

The Eurasia Proceedings of Science, Technology, Engineering and Mathematics (EPSTEM), 2026

Volume 42, Pages 184-190

ICBASSET 2026: International Conference on Basic Sciences, Engineering and Technology

Machine Learning-Based Pollution Risk Classification in Industrial Wastewater Treatment

Mustafa Ozan Ozkok
Sakarya University

Abstract: Industrial wastewater contains various pollutants that can impact water quality and the environment. It is important to identify pollution risk conditions in order to monitor wastewater treatment processes effectively. This study compares four machine learning algorithms — Random Forest, Support Vector Machine (SVM), XGBoost and LightGBM — for classifying pollution risk. The Melbourne Eastern Industrial Wastewater Plant dataset was used for this study, containing 600 observations and 14 physicochemical and operational variables. The dataset includes two predefined classes: 'Normal' and 'High Risk'. The models were evaluated using stratified five-fold cross-validation. Accuracy, precision, recall, F1-score and ROC-AUC were used to compare their performance. XGBoost achieved the highest mean accuracy of 93.5%, precision of 90.32%, recall of 91.43%, F1 score of 90.72% and ROC-AUC of 0.9803. LightGBM achieved an accuracy of 93%, followed by random forest with 91.67% and SVM with 90.33%. SHAP analysis was used to examine the contribution of input variables to XGBoost predictions. According to the mean absolute SHAP values, COD and BOD5 were the most influential variables. The results showed that the models could distinguish between the predefined pollution risk categories. However, the risk classes were derived from a composite pollution index, and further validation using independent datasets is needed.

Keywords: Industrial wastewater, Pollution risk classification, Machine learning, XGBoost, SHAP

Introduction

Industrial wastewater is an important source of water pollution due to the different pollutants it may contain. Organic compounds, heavy metals, suspended solids, and nitrogen compounds can affect water quality and aquatic ecosystems when wastewater is discharged without proper treatment (Breida et al., 2019; Zhang & Ren, 2019). For this reason, wastewater treatment and pollution control are important for reducing the environmental effects of industrial activities. Different approaches have been developed to manage industrial pollution and improve wastewater treatment processes (Singh & Mishra, 2023; Kato & Kansha, 2024).

Several physicochemical and operational parameters are used to evaluate the conditions of wastewater treatment. These include Chemical Oxygen Demand (COD), Biochemical Oxygen Demand (BODs), ammonia nitrogen (NH₃-N), pH, temperature and suspended solids, which provide information about the characteristics of the wastewater. Energy consumption, production load and system load are also related to the operation of treatment facilities. Changes in these parameters may indicate changes in wastewater quality or treatment conditions. Trubetskaya et al. (2021) presented a methodology for assessing and monitoring risks in the industrial wastewater sector. In a subsequent study, Trubetskaya et al. (2025) examined tools used to assess risks associated with wastewater and water technologies.

Wastewater quality is commonly monitored through laboratory measurements and the analysis of individual parameters. These measurements are necessary for evaluating treatment conditions, but examining several variables together can provide additional information. Machine learning methods can be used to analyze such data

- This is an Open Access article distributed under the terms of the Creative Commons Attribution-Noncommercial 4.0 Unported License, permitting all non-commercial use, distribution, and reproduction in any medium, provided the original work is properly cited.

- Selection and peer-review under responsibility of the Organizing Committee of the Conference

© 2026 Published by ISRES Publishing: www.isres.org

and identify relationships between variables. Their use in wastewater treatment has been investigated for different purposes, including water quality prediction and process monitoring. Zamfir et al. (2025) reviewed machine learning models applied to wastewater treatment optimization and discussed their applications in monitoring and operational management.

Machine learning has been used in previous studies to address various wastewater-related issues. For example, Hammond et al. (2021) investigated the use of machine learning to detect untreated sewage discharges into watercourses. Elsayed et al. (2022), meanwhile, compared 23 machine learning classification algorithms for mitigating the risk of wastewater treatment. They found that an ensemble bagged trees algorithm achieved the highest classification accuracy. Nguyen et al. (2024) used boosting algorithms to predict the toxicity of industrial wastewater effluent and examine possible control strategies. Shu et al. (2025) applied machine learning to trace the sources of wastewater using fluorescence information, achieving an identification accuracy of 94.25%. While these studies addressed different issues, they demonstrate the application of machine learning in wastewater monitoring, risk classification, toxicity prediction and source identification.

