OF WINE SPOILAGE THRESHOLDS

Data in brief 25 (2019) 104202

Contents lists available at ScienceDirect

Data in brief

journal homepage: www.elsevier.com/locate/dib

Data Article

Electronic nose dataset for detection of wine spoilage thresholds

Juan C. Rodriguez Gamboa [a, *], Eva Susana Albarracin E. [a], Adenilton J. da Silva [b], Tiago A. E. Ferreira [a]

[a] *Departamento de Estatística e Informática, Universidade Federal Rural de Pernambuco - UFRPE, Recife, PE, Brazil*
[b] *Centro de Informática, Universidade Federal de Pernambuco - UFPE, Recife, PE, Brazil*

ARTICLE INFO

Article history:
Received 28 March 2019
Received in revised form 23 April 2019
Accepted 20 June 2019
Available online 27 June 2019

Keywords:
Electronic nose
Chemical sensing
Machine learning
Beverage quality control
Wine spoilage

ABSTRACT

In this data article, we provide a time series dataset obtained for an application of wine quality detection focused on spoilage thresholds. The database contains 235 recorded measurements of wines divided into three groups and labeled as high quality (HQ), average quality (AQ) and low quality (LQ), in addition to 65 ethanol measurements. This dataset was collected using an electronic nose system (E-Nose) based on Metal Oxide Semiconductor (MOS) gas sensors, self-developed at the Universidade Federal Rural de Pernambuco (Brazil). The dataset is related to the research article entitled “Wine quality rapid detection using a compact electronic nose system: application focused on spoilage thresholds by acetic acid” by Rodriguez Gamboa et al., 2019. The dataset can be accessed publicly at the repository: https://data.mendeley.com/datasets/vpc887d53s/

DOI of original article: https://doi.org/10.1016/j.lwt.2019.03.074.
* Corresponding author.
E-mail address: juan.gamboa@ufrpe.br (J.C. Rodriguez Gamboa).

https://doi.org/10.1016/j.dib.2019.104202

Specifications table

Subject	Food Science; Computer Science Applications; Signal Processing
Specific subject area	Wine quality assessment using electronic nose technology
Type of data	Text files
How data were acquired	By using an electronic nose system (E-Nose) based on six Metal Oxide Semiconductor (MOS) gas sensors (MQ-3, MQ-4, MQ-6; two of each type).
Data format	Raw data, time series data
Parameters for data collection	In each experiment was used a 1 ml sample to amass the volatiles during 30 seconds inside the concentration chamber. The recorded data for each measurement corresponds to 180 seconds with 18.5 Hz sample rate. Then, the sensors were exposed to clean air for 600 seconds after the sample presentation.
Description of data collection	We collected wine samples categorized into three spoilage thresholds: low-quality (LQ), average-quality (AQ), and high-quality (HQ). In addition, we collected ethanol measurements in concentrations of 1%, 2.5%, 5%, 10%, 15%, and 20% (v/v).
Data source location	Institution: Universidade Federal Rural de Pernambuco City/Town/Region: Recife, PE Country: Brazil Latitude and longitude (and GPS coordinates) for collected samples/data: Latitude: 8° 1′ 2.68″ Longitude 34° 56′ 52.211" (Latitude: −8.017852 | Longitude: −34.94785)
Data accessibility	Repository name: Mendeley Data Data identification number: https://doi.org/10.17632/vpc887d53s.3 Direct URL to data: https://data.mendeley.com/datasets/vpc887d53s/
Related research article	J.C. Rodriguez Gamboa, E.S. Albarracin E., A.J. da Silva, L. L. de Andrade Lima, T.A. E. Ferreira, Wine quality rapid detection using a compact electronic nose system: application focused on spoilage thresholds by acetic acid, LWT - Food Science and Technology. 108 (2019) 377−384. https://doi.org/10.1016/j.lwt.2019.03.074.

Value of the data

- The dataset is available as a benchmark of E-Nose applications, focused on wine spoilage thresholds studies.
- This dataset is useful for testing classifiers and pattern recognition methods with comparison purposes in studies related to E-Nose applications.
- To the best of our knowledge, this dataset is the first one publicly available regarding commercial wines measurements acquired with E-Nose.
- These data are suitable to support E-Nose applications, helping in the decision-making of winemakers and consumers in routine tasks of wine quality control [2].

1. Data

The recorded time series was acquired at the sampling frequency of 18.5 Hz during 180 seconds, resulting in 3330 data points per sensor. Each file in the dataset has eight columns: relative humidity (%), temperature (°C), and the resistance readings in kΩ of the six gas sensors: MQ-3, MQ-4, MQ-6, MQ-3, MQ-4, MQ-6.

We organized the database in three folders for the wines: AQ_Wines, HQ_Wines, LQ_Wines, and one folder for the ethanol. Each folder contains text files that correspond to different measurements.

In the wines folders, each filename identifies a wine measurement as follows: the first 2 characters of the filename are an identifier of the spoilage wine threshold (AQ: average-quality, HQ: high-quality, LQ: low-quality); characters 4−9 indicate the wine brand; characters 11−13 indicate the bottle, and the last 3 characters indicate the repetition (another sample of the same bottle). For example, file LQ_Wine01-B01_R01 contains the time series recorded when low-quality wine of the brand 01, bottle 01, sample 01 was measured.

In the Ethanol folder, each filename identifies an ethanol measurement as follows: the first 2 characters of the filename are an identifier of Ethanol (Ea); characters 4−5 indicate the concentration in v/v (C1: 1%, C2: 2.5%, C3: 5%, C4: 10%, C5: 15%, C6: 20%); and the last 3 characters indicate the repetition. For example, file Ea-C1_R01 contains time series acquired when Ethanol at 1% v/v of concentration, sample 01 was measured.

In Fig. 1, we depicted the time series for several measurements collected in this work. The measurements displayed at the top of the figure are in resistance units (Ω), and at the bottom side are the same measurements in conductance units (S).

2. Experimental design, materials, and methods

2.1. Experimental setup

The dataset was collected with an E-Nose self-developed, that was named O-NOSE. We designed the datalogger for operating linked to a computer that has the proper software for data recording and processing, as shown in Fig. 2.

The operating mode of O-NOSE is depicted with more details in Fig. 3. The device contains two mini three-way solenoid valves ZHV 0519, and two mini air pumps PM201U (these actuators work with +5 VDC in the same way of all elements in the system) controlled by an embedded device: microcontroller Arduino Nano. The microcontroller takes charge of the data acquisition from the gas sensors and the temperature and humidity sensor DHT11 located into the sensors chamber. As well, of the timing control of the solenoid valves and the air pump, and the communication with the computer.

It used a 100 ml concentration chamber, where is placed the specimen to be analyzed. The sensor array of six MOS gas sensors manufactured by Hanwei Sensors (MQ-3, MQ-4, and MQ-6; two of each) is located into a 200 ml chamber connected to pneumatic hoses that carry the volatiles. The gas is sensed by its effect on the sensitive layer of tin dioxide (SnO_2), resulting from changes in conductivity brought about by chemical reactions on the surface of the tin dioxide particles [3,4].

