Socio-Economic Planning Sciences 95 (2024) 102021

Contents lists available at ScienceDirect

Socio-Economic Planning Sciences

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

Multi-class random forest model to classify wastewater treatment imbalanced data

Veronica Distefano [a,b], Monica Palma [a,c,*], Sandra De Iaco [a,c,d]

[a] Department of Economic Sciences, University of Salento, Complesso Ecotekne, Lecce, Italy
[b] European Centre for Living Technology (ECLT), Ca' Foscari University of Venice, Venice, Italy
[c] National Centre for HPC, Big Data and Quantum Computing, Bologna, Italy
[d] National Biodiversity Future Center, Palermo, Italy

ARTICLE INFO

Keywords:
Multi-classification
Data imbalance
Resampling approach
Treatment plant sections
Electronic nose
Machine learning

ABSTRACT

The odor emissions generated by treatment plants imply complex environmental and economic issues. The modern instrumental odor monitoring systems, based on an array of several sensors, continuously record the gaseous compounds. However they are characterized by poor selectivity, compromising the possibility to discriminate and identify the emission sources. In this paper, the ability of odor sensors to distinguish between the treatment plant sections generating the gaseous compounds is evaluated on the basis of the random forest classifier, and is also compared to the discriminant analysis performance. Taking into account that a multi-parametric system of sensors can be affected by the presence of a small sample size with imbalanced classes, several strategies for data balancing are proposed and analyzed. The findings show that the random forest classifier is characterized by a better capacity to distinguish the emissions sources with respect to the classical multiple discriminant analysis, in terms of all evaluation metrics. This is also confirmed for different resampling techniques, especially in the over-sampling case. The data concerning measurements from 10 sensors of multi-parametric systems of odor monitoring collected from a company specialized in environmental assistance are considered for this analysis.

1. Introduction

The accelerated urbanization and the lack of suitable sites have led to build treatment plants close to existing urban areas, raising health risks for human beings [1]. Indeed, odor emissions from industrial and environmental protection plants are often the cause of olfactory nuisance capable of generating annoyance in citizens residing in their neighborhood [2]. The annoyance is often continuous and can interfere with the state of human well-being, generating complaints and triggering conflicts that can also have repercussions on economic, commercial, and touristic activities [3]. According to [4], odor emissions are environmental pollutants that are given by the interaction of different volatile chemical species (e.g., sulfides, mercaptans), nitrogen compounds (e.g., ammonia, amines), and volatile organic compounds (VOCs). The VOCs are a group of organic chemicals characterized by a certain volatility in space and magnitude. The most common sources of odor emissions are the wastewater treatment plants (WWTPs), especially in urbanized areas [5,6]. They generate a huge amount of gaseous compounds characterized by unpleasant odors that are harmful to the environment and to the human health. For this reason, many countries proposed different strategies for odor measurements and assessment, which are essential for odor regulation and for controlling the interactions between chemical components and their dilution in the atmosphere [7]. Overall, the measurement of odor concentration in WWTP includes sensory techniques based on human examiners or analytical (instrumental) methodologies [8].

According to this classification, sensory techniques include dynamic olfactometry, field inspection, and recording from residents; on the other hand, analytical methodologies include mass spectrometry, identification of specific compounds, and electronic nose (E-nose).

In particular, the measurement of an odor concentration can be obtained in three different ways, i.e., by analytical determinations based on the mass spectrometry, by the olfactory perception from a group of panelists and by the E-nose (IOMS, Instrumental Odor Monitoring System), based on the interaction between special sensors and volatile molecules. The application of the E-nose is increased in many research areas of industrial systems, such as pharmaceutical companies, food industries, agriculture, and biotechnology, as well as for the monitoring process of the treatment plants. It is based on a set of chemical sensors,

[*] Corresponding author at: Department of Economic Sciences, University of Salento, Complesso Ecotekne, Lecce, Italy.
E-mail address: monica.palma@unisalento.it (M. Palma).

https://doi.org/10.1016/j.seps.2024.102021
Received 15 May 2024; Received in revised form 4 July 2024; Accepted 13 July 2024
Available online 20 July 2024

each of them with its specificity, and an appropriate pattern recognition system capable of identifying simple or complex odors [9]. In the case of a treatment plant, several odor sensors are installed along the various sections of the plant, and therefore they represent the key components for continuously monitoring and controlling odor emissions. However, the odors are emitted from different areas of the plant, so that the identification of the exact source of a particular odor may be difficult.

Several recent studies investigated the characterization of odor emission in WWTPs [10] and the prediction of odor concentration emitted from waste treatment, by using machine-learning models [11]. In addition, [12] applied artificial neural network and decision trees to predict the odor properties of post-fermentation sludge from two biological–mechanical treatment plants located in the North of Poland; in [13] the Random Forest (RF) was applied to recognize odor classes nearby a WWTP. The results obtained from the analysis established to what extent the unpleasant odors perceived by the citizens came from the monitored plant.

It is also worthy to mention the analysis of the relationship between odor indices measured with the so-called "triangle odor method" (Japanese standard method) [14], and odor concentration measured with dynamic olfactometry. The results suggested a strong linear correlation between the two methods, in the case of high concentrations [15].

Furthermore, [16] applied different algorithms for emission dispersion estimation in WWTPs with the use of Lagrangian atmospheric models. More recently, the work by [17] provided a review on the biochar-based odor mitigation in WWTPs that emerged in the last five years.

Within this context, the present work aims to propose a comparison between one of the most widely used learning method for classification, that is the RF classifier, and the classical multiple discriminant analysis (MDA), to investigate and evaluate the capacity of the multi-parameter sensors system (based on 10 sensors with sensitivity to different VOCs) to distinguish the waste treatment plant sections. A further innovative aspect concerns the integration of appropriate resampling techniques to process an imbalanced dataset. More specifically, this analysis provides an assessment of the possible influence of the dataset balancing on accuracy, discrimination, and identification of the emission sources. Several strategies have been proposed to generate synthetic samples for balancing datasets in the training of multi-class algorithms. It has been highlighted that the RF classifier has shown a better capacity to classify with respect to the classical MDA, in terms of the specified evaluation metrics, and the over-sampling approach has improved the classification performance of the multi-class algorithms.

Thus, differently from the works on this subject, in this contribution the odor measurements from the sensors installed on some treatment plants have been analyzed through a multi-class RF model, and the ability of the sensors to distinguish the sections of the treatment plant that produced the specific VOC at all stages of the plants has been assessed by considering also the effects of various resampling approaches used to balance the data.

As known, the sensors used in the E-noses are sensitive to a wide range of chemical compounds and therefore they are characterized by poor selectivity. The sensor matrix is generally made up of 6–10 sensors which are devoted to give a set of responses that constitutes the so-called "finger print", that is a kind of "olfactory pattern", of the odor source. Measuring at what extent the sensors are able to discriminate the odor sources could be extremely useful for managing and controlling a WWTP.

In this paper, after a short discussion on the resampling approaches, a brief review of the basic theoretical models of analysis such as the RF and MDA classifiers is presented (Section 2). Moreover, a review on classification measures to assess the performance between the two classifiers and the combination of each classifier with different resampling approaches are described (Section 3). Section 4 provides a detailed description of the dataset used in the multi-class classification problem. Thus, the case study concerning the ability of a multi-parametric system of 10 sensors to discriminate and identify the treatment plant sections generating the gaseous compounds is thoroughly discussed (Section 5). In particular, the RF classifier is implemented (Section 6) and compared with MDA traditional model (Section 7) when different resampling approaches are adopted. Finally, some concluding remarks are reported in Section 8.

