Annals of Agricultural and Environmental Medicine 2025, Vol 32, No 2, 222–229
www.aaem.pl

ORIGINAL ARTICLE

Improvement in classification capabilities of surface water samples based on analysis of multidimensional data from gas sensor array

Magdalena Piłat-Rożek[1,A-F], **Grzegorz Łagód**[2,A-F]

[1] *Department of Applied Mathematics, Lublin University of Technology, Lublin, Poland*
[2] *Department of Water Supply and Wastewater Disposal, Lublin University of Technology, Lublin, Poland*
A – Research concept and design, B – Collection and/or assembly of data, C – Data analysis and interpretation, D – Writing the article, E – Critical revision of the article, F – Final approval of the article

Piłat-Rożek M, Łagód G. Improvement in classification capabilities of surface water samples based on analysis of multidimensional data from gas sensor array. Ann Agric Environ Med. 2025; 32(2): 222–229. doi: 10.26444/aaem/206945

Abstract

Introduction and Objective. It has been proven that e-noses can successfully differentiate between drainage and river water samples. However, it was supposed that the classification accuracy in the previous article from the series could have been refined. The aim of the article was to improve the classification accuracy of surface water samples analyzed with a gas sensor array.

Materials and Method. The multidimensional data on which the machine learning models were trained was derived from river water, drainage water and synthetic air samples measured using an array comprising 17 gas sensors. In this research, the unsupervised t-SNE and k-medians were used for dimensionality reduction, visualization on 2-dimensional plane, and clustering. Subsequently, supervised classificators XGBoost and AdaBoost.M1 were trained and compared with regard to the achieved quality of classification of objects into correct classes.

Results. The visualization using t-SNE and clustering with k-medians clearly distinguished the observations from the water sample and different drainage samples. The applied supervised machine learning methods achieved 88.8% and 89.2% correct classifications on the test set for the XGBoost and AdaBoost.M1 models, respectively.

Conclusions. Despite the absence of statistical significance in differences of medians in most of the multiple comparisons between sample groups for all the classical indicators, the electronic nose allows differentiating and correctly classifying surface water samples with high accuracy.

Key words

water quality, surface water, electronic nose, multidimensional data analysis, XGBoost, t-SNE, k-medians, AdaBoost.M1

INTRODUCTION AND OBJECTIVE

With the development of new technologies and artificial intelligence, mankind has been looking for the solutions to quickly detect anomalies or distinguish between dissimilar objects. Electronic senses have become particularly helpful in this regard: the electronic nose, eye and tongue, which can operate separately or together. E-noses are used in various industries, i.e. food [1], medical [2], agricultural [3], as well as in environmental research, including those related to water quality [4] and wastewater treatment [5], among others.

The authors have dealt with the topic of classifying samples from drainages and river water in the past, and in a previous paper [6] noted that there is potential for using an electronic nose consisting of 17 MOS sensors and machine learning models for this purpose. This paper aimed at improving the classification accuracy of multivariate data by selecting unsupervised and supervised models that can handle the classification task more effectively. In view of the results obtained in the mentioned article, supervised methods based on a tree structure were selected. A thorough statistical analysis of measurements from reference methods of surface water quality testing was also performed. Improving the quality of drainage water classification by means of machine learning methods could contribute to the effective detection of the situations in which drainage water quality deviates from the reference. Such a condition is particularly problematic due to the fact that this type of water runs off, among others, from agricultural fields, often carrying nutrient elements from fertilizer leaching, as well as residues of plant protection products, or other potentially dangerous micropollutants that pose a threat to human health [7]. Nutrients entering surface waters may contribute to their eutrophication, and may also enter animal watering places. Thus, monitoring and screening of drainage water and other runoff discharged into surface waters facilitates taking preventive measures and can contribute to the prevention of water eutrophication, as well as reduce the spread of pollutants that negatively affect ecosystems and, thus, human health.

MATERIALS AND METHODS

Drainage and river water samples were taken within the People's Park in Lublin, eastern Poland. The river water for testing was taken from the Bystrzyca River, the first (1) drainage water sample at a point behind the embankment, samples labeled 2 and 3 at equidistant and adjacent points, and sample 4 from a point at the end of the park. The 5 samples (river and drainage water samples), along with

✉ Address for correspondence: Magdalena Piłat-Rożek, Department of Applied Mathematics, Lublin University of Technology, Nadbystrzycka 38, 20-618 Lublin, Poland
E-mail: m.pilat-rozek@pollub.pl

Received: 21.04.2025; accepted: 10.06.2025; first published: 25.06.2025

a reference synthetic air sample, were measured using a gas array comprising 17 Figaro sensors, constructed by researchers at Lublin University of Technology, which was used in the previous work in this series [6].

The data collected from the gas sensor array included 17 variables with measurement results (each variable represented one Figaro sensor) and a variable that contained information about the type of sample from which the observation came. These data were processed using unsupervised learning algorithms, i.e. t-SNE and the *k*-medians clustering method. Then, using the original data set containing 18 variables, which was divided in a 4:1 ratio (80% of observations in the training set, 20% in the test set) supervised models XGBoost and AdaBoost.M1 were trained. The parameters of these models were selected using hyperparameter grids, and the models were trained using a 5-fold cross-check through which validation took place and model overfitting was controlled.

