XVI International Conference on Durability of Building Materials and Components
DBMC 2023, China
K.F. Li and D.P. Fang (Eds)

Use of Machine Learning in Processes Optimization for Drinking Water Treatment Plant Butoniga (Istria, Croatia)

G. Volf[1], S. Zorko[2] and I. Sušanj Čule[1]

[1]Department of Hydraulic Engineering, Faculty of Civil Engineering, University of Rijeka, 51000 Rijeka, Croatia, goran.volf@uniri.hr(G. Volf), isusanj@uniri.hr(I. Sušanj Čule)

[2]Istarski Vodovod d.o.o. 52420 Buzet, Croatia, sonja.zorko@ivb.hr(S. Zorko)

Abstract. *Drinking water treatment plant Butoniga is located in central Istria (Croatia) downstream of the Butoniga reservoir. The Butoniga reservoir is an artificial lake created in 1987 with two main objectives: 1) protection from adverse water impacts, and 2) drinking water supply. The operation of the drinking water treatment plant is mainly related to the tourist season, and the 5.000.000 m^3 of produced and distributed drinking water annually, 3.000.000 m^3 is produced and distributed during June 15 to September 15, when the raw water quality in the Butoniga reservoir is the worst. Regarding this, main problem with Butoniga reservoir and thus drinking water treatment plant appear in summer months when water temperature is the most critical parameter and raw water for the treatment process must be captured from the lowest layer of the reservoir which have increased concentrations of Mn, Fe, NH_4 and lower pH values and thus influence on the treatment processes. To deal with this problem, model predicting Mn, seven days in advance is build using machine learning approaches. Build model have high accuracy compared to the measured data, with a good prediction of the peak values. As such, obtained model can help in optimization of the treatment processes which are depending on the quality of raw water, and overall, in sustainability and management of the drinking water treatment plant Butoniga.*

Keywords: *Machine Learning; Processes Optimization; Drinking Water Treatment Plant; Predictive Models; Sustainability; Management; Butoniga reservoir, Manganese.*

1 Introduction

Drinking water treatment plants (DWTP) use mainly mathematical models or empirical formulas as predictive models for their control systems which mainly lack a macro understanding of the overall dynamics, and the nonlinear relationships that are widely present in drinking water treatment. Accurate prediction models for drinking water treatment and production processes must be established to guide water quality processes (Godo-Pla et al. 2019). The amount of data generated along DWTP allows developing data-based models like machine learning (ML) methods which are subfield of artificial intelligence (AI) and are able to predict operational parameters which can be incorporated into environmental decision support systems (DSS) (Li et al., 2021). Unlike mathematical models or statistical methods, ML methods have ability to handle complex nonlinear relationships and an accurate understanding of the overall dynamics of water treatment processes. Therefore, ML methods have the ability to monitor the evolution of water quality, analyse and predict water quality, and also reveal the process of pollutant migration and transformation, thereby shifting the focus from solving existing problems to identifying risks in advance and dynamically optimizing the facilities (Godo-Pla et al. 2019). Also, models based on ML methods have a certain degree of

interpretability, and appropriate analysis methods are capable of mining the hidden physical meaning and chemical or some other information to deepen the understanding of water treatment processes methods (Olden and Jakson, 2002; Park et al. 2019).

Previous studies regarding functioning and problems of the Butoniga DWTP were done by Hajduk-Černeha (2021), Zorko (2017) and Volf et al. (2022). First experiences in use of the Butoniga DWTP for drinking water supply are described by Hajduk-Černeha (2021). Also, in study made by Hajduk-Černeha (2021) is analysed raw water quality from Butoniga reservoir and there are given some management guidelines regarding Butoniga reservoir and related DWTP. In study made by Zorko (2017) is considered the impact of Butoniga reservoir raw water quality on water treatment processes. This study also gives some interesting conclusions which says that main problem with Butoniga reservoir and thus related DWTP appear in summer months when water temperature is the most critical parameter, because, in order to be suitable for use and for treatment processes water mustn't exceeding the maximum allowable concentration (MAC) of 25 ^{o}C according to Croatian regulations for drinking water (Web, 2023). During this that time period water is captured from the lowest water intake which captures water from the lowest water layer in the Butoniga reservoir which have increased concentrations of manganese (Mn), iron (Fe) and ammonium (NH_4) under lower pH values. Increased concentration of Mn, Fe and NH_4 under lower pH values of water from the lowest water intake requires enhanced continuous process control and higher consumption of chemicals for the treatment process on DWTP. During these conditions the process is also stable and all water samples at the effluent are in accordance with the Croatian regulations for drinking water (Web, 2023), where on exceeding values of the water temperature due to the heating of the Butoniga reservoir cannot be influenced (Zorko, 2017). Volf et al. (2022) in his study presents and describes water quality index (WQI) for the Butoniga DWTP and associated WQI prediction models which are used for optimization of treatment processes on the DWTP.