2. Theoretical framework

Nowadays, WWTPs are considered as the main unpleasant odor sources in urbanized areas [5,6], and the reduction of their emissions is one of the most crucial aspects in air monitoring. Moreover, the identification of the emission source is a complex issue to be tackled in order to preserve possible environmental damages. This research compares two classification techniques developed to assess the ability of sensors to distinguish the emission sources (the sections of the treatment plant) using a sample dataset composed by measurements from 10 metal oxide semiconductors of the IOMS installed at various treatment plants. In the pre-processing step, a preliminary resampling has been considered and the multi-classification problem has been faced through a supervisioned learning method, that is, the RF, and its performance has been compared to the one of a traditional method such as the discriminant analysis.

According to the number of classes, there are two different classification problems, namely, binary classification in case of solely two classes or multi-classification problem when the number of classes is greater than two. In this paper, the aim is to learn from the training dataset how to categorize new unlabeled samples, in a multi-classification problem. Further, the presence of imbalanced data and possible problems related to unsuitable training observations have been treated through resampling techniques before fitting the model.

In the following, a brief description regarding the resampling methods and the classifiers adopted has been provided.

2.1. Resampling approaches for balancing classes

The imbalance classification problem occurs when the number of observations belonging to one class is significantly larger than the number of data of the other classes.

Dealing with imbalanced datasets in classification tasks became a relevant topic in data mining and machine learning [18,19]. In fact, the presence of imbalanced data might affect the learning process of a classification model. In particular, the study in [20] explored imbalanced data characteristics and [21] provided a classification of methods in the presence of class imbalance.

In the literature, several approaches were proposed to reduce the negative influence of class imbalance, that can be distinguished into data-level and algorithm-level strategies [22]. The former aims to balance the class distribution by resampling the original data in the pre-processing step. The latter attempts to develop new algorithms which are more suitable to identify the minority/majority class in imbalanced classification or adapt existing learning algorithms for this aim [23]. In this paper, the first approach has been considered, since it is generally adopted to deal with imbalanced data, and it is independent from the selected classifier. In particular, the data-level approaches are based on the idea of resampling in order to change the distribution of classes in the training dataset and to generate a new dataset with an equal ratio among the classes of the output variable, as well as to decrease the bias of the predictive algorithm [24].

The resampling techniques can be further classified into three different methods that include over-sampling, under-sampling, and hybrid methods. More specifically,

- the under-sampling approach balances the distribution of the classes by removing the cases linked to the majority class. One of the most widely-used method is the random under-sampling (RUS), which involves the random elimination of elements from the majority class [25]. The limitation of RUS is twofold: often the samples removed from the majority class might be useful, informative and if the samples within the majority class are unevenly distributed, the removal of randomly selected samples might change and worsen the classification performance. In addition, under-sampling can be achieved also applying statistical information (Informed under-sampling) as with the Tomek Links or Neighborhood Cleaning Rule;
- the over-sampling approach generates new cases linked to the minority class by producing more synthetic data or duplicating the existing ones. The most commonly adopted techniques for creating synthetic minority samples are the following: random duplication of minority samples known with the acronym ROS which stands for random over-sampling, adaptive synthetic (ADASYN) and synthetic minority over-sampling technique (SMOTE) [26]. The ROS algorithm replicates the minority class until the class distribution is balanced. The SMOTE algorithm generates synthetic data by using all samples from the minority class for uniformly over-sampling without considering the majority class distribution, which increases the overlap between classes. More specifically, in the SMOTE algorithm the synthetic data are found by using the k nearest neighbors (kNN) algorithm as linear combination of the existing data. The ADASYN is an extended version of the method SMOTE and generates a different sample using a weighted distribution for different minor class samples based on their level of learning difficulties. In particular, the algorithm generates more samples where the density of the minority class is low and fewer where the density is high. Therefore, with respect to the SMOTE algorithm, the ADASYN improves data distribution learning by reducing bias caused by class imbalance. In this paper, the ADASYN and ROS algorithms have been applied as over-sampling approaches;
- the hybrid approach combines both over-sampling and under-sampling methods at the same time to avoid over-generalization or the loss of useful data information. Within this context, multiple classifiers are combined to create a stronger and more accurate classifier in order to achieve a better performance than relying on a single classifier. Some hybrid-resampling approaches are the variants of the SMOTE algorithm, such as edited nearest neighbor (ENN) undersampling and the combination of the SMOTE and Tomek's procedures [27].

Although resampling procedures can improve the classification problem [28] they have some benefits and drawbacks. Taking into account the above-mentioned resampling strategies, it can be pointed out that the over-sampling procedure does not loose information from the training sample, but at the same time, it may lead to overfitting, since this method replicates existing observations of the minority class and increases the model-training time required to learn the process. On the other hand, under-sampling procedures reduce the samples from the original dataset potentially removing useful data that might be important for the learning process.

2.2. An overview of RF model

The RF model, introduced by Breiman [29] in 1984, has been applied in different research fields, as a supervised learning approach for classification and regression tasks.

As stated in [30], in the classification problem, the aim is to organize the dataset into classes by using predetermined class labels. In this supervised learning approach, the outcome is categorical and the aim is to "classify" new units into one of K possible classes. In this paper, K is greater than two and the tasks are referred to as multi-class classification problems.

RF is an ensemble random-decision tree with a randomized selection of variables, where each tree is generated through bagging or bootstrap sampling from the original training data. Each tree has nodes and leaves representing variables and decisions. Generally, the RF model consists of two main steps referred to as training and classification step. In the first step, RF builts trees from a randomly selected subset of the training dataset. Each tree is trained by using a bootstrap sample of cases from the data, and each split of candidate variables in the tree is randomly selected. Then, the classification step in the RF is based on the criteria known as plurality vote.

Formally, let $\mathbf{X} \in \mathbb{R}^{n\times p}$ be the training set with x_{ij} the observed value of the jth variable on the ith case (unit), and let Y be the set of responses, within K possible classes, in the training set. B sample sets are selected from the original training set using bagging sampling and the corresponding decision trees models $[f_{N_1}(x), f_{N_2}(x), \dots, f_{N_B}(x)]$ are built. Then, the validation test is used to check the classification results of each decision tree model. In the classification problem, the classifier uses the classification results from the trees and provides the decision according to the principle called "plurality vote" (multi-class response), determining the overall classification of an observation, defined as follows:

$$F(x) = \underset{y}{\operatorname{argmax}} \sum_{b=1}^{B} I(f_{N_b}(x) = y), \quad \text{with} \quad y \in \{1, \dots, K\}, \tag{1}$$

where $F(x)$ is the final classification result obtained by the plurality vote method, I is the indicator function, f_{N_b} represents a single decision tree multi-class classification model, x is the test set, and y is the prediction result.