The stages of the measurement process are concentration, data acquisition, and purge. The first stage aims to accumulate the analyte volatiles inside the concentration chamber for 30 seconds, to achieve this, the microcontroller activates the valve 1 and the air pump; simultaneously, deactivates the valve 2 for isolating the concentration chamber interior of the external environment. In Fig. 3, the dashed line indicates the airflow at this stage.

The data acquisition stage that lasts 3 min aims to collect the signals from the gas sensors, to achieve this, the microcontroller deactivates the valve 1 and activates the valve 2 and the air pump to direct the

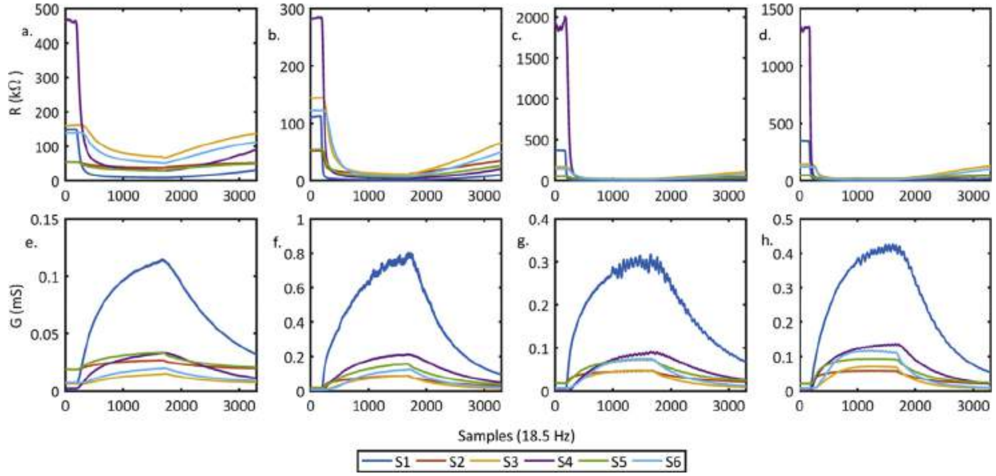


Fig. 1. Measurements acquired with our E-Nose, where S1, S2,…, S6 represent the gas sensors outputs; a. and e. correspond to the dataset file EaC1R10 (ethanol measurement); b. and f. correspond to LQWine02B01R09 dataset file; c. and g. correspond to AQWine01B01R07 dataset file; d. and h correspond to HQWine05B01R01 dataset file.

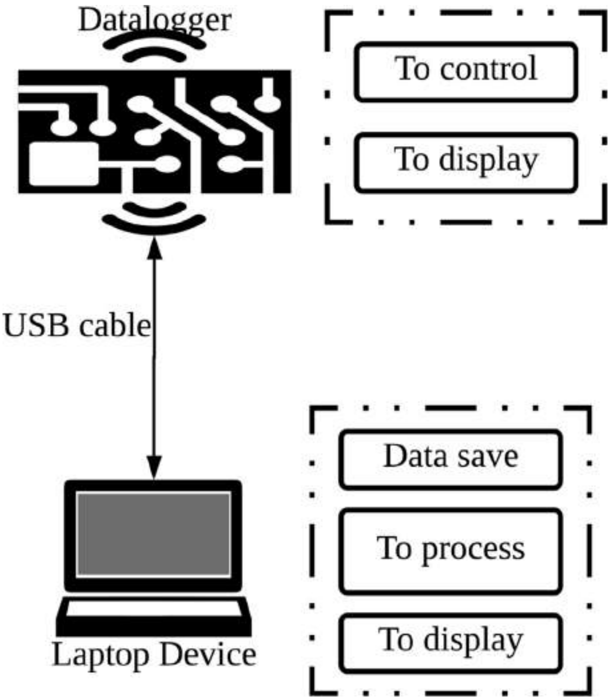


Fig. 2. Operating general diagram of O-NOSE system.

airflow from the concentration chamber dragging the volatiles towards the sensors chamber. In Fig. 3, the dotted line indicates the airflow at this stage.

The goal of the purge stage is to clean and remove volatile residues from the previous measurement during 10 minutes. Hence, the microcontroller activates the valve 1 and the air pump; simultaneously, deactivates the valve 2, the same way that for concentration stage. In Fig. 3, the dashed line indicates the airflow at this stage.

2.2. Measurement protocol

O-NOSE performs the measurement process in three stages: concentration, data acquisition (the recorded data corresponds to 180 seconds with 18.5 Hz sample rate) and purge [1]. Each measurement corresponds to the time-dependent output voltages of each gas sensor converted to resistance values according to the voltage-divider scheme [5] and the corresponding load resistor (R_L). The sensor resistance (R_S) value changes when the gas sensor is exposed to a certain specimen and was calculated as follows:

$$R_s = \frac{V_C - V_{R_L}}{V_{R_L}} \times R_L \tag{1}$$

$$V_{R_L} = \frac{ADC \times V_C}{1023} \tag{2}$$

where V_C , V_{R_L}, R_L, ADC are the standard voltage of microcontroller (5V), the output voltage, sensor load resistor, and the Analog to Digital Converter (ADC) reading, respectively [5].

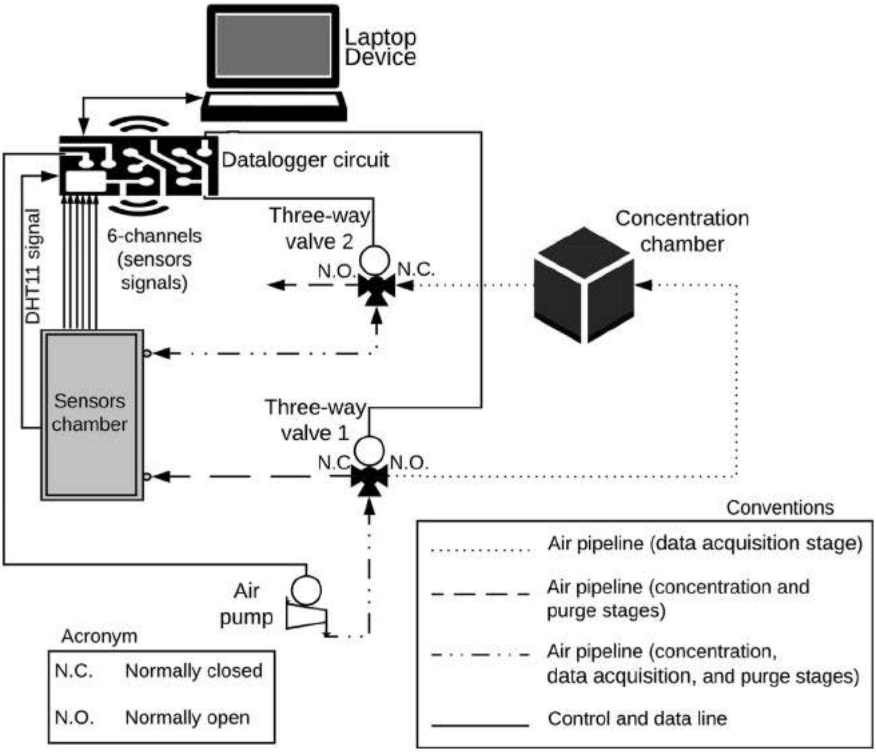


Fig. 3. Schematic diagram of O-NOSE displaying the operation stages.

