ORIGINAL ARTICLE

Annals of Agricultural and Environmental Medicine 2024, Vol 31, No 4, 513–519
www.aaem.pl

Feasibility of classification of drainage and river water quality using machine learning methods based on multidimensional data from a gas sensor array

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

[1] *Faculty of Mathematics and Information Technology, Lublin University of Technology, Lublin, Poland*
[2] *Faculty of Environmental Engineering, 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. Feasibility of drainage and river water quality classification 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

Abstract

Objective. The aim of the study is to verify whether the electronic nose system – an array of 17 gas sensors with a signal analysis system – is a useful tool for the classification and preliminary assessment of the quality of drainage water.

Materials and Method. Water samples for analysis were collected in the Park Ludowy (People's Park), located next to the Bystrzyca River, near the city center of Lublin in eastern Poland. Drainage water was sampled at 4 different points. Samples of synthetic air and river water taken from the Bystrzyca River were used for reference. All water samples were tested using an MOS gas sensor array. In order to assess how the e-nose performed in screening and discriminating/preliminarily classifying and grouping samples, their properties were tested using reference methods and assessing surface water quality. The PCA method, Kohonen's SOM with superimposed cluster boundaries by McQuitty's method, random forest and MLP neural network were used to visualize and classify the multivariate data.

Results. The visualization and multidimensionality reduction methods (PCA and SOM) did not enable to clearly distinguish the observations from different drainage water samples. The supervised random forest and MLP methods coped with the classification of samples much better, achieving 84.3% and 87.6% correct classifications on the test set, respectively.

Conclusions. Statistical analysis of the chemical properties of the samples showed that even reference tests are unable to clearly distinguish the samples in terms of a single parameter. However, the e-nose method makes it possible to distinguish these samples from a reference sample derived from river water and a clean air sample.

Keyw ords

surface water, water quality evaluation, pollution indicators, electronic nose, multidimensional data analysis, PCA, Kohonen SOM, ANN MLP, random forest

INTRODUCTION AND OBJECTIVE

Nowadays, increased emissions of harmful substances and the need to monitor them have influenced the intensive development of gas sensors and the emergence of so-called electronic noses (e-noses), i.e. devices for detection a variety of volatile organic compounds [1–3]. They have found application in the cosmetics, food, medicine and petrochemical industries, as well as in wastewater treatment and environmental analysis [4, 5] where e-noses are used alone or in combination with other e-senses, i.e. e-eye and e-tongue [6, 7].

The electronic nose, which consists of gas sensors, hardware and statistical models [8] in the case of drainage water, can be used for screening in the initial analysis of samples and their classification. This analysis enables to quickly check if there is a significant change in the parameters of the sample, and if there is no such change, it allows discontinuation of the testing of chemical indicators of pollution. Testing for chemical contamination properties/indicators is labour-intensive and costly, often generating waste that requires disposal and a skilled person to perform it. In contrast, operation of the e-nose is quick, does not require expenditures on reagents to perform each measurement, nor people with extensive laboratory experience and chemical training.

For the purposes of this study, it was assumed that it is possible to study the quality of the environment, including surface water, which includes water from draining wetlands and swamps that occur in both urban and rural areas, based on electronic sensing readings. In rural areas, drainage systems have often been used to expand farmland. The aim of the study was to verify whether the electronic nose system – an array of 17 gas sensors with a signal analysis system – is a useful tool for the classification and preliminary assessment of the quality of drainage water. This is particularly important due to the fact that drainage water discharged from intensively fertilized agricultural fields can cause eutrophication of surface water. Such waters can also enter livestock watering places at collective drainage ditches located in meadows and pastures. Thus, rapid, inexpensive and efficient detection of situations where the water in drains or the river is significantly different from the accepted level (pattern), which suggests a change in its quality, i.e. an increase in the level of pollution, necessitating taking remedial action and therefore prevent

✉ Address for correspondence: Magdalena Piłat-Rożek, Faculty of Mathematics and Information Technology, University of Technology, Lublin, Poland
E-mail: m.pilat-rozek@pollub.pl

Received: 20.09.2024; accepted: 08.11.2024; first published: 02.12.2024

the degradation of environmental quality and the spreads of animal and human diseases.

MATERIALS AND METHOD

Sampling and data description. Water samples for analysis were collected in Park Ludowy (People's Park) located near the city centre in Lublin, easstern Poland. The location makes this park a component of the city's main ecological corridor, positively affecting the biodiversity of invertebrates as well as birds and small mammals. There is a drainage system in the area of the park, built to eliminate swampland and lower the level of groundwater occurrence. The drainage ditches discharge the collected water into the Bystrzyca River. The drainage water from the Park Ludowy was sampled at four different points – the first was taken behind the dyke, at the point where the drainage flow rises to the surface, and the last at the end of the drainage in the park. The other two samples were taken at equal distances, between the aforementioned locations. Samples of both drainage and river water were taken into glass bottles, which were filled to capacity and capped below the water surface. All water samples were then tested using a MOS gas sensor array. Samples of synthetic air and river water taken from the Bystrzyca River (also collected at the river near the Park Ludowy, five meters before discharge of drainage water from the park) were used for reference.