The choice to apply this approach has been justified, on the one hand, by the possibility to identify non-linear patterns of data, and on the other hand, by the consideration that RF does not need variable scaling. Moreover, RF is robust in case of overfitting which may decrease the impact of different sample sizes. In the process of classification it assesses the impact of each variable providing the level of importance [31]. RF is also suitable for classification in the presence of an imbalanced dataset compared with other supervised methods. The RF algorithm used in this paper is defined as reported in the following Algorithm 1

Algorithm 1: RF classification

```
Data: Training step
1. $\mathbf{X}$: training set with $n$ cases and $p$ variables, and $Y$ outcome
   variable
2. $K$: number of classes in outcome variable
3. $B$: number of classifiers
Procedure
$b = 1$ to $B$
1. Generate a bootstrapped sample from the $\mathbf{X}$
2. Grow a tree using a random variable subset from bootstrapped
   sample
3. Construct trained classifiers $f_{N_b}$
Data: Classification step
Aggregate $B$ trained classifiers using the plurality vote, where the
predicted class label from classifier $f_{N_b}$ is given in (1).
```

2.3. An overview of MDA

The discriminant analysis is a popular supervised technique which is widely applied in computer-vision, machine learning, pattern recognition and in other fields of research. This approach allows the dimensionality reduction preserving as much of the class discriminatory information as possible. Among the traditional classification approaches there are Bayesian discriminant analysis, distance discriminant analysis, linear discriminant analysis (LDA), and quadratic discriminant analysis (QDA) [32]. In particular, LDA allows the definition of a linear combination of the observed variables which best separate two classes of cases. By applying LDA on a multivariate dataset, a model, called *discriminant function* for the classification of the cases under study, is defined.

When three or more classes are involved, the technique is referred to as MDA and more than one discriminant function can be computed. With respect to other techniques of multivariate analysis, such as cluster analysis, in MDA the clusters are known a priori; defining a latent discriminant function synthesizing the explanatory (quantitative) variables, new cases can be assigned to one of the groups of the primary (qualitative) variable. Hence, MDA can be used for descriptive (cases classification), as well as predictive (assignment of a new case to a group) purposes. In this paper, the use of discriminant analysis has been extended for multi-class classification problems. Thus MDA can be used only for classification, when the outcome variable presents more than two classes.

Let $\mathbf{X} = \{(X_{ij}) : i = 1,\dots,n, \quad j = 1,\dots,p\}$ be the data matrix referred to the variables that describing the cases, where x_{ij} is the value of the jth variable for the ith case. Given a $(n \times p)$ matrix $\mathbf{X}$ of $p > 2$ random variables observed on n sample cases, assume that the cases under study belong to K mutually exclusive groups defined by K different attributes of a categorical variable. Each group is composed by n_k cases, such that $\sum_{k=1}^{K} n_k = n$.

Thus, the generalization of the Fisher's procedure to K classes consists in a supervised method with the aim of constructing an objective function, as a linear transformation of $\mathbf{X}$ through $\mathbf{A}$, such that the known Rayleigh coefficient (the ratio of the between-class variance on the within-class variance) is maximized with respect to $\mathbf{A}$, as follows:

$$J(\mathbf{A}) = \text{argmax}\,\frac{\mathbf{A}'\mathbf{B}\mathbf{A}}{\mathbf{A}'\mathbf{W}\mathbf{A}} \tag{2}$$

where $\mathbf{B}$ is the between-class covariance matrix of dimension $(p \times p)$ and $\mathbf{W}$ is the within-class covariance matrix, defined as follows:

$$\mathbf{B} = \sum_{i=1}^{K} n_i(y_{i.} - \bar{y}_{..})(y_{i.} - \bar{y}_{..})'$$

$$\mathbf{W} = \sum_{i=1}^{K}\sum_{j=1}^{n_i}(y_{ij} - y_{i.})(y_{ij} - y_{i.})'$$

where $y_{i.}$ denotes the mean of the class i and $\bar{y}_{..}$ denotes the mean of all classes. Eq. (2) can be obtained from the generalized eigenvector equation $|\mathbf{B} - \lambda\mathbf{W}| = 0$, thus for the eigenvalues $\hat{\lambda}_m$, the associated vectors $\mathbf{a}_m$ for $m = 1,\dots s$, where $s = min(K-1,p)$, maximize the ratios $\mathbf{a}'_m\mathbf{B}\mathbf{a}_m/\mathbf{a}'_m\mathbf{W}\mathbf{a}_m$. The s uncorrelated functions are the linear discriminant functions and have been built to provide the maximum separation on average based upon the sample.

3. The evaluation metrics with multi-classification problem

In this section, a theoretical review on classification measures useful to compare the RF and the classical MDA, especially when different resampling methods are applied on the dataset, is provided.

In particular, the evaluation metrics have been used to measure and assess the effectiveness and adequacy of models, highlighting the advantages and disadvantages during the development of classification method, as well as they have been applied to quantify the quality of the trained classifier when validated with the unseen dataset.

For multi-class classification with imbalanced data, two main widely overall measures of model's performance are computed: macro and micro averaged metrics. The first one assigns equal importance (weight) to the classes regardless to the number of samples in a given class, whereas the second one assigns equal weight to each observation. More specifically, the macro averages provide a measure of effectiveness on classes with small observations, while the micro averages provide a measure of effectiveness on classes with large observations. For this reason, the macro-averaged metrics have been considered.

These measures, which include the accuracy, precision, recall, F1-score and balance accuracy, are based on the confusion matrix $\mathbf{C}$, since it encloses the information concerning the classification rule and classification algorithm [33]. Formally, let $\mathbf{C}$ be a confusion matrix $(k \times k)$ referred to a classifier, i.e.,

$$\mathbf{C} = \begin{bmatrix} c_{11} & c_{12} & c_{13} & c_{1k} \\ c_{21} & c_{22} & c_{23} & c_{2k} \\ \vdots & \vdots & \ddots & \vdots \\ c_{k1} & c_{k2} & c_{k3} & c_{kk} \end{bmatrix} \tag{3}$$

where c_{ij} is the number of outcomes associated to the jth observed class in column and the ith predicted class in row, with $i, j = 1,\dots,k$. The diagonal elements c_{ii} represent the number of observations of each class adequately classified, while the others constitute misclassification. Moreover,

$$c_{.i} = \sum_{j=1}^{k} c_{ji}, \quad \forall i = 1,\dots,k, \quad i \neq j, \tag{4}$$

$$c_{i.} = \sum_{j=1}^{k} c_{ij}, \quad \forall i = 1,\dots,k, \quad i \neq j, \tag{5}$$

represent, respectively, the number of observed cases and the number of predicted cases, for the ith class. Thus, the evaluation metrics generated from the confusion matrix are defined and described as follows:

- the overall accuracy

$$ACC = \frac{\sum_i c_{ii}}{\sum_{ij} c_{ij}} \tag{6}$$

  represents the ratio of the total number of data correctly predicted, averaged over all classes. The accuracy gives an overall estimate of predictive power of a model. High values of classification accuracy obtained by a classifier are considered optimal because reflecting the better classification;
- the balance accuracy

$$BalanceAccuracy = \frac{1}{k}\sum_{i=1}^{k}\frac{c_{ii}}{max(c_{i.}, c_{.i})} \tag{7}$$

  where the main difference between the two metrics is mainly due to the weighting applied on each actual class;
- the macro average precision

$$P_{macro} = \frac{1}{k}\sum_{i=1}^{k}\frac{c_{ii}}{c_{i.}} \tag{8}$$

  which measures the proportion of predicted correct outcomes to the total number of predicted outcomes, averaged over all classes;
- the macro average recall

$$R_{macro} = \frac{1}{k}\sum_{i=1}^{k}\frac{c_{ii}}{c_{.i}} \tag{9}$$

  which measures the proportion of predicted correct cases to the total number of observed cases, averaged over all classes;
- the F1-score (also known as Dice Similarity Coefficient)

$$\text{F1-score} = 2 * \frac{P_{macro}R_{macro}}{P_{macro} + R_{macro}} \tag{10}$$

  which represents the harmonic mean of P_{macro} and R_{macro} providing a balance assessment between these two metrics, as defined by [34].