2.3. Samples

We used 22 bottles of commercial wines of different varieties and vintages, elaborated in four wineries of the São Francisco valley (Pernambuco-Brazil). To obtain spoiled samples, 13 of the 22 bottles were randomly selected and left opened for six months before starting the measurements (low-quality LQ wines). Besides, four bottles were opened two weeks before beginning the data collection (average-quality AQ wines), and the remaining five bottles were opened at the starting time of each measurement (high-quality HQ wines) [1].

In addition to wines, we measured isolated ethanol in concentrations (v/v): 2, 5, 10, 20, 30, and 40 ml of ethanol diluted in distilled water to make solutions of 200 ml. These concentrations allow guaranteeing a range that covers the different possible values in wines with and without spoilage. To ensure the repeatability of the experiments using O-NOSE, we collected between 10 and 11 samples of 1mL of each wine bottles; and between 10 and 12 of the ethanol samples at their different concentrations. In this way, the database contains 235 measurements of wines divided into three groups: high quality (HQ), average quality (AQ) and low quality (LQ), with 51, 43, and 141 measurements, respectively, and 65 ethanol measurements [1].

Acknowledgments

This work was supported by the Serrapilheira Institute (grant number Serra-1709-22626), Coordenação de Aperfeiçoamento de Pessoal de Nível Superior – Brazil (CAPES) – Finance Code 001, and Conselho Nacional de Desenvolvimento Científico e Tecnológico –CNPq (grant number 420319/2016-6).

We want to thank Centro de Tecnologias Estratégicas do Nordeste (CETENE), Laboratório de Pesquisa em Infraestrutura Computacional (Juá Labs UFRPE), and Laboratório Interdisciplinar de Formação de Educadores (LIFE UFRPE) for contributing with the development of this project. Additionally, we thank Miolo Wine Group, Global wines, and Vale Do São Francisco winery for supplying the wines used in this work.

Conflict of interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

References

[1] J.C. Rodriguez Gamboa, E.S. Albarracin E., A.J. da Silva, L.L. de Andrade Lima, T.A.E. Ferreira, Wine quality rapid detection using a compact electronic nose system: application focused on spoilage thresholds by acetic acid, LWT - Food Sci. Technol. (Lebensmittel-Wissenschaft -Technol.) 108 (2019) 377–384, https://doi.org/10.1016/j.lwt.2019.03.074.

[2] M.L. Rodríguez-Méndez, J.A. De Saja, R. González-Antón, C. García-Hernández, C. Medina-Plaza, C. García-Cabezón, F. Martín-Pedrosa, Electronic noses and tongues in wine industry, Front. Bioeng. Biotech. 4 (2016) 1–12, https://doi.org/10.3389/fbioe.2016.00081.

[3] J. Rodríguez, C. Durán, A. Reyes, Electronic nose for quality control of Colombian coffee through the detection of defects in "Cup Tests, Sensors 10 (2010) 36–46, https://doi.org/10.3390/s100100036.

[4] J.C. Rodríguez-Gamboa, E.S. Albarracín-Estrada, E. Delgado-Trejos, Quality control through electronic nose system, in: Modern Approaches to Quality Control, InTech, 2011, pp. 505–522, https://doi.org/10.5772/22217.

[5] D. Rahman, R. Sarno, E. Zulaika, Data in Brief Electronic nose dataset for beef quality monitoring in uncontrolled ambient conditions, Data in Brief 21 (2018) 2414–2420, https://doi.org/10.1016/j.dib.2018.11.091.

PUBLISHED PAPER: VALIDATION OF THE RAPID DETECTION APPROACH FOR ENHANCING THE ELECTRONIC NOSE SYSTEMS PERFORMANCE, USING DIFFERENT DEEP LEARNING MODELS AND SUPPORT VECTOR MACHINES

Sensors & Actuators: B. Chemical 327 (2021) 128921

Contents lists available at ScienceDirect

Sensors and Actuators: B. Chemical

journal homepage: www.elsevier.com/locate/snb

Validation of the rapid detection approach for enhancing the electronic nose systems performance, using different deep learning models and support vector machines

Juan C. Rodriguez Gamboa [a],[*], Adenilton J. da Silva [b], Ismael C. S. Araujo [c], Eva Susana Albarracin E. [a], Cristhian M. Duran A. [d]

[a] *Departamento de Estatística e Informática, Universidade Federal Rural de Pernambuco - UFRPE, Recife, PE, Brazil*
[b] *Centro de Informática, Universidade Federal de Pernambuco - UFPE, Recife, PE, Brazil*
[c] *Departamento de Computação, Universidade Federal Rural de Pernambuco, Recife, PE, Brazil*
[d] *Facultad de ingeniería y arquitectura, GISM Group, Universidad de Pamplona, Pamplona, NdeS, Colombia*

ARTICLE INFO

Keywords:
Electronic nose
E-Nose
Rapid detection
Datasets
Deep learning
Real-time classification

ABSTRACT

Real-time gas classification is an essential issue and challenge in applications such as food and beverage quality control, accident prevention in industrial environments, for instance. In recent years, the Deep Learning (DL) models have shown great potential to classify and forecast data in diverse problems, even in the electronic nose (E-Nose) field. In this work, a Support Vector Machine (SVM) algorithm and three different DL models were used to validate the rapid detection approach (based on processing an early portion of raw signals and a rising window protocol) over diverse measurement conditions. We performed a set of experiments with five different E-Nose databases, including fifteen datasets to be used with these algorithms. Based on the obtained results, we concluded that the proposed approach has a high potential and reduces the response time for making E-nose forecasts. Because in more than 60 % of the cases, it achieved reliable estimates using only the first 30 % or fewer of measurement data (counted after the gas injection starts). The findings suggest that the rapid detection approach generates reliable forecasting models using different classification methods. Moreover, SVM seems to achieve the best accuracy and better training time.

1. Introduction

The conventional approach for data processing in the Electronic Nose implies using the entire response curves (including rising, steady-state, recovery phases, and other) of the gas sensors array. Besides, this approach includes steps such as signal pre-processing and feature generation/extraction before performing the classification tasks, which requires the selection of a suitable method for each stage, increasing the necessary time to find the appropriate classification and forecasting models [1,2]. Some works focus on reducing the steps and know-how needed for model generation, such as the works presented by Liu et al. [3] and Längkvistet al. [4]. In Liu et al. [3], a bio-inspired data processing method is proposed based on neural networks to mimic the mammalian olfactory system with high accuracy but using the entire measurement curves. In Längkvistet al. [4], the authors proposed a rapid detection system for meat spoilage using an unsupervised technique (i.e., stacked restricted Boltzmann machines and auto-encoders) that considers only the transient response. Although the obtained models offer advantages because they learn features from data instead of using hand-designed features, it may produce low suitable and inaccurate models due to the unsupervised method.

Other authors have explored another approach based on raw data treatment [5,6]. This approach reduces the steps and development time. Still, it has only been tested with the entire response curves, requiring the completion of all measurement processes and can take critical time to perform a forecast.

