agronomy

MDPI

Communication

Low-Cost Electronic Nose for Wine Variety Identification through Machine Learning Algorithms

Agustin Conesa Celdrán [1], Martin John Oates [1], Carlos Molina Cabrera [1], Chema Pangua [2], Javier Tardaguila [3] and Antonio Ruiz-Canales [1,*]

[1] Engineering Department, Miguel Hernández University of Elche, 03312 Orihuela, Spain
[2] Bodega Viñedos Alfaro (La Rioja), 26559 Aldeanueva de Ebro, Spain
[3] Altavitis Technologies SL, 26004 Logroño, Spain
* Correspondence: acanales@umh.es; Tel.: +34-649581413

Abstract: The aroma of wine is traditionally analyzed by sensory methods or by using gas chromatography; both analytical methodologies are slow and expensive and do not allow continuous monitoring. For this reason, interest in rapid methods has increased in recent times. Electronic noses (e-noses) stand out for their high sensitivity, speed, low cost, and little or no sample preparation. They present, however, low selectivity, which requires advance analytical methods to distinguish compounds. Here, we present a low-cost e-nose device for the analysis and identification of distinct varieties of wine. Chemical analysis data are compared to e-nose data through a principal component analysis (PCA) and a k-means clustering algorithm to establish relationships between varieties of wines and the e-nose classification capability. The results show that e-nose technology found significant differences between the analyzed samples, and furthermore, classifying the samples in accordance with the chemical analysis classification. The maximal accuracy obtained was 100% using the k-means algorithm for binary classification with N = 21 samples. Thus the potential of e-nose technology was shown in the wine industry for the identification and classification of wine varieties or quality.

Keywords: electronic nose; viticulture; physicochemical properties; PCA analysis; volatile organic compounds (VOCs)

Citation: Celdrán, A.C.; Oates, M.J.; Molina Cabrera, C.; Pangua, C.; Tardaguila, J.; Ruiz-Canales, A. Low-Cost Electronic Nose for Wine Variety Identification through Machine Learning Algorithms. *Agronomy* **2022**, *12*, 2627. https://doi.org/10.3390/agronomy12112627

Academic Editors: Daniel García Fernández-Pacheco, José Miguel Molina Martínez and Dolores Parras-Burgos

Received: 8 September 2022
Accepted: 19 October 2022
Published: 25 October 2022

1. Introduction

The composition of wine is mainly ethanol and water. However, over 20 compounds influence the basic flavor of wine, and many other compounds present in wines are responsible for the specific characteristics of each wine and its aromas. These aroma compounds (ACs) may or may not be present and differ in concentrations in different wines which makes them have a specific aroma profile [1]. To date, over 700 ACs have been identified, which is strong evidence for the complexity of a wine [2]. Many ACs are volatile chemical compounds (VCCs) which are present in the complex gas mixtures of some foods and beverages, including wine [3]. Thus, this gives us a way of characterizing wines by their aromas. The task of discriminating wine by its aroma is not easy, requires expert knowledge and training, and is a slow and expensive process.

Here is where e-noses play an important role, introducing an improvement to other techniques such as sensory methods or gas chromatography. E-nose technology is based on an array of non-specific sensors. They react to gases and generate distinct signals which, after been processed for feature extraction, can be used to identify compounds and for classifying odor emitting samples. Thus, e-noses are capable of detecting complex mixtures of volatile compounds present in gas samples. These VCC mixtures generate a combined response in the sensors and create an odor pattern. Employing data analyses such as principal component analysis (PCA), cluster analysis, and classification techniques such as

artificial neural networks (ANN) or support vector machines (SVM), this technology has the potential to accurately classify samples by their odor [4]. Wine is one more of the many potential uses that e-noses have.

Many studies on wine analysis using e-noses have mainly focused on the detection of wine spoilage thresholds [5–7], the discrimination of wines by their aromas [8–10], or early detection by monitoring the wine production process [5]. All the devices used to perform these studies have been designed on the same basis: a sample chamber, a gas extraction device (air pump, fan), a sensing chamber, a sensor array, and a computer. Many modifications may be applied to this configuration, such as air or sample heating steps, gas compression or expansion valves, or varying the volume of the sampling chambers.