Finally, it is worth highlighting that all metrics lie within $[0; 1]$. High values of accuracy, precision, recall, and F1-score are considered optimal.

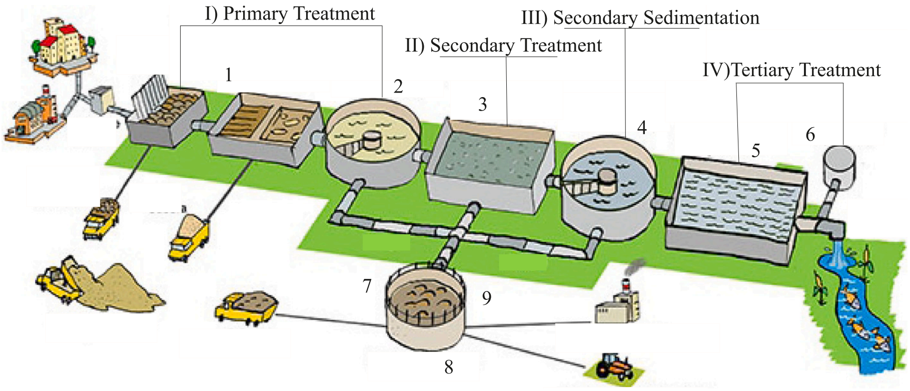

Fig. 1. Scheme of a typical WWTP, with its 9 sections and 4 macro-sections: 1-grit removal and 2-primary sedimentation, included in the first macro-section or primary treatment; 3-biological oxidation 7-stabilization of sludge, 8-sludge storage and 9-sludge thickening, included in second macro-section or secondary treatment; 4-secondary sedimentation, included in the third macro-section; 5-denitrification and 6-equalization, included in the fourth macro-section or tertiary treatment.
Source: Arpat.it.

4. Dataset description and pre-processing

A WWTP is a complex system which includes a series of treatment processes. Generally, WWTPs treatments can be distinguished into primary and secondary treatments. As a consequence, several groups of air pollutants can be generated at different stages as well as the formation of odoriferous compounds. For this reason, one should monitor in real time the odor emissions from a WWTP by employing an IOMS, also known as E-nose. In this context, the use of an E-nose provides several advantages compared to other techniques for odor measurement, such as chemical analysis, which are more expensive and discontinuous [35]. The E-nose is related to a multivariate response of the gas sensor array quantitatively characterizing odors [36].

The dataset used in this study has been collected in 2021 by a company specialized in technical and scientific environmental assistance and consultancy services for private and public enterprises. To enhance data quality and the adequacy to perform the modeling, a pre-processing phase has been faced. This phase is very important to overcome the limits related with real world dataset as in this case study.

Therefore, the filtering has been carried out, so that missing data as well as irrelevant and unnecessary data, which might decrease the accuracy of the learning approach, have been deleted. Afterwards, a data exploration analysis has been performed to detect the presence of possible extreme values. The median absolute deviation (MAD) has been applied for removing outliers from the dataset.

More specifically, after the pre-processing phase, the dataset consists of 285 measurements from 10 sensors of a IOMS, which record different types of gaseous compounds, namely W1C-Aromatic, W5S-Broad range (broad range sensitivity, react on nitrogen oxides), W3C-Aromatic (Ammonia), W6S-Hydrogen, W5C-Aromatic aliphatic (alkanes, aromatic compounds, less polar compounds), W1S-Broad methane, W1W-Sulphur organic (terpenes and sulphur organic compounds), W2S-Broad alcohol (alcohols, partially aromatic compounds), W2W-Sulphur chloride (sulphur organic compounds), and W3S-Methane aliphatic (sensitive to high concentration of methane). The sensors' measurements are expressed in mA (milliamperes).

The list of all considered sensors, jointly with their description, is presented in Table 1. An additional categorical variable, called "treatment plant section", that defines the section of the plant at which the sensor data have been recorded, has been considered in the following analysis. More specifically, the "treatment plant section" variable regards 9 sections, categorized in 4 macro-sections, of the plant (Fig. 1), that are: 1-grit removal and 2-primary sedimentation, included in the first macro-section; 3-biological oxidation 7-stabilization of sludge, 8-sludge storage and 9-sludge thickening, included in the second macro-section; 4-secondary sedimentation, included in the third macro-section; 5-denitrification and 6-equalization, included in the fourth macro-section. As already specified, the available dataset has been used to evaluate to what extent the sensor's measurements allow the separation of the sections of the treatment plant which have produced the specific odor emission. By analyzing the descriptive statistics, it has been underlined that the sensors W1C, W3C, and W5C registered, on average, the lowest values, while the sensors W1W and W2W provided the highest values. The behavior of chemical gas sensors could be attributed to two main aspects, i.e., cross-sensitivity and low selectivity. In fact, the sensors are affected by a mixture of gases with similar chemical properties [37]. In the following, the ability of the sensors to distinguish the emission sources of the treatment plant has been evaluated with respect to 4 macro-sections of the plant. To this aim, the sensors' average values have been first compared with respect to 4 macro-sections of the plants by using the non parametric Kruskal–Wallis test [38]. Table 1 highlights some statistics of the sensors values classified into the 4 plant macro-sections and shows that, at significance level of 5%, the sensors data are different on average.

Remark

The steps of the proposed multi-classification algorithm for imbalanced data are the following:

- data pre-processing removing the missing data and exploration data analysis, to detect possible extreme values (outliers) among the olfactometric measurements;
- class balancement through the resampling approaches;
- multi-classification by using RF and MDA;
- assessment of multi-classification models based on some evaluation metrics.

5. Imbalanced data analysis and resampling

In this article, a novel approach on odor source classification is provided. A machine learning classifier, such as RF, is implemented and compared with respect to the MDA; then the effect of some resampling methods on the performance of the classifiers RF and MDA is assessed.

In particular, different resampling approaches are used to improve the recognition accuracy of both majority and minority classes, taking into account the non-uniform distribution of the training data, as described in Section 2.1. For this reason, the imbalance ratio (IR) is computed to measure the degree of class-imbalance extent and its impact on the classification performance. The IR is obtained as the ratio between the majority and minority classes, where a dataset with an IR value equal to 1 can be considered perfectly balanced, while a dataset with an IR > 1 is considered imbalanced [39]. The index is a good imbalance metric for binary class data, while for multi-class data it only

Table 1
Mean values and standard deviations (in brackets) for the distributions of the sensor data, classified into 4 plant macro-sections (*p-value < 0.05).