The t-SNE (t-distributed stochastic neighbour embedding) method is an unsupervised machine learning method for reducing the dimensionality of input data. In the standard method, the user predetermined parameters are perplexity u, momentum α (in R it equals 0.8 as default) and learning rate η (in R it equals 200 as default). At the beginning of the algorithm, the conditional probabilities $p_{i|j}$ of x_i selecting observation x_j as its neighbour are calculated. In the Barnes-Hut algorithm [8], these probabilities are defined as:

$$p_{j|i} = \begin{cases} \dfrac{\exp\left(-\dfrac{d(x_i,x_j)^2}{2\sigma_i^2}\right)}{\sum_{k\in N_i}\exp\left(-\dfrac{d(x_k,x_k)^2}{2\sigma_i^2}\right)}, & j \in N_i \\ 0, & \text{otherwise.} \end{cases} \quad (1)$$

N_i denotes the set of indices $\lfloor 3u \rfloor$ of the closest observations relative to x_i, while σ_i denotes the standard deviation of the Gaussian function centred at x_i. The symmetrised conditional probability (pair-wise similarity) is then defined by the formula:

$$p_{ij} = \frac{p_{j|i} + p_{i|j}}{2n}, \quad (2)$$

where n is the number of all observations in the dataset. In the new reduced space of variables, the joint probabilities are expressed by the formula:

$$q_{ij} = \frac{\left(1 + \left\|y_i - y_j\right\|^2\right)^{-1}}{\sum_k \sum_{l\neq k}(1 + \|y_k - y_l\|^2)^{-1}}. \quad (3)$$

The t-SNE algorithm seeks to minimize the differences in p_{ij} and g_{ij} distributions by means of minimizing the cost function – Kullback-Leibler divergence:

$$D_{KL}(P\|Q) = \sum_{i\neq j} p_{ij} \cdot \log\frac{p_{ij}}{q_{ij}}. \quad (4)$$

The problem of minimizing this function is solved using the gradient descent method. The Barnes-Hut algorithm reduces the computational complexity of the problem by defining conditional probabilities $p_{j|i}$ according to equation (1) and approximating large distances between observations in a new space of variables using quadtrees. The parameter controlling how these distances are approximated is θ – the larger it is, the less accurate the approximations are (in R it has a default value of 0.5). In the R programming language, the library for training the t-SNE model using the Barnes-Hut algorithm is *Rtsne* [9].

Clustering algorithms such as *k*-means and *k*-medians were developed as early as the second half of the 20th century, making them one of the older but still widely used unsupervised machine learning algorithms. A common problem in using *k*-means is that it is less robust to the occurrence of outlier observations, and for this reason *k*-medians is more recommended. In each iteration of the algorithm operation, for each observation x_j, the number of the cluster to the median of which the distance is the smallest is sought:

$$r = \arg\min_{1\leq i\leq k}\left\| x_j - \hat{m}_i^{(t)} \right\|, \quad (5)$$

based on which the j-th observation is assigned to the C_r cluster. In addition, the Weiszfeld geometric median is updated, which for the i-th cluster ($i \in \{1,2,\ldots,k\}$) at iteration number $t + 1$ is expressed by the formula:

$$\hat{m}_i^{(t+1)} = \frac{\sum_{j\in X_i^{(t)}} \dfrac{x_j}{\left\|x_j - \hat{m}_i^{(t)}\right\|}}{\sum_{j\in X_i^{(t)}} \dfrac{1}{\left\|x_j - \hat{m}_i^{(t)}\right\|}} \quad (6)$$

where $X_i^{(t)}$ is the set of observation indices that belong to the C_i cluster in the t-th iteration, and $\hat{m}_i^{(0)}$ is the first randomly selected median of the C_i cluster [10]. Since it is usually up to the researcher to choose the parameter k, the algorithm from the *Kmedian* package of the R language [11] was chosen. The functions of this library allow finding the optimal number of clusters into which the dataset is to be divided when using at least 10 different values of the number k. This number is sought using 2 algorithms based on inheritance heuristics: dimension jump algorithm (Djump) and data-driven slope estimation algorithm (DDSE).

Both XGBoost and AdaBoost.M1 are supervised machine learning models based on classification trees. They were chosen because of the promising results of drainage water classification using the random forest model in the authors' previous work.

The XGBoost algorithm was presented in a paper [12] in 2016. Its operation begins with the selection of the parameters of the number of iterations B, the learning rate η, the regularization parameter λ, and γ corresponding to the minimum reduction of the loss function $L(y,\hat{y})$ at which the next division will be created in the tree. The XGBoost classifier starts by initializing a fixed value of the model score equal to:

$$\text{XGBoost}_0(x) = \arg\min_{\gamma} \sum_{i=1}^{n} L(y_i,\gamma), \quad (7)$$

where n is the number of observations on which the model is trained. In each iteration step $b \in \{1,2,\ldots,B\}$, the sum of the gradients G and the Hessians H for each observation is calculated. Then a classification tree is built, the optimization of which depends on G and H.

After building a tree with a certain number of terminal regions R_j, where $j \in \{1,2,\ldots,J_b\}$ for each j, a weight is calculated in the following way:

$$w_j = -\frac{G_j}{H_j + \lambda}, \quad (8)$$

where G_j denotes the sum of the gradients, while H_j denotes the hessians in leaf j. Then, in the b-th step, the model output is updated by:

$$\text{XGBoost}_b(x) = \text{XGBoost}_{b-1}(x) + \eta \sum_{j=1}^{J_b} w_j \cdot \mathbb{I}(x \in R_j), \quad (9)$$

while the final result of the model on the training set is the values of the $\text{XGBoost}_B(x)$ function.

An implementation of the XGBoost algorithm in the R programming language can be found in the *xgboost* library [12].

The AdaBoost.M1 algorithm was presented in 1996 in a paper [13]. At the beginning of the algorithm, fixed initial weights are initialized for each observation (x_i, y_i), where n is the number of observations on which the model is trained. At each iteration a weak classifier $T^{(b)}$, which is a decision tree, is fitted to the training set and its error is calculated:

$$\varepsilon_b = \sum_{i=1}^{n} w_i^{(b)} \cdot \xi_i^{(b)}, \quad (10)$$

where $\xi_i^{(b)} = \begin{cases} 0, & T^{(b)}(x_i) = y_i \\ 1, & T^{(b)}(x_i) \neq y_i \end{cases}$.