The measurements were conducted using a gas sensor array comprising 17 Figaro MOS sensors (TGS2600-B00, TGS2602-B00, TGS2610-D00, TGS2611-E00, TGS2620-C00, TGS4161, TGS2444, TGS2442-B02, TGS800, TGS825-A00, TGS813-A00, TGS821, TGS823-A00, TGS812, TGS830, TGS832-A00, TGS2106), outlined in earlier studies and conducted by a team from the Lublin University of Technology in Lublin [9, 10]. The process involved a three-to-five configuration for each port, consisting of three minutes dedicated to flushing the sensors with synthetic air, followed by five minutes for analysis of the mixture.

The data obtained consisted of 1,191 observations with 17 explanatory variables that were measurements from the gas sensor array, and one variable with six levels of Air, 1, 2, 3 and 4, and Water. Air denotes a reference sample of clean air; levels 1 – 4 denote drainage water samples taken in different parts of the Park Ludowy before being discharged into the Bystrzyca River, while Water denotes a sample of river water taken directly from the Bystrzyca River.

Reference tests with conventional methods were carried out using equipment: TOC-LCSH/CSN Shimadzu – TN, IC, TC, TOC (application note of the device); DR 6000, HACH Lange, USA – $N\text{-}NO_3$, $N\text{-}NO_2$, $P\text{-}PO_4$ (LCK tests), $N\text{-}NH_4$ (Hach method 8038), COD (predefined method), TSS (predefined method); Waterproof TN-100 Turbidimeter, EUTECH INSTRUMENTS, Singapore – turbidity (predefined method); CPC-501, ELMETRON, Poland – pH (predefined method); Orion VerseStar Thermo Scientific – conductivity (predefined method).

Elaboration of research results. The Principal Component Analysis method (PCA) was used for dimension reduction and data visualization. Another method used to visualize the data on a two-dimensional plane was a self-organizing Kohonen map, on which the boundaries of the clusters determined by the unsupervised method were superimposed. A random forest and a neural network with a single hidden layer were selected to classify objects by supervised methods in order to increase the quality of classification and compare their accuracy.

The PCA method independently presented by Pearson and Hotelling [11, 12] is a technique that seeks to represent the original data from a matrix X with n variables in a new low-dimensional space of variables. These variables, called principal components, are orthogonal to each other, and are consecutive columns of a matrix Y in the following form

$$Y_i = \gamma_i^T (X - \mu),$$

where $i \in \{1,2, \ldots ,n\}$, μ is the vector of mean variables of the set X, while γ_i is the i-th column of the orthogonal matrix Γ. This matrix is determined by the relation

$$\Lambda = \Gamma^T \Sigma \Gamma, \quad (1)$$

where Σ is the covariance matrix of X, while Λ is the diagonal matrix. If the Σ matrix is positive-definite, then for the eigenvalues of the Λ there is $\lambda_1 \geq \lambda_2 \geq \ldots \lambda_n \geq 0$ [13]. Then, due to relation (1) the eigenvalues of the Λ matrix are also the eigenvalues of the covariance matrix Σ, therefore they represent the percentage of explained variance by the subsequent principal components.

Self-organizing maps have been described by Kohonen [14] and are one type of unsupervised neural network. Such a map is formed by a one- or two-dimensional grid of l neurons with a rectangular or hexagonal topology. At the beginning of the algorithm's operation, neurons are assigned, usually randomly, n-dimensional prototypes (weights) $w_j = [w_{j1}, w_{j2}, \ldots, w_{jn}]$ for $j \in \{1,2,\ldots,l\}$. In the competition process, a winning neuron (Best Matching Unit) is found for observation x from the dataset using the function

$$c(x) = \arg\min_j ||x - w_j||,$$

where $||\,.\,||$ is the selected norm [15]. A neighborhood function is also determined, which is the distance between the i-th neuron and the winning neuron for observation x. The most commonly chosen are a Gaussian function or, as in the case of the current study, a neighborhood bubble function in the form of

$$h_{i,c(x)}(t) = \begin{cases} 1, & ||r_i - r_{c(x)}|| \leq \sigma(t), \\ 0, & \text{otherwise}, \end{cases}$$

where r_k is the position of the k-th neuron measured in the discrete output space [16], and σ is a decreasing function of the neighborhood radius. Then, at step t the neuron weights are updated for the next training steps

$$w_i(t+1) = w_i(t) + \eta(t) \cdot h_{i,c(x)}(t) \cdot (x(t) - w_j(t)),$$

where η is a function of the learning rate [15].

