Assessment of the possibility of using data mining methods to predict sorption isotherms of selected organic compounds on activated carbon

Lidia Dąbek[1,*], *Bartosz* Szeląg[1], and *Anna* Picheta-Oleś[2]

[1]Kielce University of Technology, Faculty of Environmental, Geomatic and Energy Engineering, al. Tysiąclecia Państwa Polskiego 7, Kielce, Poland
[2]Marshal's Office of the Świętokrzyskie Voivodeship, al. IX Wieków 4, Kielce, Poland

Abstract. The paper analyses the use of four data mining methods (Support Vector Machines, Cascade Neural Networks, Random Forests and Boosted Trees) to predict sorption on activated carbons. The input data for statistical models included the activated carbon parameters, organic substances and equilibrium concentrations in the solution. The assessment of the predictive abilities of the developed models was made with the use of mean absolute error (MAE), mean absolute percentage error (MAPE), and root mean squared error (RMSE). The computations proved that methods of data mining considered in the study can be applied to predict sorption of selected organic compounds on activated carbon. The lowest values of sorption prediction errors were obtained with the Cascade Neural Networks method (MAE = 1.23 g/g; MAPE = 7.90% and RMSE = 1.81 g/g), while the highest error values were produced by the Boosted Trees method (MAE=14.31 g/g; MAPE = 39.43% and RMSE = 27.76 g/g).

1 Introduction

Ever increasing industrialization is associated with the generation of wastewater that has complex chemical composition, and contains hazardous, carcinogenic and mutagenic substances. Due to enormous quantitative and qualitative diversity of industrial wastewater, its treatment involves the application and combination of many different methods. They include biodegradation, coagulation, sorption, oxidation and filtration [1–3]. Literature data [3–6] indicate that sorption on activated carbons is one of the most effective means of the removal of organic compounds. That is due to carbons extended porous structure and specific chemical properties of their surface. To efficiently remove given pollutants from aqueous solution by means of sorption, it is necessary to choose appropriate process conditions. That concerns the physicochemical properties of both the adsorbate and activated carbons

* Corresponding author: lidiadabek@tu.kielce.pl

(magnitude of specific surface area, pore distribution, size and volume, surface chemical properties). Sorption capacity is determined on the basis of experimental investigations and the fitting of the theoretical curves (Langmuir, Freundlich isotherms and others) to the results obtained in measurements. It is a complex task that requires substantial labour input and generates high costs while the results obtained describe only one particular adsorbent – adsorbate system. However, both the literature on the subject [1–12] and technological practice provide databases of sorption results for activated carbon-adsorbate systems that have already been investigated. Therefore, it seems reasonable to make use of those data to predict sorption in the systems that show comparable properties. Literature data [5, 11–15] indicate that mathematical models using multiple regression equations, artificial neural networks and their numerous modifications could be employed to this end. The models are developed on the basis of the dataset, which is partitioned into the training set, and the testing and validating set. That is intended to ensure satisfactory predictive abilities of the model [16–18]. At the training stage, parameters in the model structure are estimated, the next two stages involve the assessment of the predictive powers of the developed mathematical models. The condition of model usability is met if the model has satisfactory predictive ability with respect to all three sets.

The application of mathematical models to sorption predictions was discussed in many studies, which concerned both metal sorption on active carbons and that of organic compounds [5, 12–14]. However, it should be noted that in a majority of cases, the models presented by researchers described the sorption of single pollutants on selected activated carbons. In addition, more attention was given to the sorption process conditions than to physicochemical properties of activated carbons. As a result, the practical applicability of such models is limited as they cannot be used for predicting sorption amount of another adsorbate group on different activated carbons. Consequently, it is reasonable to develop statistical models for the prediction of sorption on activated carbons, which account for the diversity of carbon physicochemical properties, and also different adsorbates.

On the basis of previous results [10] on the sorption of dyes (crystal violet and phenol red) and para – chlorophenol on four different activated carbons that have brand names of WDex, WG12, F200S and F200R, in this study an attempt was made to develop a mathematical model for the prediction of sorption of the same adsorbates, but on different activated carbons that have parameters similar to those used while developing the model.

