Applying Anomaly Detection Models in Wastewater Management: A case study of nitrates concentration in the effluent

Pedro Oliveira[1][0000−0001−7143−5413], M. Salomé Duarte[2,3][0000−0003−4645−908X], and Paulo Novais[1][0000−0002−3549−0754]

[1] ALGORITMI Centre, University of Minho, Braga, Portugal
poliveira199208@gmail.com,pjon@di.uminho.pt
[2] CEB - Centre of Biological Engineering, University of Minho, Braga, Portugal
salomeduarte@ceb.uminho.pt
[3] LABBELS – Associate Laboratory, Braga, Guimarães, Portugal

Abstract. With an increase in the diversity of data that companies in our society produce today, extracting insights from them manually has become an arduous task. One of the processes of extracting knowledge from the data is the application of anomaly detection models, which allows for finding unusual patterns in a given dataset. The application of these models in the context of Wastewater Treatment Plants (WWTPs) can improve water quality monitoring in these facilities, alerting decision-makers to act more quickly and effectively on anomalous events. Hence, this study aims to conceive and evaluate several candidate models based on Isolations Forest and Long Short-Term Memory-Autoencoders (LSTM-AE) to detect anomalies in the WWTP effluent, namely in the concentration of nitrates. Considering the obtained results, the best candidate was the LSTM-AE-based model, which had the best performance with an F1-Score of 97% and an AUC-ROC of 98%.

Keywords: Anomaly Detection · Isolation Forests · Long Short-Term Memory-Autoencoders · Nitrates · Wastewater Treatment Plants.

1 Introduction

With the growth in the number of computing devices and their sensing capacity, the diversity of data available in different areas of our society is increasing [1]. Hence, an anomaly detection process, which aims to discover unusual deviations in a given dataset, has become arduous to perform manually [2, 3]. By applying Machine Learning (ML) models to detect anomalies, it is possible to make this process automatic [4]. Within the scope of Wastewater Treatment Plants (WWTPs), anomaly detection models aim to help, in a better and more effective response, the decision-making process.

In the process of treatment in a WWTP, there are a variety of substances that can be taken into account in the context of anomaly detection. In this study, the focus was on the concentration of nitrates present in effluent. Nitrate contamination in water bodies is considered a serious environmental problem since nitrate can cause water quality deterioration and river eutrophication. Additionally, the presence of nitrite due to nitrate reduction is also a threat to humans' health [5]. Therefore, it is of extreme importance the detection of anomalies regarded to nitrate present in effluents, due to the impact that this compound can have on the ecosystems.

Therefore, this study aims to design, evaluate and tune different candidate models for each ML model to detect anomalies in a WWTP, specifically in the nitrate levels present in the effluent released by facilities. For this, we conceived models based on Isolation Forests and Long Short-Term Memory-Autoencoders (LSTM-AE). This manuscript is structured as follows: the next section presents the literature review carried out in the detection of water quality anomalies in WWTPs. The third section focuses on the description of the data exploration and preparation process carried out in this study and a brief explanation of the models and evaluation metrics used. The fourth section presents the experiments carried out throughout this study. The fifth section presents the obtained results and its respective discussion. The conclusions drawn from this study and future work are presented in the last section.

2 State of the Art

In comparison with other areas, in the context of WWTPs, the application of anomaly detection models is not yet fully explored. Although, some studies have already applied an anomaly detection process, especially concerning water quality [6–8].

A study by Mamandipoor et al. [6] focused on using a deep learning model, namely LSTMs, to detect anomalies regarding ammonia in a WWTP. To compare the performance of their model, the authors also used a statistical analysis method and a model based on the Principal Component Analysis-Support Vector Machine (PCA-SVM) for the anomaly detection process. The study was based on a WWTP in northern Italy, with data collection between January and December 2017, labelled as anomalous or non-anomalous by experts in the field. After that, the authors used a random search approach to select some of the best hyperparameters for the LSTM model. The data used were 70% and 30% for training and testing, respectively. To evaluate the performance of the models, the authors used different metrics, one of them being the F1-Score. Through the analysis of the results obtained, it was possible to

verify that the LSTM model presented the best performance, compared to the others, with an F1-Score of 93%. The authors concluded that these results were obtained due to the high capacity that this model has to model temporal dependencies.