This study deals with the investigations made by Zorko (2017) regarding increased concentrations of Mn, Fe and NH_4 during summer months. To deal with the problem of increased concentrations of Mn, Fe and NH_4 during summer months when higher consumption of chemicals is needed in treatment processes, in this research is obtained prediction model for Mn seven days in advance and are also given relevant conclusions regarding functioning of the Butoniga DWTP using ML technique in form of model trees incorporated in Weka modelling software (Witten and Frank, 2005). The model trees used for numeric prediction use linear equation in the terminal nodes (leaves) which allow a more accurate prediction of the target attribute.

Therefore, specific objective of this study is to develop prediction model for Mn seven days in advance that can be used in performance and optimization of treatment processes which are depending on the quality of raw water on the Butoniga DWTP.

2 Study Area and Data Description

Raw water for treatment processes and obtaining drinking water on Butoniga DWTP is taken from the Butoniga reservoir which is located upstream from the DWTP. Butoniga DWTP is located about 600 m downstream from the dam of the Butoniga reservoir on an area of 80.000 m^2. First phase of Butoniga DWTP is designed to process 1000 l/s or 3600 m^3/h. Parts of the process are designed for a final capacity of 2000 l/s, which is planned in the second phase. All

process units are designed for 24-hour full capacity with a hydraulic reserve of 25 %. The plant can operate flexibly by changing the capacity from 20 to 100 % of the nominal capacity. The main drinking water treatment process (Figure 1) consists of the following units: raw water intake, pre-ozonation, coagulation-flocculation, flotation, rapid filtration, main ozonation, slow sand filter, disinfection, final pH correction, pressure pumping and chlorination. The auxiliary process (Figure 1) of drinking water treatment consists of the following units: station for cleaning sand from slow sand filters, treatment of water from washing filters, sludge treatment, neutralization of wastewater from chemicals. The plant was completed and put into operation in June 2002, while it has been in continuous operation since the spring of 2004 (Zorko, 2017).

The operation of the DWTP is mainly related to the tourist season, and of the 5.000.000 m^3 of produced and distributed water annually, 3.000.000 m^3 is produced and distributed during June 15 to September 15, when the water quality in the Butoniga reservoir is the worst (Zorko, 2017).