We proposed a novel approach on Rodriguez Gamboa et al. [7], based on processing an early portion of signals (while the measurement process is still running.) The proposed method was also tested in a wine quality application, obtaining excellent results against the traditional

[*] Corresponding author at: Departamento de Estatística e Informática, Universidade Federal Rural de Pernambuco - UFRPE, Rua Dom Manoel de Medeiros s/n, Dois Irmãos, CEP: 52171-900, Recife, PE, Brazil.
E-mail address: juan.gamboa@ufrpe.br (J.C. Rodriguez Gamboa).

https://doi.org/10.1016/j.snb.2020.128921
Received 8 May 2020; Received in revised form 8 September 2020; Accepted 14 September 2020
Available online 21 September 2020

methodology. A deep MLP classifier was trained with the raw data acquired from an E-Nose composed of six Metal Oxide (MOX) gas sensors. We achieved results around 63 times faster (Eq. 1) compared with a traditional method (using the entire response curves, applying pre-processing techniques to extract the features and later processing them using an SVM algorithm.)

$$\textit{relation of measurement time} = \left(\frac{\textit{measurement time from the starting gas injection to the finish}}{\textit{necessary time for making a forecast or window size}} \right) \tag{1}$$

Support Vector Machines (SVM) is one of the most applied methods for classification in E-Nose. Other used methods are K-Nearest Neighbors (KNN), Naive Bayes (NB), Linear Discriminant Analysis (LDA), and Adaptive Resonance Theory Map (ARTMAP) [8]. More recent approaches have used deep learning models [1,2,5,6], where the authors have also explored Convolutional Neural Networks (CNN).

The present study focuses on validating if the rapid detection approach is suitable for diverse E-Nose settings (five different databases) [7]. Additionally, we test Deep Learning (DL) techniques such as Convolutional Neural Networks (CNN) against a more classical method like SVM for classification tasks in E-Nose using the proposed approach.

2. Materials and methods

In this work, we used five E-Nose databases that include fifteen datasets to test our approach. The tested databases correspond to different E-nose systems with diverse configuration and experimental setups, guaranteeing varied conditions. We intend to make this work as a reference for further research, encouraging the research community to perform more studies or analysis by using E-Noses with numerous databases and making public the collected databases.

2.1. Databases

2.1.1. Database 1: electronic nose dataset for the detection of wine spoilage thresholds

This public database consists of time series collected through an electronic nose for a wine quality control application focused on spoilage thresholds. This database has two datasets, one of them composed of only wines (three-class classification problem), and the other comprises wines and ethanol (four classes). The database contains 235 recorded measurements of wines divided into three groups, labeled as high quality (HQ), average quality (AQ), and low quality (LQ), in addition to 65 ethanol measurements. The time series acquired at 18.5 Hz of sampling frequency during 180 s correspond to 3330 data points per sensor. Each file in the dataset has eight columns: relative humidity (%), temperature (°C), and the resistance readings in kΩ of the six MOX gas sensors: MQ-3, MQ-4, MQ-6, MQ- 3, MQ-4, MQ-6. More details are available in [7,9].

2.1.2. Database 2: electronic nose for quality control of colombian coffee through the detection of defects in "Cup tests"

This dataset consists of time-series recorded by an electronic nose used for the coffee quality control to detect defects in the grain [10,11]. The dataset contains 58 measurements of coffee samples divided into three groups and labeled as high quality (HQ), average quality (AQ), and low quality (LQ), inducing a three classes classification problem. The time series acquired at 1 Hz of sampling frequency during 300 s correspond to 2400 data points for each measurement. Where, each file in the dataset has eight columns with the resistance readings in kΩ of the gas sensors: SP-12A, SP-31, TGS-813, TGS-842, SP-AQ3, TGS-823, ST-31, TGS-800.

2.1.3. Database 3: gas sensor arrays in open sampling settings data set

The authors compiled an extensive database through a chemical detection platform for detecting potentially hazardous substances at different concentrations, composed of nine portable sensor array modules (72 metal-oxide chemical sensors in a wind tunnel facility.) Each module had eight MOX gas sensors and positioned at six different line locations normal to the wind direction. Thus, creating thereby a total number of 54 measurement locations, uniformly distributed for a total of 18,000 measurements. We split this database into six datasets (each dataset corresponds to one line location: L1, L2, …, L6). Different compounds, such as acetone, acetaldehyde, ammonia, butanol (butyl-alcohol), ethylene, methane, methanol, carbon monoxide, benzene, and toluene (ten classes) were measured to generate the database [12].

2.1.4. Database 4: gas sensor array exposed to turbulent gas mixtures Data set

The generation of this dataset used the same wind tunnel mentioned in Section 2.1.3, but the wind tunnel was adapted from the previous setup to include two independent gas sources. Besides, only one module (eight MOX gas sensors array) was used in a fixed location in the wind tunnel. The sensors array was exposed to binary mixtures of ethylene with either methane or carbon monoxide. Volatile Organic Compounds (VOCs) were released at four different rates to induce different concentration levels in the module vicinity. Each configuration was repeated six times, for a total of 180 measurements. See [13,14] for additional details. In this work, we split the dataset to generate a four-class classification problem, including the followings categories (high ethylene concentration, medium ethylene concentration, low ethylene concentration, and without ethylene.) Hence, this is a challenging problem because the measurements were performed using two interfering gases (Methane, carbon monoxide) at different concentrations, and all groups of measurements include binary mixtures of ethylene with combinations of the mentioned interfering VOC.

2.1.5. Database 5: twin gas sensor arrays data set

This database comprises the recordings of five twin devices (detection units) composed of eight gas sensors. This database has five datasets (B1, B2, …, B5) where each dataset corresponds to the measurements of one twin system (authors followed the same measuring experimental protocol in the five twin units). Every day, a different detection unit was tested using 40 distinct gas conditions, presented in random order, exposing each device to 10 concentration levels of Ethanol, Methane, Ethylene, and Carbon Monoxide (four classes). Each sensor's conductivity for 600 s in each experiment was acquired by using a sample rate of 100 Hz. The authors tested the detections platforms for 22 days, but only 16 days were collected. Hence, the complete dataset comprises 640 records [15].

2.2. Deep learning models

The models generated were implemented by using the *Python* programming language. In this study, three DL architecture implementations types were used for the classification tasks. The first type was a set

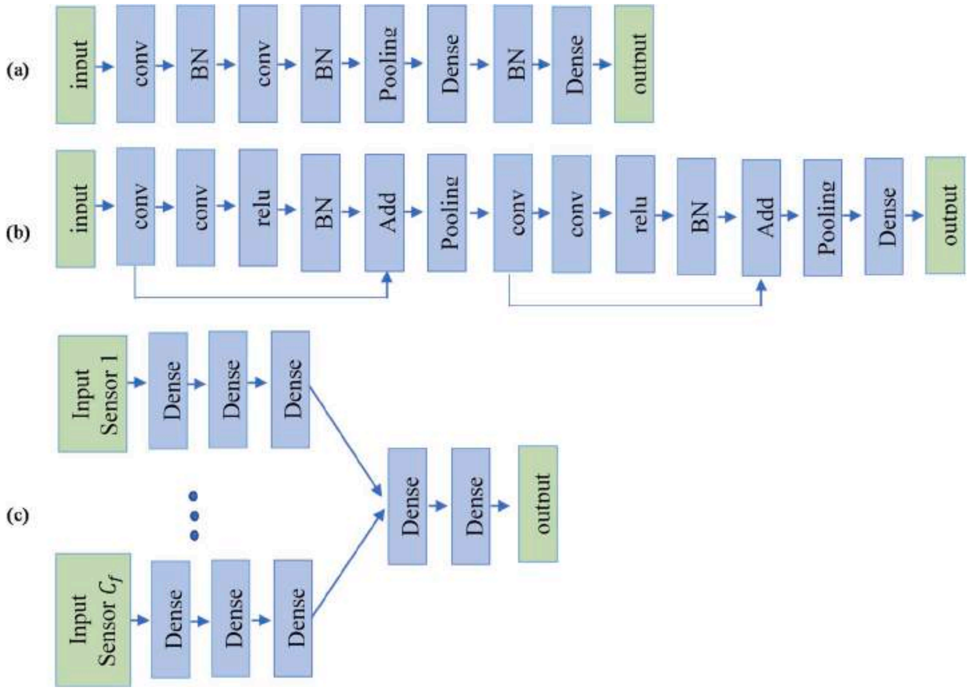


Fig. 1. The architecture of the neural networks models used in the experiments: (a) Sniff ConvNets, (b) Sniff ResNet, and (c) Sniff Multinose.