In this report, an e-nose prototype based on eight metal oxide MQ sensors was used to analyze and characterize wine samples from La Rioja (Spain). Repeatability and reproducibility of the device was also studied, as well as long-term device drift. A PCA analysis was used to group the wine samples, and a comparative study was conducted using basic chemical analysis of the wine to back up the results from the e-nose prototype. We show the potential capability of this device to discriminate wine samples similar to a chemical analysis, decreasing the analysis cost and expertise needed to carry out the analysis and characterization.

2. Materials and Methods

The wine samples were kept at temperatures between 4 and 8 °C. To prevent deterioration, the samples were sealed in dark glass bottles to prevent oxidation, light damage, and contamination by recipient material. A total of 10 samples were analyzed and coded as follows: 43D, F1F, 200D, 203D, 9F, P2F, 248D, 252D, 11F, and 42F. In this case, three different varieties of Rioja Spanish wine types were selected: 43D and F1F are corresponding to the Graciano variety; 200D, 203D, 9F, and P2F corresponded to the Grenache variety; 248D, 252D, and 11F corresponded to the Tempranillo variety; and the sample with the code 42F corresponded to Mazuelo variety. All the samples were obtained from the harvest of 2019. For each variety, 5 and 6 samples were selected. The main objective was to detect the differences of varieties among wines by means of the described e-nose prototype. To compare the analysis carried out by the e-nose, a ground truth knowledge about the samples was obtained by performing a basic chemical analysis. From this data, information could be extracted that provided independent information about the samples but could also be used for a comparative analysis by performing cross-correlations with the e-nose data. The main chemical analyses were as follows: First, the quantity of ethanol in the wine (percentage) was determined by distillation/densitometry. Distillation of a wine sample separates the volatile from the non-volatile components. The alcohol content of the resulting distillate can then be readily measured using a density meter. Moreover, volatile acidity determination ($gr \cdot L^{-1}$) can be determined using the Cash still method. Next, the total acidity in the wine ($gr \cdot L^{-1}$ of tartaric acid) was determined. Then, the pH values of the obtained samples were determined. The next analysis was the obtention of reducing sugars by means of a refractometer or a hydrometer. The obtention of malic acid was another analysis of the samples ($gr \cdot L^{-1}$) performed by using enzymatic measurement, which involved the conversion of malic acid using a specific enzyme. It can be monitored directly by measuring the absorbance (344 nm) resulting from the generation of a by-product of the reaction (NADH). The quantity of CO_2 in wine was also obtained ($mg \cdot L^{-1}$) by employing a carbodoseur. The quantity of anthocyanin was calculated ($mg \cdot L^{-1}$). In addition, the temperature of saturation (TSAT, in °C) was determined to test wine stability. It uses conductivity to measure the amount of tartrate that can be absorbed by a wine at room temperature. These analyses are very common in viticulture. All the previous cited analyses have been done by means of FTIR, a multiparametric analytical technique. Fourier transform infrared (FTIR) is a spectroscopic analysis technique that uses a part of the electromagnetic spectrum. Specifically, wavelengths between 2500 nanometers (nm) and 25,000 nm, which is the "mid-infrared" region, and thus, the term FTIR. Although

the abbreviation FTIR is, generally, the name of a mathematical technique used to convert numerical data into useful results, this term has also been popularized to describe this analytical technique.

We based our knowledge of the samples on how they correlated to this information. As our goal was to compare the samples and study similarities and differences between them. To do so we use as a baseline a chemical analysis. We compared the data from the e-nose by means of clustering and the data form chemical analysis. By doing so we can study how the samples are related by the enose and see if these rela-tions has correspondence with the chemical relation between them.

An electronic nose (EN) device was used to perform the analysis of each sample.