Another work carried out by Li et al. [7] aimed to detect contamination events, focusing on various water quality parameters in a water distribution system. The authors propose a stack-based learning model to detect anomalies in substances such as pH or turbidity. This model consisted of a first phase of predicting each parameter using a stacking model. In a second phase, anomalous and non-anomalous events are classified through the residuals between the predicted and measured data and the threshold obtained by the Sequential model-based optimization (SMBO) algorithm. The data used was based on the CANARY dataset, with data on water quality over a period of 4 months. These data were divided into 67% for training and 33% for tests, and cross-validation was used in the training set. The authors also used an Artificial Neural Network (ANN) to compare the designed model for anomaly detection. The F1-score value was higher for the stacked model than for the ANN model in all water quality parameters. For example, in the case of turbidity, the stacked model reached an F1-score of 75%.

Farhi et al. [8] carried out a study where they developed several models to detect anomalies in water quality in a WWTP, namely in ammonia concentration. The authors designed different ML models, such as LSTM-AE, LSTM or Gated Recurrent Units (GRUs). The data used in this study were based on a WWTP installed in Israel using the SCADA platform. The data were divided into 60% for training, 20% for testing and 20% for validation. Also, the data were normalized in an interval between 0 and 1. In the study, the authors performed only the number of epochs, the optimizer to be used, and the number of batch sizes considered in optimizing the hyperparameters. Before the anomaly classification process, the authors used the models designed to predict the values of substances for the next two days. After that, they defined a threshold, and if the predicted values were above the specified threshold, they were classified as anomalies. Concerning the evaluation metrics, the authors used the Accuracy and F1-Score. Through the results obtained, the LSTM-AE model obtained the best performance, in the case of ammonia concentration, with an F1-Score of 88%.

As mentioned, the use of anomaly detection models in WWTPs is still a topic that has not been much explored in the literature. From the analyzed studies, it is possible to verify that the process of searching for the best hyperparameters in the models used should be a more comprehensive topic. The studies that searched for the best hyperparameters were very contained in their scope, which may have achieved better results. The use of LSTM-AE in this context is also somewhat unexplored, and critical issues are not mentioned, such as the prevention of overfitting or the use of cross-validation appropriate to the cases of temporal sequences.

3 Material and Methods

Throughout this section, we will explain the steps developed in collecting, exploring and processing the data used in this study. The ML models used to detect anomalies are also presented as their evaluation metrics.

3.1 Data Collection

Concerning the data collection process, the data used in this study was provided by a multi-municipal Portuguese company responsible by the management of several WWTPs.The data provided based on one of its WWTPs, was collected between August 6^{th}, 2018 and September 28^{th}, 2019.

3.2 Data Exploration

The dataset used in the various experiments carried out in this study is based on the values of nitrates collected in a Portuguese WWTP. This dataset, containing 207 observations, presents a total of two features, namely the value of nitrates and the date on which this value was collected, which are described in Table 1. The data present in the dataset have a periodicity of every two days.

Table 1. Available features in the used datasets.

#	Features	Unit	Description
1	*date_time*	Timestamp	date & time
2	*nitrates_value*	mg/L	nitrates concentration in the effluent

After verifying the number of features and observations in the dataset, the next step was to check for missing values in the entire dataset. When analyzing the data provided, it was possible to verify the existence of 3 missing timesteps, whose treatment is presented in the sub-section corresponding to the data preparation.

The next step focused on the statistical analysis of the column corresponding to the nitrate concentration. With this in mind, we analyzed different metrics such as the mean, the value of Kurtosis and Skewness. The mean value presented by nitrates throughout the dataset was 11.25. Regarding the Skewness, a value of 1.38 was obtained, thus

affirming that the data related to nitrates present a positive asymmetric distribution. Finally, with a Kurtosis value of 1.08, it was possible to conclude that the nitrate data followed a leptokurtic distribution.