Hierarchical cluster analysis is another unsupervised method used in an agglomerative approach, which relies on the fact that at the beginning, each observation forms a one-element cluster, which successively combine until all the data are in one cluster. There are a number of agglomeration methods that differ in their methods of binding. For this purpose, the Weighted Average linkage McQuitty method

(WPGMA) was applied where, in each step, the 2 closest k and l clusters are combined in $k \cup l$. The distance between cluster $k \cup l$ and m is determined as the average of the distances between clusters k and l, as well as l and m clusters

$$d(k \cup l, m) = \frac{d(k,l) + d(l,m)}{2},$$

where d is the distance in the chosen metric. In this case, the Euclidean distance was used [17].

The random forest presented by Breiman [18], among others, is a supervised machine learning algorithm based on training a fixed number (m) of decision trees. For the k-th tree, where $1 \leq k \leq m$ a p-element sample is drawn with return, based on which the tree is trained. While building the tree, at each split node, v explanatory variables are randomly drawn ($v < n$, where n is the number of predictors), from which the best variable is selected to make the split. In the case of a random forest used to classify objects into appropriate classes, each decision tree classifies observations based on its own splitting rules into one class. The one into which the object under consideration has been classified most often becomes the observation class predicted by the random forest [19].

One-way neural network with one hidden layer is a type of MLP (MultiLayer Perceptron) network with the simplest structure. The network to classify objects into classes has the following structure:
- the input layer receives data from n predictors;
- the hidden layer has a total number (H) of neurons;
- in the output layer there are N neurons, to which the probabilities of belonging to each class are passed.

The probability of each class is predicted as a linear combination of signals from H hidden neurons, which have values in the interval [0,1], which is then transformed using the softmax function. In this way, the values of the probabilities in the output layer add up to 1 and have the following form:

$$f_{il}^{*}(x) = e^{f_{il}(x)} \cdot \left(\sum_{l=1}^{N} e^{f_{il}(x)} \right)^{-1},$$

where $f_{il}(x)$ is the prediction of the probability of occurrence of the l-th class for i-th observation before applying the transformation function. Optimization of the parameter coefficients in such a network is done by minimizing the following equation:

$$\sum_{l=1}^{N} \sum_{i=1}^{s} (y_{il} - f_{il}^{*}(x))^2,$$

where s is the sample size, while y_{il} is the index of the occurrence of the l-th class for i-th observation [20].

Analysis, data processing and plotting were performed in the statistical computing system R version 4.4.0 [21]. Using the kohonen package [22], Kohonen's self-organizing map was trained and plotted, using the functions of the tidymodels library [23], supervised models were trained, and other graphs developed using the tidyverse package [24].

RESULTS AND DISCUSSION

Figure 1 shows a visualization of the multivariate data obtained from the gas sensor array using the PCA method on a two-dimensional and three-dimensional plane. Observations from different samples are marked with different colours in the graph. The first two principal components explain 47% of the variance in the data set, while the first three explain 53.6%. The ranges of variation of the variables denoting the first three principal components are small. In the two-dimensional graph, however, one can see a prominent separated cluster of a large number of observations from the reference air sample (marked in green), with single observations from the drainage water samples. In addition, a cluster of Sample 1 taken from the drainages is separated in the lower right corner of the graph. The cluster of observations from the air samples and the point clouds from the drainage water and river water samples are located on opposite sides of the graph. In contrast, the three-dimensional graph shows that the observations from the different types of samples are arranged in vertical stripes. The data points from the drains are located in the point cloud between the clusters of clean air and river water observations.

Adding more principal components, although impossible to visualize, will not be optimal. This fact can be seen in the scatter plot in Figure 2. It can be seen that from the third principal component onward, the percentage of explained variance by subsequent components remains almost the same. Another drop occurs at the tenth component, which is an insufficient reduction in dimensionality for 17 explanatory variables.

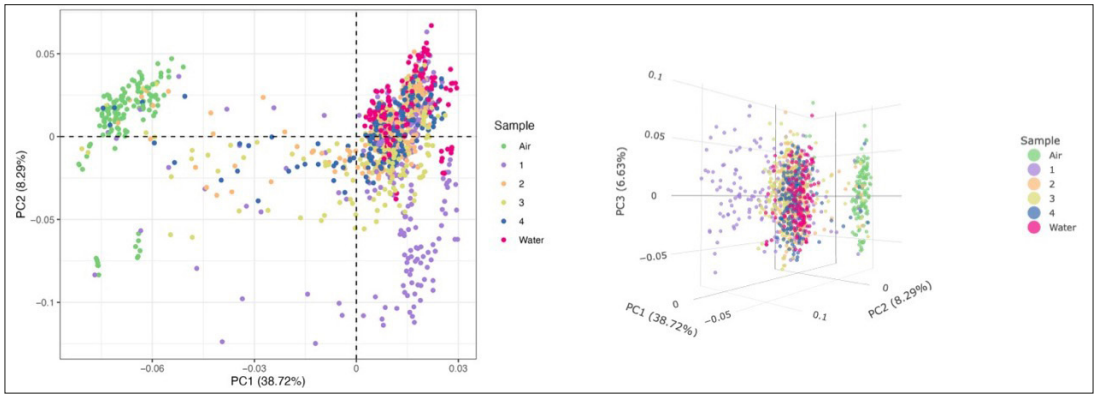

Figure 1. Data visualization using the PCA method in two-dimensional (left) and three-dimensional (right) space