With its help, the learning rate α_b is calculated, which in the *adabag* library of the R language can be expressed by the Breiman formula of the form $\alpha^{(b)} = \frac{1}{2} \cdot \ln\left(\frac{1-\varepsilon_b}{\varepsilon_b}\right)$ or Freund's $\alpha^{(b)} = \ln\left(\frac{1-\varepsilon_b}{\varepsilon_b}\right)$ [14]. The weights for each observation are then updated in the next step

$$w_i^{(b+1)} = \frac{w_i^{(b)} \cdot e^{\alpha^{(b)} \cdot \xi_i^{(b)}}}{\sum_{i=1}^{n} w_i^{(b)} \cdot e^{\alpha^{(b)} \cdot \xi_i^{(b)}}}. \quad (11)$$

Then, after all iterations for each observation, the model response is calculated:

$$\text{AdaBoost.M1}(x) = \arg\max_{y_j} \sum_{b=1}^{B} \alpha^{(b)} \cdot \delta(T^{(b)}(x), y_j), \quad (12)$$

where δ is the Kronecker delta function.

In this study, statistical analysis of classical physicochemical parameters was carried out, data for analysis were obtained using measurements from equipment: DR 6000, HACH Lange, USA (TSS, COD), Orion VerseStar Thermo Scientidic (conductivity), TOC-LCSH/CSN Shimadzu (TN, IC, TC, DOC), Waterproof TN-100 Turibidimeter, EUTECH INSTRUMENTS, Singapore (turbidity) [6].

Statistical analysis was performed with the Kruskal-Wallis test using the kruskal_test function from the *rstatix* library [15]. This is a non-parametric test based on the ranks assigned to the observations. It is used to test the null hypothesis that the distributions of a variable from each of $k \geq 3$ groups are identical, with the alternative hypothesis that there is at least one pair of groups in which the distributions are different from each other. When the null hypothesis is rejected in the Kruskal-Wallis test, *post-hoc* tests are used. One such test is the Dunn test used in this study, and is used to compare medians or means in distributions of 2 groups. Its null hypothesis is that the parameters being compared in both groups are the same, while the alternative hypothesis can be either 2-sided or 1-sided. The p-value of the test, due to the fact that it is a repeated measures test, is corrected for using a matched correction to control for type I error. In this case, Holm's correction was used by sequentially multiplying the p-value in each of the multiple comparison tests by an increasingly smaller number. Dunn's test was performed using dunnTest function from the *FSA* library [16].

The graphs and statistical calculations were prepared and the models trained using the R statistical computing package version 4.4.0 [17].

RESULTS AND DISCUSSION

In order to visualize and show the classification potential, the input data space from the gas sensor array was mapped to 2 variables using the Barnes-Hut t-SNE algorithm. For this purpose, the parameters $u = 35$ and $\theta = 0.75$, were set, the other parameters listed in the algorithm description were not changed in value. Figure 1 (on the left) presents the result of this algorithm by means of a dot plot of observations in a 2-dimensional space of variables. In it, it can be observed that the 6 types of samples from which the observations are derived are arranged in distinct clusters located far away from each other. The points, although from a different sample type, and not next to their group, are to the right of and down from the cluster of Air observations (shown in red). Among them are at least one observation from each of the drainage

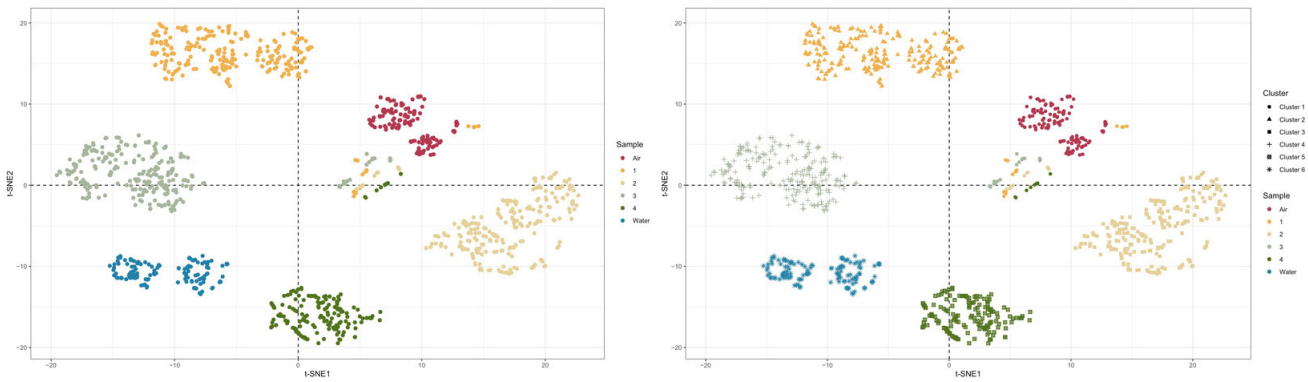

Figure 1. Two-dimensional t-SNE mapping of original data space: dot-plot of original points – different colors denote sample type of each observation (on the left) and 6-medians clustering algorithm on t-SNE mapping – colors denote sample types and shapes of points the clusters which the observations were assigned to (on the right)

water samples. All observations from synthetic air (Air) and river water (Water – highlighted in blue) samples are in unseparated clusters and as observations from reference samples are located on opposite sides of the graph.

The t-SNE algorithm in the Barnes-Hut method performs significantly better at visualizing data from drainage water and river water samples than the PCA shown in a previous study. However, because it is an unsupervised method, when the true classes of objects are unknown, this method and the selection of hyperparameters in it rely on observing point clouds that move away from or closer to each other [18].