2 Methodology

To develop a mathematical model for predicting sorption (output), the data (discrete variables) that constitute the characteristics of WDex, WG12, F200S and F200R activated carbons were used [3,8-10]. They include the following: dechlorination ability (Ab. De.), iron content (Cont. Fe.), methylene number (L. Met.), iodine number (L. J.), detergent number (L. Det.), ash content (Cont. ash), soluble substance content (Cont. s. sol.), specific surface area (S), and values of initial and equilibrium concentrations for the following adsorbates: crystal violet, phenol red and p-chlorophenol. The variables mentioned above were inputs for the mathematical model. The methodology followed in sorption experiments and in the determination of physico- chemical properties of activated carbons was discussed in details in the study by Dąbek et al. [3, 10]. The characteristics of activated carbons selected for investigations are presented in Table 1. Equilibrium concentrations of the examined adsorbates ranged 1.32–64.41 mg/dm^3 for phenol red, 0.42–15.85 mg/dm^3 for crystal violet and 4.52–84.52 mg/dm^3 for p-chlorophenol, respectively. Additionally, because of their diversified composition and chemical properties, the adsorbates selected for the investigations were modelled as linguistic variables.

Table 1. Physicochemical characteristics of activated carbons selected for investigations [3, 8–10].

Carbon	**Ab.De**	**Cont.Fe.**	**L.Met.**	**L.J.**	**L.Det.**	**Cont. ash**	**Cont.s. sol.**	**S**	**V**
	cm	mg/g	cm^3	mg/g		%	%	m^2/g	cm^3/g
WDex	5	0.50	36	990	20.6	20.64	1.07	1050	0.95
WG12	3	0.44	32	1230	32.0	8.85	2.48	980	0.89
F200S	7	0.43	24	710	12.7	7.28	1.16	720	0.72
F200R	6	0.50	16	770	15.7	7.84	1.74	800	0.86

In this study, Support Vector Regression (SVR), Boosted Trees (BT), Random Forests (RF), and Cascade Neural Networks (CNN) methods were applied for predicting sorption on selected activated carbons. The methods selected for the study were divided into two model groups. In complex models, complex structures (regression support vectors, neural networks) are generated at the training stage, and also it is necessary to compute many parameters. The other group comprises much simpler models, for the implementation of which a much smaller number of parameters is needed. Therefore, one of the objectives of this study was to compare the results of sorption simulations obtained with complex models (CNN, SVM) with those produced by models that have a simpler structure (BT, RF).

Prior to the development of the models, input data and discrete output data were normalised using the min-max transformation. Additionally, to ensure the correctness of the training process, the data (N = 50) were partitioned into three sets, namely the training (75%), testing (25%) ones. In the study, the measurement results of crystal violet sorption on F 200R activated carbon were employed to validate the statistical model.

Support Vector Machines (SVM) comprise a group of methods developed by Vapnik [21] originally for entirely classification purposes, and later on also for regression problems (SVR). Due to the fact that dependence holding between the model output and input variables could be non-linear, in the method, N – dimensional space is non-linearly transformed into K – dimensional space of features, which has a greater size. In this study, to predict sorption, Support Vector Regression method was applied with the radial kernel function intended to minimise the functional having the following form:

$$\sum_{i=1}^{m} \frac{C}{m} \cdot |y_i - f(x_i)|_{\varepsilon} + \frac{1}{2} \cdot \|f\|_k^2 \tag{1}$$

where: $|y_i - f(x_i)|_{\varepsilon} = \max\{0, |y_i - f(x_i)| - \varepsilon\}$, ε – permissible error value, $\|f\|$ - rule f in the Hilbert space, c – constant selected by the user depending on value ε [19], m – size of the training set, $f(x_i)$ – value of function $f(x)$ at point x_i, described by the equation:

$$f(x) = \sum_{i=1}^{N_{sv}} (\alpha_i - \alpha_i') \cdot K(x, x_i) + w_0 \tag{2}$$

in which: w_0 – deviation, N_{sv} – number of support vectors, which depends on C and ε, α_i, α_i' – Lagrangian multipliers, $K(x,x_i)$ – kernel function with radial basis functions [19]

Boosted Trees (BT) are an implementation of the stochastic gradient boosting method applied to classification and regression problems [17]. The main concept behind the method is the creation of a series of decision trees, with each successive tree used to identify the remainders generated by the previous one. The computations demonstrated that for some estimation and prediction problems, boosted trees produced predictions closer to actual values than it was the case with solutions obtained while using single regression trees.