Nr.	Gaseous compound	Description	Sensitive substance	Plant macro-section				
				I	II	III	IV	Test
1	W1C	Affects the skin, eyes, and respiratory tract. Used in the production of paints and rubber	Aromatic	0.7 (0.2)	0.7 (0.1)	0.7 (0.1)	0.7 (0.2)	8.4*
2	W5S	Produced by biogenic sources such as plants and yeasts. Some are toxic, and all contribute to ozone.	Broadrange	4.9 (7.6)	7.7 (13.9)	2.9 (0.7)	0.7 (12.8)	13.9*
3	W3C	A liquid that smells like gasoline and boils at 80 °C	Aromatic	0.7 (0.2)	0.8 (0.1)	0.7 (0.7)	0.7 (0.2)	0.7*
4	W6S	It is found in oil, human and animal waste, and sewage treatment. Used for producing chemicals	Hydrogen	1.7 (0.2)	1.6 (1.2)	1.6 (0.1)	1.7 (0.2)	43.2*
5	W5C	Produced from crude oil refinement. Causes headaches, dizziness, and even death.	Aromatic aliphane	0.7 (0.2)	0.8 (0.1)	0.8 (0.1)	0.7 (0.2)	27.9*
6	W1S	A colorless gas with a pungent odor. Toxic to human and aquatic organisms.	Broad methane	8.5 (5.8)	6.5 (5.8)	5.9 (3.3)	8.1 (6.4)	18.6*
7	W1W	A compound gas with distinctive aromatic flavors like citrus. Prevent inflammatory diseases.	Sulphur organic	35.2 (49.2)	19.8 (30.9)	10.8 (10.7)	29.1 (46.3)	14.7*
8	W2S	A poisonous gas. Originates from vehicle engines, waste burning and forest wildfires.	Broad alcohol	10.8 (9.1)	7.4 (6.9)	9.0 (7.6)	10.2 (8.5)	12.6*
9	W2W	It is found in garlic and in crude oil. Causes extreme global warming and acid rain	Sulphur chloride	29.7 (47.3)	14.3 (29.4)	8.9 (8.2)	23.0 (41.2)	14.8*
10	W3S	A gas with important greenhouse gas properties. Fuel production and engines.	Methane aliphatic	7.5 (5.3)	4.9 (3.3)	3.9 (1.3)	5.6 (3.0)	21.2*

Table 2
Descriptive statistics for the dataset sensor system, classified in 4 plant macro-sections used in the classification process. IR measure, N: the sample size, class+ and class−: the size of the majority and minority class, respectively.

dataset	IR	N	class+	class−
original database	4.5	285	109	22
training set	5	213	85	17
under-sampling (RUS)	1	76	15	15
over-sampling (ROS)	1	328	88	88
over-sampling (ADASYN)	1.07	325	83	77
hybrid-resampling	1.5	105	33	22

Table 3
Class sizes by datasets and resampling techniques.

dataset	class 1	class 2	class 3	class 4	N
original dataset	77	109	22	77	285
training set	58	82	15	58	213
test set	19	27	7	19	72
under-sampling (RUS)	15	15	15	15	60
over-sampling (ROS)	82	82	82	82	328
over-sampling (ADASYN)	77	82	83	83	325
hybrid-resampling	22	33	30	20	105

considers the information of the majority class and the minority class and ignores the information of classes in between. In Table 2 is reported the IR index, the sample size for each dataset with indication about the size of majority and minority classes with respect to the study dataset where the categorical variable has been defined by grouping the WWTP into 4 macro-sections, called indistinctly classes hereafter. According to Table 2, the IR shows values greater than one, equal to 4.5 and 5, for the original dataset and training dataset, respectively; on the other hand, all the resampling approaches have reduced significantly the value of this index, achieving datasets with less imbalanced classes.

It is important to specify that the original dataset has been divided into two sets: 75% for training and 25% for independent testing. To correctly compute the RF classifier, the Gridsearch has been used to tune the hyper-parameters, and the model's performance has been evaluated on the 25% test set.

Hence, the training sample has been used to generate new balanced datasets for improving the classification model. The class frequencies of the datasets involved in the analysis are reported in Table 3. It is evident that in the RUS method, the size of classes has been reduced by randomly removing some cases in order to match the size of the minority class (class 3).

More specifically, all wastewater treatment sections have been resampled with a class numerosity corresponding to minority class, that is the plant section of the secondary sedimentation (class 3). Note that the purpose of this section is to remove the biological solids, referred to as biological sludge, that are normally combined with primary sludge for its processing. On the other hand, the over-resampling with the ROS method has increased the size of all classes (the three macro sections relating to primary treatment, tecondary treatment and tertiary treatment) by randomly resampling some cases to match the size of the majority class (secondary treatment). The secondary treatment, corresponding to the majority class, includes biological activities (divided into aerobic and anaerobic parts) which are used to remove biodegradable, soluble, organic and nutrient substances from wastewater. Indeed, the main widespread treatment is the activated sludge. In the over-resampling with ADASYN, the number of synthetic samples generated for each minority sample has been decided through a density distribution. For this reason, the wastewater treatment sections have been resampled by considering as a reference the class numerosity corresponding to the majority class (class 2).

Similarly, in the hybrid-resampling, a combination of the data-based and algorithmic approaches has been used to handle the imbalanced datasets. The R statistical software has been used for all the computational aspects concerning cleaning, exploratory data analysis, classification-model fitting, and performance evaluation.

6. RF classifier and resampling approaches

In this study, the multi-classification RF models have been implemented on the no-resampling dataset (training dataset) and on different resampled training datasets, obtained through over-sampling, under-sampling, and hybrid-resampling. The resampling has been performed at the pre-processing step by changing the class distribution of the training set and the influence of the IR on the classification capabilities of the RF model has been measured. In other terms, it has been assessed the influence of different resampling approaches on the RF algorithm used to detect the ability of sensors array to discriminate and identify the emission source related to WWTP.

Therefore, the RF classifier has been trained on the basis of different training datasets, as described in the following:

- Model 1 is built by using the no-resampling dataset as a training set;
- Model 2 is built by using the under-resampling dataset where samples of the majority class are randomly generated to be excluded from the database for training. This process is repeated until the class distribution is balanced, and each class has the same size, equal to the size of the minority class;
- Models 3 and 4 are built by using two different over-sampling approaches. Model 3 is built by using the random over-sampling with the ROS method where this strategy replicates the instances present in the dataset and these replications are randomly selected from the minority classes. Model 4 is built by using ADASYN, where this strategy generates minority data samples according to their distributions. The latter approach allows the reduction of the learning bias introduced by the original imbalanced data distribution;
- Model 5 is built based on the hybrid-resampling approach combining the two previous methods, by removing the instances of the majority class and increasing the instances of the minority class at the same time to create a balanced synthetic dataset.

The comparison among classification models has been assessed through the confusion matrix. More specifically, Fig. 2 shows the confusion matrix for each RF classifier model from which the different evaluation measures have been generated. The diagonal elements represent the number of observations correctly classified, while the off-diagonal elements represent misclassification. In this way, one can immediately find out both the correct and wrong recognition of each type, where the class is referred to a section of the treatment plant. Focusing on the diagonal elements of the confusion matrices, it is evident that 41 out of 72 sections used for testing, have been detected correctly from the RF algorithm, trained on the dataset balanced through ADASYN (Fig. 2d). In the other cases, the RF classifier has correctly detected a lower number of sections.

Hence, according to the confusion matrix in Fig. 2d, the performance of the model might be improved by focusing on the predictive results of class 1 and class 4, namely the sections of the treatment plants which are referred to primary treatment and tertiary treatment, respectively. Class 2, referred to as “secondary treatment”, is the one with the greatest accuracy in prediction. This means that, for this section, the sensor’s measurements allow the separation of the sections of the treatment plant that have produced a specific odor emission.

The comparison among the goodness of the classification for each type of models has been evaluated by using the balance accuracy, as shown in Fig. 2. Such a measure highlights that the RF, together with ADASYN, significantly outperforms the others on each class/sections, namely this model shows a greater ability of odor sensors to distinguish the treatment plant sections. In addition, for all models, class 4 is the one that has achieved the lowest value of balance accuracy. This class/plant section is referred to as tertiary treatment, which is intended to remove the specific wastewater constituents that have not been eliminated with the secondary treatment (class 2). This result might be justified considering that this treatment is sometimes combined with the primary or secondary treatment to remove the smaller particles not captured by primary sedimentation.

