A Deep Learning Model, interpreted with an XAI technique, to simulate and optimize the remediation of oil-drilling cuttings in bubble flow reactors

KALARI K.[1,2], CHRISTODOULIS K.[1,2], BALI N.[1], THEODOROPOULOU M.[1] , TSAKIROGLOU D. C.[1,*]

[1]Foundation for Research and Technology Hellas, Institute of Chemical Engineering Sciences (FORTH/ICE-HT), 26504 Patras, Greece

[2]University of Patras, Department of Physics, 26504 Patras, Greece

*corresponding author:
e-mail: ctsakir@iceht.forth.gr

Abstract A multitask deep neural network (DNN) is developed to simulate the ozonation of oil-drilling cuttings (ODC) and is interpreted through a technique of explainable artificial intelligence (XAI) to provide knowledge about the experimental conditions that will maximize the decontamination of ODC. On a semi-batch bubble flow column, ozonation experiments of ODC are carried out after pretreatment with synthetic seawater (SW) and the anionic surfactant sodium dodecyl sulphate (SDS). The performance of ozonation experiments is evaluated by measuring the removal efficiency of the total organic carbon (TOC). The experimental data are used for training and testing an DNN that can predict accurately the TOC removal efficiency of the ozonation process as well as the values of different variables such as pH, oxidation-reduction potential (ORP), temperature (T), pressure drop (ΔP), based on the values of the input variables of the model. The acquired model is interpreted through the Shapley Additive explanations (SHAP) method, an important advancement in the field of machine learning interpretation provided by XAI, regarding the significance of the models' input variables in the TOC removal efficiency. This step aims at establishing the experimental conditions that lead to the highest remediation rate.

Keywords: Ozonation, Soil Remediation, Bubble Column Reactor, Deep Neural Networks, Multitask Learning, Explainable Artificial Intelligence.

1. Introduction

Oil drilling cuttings (ODC) are the main wastes generated during the oil/gas exploration and extraction (Talbi et al, 2009), where the oil content may reach up to 5-20% by weight, in dry basis. The most common methods used for the removal of oil from ODC are: the surfactant-enhanced washing (Yan et al,2011), the thermal (Falciglia et al, 2011) and microwave (Robinson et al, 2011) desorption / evaporation the solidification / stabilization (Leonard et al, 2010), the phytoremediation (Ji et al, 2004), and the bioremediation (Azubuike et al, 2016).

During the last years, the ozonation has emerged as a promising advanced oxidation process (AOP) for the degradation of petroleum products from heavily contaminated soils and sludges (Wang et al, 2017). One type of reactors used for the ozonation of wastewater or aqueous suspensions of solids (slurries), because of its simple construction and maintenance, is the semi-batch bubble columns (Sun et al, 2020). The inherent complexity of the fluid dynamics and chemical reactions occurring within those multiphase reactors renders their simulation a complex and challenging task. Traditional models, such as analytical and computational fluid dynamics (CFD) models, have been widely used to describe the multiphase flow behavior and reaction kinetics in these systems (An et al, 2020). However, these models often have limitations in accurately predicting the complex behavior of the system due to the assumptions and simplifications made during their development.

Data-driven models, such as artificial neural networks (ANNs) or more intricate ones such as deep neural networks (DNNs), offer an alternative approach on simulating multiphase reactors. These algorithms provide a non-linear mapping between input and output variables and are also useful in providing cross-correlations among these variables. Moreover, compared to the empirical curve-fitted models, ANNs are relatively less sensitive to noise and incomplete information. Because of their advantages, during the last years, several attempts have been devoted to the application of

ANN to numerical modeling and control of multi-phase reactors, such as bubble column reactors (Shaikh et al, 2007).

In the present work, a multitask deep neural network (DNN) is developed to simulate the ozonation of ODC. Specifically, the DNN evolved is a feedforward ANN with multiple layers that is trained through multi-task learning. Multi-task learning is a training paradigm in which machine learning models are trained with data from multiple tasks simultaneously, using shared representations to learn the common ideas between a collection of related tasks. The DNN has undergone training and testing by using experimental data, and it is capable of making highly accuracy predictions (> 99%). The model can perform predictions in five distinct tasks, which include the prediction of: TOC Removal Efficiency in solid phase, pressure drop, pH, reduction-oxidation (redox) potential and temperature. Finally, the issue of interpretability in neural networks, often referred to as their "black box" character, has been tackled using the SHAP method. This technique is a part of the XAI (eXplainable Artificial Intelligence) toolbox and is employed to assess the influence and effect of input variables to the TOC removal efficiency.

