Research Article

International Journal of Distributed Sensor Networks

International Journal of Distributed Sensor Networks
2019, Vol. 15(10)

DOI: 10.1177/1550147719881601
journals.sagepub.com/home/dsn
SAGE

A data-driven methodology for the classification of different liquids in artificial taste recognition applications with a pulse voltammetric electronic tongue

Jersson X Leon-Medina[1,2], Leydi J Cardenas-Flechas[1,2] and Diego A Tibaduiza[3]

Abstract

Electronic tongue-type sensor arrays are devices used to determine the quality of substances and seek to imitate the main components of the human sense of taste. For this purpose, an electronic tongue-based system makes use of sensors, data acquisition systems, and a pattern recognition system. Particularly, in the latter, machine learning techniques are useful in data analysis and have been used to solve classification and regression problems. However, one of the problems in the use of this kind of device is associated with the development of reliable pattern recognition algorithms and robust data analysis. In this sense, this work introduces a taste recognition methodology, which is composed of several steps including unfolding data, data normalization, principal component analysis for compressing the data, and classification through different machine learning models. The proposed methodology is tested using data from an electronic tongue with 13 different liquid substances; this electronic tongue uses multifrequency large amplitude pulse signal voltammetry. Results show that the methodology is able to perform the classification accurately and the best results are obtained when it includes the use of K-nearest neighbor machine in terms of accuracy compared with other kinds of machine learning approaches. Besides, the comparison to evaluate the methodology is made with different classification performance measures that show the behavior of the process in a single number.

Keywords

Electronic tongue, sensor array, pattern recognition, data compression, principal component analysis, machine learning

Date received: 20 March 2019; accepted: 30 August 2019

Handling Editor: Florentino Fdez-Riverola

Introduction

A taste recognition system based on electronic tongues is used to identify, classify, and analyze qualitative and quantitatively mixtures of multiple components in liquids by applying the perspective of pattern recognition (PARC), that is, comparing the profiles of the mixture with a defined pattern. The analysis of the samples in liquid phase is carried out directly, while the samples in solid phase must be dissolved before being analyzed.[1] These sensors are characterized by low

[1]Departamento de Ingeniería Mecánica y Mecatrónica, Universidad Nacional de Colombia, Bogotá, Colombia
[2]Escuela de Ingeniería Electromecánica, Facultad seccional Duitama, Universidad Pedagógica y Tecnológica de Colombia, Duitama, Colombia
[3]Departamento de Ingeniería Eléctrica y Electrónica, Universidad Nacional de Colombia, Bogotá, Colombia

Corresponding author:
Diego A Tibaduiza, Departamento de Ingeniería Eléctrica y Electrónica, Universidad Nacional de Colombia, Cra 45 No. 26-85, Bogotá 111321, Colombia.
Email: dtibaduizab@unal.edu.co

selectivity because they generate information in a wide range of different substances in a liquid mixture. The signal from the chemical sensors is transformed into a data matrix. The identification and classification are carried out in the stage of multivariate data analysis, which can be performed through statistical techniques using machine learning algorithms.[2]

Through taste, it is possible to ensure the constant flavor of a product in a production process, that is, to know if the product is the same, if it is in good condition, or if any modification has occurred. For example, depending on the taste of the product, it can be determined if it is a good crop or not, discriminate between several types, detect if the product has defects, among other aspects.[3] Currently, this process is developed by a panel of highly trained people who make use of their sense of taste to certify the quality of the process.

In the food industry, the use of measuring instruments is a very important task in the development of new products, since it allows to implement or automate processes such as quality control, life control, degradation monitoring during transportation, highly perishable food monitoring, among others. The control of food products is normally carried out using physicochemical measurements, such as pH value, color, concentration of chemical substances, or biomolecules.[4] This is due to the lack of reliable instruments to evaluate the smell and taste of the products and the practical problems associated with the use of highly trained people, forming sensory panels for the continuous control of the aroma or flavor. Gas chromatography is another method used in the development of new products, but it is time-consuming and expensive, which in many cases discourages companies from using it.[5]

Data analysis in an electronic tongue can be performed from the point of view of chemometrics, which is a discipline of chemistry that uses mathematics, statistics, and formal logic to provide relevant information by analyzing chemical data.[6] Regarding the analysis by PARC, the approach is oriented to the use of data and the definition of a pattern to compare and thus obtain relevant information about the monitored process. Currently, this approach along with the use of chemometric techniques are used for analyzing data from electronic tongues. These techniques are schematized in Figure 1. According to Oliveri et al.,[7] the three possible results obtained by analyzing a data set are as follows: (1) recognition of the presence of structures (groupings, correlation) between the objects and/or variables studied (exploratory, unsupervised analysis); (2) development of mathematical models for the prediction of qualitative responses (classification and analysis of class models, supervised); (3) development of mathematical models for the prediction of quantitative responses (regression analysis, supervised).

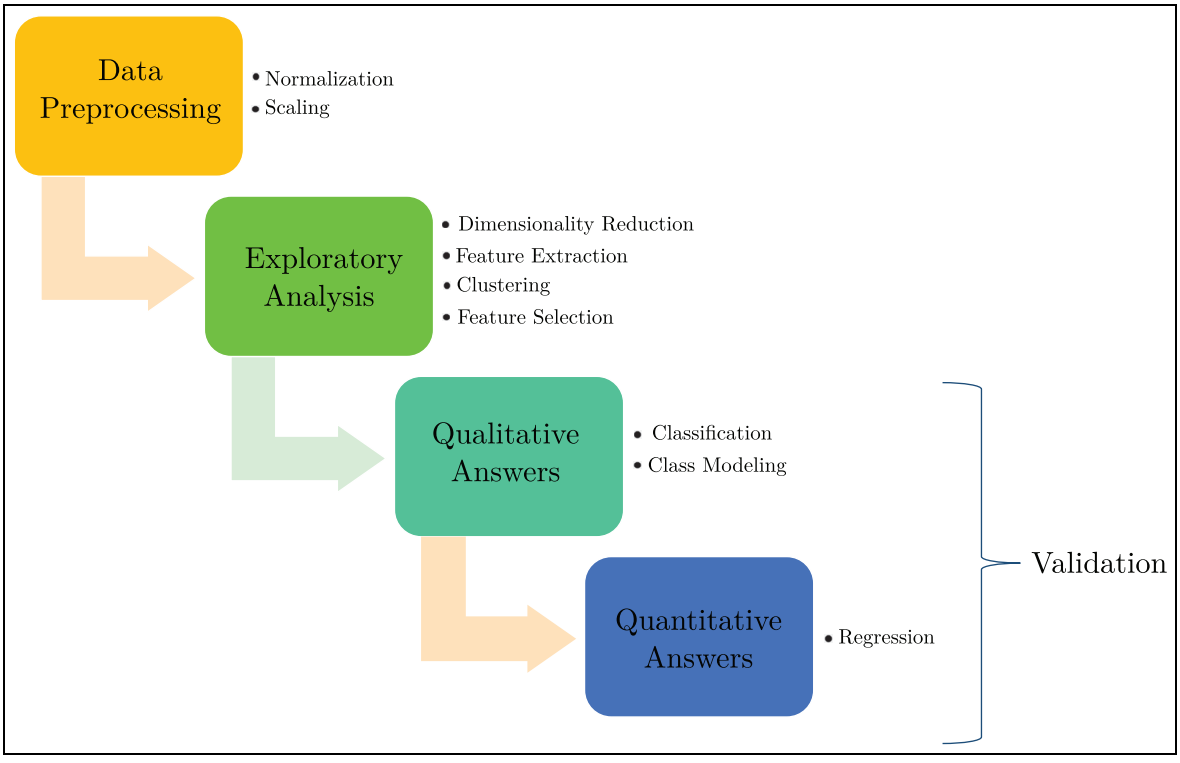

Figure 1. Main steps of the pattern recognition process.

The study by Śliwińska et al.[2] shows an exhaustive compilation of more than 75 publications related to electronic tongues, where the application, the type of analyte, the research object, the type of sensor, and the method for data analysis are distinguished. This work allows us to conclude that the principal component analysis (PCA) is one of the most used techniques by the scientific community to analyze the data coming from nonspecific sensor arrays ("electronic tongue") for chemical analysis of liquids.

In 2010, Oliveri et al.[7] performed a study using the SciFinder Scholar database, which indicated that the majority of signal processing works of electronic tongue-type sensors mostly uses the PCA in the feature extraction stage; artificial neural networks (ANNs) for classification stage; and partial least squares (PLS) regression for the regression stage. This can also be evidenced in the study by Del Valle[8] in 2017. Therefore, there is an opportunity to explore different ways to use these well-known strategies or new possibilities using different techniques for multivariate data analysis in this type of sensors.

Different researches related to classification problems in sensor arrays have been produced in recent years; among them are in 2014, the study by Zhang and Tian[9] combines the nonlinear method of the feature extracting kernel PCA with a new discriminant analysis (NDA) framework, which, through the construction of a Laplacian dispersion matrix between classes and a Laplacian dispersion matrix within the classes, solve an optimization problem. It makes samples between classes more separable and samples within classes are more compactable, thus improving the process of reducing the dimensionality of the data to in turn enhance the results of the classification algorithm. As a main result, the new kernel PCA plus NDA (KNDA) method achieved the same accuracy of the kernel support vector machine (KSVM) method with 95.06%; however, when comparing the running time of KNDA algorithms, it was 18 times faster than KSVM.

In 2018, the study by Zhang et al.[10] shows an electronic tongue data processing methodology that comprises the following stages: it starts with filter and feature selection via sliding window–based smooth filter, later a stage of feature extraction with the local discriminant preservation projection (LDPP) algorithm to finally compare different classifier algorithms like support vector machines (SVMs), extreme learning machines (ELMs), and kernelized extreme learning machine (KELM). The process was validated using the fivefold cross-validation technique, achieving the best average accuracy of 98.22% for the KELM algorithm. Besides, computational time studies were carried out comparing KELM, ELM, and SVM classification algorithms concluding that the KELM algorithm was the fastest of all.