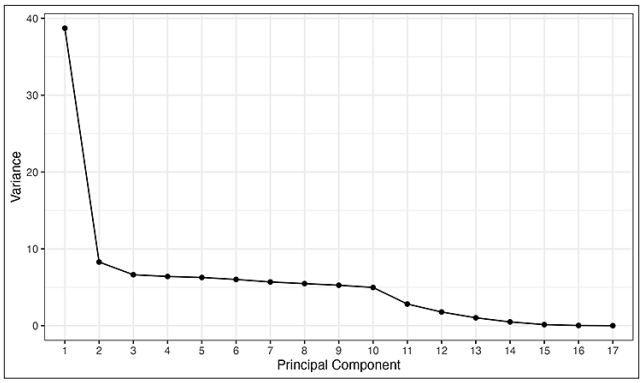

Figure 2. Scree plot showing the percentage of explained variance by each principal component of the PCA method

The PCA method was used to represent the considered observations on a plane or low-dimensional space, while the self-organizing Kohonen map was used to visually represent their similarity by means of their distribution in neurons. Figure 3 shows a visualization of a self-organizing Kohonen map of 15x15 neurons in a hexagonal topology, with bubble neighbourhood function. The figure on the left indicates the assignment of the original points to each neuron of the map. Again, it can be seen that the observations from the clean air sample cluster in the neurons located in the lower left corner of the map. The neurons in the upper left corner of the map contained a large group of observations originating from the first drainage water sample. The remaining observations do not group in clusters on the Kohonen map that can be distinguished by sample type. An unsupervised method was used to assign observations to clusters that are as homogeneous as possible within a group. The figure to the right shows the boundaries of the six clusters found using hierarchical cluster analysis using the McQuitty method (WPMA). Data were assigned to six clusters to represent sample types. Thus, it can be seen that the largest, but also the most diverse clusters are four and five, which contain observations from drainage water samples. The smallest clusters are numbered 1 and 3, but they do not contain samples that belong to a single type. The most homogeneous in terms of sample type are cluster 2, whose neurons contain observations from clean air samples, and cluster 6 with observations from drainage water sample 1.

The unsupervised method, when adopting the division of the set into 6 clusters, is not able to extract heterogeneous clusters, which can also be seen when applying the same method for a smaller number of clusters in Figure 4. This is due to the fact that many neurons contain more than one type of sample. When the number of groups is reduced, the largest, most heterogeneous clusters merge into one.

Since unsupervised methods did not yield satisfactory results and did not allow dividing observations into samples heterogeneous with respect to water quality, supervised methods were used. For this purpose, the analyzed data were divided into a training set and a test set in a ratio of 3:1. The selection of hyperparameters of the models was performed by means of a five-fold cross-check.

The random forest model was trained using 500 decision trees, the *mtry* and *min_n* parameters were selected based on the tuning of hyperparameters. The *mtry* parameter denotes the number of random predictors at each node, while *min_n* corresponds to the minimum number of observations that are required for further node partitioning. The grid of tested hyperparameters consisted of the values rearranged in Table 1.

Table 1. Results of random forest hyperparameters tuning

Parameter	Tested values	Best
mtry	1, 5, 10, 11, 15	5
min_n	7, 15, 18, 26, 39	7

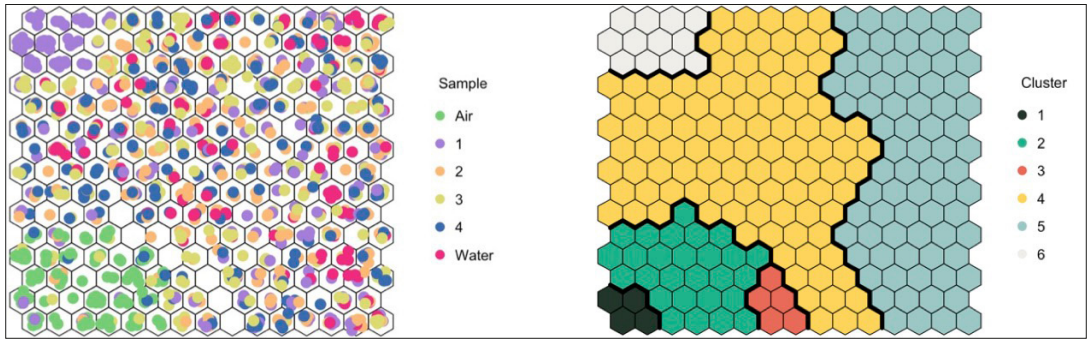

Figure 3. Self-organizing Kohonen map – mapped points from different samples (left), boundaries of 6-clusters determined by McQuitty method superimposed on the map (right)