By the shapes and homogeneity of the groups that form in Figure 1 (on the right), it can be purported that the *k*-medians algorithm would perform well with clustering such data. The mapped, 2-dimensional data derived from the results of the t-SNE algorithm were subjected to *k*-medians clustering. Since the *Kmedians* function from the package of the same name allows finding the optimal number of clusters, the performance of the algorithm was tested on between 3 and 13 groups, to which the observations of the dataset were assigned. The critical number of clusters obtained by calibration with both DDSE and Djump inheritance methods was 6. Figure 1 (on the right) presents the results of clustering with the *k*-medians method for the optimal number *k*. As expected, the points closest to the observations from the Air sample were assigned with them to Cluster 1. Moreover, Cluster 1 was the only cluster to which more than one class of observations was assigned.

The XGBoost model was trained using the grid of hyperparameters (Tab. 1). The model achieved classification correctness on the training set equal to 87.3% for parameter values:

Table 1. Values used in hyperparemeters grid search for the XGBoost and AdaBoost.M1 model training

Model	Hyperparameter	Possible values	Optimal value
XGBoost	*nrounds*	200, 300, 400, 450	400
	max_depth	3, 6, 9, 10, 11, 12	9
	eta	0.01, 0.1, 0.15, 0,3	0.01
AdaBoost.M1	*coeflearn*	Breiman, Freund	Breiman
	mfinal	100, 125, 150, 200, 500	200
	maxdepth	5, 6, 7, 8, 9, 10, 11	10

- *nrounds*, which denotes the maximum number of iterations equal to 400;
- *max_depth*, which corresponds to the maximum depth of a single tree equal to 9;
- *eta*, which is the learning rate of the model equal to 0.01.

In addition, the *colsample_bytree* parameter, denoting the proportion of explanatory variables randomized to single learning, was set to a fixed number equal to 0.8. The values of the other parameters were left at default.

The results of the XGBoost model on the test set are shown in Figure 2. The Accuracy of the model on the test set was 88.8%, where the best predicted classes were Air (synthetic air) and Water (water from the Bystrzyca River), where only one observation was incorrectly classified as Sample 2 from the drainage.

The AdaBoost.M1 model was trained using the grid of hyperparameters (Tab. 1). The model achieved a classification accuracy on the training set of 91% for the parameter values:

- *coeflearn*, being the choice of the formula with which the learning rate was calculated, the optimal one was Breiman's coefficient;
- *mfinal,* corresponding to the number of iterations equal to 200;
- *maxdepth*, which denotes the maximum depth of a single tree equal to 10.

The results of the AdaBoost.M1 model on the test set are shown in Figure 3. Accuracy of the model on the test set was 89.2%, where again the best predicted classes were Air and Water. In this case, the model performed better at predicting observations from the Sample 3 class (3 more correct classifications than in the XGBoost model), but one less observation was predicted for the Sample 2 and Water classes. The improvement, although small compared to the previous model, occurred for both the training and test sets.

The t-SNE algorithm was used in a study [19] that predicted groundwater quality in different administrative regions of Mexico. The model was used for 2-dimensional visualization and to show relationships between different samples. In turn, in the study [20] it was used for visualization and *k*-means clustering of observations from groundwater and surface water samples contaminated with manure, inorganic fertilizers, or wastewater. These applications are similar to the use of the t-SNE algorithm in the current study.

The *k*-medians clustering algorithm was used in the study [21] to divide the dataset into clusters that were to contain observations with different values of hydraulic flow units in different parts of the Williston Basin. Separate machine

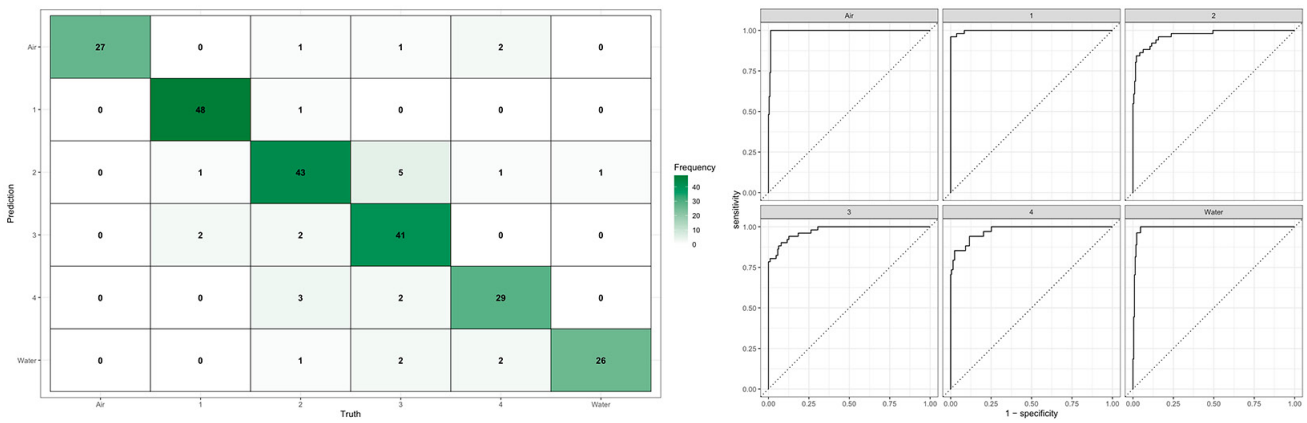

Figure 2. Results of XGBoost model on the test set: confusion matrix (on the left) and ROC curves for each of the classes (on the right)