2. Materials and methods

2.1 Ozonation in a bubble flow reactor: experimental

Water-saturated ODC samples collected from off-shore wells was sealed in a container and used for testing. Tri-distilled water and chemicals of analytical grade were utilized for the preparation of the solutions used for the pre-treatment of ODC and chemical analyses: magnesium sulphate, $MgSO_4$ (Carlo Erba), calcium chloride, $CaCl_2$ (Merck), potassium chloride, KCl (Merck), sodium chloride, NaCl (Merck), sodium dodecyl sulphate, SDS (Merck), potassium dichromate, $K_2Cr_2O_7$, sulfuric acid H_2SO_4, Mercury sulphate (Penta), Silver sulphate (Penta), Hydrochloric acid (Sigma-Aldrich), Acetone (Honeywell), Hexane (Honeywell), Tetracontane (Merck), Florisil (Sigma-Aldrich). Artificial seawater (SW) containing 12.85 g/L $MgSO_4$, 1.0154 g/L $CaCl_2$, 2.6181 g/L KCl, 27.5419 g/L NaCl was prepared and enriched with SDS in three concentrations: 0 %, 0.2%, and 0.5% w/w. Samples of ODC were mixed with the abovementioned SW solutions in various ratios (1:2, 1:3, 1:4, 1:5), placed in an ultrasonic bath (Witeg WUC-D 3.3 40 kHz), and sonicated for 30 min at 30oC.

The pre-treated suspension of ODC diluted in SW is placed inside a poly methyl methacrylate (PMMA) cylindrical column of inner diameter 3 cm, height 50 cm, and equipped with inlet/outlet ports and four side ports for sampling. For ozone production, bottled oxygen is fed to an ozone generator at a flow rate regulated (0-5 L/min) through a mass flow controller. The ozone-rich gas is humidified by passing through a gas washing bottle and then injected into the PMMA column through a cylindrical porous diffuser. The concentration of the produced ozone is detected at the inlet port with an ozone analyzer. The gas pressure is measured at the inlet and outlet ports with absolute pressure transducers. The outflowing gas is injected through a column packed with silica gel to remove moisture and then through an ozone destructor for safety reasons. There are also sensors that measure the temperature, the pH and the redox potential every second during the ozonation time. After every 15 min, 5 g of suspension was collected from a sampling port and centrifuged at 11000 rpm for 15 min. The supernatant was then removed and the residual ODC was left under a hood for 16 h to air dry. The total organic carbon was measured in dried samples with an organic carbon analyzer (Multi N/C 2100 S with auto-sampler AS 60, Analytik Jena).

2.2. Development of a multitask neural network

In general, developing an ANN involves several steps, from data preparation to model architecture, training, validation, and testing. In our case, the data preparation includes selecting the appropriate dataset, cleaning and formatting the data, splitting it into training and testing sets. The dataset utilized in the multitask model, is composed of results and data recordings from 18 experiments. Each experiment involves 3335-4002 data points. Each datapoint distinguishes itself for the treatment time (t), the ozone concentration (C_{O_3}), the pressure drop (ΔP), the temperature (T), the pH, the redox potential (ORP), the TOC Removal Efficiency in solid phase (R_{ES}) and there are also the values of the initial TOC concentration of the ODC $(C_{TOC,init})$, the SDS concentration used (C_{SDS}), the flow rate of oxygen (F_{O_2}), the mass of treated ODC (m_{ODC}) and the volume of artificial sea water (V_{sw}) mixed with ODC, the initial temperature (T_{init}), the dilution ratio (DR) as well as the liquid height before (H_L) and after injecting gas (H_L'). All the values of these parameters of the experiment are utilized in the multitask neural network in order to help us in elucidating the physicochemical properties governing the ozonation process and consequently, in establishing the experimental conditions that lead to the highest degree of ODC remediation. It is significant to note that during the very first experimental trials, the absence of certain sensors from the experimental setup resulted in the existence of NaN values in the pH, ORP variables.

The dimensions of the experimental dataset are (62031,16), with eleven of the sixteen variables designated as input variables for the model $(t, C_{O_3}, C_{TOC,init}, C_{SDS}, m_{ODC}, V_{sw}, T_{init}, H_L, H_L', F_{O_2}, DR)$, while the remaining five variables $(R_{ES}, \Delta P, T, pH, ORP)$ concern the output variables that are expected to be predicted by the neural network. The available dataset is divided, randomly, into training and testing sets. The training set is composed of 55827 data points, whereas the testing set includes 6204 data points.

The architecture of the model chosen for simulating the ozonation experiment is a feed-forward multilayer neural network, which is reported as universal approximators (Hornik et al, 1989), trained with multitask learning. Multitask learning is a technique which aims at improving machine learning efficacy by simultaneously co-modelling multiple properties within a single model. This approach can provide several advantages, such as improved performance on each task, reduced training time, better generalization to new tasks, and it can contribute to reduce the overfitting (Zhang et al, 2018). In our model, the multitask learning was performed by hard parameter sharing of hidden layers. Specifically, hard parameter sharing is applied by sharing the hidden layers between all tasks, while keeping several task specific output layers. The number of layers, as well as the number of nodes are chosen by the trial and error method.

