AI-aided e-nose calibration

A Malec¹, Z Suchorab¹, L Guz¹ and C Caserta²

¹ Lublin University of Technology, Faculty of Environmental Engineering, Nadbystrzycka 40B, 20-618 Lublin, Poland
² Mediterranean University of Reggio Calabria, Via dell'Università, 25, 89124 Reggio Calabria, Italy

aga_malec@interia.pl

Abstract. The development of the semiconductor technology and computer methods contributes to the improvements in measurement techniques of air pollution. Besides traditional techniques of air pollution measurement, such as chromatography or dynamic olfactometry, the electronic nose (EN) is attracting an increasing amount of attention in this area. EN is a device which imitates the human sense of smell. The main component of the device is an array of many non-specific gas sensors. The aim of this article is to present the result of calibration of EN composed of MOS (metal-oxide semiconductor) sensor with the use of the artificial neural networks. For data analysis, linear and artificial intelligence algorithms were employed, whose accuracy was compared.

1. Introduction

An artificial neural network is a general name for mathematical structures and their software or hardware models that perform calculations or signal processing through rows of processing elements. These elements are known as artificial neurons and they perform certain basic operations at their input. The inspiration for such a structure was the construction of natural neurons, the synapses linking them, and the nervous system, particularly the brain. The common feature of all neural networks is that their structure consists of neurons connected by synapses. There are weights connected with synapses, which are numerical values whose interpretation depends on the model [1, 2].

The mathematical model describing the principle of information processing is based on the determination of the block that weights the input signals [3, 22], which resembles the biological synaptic connection. Low-capacity dendrites are represented in the model by low-capacity scales. The summing element sums up the weighted signals. If the sum of weighed signals is greater than the activation threshold set at the input, the signal is transferred by activation functions [4, 15]. As a result of this operation, we obtain an output signal which corresponds to the signal flowing through the cellular axon.

Artificial Neural Networks (ANNs) are relatively new computational tools which have been widely used to solve many complex problems in the real world. The attractiveness of ANNs lies in their information processing characteristics, mainly related to nonlinearity, high parallelism, resistance to errors and noise, and learning and generalisation capabilities [5, 20].

In order for the network to function properly, it must go through a learning process [6, 24]. Teaching a neural network consists in selecting weights so that the output signal is within the limits of the accepted error with respect to the expected signal. The object to be taught in such a network is a matrix, the values of which should be selected so that the network realises the assumed form of mapping the input signals
to the output signals. Network training can be carried out with or without supervision by an algorithm. One of the teaching algorithms is the backpropagation (BP) algorithm of error, which consists in the fact that having determined the error occurring during the implementation of this step of the learning process in the neuron with the number of \( m \) - we can “project” this error back to all those neurons whose signals were inputs for the \( m \)-th neuron. This algorithm is used in multi-layer networks with the so-called hidden layers [7].

Considering the cases of such networks, it can be concluded that in many of them, with simple input-output relationships, one layer of hidden neurons is sufficient.

In other cases, the use of additional hidden layers might be necessary. There is a general suggestion from Nielsen [8, 25] that the number of hidden neurons should be less than \( 2n + 1 \), where \( n \) is the number of inputs. A network with a small number of hidden neurons does not allow the modelling of complex input-output data, but on the other hand networks with many neurons in hidden layers have weak generalising capabilities [9, 23].

The neural networks have been used in many areas of life, due to the constant development of the field of knowledge they build. So far, they have found application in: electronic system diagnostics, psychiatric research, stock market forecasts, sales forecasts, oil exploration, interpretation of biological and medical research, price forecasts, spectral analysis, waste disposal optimisation, selection of raw materials, selection of investigation targets in forensics and industrial process control [10, 21].

Currently, much attention is paid in the scientific literature to the identification of samples emitting volatile air pollutants by means of different sensor matrices, for example for the identification of gaseous pollutants, poor quality food products or cereals infected with fungi [11, 12, 13]. Despite its usefulness, ANN also has some disadvantages. Network learning usually requires large amounts of data. The output information is not very accurate, since they are only approximations, so it is not suitable for precise engineering calculations. From the practical point of view, networks are treated as "black boxes" because it is difficult to trace the transformation of the input information into a useful output result. For example, they were used to identify pollutants from the group: four [2, 5, 14], five [7, 27] or eight compounds [14, 19]. Due to the limited number of components, the effectiveness of identification was high.