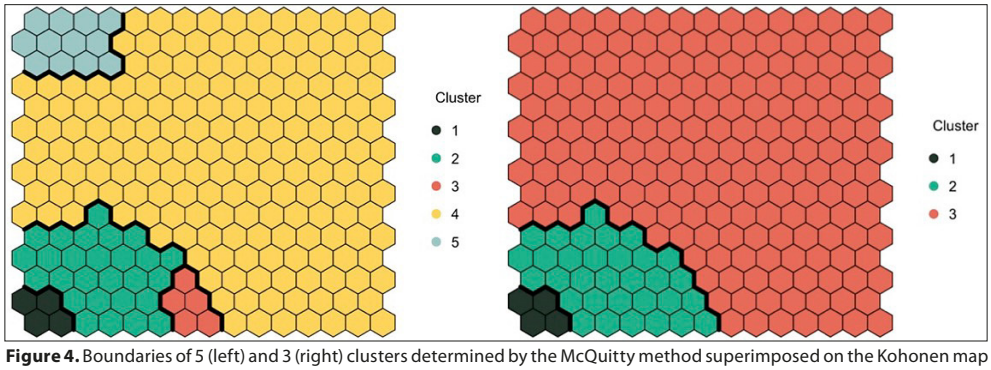

Figure 4. Boundaries of 5 (left) and 3 (right) clusters determined by the McQuitty method superimposed on the Kohonen map

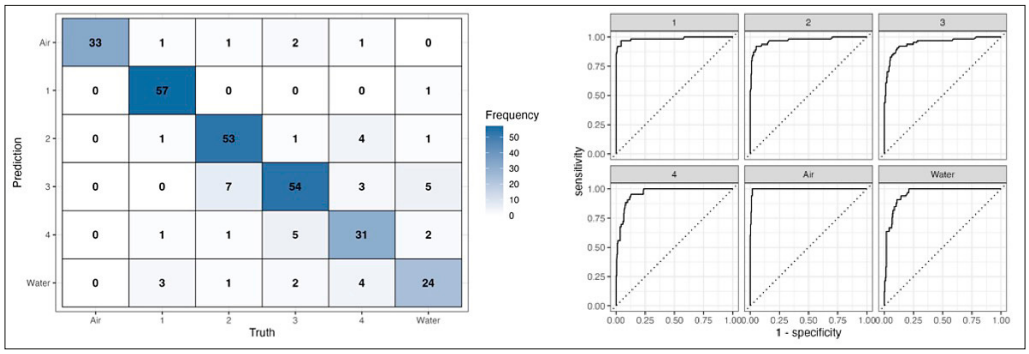

Figure 5. Contingency matrix (left) and ROC-AUC curves (right) for the test data in the random forest model

The best of the models turned out to be a random forest trained with *mtry* = 5 and *min_n* = 7 parameters. Figure 5 shows a contingency table comparing the true and model-predicted labels, as well as ROC-AUC curves for each level of the explanatory variable on the test set. In both graphs it can be seen that the model best classifies the observations from the clean air sample, as each of the observations was classified into this type of sample. The observations from sample 1 were also mostly matched to the correct class, while the model did worst in classifying the river water sample and the four drainage water samples. The random forest model on the test set achieved 84.3% of correct classifications.

Another model trained was a one-way MLP neural network with one hidden layer. The tuned parameters included *hidden_units*, *epochs* and *penalty*. The average accuracy results with 5-fold cross-validation for each of the tested sets of hyperparameters are shown in Figure 6. The best average classification accuracy was achieved for 20 neurons in the hidden layer, 300 epochs of network training and regularization at the 0.01 level.

Figure 7 shows the contingency matrix of the true and model-predicted values of the sample variable for the test data, as well as the ROC-AUC curves for each level of this variable. Again, as in the case of the random forest, the neural network model performed well in classifying observations from the clean air sample and Sample 1 of drainage water. Despite the errors in classifying river water, there is a noticeable improvement in the quality of classification for the test data, as the model achieved an accuracy of 87.6%.