To understand the value of nitrates throughout the dataset, we conceived a graph with a set of boxplots by quarters. Through Figure 1, it is possible to verify that the only quart that does not present outliers in the value of nitrates is the third quarter. However, in this quarter, the set of box plots illustrates the greatest dispersion of the data. In addition, the first and third quarters have a non-symmetrical distribution, unlike the second quarter, where this symmetry is evident. The highest outlier is presented in the first quarter, above 40 mg/L of nitrate, while in quantity, the second and fourth quarters have three outliers each.

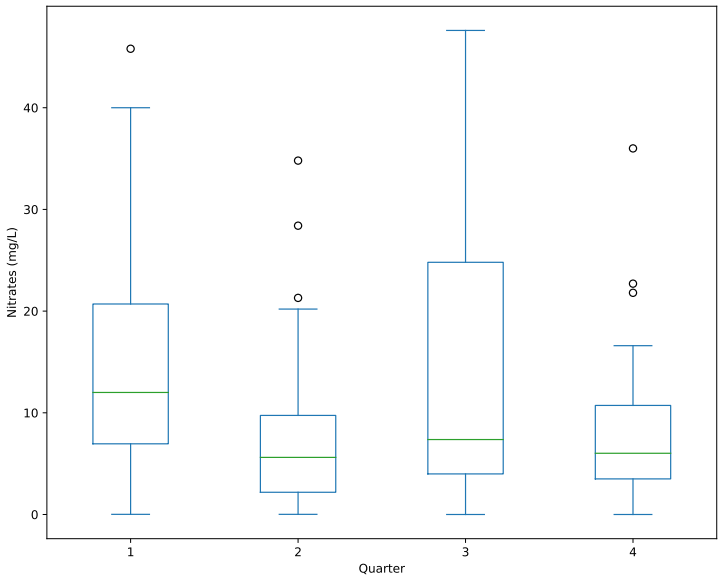

Fig. 1. Distribution of nitrate values per quarter.

The last step in data exploration was verifying if the data followed a Gaussian distribution. This is an essential process for deciding which data correlation analysis approach is used in the data processing phase. For this, the Kolmogorov-Smirnov test was used, with $p < 0.05$. Through the analysis of the obtained results, it was possible to conclude that the data did not follow a Gaussian distribution.

3.3 Data Preparation

The first step in data processing was the treatment of the missing timesteps identified earlier. In the exploratory analysis of the data, three missing timesteps were identified, and as in the case of the LSTM-AE model, the temporal sequence is essential. So it was necessary to insert these three observations into the data. The introduction of these missing timesteps resulted in missing values associated with these observations. To deal with these missing values, linear interpolation was used to fill them in.

Then, we applied a feature engineering process to create new resources to verify the possibility of correlation with the value of nitrates. From the *date_time* attribute, four new functionalities were created: the year, the month, the day of the month and the day of the week.

The next step developed was the correlation analysis. Considering the obtained results in the analysis of the Gaussian distribution of the data, as they did not follow a normal distribution, the nonparametric Spearman's rank correlation coefficient was chosen for correlation analysis. The correlation of the different features with the target was carried out, in this case, the value of nitrates. Through the results obtained, it was possible to verify that none of the features presented a strong correlation with the value of nitrates. Consequently, we removed all attributes, except the target, with the final dataset having 210 observations, sorted ascending by the index (*date_time*). Then, to label the data as anomalous and non-anomalous, to evaluate the performance of the developed models, the data from the final dataset were labelled by specialists. It should be noted that the feature with data labelling only served to evaluate the performance of the models, thus not being used in their training.

Finally, taking into account the use of a model based on LSTMs, the data that feed the LSTM-AE model were normalized considering the *MinMaxScaler*, between -1 and 1.

3.4 Evaluation Metrics

Two evaluation metrics were considered to assess the performance of the different candidate models conceived. Considering that we face a classification problem, the metrics chosen to evaluate the performance were the F-Score and the Area Under the Curve-Receiver Operating Characteristics (AUC-ROC).