The random forest algorithm was proposed by Leo Breiman [16] and it originated from the bootstrap method. At the first stage, k – fold sampling for the training dataset of size n is performed, in which repetitions are allowed. Next, based on the sets obtained, regression

trees are formed. The process of their construction was modified compared with the original algorithm, so that every node of the tree could be split in the best manner. That is done not on the basis of all attributes, but those randomly selected ones (independent variables). In this way, k-number of regression trees are obtained, which form a forest. The later provides a basis to determine prediction, which comes as the model output. The overall prediction is computed as an arithmetic mean of individual predictions for individual trees.

Neural networks could be used for optimisation, classification, the simulation of linear and non-linear processes and also for control [18, 20, 21]. As regards the modelling of the processes taking place in the natural environment, the network termed as multilayer perceptron (MLP) is used. One of the modifications of ANN of the multilayer perceptron-type is a cascade neural network (CNN). The latter has additional weighted connections from each layer input, and from each layer to a successive one. Cascade network having a higher number of layers can be used to model complex non-linear dependences. The optimal structure of the neural network for sorption amount predictions was sought computing the measures of fit of simulation results to measurements (MAE, MAPE). That was done for a number of neurons in the hidden layer (3–20) adopted in the study and for linear, exponential (exp), sine (sin), sigmoidal (sigm) and hyperbolic tangent activation functions. In the output layer, linear activation function was assumed. Based on studies [22] on cascade neural networks, three connections linking inputs to successive CNN layers were adopted.

In the analyses, to assess the predictive powers of the models for sorption value prediction, mean absolute error (MAE), mean absolute percentage error (MAPE), coefficient of correlation (R), and root mean squared error (RMSE) described in studies by Szeląg and Gawdzik [19, 20] were employed.

3 Results

On the basis of the determined parameters that describe physicochemical properties of activated carbons and the values of equilibrium concentrations for individual adsorbates, statistical models were developed to predict sorption using Support Vector Machines, Boosted Trees, Random Forests, and Cascade Neural Networks methods. Table 2 shows the computed parameters of fit (MAE, MAPE, RMSE, R) of computational results to sorption value measurements. Table 3 lists the parameters that describe the structure of the five selected statistical models, developed using the CNN method, which predict sorption in the best way. Additionally, to visualise the results, a comparative analysis of experimentally determined sorption isotherms and values computed using statistical models is presented in Figs. 1–3.

The analyses indicate that the value of parameter C in eq. (1) in the model obtained with the SVM method was 7. As regards both BT and RF methods, the number of additive trees, on the basis of which the regression models were devised, was not higher than 150, which means the models were not over-trained. The data presented in Tables 2 and 3 show that the lowest errors in the fit of computational results to sorption measurements were obtained for the CNN method, whereas the highest error values were found when the BT method was used. With respect to CNN-based models (Table 3), it was observed that the number of neurons in the individual hidden layers ranged from 6 to 10. Also, the lowest error values (MAE, MAPE, RMSE) were produced using the statistical model, in which the neuron number in individual hidden layers was 7, and the activation function was a hyperbolic tangent dependence. The analyses performed for the study show the values of mean absolute and percentage errors are higher for the models based on BT and RF methods than it is the case for SVM and CNN methods.

Table 2. Comparison of predictive abilities of sorption prediction models developed using BT, RF, and SVM methods.

Method	training				test			
	MAE	MAPE	RMSE	R	MAE	MAPE	RMSE	R
BT	12.29	38.46	22.73	0.812	14.32	39.43	27.76	0.808
RF	5.32	24.28	8.26	0.991	5.7	25.49	8.93	0.989
SVM	2.78	14.94	4.06	0.995	2.91	15.45	4.26	0.994