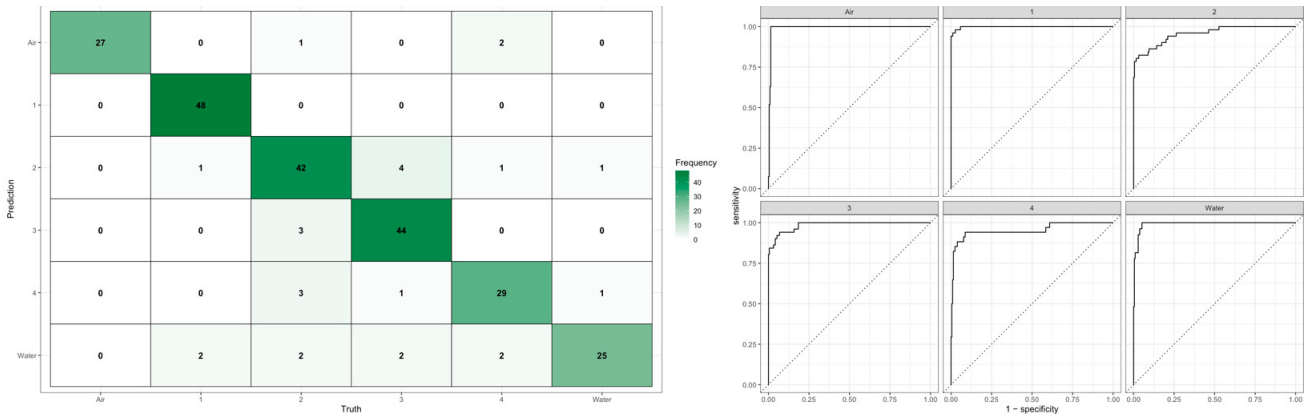

Figure 3. Results of AdaBoost.M1 model on the test set: confusion matrix (on the left) and ROC curves for each of the classes (on the right)

learning models were then trained on the clustered data. In contrast, the study [22] used the *k*-medians method to classify storms at 2 locations. The data were initially transformed using the PCA method, this can therefore be considered a similar procedure to the present study.

In the study [23], the authors used the XGBoost algorithm to predict the value of the α-factor, which accounts for the dependence of oxygen transfer efficiency on water quality parameters in a wastewater treatment plant. Although the model achieved high accuracy, it was discarded for further consideration due to the long inference time. In contrast, in Pakistan, the XGBoost classifier was used in a study [24] to predict Water Quality Class of water samples from field stations monitoring water quality. The data came from IoT sensors examining temperature, pH, turbidity and total dissolved solids. XGBoost was the model with the second best number of correct classifications, right after random forest. The percentage of correct classifications for each class ranged from 88.3% – 92.3%. IoT sensor data was also used to train the XGBoost and AdaBoost models in the study [25] to predict the values of biological oxygen demand and chemical oxygen demand in data sets consisting of observations that are measurements for samples of different types of wastewater. In most cases, in terms of the R^2 coefficient, the AdaBoost model performed better, but not once was it the best of the trained models. The study [26] used the AdaBoost algorithm to classify readings from a matrix of gas sensors. Using the 3-input and 3-output (TITO) technique for obtaining efficient virtual sensor responses, a classification accuracy of 95% was obtained for this model.

In addition to training machine learning models to classify observations into groups denoting the types of samples from which they originated, classical physical and chemical indicators were measured in a river water sample and 4 drainage water samples. These indicators are Conductivity [µS], Total Suspended Solids TSS [mg/l], Chemical Oxygen Demand COD [mg/l], Total Carbon TC [mg/l], Total Nitrogen TN [mg/l], Turbidity [NTU], Inorganic Carbon IC [mg/l] and Dissolved Organic Carbon DOC [mg/l]. Due to the fact that the measurements of all these indicators were made in 3 repetitions for each type of sample, and by not meeting the assumption of normality of distribution, it was necessary to use non-parametric statistical tests. Namely, the Kruskal-Wallis rank-sum test was used for each variable. The test summaries for each of the indicators – test statistics and p-value – are included in Table 2. As can be seen by assuming a significance level of $\alpha = 0.05$ in each test, the null hypothesis of no difference between groups should be rejected.

Table 2. Summary of Kruskal-Wallis test of differences between groups denoting the type of sample for each of the chemical indicators tested

Variable	Statistic	p
Conductivity	13.500	0.009
TSS	13.919	0.008
COD	12.088	0.017
TC	13.233	0.010
TN	11.567	0.021
Turbidity	13.033	0.011
IC	11.033	0.026
DOC	13.500	0.009

The Kruskal-Wallis test does not provide an answer about which groups have statistically significant differences, for this reason the Dunn's multiple comparisons test with Holm's correction was used, a summary of which can be seen in Table 3. Despite obtaining a p-value of 0.026 for the IC index, the Dunn's test did not detect significant differences between groups determined by different sample types. This is because the Kruskal-Wallis test covers all differences and has a much higher sensitivity than the Dunn test, which with Holm's correction controls for the FWER (family-wise error rate). Moreover, the Kruskal-Wallis test compares the distributions of each of the groups, while the Dunn's test checks only whether the medians in the pair of the groups tested is the same. For each of the other physical-chemical indicators, there is exactly one pair-wise comparison in which the corrected p-value is less than 0.05.

Table 3. List of all statistically significant comparisons for every chemical indicator between groups in Dunn's multiple comparisons test with Holm's correction

Variable	Comparison	Z Statistic	p unadjusted	p adjusted
Conductivity	1 – Water	3.286	0.001	0.010
TSS	2 – Water	−3.424	0.001	0.006
COD	3 – Water	−2.832	0.005	0.046
TC	1 – Water	3.195	0.001	0.014
TN	3 – Water	2.830	0.005	0.047
Turbidity	2 – Water	−3.104	0.002	0.019
DOC	1 – Water	3.286	0.001	0.010

Identical information can be read from the graphs in Figures 4 and 5, where the statistically significant multiple