2. Materials and methods

Sensors
The device used in tests was a gas sensor array consisting of 17 MOS (metal oxides semiconductor) sensors manufactured by TGS Figaro producer:

- TGS-2600-B00 – general air pollution sensor 1-30 ppm (for \( H_2 \))
- TGS-2602-B00 – toxic air pollutants (\( H_2S \), \( CH_4 \), \( CH_3OH \), \( C_6H_5-CH_3 \)) 1-30 ppm (for \( CH_3OH \))
- TGS-2610-D00 – propane and butane sensor with carbon filter (\( CH_4 \), \( C_4H_{10} \) 500-10000 ppm
- TGS-2611-E00 – methane sensor with carbon filter (\( CH_4 \) 500-10000 ppm
- TGS-2620-C00 – ethanol sensor, 50-5000 ppm
- TGS-4161 – carbon monoxide 350-10000 ppm
- TGS-2444 – ammonia 10-100 ppm
- TGS-2442-B02 – carbon monoxide sensor CO 30-1000 ppm
- TGS-800 – general air pollution 1-1000 ppm (for \( C_4H_{10} \))
- TGS-825-A00 – hydrogen sulphide 5-100 ppm
- TGS-813-A00 – flammable gas sensor 500-10000 ppm
- TGS-821 – hydrogen 10-5000 ppm
- TGS-823-A00 – solvent vapours 50-5000 ppm
- TGS-812 – sensor for explosive and toxic gases 100-5000 ppm (for \( CH_4 \), \( C_4H_{10} \), \( H_2 \))
- TGS-830 – chlorofluorocarbons 100-3000 ppm
- TGS-832-A00 – chlorofluorocarbons 100-3000 ppm
- TGS-2106 – exhaust gas sensor, diesel exhaust gas sensor 0.1-10 ppm (for \( NO_2 \)).
The general scheme of the stand is presented in figure 1. Switching between synthetic air and the connected sample took place automatically according to the set protocol.

![Diagram of the measuring station](image)

**Figure 1.** General diagram of the measuring station: 1) sensor matrix, 2) flow meter, 3) micro membrane pump, 4) switchover valve, 5) control valve, 6) reducer, 8) air cylinder, 7) bag.

Three measurements were taken for each sample of a specific concentration. A single measurement lasted 90 seconds and consisted of a 420-second phase of clean air rinsing and 480-second measurement. The flush of airflow and sample collection were set to 0.5 l/min. An example of voltage change for the TGS-800 sensor during butanol measurements is shown in figure 2.

![Graph of voltage change](image)

**Figure 2.** Voltage change for TGS-800 sensor during 1-butanol measurements.

**Samples**
Specific doses of chemical compounds: n-1-butanol, acetone, benzaldehyde, dimethylamine, toluene were dosed with a chromatographic syringe (10 μL) into a 16-litre Tedlar bag with synthetic air, with a hydrocarbon content of less than 0.1 ppm. The doses were respectively: 0 (blank samples), 1, 2, 4 and 8 μL. After evaporation of the dosed substance, the bag was connected to the sampling port of the e-nose. The detected concentrations are given in the table below.
Table 1. Concentrations of chemical compounds.

| Compound       | Dispensed volume (μl) | Concentration (ppm) |
|----------------|-----------------------|---------------------|
| benzaldehyde   | 1                     | 0.0666              |
|                | 2                     | 0.1341              |
|                | 4                     | 0.2682              |
|                | 8                     | 0.5366              |
| acetone        | 1                     | 0.0608              |
|                | 2                     | 0.1341              |
|                | 4                     | 0.2430              |
|                | 8                     | 0.4860              |
| toluene        | 1                     | 0.0671              |
|                | 2                     | 0.1341              |
|                | 4                     | 0.2683              |
|                | 8                     | 0.5366              |
| dimethylamine  | 1                     | 0.0671              |
|                | 2                     | 0.1342              |
|                | 4                     | 0.2684              |
|                | 8                     | 0.5367              |
| 1-butanol      | 1                     | 0.0671              |
|                | 2                     | 0.1341              |
|                | 4                     | 0.2683              |
|                | 8                     | 0.5365              |

Analysis

Strict results of measurements obtained with the use of electronic nose contain signals from individual sensors. Finding important relationships in such multidimensional sets is not easy because of their multidimensionality and often redundancy of information. This requires the use of appropriate statistical tools which, for instance, reduce the multidimensionality of the data set, such as the analysis of the main components, the results of which must be further analysed. Artificial neural networks, on the other hand, immediately provide useful information about the tested sample. For this reason, artificial neural networks were used to analyse the data presented in this paper.