The book by Zhang et al.[11] becomes the first one that deals with the problems of electronic noses from the point of view of algorithms and divides them into four main challenges: (1) the algorithmic challenge in PARC and machine learning including the odor recognition algorithms and concentration estimation algorithms, (2) the sensor drift problem, (3) the disturbance issue, and (4) the discreteness issue. The solutions to these problems make possible even more the real application of the sensor arrays at the industrial level, making them more robust and improving their accuracy.

As a contribution, this work presents a taste recognition methodology that uses data from a sensor array to classify different substances. The methodology is validated with a public data set from a multifrequency large amplitude pulse voltammetry (MLAPV) electronic tongue. The remainder of this article is organized as follows. Section "Theoretical background" briefly describes the most relevant concepts of the methods used in the proposed methodology. Section "Artificial taste recognition methodology" is devoted to present the different stages that compose the developed methodology. Section "Data set for validation" introduces the MLAPV electronic tongue data set used to validate the methodology. Then, section "Experimental results and discussion" identifies the effectiveness of the proposed methodology. Finally, the "Conclusion" section outlines the main conclusions of this article.

Theoretical background

This section includes information about the methods that will be used in the taste recognition methodology. It is recommended to review the cited works in this section in order to expand the information about each method.

Electronic tongue-type sensor arrays

According to the International Union of Pure and Applied Chemistry (IUPAC) definition,[12] an Electronic Tongue is "a multisensor system, which consists of a number of low-selective sensors and uses advanced mathematical procedures for signal processing based on PARC and/or multivariate data analysis (ANNs, PCA, etc.)." The main parts of an electronic tongue system are illustrated in Figure 2; in general, first an array of sensors is located at an electrochemical cell to produce the electronic tongue sensor which is the element introduced in the liquid substance, and then the electrical signals produced by these kinds of sensors are acquired by a data acquisition module that control the electrochemical technique to send the results signals toward a computer or a processor-based system that performs the signal processing and PARC stage in order to produce the analysis of the data.

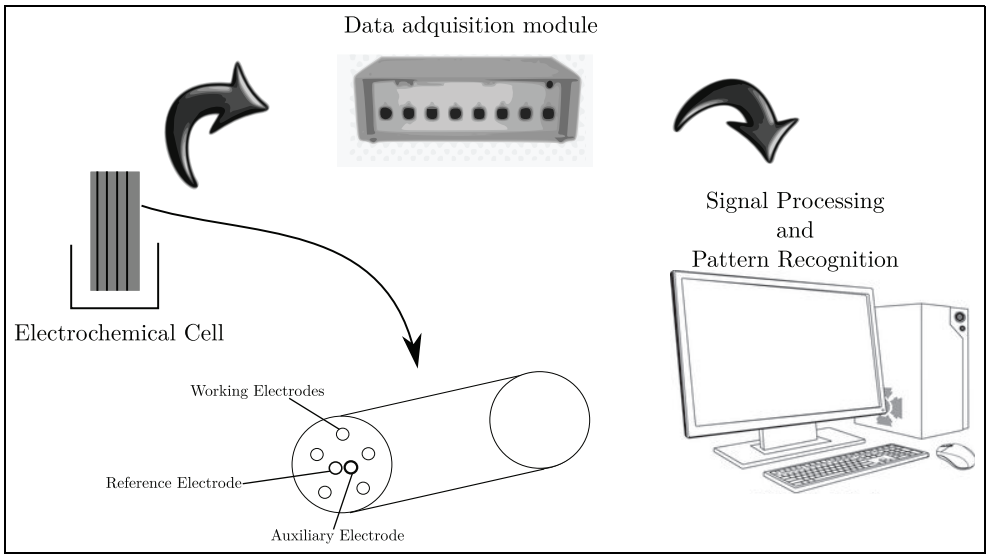

Figure 2. Electronic tongue system.

There are mainly two types of electronic tongue systems which are distinguished between them by the nature of the electrochemical sensors. First, systems are potentiometric when an ion-selective electrode (ISE) array is used. Second, systems are voltammetric when a number of metal electrodes of different nature or a number of modified electrodes are used. When both types of sensors compose an electronic tongue, it is called a hybrid electronic tongue.[8] For further information about potentiometric electronic tongues, the authors recommend the detailed review of Bratov et al.,[13] as well as the work by Wei et al.,[14] for voltammetric electronic tongues. The type of electrochemical sensor for this work was selected according to the substance that is to be detected. In this case, different substances are classified using a matrix of non-selective sensors, composed of noble metal electrodes, to use voltammetric techniques

MLAPV

One of the first voltammetric techniques used to analyze signals coming from electronic tongue-type sensor arrays is the large amplitude pulse voltammetry (LAPV). Pulse voltammetry is of special interest due to its advantage of greater sensitivity and resolution.[15] Subsequently, the MLAPV technique was used, which is a pulse voltammetry composed of a series of individual waveforms of LAPV with different step lengths.[16] Each cycle of the potential pulse step starts from the transient state of reaction process of the last cycle step, so that extra information is obtained by the MLAPV.[14] The MLAPV family includes multifrequency hackle pulse voltammetry (MHPV), multifrequency rectangle pulse voltammetry (MRPV), and multifrequency staircase pulse voltammetry (MSPV). As an example, the MSPV is proved to be the best waveform for classifying different yogurts.[17]

In 2018, the study by Zhang et al.[10] showed an electronic tongue based on an MLAPV technique. This electronic tongue followed the electrodes setup of Tian et al.[16] In the e-tongue, five electrodes, made of gold, platinum, palladium, tungsten, and silver, were chosen as working electrodes, whose responses are illustrated in Figure 3. The materials (Au, Pt, Pd, W, Ag) have been used in previous electronic tongues.[16,17] The pillar platinum is used as the auxiliary electrode and the Ag/AgCl is used as the reference electrode. Their experiments indicate that the electrodes sensor array shows different patterns to different kinds of substances in the electrochemical cell, which preliminarily shows the feasibility of the e-tongue. Further details of the data set used to evaluate the current taste recognition methodology will be presented in the data set for validation section.

Preprocessing

An important step for multivariate data analysis is data preprocessing. It is used to reduce random sources of variation in the data set. Some of the principal preprocessing techniques are centering of the average and autoscale. However, the use of these techniques can generate results with different possible forms of interpretation. In this sense, regarding the use of chemometric tools, the sequential publications of some authors have evidenced the lack of study in the previous treatment of the data, that is, only one method is

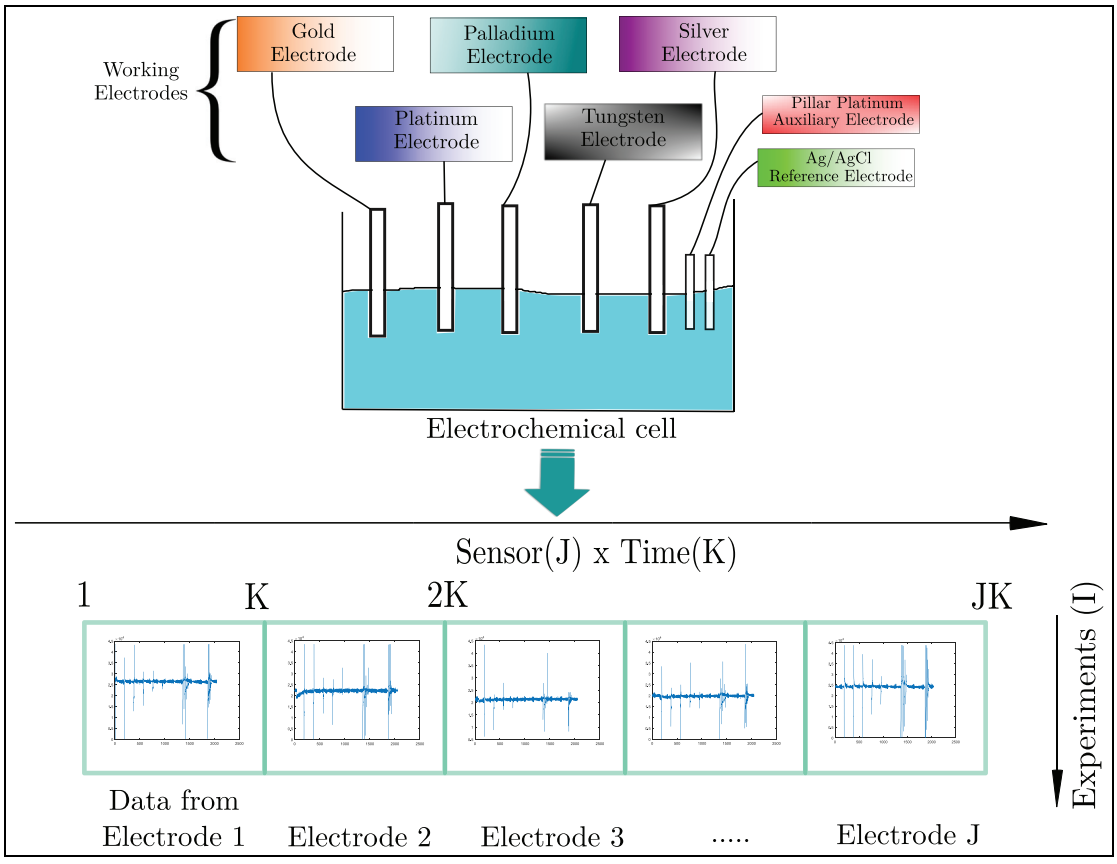

Figure 3. Data collection process: top: electrochemical cell containing the electrodes in the sensor array-type electronic tongue; bottom: three-dimensional matrix $I \times K \times J$ unfolded into a two-dimensional matrix $I \times (J \cdot K)$.

used and then results are discussed. To identify the best preprocessing method, it is advisable to try at least three different approaches (i.e. centering the average, autoscaling, smoothing, among others), compare results and choose the option that best suits the objectives of the work.[18]

Data preprocessing involves techniques that transform a data set to provide better input to the PARC engines.[19] In 2010, Palit et al.[20] studied the influence of several mathematical techniques for preprocessing multivariate data obtained from an array of electronic tongue-type sensors. To select the appropriate preprocessing technique, it is important to determine the effectiveness of these techniques in terms of the measure of the separability of the groups and the accuracy of the classification of the pattern in each case. The new data set is, therefore, a better substitute for the former in terms of data analysis. Palit et al. applied a cluster separability criterion for two analyses: one to determine the optimal level of wavelet decomposition and the other to find the best preprocessing method. The performance and separability of the different preprocessing techniques were compared using the class separability measure without compressing the data. The level of decomposition had been chosen based on two parameters: (1) percentage of accurate classification rate and (2) cluster separability criterion. They observed that selecting the appropriate preprocessing technique resulted in the improvement of the separability criterion. The preprocessing techniques that show the highest separability measure were combined to obtain a technique that produces an even better separability measurement.[21] In our work, group scaling normalization is used. This method will be further explained in this section.