This electronic nose prototype included a simple sample delivery system (sample chamber where samples are deposited, along with an air pump or fan), a sensor array, and a data processing unit or microcontroller (Arduino Nano microcontroller with USB serial connection). This EN prototype was based on an array of eight MOS sensors (MQ-135, MQ-2, MQ-3, MQ-4, MQ-5, MQ-7, MQ-8, and MQ-9) fabricated by Hanwei Electronics Co., Ltd. (Zhengzhou, China). These sensors have resistances (RL) which change their values depending on the gas mixture present in them. Varying the voltage across these resistances introduces modulation of the heating of the sensor. The device works with a microcontroller Arduino Nano to generate the voltage signals introduced in the sensors and measure the responses from these, along with an analog circuit comprising a DAC and operational amplifiers to effectively control the heating of the sensors. Datasheets for some sensors (MQ7 and MQ9 in particular) recommend switching between two heater voltages (5.0 V and 1.4 V) on a 60 + 90 s cycle, with the sensor response at the end of the 90 s interval. This can significantly improve sensitivity as it is now possible to perform sensor detection at different operating temperatures. This device is built to be capable of varying the voltage introduced to the sensors. In this experiment, the voltage was varied sinusoidally with a period of 128 s, with values ranging from 1.6 V to 4.8 V, and a total of 256 steps in each period. The device has a glass chamber where the samples are introduced, connected by 6 mm PVC tubing through PG7 nylon glands to a separate PP5 (food grade polypropylene) detection chamber containing the sensor assembly. Another tube returns to the sample chamber, completing a hermetically sealed circuit. To normalize the outputs of the MQ sensors, 50 kΩ trim potentiometers were used as load resistors, with potentiometer values adjusted until sensor channels gave a voltage difference smaller than 100 mV, at least half an hour after introducing the sample. In addition to balancing the impedance characteristics that each sensor has, this also provides a degree of normalization to counteract the variability that occurs in sensor manufacturing.

A calibration process was needed in order to define the parameters for the experiment, which included: stabilization time (SBT) of the sensor previous to the start of the experiments, sensing time (ST) which represents the time that the sensors will be exposed to the sample, cleaning time (CT) which represents the time that passes between the extraction of a sample and the introduction of another, overall time (OVT = ST + CT) which is the time needed for the analysis of a whole sample, and the maximum number of analysis (MNA) which we have estimated to be: MNA = (300−SBT)/OVT. This last value is established for the overall time of the experiment to be less than five hours so that high drift processes do not strongly affect the results. Table 1 shows the mean values of the parameters used for each of the 5 experiments carried out.

Table 1. Values of the parameters used for each of the 5 experiments carried out.

Experiment Number	SBT	ST	CT	OVT	MNA	Number of Analysed Samples
1	16.67	8	30	38	5	4
2	26.67	8	30	38	5	3
3	18.33	8	30	38	5	3
4	23.33	8	30	38	5	7
5	28.45	8	30	38	5	5
Mean and satardard values	22.69 ± 4.58	-	-	-	-	4.4 ± 1.5

A total of 5 experiments were carried out on different dates. One experiment was on 14 April 2021, two experiments were on 19 April 2021, another experiment was on 20 April 2021, and the last experiment was on 23 May 2021. The dates were chosen to be able to compare the performance of the device when: a few days had passed, it was used more than once in one day, when used on consecutive days, and when left for long periods of time (one month) without being used. A total of 21 analyses were performed, both the order and the number of repetitions per sample were randomly chosen; it was assured that all 10 samples were analyzed at least once.

The experiments were carried out in a clean and disinfected environment. Aliquot samples of 1 mL of wine were taken from the bottles, placed in a 135 mL glass chamber, and heated at room temperature for 15 min. The bottles were rapidly closed and reintroduced in the refrigerator to avoid deterioration of the samples. The chambers were then introduced in the e-nose to do the analysis.

The device includes self-developed software capable of connecting the device to the computer to: realize both the data acquisition during the analysis, generate the Excel files containing the data of each of the experiments, and to perform the analysis of the raw data for feature extraction.

Figure 1 shows the response for sensors MQ2 and MQ135 to wine sample 43D along with the voltage signal introduced into the sensors. This graph includes the heating time from Point 1 to Point 2025, the analysis time from 2026 to 3129, and the cleaning time from 3130 until the end. It can be seen how the sensor responses are deformed following the introduction of the sample, and how they recover after its extraction. Similar graphs are observed across all samples.