The statistical package Statistica 13.1 was used for the analysis. Individual variables (results from sensors) were standardised according to their dependence: \( s = (S - \bar{S}) / \delta \) where: \( s \) - standard value, \( S \) - raw value, \( \bar{S} \) - mean value of the sensor, \( \delta \) - standard deviation.

PCA analysis based on covariance matrix was used for initial visualisation. The number of main components was limited to two.

For further analysis, (ANN) artificial neural networks were used. First of all, supervised classification networks were used to identify chemical compounds. Five different MLP (Multilayer Perceptron) networks were determined (table 2), each with five outputs, corresponding to particular compounds.

Afterwards, upon the chemical compound detection, subsequent neural networks were used in the tested sample to estimate the concentration of the gas sample. For each chemical compound, five different regression neural networks were also determined. The use of several neural networks in each application was intended to exclude the possibility of obtaining random results of the analysis.

In the Statistica software, the Automated Neural Networks module was used. For neural network analysis, all measurements from the range of 800-900 sec. were selected. The data was randomly divided into a training subset (75%), a test subset (15%) and a validation subset (15%). These samples are used to 1) train the network, 2) verify (test) the performance of the networks while under training, and 3) perform final validation test to determine how well the network predicts “new” data that was neither used to train the model or to test its performance when trained. The networks used had one layer of hidden neurons with the minimum and the maximum number of neurons, respectively 17 and 34. With fewer neurons, network error increases, and with a much larger number of neurons, networks lose their ability to generalise. All the neural networks used had 17 inputs. As the training algorithm BFGS
(Broyden–Fletcher–Goldfarb–Shanno) was used. For error evaluation Entropy and SOS (Sum of Squares) functions were applied.

3. Results

PCA analysis

The cumulative value of the first two main components is 83.19%. The use of subsequent main components makes a much smaller contribution to the description of the variance of the data set, which is visible on the landfill slope diagram (figure 3a).

Figure 3. Analysis of the obtained data using PCA method. a) change in percentage contribution to the description of the total variance of the dataset depending on the number of main components; b) PCA diagram of the analysed compounds. On the left: n–1-butanol, Ac–acetone Be–benzaldehyde, Di–dimethylamine, To–toluene. The number on the right denotes dose.

The two-dimensional PCA analysis graph is shown in figure 3b. The graph shows the average values from the standardised data. The separation of individual chemicals is visible. The points representing individual concentrations tend to be arranged along the line, and ordered according to increasing concentration. Measuring points for clean air are grouped in the right part of the graph. The lowest regression coefficient was observed for benzaldehyde.

Identification of chemical compounds using neural networks

The data computed by means of the artificial neural networks for detection of the particular chemicals, regardless of sample concentration are presented in table 2.

Table 2. Neural networks applied for compound identification.

| Net. name | Training accuracy (%) | Test accuracy (%) | Validation accuracy (%) | Training algorithm | Error function | Hidden activation | Output activation |
|-----------|-----------------------|-------------------|-------------------------|--------------------|---------------|------------------|------------------|
| MLP 17-29-5 | 97.8716 | 95.9142 | 88.2410 | BFGS 21 | Entropy | Logistic | Softmax |
| MLP 17-31-5 | 94.2857 | 87.6857 | 81.5477 | BFGS 83 | SOS | Exponential | Exponential |
| MLP 17-34-5 | 99.2764 | 92.3271 | 86.7874 | BFGS 22 | Entropy | Linear | Softmax |
| MLP 17-32-5 | 92.2044 | 90.3603 | 89.4567 | BFGS 20 | Entropy | Linear | Softmax |
| MLP 17-21-5 | 95.9478 | 89.2315 | 82.9853 | BFGS 28 | Entropy | Logistic | Softmax |

The graph below shows the result of the identification of chemical compounds. On the cut-off axis, there are chemical compounds that have been measured with e-nose and then classified with ANN.