Data compression

The treatment of signals of sensors arrays that uses voltammetric electronic tongues requires preprocessing techniques for data reduction due to the size of the information. Some of these methods are PCA, the discrete Fourier transform (DFT), or the wavelet transform (WT). They allow to obtain advantages over time of training, avoid redundancy, and obtain a model with better generalization capacity. Preprocessing provides an accurate interpretation of data from a data set.[22]

Voltammetric signals contain hundreds of measurements and usually have overlapping regions with nonstationary characteristics.[23] Therefore, the objective of feature extraction is to fully exploit the information obtained from each sensor. To achieve this objective, some representative characteristics must be extracted from the data obtained by each sensor of the matrix.[24] The artificial taste recognition methodology uses information from the MLAPV electrochemical technique, which relates the variables of current versus time; this technique was first shown by Tian et al.,[16] for be used in electronic tongue-type sensor arrays. Particularly, in Tian et al.'s[16] work, PCA, a kind of multivariate data analysis, was used for processing the data from the electronic tongue. six Chinese distilled spirits and seven Longjing teas were analyzed by the electronic tongue based on MLAPV and were successfully discriminated by the working electrodes at different frequency segments using the PCA algorithm.

Another electronic tongue study using pulse voltammetry was developed in 2013 by Wei et al.[17] The voltammetric electronic tongue based on MSPV with PCA presented the best separation ability in classifying the six categories of set yogurt. The separation ability of voltammetric electronic tongue based on the classification data set was presented by PCA. These two works showed that PCA has a successful behavior when it was applied in electronic tongues with pulse voltammetry.

In 2018, Zhang et al.[10] conducted a research, where they showed that due to the noise of the system and a variety of environmental conditions, the data acquired by an electronic tongue show inseparable patterns. To solve this phenomenon, from the point of view of the algorithm, they proposed an LDPP model. The LDPP algorithm of dimensionality reduction and feature extraction is a poorly studied subspace learning algorithm, which refers to local discrimination and the preservation of neighborhood structure. Other applications of subspace learning algorithms can be highlighted for transferring odor recognition model, such as cross-domain discriminative subspace learning (CDSL).[25]

PCA

The PCA analyzes a data set that represents observations described by several dependent variables, which are, in general, interrelated. Its objective is to extract important information from the data table and express this information as a set of new orthogonal variables called principal components (PCs).[26] The objectives of PCA are (1) to extract the most important information from the data table, (b) to compress the size of the data set keeping only important information, (c) to simplify the description of the data set, and (d) to analyze the structure of the observations and the variables.

It is quite common to use PCA as a preprocessing step in order to get a nice compact representation of a data set. Instead of the original many (J) variables, the data set can be expressed in terms of the few PCs.[27] Scores can be used to build a classification model, although there is a risk that the components that could help to improve the classification are excluded. It will often make more sense to use the number of components that provide the best classification results given by an appropriate metric if PCA is used in conjunction with a classification model such as K-nearest neighbor (KNN).[27]

Multiway principal component analysis (MPCA) is a conventional PCA extension for managing data in multidimensional arrays. A typical two-dimensional (2D) data matrix can be considered as a matrix with experiments and variables. For the application described in this article, it is necessary to extend this scheme to multiway matrices, since the electronic tongue needs different experimental tests, and several sensors measuring at different instants of time. MPCA is equivalent to performing a normal PCA for an unfolded version of the original multiway array.[28]

Organization of the recorded data

In an MLAPV electronic tongue, the measures are currents, and the measurements of only one sensor in a given number of moments of discretization are repeated several times (experimental tests), considering that each measurement is an individual experiment in the data set. The collected data are ordered as follows[28]

$$X = \begin{pmatrix} x_{11} & x_{12} & \cdots & x_{1K} \\ \vdots & \vdots & \ddots & \vdots \\ x_{i1} & x_{i2} & \cdots & x_{iK} \\ \vdots & \vdots & \ddots & \vdots \\ x_{I1} & x_{I2} & \cdots & x_{IK} \end{pmatrix} \in \mathcal{M}_{I\times K}(\mathbb{R}) \qquad (1)$$

This matrix $\mathbf{X} \in \mathcal{M}_{I\times K}(\mathbb{R})$, where $\mathcal{M}_{I\times K}(\mathbb{R})$ is the vector space of $I \times K$ matrices over $\mathbb{R}$, which contains information from $I \in \mathbb{N}$ experimental trials and $K \in \mathbb{N}$ time instants.

Considering that in the case of electronic tongue, an array of sensors is used, $J \in \mathbb{N}$ is the number of sensors (working electrodes) at each experiment, and there is a number J of the aforementioned matrix (equation (1)). Therefore, the whole set of the data collected in each phase can be organized in a three-dimensional (3D) matrix $I \times K \times J$, as shown in Figure 3.

Matrix unfolding

Westerhuis et al.[29] describe six different ways of unfolding this three-way matrix X into a 2D data matrix X. In

this work, we select unfolding as type E that produces a matrix where the rows are the signals from all sensors (from the same experiment) arranged one after the other, as observed in Figure 3.

$\mathcal{M}_{I\times(J\cdot K)}(\mathbb{R})$ is the vector space of $I\times(J\cdot K)$ matrices over $\mathbb{R}$ which includes data from J sensors at K discretization instants and I experimental trials.[30] The goal of PCA is to find out which dynamics are more important in the system, which are redundant, and which are just noise.[31,32] This goal is fundamentally achieved by defining a new coordinate space to re-express the original one, maximizing the variance, and minimizing the correlation between variables in the new space.[31] In other words, the objective is to find a linear orthogonal transformation $\mathbf{P}\in\mathcal{M}_{(J\cdot K)\times(J\cdot K)}(\mathbb{R})$ that will be used to convert the original data matrix X into the following form

$$\mathbf{T}=\mathbf{XP}\in\mathcal{M}_{I\times(J\cdot K)}(\mathbb{R}) \tag{2}$$

Matrix $\mathbf{P}$ is commonly named the loading matrix containing the PCs of the data set, while matrix $\mathbf{T}$ is named the score matrix which is the transformed or projected matrix onto the PC space.

Group scaling

The data of sensor array come from several sensors and these measures could have different magnitudes. It is necessary to highlight that the PCA is not invariant to scale; thus, the data must be rescaled applying a preprocessing stage.

For this preprocessing stage, group scaling normalization was selected. Comparable to autoscaling, group scaling, also called variable scaling, is performed according to standard deviations. Group scaling is frequently used when the data have several blocks of equal variables. Each block comprises variables in some given units of measure, but different blocks use different units. This type of situation occurs in the MPCA. Particularly, in this case, each block represents the information obtained by a single sensor.[33] In this type of unfolding matrices, group scaling presents better performance than other kinds of normalizations. The reason is that group scaling considers changes between sensors and does not process them independently.[30,34]

In group scaling, the variables are divided into a predefined number of blocks of equal size and each block is scaled by the large mean of their standard deviations. For each block, the standard deviations of each variable in a block are calculated and the average of these standard deviations is used to scale all the columns of the block. The same procedure is repeated for each block of variables.[33] In this sense, for $t=1,\ldots,J$ sensors(groups), and r,s the index in the unfolded 2D data matrix, we define[30,35]

$$\mu_s^t=\frac{1}{I}\sum_{r=1}^{I}x_{rs}^t,\ s=1,\ldots,K \tag{3}$$

$$\mu^t=\frac{1}{IK}\sum_{r=1}^{I}\sum_{s=1}^{K}x_{rs}^t \tag{4}$$

$$\sigma^t=\sqrt{\frac{1}{IK}\sum_{r=1}^{I}\sum_{s=1}^{K}\left(x_{rs}^t-\mu^t\right)^2} \tag{5}$$

where μ_s^t is the mean of the measures placed at the same column, that is the mean of the I measures of sensor t in matrix $\mathbf{X}^t$ at time instants $(j-1)\,\Delta$ seconds; μ^t is the mean of all of the elements in matrix $\mathbf{X}^t$, that is, the mean of all of the measures of sensor t; and σ^t is the standard deviation of all of the measures of sensor t. Then, the elements x_{rs}^t of matrix $\mathbf{X}$ are scaled to define a new matrix $\breve{\mathbf{X}}$ as

$$\breve{x}_{rs}^t:=\frac{x_{rs}^t-\mu_s^t}{\sigma^t},\ r=1,\ldots,I,\ s=1,\ldots,J,\ t=1,\ldots,K \tag{6}$$