From Fig. 2, it is clear that class 2 (associated with biological oxidation processes in the WWTP such as activated sludge) presents a higher recall index value than the precision index for each assessed model (with classified values ranging from 15 to 20). Note that, in this case, the biological treatments are considered as artificial ecosystems where that sludge system has been reproduced in a limited space denominated “biological reactor”. Then, class 1, namely primary treatment, is the one with the highest number of misclassifications particularly with respect to model 5.

On the other hand, the RF has been used to rank the importance of the variables in classification problems by using a measure of significance, referred to as Mean Decrease Accuracy. For each variable, the Mean Decrease Accuracy is obtained by averaging the difference in out-of-bag errors before and after the permutation over all trees. Fig. 3 shows the Mean Decrease Accuracy score for each variable over the RF classifier, where they are sorted according to increasing importance. Three of the top-ranked features such as W6S, W3C, and W5S, have played the most important roles in the RF model based on over-resampling with ADASYN. In addition, it is evident that all models have identified W6S as the first most important variable for the classification model, compared with the rest of the sensors. However, the second variable has been different for the applied models. In particular, the RF with original training set has identified W5S as the second most relevant variable. W3S has resulted as the second most relevant variable for the RF model with under-sampling and over-sampling with ROS, whereas the hybrid model has identified the sensor W1W. Conversely, the sensors W2W and W1S have been the ones considered less important in the multi-classification process. This methodology for selecting the most frequent variables in the construction of the most accurate trees has been powerful in identifying the sensors that best contribute to classification models. From the applied variable-selection method, it is possible to check and explore sensor array optimization of the E-nose to identify different sections in WWTP. However, external environmental conditions and the odor emitted from many different areas of the plant diminish the classification accuracy; this issue will be addressed in a future study.

7. Comparison of different sampling approaches for RF and MDA

In the previous section, the ability of the sensors to distinguish the emission sources of the treatment plant has been evaluated. To this aim, the supervised machine-learning RF has been also compared with MDA, where the target variable is related to the 4 macro-sections of the plant. In particular, in the following it is presented the performed analysis of the benchmarks of the two classifiers, and the combination of each classifier with different resampling approaches including data-level over-sampling, under-sampling and hybrid-resampling. Thus, the effects of IR on the performance of the multi-class classification models have been highlighted. The evaluation measures have been computed on the basis of the confusion matrices to assess and test classification efficiency, as reported in Table 4. The first column in Table 4 lists the classifiers and models used, then the obtained values of evaluation metrics.

Overall, the results of the metrics computed for this case study do not present very high values. This might be justified by taking into account that the VOC concentrations are characterized by a significant variability over space. However, it is evident that for both classifiers the use of resampling methods allowed to achieve a higher classification balance accuracy than on unbalanced datasets. The RF classifier based on the given training sets has provided a better performance than the MDA classifier. In particular, the best balance accuracy (with a value equal to 0.70) has been achieved through the RF with the ADASYN algorithm; this is also confirmed by the F1-score, which is greater than 0.5. Note that all the metrics go down when the data are balanced through the under-sampling approach. Compared to the use of the un-resampled training dataset, the RF classifier, trained on the best resampled data set (over-sampled with the ADASYN algorithm), has showed an increase of 0.05 for the balance accuracy and of 0.07 for the F1-score. However, it is worth pointing out that the balance accuracy is a preferable index in presence of imbalanced data as performance metric.

In terms of precision and recall measures, the RF significantly outperforms the MDA. Both recall and precision represent measures of trustworthiness and completeness. As a consequence, these metrics can be useful to measure the performance of models trained on imbalanced data. According to these considerations, the RF model based on over-sampling with ADASYN presents the highest values in terms of the aforementioned measures. According to these findings,

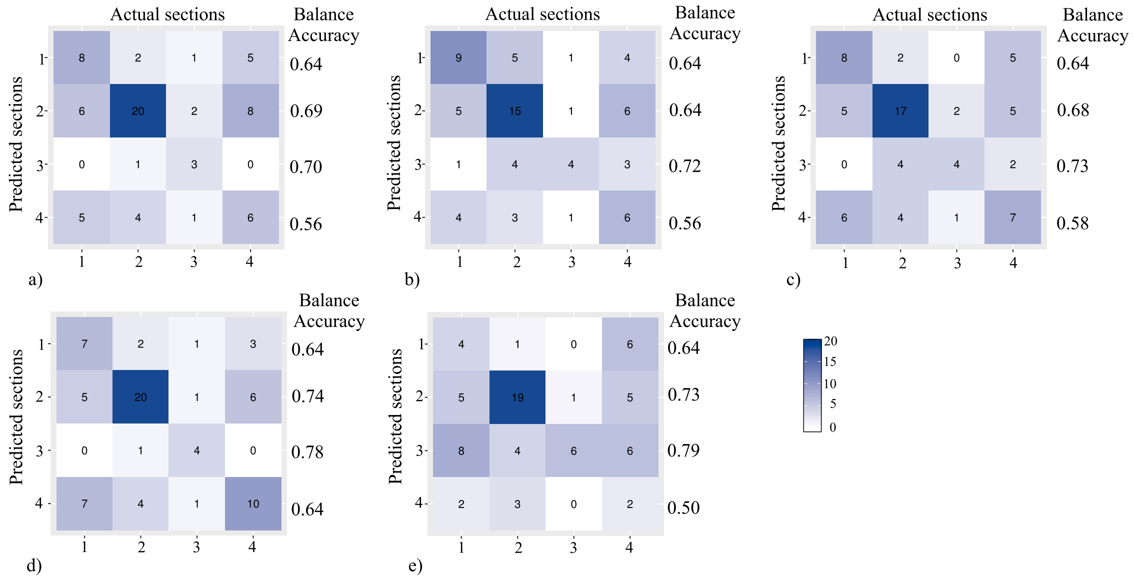

Fig. 2. Colormaps of the confusion matrices achieved with RF with training sets based on different class distributions: (a) Model 1 based on original dataset, (b) Model 2 based on RUS, (c) Model 3 based on ROS, (d) Model 4 based on ADASYN, (e) Model 5 based on hybrid-resampling. A darker color demonstrates more accurate prediction, and the diagonal shows the labels predicted correctly.