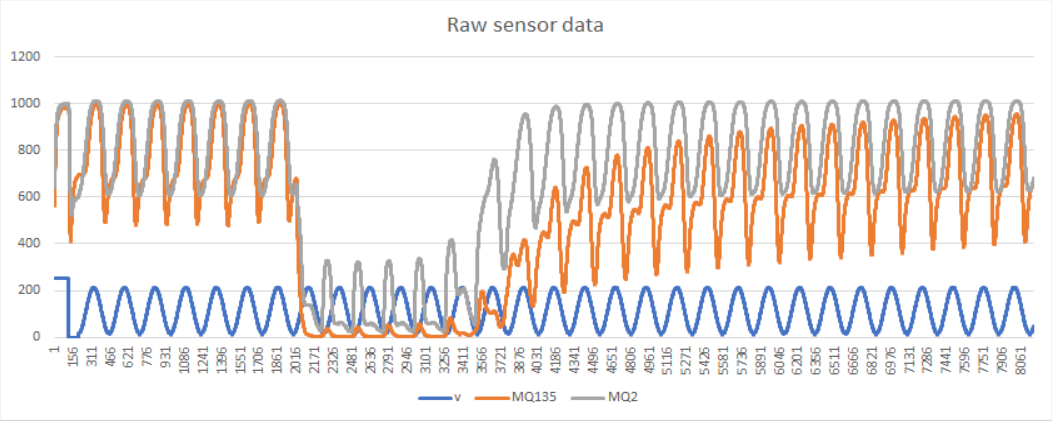

Figure 1. Response for sensors MQ2 and MQ135 to wine sample 43D along with the voltage signal introduced into the sensors.

To analyze the raw data, the first step was to perform a discrete Fourier transform (DFT) for each one of the cycles of the signal introduced. To do so, we used the software we developed. The signal introduced had a period of 128 s and the sampling frequency of the

device was set to 2 Hz. Thus, we obtained 256 values for each cycle which were then passed through the Matlab FFT function to obtain the cosine and sine coefficients, along with 5 harmonics including the DC component for each coefficient. This calculation was carried out for the response of each one of the sensors, obtaining 2 coefficients × 5 harmonics × 8 sensor = 80 values representing each cycle. A typical representation of the coefficients is shown in Figure 2, where each value in the plot corresponds to the calculation of the DFT for a single cycle of a specific sensor and harmonic. The names for each coefficient are constructed in the following way: sensor name (MQX) + coefficient (sine or cosine) + harmonic (1–5).

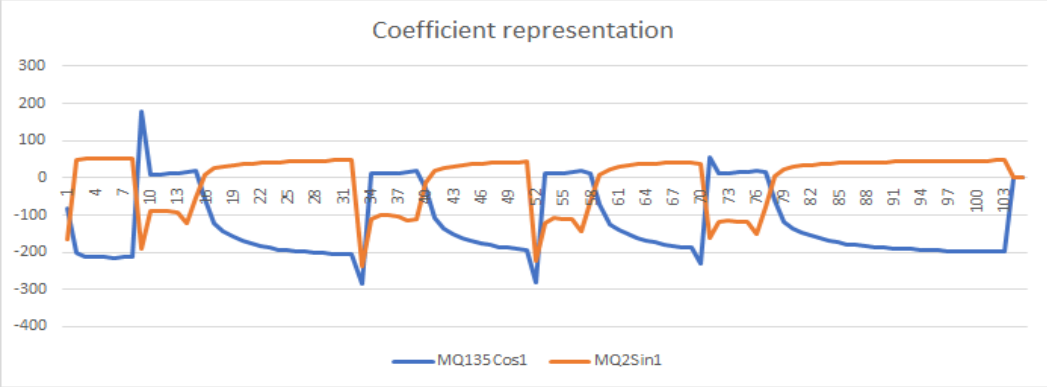

Figure 2. Typical representation of the coefficients of each sensor.

A PCA was carried out to observe and to cluster the samples attending to this data. The PCA and representation were executed using the MATLAB software version R2020a. The data were stored in a matrix that was then normalized and decomposed into principal components (PCs). To perform further analysis using PCA, we could only use one of the cycles; therefore, we had to select which cylce should be used. The hypothesis that initially occurred to us was that, during the analysis time, the sensors were saturated due to the strong smell of the wine and the high presence of ethanol, thus, we thought that the cycles in this phase would not carry relevant information. By observation of the raw data, we presented the idea that the start of the cleaning time would indeed carry much more information as the odor was slowly diluted when the sample was withdrawn. Therefore, we thought that this phase would be the one carrying most of the relevant information.

3. Results

3.1. Data Clustering

3.1.1. Chemical Clustering

Results of the chemical analysis are shown in Table 2.

Table 2. Results of the chemical analysis.