The first metric, the F-Score, is a measure used to evaluate binary classification systems. The F-Score is a metric that combines two metrics, in this case, Precision and Recall, and is, therefore, a weighted average of Precision and Recall. When the F-Score value is 1, there is a perfect Precision and Recall. On the other hand, when it has a value of 0, it indicates that the other two metrics are 0 [9]. In this study, the F1-Score was determined according to the following equation:

$$\text{F1-Score} = 2 \times \frac{Precision \times Recall}{Precision + Recall} \tag{1}$$

The second metric, the AUC-ROC, is used to evaluate the performance of classification models. This metric helps to determine the ability of a given model to distinguish classes, where ROC is the probability curve, and AUC represents the degree of separability. A model with an excellent separability measure has an AUC-ROC close to 1, whereas a value close to 0 means worse separability. If the value assigned to the AUC-ROC is 0.5, the model could not separate the classes present in the data [10].

3.5 Isolation Forests

Isolation Forests are a type of anomaly detection model based on decision trees, showing similarities with Random Forests. This model is based on the fact that the anomalies in a given dataset are generally found in small numbers, which are different from most of the data. Unlike other anomaly detection models, Isolation Forests isolate anomalies rather than profiling what non-anomalous instances are [11].

The isolation process carried out in an Isolation Forest model is based on creating several Isolation Trees (iTrees) for a given data set. The instances defined as anomalies are those that are not found in the depths of the iTrees because it was easier for the tree to separate it from the other instances. On the other hand, the instances with greater depth in iTrees are those where there were more cuts to isolate them, thus becoming less likely to be classified as anomalies [11]. Figure 2 presents an example of the architecture of the Isolation Forest model.

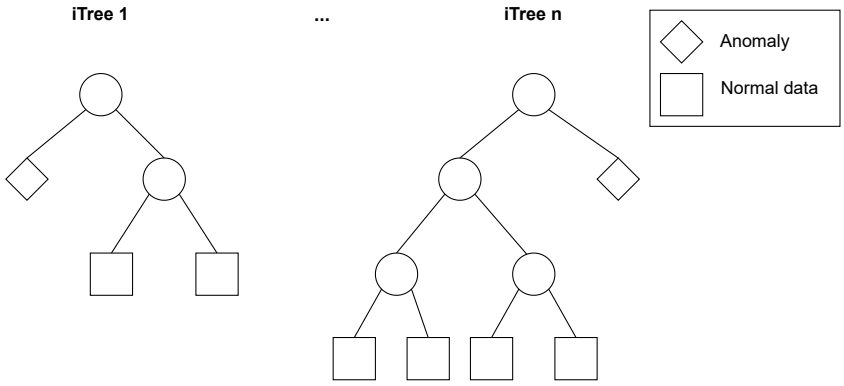

Fig. 2. Example of an Isolation Forest architecture.

3.6 LSTM-Autoencoder

A LSTM-AE is a network that implements an autoencoder for a given sequence of data using LSTMs. An autoencoder aims to reduce the dimensions of the data without changing the main information present in the structure of the data. This type of network is characterized by an input and output layer, encoding and decoding neural network, and a latent space. Through the encoding network, the objective is to compare the data in the latent space. On the other hand, the decoding network focuses on decompressing the encoded representation at the output layer [12].

Using a LSTM-AE, both the encoder and decoder are part of an LSTM network. Taking into account the ability of LSTMs to learn the temporal sequences existing in the data, the combination of an autoencoder with an LSTM makes it possible to perceive patterns of sequential data and recreate the input sequence. The performance of these models is evaluated on the ability of the trained model to recreate the input sequence. In the case of using these models for an anomaly detection process, the critical point is to determine the threshold of the reconstruction error. This threshold can be defined as the maximum mean absolute error loss value. For a given point in the test data, if its reconstruction error is greater than the defined threshold, it is labelled an anomaly [12]. Figure 3 presents an example of a LSTM-AE. In our study, at the end of the LSTM encoder there is a Dropout layer, as well as at the end of the LSTM decoder.