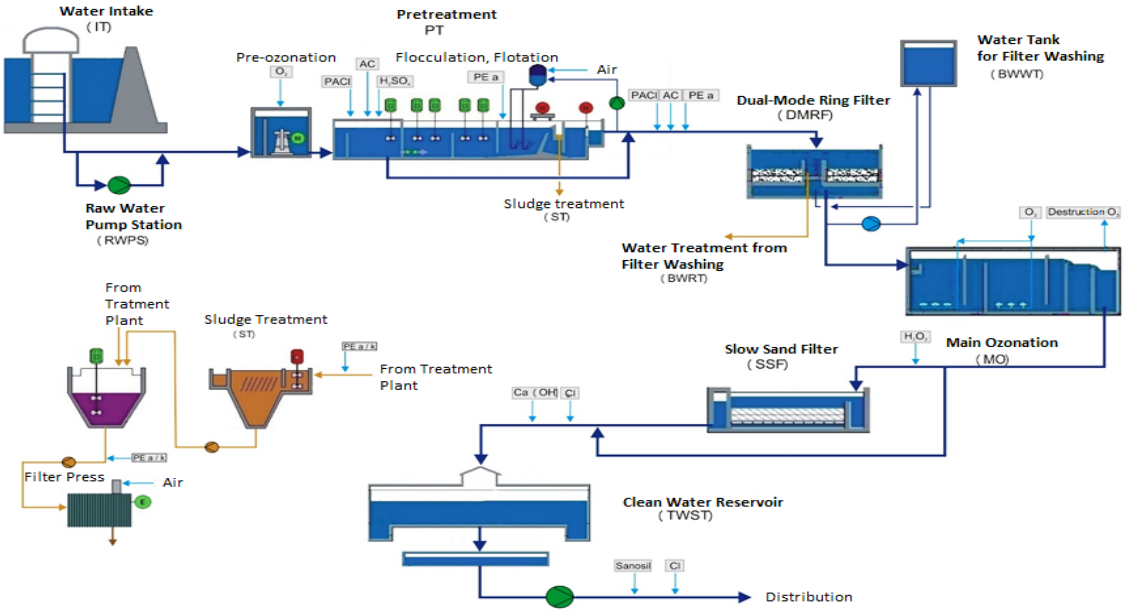

Figure 1. Treatment processes scheme for drinking water treatment plant Butoniga .

The data set for building the prediction model of Mn concentration consists of physical and chemical parameters measured once a day at the input of raw water to the DWTP, from 2011 to 2020 (see Table 1). Physiochemical parameters include reservoir temperature (Temp), pH, turbidity (Tur), oxygen concentration (O_2), total organic carbon (TOC), potassium permanganate ($KMnO_4$), ammonia (NH_4), manganese (Mn), aluminum (Al), iron (Fe) and amount of organic substances (UV254) whose concentration were determinate in internal laboratory of the Butoniga DWTP by standard analytical methods according to ISO standards (2023) and Standard methods for the examination of water and wastewater (2017). With data measured on the DWTP, for prediction models were also used reservoir water level data (Lake level) and data from nearby meteorological station obtained from Croatian Hydro-Meteorological Service (CHMS). This data contain daily precipitations (Prec) and air

temperatures (Air temp). Also, for better prediction of the models sum of 30, 25, 20, 15, 10 and 5 days precipitations were used for modelling purposes which are significant from a hydrological aspect due to precipitation concentration runoff.

All data were pre-processed regarding to modelling and research goal. For the prediction models the entire span of the measured data was used; from 2011 to 2020. Missing data were managed with a cubic spline interpolation.

Table 1. Data used for prediction models.

Symbol	Description	Unit
Temp	Water temperature	^oC
O_2	Oxygen concentration	mg/l
pH	pH	-
Tur	Turbidity	NTU
TOC	Total organic carbon	mg/l
$KMnO_4$	Potassium permanganate	mg/l
UV254	Organic matter in water	1/cm
Al	Aluminium	mg/l
NH_4	Ammonium	mg/l
Mn	Manganese	mg/l
Fe	Iron	mg/l
Prec 30 days	30 days sum of precipitations	mm
Prec 25 days	25 days sum of precipitations	mm
Prec 20 days	20 days sum of precipitations	mm
Prec 15 days	15 days sum of precipitations	mm
Prec 10 days	10 days sum of precipitations	mm
Prec 5 days	5 days sum of precipitations	mm
Prec	Precipitations	mm
Air temp	Air temperature	^oC
Lake level	Lake level	m
Mn_pred	Predicted Manganese 7 days in advance	mg/l

3 Modelling Methods and Design of the Experiment

Model trees are hierarchical structures composed of nodes and branches. Internal nodes contain tests on the input attributes while each branch of an internal test corresponds to an outcome of the test and the predictions for the values of the target variable (i.e. the class) are stored in the leaves which are the terminal nodes in the tree. If the leafs contain a single value for the class prediction, then it is talked about simple regression trees, while if a linear equation is used for prediction in the leaf, then it is talked about model trees (Witten and Frank, 2005; Quinlan, 1993). Induction of model trees is presented on Figure 2.