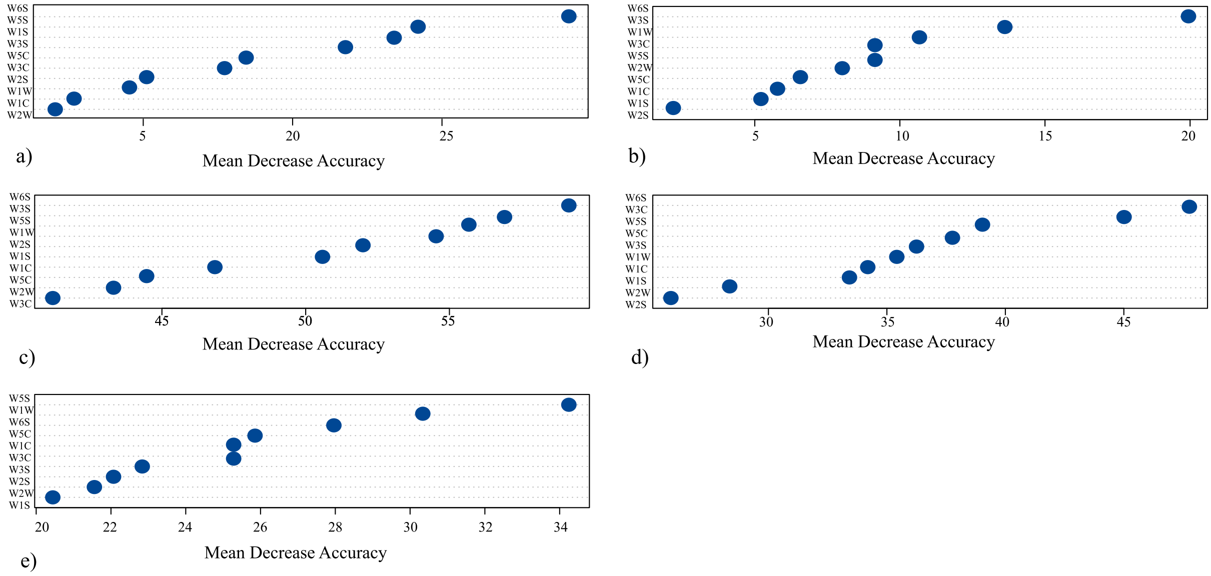

Fig. 3. The variable importance based on Mean Decrease Accuracy achieved with RF with training sets based on different class distributions: (a) Model 1 based on original dataset, (b) Model 2 based on RUS, (c) Model 3 based on ROS, (d) Model 4 based on ADASYN, (e) Model 5 based on hybrid-resampling.

Table 4
Comparison of the proposed models for multi-class classification.

	Balance accuracy	Accuracy	P_{macro}	R_{macro}	F1-score
RF model					
training set	0.65	0.51	0.55	0.48	0.50
under-sampling (RUS)	0.64	0.48	0.44	0.48	0.45
over-sampling (ROS)	0.66	0.50	0.48	0.50	0.48
over-sampling (ADASYN)	0.70	0.57	0.61	0.55	0.57
hybrid-resampling	0.61	0.50	0.48	0.50	0.48
MDA model					
training set	0.59	0.48	0.42	0.38	0.36
under-sampling (RUS)	0.65	0.47	0.47	0.50	0.45
over-sampling (ROS)	0.60	0.39	0.36	0.40	0.36
over-sampling (ADASYN)	0.58	0.45	0.27	0.39	0.25
hybrid-resampling	0.58	0.28	0.35	0.36	0.33

the macro-level metric for precision (P_{macro}) is greater than the macro-level recall (R_{macro}) for the RF with ADASYN resampling. This result might be influenced by the poor selectivity of the sensors of the WWTP, regardless of the number of observations. However, the recall of 0.55% and precision of 0.61% obtained with the ADASYN algorithm might be adequate to be implemented in a decision-support system. This highlights the impact of including sampling approaches especially by using the RF classifier.

8. Conclusions

This study focused on the ability of the RF classification method to evaluate the extent to which the sensors of a treatment plant can separate its macro-sections. The goodness of the proposed method was highlighted in terms of balance accuracy, precision, recall, and F1-score measures. Then, it was pointed out that the RF algorithm out-performs the classical MDA. The main contribution of this paper concerned the comparison among classification models based on different resampling methods. Indeed, the resampling approaches may help to handle multi-class imbalance, which can characterize sensor system data, as in the proposed application. The results are noteworthy because the over-sampling with ADASYN method for the RF classifier performs better with respect to the use of the original training set. Moreover, on the basis of the findings, it is possible to learn not only which method could be reasonably chosen according to the context, but also which variables are more relevant in the classification for each method. However, the evaluation metrics obtained for the RF, based on hybrid-resampling, might be adequate to be implemented in a decision-support system, taking into account that these models are more complex to train a multi-class classification algorithm.

Further developments will evaluate an improved RF classifier approach to better separate the odor emission sources by considering a combination of RF machine learning approach and a filter method for variable selection. Furthermore, it might also be interesting to assess classification algorithm for cost-sensitive learning to improve classification accuracy on an imbalanced dataset. Finally, an additional aspect that can help to improve the results is the expansion of the dataset, in particular for less frequent sections, and the combination of a binary classification with a multi-class classifier. In this way, the binary classification might identify the events and increase the values of some evaluation metrics such as recall and precision, while the multi-classification algorithm helps to identify in a more detailed manner the VOCs emitted by WWTP and the section of treatment plant.

Funding

This research was supported by:

- the National Biodiversity Future Center-NBFC, Spoke 4, Activity 4.1. Funder: Project funded under the National Recovery and Resilience Plan (NRRP), Mission 4 Component 2 Investment 1.4 - Call for tender No. 3138 of 16 December 2021, rectified by Decree n.3175 of 18 December 2021 of Italian Ministry of University and Research funded by the European Union – Next GenerationEU. Award Number: Project code CN_00000033, Concession Decree No. 1034 of 17 June 2022 adopted by the Italian Ministry of University and Research, CUP F87G22000290001, Project title "National Biodiversity Future Center - NBFC";

- ICSC–Centro Nazionale di Ricerca in High Performance Computing, Big Data and Quantum Computing, funded by European Union-Next GenerationEU. Project name: PNRR-HPC; Project No.: CN00000013;

CUP: C83C22000560007

CRediT authorship contribution statement

Veronica Distefano: Writing – review & editing, Writing – original draft, Validation, Methodology, Formal analysis, Data curation, Conceptualization. **Monica Palma:** Writing – review & editing, Writing – original draft, Conceptualization. **Sandra De Iaco:** Supervision.

Declaration of competing interest

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

Data availability

Data will be made available on request.

Acknowledgments

The authors sincerely thank the editor and two reviewers for their valuable comments and suggestions, which helped them to improve the paper.

The authors also express our gratitude to Dr. Lucrezia de Gennaro of the spin-off company of the University of Bari "LEnviroS" (https://www.lenviros.it/), for providing the dataset and for her support in the research process.

References

[1] Peters GM, Murphy K, Adamsen A, et al. Improving odour assessment in LCA—the odour footprint. Int J Life Cycle Assess 2014;846:1891–900.

[2] Zarra T, Galang MGK, Ballesteros FC, Belgiorno V, Naddeo V. Environmental odour management by artificial neural network - A review. Environ Int 2019;133 Pt B:105–89.

[3] Oliva G, Zarra T, Massimo R, Senatore V, Buonerba A, Belgiorno V, Naddeo V. Optimization of classification prediction performances of an instrumental odour monitoring system by using temperature correction approach. Chemosensors 2021.

[4] Barth CL, Elliott LF, Melvin SW. Using odor control technology to support animal agriculture. Trans ASABE 1984;27:859–64.

[5] Gostelow P, Parsons S, Stuetz R. Odour measurements for sewage treatment works. Water Res 2001;35:579–97.

[6] Giuliani S, Zarra T, Naddeo V, Belgiorno V. A novel tool for odor emission assessment in wastewater treatment plant. Desalin Water Treat 2015;55:712–7.

[7] Carrera-Chapela F, Donoso-Bravo A, Souto JA, Ruiz-Filippi G. Modeling the odor generation in WWTP: An integrated approach review. Water Air Soil Pollut 2014;225:1–15.