There are several machine learning algorithms available for classification tasks. Random Forest, for example, combines the predictions of multiple decision trees (Breiman, 2001). A Support Vector Machine (SVM) determines a decision boundary between classes and can handle non-linear classification problems through kernel functions (Boser et al., 1992). XGBoost and LightGBM use gradient boosting to build decision trees sequentially (Chen & Guestrin, 2016; Ke et al., 2017). The performance of these algorithms can be compared using the same dataset and evaluation procedure. However, classification metrics alone do not explain which variables influence model predictions. SHapley Additive exPlanations (SHAP), introduced by Lundberg and Lee (2017), can be used to analyse the impact of individual input variables on model predictions.

This study compared Random Forest, SVM, XGBoost and LightGBM for the classification of pollution risk in industrial wastewater treatment. The Melbourne Eastern Industrial Wastewater Plant dataset, containing 600 observations and 14 physicochemical and operational variables, was used for this comparison. The target variable includes two predefined classes, Normal and High Risk, obtained from a composite pollution risk index. The models were evaluated using stratified five-fold cross-validation. Accuracy, Precision, Recall, F1-score, and ROC-AUC were calculated to compare their performance. SHAP analysis was also performed for the XGBoost model to examine the contribution of the input variables. The study focuses on comparing the four algorithms and identifying the parameters associated with the predefined pollution risk categories.

Method

Four machine learning algorithms — Random Forest, Support Vector Machine (SVM), XGBoost and LightGBM — were used to classify pollution risk conditions in industrial wastewater treatment. The dataset was prepared and preprocessed prior to model training. All four models were evaluated using the same dataset and cross-validation procedure. Their classification performance was then compared using different evaluation metrics. SHAP analysis was also applied to the XGBoost model to examine the contribution of the input variables to its predictions.

Dataset

For the classification experiments, data from the Melbourne Eastern Industrial Wastewater Plant were used (Melbourne Eastern Industrial Wastewater Plant, Kaggle). The dataset is available on Kaggle and includes 600 records. Each record contains 14 measurements related to wastewater quality and plant operation. These measurements are heavy metals, COD, BOD5, suspended solids, NH₃-N, pH, temperature, conductivity, flow velocity, energy consumption, production load, effluent flow, daily output, and system load.

The 'Pollution_Risk' column was used as the target variable. Of the 600 records, 390 were classified as 'Normal' and 210 as 'High Risk'. In the original dataset, 'Normal' was coded as 1 and 'High Risk' as 0, and these labels were determined using a pollution index based on COD, BOD5, NH₃-N, energy consumption and suspended solids. The 65th percentile of the index was used as the threshold. Records exceeding this threshold were assigned to the 'High Risk' class, while the others were assigned to the 'Normal' class.

Data Preprocessing

The dataset contained no missing values. For the classification task, the original class labels were changed to 1 for High Risk and 0 for Normal. All 14 numerical features were used in the experiments. Feature standardization was performed only for SVM using StandardScaler. In each fold, the scaling parameters were obtained from the training set and then applied to the test set. This ensured that the test data did not influence the preprocessing stage.

Machine Learning Models

In this study, three tree-based machine learning algorithms and SVM were used to classify pollution risk conditions in industrial wastewater treatment. The models were evaluated using the same input variables and cross-validation procedure to compare their classification performance. Random Forest combines the predictions of multiple decision trees (Breiman, 2001). In this study, the model was trained with 300 trees. For SVM, the radial basis function (RBF) kernel was selected, and the regularization parameter was set to 1.0. The algorithm uses the maximum margin principle to separate observations belonging to different classes (Boser et al., 1992).

Two gradient boosting algorithms, XGBoost and LightGBM, were also included in the experiments. XGBoost constructs decision trees sequentially, with each new tree contributing to the correction of previous prediction errors (Chen & Guestrin, 2016). A total of 300 estimators were used, together with a maximum tree depth of 4 and a learning rate of 0.05. LightGBM was included as another tree-based boosting method with an emphasis on computational efficiency (Ke et al., 2017). Its configuration consisted of 300 estimators, a learning rate of 0.05, and 15 leaves. All experiments used a random seed of 42.

Experimental Setup and Performance Evaluation

The models were evaluated using stratified five-fold cross-validation. The 600 observations were divided into five folds with similar class distributions. Four folds were used to train the models, and the remaining fold was used for testing. The training and testing folds were changed in each iteration. After five iterations, every observation had been used once for testing. Model performance was evaluated using accuracy, precision, recall, F1-score and ROC-AUC. Precision, recall and F1 score were calculated by considering high risk as the positive class. The mean and standard deviation of the five folds were calculated for each performance metric. Additionally, out-of-fold predictions were combined to construct confusion matrices and ROC curves.