Neural network responses are indicated on the ordinate axis. 1-butanol was correctly identified as 1-butanol, but also as acetone and dimethylamine. All the networks correctly identified benzaldehyde, despite the fact that the PCA analysis graph showed the highest data scattering for a given chemical compound.
Evaluation of particular compounds concentration using neural networks

The potential of multisensory arrays to evaluate particular compounds concentrations (1-butanol, acetone, benzaldehyde, dimethylamine, toluene) is presented in the tables and diagrams below. The tested chemical compounds were selected because of their characteristic, perceptible odour. In addition, 1-butanol is the reference substance in dynamic olfactometry (EN-13725).

Table 3. Identification of tested substances concentration using particular networks.

| Net. name | Training accuracy | Test accuracy | Validation accuracy | Training error | Test error | Validation error | Training algorithm | Error function | Hidden activation | Output activation |
|-----------|-------------------|--------------|---------------------|----------------|-----------|-----------------|-------------------|--------------|------------------|------------------|
| MLP 17-19-1 | 0.999372 | 0.999621 | 0.999955 | 0.006487 | 0.005904 | 0.004877 | BFGS 24 | SOS | Linear | Tanh |
| MLP 17-24-1 | 0.999533 | 0.999042 | 0.999745 | 0.005605 | 0.006511 | 0.005779 | BFGS 30 | SOS | Linear | Tanh |
| MLP 17-17-1 | 0.999979 | 0.999911 | 0.999984 | 0.000175 | 0.000539 | 0.000149 | BFGS 97 | SOS | Tanh | Logistic |
| MLP 17-18-1 | 0.999997 | 0.999982 | 0.999997 | 0.000023 | 0.000100 | 0.000032 | BFGS 220 | SOS | Tanh | Tanh |
| MLP 17-24-1 | 0.999288 | 0.998739 | 0.999678 | 0.006829 | 0.007918 | 0.005385 | BFGS 25 | SOS | Logistic | Tanh |
| MLP 17-24-1 | 0.999965 | 0.996825 | 0.996412 | 0.000277 | 0.033676 | 0.034449 | BFGS 39 | SOS | Linear | Exponential |
| MLP 17-22-1 | 0.998054 | 0.995334 | 0.993337 | 0.015533 | 0.043683 | 0.059666 | BFGS 6 | SOS | Exponential | Linear |
| MLP 17-23-1 | 0.965803 | 0.967476 | 0.967635 | 1.492294 | 1.867394 | 1.519765 | BFGS 4 | SOS | Logistic | Linear |
| MLP 17-17-1 | 0.999911 | 0.948048 | 0.942155 | 0.000712 | 0.507325 | 0.508122 | BFGS 27 | SOS | Exponential | Logistic |
| MLP 17-26-1 | 0.999996 | 0.974881 | 0.972070 | 0.000033 | 0.245970 | 0.245975 | BFGS 179 | SOS | Tanh | Tanh |
| MLP 17-24-1 | 0.999989 | 0.999863 | 0.999967 | 0.000098 | 0.000158 | 0.000164 | BFGS 167 | SOS | Tanh | Linear |
| MLP 17-22-1 | 0.999999 | 0.999999 | 0.999981 | 0.000009 | 0.000077 | 0.000114 | BFGS 267 | SOS | Tanh | Tanh |
| MLP 17-27-1 | 0.999994 | 0.999990 | 0.999977 | 0.000055 | 0.000133 | 0.000107 | BFGS 105 | SOS | Exponential | Logistic |
| MLP 17-20-1 | 1.000000 | 0.999995 | 0.999959 | 0.000002 | 0.000034 | 0.000176 | BFGS 369 | SOS | Exponential | Tanh |
| MLP 17-25-1 | 0.999989 | 0.999662 | 0.999956 | 0.000100 | 0.000250 | 0.000263 | BFGS 122 | SOS | Logistic | Logistic |
| MLP 17-26-1 | 0.999974 | 0.999996 | 0.999021 | 0.000200 | 0.000046 | 0.003301 | BFGS 86 | SOS | Logistic | Logistic |
| MLP 17-30-1 | 0.999995 | 0.999999 | 0.986592 | 0.000036 | 0.000010 | 0.129585 | BFGS 122 | SOS | Logistic | Logistic |
| MLP 17-28-1 | 0.999986 | 0.999972 | 0.986598 | 0.000106 | 0.000257 | 0.129745 | BFGS 76 | SOS | Tanh | Exponential |
| MLP 17-31-1 | 0.999981 | 0.999975 | 0.992708 | 0.000158 | 0.000268 | 0.067533 | BFGS 52 | SOS | Exponential | Tanh |
| MLP 17-23-1 | 0.999991 | 0.999997 | 0.990497 | 0.000067 | 0.000030 | 0.083596 | BFGS 88 | SOS | Exponential | Logistic |
| MLP 17-20-1 | 0.999974 | 0.999998 | 0.997244 | 0.000197 | 0.000256 | 0.029879 | BFGS 105 | SOS | Exponential | Logistic |
| MLP 17-25-1 | 0.999903 | 0.999995 | 0.994743 | 0.000050 | 0.000047 | 0.057111 | BFGS 268 | SOS | Tanh | Linear |
| MLP 17-21-1 | 0.999991 | 0.999974 | 0.970746 | 0.000066 | 0.000078 | 0.033534 | BFGS 141 | SOS | Logistic | Logistic |
| MLP 17-17-1 | 0.999992 | 0.999989 | 0.986964 | 0.000062 | 0.000084 | 0.033414 | BFGS 84 | SOS | Tanh | Logistic |
| MLP 17-18-1 | 0.999978 | 0.999988 | 0.996962 | 0.000164 | 0.000169 | 0.032799 | BFGS 92 | SOS | Tanh | Exponential |
Figure 5. Efficiency of the particular networks in particular substance concentration detection.
4. Discussion
The PCA diagrams show that it is possible to distinguish particular compounds. The position in the PCA space is different for each tested substance. It is also possible to distinguish substance concentration. In the applied reference system, it was noticed that the increase in the concentration level was inversely proportional to the PCA component.