of three simple DL models named SniffNets in [16]. The SniffNets were implemented employing the machine learning framework: Keras [17]. The second architecture implementation type was a DL model to perform meta-learning, adjusting the connections between different computing cells by differentiable search to obtain the best graph configuration while training. The authors called that methodology as Differentiable Architecture Search (DARTS) [18]. The DARTS implementation has been made available by its authors, but we adapted it so that the model could fit the shape of the target data. This implementation was created using the PyTorch library [19]. Finally, the third model corresponds to a simple Deep MLP model with only fully connected layers based on the model used in Rodriguez Gamboa et al. [7].

The input format used for the models with convolutional layers was a feature matrix with dimensions R_f x C_f, where R_f corresponds to the rows (represents the time interval) and the columns C_f that corresponds to the gas sensors used to detect the specimens. And N_n denotes the number of neurons of a given fully connected (Dense) layer.

2.2.1. SniffNets

Three models with different architecture implementations were generated based on [16], adapting the architectures to test the proposed rapid detection approach [7]. On these models, the *softmax* was used as the activation function for the output layer.

The first model is a convolutional network [20] named Sniff ConvNet in this document, which consists of two layers that apply a bi-dimensional convolution (Conv2D), followed by two fully connected (FC) layers. We used the activation function *ReLU* in both Conv2D and FC layers. The second model is a residual network [21] named Sniff ResNet composed of two residual blocks, each with two Conv2D layers. In each block, the first convolutional layer has a skip connection joined to the second convolutional layer output. Two FC layers follow the two residual blocks. Like the Sniff ConvNet model, we used the activation function *ReLU* in the Conv2D and FC layers. The third model is a fusion neural network called Sniff Multinose. In this case, we adopted a different approach, where the feature matrix has a shape R_f X C_f. We split the feature matrix by columns C_f, and each one was used as an input of a Multilayer Perceptron (MLP) model. Then, we concatenated the outputs of all MLP models and utilized them as inputs of another MLP network to complete the classification model. Fig. 1 depicts the basic configuration of the Sniff ConvNet, Sniff ResNet, and Sniff Multinose.

2.2.2. DARTS: differentiable architecture search

Searching for optimal neural network architectures is a task that can be both difficult and time-consuming. DARTS [18] algorithm uses differentiation to perform this search. The algorithm performs the search in a network considered as a directed acyclic graph. Each node xi represents the output of a subnetwork in the chart. For example, xi can be a feature vector from a fully connected Multilayer perceptron or a feature map from a convolutional layer. Let O be the set of all the arcs(i,j) being an operation between the i-th node to the j-th node pondered by a factor α(i,j). The arc(i, j) represents the connection between nodes xi and xj. This connection is the o(i,j) operation with inputs Xi and outputs Xj. Afterward, the initialization of a set of candidate operations between all nodes (i, j) of the graph, the search task is then performed by first computing the gradient of the loss function for the factors α(i,j) and then concerning the weights of the model. Thus, after computing the minimal

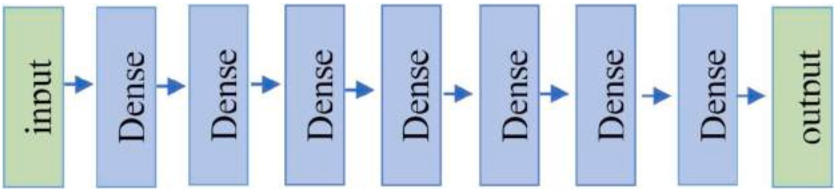


Fig. 2. The architecture of the neural network models used in the MLP experiments.

Table 1
Summary of experiments, showing the test data accuracy of the best window (b-win) and the accuracy of the last window (l-win), i.e., using all information after the gas injection. The most outstanding performance for each dataset displays highlighted in red color and green color for b-win and l-win columns, respectively. The means values obtained using the holdout cross-validation method as pointed in Section 2.2.4.

Dataset Method	Sniff-ConvNet		Sniff-Resnet		Sniff-Multinose		DARTS		MLP		SVM	
	b-win	l-win	b-win	l-win	b-win	l-win	b-win	l-win	b-win	l-win	b-win	l-win
Wines	97.5	97.5	97.5	97.5	99.1	93.2	100	87	100	95.7	100	100
	w3	w10	w3	w10	w4	w10	w1	w10	w4	w10	w2	w10
Wines & ethanol	98	95.3	94	94	96.7	94.7	98.3	95	100	95	100	98.3
	w2	w10	w10	w10	w4	w10	w9	w10	w2	w10	w1	w10
Coffee	79.3	65.5	79.3	44.8	89.7	69	92	67	91.6	66.6	100	100
	w1	w10	w8	w10	w3	w10	w3	w10	w5	w10	w1	w10
Wind tunnelL1	79.6	79.6	85.4	69.9	92.2	91.3	96.77	95.16	82.9	73.2	90.2	73.2
	w2	w10	w2	w10	w5	w10	w5	w10	w5	w10	w2	w10
Wind tunnelL2	95.8	83.2	97.9	81	97.9	92.6	98.4	98.4	92.1	78.9	97.4	89.5
	w4	w10	w3	w10	w2	w10	w7	w10	w6	w10	w6	w10
Wind tunnelL3	96.1	90.3	92.2	87.4	99	94.2	98.4	98.4	90.2	92.7	100	97.6
	w6	w10	w7	w10	w4	w10	w7	w10	w3	w10	w5	w10
Wind tunnelL4	91.2	95.1	96.1	81.4	100	96.1	98.4	98.4	95.1	87.8	97.6	95.1
	w9	w10	w6	w10	w9	w10	w7	w10	w7	w10	w2	w10
Wind tunnelL5	94.2	91.3	89.3	85.4	93.2	98	98.4	98.4	90.2	87.8	95.1	80.5
	w7	w10	w5	w10	w6	w10	w7	w10	w4	w10	w5	w10
Wind tunnelL6	95.7	88.2	89.2	90.3	100	95.7	98.4	98.4	91.9	89.2	94.6	94.6
	w4	w10	w6	w10	w2	w10	w7	w10	w8	w10	w4	w10
Turbulent gas mixtures	60	52.2	53.3	34.4	57.7	30	73.3	60.1	72.2	72.2	77.7	86.1
	w3	w10	w4	w10	w7	w10	w3	w10	w10	w10	w4	w10
Twin gas sensorB1	100	97.5	97.5	97.5	97.5	97.5	100	93	100	100	100	100
	w2	w10	w7	w10	w10	w10	w9	w10	w3	w10	w3	w10
Twin gas sensorB2	97.5	91.2	96.2	100	100	80	96.9	92.1	100	90.6	100	96.9
	w3	w10	w6	w10	w1	w10	w10	w10	w1	w10	w1	w10
Twin gas sensorB3	100	92.5	100	90	98.7	88.7	96.9	94.4	100	93.7	100	100
	w8	w10	w8	w10	w2	w10	w10	w10	w3	w10	w2	w10
Twin gas sensorB4	92.5	52.5	100	92.5	100	90	87.5	77.6	100	62.5	100	100
	w3	w10	w8	w10	w3	w10	w9	w10	w5	w10	w2	w10
Twin gas sensorB5	100	77.5	100	92.5	100	87.5	100	80.1	100	87.5	100	100
	w3	w10	w9	w10	w1	w10	w10	w10	w1	w10	w1	w10