One of the mostly used algorithm for induction of model trees is the M5 algorithm (Quinlan, 1993), based on the TDIDT top-down induction of decision trees (TDIDT) algorithm (Quinlan, 1986). For the experiments conducted in this research a variation of the M5 algorithm was used, called M5P, implemented in the software package Weka (Witten and Frank, 2005).

After the tree is constructed from the training (learning) set of data, it is necessary to assess the model quality, i.e., the accuracy of prediction. This can be done by simulating the model on

a testing set of data and comparing the predicted values of the target with the actual values. Another option is to employ cross-validation. The given (training) data set is partitioned on a chosen number of folds (n), usually 10. In turn, each fold is used for testing, while the remainder (n-1 folds) is used for training. The final error is the averaged error of all the models throughout the procedure.

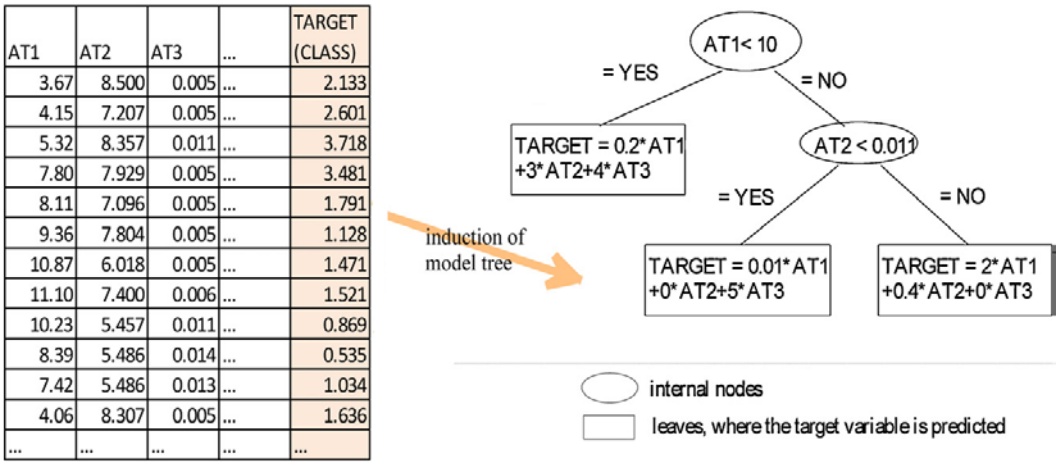

Figure 2. Induction of model trees.

The size of the error between the actual and the predicted values can be calculated by several measures to evaluate the model accuracy: root mean-squared error (RMSE), mean absolute error (MAE), root relative squared error (RRSE), relative absolute error (RAE), and correlation coefficient (R) (Witten and Frank, 2005). In the performed experiments the accuracy of the models is evaluated through all the measures of accuracy.

The data used for building prediction model of Mn are depicted in Table 1. Model is built to predict concentrations of Mn seven days in advance with purpose to improve treatment processes on DWTP regarding to changes of raw water quality in the Butoniga reservoir.

For the experiment the ML algorithm M5P for induction of model trees integrated in the Weka modelling software (Witten and Frank, 2005) was used. Predicted concentrations of Mn seven days in advance were set as a target (dependent) variable, whereas water temperature, pH, turbidity, $KMnO_4$, NH_4, Mn, Al, Fe, O_2, TOC, UV254, Prec 30 days, Prec 25 days, Prec 20 days, Prec 15 days, Prec 10 days, Prec 5 days, Prec, Air temp and Lake level (Table 1) were set as independent variables (descriptors) from which the predicted values of Mn were modelled. The above parameters were mainly used because they best represent the parts of the system (DWTP and Butoniga reservoir) on top of which the target variable relays.

The aim of obtained prediction model is to be as much as possible applicable and valid for the prediction of Mn, meaning that they should perform as accurately as possible. To achieve this, the most commonly used procedures of building and testing models was applied; the entire data set was taken for training while validating with 10-fold cross-validation. To achieve the highest correlation coefficient (R) and the optimal number of rules default values of parameters for building models were used in Weka modelling software (Witten and Frank, 2005).

The model performing most accurately according to the validation method was selected as a

representative model for the prediction purposes.