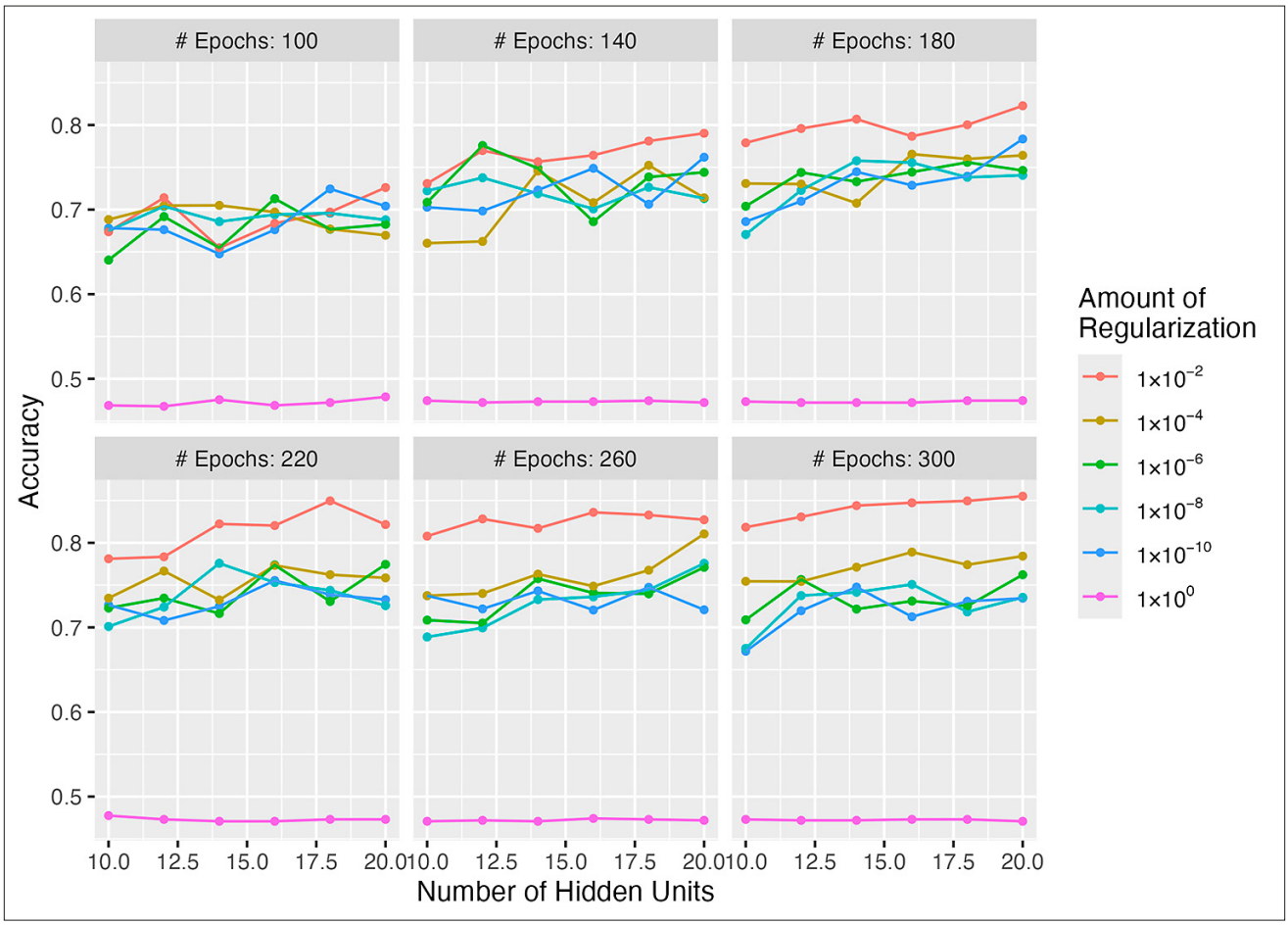

Figure 6. Accuracy of the single-layer neural network when tuning its hyperparameters on the learning set

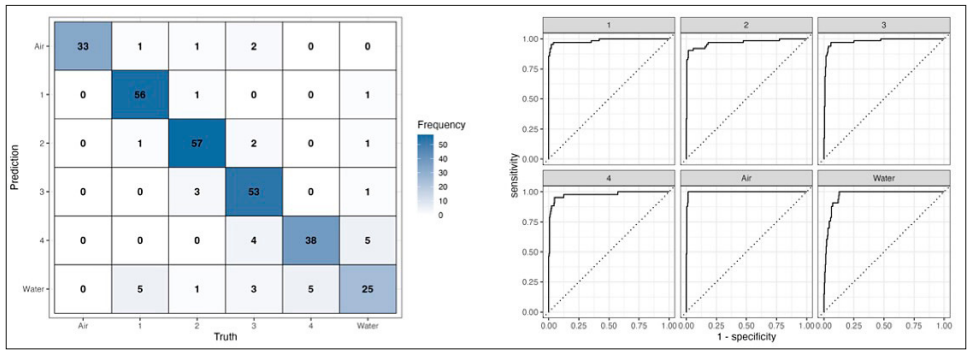

Figure 7. Contingency matrix (left) and ROC-AUC curves (right) for the test data in the single-layer neural network model