The use of ANN also enabled to achieve high values of accuracy in the detection of particular substances and their concentrations detection. Five output neurons (one output for each compound) were applied there for the compound identification. The training accuracy was between 92.2% and 99.3%, test accuracy: 87.7% and 95.9%, and finally validation accuracy was between 81.5% and 89.5%. For 1-butanol the best match (0.999997) and the lowest error (0.000032) for validation data were obtained for MLP 17-18-1 network, which has tanh as a function of activation of hidden and output neurons. For acetone, MLP 17-24-1 with linear and exponential functions, matching and error of 0.996412 and 0.034449 respectively. Benzaldehyde: MLP 17-27-1 with exponential and logistic functions, matching and error of 0.999977 and 0.000107 respectively. Dimethylamine: MLP 17-31-1 with exponential and tanh functions, matching and error of 0.992708 and 0.067533 respectively. For toluene, MLP 17-25-1 network with tanh and linear activation functions with 0.999473 matching and 0.005711 error. The obtained results did not indicate any activation function that would be optimal for hidden and output neurons.

In the case of the concentration detection, these coefficients were higher for the particular compounds and achieved values between 0.94 and 0.99.

The results from the study prove that the network has good generalisation capabilities, i.e. it correctly interprets new data, not used previously in ANN training. It is a characteristic feature of a network with an optimal number of hidden neurons. Networks with too many hidden neurons tend to remember the learners' datasets, so they are unable to interpret new data correctly. Too few hidden neurons make the network incapable of recognising complex relationships between input and output data. According to Basheera and Hajmeerb, ANNs are more effective at classifying and determining complex relationships than other methods of analysis [9]. The efficiency of ANN was also described by Fu et al., Hetch-Nielsen and Jiang et al. in the articles concerning architectural design with the use of artificial neural networks and their capacity for learning [3, 10, 16]. Comparable effectiveness of chemical compound identification was achieved by Srivastava [14].