4 Results and Discussion

Prediction model for Mn concentration is presented on Figure 3, while related model equations are given in Table 2. Model is consisted of nine leaves, i.e. equations were each equation is used to predict Mn concentrations seven days in advance using parameters given in model tree nodes (Figure 3). As can be seen from Figure 3 in model tree nodes, the most appearing parameters are current concentration of Mn, as expected, pH values, sum of 5 days precipitations, water temperature and Fe concentration, while in related equations (Table 2) are contained parameters such as temperature, pH, Mn, Fe and sum of 5 day precipitations. The equation selection in the leaves depends on the values of the variables in the tree nodes, and when selection according to values of the variables in the tree nodes is done, a corresponding equation is applied to calculate the Mn concentration seven days in advance.

Model has very high correlation coefficient of 0.92, while MAE is 0.049, RMSE is 0.084, RAE is 33.71 % and RRSE is 41.12 %.

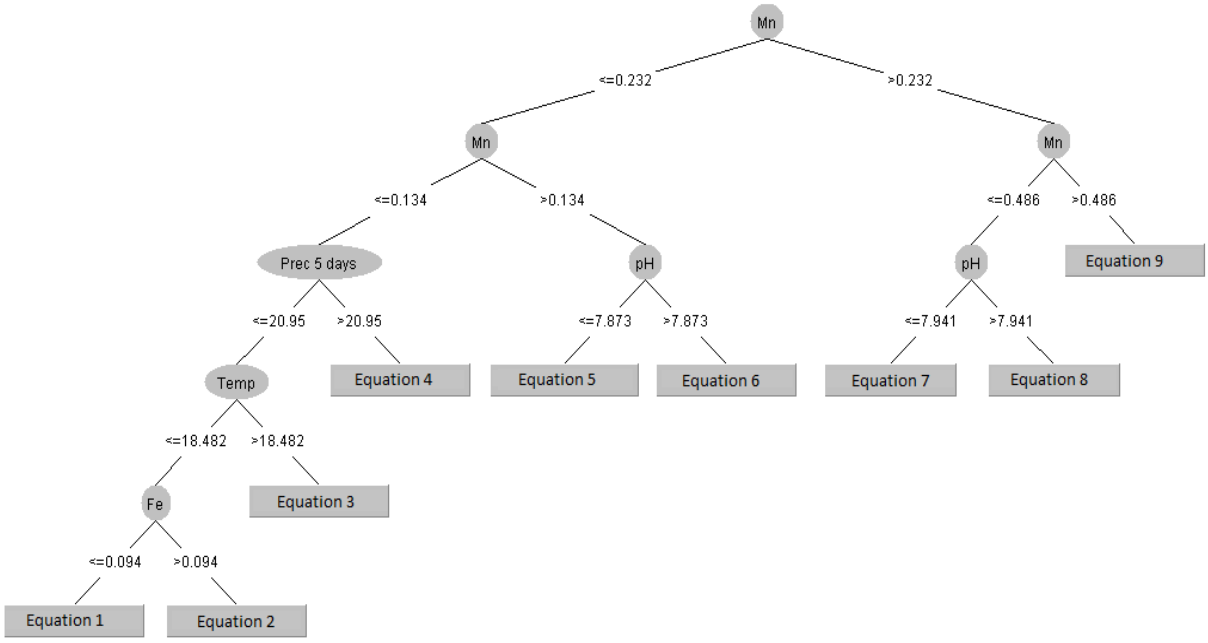

Figure 3. Model tree for manganese prediction 7 days in advance.

The performance of the prediction model is presented on Figure 4. Figure 4 represents the modelled (predicted) vs. measured values of Mn for the given time period from 2011 to 2020 seven days in advance. Figure 3 also indicate very high accuracy of the prediction model with a relatively good prediction of the peak values included.

Overall, obtained results i.e., prediction model is at an acceptable level, looking at the correlation coefficients and prediction of the peak values. As such, model can be used in prediction purposes for optimization of the treatment processes on the DWTP which are depending on the quality of raw water in the Butoniga reservoir.

Table 2. Equations for model tree presented on Figure 6 (manganese prediction).