loss concerning the α and the weights in the arcs between (i,j), the algorithm determines the optimal architecture according to the values of α [18]. We do not display the architecture scheme for the DARTS algorithm because it is too large and increase the number of pages in this document.

2.2.3. Deep MLP model

We also used a Deep MLP model presented in [7]. The configuration of the model consists of eight layers with *Tanh* as the activation function except for the output layer, in which we used *softmax*. The input layer has 100 neurons, and all the hidden layers have 30 neurons. Fig. 2 depicts the basic configuration of the MLP architecture.

2.2.4. Training configurations

Three sets of DL models were trained until to reach 20 epochs by using the Stochastic Gradient Descent (SGD) algorithm for optimization with a learning rate of 0.001 and a momentum of 0.9. Besides, we used the categorical cross-entropy loss function.

Regarding the training process in all tested classification methods, all datasets were randomly split as follows, training group including 80 % of measurements and the validation group with 20 %. Besides, we used the holdout cross-validation method.

2.2.5. Configurations for the SVM model

An SVM model available in the scikit-learn library was used. Furthermore, we defined the following parameters to optimize the model: A Radial Bayes Function (RBF) as the kernel, the regularization parameter C as 10, and the other settings as the default value. Given a dataset *D* with vectors of *n* features, the value computed for the *gamma* parameter is $(n \bullet variance(D_{flat}))^{-1}$. Where *variance* (D_{flat}) is the variance over the flattened dataset. The algorithm computed the *gamma* value over the normalized data, using standardization or *z-score normalization*.

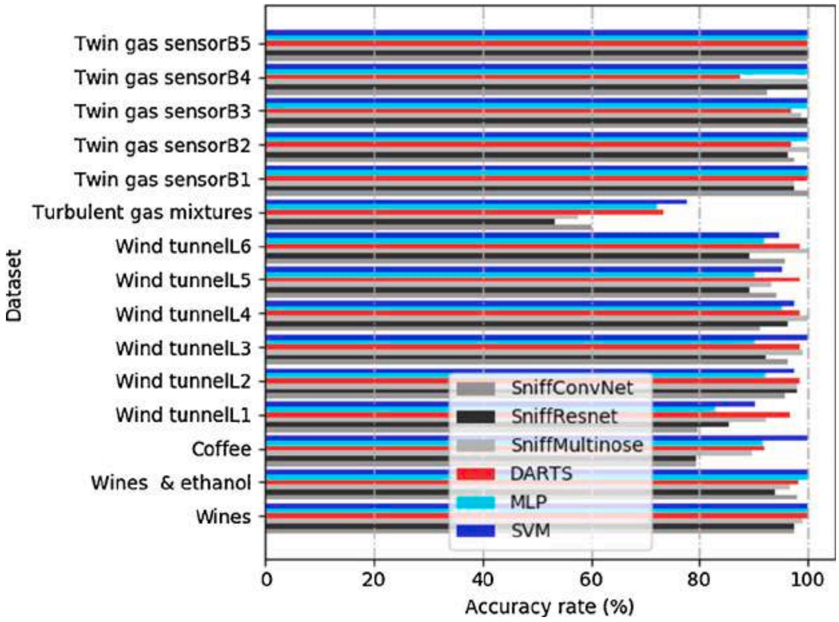


Fig. 3. The classification accuracy rate of the test data over the 15 datasets for the tested methods.

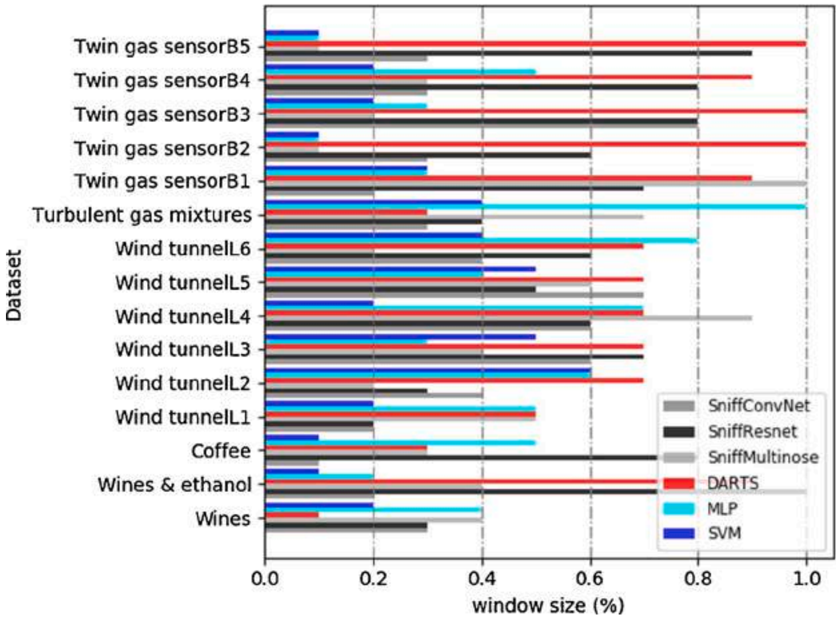


Fig. 4. The window size in percentage regarding the complete measurements over the 15 datasets for the tested methods.

3. Results and discussion

According to the rapid detection approach, the rising window protocol was applied to find the early portion with the best validation accuracy in each dataset. We used different methods such as the DARTS search model architecture, a deep MLP, three DL models called Sniff (ConvNet, Resnet, and Multinose), and SVM, to validate the proposed approach and determine whether it could be applied independently to the classification method. Table 1 summarizes the experimental results to compare the test accuracy on the best and the last window. The results let infer that the models generated with the best windows are similar or outperform in the majority of cases (94.44 %) the models obtained using the complete information of the measurements. Results using other window sizes were not displayed in this document to not increase the number of sheets. Although, in some cases, the accuracy could be similar to the obtained with the best window (best accuracy), the method focuses on using the smaller early portion of information to perform a reliable forecast. Fig. 3 depicts the accuracy of the test data in the best windows for easy visualization.