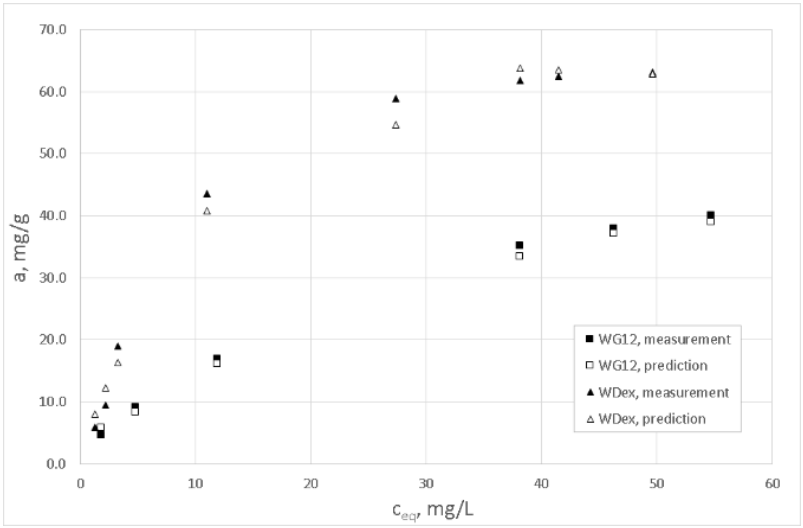
Table 3. Comparison of predictive abilities of sorption prediction models developed using the CNN method.

Neurons	Activation function	training				test			
		MAE	MAPE	RMSE	R	MAE	MAPE	RMSE	R
9	tanh	1.77	8.74	2.61	0.997	1.88	9.3	2.72	0.996
6	sigm	2.02	9.9	2.95	0.995	2.15	10.53	3.07	0.994
10	tanh	2.05	9.92	2.95	0.995	2.18	10.55	3.07	0.994
7	sin	1.84	9.06	2.75	0.996	1.96	9.64	2.87	0.995
8	exp	1.85	9.07	2.76	0.996	1.97	9.65	2.88	0.995

It should be noted that values of errors in the fit of computational results to measurements, for the RF method are lower than those determined using the BT method. Additionally, it was concluded that for the prediction of the sorption of selected organic compounds, the models that have more complex structure produce lower values of MAE, MAPE, RMSE errors (Tables 2 and 3) than those obtained using the models based on the modified method of regression trees.

The predictive powers of the statistical models developed for sorption prediction were confirmed in the investigations of Meena et al [13]. The researchers applied artificial neural networks to model the phenomenon of crystal violet dye removal and obtained the value of R = 0.965. The results reported by Yang et al. [11], who received a good fit of computational results to measurements (R = 0.997), support the previous findings.

a)



b)

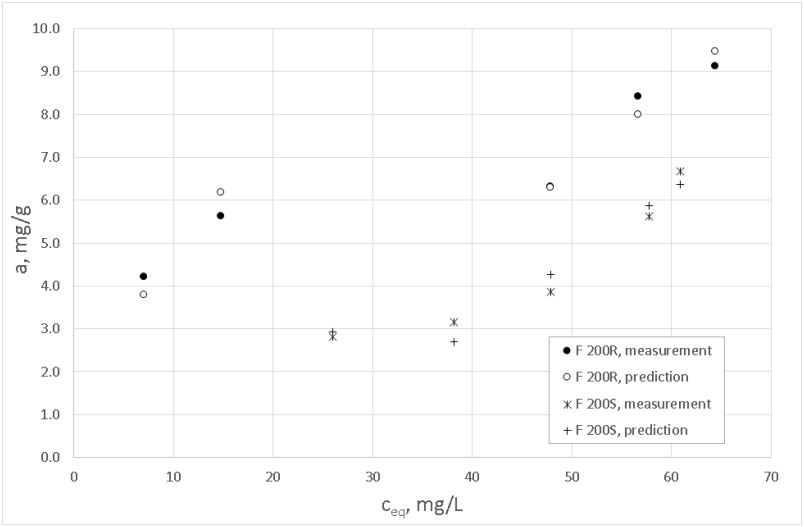


Fig. 1. Comparison of the results of measurements and computations of phenol red sorption: a) on WG12 and WDex activated carbons ; b) on F 200R and F 200S activated carbons.