The scaled matrix $\breve{\mathbf{X}}$ is renamed again as $\mathbf{X}$. The columns of the loading matrix $\mathbf{P}\in\mathcal{M}_{(J\cdot K)\times(J\cdot K)}(\mathbb{R})$ are the eigenvectors of the covariance matrix $\mathbf{C_X}$. They are also defined as the PCs organized according to its associated eigenvalue in descending order. In this way, the subspaces in PCA are defined by the eigenvectors and eigenvalues of the covariance matrix as follows

$$\boldsymbol{C_X P}=\mathbf{P\Lambda} \tag{7}$$

where

$$\mathbf{C}_X=\frac{\mathbf{X^T X}}{J\cdot K-1}\in\mathcal{M}_{(J\cdot K)\times(J\cdot K)}(\mathbb{R}) \tag{8}$$

and the diagonal terms of matrix $\mathbf{\Lambda}\in\mathcal{M}_{(J\cdot K)\times(J\cdot K)}(\mathbb{R})$ are the eigenvalues λ_i, $i=1,2,\ldots,J\cdot K$. PCA suppresses the redundancies transforming the original data through the projection definded by matrix $\mathbf{P}$ in equation (7). In addition, it reduces the dimensionality of the data set, which is achieved by selecting only a limited number $\ell<J\cdot K$ of PCs related to the ℓ highest eigenvalues. In this manner, given the reduced matrix

$$\hat{\mathbf{P}}=(p_1|p_2|\cdots|p_\ell)\in\mathcal{M}_{(J\cdot K)\times\ell}(\mathbb{R}) \tag{9}$$

matrix $\hat{\mathbf{T}}$ is defined as

$$\hat{\mathbf{T}}=\mathbf{X}\hat{\mathbf{P}}\in\mathcal{M}_{I\times\ell}(\mathbb{R}) \tag{10}$$

Classification approaches

The dimensionality of the pattern for an electrochemical sensor array is considerably small, usually ranging from 3 to 16, which reduces the calculation load in the

classification algorithm.[36] In addition, for the implementation of a portable sensor system, the algorithm must be transferable to an electronic system. Therefore, many of the accepted practices for the recognition of patterns in general may not be relevant for the recognition of patterns of chemical sensor matrices. From the types of environments and situations in which the chemical sensor arrays are expected to work, Shaffer et al.[36] highlight six qualities that the ideal PARC algorithm must have in an array of chemical sensors: (1) high precision, (2) quick, (3) simple to train, (4) low memory requirements, (5) robust to the outliers, and (6) produce a measure of uncertainty.

In 2018, Wesoły and Ciosek-Skibińska[37] affirm that as far as they know, such a profound study has not been presented previously on the chemometric treatment of sensor responses for taste discrimination using an electronic tongue. For the most commonly applied data extraction, the best classification results were obtained using SVM-DA and PCA + back propagation neural network (BPNN), while principal component regression (PCR) and KNN provided the worst results. Their paper concludes that it is not feasible to develop a universal methodology for the analysis of data from an electronic tongue, but some general recommendations can be provided based on the results obtained.

In this work, six machine learning algorithms will be used to perform the classification process: KNN, linear discriminant analysis (LDA), classification trees, Naive Bayes (NB), random forest (RF), and SVMs. The KNN algorithm[38] or KNN is a non-parametric supervised classification system based on neighborhood criteria. In particular, KNN is based on the idea that the new samples will be classified into the class of the closest number of closest neighbors of the closest training set.[39]

LDA[40] fits a parametric model to the training data and interpolates to classify test data. It assumes that different classes generate data based on different Gaussian distributions. To train a classifier, the fitting function estimates the parameters of a Gaussian distribution for each class. To predict the classes of new data, the trained classifier finds the class with the smallest misclassification cost.[41] LDA is based on the estimation of multivariate probability density functions, which are entirely described by a minimum number of parameters (means, variances, and covariances), as in the case of the well-known univariate normal distribution.[7]

Classification and regression tree (CART)[42] is a non-parametric supervised classification method that predicts responses to data. To predict a response, it follows the decisions in the tree from the root node (beginning) to a leaf node that contains the response. Classification trees give nominal responses, such as "true" or "false."[41] Trees are grown to a maximal size without the use of a stopping rule and then pruned back (basically split by split) to the root via cost-complexity pruning. The next split to be pruned is the one that contributes the least to the overall performance of the tree in training data. The CART mechanism is intended to produce not one, but a sequence of nested pruned trees, all of which are candidate optimal trees. The right sized tree is identified by evaluating the predictive performance of every tree in the pruning sequence.[39]

NB is a parametric classification approach designed for using when predictors are independent of one another within each class. It appears to work well in practice even when that independence assumption is not valid. It classifies data in two steps. First, the training step using the training data, that is, the method estimates the parameters of a probability distribution, assuming predictors are conditionally independent given the class. Second, the prediction step, that is, for any unseen test data, the method computes the posterior probability of that sample belonging to each class. The method then classifies the test data according to the largest posterior probability.[41]

RF is a PARC method and it has been proposed by Breiman.[43] RF is a method based on a kind of learning strategy "ensemble learning" with generating many classifiers and aggregating their results. RF operates by constructing a multitude of decision trees at training time and outputting the class that is the mode of the classes of the individual trees.[44]

SVMs[45] are supervised learning models with associated learning algorithms that analyze data used for classification and regression analysis. SVM offers one of the most robust and accurate methods among all well-known algorithms.[39] In addition to performing linear classification, SVMs can achieve a nonlinear classification using the kernel trick, implicitly mapping their inputs into high dimensional feature spaces. Multiclass SVM purposes to assign labels to instances using SVMs, where the labels are drawn from a finite set of several elements. The main approach for doing so is to reduce the single multiclass problem into multiple binary classification problems.[46] Common approaches for such reduction include building binary classifiers that discriminate between one of the labels and the rest (one-versus-all) or between every pair of classes (one-versus-one).

The parameters used in each of the classifiers were those found by default in MATLAB. For KNN, the number of neighbors was set to one, converting it into fine KNN, also using a Euclidean distance. For LDA, a type of linear discriminant is used and normal distributions are used for NB.

Artificial taste recognition methodology

The artificial taste recognition methodology developed in this work consists of the reunion of different stages of signal processing and PARC; all the steps are based on the use of different well-known methods in the literature. As described before, the first step unfolds the data and defines a new 2D matrix. Consequently, in the second step, data will be normalized using the group scaling method. The third step is the dimensionality reduction process executed by PCA, in which the size of the data is decreased to a new matrix $\hat{\mathbf{T}} = \mathbf{X}\hat{\mathbf{P}} \in \mathcal{M}_{I\times\ell}(\mathbb{R})$. Next, the fourth step is done by the classification algorithm. For the present methodology, we use the KNN classification approach as the best result, comparing it with NB, LDA, classification trees, RF, and SVMs. The classification models discriminate between the 13 different classes of the data set. This process is validated using stratified fivefold cross-validation to obtain a confusion matrix. The confusion matrix will be analyzed by different performance measures to determine the quality of the classification using a set of unique numbers. The taste recognition methodology is summarized in Figure 4.

Stratified fivefold cross-validation

Stratification seeks an intelligent grouping of the elements that constitute the population of a certain subset. When a data set is divided into homogeneous groups using stratification, an exchange of elements between these groups will occur in order to make a heterogeneous distribution. For a small group of samples and with a high number of classes, the use of stratification is indispensable, as with the current data set.

Breiman et al.[42] propose the K-fold cross-validation method as fold or iterations, where the original sample set is divided into K subsets. Since data are often scarce, this is usually not possible. For this reason, this variant of the cross-validation uses part of the available data to adjust the model, and a different part to verify it. This technique is considered a reliable media that offers good complexity in real time. The model, by repeating the training process, produces a resulting percentage of success that is used as a measure of the goodness of the evaluated algorithm. The stratified fivefold cross-validation[47] is used to evaluate the classification results performed for the machine learning algorithms in the current methodology of taste recognition.

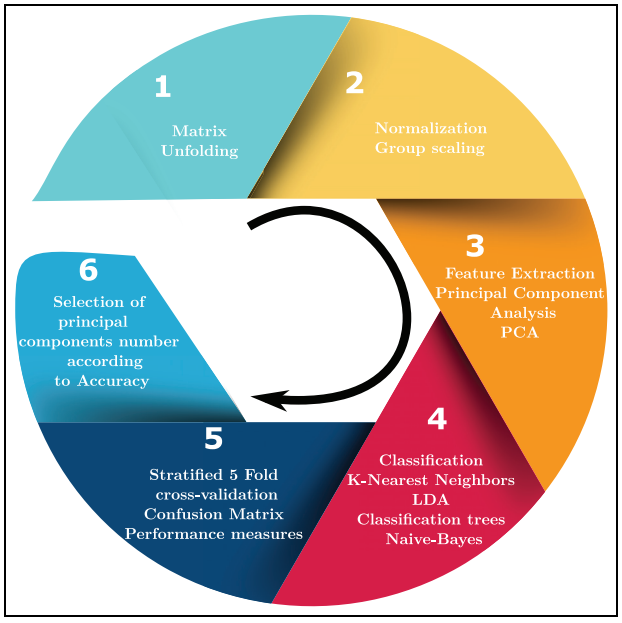

Figure 4. Main steps of the taste recognition methodology.

Performance measures

The overall classification performance of the taste recognition methodology is stated from different metrics represented by a single number. In the literature, among the different metrics to measure the goodness of classification, accuracy is one of the best knowns. However, after performing the K-fold cross-validation process, the taste recognition methodology results in a confusion matrix that encodes the number of both correct and incorrect predictions for each class.