The detection effectiveness measured at this stage of research showed very good results in comparison with the results of other studies published in the scientific literature. The reported high identification efficiency was achieved only in the case of studies with the number of compounds below five [8, 17, 18]. With 6 chemical compounds, the classification accuracy reported by J.F.et al. [3] was between 92.9% and 100%, [3], while for 10 chemical compounds in the experiment conducted by Santos et al. [11], the average efficiency of classification was only 71.9%.

5. Conclusions
This paper shows the potential of the Artificial Neural Networks in detection of certain chemical compounds present in the tested air and, in addition, their concentration. According to the conducted measurements and the ANN analysis, the following conclusions may be formulated:
- Artificial Neural Networks provide better results compared to the PCA analysis.
- The multi-sensor device in use enables the detection of individual substances with validation accuracy in the range of 81.5 and 89.5%.
- The applied device evaluates particular substance concentration using the ANN with validation accuracy in the range of 0.94 and 0.99%.

6. References
[1] Gardner J W and Bartlett P 1994 Sensor Actuat B-Chem 18(1–3) 210–11
[2] Villarubia G, De Paz J F, Chamoso P and De la Prieta F 2018 Neurocomputing 272 10-6
[3] Fu J, Li G, Qin Y and Freeman W J 2007 Sensor Actuat B-Chem 125(2) 489–97
[4] Haugen J E and Kvaal K 1998 Meat Sci 49(98) 273-86
[5] Barisci J N, Wallace G G, Andrews M K, Partridge A C and Harris P D 2002 Sensor Actuat B-Chem 84(2) 252–57
[6] Guz L, Łagód G, Jaromin-Gleń K, Suchorab Z, Sobczuk H, Bieganowski A 2015 Sensors 15(1), 1-21
[7] Hamilton S, Hepher M J and Sommerville J 2005 Sensor Actuat B-Chem 107(1) 424–32
[8] Ding H, Ge H and Liu J 2005 Sensor Actuat B-Chem 107(2) 749–55
[9] Basheera A and Hajmeerb M 2000 J Microbiol Meth 43(1) 3-31
[10] Hecht-Nielsen R 1987, Proc. of the IEEE (San Diego, USA) p 11–3
[11] Negri R M and Reich S 2009 Sensor Actuat B-Chem 75(3) 172–78
[12] Peris M and Escuder-Gilabert L 2009 Anal Chim Acta 638(1) 1–15
[13] Sardi S, Vardi R, Sheinin A, Goldental A and Kanter I 2017 Sci Rep-UK 7(1) 18036
[14] Lawson D and Marion G 2008, Bioinf Stat 53(4) 543-56
[15] Tadeusiewicz R 1993 Sieci neuronowe (Kraków: Akademicka Oficyna Wydawnicza)
[16] Jiang J, Trundle P and Ren J 2010 Comp Med Imag Grap 34(8) 617-631 10.1016/2007.07.003,
[17] Llobet E, Brezmes J, Vilanova X, Sueiras J E and Correig X 1997 Sensor Actuat B-Chem 41(1–3) 13–21
[18] Szczurek A and Maciejewska M 2004 Talanta 64(3) 609–17
[19] Santos J P, Aleixandre M and Cruz C 2012 Chem Engineer Trans 30 181-6
[20] Schaller E, Bosset J O and Escher F 1998 LWT-Food Sci Technol 31(4) 305–16
[21] Stetter J R, Findlay M W, Schroeder K M, Yue C and Penrose W R 1993 Anal Chim Acta 284(1) 1–11
[22] Srivastava A K 2003 Sensor Actuat B-Chem 96(1–2) 24–37
[23] Zhang Z 2018, Multivariate Time Series Analysis in Climate and Environmental Research (Cham: Springer)
[24] Yang Y M, Yang P Y and Wang X R 2000 Sensor Actuat B-Chem 66(1–3) 167–170
[25] Lee D S, Jung J K, Lim J W, Huh J S and Lee D D 2001 Sensor Actuat B-Chem 77(1) 228–236
[26] Kadmon J and Sompolinski H 2016, Adv Neur In, 29 1-5
[27] Johnson J L, Schamschula M P, Inguva R and Caulfield H J 1998 Pulse-Coupled Neural Networks Neural Networks and Pattern Recognition, ed Omidvar O., Dayhoff J (San Diego: Akademic) 1 p 1-56.