[8] Munoz R, Munoz R, Sivret EC, Parcsi G, Lebrero R, Wang X, Suffet IHM, Stuetz RM. Monitoring techniques for odour abatement assessment. Water Res 2010;44(18):5129–49.

[9] Gardner JW, Bartlett PN. A brief history of electronic noses. Sensors Actuators B 1994;18(1):210–1.

[10] Burgués J, Doñate S, Esclapez MD, Saúco L, Marco S. Characterization of odour emissions in a wastewater treatment plant using a drone-based chemical sensor system. Sci Total Environ 2022;846:157–290.

[11] Kang JH, Song J, Yoo SS, Lee B-J, Ji HW. Prediction of odor concentration emitted from wastewater treatment plant using an artificial neural network (ANN). Atmosphere 2020;11(8).

[12] Byliński H, Sobecki A, Gebicki J. The use of artificial neural networks and decision trees to predict the degree of odor nuisance of post-digestion sludge in the sewage treatment plant process. Sustainability 2019;11:4407.

[13] Cangialosi F, Bruno E, Santis G. Application of machine learning for fenceline monitoring of odor classes and concentrations at a wastewater treatment plant. Sensors 2021;21:4716.

[14] Iwasaki Y, Fukushima H, Nakaura H, Yajima T, Ishiguro T. A new method for measuring odors by triangle odor bag method. J Jpn Soc Air Pollut 1978;13(6):246–51.

[15] Naddeo V, Zarra T, Kubo A, Uchida N, Higuchi T, Belgiorno V. Odour measurement in wastewater treatment plant using both european and japanese standardized methods: correlation and comparison study. 2016;18:728–33.

[16] Ravina M, Panepinto D, Mejia J, Giorgio L, Salizzoni P, Zanetti M, Meucci L. Integrated model for estimating odor emissions from civil wastewater treatment plants. Environ Sci Pollut Res 2020;27:3992–4007.

[17] Lee J, Lee S, Lin K, Jung S, Kwon EE. Abatement of odor emissions from wastewater treatment plants using biochar: Review of the state-of-the-art approaches. Environ Pollut 2023;122426.
[18] Krawczyk B. Learning from imbalanced data: open challenges and future directions. Prog Artif Intell 2016;5(4):221–32.
[19] Yang Q, Wu X. 10 Challenging problems in data mining research. Int J Inf Technol Decis Mak (IJITDM) 2006;5(4):597–604.
[20] López V, Fernández A, García S, Palade V, Herrera F. An insight into classification with imbalanced data: Empirical results and current trends on using data intrinsic characteristics. Inform Sci 2013;250:113–41.
[21] Sun Y, Wong AKC, Kamel MS. Classification of imbalanced data: a review. Int J Pattern Recognit Artif Intell 2009;23:687–719.
[22] Fernández A, García S, Galar M, Prati R, Krawczyk B, Herrera F. Learning from imbalanced data sets. 2018.
[23] Stefanowski J. Dealing with data difficulty factors while learning from imbalanced data. In: Matwin S, Mielniczuk J, editors. Challenges in computational statistics and data mining. Cham: Springer International Publishing; 2016, p. 333–63.
[24] Johnson JM, Khoshgoftaar TM. Survey on deep learning with class imbalance. J Big Data 2019;6:1–54.
[25] Tahir M, Kittler J, Mikolajczyk K, Yan F. A multiple expert approach to the class imbalance problem using inverse random under sampling. In: Multiple Classifier Systems. Berlin, Heidelberg: Springer Berlin Heidelberg; 2009, p. 82–91.
[26] Chawla N, Bowyer K, Hall L, Kegelmeyer W. SMOTE: Synthetic minority over-sampling technique. J Artif Intell Res (JAIR) 2002;16:321–57.
[27] Hairani H, Priyanto D. A new approach of hybrid sampling SMOTE and ENN to the accuracy of machine learning methods on unbalanced diabetes disease data. International Journal of Advanced Computer Science and Applications 2023;14(8):585–90.
[28] He H, Bai Y, Garcia EA, Li S. ADASYN: Adaptive synthetic sampling approach for imbalanced learning. In: 2008 IEEE international joint conference on neural networks (IEEE world congress on computational intelligence). 2008, p. 1322–8.
[29] Breiman L, Friedman JH, Olshen RA, Stone CJ. Classification and regression trees. Biometrics 1984;40:874.
[30] Fayyad U, Piatetsky-Shapiro G, Smyth P. From data mining to knowledge discovery in databases. AI Mag 1996;17(3):37–54.
[31] Al-Behadili H. Decision tree for multiclass classification of firewall access. Int J Intell Eng Syst 2021;14:294–302.
[32] Fisher RA. The use of multiple measurements in taxonomic problems. Ann Eugen 1936;7(2):179–88.
[33] Mosley L. A balanced approach to the multi-class imbalance problem. 2013.
[34] Sokolova M, Lapalme G. A systematic analysis of performance measures for classification tasks. Inf Process Manage 2009;45(4):427–37.
[35] Capelli L, Sironi S, Del Rosso R. Electronic noses for environmental monitoring applications. Sens (Basel Switz) 2014;14:19979–20007.
[36] Bax C, Capelli L. Odour nuisance monitoring. In: De Vito S, Karatzas K, Bartonova A, Fattoruso G, editors. Air Quality Networks: Data Analysis, Calibration & Data Fusion. Cham: Springer International Publishing; 2023, p. 95–113.
[37] Yaqoob U YM. Chemical gas sensors: Recent developments, challenges, and the potential of machine learning. A review. Sensors 2021;21(8).
[38] Kruskal WH, Wallis WA. Use of ranks in one-criterion variance analysis. J Amer Statist Assoc 1952;47(260):583–621.
[39] Verbiest N, Ramentol E, Cornelis C, Herrera F. Improving SMOTE with fuzzy rough prototype selection to detect noise in imbalanced classification data. 2012, 7637.

Veronica Distefano is a researcher in Statistics at the Department of Economic Sciences of the University of Salento (Lecce, Italy) since July 2022. Previously at the same University, she held a Ph.D. degree in Statistics and she worked as a PostDoc in several research projects aiming to develop statistical models for reducing the dimensionality and forecasting of spatial variables. Afterwards, she was PostDoc research fellow at ECLT (European Centre for Living Technology) at Ca' Foscari University of Venice, where her research interests have regarded the development of new analytical methodologies, included random forest, cluster analysis and dimension reduction techniques, to analyze the information contained in different type of data (environmental, social and economic) for big data sets.

Monica Palma She is Associate Professor in Statistics at the University of Salento (Lecce, Italy) since 2015. Her research interests are referred to: (a) Multivariate Geostatistics for environmental data, (b) Space–time covariance modeling, (c) Stochastic conditional and unconditional simulation, (d) GIS and WebGIS for big data. She is author of several publications in scientific international journals, as well as of statistical books.

Sandra De Iaco is a Full professor in Statistics at the University of Salento (Lecce, Italy). She gained her Phd in Statistics, with a dissertation in “Space–time covariance models” at the University of Chieti-Pescara, (Italy). Her research interests are referred to: (a) multilevel regression model and multinomial logit model, (b) time series analysis (c) space–time covariance modeling, (d) multivariate Geostatistics for environmental data, (e) stochastic conditional and unconditional simulation. She has given numerous short courses and lectures on spatio-temporal modelling and computational tools for the analysis of water and air quality monitoring. She serves as member of the editorial board and reviewer of various international scientific journals and has been member of scientific committees of renowned conferences. She is author of more than 150 scientific publications.