Sample	Ethanol	Volatile	Total Ac.	PH	Reductor	Malic Acid	CO_2	Anthocyane	TSAT
44D 111	14.15	0.54	5.52	3.75	1.87	−0.14	848.74	14.3	47.02
43D 112	13.97	0.5	5.48	3.69	2.15	−0.22	772.37	14.39	47.3
F1F 131	14.87	0.51	5.45	3.47	3.56	−0.09	403.42	14.34	45.37
200 D 031	15.68	0.36	5.88	3.57	1.64	0.2	569.9	14.07	40.14
203 D 032	15.15	0.45	5.57	3.61	1.72	0.03	682.5	14	38.56
9 F 033	15.16	0.46	6.44	3.27	1.25	0.2	360.8	14.08	42.41
P2 F 134	15.7	0.4	7.05	3.15	1.59	0.37	472.33	13.85	38.74
248 D 021	14.98	0.52	5.35	3.93	1.97	0.11	677.38	14.15	45.98
252 D 022	14.69	0.51	5.28	3.85	1.78	−0.07	696.93	14.23	43.11
11F 023	14.6	0.51	4.88	3.53	1.79	−0.12	387.77	14.03	34.07
42 F 041	15.39	0.39	5.89	3.42	1.98	0.05	541.47	14.13	44.13

PCA data are represented using a 2D plot in Figure 3. It uses the two first PCs and the samples are represented by a different color and shape according to the group they belong to. This clustering is observed with a chemical analysis, and it was compared to the results of the e-nose analysis.

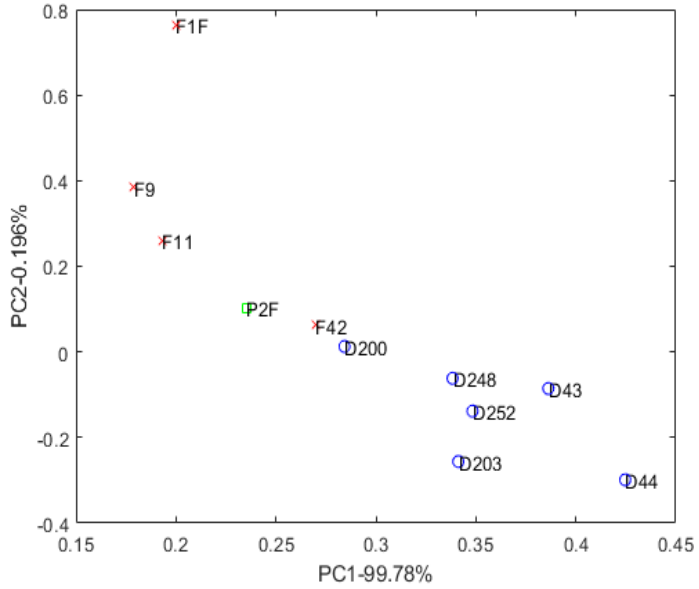

Figure 3. 2D PCA data.

It can be seen from Figure 3 that the samples can be classified into two separate groups, corresponding to the letter they carry in their names, namely Groups F and D. The sample P2F should be another group but data shows that it is very similar to Group F group.

3.1.2. Enose Data Clustering

With the data from the DFT analysis, a PCA was realized to observe the clustering that the e-nose could find between the samples. To do so, the 80 coefficients of a whole cycle were selected to represent each sample. A 21 × 80 matrix containing the 21 analyses was constructed to apply PCA using the MATLAB software version R2020a. To respond to the hypothesis of which cycle should be selected for the analysis, we generated the PCA of the fourth cycle after the introduction of the sample, which was thought to be saturated, and four more PCAs for each of the following cycles which are the cycles in the cleaning time phase. The results of the PCA for each cycle k are shown in Figures 4–8.

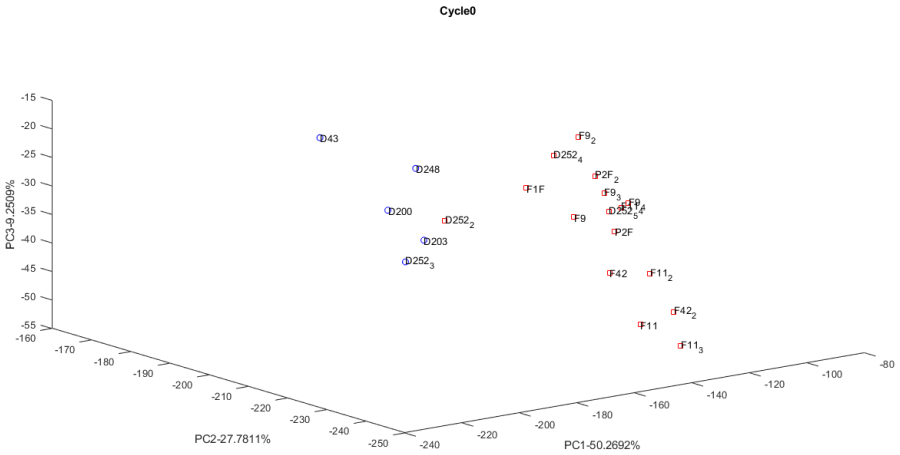

Figure 4. 3D PCA data. It is corresponding to Cycle 0, which will be the fourth after the introduction of the sample.