Equation number	Equations
Equation 1	Mn_pred = 0.0001*Temp - 0.0018*pH + 0.0065*Mn + 0.0276*Fe + 0*Prec 5 days + 0.0742
Equation 2	Mn_pred = 0.0001*Temp - 0.0018*pH + 0.0065*Mn + 0.0273*Fe + 0*Prec 5 days + 0.1071
Equation 3	Mn_pred = 0.0002*Temp - 0.0018*pH + 0.0065*Mn + 0.0345*Fe + 0*Prec 5 days + 0.1365
Equation 4	Mn_pred = 0*Temp - 0.0018*pH + 0.0065*Mn + 0.022*Fe + 0.0001*Prec 5 days + 0.1447
Equation 5	Mn_pred = −0*Temp - 0.0157*pH + 0.0094*Mn + 0.0077*Fe + 0*Prec 5 days + 0.378
Equation 6	Mn_pred = −0*Temp - 0.007*pH + 0.0094*Mn + 0.0077*Fe + 0*Prec 5 days + 0.2279
Equation 7	Mn_pred = −0*Temp - 0.0108*pH + 0.0189*Mn - 0.0011*Fe + 0*Prec 5 days + 0.43
Equation 8	Mn_pred = −0*Temp - 0.015*pH + 0.0189*Mn - 0.0011*Fe + 0*Prec 5 days + 0.3625
Equation 9	Mn_pred = −0*Temp - 0.008*pH + 0.0298*Mn - 0.0011*Fe + 0*Prec 5 days + 0.6041

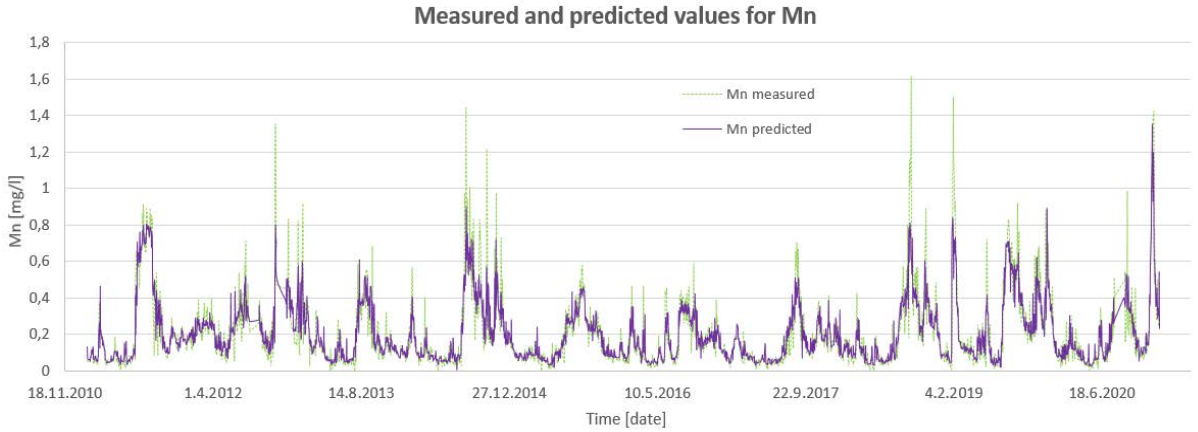

Figure 4. Comparison of measured and predicted values of manganese concentration for modelled period.

In this study, one of the research contributions is obtained prediction model for Mn concentration which was built with use of ML method in form of model trees. This method for numerical predictions use linear equations incorporated in leaves, i.e. the terminal nodes in the tree. Use of this method is very simple, and can be incorporated in DSS for the observed DWTP. Also, model trees in their "decision" nodes use parameters which are important for the predicted variable, and from them can be also seen on which parameters the target variable depends. So, unlike other ML based methods which provide very good predictions, but sometimes are limited in terms of interpretability (black box models), the model trees tend to be more descriptive and interpretable (white box models) which also falls under second contribution.

As mentioned, this research is focused mainly on the predictions of Mn concentration in order to be able to respond on time and to manage the Butoniga DWTP in summer months which in this period have increased concentrations of Mn, Fe and NH4 and it is required enhanced and continuous process control and also higher consumption of chemicals so that treatment process remains stable and all water samples in the effluent stay below MAC.