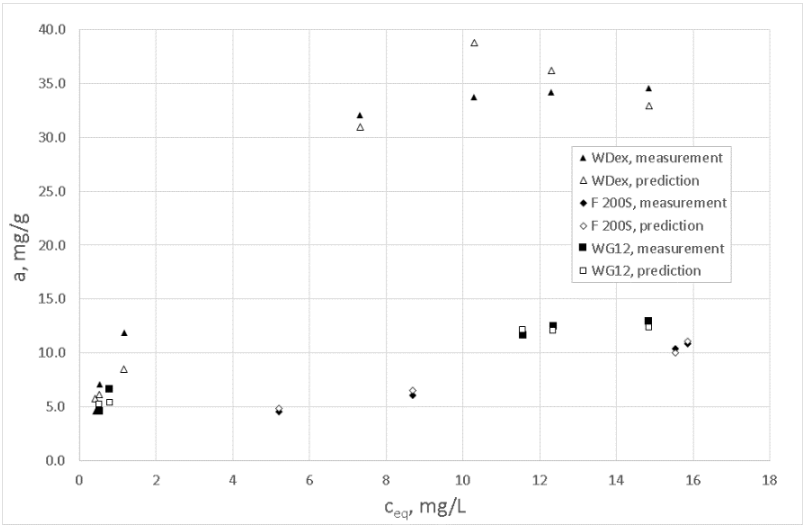


Fig. 2. Comparison of the results of measurements and computations of crystal violet sorption on WDex, F 200S, and WG12 activated carbons.

The values of errors (Tables 2 and 3) are comparable with the results (0.777–0.999) obtained with classic methods (Langmuir, Freundlich isotherms and others) reported by Dąbek et al. [10]. The researchers used Langmuir model to predict sorption on WDex, F200S, F200R and WG12 activated carbons. The analyses conducted by Depci et al. [12] produced similar results.

The results of sorption computations with the use of the CNN method for the validating set, presented in Fig. 4, confirm that the developed statistical model shows satisfactory predictive abilities. In the case under consideration, error values are MAE = 0.85 mg/g;

MAPE = 12.68%; RMSE = 0.86 mg/g. The results of computations presented above indicate it is possible, using data mining methods, to predict sorption on selected activated carbons in the presence of different organic substances.

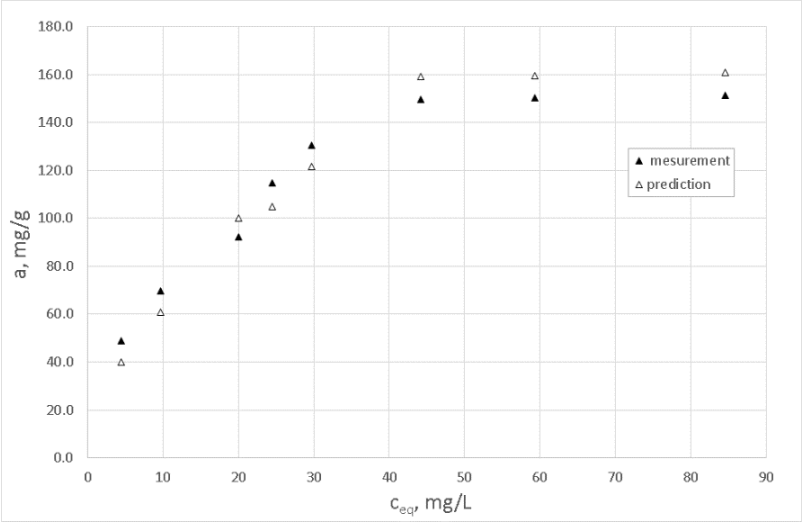


Fig. 3. Comparison of the results of measurements and computations of para-chlorophenol sorption on WDex activated carbon.

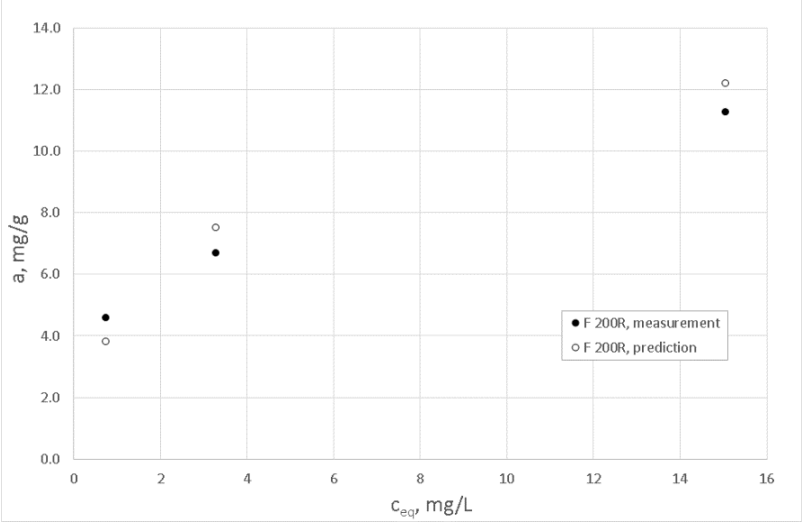


Fig. 4. Comparison of the results of measurements and computations of crystal violet sorption on F 200R activated carbon.