From the analysis of the confusion matrix, three different measures are defined: sensitivity, specificity, and precision. These measures describe the classification results achieved in each class and it is incorrect to take them into account individually to evaluate in a general way the classification performance of a model. Each of these three measures provides an index that relates the true positive (TP), true negative (TN), false negative (FN), and false positive (FP) cases. The classification quality must be evaluated holistically grouping all these measures.[48]

The global classification indexes can be useful when the evaluation of the model must be represented in a single numerical value, as in the case of this work, to determine the optimal number of main components that yield the maximum accuracy in the classification. Global classification measures may behave differently when dealing with unbalanced data sets, whose classes contain a significantly different number of samples.[49] Recalling the properties of the data set that evaluates this methodology of taste recognition, there are 13 different classes whose number of samples varies from beer with 19 samples to salt with 6 samples. In this taste recognition methodology, some measures described in the study by Ballabio et al.,[48] will be used to evaluate the classification performance. The measures and a brief description of each one of them are included in

Table 1. Classification performance measures.

Acronym	Measure name
Accuracy	Accuracy measure
Kappa	Kappa coefficient
NER	Arithmetic mean of class sensitivities (non-error rate)
Pr	Arithmetic mean of class precisions
GmM	Macro-average of sensitivity and specificity geometric mean
V	Micro-average of geometric mean of class sensitivity and precision
VM	Macro-average of geometric mean of class sensitivity and precision
F	Micro-average of F measure (F)
FM	Macro-average of F measure (F)
JM	Macro-average of Youden J index
sInd	Similarity in ROC space
EMCC	Extended Matthew correlation coefficient
AUNU	Global measure based on the ROC approach as the average of single-class measures
AUNP	Area under ROC curve (AUC) with weights proportional to the number of samples experimentally belonging to each class
AU1U	Averaged AUC of each class against each other

ROC: receiver operating characteristic.

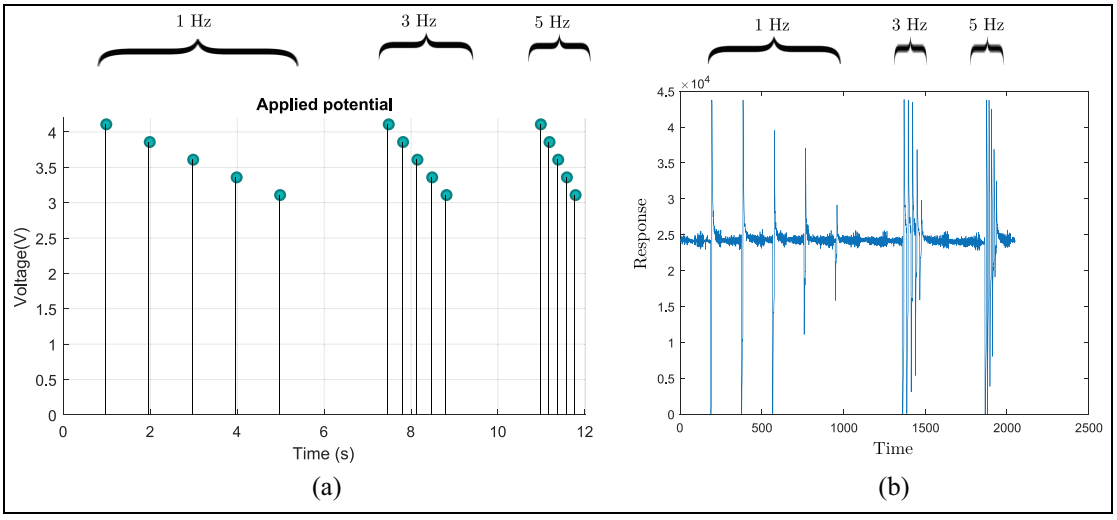

Figure 5. Multifrequency large amplitude pulse voltammetry: (a) applied potential and (b) responding current of one working electrode.

Table 1. For further information about each measure, see the work by Ballabio et al.[48]

Data set for validation

In the study by Zhang et al.,[10] an electronic tongue based on an array of five different sensors applies the technique of MLAPV. The pulse signal (excitation) comprises three individual frequencies: 1, 3, and 5 Hz, and five amplitudes of voltage: 4.10, 3.85, 3.60, 3.35, and 3.10 V for each frequency. After applying the described potential, each working electrode measures a responding current. The multiple frequency pulse signal and its response are shown in Figure 5.

The proposed methodology is validated using a public data set published in the study by Zhang et al.[10] The data set contains 114 measurements from five chemical sensors of E-Tongue based on MLAPV in a discrimination task of 13 liquid samples at various levels of concentrations. The data set contains samples of liquid substances, such as beer, red wine, white alcohol, black tea, Maofeng tea, pu-erh tea, Oolong tea, coffee, milk, cola, vinegar, medicine, and salt. The number of samples per liquid is provided in Table 2.

The system used for measuring the data set contains five electrode sensors, and the response of each sensor has 2050 sampling points. Each sample includes data points; thus, it has 114 instances and 10,250 attributes.

Table 2. Number of samples per liquid in the data set.

ID	Liquid type	Samples
1	Beer	19
2	Black tea	9
3	Coffee	9
4	Cola	6
5	Maofeng tea	9
6	Medicine	6
7	Milk	9
8	Oolong tea	9
9	Pu er tea	9
10	Redwine	8
11	Salt	6
12	Vinegar	9
13	Whitespirit	6

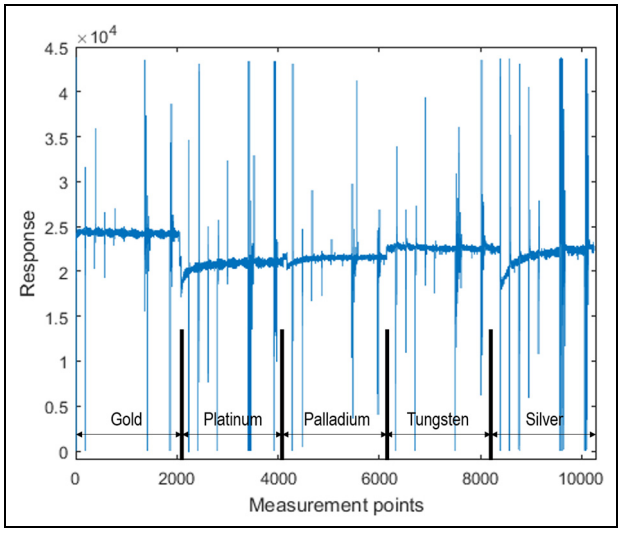

Figure 6. MLAPV signals of each electrode arranged one after another.

The electrode from the first to the fifth row in each sample denotes the gold, platinum, palladium, tungsten, and silver electrodes, respectively. Figure 6 shows the response of the five electrodes in 10,250 sampling times as an example of how the signals of each sensor are arranged one after another, as described in the unfolding section.

Experimental results and discussion

Figure 7 includes a resume of the steps followed by the proposed methodology. It starts with the data unfolding for each substance from different sensors. Afterward, unfolded data are normalized by group scaling and data reduction is performed using PCA to obtain the components and define the PCA modeling or pattern. As a result, it is possible to plot the first two components, as shown in Figure 8.

Figure 8 shows the PCA score plot of the first two PCs, where some classes overlap close to the coordinate (1,0), evidencing the need for using a machine learning approach for the classification. The next step of the taste recognition methodology is to organize a group of PCs to create a featured vector, which is used as input in the machine learning algorithms for classification. The resulting featured matrix is formed by the projections or scores of the original data into the PCA model created, as described in the PCA section and illustrated in Figure 7.

The study by Zhang et al.[10] uses LDPP, which is a subspace learning method, as feature extraction. In contrast, this work uses PCA. Another difference is that the current taste recognition methodology does not use the sliding window–based smooth filter for denoising, thus the feature selection stage is avoided.

The stratified K-fold cross-validation allows to evaluate the classification process. In the current methodology of taste recognition, stratified fivefold cross-validation is used. It evaluates five times the data set of 114 measurements. This is done by randomly partitioning the data set and using a subset to train the algorithm and the remaining data for testing. As cross-validation does not use all the data to build a model, it is a commonly used method to prevent overfitting during training. The training set is then used to train a supervised learning algorithm, and the testing set is used to evaluate its performance.[41] This process is repeated several times and the average cross-validation error is used as a performance indicator. Specifically, the data set of 114 number of observations is divided into 80% for train subset and the remaining 20% for the test subset. In particular, the sizes of the fivefold groups are shown in Table 3.

The fine tuning of the KNN algorithm is done by finding the best K-hyperparameter of the classification model. For this purpose, the parameter K was varied in a range of 1–15. Figure 9 indicates that the best value for K (number of neighbors) was 1, since it yields the highest value of accuracy when using the first eight PCs.

From previous works that use PCA,[30,35] the influence of the number of scores is well known on the analysis of data. The reason is that the variability of the data cannot be concentrated in the first components. Figure 10 shows the percentage of cumulative variance for the first eight components. To validate the influence of the number of components, this number will be varied for the classification. In general, the number of scores that has to be used depends on the cumulative contribution of variance that it is accounted for. More precisely, if the ith score is related to the eigenvector p_i, defined in equation (9), and the eigenvalue λ_i, the cumulative contribution rate of variance accounting for the first $\sigma \in \mathbb{N}$ scores is defined as[35]