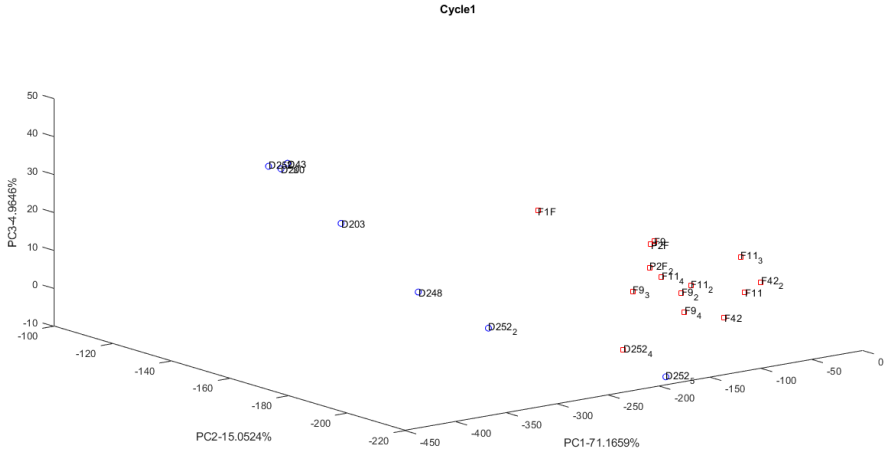

Figure 5. 3D PCA data. It is corresponding to Cycle 1.

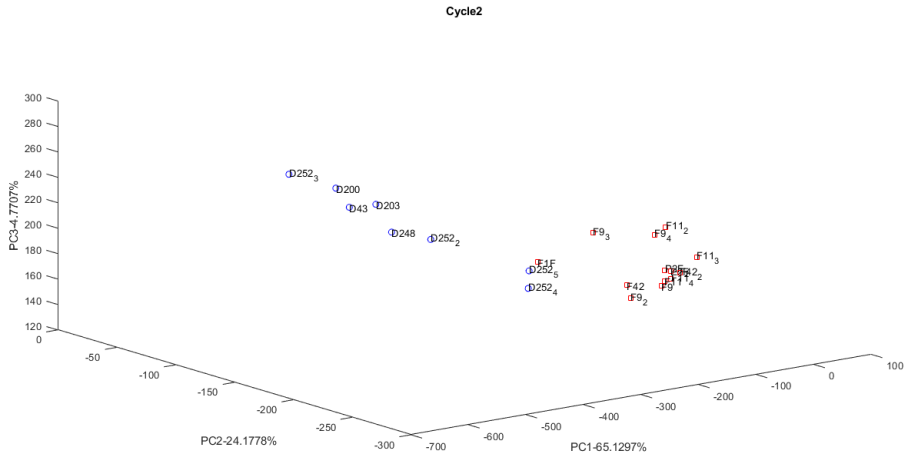

Figure 6. 3D PCA data. It is corresponding to Cycle 2. As cycles advance, both the sample grouping and separation between groups F and D continue to be more clear.

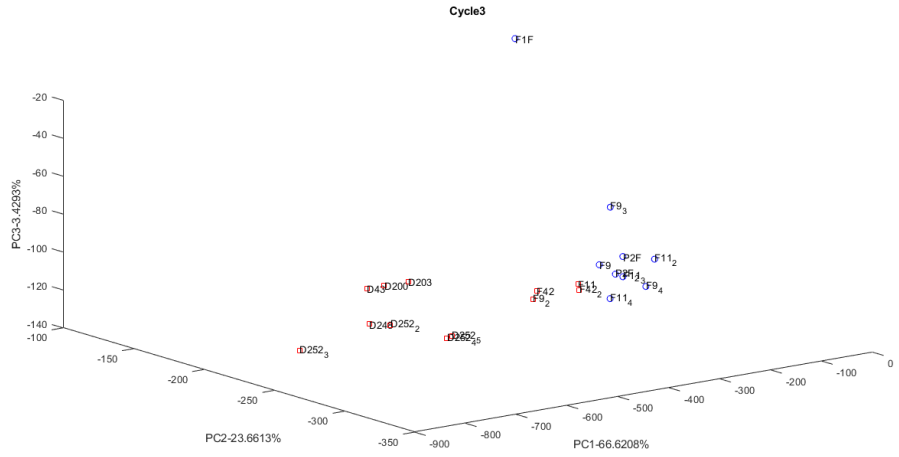

Figure 7. 3D PCA data. Sample grouping becomes less relevant as cycles pass.