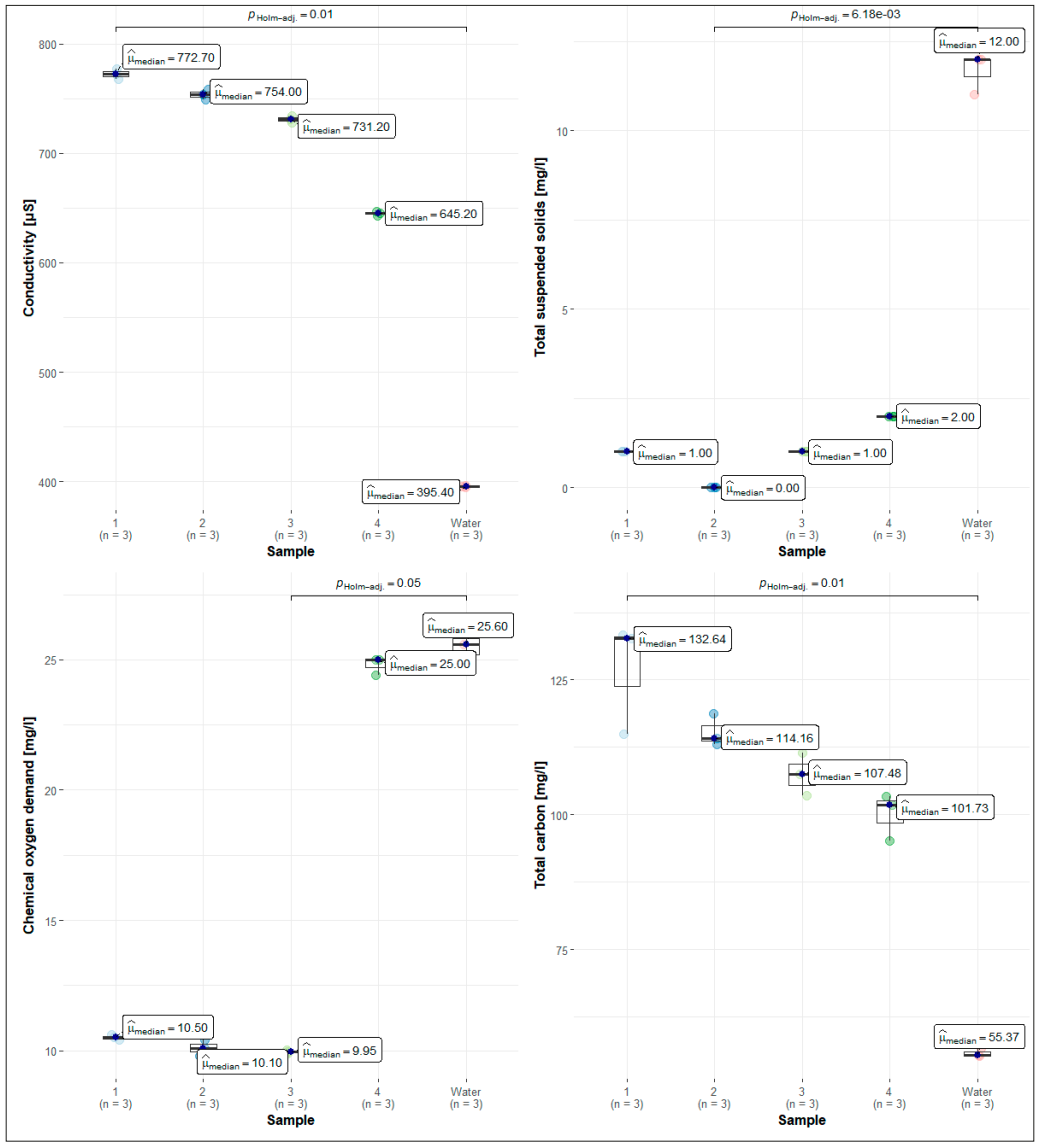

Figure 4. Boxplots of conductivity, TSS, COD and TC in each of the samples' groups. At the top of each graph are marked the statistically significant adjusted p-values in Dunn's test after Holm's correction

comparisons in the Dunn test with Holm's correction and their corrected p-values are marked at the top of each with horizontal dashes. In addition, these images include boxplots of the variability of the measurements of each of the indicators tested as determined by the types of samples from which the observations were taken. In Figure 4, there are graphs plotted for Conductivity, TSS, COD and TC measurements, while Figure 5 includes the graphs plotted for TN, Turbidity, IC and DOC.

The Kruskal-Wallis test was applied previously in study [27] for the analysis of physicochemical parameters, e.g. dissolved organic nitrogen and total nitrogen concentrations in different wastewater samples. In this article *post-hoc* tests were not performed. Whereas in the study [28], the Kruskal-Wallis test with subsequent Dunn's multiple comparisons tests were performed for measurements of electrical conductivity, pH, chloride, sulfate, nitrate, water hardness, turbidity, and colour in samples from the Dunajec River and row wells in southern Poland.

CONCLUSIONS

The t-SNE algorithm is a superior method for visualizing multivariate data from drainage water samples than PCA. However, this is only because information is available that the point clouds in Figure 1 visualized groups of different classes of samples quite well which, in turn, allowed applying the clustering algorithm on 2-dimensional data derived from t-SNE. Thus, it can be concluded that the visualization obtained using a probabilistic method, such as t-SNE, although it turned out to be better, is random even with identical model parameters. For this reason, in the absence of knowledge of the true classes, it can lead to erroneous inferences about the clustering abilities of the data when algorithms such as *k*-medians or *k*-nearest neighbours are applied to them.