In order to assess how the e-nose performed in screening and discriminating/preliminarily classifying and grouping samples, their properties were tested using reference methods used for analyzing and assessing surface water quality – conductivity [μ*S*], pH, content of P-PO_4, N-NH_4, N-NO_3, N-NO_2, [mg/l], total suspended solids [mg/l], total nitrogen TN [mg/l], total carbon TC [mg/l], inorganic carbon IC [mg/l], turbidity [NTU], chemical oxygen demand COD [mg/l] and total organic carbon TOC [mg/l]. The results obtained [25] showed that there was a high correlation between the readings from the gas sensor arrays and some of the chemical parameters studied, i.e. COD, pH, N-NO_3, N-NO_2. However, the situation in the cited study was simple insofar as the differences in the level of contamination of the samples tested were large.

Analyzing the values of pollution indicators, it can be concluded that the COD parameter values oscillated within 25 [mg/l], being lower than the averages for a river of this type [26]. The COD parameter values of drainage waters were much lower than for rainwater, within the range of 10–25 [mg/l] [26]. Thus, it can be concluded that there was an elevated COD value in the river compared to the drainage water. In contrast, the opposite was observed for the carbon content TC, with higher contents in the drainage water samples compared to the river. Total suspended solids and turbidity were highest in the river samples.

A non-parametric Kruskal-Wallis test comparing the distributions of the variables under study for each type of sample yielded a p-value of 0.009. This means that the hypothesis that the multivariate distributions are identical for all types of samples was rejected. On the other hand, Dunn's paired difference tests, as amended by Bonferonni, showed that there were significant differences between:

- Samples 1 and 3 in the distribution of N-NH_4 measurements (p-value=0.008);
- Samples 3 and 4 in the distribution of N-NO3 measurements (p-value=0.008);
- Sample 1 and river water in the distribution of measurements of conductivity (p-value=0.008), TC (p-value=0.014) and TOC (p-value=0.010);
- Sample 2 and river water in the distributions of measurements of P-PO_4 (p-value=0.008), suspended solids (p-value=0.006) and turbidity (p-value=0.008);
- Sample 3 and river water in the distributions of measurements of pH (p-value=0.008), TN (p-value=0.04), N-NO_2 (p-value=0.008), COD (p-value=0.008).

Thus, only in the case of the amount of inorganic carbon (IC) was there no basis for rejecting the hypothesis that there were no differences between groups determined by sample type. For the other parameters, there were differences, but for each of the variables there were differences only for one pair of sample types.

CONCLUSIONS

On the basis of the analysis, it can be concluded that the visualization and multidimensionality reduction methods did not enable to clearly distinguish observations from different drainage water samples. Also, the distribution of points in the neurons of the Kohonen map and the distinction of boundaries on the map established by hierarchical cluster analysis using McQuitty's method, did not allow distinguishing heterogeneous clusters. The supervised random forest and MLP methods coped with the classification of samples much better, achieving 84.3% and 87.6% correct classifications on the test set, respectively. Across all sample types, correct classifications were, respectively: 90.5% and 88.9% for Sample 1; 84.1% and 90.5% for Sample 2; 84.4% and 82.8% for Sample 3; 72.1% and 88.4% for Sample 4; 72.7% and 75.8% for the river water sample; 100% for the clean air sample for both models.

Statistical analysis of measurements pertaining to the chemical properties of the samples enabled the conclusion that even time- and cost-consuming reference tests are unable to distinguish samples in terms of a single parameter. Thus, although the unsupervised methods did not provide a basis for a clear division of the collection into intra-group homogeneous clusters, with the help of the supervised methods used, it was possible to achieve a level of accuracy that allowed the initial differentiation of samples from drainage waters. The method also makes it possible to distinguish these samples from a reference sample derived from river water and a clean air sample. Therefore, the use of an electronic nose can improve the monitoring of surface runoff from drainage waters and the surface waters of the receiver, which in the case described were river waters. This method can be used in practice for screening the quality of the aquatic environment, e.g. changes over time at the same point or at the same time at neighboring points.

The next step will be the search and selection of more efficient clustering and classification methods and, in parallel, how these described methods perform for cases of other types of contaminated water.