4 Conclusions

The computations performed for the study showed that the course of sorption on selected activated carbons with diversified properties, in the presence of various organic substances can be modelled using Boosted Trees, Random Forests, Support Vector Machines, and

Cascade Neural Networks. In the models under consideration, the lowest error values in sorption prediction were obtained using the CNN method, whereas the BT method produced the highest error values. The computations also indicated that lower error values were found for complex models (Neural Networks, Support Vector Machines) than for the models that required estimation of a smaller number of parameters in their structure.

The results of introductory investigations and satisfactory predictive abilities of the mathematical model developed for predicting sorption of various adsorbates on carbons of diversified properties seem promising. However, it is necessary to verify the model, and to evaluate its usability for different active carbons. Due to a limited amount of data available for analysis, it is also necessary to work out methodology that would account for random selection of data for the training and testing sets, respectively.

References

1. V.K Gupta, V.Suhas, J. Environ. Manage. **90**, 8 (2009)
2. V.P. Santos, M.F.R. Pereira, P.C.C. Faria, J.J.M. Órfão, J. Hazard. Mater. **162**, 2–3 (2009)
3. L. Dąbek, E. Ozimina, A. Picheta-Oleś, Ecol. Chem. Eng. A. **19**, 3 (2012)
4. R.Ch. Bansal, M. Goyal, *Activated Carbon Adsorption* (Taylor & Francis Group, Baca Raton-London-New York-Singapore, 2005)
5. S. Azizian, M. Haerifar, H Bashiri, Chem. Eng. J. **146**, 36 (2009)
6. A.P. Terzyk, G. Rychlicki, P.A. Gauden, P. Kowalczyk, Water Encyclopedia, Volume Oceanography; Methodology; Physics and Chemistry; Water Law; and Water History Art and Culture, 404–408, Willey (2005)
7. N. Daneshvar, S. Aber, F. Hosseinzadeh, Global Nest J. **10**, 1 (2008)
8. L. Dąbek, E. Ozimina, A. Picheta-Oleś, Environ. Prot. Eng. **38**, 1 (2012)
9. L. Dąbek, E. Ozimina, A. Picheta-Oleś, Ecol. Chem. Eng. A. **17**, 11 (2010)
10. L. Dąbek, E. Ozimina, A. Picheta-Oleś, Rocznik Ochrona Środowiska **13** (2011)
11. Y. Yang, X. Lin, B. Wei, Y. Zhao, J. Wang, Int. J. Environ. Sci. Te. **11**, 4 (2013)
12. T. Depci, A. R. Kul, Y. Onal, E. Disli , S. Alkan, Z. F. Turkmenoglu, Physicochemical Problems of Mineral Processing **48**, 1 (2011)
13. M. Meena, K.Srinivasan, Pollution **50**, 10414 (2012)
14. A. Khataee, A.Khani, Inter. J. Chem. React. Engin. **7**, A5 (2005)
15. N. Prakash, S.A.Monikandan, J. Hazard. Mater. **152**, 1268 (2008)
16. L. Breiman, Journal Machine Learning **45**, 1 (2000)
17. J. Friedman, Comput. Stat. Data An. **38**, 4 (2002)
18. S. Dellana, D. West, Environ Modell Softw. **24**, 1 (2009)
19. B. Szeląg, J Gawdzik, Pol. J. Environ. Stud. **25**, 4 (2016)
20. B. Szeląg, J Gawdzik, Pol. J. Environ. Stud. **26**, 1 (2017)
21. V. Vapnik, *Statistical Learning Theory* (John Wiley and Sons, New York 1998)
22. S. B. Al-Batah, M.S Alkhasawneh, L.T Tay, U.K. Ngah, H. H. Lateh, M.T.A Isa, Math. Probl. Eng. **2015**, 1 (2015)