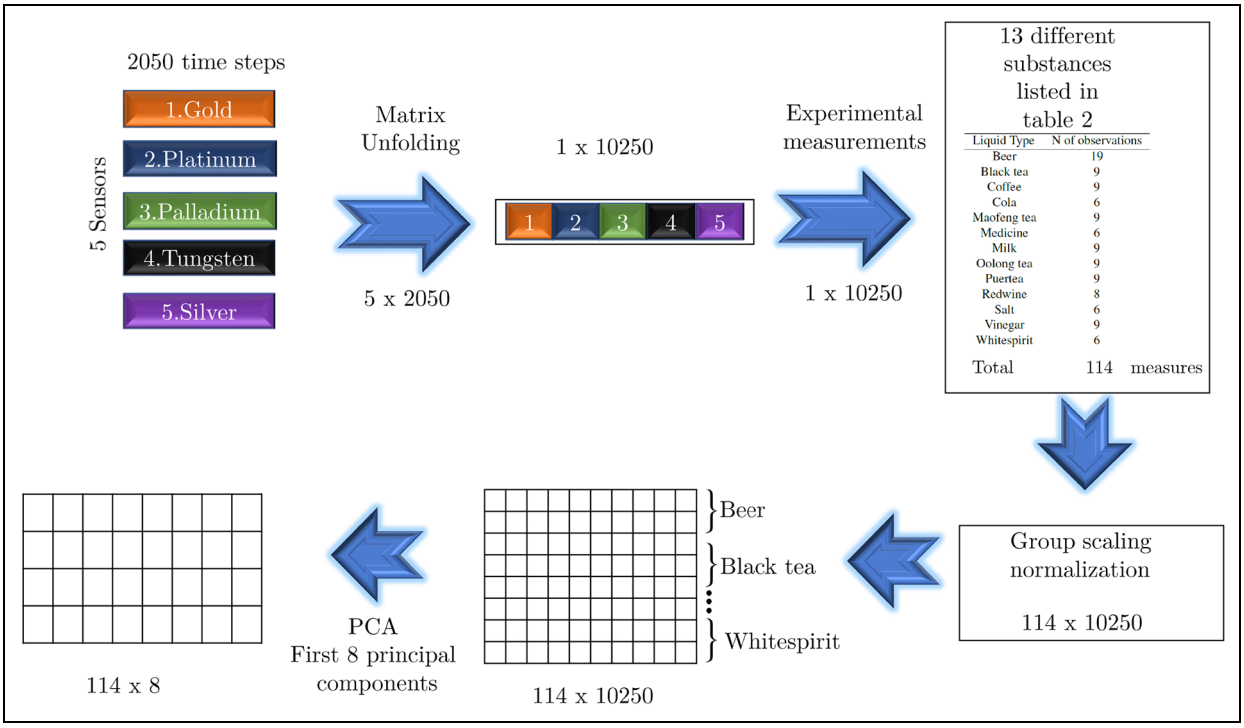

Figure 7. Experimental setup of the taste recognition methodology.

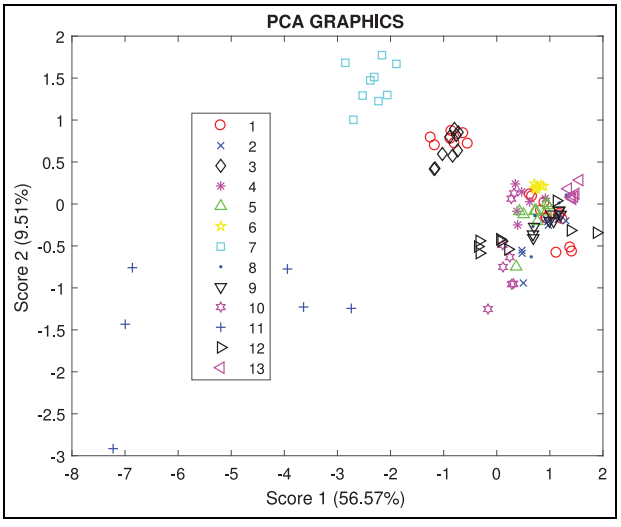

Figure 8. PCA with the first two PCs, the caption is ordered according to Table 2.

$$\frac{\sum_{i=1}^{\sigma} \lambda_i}{\sum_{i=1}^{\ell} \lambda_i} \tag{11}$$

where $\ell \in \mathbb{N}$ is the number of PCs. In this sense, the cumulative contribution of the first eight scores is depicted in Figure 10. It can be seen that the first two

Table 3. Sizes of the train ad test subsets in the stratified fivefold cross-validation.

Fold number	Train size	Test size
1	92	22
2	91	23
3	91	23
4	91	23
5	91	23

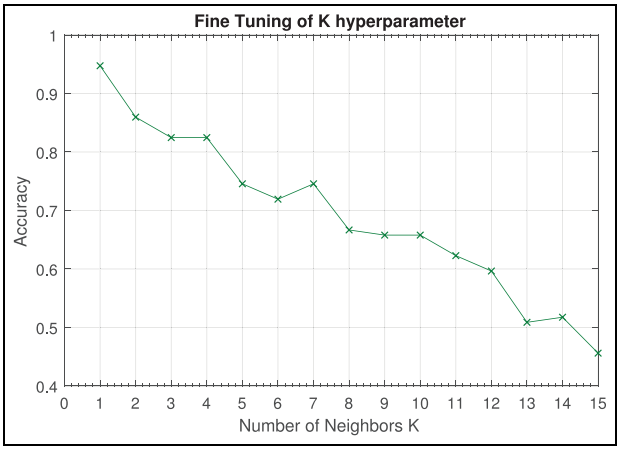

Figure 9. Fine tuning of K-hyperparameter of the KNN classification algorithm.

Table 4. Performance measures of the classification models considering two components.

Measure	KNN	NB	C trees	LDA	SVM	RF
Accuracy	0.6404	0.4912	0.4825	0.4035	0.3684	0.5044
Kappa	0.6062	0.4502	0.4325	0.3357	0.2728	0.4562
NER	0.6360	0.5694	0.5067	0.4057	0.3283	0.5677
Pr	0.6505	0.4903	0.5172	0.4415	0.2859	0.5460
GmM	0.7688	0.6550	0.6405	0.5199	0.3471	0.7015
V	0.6432	0.5284	0.5120	0.4232	0.3063	0.5567
VM	0.6408	0.5215	0.5069	0.4065	0.2926	0.5530
F	0.6432	0.5269	0.5119	0.4228	0.3056	0.5566
FM	0.6384	0.5137	0.5019	0.3921	0.2805	0.5492
JM	0.6057	0.5269	0.4628	0.3538	0.2711	0.5251
sInd	0.7405	0.6890	0.6487	0.5719	0.5035	0.6908
EMCC	0.6069	0.4611	0.4338	0.3428	0.3270	0.4572
AUNU	0.8029	0.7634	0.7314	0.6769	0.6355	0.7625
AUNP	0.8036	0.7232	0.7143	0.6626	0.6283	0.7231
AU1U	0.9482	0.7626	0.8325	0.6427	0.3402	0.9232

KNN: K-nearest neighbor; NB: Naive Bayes; LDA: linear discriminant analysis; SVM: support vector machine; RF: random forest; NER: non-error rate; EMCC: extended Matthew correlation coefficient.

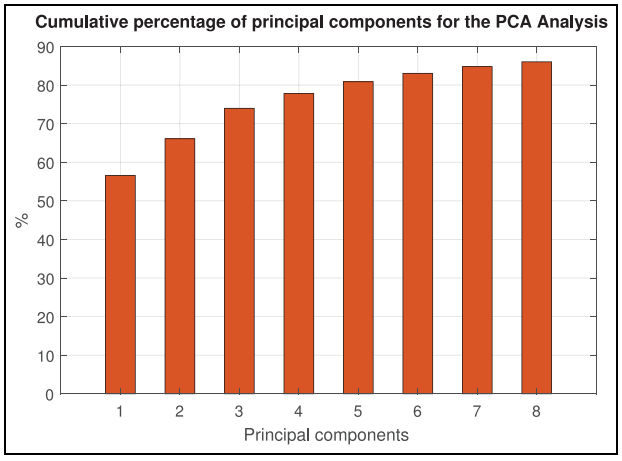

Figure 10. Percentage of cumulative variance of the first eight principal components.

PCs account for 66% of the variance, while the first three accounts for almost 74% and the first eight accounts for 86%. Better results a priori should be obtained if as many PCs as possible are used. Nevertheless, when nine components were used, it was observed that the behavior of the accuracy decreased. Therefore, Figure 14 shows the improvement of the accuracy taking into account the addition of a number of components to the classification model. The best accuracy of all was obtained using the first eight PCs. In the methodology, this procedure should be performed with each new data set because of the cumulative variance changes depending on the data and its characteristics in the data acquisition process.

Figure 11 shows the confusion matrix obtained after the classification process made by KNN, where only the first two scores are used in the training process. Data shown in Table 4 corresponds to the performance measures of the confusion matrix illustrated in Figure 11. The best results obtained were those of the KNN classifier compared with NB, C trees, LDA, SVM, and RF.

Figure 12 shows the confusion matrix obtained after the classification process made by KNN, where the first three scores are used in the training process. Data shown in Table 5 corresponds to the performance measures of the confusion matrix illustrated in Figure 12. The best results obtained were those of the KNN classifier compared with NB, C trees, LDA, SVM, and RF.

Figure 13 shows the confusion matrix obtained after the classification process made by KNN, where the first eight scores are used in the training process. Data shown in Table 6 correspond to the performance measures of the confusion matrix illustrated in Figure 13. The best results obtained were those of the KNN classifier compared with NB, C trees, LDA, SVM, and RF.

Accuracy generally provides higher values than the non-error rate (NER), because the most represented class is usually the most easily identifiable by classification algorithms and it typically results to be the best-modeled class, thus having the highest sensitivity.[48] In this case, after the application of the taste recognition methodology, the beer class with 19 samples is the most represented and that is why it is 100% predicted.

Results shown in Tables 4–6 express the improvement in accuracy obtained by the taste recognition methodology as PCs are added to the classification model. Particularly, for the KNN classification, accuracy obtained values of 0.6404 when taking into account the first two PCs, 0.7193 when considering the first three PCs and 0.9474 with the first eight PCs.