REFERENCES

1. Lvova L, Di Natale C, Paolesse R. Chemical Sensors for Water Potability Assessment. In: Bottled and Packaged Water. Elsevier; 2019:177–208. doi:10.1016/B978-0-12-815272-0.00007-6
2. Arroyo P, Meléndez F, Suárez JI, Herrero JL, Rodríguez S, Lozano J. Electronic Nose with Digital Gas Sensors Connected via Bluetooth to a Smartphone for Air Quality Measurements. Sensors. 2020;20(3):786. doi:10.3390/s20030786
3. Herrero JL, Lozano J, Santos JP, Suárez JI. On-line classification of pollutants in water using wireless portable electronic noses. Chemosphere. 2016;152:107–116. doi:10.1016/j.chemosphere.2016.02.106
4. Ye Z, Li Y, Jin R, Li Q. Toward Accurate Odor Identification and Effective Feature Learning With an AI-Empowered Electronic Nose. IEEE Internet Things J. 2024;11(3):4735–4746. doi:10.1109/JIOT.2023.3299555
5. Zhang Y, Askim JR, Zhong W, Orlean P, Suslick KS. Identification of pathogenic fungi with an optoelectronic nose. Analyst. 2014;139(8):1922–1928. doi:10.1039/C3AN02112B
6. Savio S, di Natale C, Paolesse R, Lvova L, Congestri R. Keeping Track of Phaeodactylum tricornutum (Bacillariophyta) Culture Contamination by Potentiometric E-Tongue. Sensors. 2021;21(12):4052. doi:10.3390/s21124052
7. Jamka K, Wróblewska-Łuczka P, Adamczuk P, Zawadzki P, Bojar H, Raszewski G. Methodology for preparing a cosmetic sample for the development of Microorganism Detection System (SDM) software and artificial intelligence learning to recognize specific microbial species. Ann Agric Environ Med. 2021;28(4):681–685. doi:10.26444/aaem/144696
8. Persaud K, Dodd G. Analysis of discrimination mechanisms in the mammalian olfactory system using a model nose. Nat 1982 2995881. 1982;299(5881):352–355. doi:10.1038/299352a0
9. Piłat-Rożek M, Łazuka E, Majerek D, Szeląg B, Duda-Saternus S, Łagód G. 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
10. Łagód G, Duda SM, Majerek D, Szutt A, Dołhańczuk-Śródka A. Application of Electronic Nose for Evaluation of Wastewater Treatment Process Effects at Full-Scale WWTP. Processes. 2019;7(5):251. doi:10.3390/pr7050251
11. F.R.S KP. LIII. On lines and planes of closest fit to systems of points in space. London, Edinburgh, Dublin Philos Mag J Sci. 1901;2(11):559–572. doi:10.1080/14786440109462720
12. Hotelling H. Analysis of a complex of statistical variables into principal components. J Educ Psychol. 1933;24(7):498–520. doi:10.1037/h0070888
13. Mardia KV, Kent T, Bibby J. Multivariate Analysis. Academic Press Limited; 1979.
14. Kohonen T. Self-organized formation of topologically correct feature maps. Biol Cybern. 1982;43(1):59–69. doi:10.1007/BF00337288
15. Haykin S. Neural Networks and Learning Machines. Pearson Education; 2009.
16. Ponmalai R, Kamath C. Self-Organizing Maps and Their Applications to Data Analysis; 2019. doi:10.2172/1566795
17. Everitt BS, Landau S, Leese M, Stahl D. Hierarchical Clustering. In: Cluster Analysis, 5th Edition. John Wiley & Sons; 2011. pp. 71–110. doi:10.1002/9780470977811.ch4
18. Breiman L. Using adaptive bagging to debias regressions. Stat Dept UCB. 1999;547.
19. Kuhn M, Johnson K. Regression Trees and Rule-Based Models. In: Applied Predictive Modeling. New York: Springer; 2013. pp. 173–220. doi:10.1007/978-1-4614-6849-3_8
20. Kuhn M, Johnson K. Nonlinear Classification Models. In: Applied Predictive Modeling. New York: Springer; 2013. pp. 329–367. doi:10.1007/978-1-4614-6849-3_13
21. R Core Team. R: A Language and Environment for Statistical Computing. Published online 2024. http://www.r-project.org/
22. Wehrens R, Buydens LMC. Self- and Super-organizing Maps in R: The kohonen Package. J Stat Softw. 2007;21(5). doi:10.18637/jss.v021.i05
23. Kuhn M, Wickham H. Tidymodels: a collection of packages for modeling and machine learning using tidyverse principles. Published online 2020. https://www.tidymodels.org
24. Wickham H, Averick M, Bryan J, et al. Welcome to the Tidyverse. J Open Source Softw. 2019;4(43):1686. doi:10.21105/joss.01686
25. Guz Ł, Łagód G, Jaromin-Gleń K, Suchorab Z, Sobczuk H, Bieganowski A. Application of gas sensor arrays in assessment of wastewater purification effects. Sensors. 2015;15(1):1–21. doi:10.3390/S150100001
26. Babko R, Szulżyk-Cieplak J, Danko Y, Duda S, Kirichenko-Babko M, Łagód G. Effect of Stormwater System on the Receiver. J Ecol Eng. 2019;20(6):52–59. doi:10.12911/22998993/109433