The best windows size for each dataset and model are depicted in Fig. 4. It is important to remark that the first-window (w1) corresponds to the first 10 % of measurement data, the second-window (w2) has 20 % of the information, continuing until the tenth-window (w10) that has 100 % of measurements data. The results suggest that 42.22 % of the cases, we can obtain suitable models using only 30 % or even less information, and 66.66 % of the time by applying the SVM classifier. The necessary time to get the appropriate models for each dataset and classification method is detailed in Table 2. Those times are only a reference to report which methods work faster in the training process when is adopted the rapid detection approach.

Table 2
Training time in seconds for the tested classifications methods on the 15 datasets, only for reference.

Dataset Method	Sniff-ConvNet	Sniff-Resnet	Sniff-Multinose	DARTS	MLP	SVM
Wines	1475.2	148.93	83.55	5940	23.63	0.98
Wines & ethanol	1885.66	191.22	114.2	7380	33.28	1.65
Coffee	237.16	134.23	39.74	960	99.9	0.0375
Wind tunnelL1	509.87	143.09	228.52	3360	65.676	6.22
Wind tunnelL2	492.57	143.75	228.69	3120	71.746	3.44
Wind tunnelL3	527.76	162.31	256.22	3360	82.245	5.65
Wind tunnelL4	546.04	173.22	169.41	3300	90.31	5.82
Wind tunnelL5	560.77	184.49	201.69	3300	99.32	6.17
Wind tunnelL6	545.27	185.97	285.59	2820	104.273	5.11
Turbulent gas mixtures	3119.55	105.13	115.94	3120	55.734	2.83
Twin gas sensorB1	5137.12	453.2	105.99	3300	28.692	1.1
Twin gas sensorB2	5035.33	79.2	124.11	3360	24.93	1.2
Twin gas sensorB3	5027.64	80.42	110.2	3360	31.236	1.1
Twin gas sensorB4	2588.93	67.41	68.47	1800	32.4921	0.4
Twin gas sensorB5	2574.06	79.88	84.79	1860	40.126	0.4

Results showed similar accuracy for each dataset comparing the tested classification methods. The main difference is presented in the size of the best window. By comparing the Sniff models (gray bars in Figs. 3 and 4), SniffMultinose reached the best-combined performance on the 15 datasets because it has better average accuracy, shorter average training time, and an average size for the best window of 42 %, comparable to Sniff-ConvNet 40 %. The DARTS algorithm (red bars in Figs. 3 and 4) generates models that outperform the models created by the Sniff and MLP architectures tested in this work. Still, it is the method that needs more time in the training process to generate reliable models and reached the worst average size for the best window 71.33 %, more significant than the obtained with Sniff-Resnet 61.33 %, MLP 44.66 %, and SVM 27.33 %.

Analyzing the window size with the best accuracy on the test data, we conclude that based on the tested classification methods (DL techniques and SVM), the rapid detection approach is reliable for electronic nose applications. The results achieved in this study validate the use of the proposed method in E-Nose datasets. Besides, it is relevant to remark that in the majority of cases, the SVM classifier (blue bars in Figs. 3 and 4) generates models that use only an early portion of information, which entails faster forecasts. Additionally, the training time is relatively less, reducing the computational cost. Therefore, the mentioned findings suggest that in the electronic nose field, SVM is competitive with deep learning techniques for classification tasks. DL techniques increase the necessary time to generate reliable models and, in some E-nose datasets do not reach better results.

4. Conclusions

In this research, we validated the rapid detection approach [7] in several datasets with diverse electronic nose settings, showing that it is suitable in this field, with better or similar accuracy in 14 of 15 datasets, when compared with a conventional approach that needs the complete information of the measurements.

The investigation allowed finding that in the majority of times is possible to obtain a reliable forecast using only the first 30 % (even less) of the measure after the gas injection started. Therefore, subsequent investigations could focus on generating models using only this portion of the gas sensors signals, which entails reducing the time to produce models and make the forecasts (accelerating response).

In this work, the proposed approach was also validated by using several classification methods; the SVM algorithm and three different DL architectures: (i) the Differentiable Architecture Search (DARTS) algorithm, (ii) three deep learning models based on SniffNets, and (iii) a Deep MLP. Although deep learning models are useful when there is a large volume of data, and it can automatically identify patterns. The results showed that using SVM models in the majority of cases, the results are similar or even better and were consistent concerning the early portion of signals needed to make reliable forecasts. Therefore, SVM still is an option in the electronic nose field and could be used to apply the rapid detection approach, as well, the tested deep learning techniques. Still, SVM needs less time for the training process against the other tested classifications methods.

Credit author statement

Author A: Juan C. Rodriguez Gamboa
Author B: Adenilton J. da Silva,
Author C: Ismael C. S. Araujo,
Author D: Eva Susana Albarracin E.,
Author E: Cristhian M. Duran A.

A, B and D conceived the presented idea. A, B and D developed the hypothesis and performed the initials experiments. A and D performed the computations and verified the methods. B and E encouraged A and D to investigate related works and refine the proposed approach. A, B, and C carried out the experiments, and A and D wrote the manuscript with support from B, and C. All authors discussed the results and contributed to the final manuscript.

Data availability

All the code and data used in this work are publicly available at https://github.com/IsmaelCesar/SniffNets.

Declaration of Competing Interest

The authors report no declarations of interest.

Acknowledgments

We gratefully acknowledge the support of the following Brazilian agencies: CAPES (Finance Code 001), CNPq (grant number 308730/2018-6), and FACEPE (grant number BIC-1528-1.03/18), who promoted and financed our research project. We gratefully acknowledge the support of NVIDIA Corporation with the donation of the Titan Xp GPU used for this research.

References

[1] P.-F. Qi, Q.-H. Meng, M. Zeng, A CNN-based simplified data processing method for electronic noses, 2017 ISOCS/IEEE Int. Symp. Olfaction Electron. Nose, IEEE (2017) 1–3, https://doi.org/10.1109/ISOEN.2017.7968887.

[2] Y.-J. Liu, Q.-H. Meng, X.-N. Zhang, Data processing for multiple electronic noses using sensor response visualization, IEEE Sens. J. 18 (2018) 9360–9369, https://doi.org/10.1109/JSEN.2018.2871599.

[3] Y.-J. Liu, M. Zeng, Q.-H. Meng, Electronic nose using a bio-inspired neural network modeled on mammalian olfactory system for Chinese liquor classification, Rev. Sci. Instrum. 90 (2019), 025001, https://doi.org/10.1063/1.5064540.

[4] M. Längkvist, S. Coradeschi, A. Loutfi, J. Rayappan, M. Längkvist, S. Coradeschi, A. Loutfi, J.B.B. Rayappan, Fast classification of meat spoilage markers using nanostructured ZnO thin films and unsupervised feature learning, Sensors 13 (2013) 1578–1592, https://doi.org/10.3390/s130201578.

[5] P. Peng, X. Zhao, X. Pan, W. Ye, P. Peng, X. Zhao, X. Pan, W. Ye, Gas classification using deep convolutional neural networks, Sensors 18 (2018) 157, https://doi.org/10.3390/s18010157.

[6] G. Wei, G. Li, J. Zhao, A. He, G. Wei, G. Li, J. Zhao, A. He, Development of a LeNet-5 gas identification CNN structure for electronic noses, Sensors 19 (2019) 217, https://doi.org/10.3390/s19010217.