Actual Class	Predicted Class												
	1	2	3	4	5	6	7	8	9	10	11	12	13
1	10	1	4	0	2	0	0	1	1	0	0	0	0
2	2	5	0	0	0	0	0	0	1	1	0	0	0
3	4	0	5	0	0	0	0	0	0	0	0	0	0
4	2	0	0	2	1	0	0	0	0	1	0	0	0
5	0	0	0	1	6	0	0	0	1	1	0	0	0
6	0	0	0	0	0	6	0	0	0	0	0	0	0
7	0	0	0	0	0	0	9	0	0	0	0	0	0
8	2	1	0	0	1	0	0	2	2	0	0	0	1
9	0	1	0	0	0	0	0	2	6	0	0	0	0
10	0	0	0	0	1	0	0	0	0	6	0	1	0
11	0	0	0	0	0	0	0	0	0	0	6	0	0
12	0	1	0	0	0	0	0	0	1	1	0	6	0
13	0	0	0	0	0	0	0	1	0	0	0	0	5

Figure 11. Confusion matrix results of KNN classification considering two components.

Actual Class	Predicted Class												
	1	2	3	4	5	6	7	8	9	10	11	12	13
1	12	0	2	2	0	0	0	0	2	0	0	1	0
2	0	7	0	0	1	0	0	0	1	0	0	0	0
3	1	0	8	0	0	0	0	0	0	0	0	0	0
4	3	0	0	0	1	0	0	0	1	1	0	0	0
5	1	1	0	3	4	0	0	0	0	0	0	0	0
6	1	0	0	0	0	5	0	0	0	0	0	0	0
7	0	0	0	0	0	0	9	0	0	0	0	0	0
8	0	1	0	0	0	0	0	6	2	0	0	0	0
9	0	0	0	0	0	0	0	2	7	0	0	0	0
10	0	0	0	0	0	0	0	0	0	7	0	1	0
11	0	0	0	0	0	0	0	0	0	0	6	0	0
12	3	0	0	0	0	0	0	0	0	1	0	5	0
13	0	0	0	0	0	0	0	0	0	0	0	0	6

Figure 12. Confusion matrix results of KNN classification considering three components.

Table 5. Performance measures of the classification models considering three components.

Measure	KNN	NB	C Trees	LDA	SVM	RF
Accuracy	0.7193	0.5088	0.5965	0.4298	0.4649	0.5877
Kappa	0.6917	0.4684	0.5587	0.3693	0.3920	0.5471
NER	0.7225	0.5637	0.6431	0.4532	0.4514	0.6247
Pr	0.7067	0.5440	0.6244	0.4015	0.4700	0.6329
GmM	0.8016	0.6757	0.7450	0.5455	0.4992	0.7325
V	0.7146	0.5538	0.6337	0.4266	0.4606	0.6288
VM	0.7133	0.5420	0.6318	0.4199	0.4450	0.6265
F	0.7145	0.5537	0.6336	0.4258	0.4605	0.6288
FM	0.7119	0.5317	0.6299	0.4127	0.4310	0.6242
JM	0.6986	0.5226	0.6086	0.4038	0.4032	0.5892
sInd	0.8025	0.6865	0.7456	0.6050	0.5983	0.7327
EMCC	0.6930	0.4781	0.5601	0.3753	0.4304	0.5481
AUNU	0.8493	0.7613	0.8043	0.7019	0.7016	0.7946
AUNP	0.8443	0.7334	0.7760	0.6794	0.6870	0.7692
AU1U	0.8765	0.8308	0.8606	0.6416	0.5003	0.8551

KNN: K-nearest neighbor; NB: Naive Bayes; LDA: linear discriminant analysis; SVM: support vector machine; RF: random forest; NER: non-error rate; EMCC: extended Matthew correlation coefficient.

Actual Class	Predicted Class												
	1	2	3	4	5	6	7	8	9	10	11	12	13
1	19	0	0	0	0	0	0	0	0	0	0	0	0
2	0	9	0	0	0	0	0	0	0	0	0	0	0
3	0	0	9	0	0	0	0	0	0	0	0	0	0
4	0	0	0	6	0	0	0	0	0	0	0	0	0
5	0	1	0	0	7	0	0	1	0	0	0	0	0
6	0	0	0	0	0	6	0	0	0	0	0	0	0
7	0	0	0	0	0	0	9	0	0	0	0	0	0
8	0	1	0	0	2	0	0	6	0	0	0	0	0
9	0	0	0	0	0	0	0	0	9	0	0	0	0
10	0	0	0	0	0	0	0	0	0	8	0	0	0
11	0	0	0	0	0	0	0	0	0	0	6	0	0
12	1	0	0	0	0	0	0	0	0	0	0	8	0
13	0	0	0	0	0	0	0	0	0	0	0	0	6

Figure 13. Confusion matrix results of KNN classification considering eight components.

Table 6. Performance measures of the classification models considering eight components.

Measure	KNN	NB	C trees	LDA	SVM	RF
Accuracy	0.9474	0.7018	0.7018	0.7719	0.7632	0.7719
Kappa	0.9423	0.6736	0.6736	0.7510	0.7399	0.7502
NER	0.9532	0.7249	0.6928	0.7898	0.7684	0.7832
Pr	0.9596	0.7386	0.7184	0.7907	0.8276	0.7917
GmM	0.9735	0.8251	0.8058	0.8729	0.8605	0.8649
V	0.9564	0.7317	0.7055	0.7903	0.7975	0.7874
VM	0.9553	0.7263	0.6919	0.7870	0.7882	0.7821
F	0.9564	0.7316	0.7054	0.7903	0.7969	0.7874
FM	0.9543	0.7209	0.6796	0.7838	0.7794	0.7769
JM	0.9487	0.6996	0.6677	0.7707	0.7481	0.7639
sInd	0.9666	0.8033	0.7797	0.8500	0.8348	0.8432
EMCC	0.9427	0.6755	0.6757	0.7523	0.7419	0.7519
AUNU	0.9744	0.8498	0.8339	0.8854	0.8741	0.8819
AUNP	0.9709	0.8359	0.8369	0.8758	0.8681	0.8744
AU1U	0.9959	0.9666	0.9610	0.9790	0.9779	0.9733

KNN: K-nearest neighbor; NB: Naive Bayes; LDA: linear discriminant analysis; SVM: support vector machine; RF: random forest; NER: non-error rate; EMCC: extended Matthew correlation coefficient.

Figure 14 shows the increase in accuracy from the inclusion of the first two main components to the first eight, and from the first nine onward, as well as an oscillatory behavior of the accuracy obtained by the KNN classification model. The cross-validation loss (cvmdlloss) is the average loss of each cross-validation model when predicting data that are not used for training. When taking into account the first eight PCs, the KNN yielded an accuracy value of 0.9474, which refers to a cvmdlloss of 0.0526. The cross-validated classification accuracy resembles resubstitution accuracy. Therefore, it can be expected that the obtained KNN model misclassifies approximately 5% of new data, assuming that those data have about the same distribution as the training data.

Conclusion

This work described a taste recognition methodology that consists of several stages of signal processing and PARC methods. The methodology was tested with an MLAPV data set of 13 different substances. The data set presents measurements at different concentration levels for each substance in order to improve their discrimination. Results showed that the proposed methodology allows to classify all the substances with high accuracy.

The data obtained by the MLAPV technique in the electronic tongue-type sensor array were organized as an unfolding matrix. This configuration along with the use of group scaling for preprocessing allowed to have a comparable data scaled and normalized and

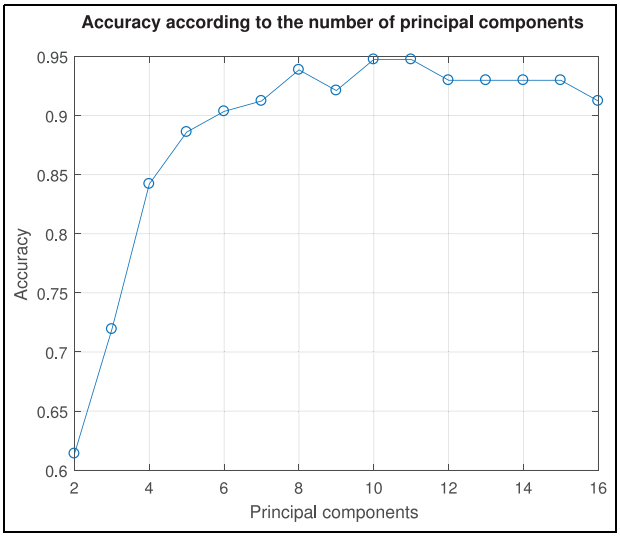

Figure 14. Accuracy according to the number of principal components used in the KNN classification model.

decreased the existence of outliers in the classification process.

PCA was used as a compression tool to reduce the dimensionality of the data. As expected, it allowed to identify relevant information on each measure using the PCs. In particular, a comparative study was made on the use of the number of components at the entrance to the classification process carried out by the machine learning algorithms. When considering the first eight PCs, the classification accuracy made by KNN reached 0.9474. It was found that results are associated with the number of components.

Different machine learning algorithms were compared: KNN, LDA classification trees, NB, RF, and SVM. The method of fine KNN was the one with the best behavior. The performance measures used to determine the classification behavior from results in the confusion matrices and provide important information in a single number that reflects the behavior. In total, 15 different performance measures were described, which allows to compare the classification performance from different points of view by having a set of performance indicators.

Future work can study the influence of changes in the preprocessing normalization stage, the use of different subspace learning algorithms for the feature extraction process, and the use of different classifiers, such as non-supervised, semi-supervised, and supervised algorithms with a comparison between their computational time and accuracy. Since PCA is a stationary technique that preserves the stationary information from the original data, more tests are required to evaluate the methodology in non-stationary conditions to evaluate the effectiveness of the use of PCA for data reduction.

Acknowledgements

The authors thank Dr Lei Zhang at the Chongqing University for releasing the data of the experiments carried out with his MLAPV electronic tongue to the public, which are published by Tian et al. and available in http://www.leizhang.tk/tempcode.html. The authors express their gratitude to the Administrative Department of Science, Technology and Innovation (Colciencias), "Convocatoria para la Formación de Capital Humano de Alto Nivel para el Departamento de Boyacá 2017," for sponsoring the research presented herein (grant 779). J.X.L.-M. is grateful to Colciencias and Gobernación de Boyacá for the PhD fellowship.