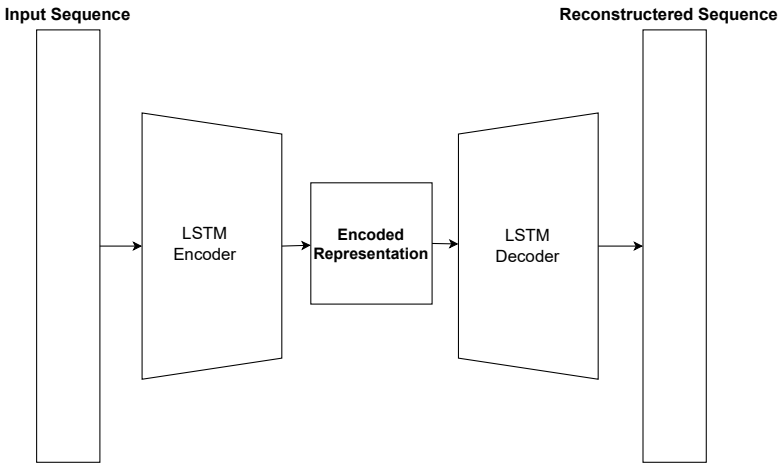

Fig. 3. Example of a LSTM-AE architecture.

4 Experiments

Several experiments were developed following a uni-variate approach to obtain the best candidate model for detecting anomalies in the nitrates present in the effluent of a WWTP. All experiments carried out were evaluated by considering the previously mentioned metrics to discover the best combination of hyperparameters to select the best one. For this search, the grid search technique was used.

Table 2 describes the hyperparameters, and the different values tested, with various combinations thereof, in the experiments carried out.

Table 2. Isolation Forests vs LSTM-AE hyperparameters' searching space.

Parameter	**Isolation Forest**	**LSTM-AE**
Neurons	-	[32,64,128]
Batch size	-	[10,20]
Drop-Out	-	[0.0, 0.2, 0.5]
LSTM layers	-	[1,2,3]
Activation	-	[ReLU, tanh]
Nº Estimators	[100,150,200]	-
Contamination	[0.02, 0.04, 0.06, 0.08, 0.1]	-
Bootstrap	[True, False]	-
Max Samples	[80, 100, 120]	-

In all the experiments, the models were trained without the labels that identify if an observation is an anomaly. These labels were only used in the test data to evaluate the performances of the different candidate models. Data were divided into 70% for training and 30% for testing.

In the case of candidate models based on LSTM-AE, the learning curves were analyzed so that the models did not suffer from overfitting or underfitting. For the models not to go through an overfitting process, the epoch value used was 100. In addition, specific cross-validation for time series was used, namely the *TimeSeriesSplit*, with a value of k equal to 3. The threshold used in the LSTM-AE was defined based on the maximum value of the Mean Absolute Error obtained in the training phase.

Regarding the technologies used in the development of this study, Python 3.9 was the programming language selected for the exploration and processing of data and the conception of the different candidate models. Different libraries were used, such as pandas, scikit-learn and TensorFlow v2.0. The experiments carried out were developed on the hardware made available to Google's Collaboratory.

5 Results and Discussion

With all the experiments carried out, the next step was to analyze their results. In Table 3 and Table 4, it is possible to observe the top-3 of the best candidate for the models based on Isolation Forest and LSTM-AE, respectively. In

these tables, it is possible to verify the value of each hyperparameter for each candidate model and the respective value of the two evaluation metrics taken into account. In addition, the training time of each candidate model is also illustrated.

Table 3. Isolation Forest top-3 candidate models. Legend: a. n_estimators; b. contamination; c. max_samples; d. bootstrap; e. F1-Score; f. AUC-ROC; g. time (in seconds).

a.	**b.**	**c.**	**d.**	**e.**	**f.**	**g.**
100	0.02	80	True	**0.91**	**0.92**	0.243
100	0.08	80	True	0.87	0.88	0.248
100	0.08	80	False	0.83	0.87	0.196

Table 4. LSTM-AE top-3 candidate models. Legend: a. layers; b. neurons; c. activation function; d. dropout-rate; e. batch size; f. F1-Score; g. AUC-ROC; h. time (in seconds).