[7] J.C. Rodriguez Gamboa, E.S. Albarracin E, A.J. da Silva, L.L. de Andrade Lima, T.A. E. Ferreira, Wine quality rapid detection using a compact electronic nose system: application focused on spoilage thresholds by acetic acid, LWT 108 (2019) 377–384, https://doi.org/10.1016/J.LWT.2019.03.074.

[8] S.K. Jha, R.D.S. Yadava, K. Hayashi, N. Patel, Recognition and sensing of organic compounds using analytical methods, chemical sensors, and pattern recognition approaches, Chemometr. Intell. Lab. Syst. 185 (2019) 18–31, https://doi.org/10.1016/J.CHEMOLAB.2018.12.008.

[9] J.C. Rodriguez Gamboa, E.S. Albarracin E, A.J. da Silva, T.A. Tiago, Electronic nose dataset for detection of wine spoilage thresholds, Data Br. 25 (2019), 104202, https://doi.org/10.1016/j.dib.2019.104202.

[10] J.C. Rodriguez Gamboa, C.M. Duran Acevedo, Dataset: Electronic Nose for Quality Control of Colombian Coffee Through the Detection of Defects in “Cup Tests”, 2009, https://doi.org/10.17632/7spd6fpvyk.1.

[11] J. Rodríguez, C. Durán, A. Reyes, Electronic nose for quality control of Colombian coffee through the detection of defects in «Cup Tests», Sensors (Basel) 10 (2010) 36–46, https://doi.org/10.3390/s100100036.

[12] A. Vergara, J. Fonollosa, J. Mahiques, M. Trincavelli, N. Rulkov, R. Huerta, On the performance of gas sensor arrays in open sampling systems using Inhibitory Support Vector Machines, Sens. Actuators B Chem. 185 (2013) 462–477, https://doi.org/10.1016/J.SNB.2013.05.027.

[13] J. Fonollosa, I. Rodríguez-Luján, M. Trincavelli, A. Vergara, R. Huerta, Chemical discrimination in turbulent gas mixtures with MOX sensors validated by gas chromatography-mass spectrometry, Sensors 14 (2014) 19336–19353, https://doi.org/10.3390/s141019336.

[14] J. Fonollosa, I. Rodríguez-Luján, M. Trincavelli, R. Huerta, Data set from chemical sensor array exposed to turbulent gas mixtures, Data Br. 3 (2015) 216–220, https://doi.org/10.1016/J.DIB.2015.02.022.

[15] J. Fonollosa, L. Fernández, A. Gutiérrez-Gálvez, R. Huerta, S. Marco, Calibration transfer and drift counteraction in chemical sensor arrays using Direct Standardization, Sens. Actuators B Chem. 236 (2016) 1044–1053, https://doi.org/10.1016/J.SNB.2016.05.089.

[16] I.C.S. Araujo, A.J. Silva, J.C. Rodriguez Gamboa, Modelos de deep learning para classificação de gases detectados por matrizes de sensores nariz artificial, in: An. Do encontro Nac. Inteligência Artif. e Comput. (eNIAC 2019), Sociedade Brasileira de Computação - SBC, Salvador - Brasil, 2019, pp. 844–855.

[17] L. Buitinck, G. Louppe, M. Blondel, F. Pedregosa, A. Mueller, O. Grisel, V. Niculae, P. Prettenhofer, A. Gramfort, J. Grobler, R. Layton, J. VanderPlas, A. Joly, B. Holt, G. Varoquaux, API Design for Machine Learning Software: Experiences From the Scikit-Learn Project, CoRR. abs/1309.0, 2013, http://arxiv.org/abs/1309.0238.

[18] H. Liu, K. Simonyan, Y. Yang, DARTS Differentiable Architecture Search, CoRR. abs/1806.0, 2018, http://arxiv.org/abs/1806.09055.

[19] A.L. Adam Paszke, Sam Gross, Soumith Chintala, Gregory Chanan, Edward Yang, Zachary DeVito, Zeming Lin, Alban Desmaison, Luca Antiga, Automatic differentiation in PyTorch. NIPS 2017 Work. Autodiff Decis. Progr. Chairs, 2017.

[20] A. Krizhevsky, I. Sutskever, G.E. Hinton, ImageNet Classification With Deep Convolutional Neural Networks, 2012.

[21] K. He, X. Zhang, S. Ren, J. Sun, Deep Residual Learning for Image Recognition, 2016.

Juan C. Rodríguez-Gamboa was born on December 20, 1982, in Cúcuta - Colombia. He received an undergraduate degree in electronic engineering at Universidad de Pamplona (2007), Colombia. He has a master’s degree in automation and industrial control from Instituto Tecnológico Metropolitano (2014), Colombia. Currently is a doctoral student in biometrics and applied statistics at Universidade Federal Rural de Pernambuco, Brazil, and he had defended his doctoral thesis in February 2020. He has worked as a professor of the engineering faculty at Instituto Tecnológico Metropolitano, Tecnológico de Antioquia, Universidad de San Buenaventura, and Universidad de Pamplona, Colombia. His current research interests include electronic nose systems, pattern recognition, machine learning, digital signal processing, virtual sensing, and industrial automation.

Adenilton J. da Silva received the Licentiate degree in mathematics from Universidade Federal Rural de Pernambuco, Brazil, and received the pH.D. degree in Computer Science from Universidade Federal de Pernambuco (UFPE), Brazil. He joined the Centro de Informática at UFPE in 2019 where he is now Adjunct Professor. His current research interests include quantum computation, artificial neural networks, machine learning, and hybrid neural systems.

Ismael C. S. de Araujo is a Computer Science Bachelor’s student at Universidade Federal Rural de Pernambuco (UFRPE). He has worked on projects such as "Quantum Machine Learning: models and learning algorithms" and "Quantum Neural Networks: models, architecture selection and learning algorithms," both in the field of quantum computing and quantum machine learning. Other of his interests are related to machine learning applications, and currently is working on several related projects.

E. Susana Albarracín-Estrada received an undergradute degree in electronic engineering at the Universidad Francisco de Paula Santander (UFPS), Cúcuta, Colombia in 2005; and the master’s degree in automation and industrial control from the Instituto Tecnológico Metropolitano (ITM) in 2014, Colombia. She has worked as a professor at the UFPS, ITM, Universidad de San Buenaventura (USB) and Tecnológico de Antioquia (TdeA) in telecommunications, electronic and electrical circuits area. She is currently pursuing the PhD degree in the Biometria e Estatística Aplicada program, Universidade Federal Rural de Pernambuco (UFRPE), Brazil. Her research interests include pattern recognition, electronic nose systems, chemical sensors, drift compensation, automation and industrial control.

Cristhian M. Durán was born on July 12 of 1973, in Pamplona, Colombia. He received an undergraduate degree as an electronic engineer at the Universidad de Pamplona in 2000, a pH.D. degree in electronic engineering in 2005 from the University Rovira I Virgili, Spain. He is currently working as a full professor at the Universidad de Pamplona, Colombia, leading the Multisensory Systems and Pattern Recognition research group, and is presently coordinating two European projects (RISE, H2020) from Universidad de Pamplona, Colombia. His research background is related to multisensory systems (i.e., design and construction of electronic noses and tongues), pattern recognition methods, control and industrial automation, data acquisition, and artificial intelligence. He has published around 45 papers and has attended more than 50 conferences (oral and poster presentation).