Supervised machine learning algorithms based on classification trees, as expected, handled the task of classifying observations from drainage water, river water and synthetic air samples better than the algorithms used in previous research The XGBoost algorithm and AdaBoost.M1

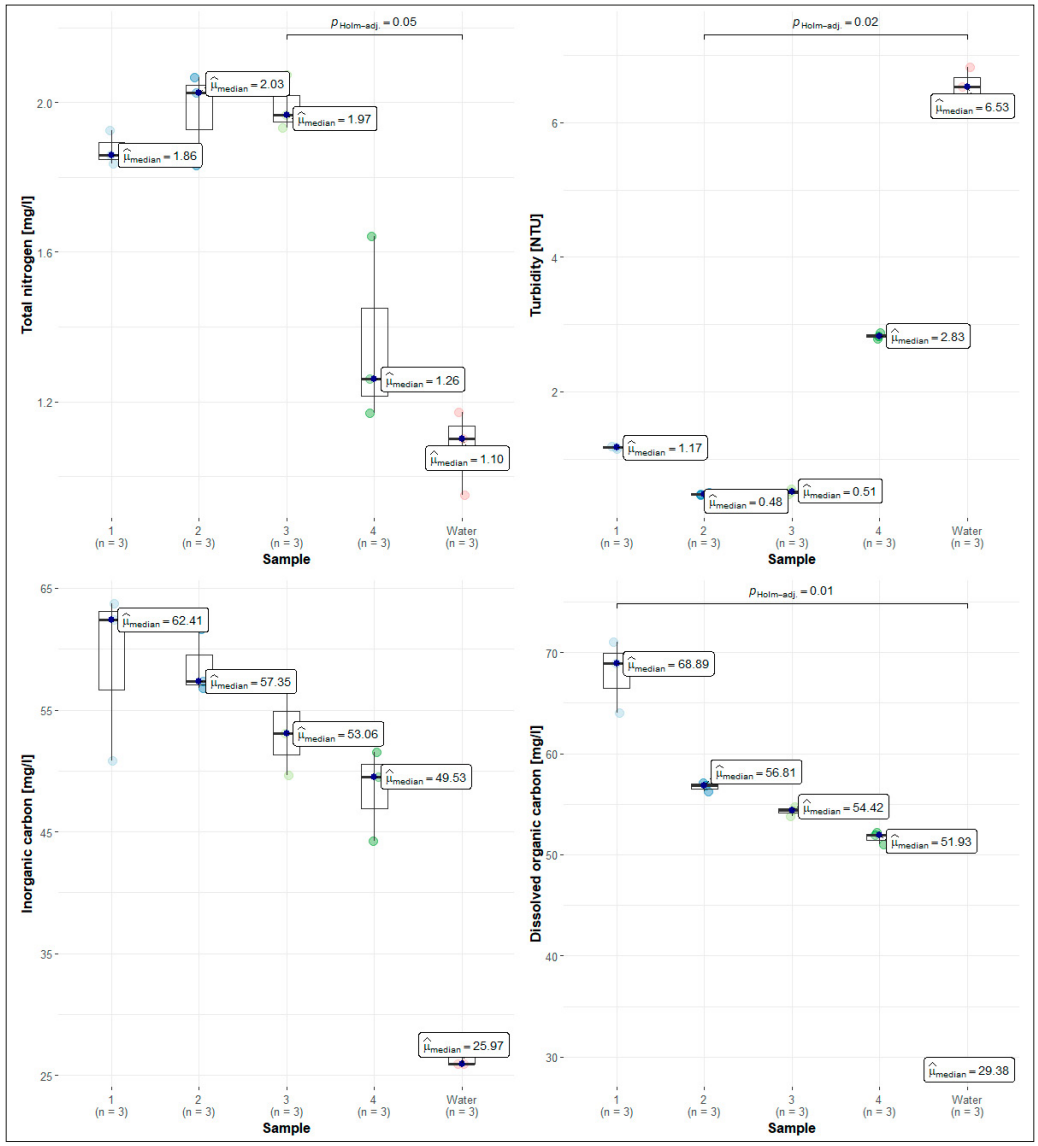

Figure 5. Boxplots of TN, turbidity, IC and DOC in each of the samples' groups. At the top of each graph are marked the statistically significant adjusted p-values in Dunn's test after Holm's correction

achieved 88.8% and 89.2% correct classifications on the test set, respectively, while previously 87.6% for MLP and 84.3% for the random forest model were obtained [6]. Although the current classification is not yet perfect, previous work allows the assumption that ensemble model building and deep learning models could significantly improve the classification quality of these objects.

Statistical analysis of measurements of classical indicators using the Kruskal-Wallis test confirms that there are significant differences in the distributions of different samples classes for each indicator. However, the results of Dunn's multiple comparison tests with Holm's correction show that even in terms of physical and chemical measurements, if there is a statistically significant comparison, it is in a single pair of sample types. These comparisons occur between river water samples and different drainage water samples (Tab. 3). For the IC (Inorganic Carbon) variable, not a single adjusted p-value less than the accepted level of significance occurred. This allows the conclusion that these samples are not easily distinguishable from each other, even in terms of the classic and widely used indicators that allow testing to assess the level of water pollution.

Therefore, it can be concluded that the gas sensor array, together with a set of models that classify observations into the appropriate samples, is not only more effective, but also enables quicker assignment of samples to the appropriate group than in the case of classical methods.

REFERENCES

1. Qin Y, Zhao Q, Zhou D, et al. Application of flash GC e-nose and FT-NIR combined with deep learning algorithm in preventing age fraud and quality evaluation of pericarpium citri reticulatae. Food Chem X. 2024;21:101220. doi:10.1016/j.fochx.2024.101220
2. Paleczek A, Rydosz A. The effect of high ethanol concentration on E-nose response for diabetes detection in exhaled breath: Laboratory studies. Sensors Actuators B Chem. 2024;408:135550. doi:10.1016/j.snb.2024.135550
3. Borowik P, Dyshko V, Tkaczyk M, et al. Analysis of Wheat Grain Infection by Fusarium Mycotoxin-Producing Fungi Using an Electronic Nose, GC-MS, and qPCR. Sensors. 2024;24(2):326. doi:10.3390/s24020326
4. Nam SH, Lee J, Kim E, et al. Electronic tongue and nose sensor coupled with fluorescence spectroscopy to analyze aesthetic water quality