a.	**b.**	**c.**	**d.**	**e.**	**f.**	**g.**	**h.**
2	64	ReLu	0.5	10	**0.97**	**0.98**	12.787
3	128	ReLu	0.2	20	0.95	0.97	23.469
3	128	tanh	0.2	20	0.94	0.96	24.436

Analyzing the results obtained, expressed in the previous tables, it is possible to verify that the best candidate model is a model based on LSTM-AE with an F1-Score of 0.97 and an AUC-ROC of 0.98. Compared with the other two best candidate models, it is possible to verify that this one needs a smaller number of layers and fewer neurons per layer. On the other hand, the best model based on LSTM-AE required a higher drop-out value than the others. Regarding the activation function, in comparing the three models, there is a higher prevalence of the ReLu function. Also, the best model needed a lower value than the other two models in terms of batch size.

Regarding models based on Isolation Forests, the best candidate model obtained an F1-Score of 0.91 and an AUC-ROC of 0.92. It is, therefore, possible to verify a certain homogeneity in the value of the hyperparameters present in the three best candidate models, mainly in terms of *n_estimators* and *max_samples*. Considering the best model, it is possible to verify that it needs a lower dataset contamination value, in this case, 0.02, than the other two that need a value of 0.08. In terms of the bootstrap hyperparameter, there was a prevalence of the value True, as far as the three best candidate models are concerned.

When comparing the two types of models, as expected, in terms of training time, LSTM-based models have a higher value, as this type of model has a higher computational cost. Considering the evaluation metrics (F1-Score and the AUC-ROC), it is possible to verify that the candidate models based on LSTM-AE are always superior when compared to the Isolation Forests-based models. Focusing on the best candidate models of both models, there is a 6% improvement over the F1-Score and the AUC-ROC. It is also important to note a more pronounced decrease of both metrics in Isolation Forest-based models.

Considering the best candidate model, LSTM-AE-based, Figure 4 illustrates the anomalies detected in nitrate concentration by this model. It is possible to verify that the model detected as anomalies the values above 20 mg/L. At the bottom of the graph, it is clear that the anomalies detected are at values close to 0 mg/L.

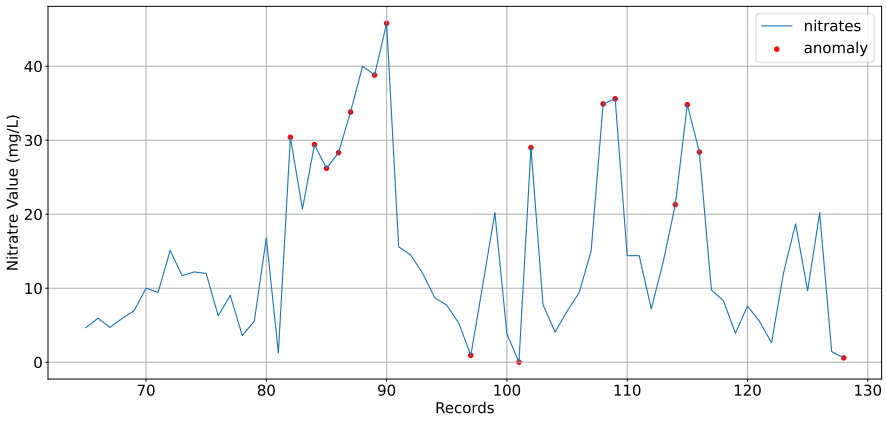

Fig. 4. Anomalies detected by the best candidate model.

6 Conclusions

Automating an anomaly detection process aims to help and alert decision-makers more quickly. In the case of WWTPs, the alert of a possible anomaly in water quality has the consequence that it is possible to act in advance on it, which can lead decision-makers to prevent events throughout the decision process. Therefore, through this study, the objective was to design ML models to detect anomalies in terms of nitrates in the effluent of a WWTP, namely through Isolation Forests and LSTM-AE.

To understand which candidate model had the best performance, considering the focus of the study, we developed several experiments in this sense. The results verified that the best candidate model was LSTM-AE-based, with an F1-Score of 0.97 and an AUC-ROC of 0.98. As expected, the LSTM-AE-based models had a higher training time than the Isolation Forests-based ones. Another conclusion is that in the three best candidate LSTM-AE-based models, there is no marked decrease in their performance, taking into account the two evaluation metrics. On the contrary, this decrease is more evidently verified in the Isolation Forest-based models.