The training of the multitask neural network is performed through backpropagation algorithm. The process involves passing the input data of the training set forward through the network to produce the five outputs of the tasks defined. Each output is associated with its own loss function which measures the error noted between the produced output and the desired value. For each task, the loss function is the mean squared error, which in the context of a task with NaN values is modified to ignore these examples. The total loss of the network is formed as a weighted sum of the individual losses of the tasks, to ensure that during training the model prioritizes the tasks that require more attention. During backpropagation, the gradient of the total loss is computed with respect to the weights of the network and then used to update the weights by using the optimization algorithm Adam. These feed-forward propagation and the back propagation steps are repeated iteratively (epochs) until a global (or an accepted local) minimum of total loss function is obtained. After a sufficiently large number of trainings cycles, the network will usually converge to some state where the error of the calculations is quite small. The multitask neural network is then utilized to make predictions using the final weights related to the network's neurons.

Once the model has been trained with the training data, the next step is to evaluate its performance using the testing data. The purpose of testing is to evaluate the performance of the model on data that have not been used before, and ensure that the model generalizes well to new examples. The DNN is tested on 6204 new examples that comprise the testing set, and its performance is evaluated through the developed loss functions and the coefficient of determination R^2. The examples in the test set that contain NaN values for the variables pH, ORP are ignored from the metric R^2.

The last step in our approach is the interpretation of the performance of the multitask model. This analysis is accomplished by using an agnostic interpretation technique. Specifically, the analysis is carried out by investigating a small part of training set using the SHAP values (Lundberg et al, 2017). With the aid of these values, the importance of each input variable, along with its effect on the output variables is obtained. In this manner, the significance of the effect of each operational variable on the remediation efficiency is quantified by providing knowledge that could contribute to the ozonation optimization. For our neural network, we used KernelShap that is based on LIME (Ribiero et al, 2016).

3. Results and discussion

3.1 Performance of the multitask neural network

The neural network that attains the best performance in the testing procedure is a multitask neural network, with 3 shared hidden layers with 1024,512,256 nodes each layer and activation function ReLu. The task specific layers are 4 layers for the first and fifth task, 3 layers for the second task and 2 layers for the third and fifth tasks. The last layer for each task is composed of one node, characterized by the linear activation function, and correspond to the outputs of the model. Whereas, the nodes of the rest task specific layers are 1024,624,124 for the first task, 624,182 for the second task, 162 for the third task, 162 for the fourth task and 1024,824,624 for the fifth task. Its performance in new data of the testing set is illustrated in Table 1.

Table 1. The performance of the multitask neural network in the testing process.

	Loss Function	Coefficient of determination
Task 1	1.5e-05	0.9997
Task 2	0.00016	0.9938
Task 3	9.46e-06	0.9995
Task 4	2.87e-05	0.9991
Task 5	0.00026	0.9956

3.2 Interpretation of the ANN model with regard to TOC removal efficiency

Following the overall model performance, attention was paid on the estimated SHAP values from the ANN model for the first task (Figure 1). Each point on the summary plots a SHAP value of an instance per variable and is coloured according to the value of the variable, going from low (blue) to high (red) values. Their position to the right or left of the vertical axis indicates if the corresponding values favour (positive SHAP values) or decrease (negative SHAP values) the values of TOC removal efficiency. Also, the variables are ranked in accordance to their significance and stacked vertically (Figure 1).

According to the neural network predictions, the most significant input variable that affect strongly the prediction is the treatment time. Over time, the TOC removal efficiency is increasing. The second variable, in terms of importance, according to the SHAP method, is the concentration of SDS. It appears that, the presence of SDS does not mean that the R_{ES} will shift to higher values. In contrast, the higher the SDS concentration in seawater leads to lower prediction of R_{ES} (Figure 1). The total liquid height after the gas injection (namely the gas holdup) is the next more important variable, and its higher value leads to lower level of TOC removal efficiency. The opposite relation seems to be valid between the initial temperature and R_{ES}. The higher the T_{init} , the higher the TOC removal efficiency from solid phase. On the other hand, for small values of the sea water volume, low oxygen flow rates favor the remediation of the ODC. Regarding the mass of ODC, its higher values lead to higher predictions of R_{ES}. The same effect appears for the variables $C_{TOC,init}$, C_{O_3} on the TOC removal efficiency in solid phase. On the other hand, no clear picture is evident for the effect of DR, H_L on R_{ES}.