5 Conclusions

In this study prediction model for Mn concentration seven days in advance was built with use of ML method in form of model trees which was applied on measured data. Model was built to cope with problem of high concentrations of Mn, Fe and NH4 which occurs during summer months when raw water is captured from the lowest water intake in the Butoniga reservoir. Prediction of Mn concentration is done according to current values of measured parameters at the intake of raw water in the Butoniga reservoir. Obtained model has high correlation coefficient and thus provide accurate predictions, correctly predicting the peak values when compared to measured data and can be incorporated in DSS of the Butoniga DWTP. As such, obtained prediction model can also help with the optimization and management of treatment processes at the DWTP, especially during the summer months when the quality of raw water in the Butoniga reservoir is the worst, and where changes in the raw water quality can result in direct action and optimization of the operation of the DWTP.

Acknowledgements

This work was supported by the projects: "Sustainable river basin management by implementation of innovative methodologies, approaches and tools" (uniri-tehnic-18-129) and "Hydrology of water resources and identification of flood and mudflow risk in karst" (uniri-tehnic-18-54). This paper was also funded under the project line ZIP UNIRI of the University of Rijeka, for the project ZIP-UNIRI-1500-3-22.

References

Baird, R.B., Eaton, A.D. and Rice, E.W. (2017). *Standard Methods for the Examination of Water and Wastewater*, 23rd ed., AWWA, USA.

Croatian regulations for drinking water, *Pravilnik o parametrima sukladnosti, metodama analize, monitoringu i planovima sigurnosti vode za ljudsku potrošnju te načinu vođenja registra pravnih osoba koje obavljaju djelatnost javne vodoopskrbe*, available online https://narodne-novine.nn.hr/clanci/sluzbeni/2017_12_125_2848.html. (Accessed: 21 March 2023).

Godo-Pla, L., Emiliano, P., Valero, F., Sin, G., and Monclus, H. (2019). *Predicting the oxidant demand in full-scale drinking water treatment using an artificial neural network: Uncertainty and sensitivity analysis*, Process Saf. Environ. Prot., 125, 317-327.

Hajduk Černeha, B. (2021). *Akumulacija Butoniga u Istri - Prva iskustva u korištenju za vodoopskrbu*, Proceedings of Vodni dnevi, Rimske Toplice, Slovenia, 7.-8. October 2021.

Li, L., Rong, S., Wang, R. and Yu, S. (2021). *Recent advances in artificial intelligence and machine learning for nonlinear relationship analysis and process control in drinking water treatment: A review*, J. Chem. Eng., 405.

Olden, J.D. and Jackson, D.A. (2002). *Illuminating the "black box": a randomization approach for understanding variable contributions in artificial neural networks*, Ecol. Model., 154, 135-150.

Park, S., Baek, S.S., Pyo, J., Pachepsky, Y., Park, J. and Cho, K.H. (2019). *Deep neural networks for modeling fouling growth and flux decline during NF/RO membrane filtration*, J. Membr. Sci., 587.

Quinlan, J.R. (1993). *C4.5: Programs for Machine Learning*, Morgan Kaufmann.

Quinlan, J.R. (1986). *Induction of decision trees*, Machine Learning, 1, 81-106.

Technical Committees, *ISO/TC, Water quality*. Available online: https://www.iso.org/committee/52834/x/catalogue F. (Accessed 4 Feb. 2022).

Volf, G., Sušanj Čule, I., Žic, E. and Zorko, S. (2022). *Water Quality Index Prediction for Improvement of Treatment Processes on Drinking Water Treatment Plant*. Sustainability, 14, 11481.

Witten, I.H. and Frank, E. (2005). *Data Mining: Practical Machine Learning Tools and Techniques*, 2nd ed., Elsevier.

Zorko, S. (2017). *Akumulacija Butoniga - pritisci u slijevu i zaštita voda*, Zbornik radova-Upravljanje jezerima i akumulacijama u Hrvatskoj i Okrugli stol o aktualnoj problematici Vranskog jezera kod Biograda na Moru, Biograd na Moru, Croatia, 4.-6. May 2017.