Declaration of conflicting interests

The author(s) declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

Funding

The author(s) disclosed receipt of the following financial support for the research, authorship, and/or publication of this article: The Administrative Department of Science, Technology and Innovation (Colciencias), "Convocatoria para la Formación de Capital Humano de Alto Nivel para el Departamento de Boyacá 2017," sponsored the research presented herein (grant 779).

ORCID iD

Diego A Tibaduiza https://orcid.org/0000-0002-4498-596X

References

1. Wang L, Niu Q, Hui Y, et al. Discrimination of rice with different pretreatment methods by using a voltammetric electronic tongue. *Sensors* 2015; 15(7): 17767–17785.
2. Śliwińska M, Wiśniewska P, Dymerski T, et al. Food analysis using artificial senses. *J Agric Food Chem* 2014; 62(7): 1423–1448.
3. Jiang H, Zhang M, Bhandari B, et al. Application of electronic tongue for fresh foods quality evaluation: a review. *Food Rev Int* 2018; 34(8): 746–769.
4. Costa C, Taiti C, Strano MC, et al. Multivariate approaches to electronic nose and PTR-TOF-MS technologies in agro-food products. In: ML Rodríguez Méndez (ed.) *Electronic noses and tongues in food science*. London: Academic Press, 2016, pp.73–82, http://www.sciencedirect.com/science/article/pii/B9780128002438000081
5. Ha D, Sun Q, Su K, et al. Recent achievements in electronic tongue and bioelectronic tongue as taste sensors. *Sensor Actuat B-Chem* 2015; 207: 1136–1146.
6. Esteban M, Ariño C and Díaz-Cruz JM. Chemometrics in electroanalytical chemistry. *Crit Rev Anal Chem* 2006; 36: 295–313.
7. Oliveri P, Casolino MC and Forina M. Chemometric brains for artificial tongues. In: S Taylor (ed.) *Advances*

in food and nutrition research, vol. 61. 1st ed. London: Elsevier Inc., 2010, pp.57–117.
8. Del Valle M. Materials for electronic tongues: smart sensor combining different materials and chemometric tools. In: Cesar Paixão TRL and Reddy SM (eds) *Materials for chemical sensing*. Berlin: Springer, 2017, pp.227–265.
9. Zhang L and Tian FC. A new kernel discriminant analysis framework for electronic nose recognition. *Anal Chim Acta* 2014; 816: 8–17.
10. Zhang L, Wang X, Huang GB, et al. Taste recognition in E-tongue using local discriminant preservation projection. *IEEE Trans Cybern* 2019; 49: 947–960, http://ieeexplore.ieee.org/document/8262615/
11. Zhang L, Tian F and Zhang D. *E-nose algorithms and challenges*. Singapore: Springer, 2018.
12. Vlasov Y, Legin A, Rudnitskaya A, et al. Nonspecific sensor arrays ("electronic tongue") for chemical analysis of liquids (IUPAC technical report). *Pure Appl Chem* 2005; 77(11): 1965–1983.
13. Bratov A, Abramova N and Ipatov A. Recent trends in potentiometric sensor arrays—a review. *Anal Chim Acta* 2010; 678: 149–159.
14. Wei Z, Yang Y, Wang J, et al. The measurement principles, working parameters and configurations of voltammetric electronic tongues and its applications for foodstuff analysis. *J Food Eng* 2018; 217: 75–92.
15. Winquist F, Wide P and Lundström I. An electronic tongue based on voltammetry. *Anal Chim Acta* 1997; 357: 21–31.
16. Tian SY, Deng SP and Chen ZX. Multifrequency large amplitude pulse voltammetry: a novel electrochemical method for electronic tongue. *Sensor Actuat B-Chem* 2007; 123(2): 1049–1056.
17. Wei Z, Wang J and Jin W. Evaluation of varieties of set yogurts and their physical properties using a voltammetric electronic tongue based on various potential waveforms. *Sensor Actuat B-Chem* 2013; 177: 684–694.
18. Nunes CA, Alvarenga VO, de Souza Sant'Ana A, et al. The use of statistical software in food science and technology: advantages, limitations and misuses. *Food Res Int* 2015; 75: 270–280.
19. Jurs PC, Bakken GA and McClelland HE. Computational methods for the analysis of chemical sensor array data from volatile analytes. *Chem Rev* 2000; 100(7): 2649–2678.
20. Palit M, Tudu B, Bhattacharyya N, et al. Comparison of multivariate preprocessing techniques as applied to electronic tongue based pattern classification for black tea. *Anal Chim Acta* 2010; 675(1): 8–15.
21. Duda RO, Hart PE and Stork DG. *Pattern classification*. New York: John Wiley & Sons, 2012.
22. Gutés A, Céspedes F, Cartas R, et al. Multivariate calibration model from overlapping voltammetric signals employing wavelet neural networks. *Chemometr Intell Lab Syst* 2006; 83(2): 169–179.
23. Gutiérrez JM, Moreno-Barón L, Céspedes F, et al. Resolution of heavy metal mixtures from highly overlapped ASV voltammograms employing a wavelet neural network. *Electroanalysis* 2009; 21(3–5): 445–451.
24. Haddi Z, Alami H, El Bari N, et al. Electronic nose and tongue combination for improved classification of Moroccan virgin olive oil profiles. *Food Res Int* 2013; 54(2): 1488–1498.
25. Zhang L, Liu Y and Deng P. Odor recognition in multiple E-nose systems with cross-domain discriminative subspace learning. *IEEE Trans Instrum Meas* 2017; 66(7): 1679–1692.
26. Abdi H and Williams LJ. Principal component analysis. *Wiley Interdiscip Rev Comput Stat* 2010; 2(4): 433–459.
27. Bro R and Smilde AK. Principal component analysis. *Anal Methods* 2014; 6(9): 2812–2831.
28. Pozo F, Vidal Y and Salgado Ó. Wind turbine condition monitoring strategy through multiway PCA and multivariate inference. *Energies* 2018; 11(4): 1–19.
29. Westerhuis JA, Kourti T and MacGregor JF. Comparing alternative approaches for multivariate statistical analysis of batch process data. *J Chemom* 1999; 13(3–4): 397–413, https://onlinelibrary.wiley.com/doi/abs/10.1002/(SICI)1099-128X(199905/08)13:3/4%3C397::AID-CEM559%3E3.0.CO;2-I
30. Anaya M, Tibaduiza DA and Pozo F. Detection and classification of structural changes using artificial immune systems and fuzzy clustering. *Int J Bio-Inspir Com* 2017; 9(1): 35–52, http://www.inderscience.com/link.php?id=81843
31. Tibaduiza D, Mujica L and Rodellar J. Damage classification in structural health monitoring using principal component analysis and self-organizing maps. *Struct Control Hlth* 2013; 20(10): 1303–1316.
32. Anaya Vejar M. *Design and validation of structural health monitoring system based on bio-inspired algorithms*. PhD Thesis, Universitat Politècnica de Catalunya, España, 2016, http://www.tesisenred.net/handle/10803/396370
33. Wise BM, Gallagher NB, Bro R, et al. *PLS toolbox 3.5 for use with Matlab*. Manson, WA: Eigenvector Research Inc., 2005.
34. Anaya M, Tibaduiza D and Pozo F. Data driven methodology based on artificial immune systems for damage detection. In: *7th European workshop on structural health monitoring*, Nantes, 8–11 July 2014. IFFSTTAR, Inria, Université de Nantes, https://hal.inria.fr/hal-01021192
35. Vitola J, Pozo F, Tibaduiza D, et al. A sensor data fusion system based on k-nearest neighbor pattern classification for structural health monitoring applications. *Sensors* 2017; 17: 417.
36. Shaffer RE, Rose-Pehrsson SL and McGill RA. A comparison study of chemical sensor array pattern recognition algorithms. *Anal Chim Acta* 1999; 384(3): 305–317.
37. Wesoły M and Ciosek-Skibińska P. Comparison of various data analysis techniques applied for the classification of pharmaceutical samples by electronic tongue. *Sensor Actuat B-Chem* 2018; 267: 570–580.
38. Dhanabal S and Chandramathi S. A review of various k-nearest neighbor query processing techniques. *Int J Comput Appl* 2011; 31(7): 14–22.
39. Wu X, Kumar V, Quinlan JR, et al. Top 10 algorithms in data mining. *Knowl Inf Syst* 2008; 14(1): 1–37.
40. Fisher RA. The use of multiple measurements in taxonomic problems. *Ann Eugen* 1936; 7(2): 179–188.
41. MathWorks. *Statistics and machine learning toolbox for Matlab*. Natick, MA: The MathWorks, Inc., 2015, https://es.mathworks.com/help/pdf_doc/stats/stats.pdf

42. Breiman L, Friedman JH and Olshen RA. *Classification and regression trees*. New York: Routledge, 1984.
43. Breiman L. Random forests. *Mach Learn* 2001; 45: 5–32.
44. Ho TK. The random subspace method for constructing decision forests. *IEEE Trans Pattern Anal Mach Intell* 1998; 20(8): 832–844.
45. Cortes C and Vapnik V. Support-vector networks. *Mach Learn* 1995; 20(3): 273–297.
46. Hsu CW and Lin CJ. A comparison of methods for multiclass support vector machines. *IEEE Trans Neural Netw* 2002; 13(2): 415–425.
47. Liu M, Wang M, Wang J, et al. Comparison of random forest, support vector machine and back propagation neural network for electronic tongue data classification: application to the recognition of orange beverage and Chinese vinegar. *Sensor Actuat B-Chem* 2013; 177: 970–980.
48. Ballabio D, Grisoni F and Todeschini R. Multivariate comparison of classification performance measures. *Chemometr Intell Lab Syst* 2018; 174: 33–44.
49. Varmuza K and Filzmoser P. *Introduction to multivariate statistical analysis in chemometrics*. Boca Raton, FL: CRC Press, 2009.