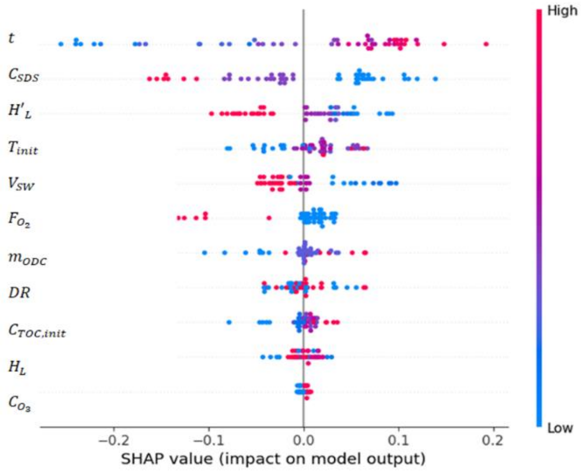

Figure 1. Interpretation of the significance and individual effect of each input variable on the TOC Removal Efficiency in solid phase.

Acknowledgements:
This research has been co-financed by the European Regional Development Fund of the European Union and Greek national funds through the Operational Program Competitiveness, Entrepreneurship, and Innovation, under the call RESEARCH – CREATE – INNOVATE (project code: MIS 5069890 / T2EΔK-01447; project title: “Development of a mobile ozonation unit for the remediation of polluted soils and oil-drilling cuttings”).

References

An M., Guan X. and Yang, N. (2020), Modeling the effects of solid particles in CFD-PBM simulation of slurry bubble columns. Chemical Engineering Science **223**, 115743.

Azubuike C.C., Chikere C.B. and Okpokwasili G.C. (2016), Bioremediation techniques–classification based on site of application: principles, advantages, limitations and prospects. World Journal of Microbiology and Biotechnology **32**, 1-18.

Barwain M., S., Mohamed G. and Daniel, W.S., (2003), Artificial Neural Networks Modeling of Ozone Bubble Columns: Mass Transfer Coefficient, Gas Hold-Up, and Bubble Size. Ozone: Science & Engineering **29**,343-352.

Falciglia P.P., Giustra M.G. and Vagliasindi F.G.A. (2011), Low-temperature thermal desorption of diesel polluted soil: Influence of temperature and soil texture on contaminant removal kinetics, Journal of Hazardous Materials **185**, 392–400.

Hornik K., Stinchcombe M. and White H., (1989), Multilayer feedforward networks are universal approximators. Neural Networks **2**, 359–366.

Ji G.D., Yang Y.S., Zhou Q., Sun T. and Ni J.R. (2004), Phytodegradation of extra heavy oil-based drill cuttings using mature reed wetland: an in situ pilot study. Environment International **30**, 509–517.

Leonard S.A. and Stegemann, J.A., (2010), Stabilization/solidification of petroleum drill cuttings. Journal of Hazardous Materials **174**, 463–472.

Lundberg S.M. and Lee S.I., (2017) A Unified Approach to Interpreting Model Predictions, in: 31st Conference on Neural Information Processing Systems (NIPS 2017), Long Beach, CA, USA. 1–10.

Ribeiro M.T, Singh S. and Guestrin C., (2016). "Why Should I Trust You?": Explaining the Predictions of Any Classifier. 22nd ACM SIGKDD International Conference of Knowledge Discovery and Data Mining.1135-1144.

Robinson J.P., Kingman S.W., Lester E.H. and Yi C., (2012), Microwave remediation of hydrocarbon-contaminated soils – Scale-up using batch reactors. Separation and Purification Technology **96**, 12–19.

Shaikh A. and Al-Dahhan M., (2007), Development of an artificial neural network correlation for prediction of overall gas holdup in bubble column reactors. Chemical Engineering and Processing: Process Intensification **42**, 599-610.

Sun Z., Xia F., Lou Z., Chen X., Zhu N., Yuan H. and Shen, Y., (2020). Innovative process for total petroleum hydrocarbons reduction on oil refinery sludge through microbubble ozonation. Journal of Cleaner Production **256**, 120337.

Zhang Y. and Yang Q., (2018) An overview of multi-task learning, National Science Review **5**, 30–43

Talbi Z., Haddou B., Bouberka Z. and Derriche, Z., (2009). Simultaneous elimination of dissolved and dispersed pollutants from cutting oil wastes using two aqueous phase extraction methods, Journal of Hazardous Materials **163**, 748–755.

Wang H., Wang Y., Lou Z., Zhu N. and Yuan H. (2017), The degradation processes of refractory substances in nanofiltration concentrated leachate using micro-ozonation. Waste Management **69**, 274–280.

Yan P., Lu M., Guan Y., Zhang W. and Zhang Z., (2011), Remediation of oil-based drill cuttings through a biosurfactant-based washing followed by a biodegradation treatment, Bioresource Technology **102**, 10252–10259.