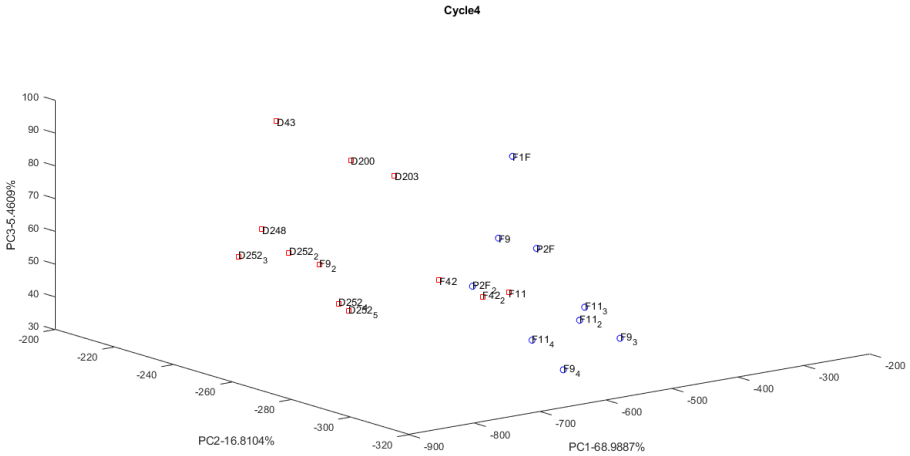

Figure 8. 3D PCA data. Later cycles are thought not to be very relevant.

Figure 4 corresponds to Cycle 0, which is the fourth cycle after the introduction of the sample. A k-means clustering algorithm was used to divide the dataset into two groups, a red group and a blue group. Considering that the analysis of each sample was carried out more than once, the samples were labeled with a subindex according to the number of the repetition they correspond to. This cycle does not group samples in an accurate manner, neither does a clear differentiation between groups F and D, as the accuracy obtained is 85.71%.

Continuing with Cycle 1 and following the same methodology as for the previous figure, Figure 5 was generated. Here, groups D and F are now more clearly separated but there is still some overlap between some repetitions of samples D252 and F42. Figure 3, with the chemical data, also shows that the overlap between groups is produced by sample F42. Sample grouping is still not very clear and the accuracy obtained is 90.48%.

Continuing with Cycle 2 and following the same methodology as for the previous figure, Figure 6 was generated. As cycles advance, both the sample grouping and separation between groups F and D continue to be clearer. This was selected as the best cycle, which obtained 100% accuracy in the classification. It can also be seen that there is more variation carried by the PCs.

Continuing with Cycle 3 and following the same methodology as for the previous figure, Figure 7 was generated. Even though groups overlap by using k-means clustering algorithms, giving a 89.50% accuracy, visual differentiation between the groups can be established. In addition, sample grouping becomes more relevant, but still not completely clear.

Later cycles, as in Figure 8, are thought not to be very relevant as the sample's odor would be too diluted by the passage of time, which could lead to air interference. The accuracy obtained is 76.19%.

4. Discussion

Figure 1, which represents the PCA grouping of the chemical data, shows that the samples can be separated into group F which is colored red and group D which is colored blue. It can also be observed that group D shows more similarities between its members, while group F is sparser. In general, the similarity between all the samples is high, which makes it difficult to observe significant differences between the samples.

Even though samples could not be classified independently, the e-nose classified the samples in a similar way that the chemical analysis would with a maximal accuracy of 100% obtained through the k-means clustering algorithm. It is to be to expected that if the samples are chemically similar, the e-nose device will not be able to differentiate them. For example if we look at group D, we can see that the data points are more concentrated in the

enose clustering, which can also be observed in the chemical clustering. Thus this will be seen as a group and we can not distinguish each sample from others inside the group.