parameters in drinking water distribution system. Process Saf Environ Prot. 2024;188:1201–1210. doi:10.1016/j.psep.2024.05.134
5. Piłat-Rożek M, Łazuka E, Majerek D, et al. Application of Machine Learning Methods for an Analysis of E-Nose Multidimensional Signals in Wastewater Treatment. Sensors. 2023;23(1):487. doi:10.3390/s23010487
6. Piłat-Rożek M, Łagód G. Feasibility of classification of drainage and river water quality using machine learning methods based on multidimensional data from a gas sensor array. Ann Agric Environ Med. 2024;31(4):513–519. doi:10.26444/aaem/196101
7. Raszewski G, Jamka K, Bojar H, et al. Endocrine disrupting micropollutants in water and their effects on human fertility and fecundity. Ann Agric Environ Med. 2022;29(4):477–482. doi:10.26444/aaem/156694
8. van der Maaten L. Accelerating t-SNE using Tree-Based Algorithms. J Mach Learn Res. 2014;15(93):3221–3245.
9. Krijthe JH. Rtsne: T-Distributed Stochastic Neighbor Embedding using a Barnes-Hut Implementation. Published online 2015. https://github.com/jkrijthe/Rtsne (access: 2025.05.19).
10. Godichon-Baggioni A, Surendran S. A penalized criterion for selecting the number of clusters for K-medians. Published online September 8, 2022. doi:10.48550/arXiv.2209.03597
11. Godichon-Baggioni A, Surendran S. Kmedians: K-Medians. CRAN Contrib Packag. Published online September 6, 2022. doi:10.32614/CRAN.package.Kmedians
12. Chen T, Guestrin C. XGBoost: A Scalable Tree Boosting System. In: Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. KDD '16. Association for Computing Machinery; 2016. p. 785–794. doi:10.1145/2939672.2939785
13. Freund Y, Schapire RE. Experiments with a new boosting algorithm. In: Proceedings of the Thirteenth International Conference on International Conference on Machine Learning. Inc., Morgan Kaufmann Publishers; 1996. p. 148–156.
14. Alfaro E, Gamez M, Garcia N. adabag: Applies Multiclass AdaBoost. M1, SAMME and Bagging. CRAN Contrib Packag. Published online June 6, 2006. doi:10.32614/CRAN.package.adabag
15. Kassambara A. rstatix: Pipe-Friendly Framework for Basic Statistical Tests. CRAN Contrib Packag. Published online May 27, 2019. doi:10.32614/CRAN.package.rstatix
16. Ogle DH, Doll JC, Wheeler AP, et al. FSA: Simple Fisheries Stock Assessment Methods. CRAN Contrib Packag. Published online October 8, 2015. doi:10.32614/CRAN.package.FSA
17. R Core Team. R: A Language and Environment for Statistical Computing. Published online 2024. http://www.r-project.org/ (access: 2025.05.19).
18. Chari T, Pachter L. The specious art of single-cell genomics. Papin JA, ed. PLOS Comput Biol. 2023;19(8):e1011288. doi:10.1371/journal.pcbi.1011288
19. Díaz-González L, Rosales-Rivera M, Chávez-Almazán LA. Comprehensive assessment of groundwater quality in Mexico and application of new water classification scheme based on machine learning. Rev Mex Ing Química. 2023;22(2):1–30. doi:10.24275/rmiq/IA235
20. Christiaens L, Orban P, Brouyère S, et al. Tracking the sources and fate of nitrate pollution by combining hydrochemical and isotopic data with a statistical approach. Hydrogeol J. 2023;31(5):1271–1289. doi:10.1007/s10040-023-02646-1
21. Koray AM, Gyimah E, Metwally M, et al. Leveraging machine learning for enhanced reservoir permeability estimation in geothermal hotspots: a case study of the Williston Basin. Geotherm Energy. 2025;13(1):8. doi:10.1186/s40517-024-00323-4
22. Arrueta L, King K, Hanrahan B, et al. The Effect of Alfalfa on Subsurface Discharge and Nutrient Losses Mediated by Precipitation and Antecedent Moisture Conditions. JAWRA J Am Water Resour Assoc. 2025;61(2). doi:10.1111/1752-1688.70018
23. Tenneti S, Divya PD, Tejaswini ESS, et al. Interpretability and performance assessment of advanced machine learning models for α-factor prediction in wastewater treatment plants. J Water Process Eng. 2025;72:107637. doi:10.1016/j.jwpe.2025.107637
24. Rahu MA, Chandio AF, Aurangzeb K, et al. Toward Design of Internet of Things and Machine Learning-Enabled Frameworks for Analysis and Prediction of Water Quality. IEEE Access. 2023;11:101055–101086. doi:10.1109/ACCESS.2023.3315649
25. Soetedjo A, Hendriarianti E, Prasetya RP. Biological Oxygen Demand (BOD) and Chemical Oxygen Demand (COD) measurement of wastewater using Machine Learning regression techniques implemented on the embedded system. Int J Innov Comput Inf Control. 2023;19(5). doi:10.24507/ijicic.19.05.1407
26. Srivastava S, Chaudhri SN, Rajput NS, et al. A novel data-driven technique to produce multi- sensor virtual responses for gas sensor array-based electronic noses. J Electr Eng. 2023;74(2):102–108. doi:10.2478/jee-2023-0013
27. Liao K, You J, Han C, et al. Dissolved organic nitrogen depresses the expected outcome of wastewater treatment upgrading on effluent eutrophication potential mitigation: Molecular mechanistic insight. Water Res. 2024;267:122535. doi:10.1016/j.watres.2024.122535
28. Janik K, Ślósarczyk K, Sitek S. A study of riverbank filtration effectiveness in the Kępa Bogumiłowicka well field, southern Poland. J Hydrol Reg Stud. 2024;53:101834. doi:10.1016/j.ejrh.2024.101834