Regarding the following steps to be taken, as future work, the objective is to design more anomaly detection models, such as the One-Class Support Vector Machines, in addition to the development of hybrid models, such as the conjunction of the LSTM-AE with the Isolation Forest. In the case of LSTM-AE, in which the threshold is static in this study, the objective is to apply a threshold moving technique to verify the model's performance with this new approach. In addition, the next step is to use the anomaly detection models for other aspects of WWTPs, such as energy consumption or volumetric flows.

Acknowledgments. This work is financed by National Funds through the Portuguese funding agency, FCT - Fundação para a Ciência e a Tecnologia within project DSAIPA/AI/0099/2019.

References

1. Pang, G., Shen, C., Cao, L., Hengel, A. V. D. (2021). Deep learning for anomaly detection: A review. ACM Computing Surveys (CSUR), 54(2), 1-38. doi: 10.1145/3439950
2. Zenati, H., Romain, M., Foo, C. S., Lecouat, B., Chandrasekhar, V. (2018, November). Adversarially learned anomaly detection. In 2018 IEEE International conference on data mining (ICDM) (pp. 727-736). IEEE. doi:10.1109/ICDM.2018.00088
3. Cook, A. A., Mısırlı, G., Fan, Z. (2019). Anomaly detection for IoT time-series data: A survey. IEEE Internet of Things Journal, 7(7), 6481-6494. doi:10.1109/JIOT.2019.2958185
4. Nassif, A. B., Talib, M. A., Nasir, Q., Dakalbab, F. M. (2021). Machine learning for anomaly detection: a systematic review. IEEE Access. doi: 10.1109/ACCESS.2021.3083060
5. Yun, Y., Li, Z., Chen, Y. H., Saino, M., Cheng, S., Zheng, L. (2018). Elimination of nitrate in secondary effluent of wastewater treatment plants by Fe0 and Pd-Cu/diatomite. Journal of Water Reuse and Desalination, 8(1), 29-37. doi: 10.2166/wrd.2016.122
6. Mamandipoor, B., Majd, M., Sheikhalishahi, S., Modena, C., Osmani, V. (2020). Monitoring and detecting faults in wastewater treatment plants using deep learning. Environmental monitoring and assessment, 192(2), 1-12. doi: 10.1007/s10661-020-8064-1
7. Li, Z., Zhang, C., Liu, H., Zhang, C., Zhao, M., Gong, Q., Fu, G. (2022). Developing stacking ensemble models for multivariate contamination detection in water distribution systems. Science of The Total Environment, 828, 154284. doi: 10.1016/j.scitotenv.2022.154284
8. Farhi, N., Kohen, E., Mamane, H., Shavitt, Y. (2021). Prediction of wastewater treatment quality using LSTM neural network. Environmental Technology & Innovation, 23, 101632. doi: 10.1016/j.eti.2021.101632
9. Tharwat, A. (2020). Classification assessment methods. Applied Computing and Informatics. doi: 10.1016/j.aci.2018.08.003
10. Muschelli, J. (2020). ROC and AUC with a binary predictor: a potentially misleading metric. Journal of classification, 37(3), 696-708. doi:10.1007/s00357-019-09345-1
11. Al Farizi, W. S., Hidayah, I., Rizal, M. N. (2021, September). Isolation Forest Based Anomaly Detection: A Systematic Literature Review. In 2021 8th International Conference on Information Technology, Computer and Electrical Engineering (ICITACEE) (pp. 118-122). IEEE. doi: 10.1109/ICITACEE53184.2021.9617498
12. Tran, P. H., Heuchenne, C., Thomassey, S. (2020). An anomaly detection approach based on the combination of LSTM autoencoder and isolation forest for multivariate time series data. In Developments of Artificial Intelligence Technologies in Computation and Robotics: Proceedings of the 14th International FLINS Conference (FLINS 2020) (pp. 589-596). doi: 10.1142/9789811223334_0071