Moreover, it was shown that the analysis time of the samples influences the results of the classification. Thus, showing the need for a calibration step in order to establish analysis times and the influence of the sample on the device. It was seen that, for wine samples, the use of cycles obtained minutes after the retrieval of the sample improved the classification results. However, the accuracy decreased again in later cycles, thus, a selection method for the correct cycle was needed.

These results show the potential that e-noses could have in the wine industry. Some of its purposes could be, for example, as a tool to detect fraudulent wine labeling. By having enough data of the different types of wine to be checked or to be classified attending to their quality, among other factors, we could perform this kind of controls.

This method introduces a reliable way for classifying wines. As compared with other methods such as sensory analysis, it brings an enormous improvement in objectivity, reduces costs, and opens the door to the possibility of analysis of a large number of samples due to its reduce time and cost expenses. In addition, by comparing it with chemical analysis or more sophisticated techniques such as gas chromatography, an e-nose analysis is much simpler, reduces the costs, does not require trained or qualified specialists, and can be use for monitoring processes.

In general, this method provides a tool to complement the use of other techniques as well as covering aspects that others cannot achieve.

Author Contributions: Conceptualization, A.R.-C. and A.C.C.; methodology, A.C.C. and M.J.O.; software, C.M.C. and A.C.C.; validation, A.C.C. and M.J.O.; formal analysis, A.C.C. and M.J.O.; investigation, A.C.C. and A.R.-C.; resources, A.R.-C.; data curation, A.C.C., C.P. and J.T.; writing—original draft preparation, A.C.C. and A.R.-C.; writing—review and editing, A.R.-C., A.C.C. and M.J.O.; visualization, A.C.C.; supervision, A.R.-C.; project administration, A.R.-C.; funding acquisition, A.R.-C. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Data Availability Statement: Not applicable.

Acknowledgments: All the authors would like to express our special thanks to the TeleNatura EBT, S.L. company who collaborated with us for this study.

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Rapp, A.; Linskens, H.F.; Jackson, J.F. Wine Aroma Substances from Gas Chromatographic Analysis. In *Wine Analysis*; Springer: Berlin, Germany, 1998; pp. 29–66.
2. Gilberto, V.; de Melo Pereira, C.; Soccol, R. *The Role of Alternative and Innovative Food Ingredients and Products in Consumer Wellness*; Elsevier Inc.: Amsterdam, The Netherlands, 2019; pp. 315–318.
3. Fischer, U. "Wine Aroma". In *Flavours and Fragrances*; Springer: Berlin/Heidelberg, Germany, 2007; pp. 241–267.
4. Gutierrez-Osuna, R. Pattern analysis for machine olfaction: A review. *IEEE Sens. J.* **2002**, *2*, 189–202. [CrossRef]
5. Gamboa, J.C.R.; da Silva, A.J.; de Andrade Lima, L.L.; Ferreira, T.A.E. Wine quality rapid detection using a compact electronic nose system: Application focused on spoilage thresholds by acetic acid. *LWT* **2019**, *108*, 377–384. [CrossRef]
6. Aleixandre, M.; Lozano, J.; Gutierrez, J.; Sayago, I.; Fernández, M.; Horrillo, M. Portable e-nose to classify different kinds of wine. *Sens. Actuators B Chem.* **2008**, *131*, 71–76. [CrossRef]
7. Gamboa, J.C.R.; Albarracin, E.S.A.; da Silva, A.; Ferreira, T.A.E. Electronic nose dataset for detection of wine spoilage thresholds. *Data Brief* **2019**, *25*, 104202. [CrossRef] [PubMed]
8. Berna, A.Z.; Trowell, S.; Cynkar, W.; Cozzolino, D. Comparison of Metal Oxide-Based Electronic Nose and Mass Spectrometry-Based Electronic Nose for the Prediction of Red Wine Spoilage. *J. Agric. Food Chem.* **2008**, *56*, 3238–3244. [CrossRef] [PubMed]
9. García, M.; Aleixandre, M.; Gutiérrez, J.; Horrillo, M.C. Electronic nose for wine discrimination. *Sens. Actuators B Chem.* **2006**, *113*, 911–916. [CrossRef]
10. Lozano, J.; Santos, J.; Horrillo, M. Classification of white wine aromas with an electronic nose. *Talanta* **2005**, *67*, 610–616. [CrossRef] [PubMed]