SHAP Analysis

SHAP (SHapley Additive exPlanations) was used to examine how the input variables affected the predictions of XGBoost. Following the five-fold cross-validation experiments, XGBoost was trained again on all 600 observations. SHAP values were then calculated for the input variables, and their mean absolute values were used to compare feature importance. This analysis focused on the physicochemical and operational variables associated with the predicted pollution risk classes.

Results and Discussion

The classification performances of Random Forest, SVM, XGBoost, and LightGBM were evaluated using stratified five-fold cross-validation. Accuracy, Precision, Recall, F1-score, and ROC-AUC were used to compare the models. The mean and standard deviation values obtained from the five folds are presented in Table 1.

Table 1. Classification performance of machine learning models (mean $\pm$ standard deviation)

Model	Accuracy	Precision	Recall	F1-score	ROC-AUC
Random Forest	0.9167 $\pm$ 0.0156	0.8957 $\pm$ 0.0456	0.8667 $\pm$ 0.0643	0.8786 $\pm$ 0.0258	0.9692 $\pm$ 0.0146
SVM	0.9033 $\pm$ 0.0095	0.8604 $\pm$ 0.0337	0.8667 $\pm$ 0.0361	0.8625 $\pm$ 0.0129	0.9702 $\pm$ 0.0033
XGBoost	0.9350 $\pm$ 0.0137	0.9032 $\pm$ 0.0287	0.9143 $\pm$ 0.0598	0.9072 $\pm$ 0.0232	0.9803 $\pm$ 0.0047
LightGBM	0.9300 $\pm$ 0.0095	0.9025 $\pm$ 0.0344	0.9000 $\pm$ 0.0543	0.8996 $\pm$ 0.0170	0.9783 $\pm$ 0.0068

As shown in Table 1, all four models achieved high classification performance. XGBoost obtained the highest mean Accuracy of 0.9350, followed by LightGBM with 0.9300, Random Forest with 0.9167, and SVM with

0.9033. XGBoost also achieved the highest Precision (0.9032), Recall (0.9143), F1-score (0.9072), and ROC-AUC (0.9803). These results indicate that XGBoost provided the highest overall classification performance among the evaluated models.

The performance of LightGBM was close to that of XGBoost, with a difference of 0.0050 in mean Accuracy. LightGBM also achieved an F1-score of 0.8996. Although Random Forest had a higher Accuracy than SVM, SVM achieved a slightly higher ROC-AUC. This difference was also observed in the comparison of the evaluation metrics. Recall was evaluated by considering High Risk as the positive class. LightGBM achieved a mean Recall of 0.9000, while Random Forest and SVM both obtained 0.8667. Compared with these models, XGBoost achieved a higher Recall and identified a larger proportion of high-risk observations. The mean ROC-AUC values were 0.9783 for LightGBM, 0.9702 for SVM, and 0.9692 for Random Forest. All four models achieved ROC-AUC values above 0.96. The ROC curves of the models are presented in Figure 1.

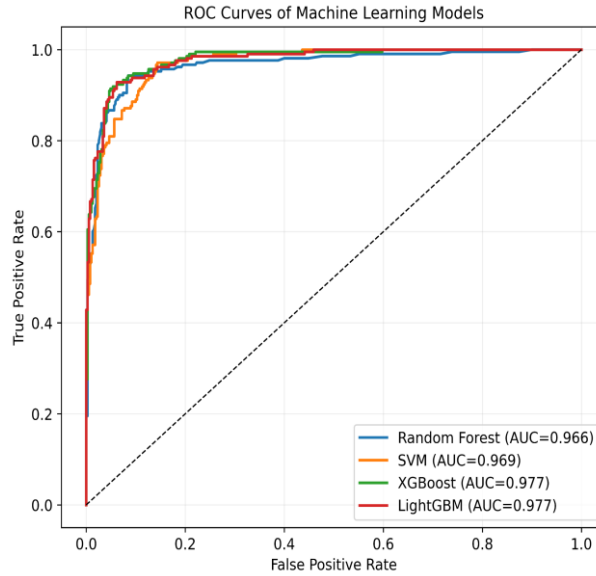


Figure 1. ROC curves of the four machine learning models based on combined out-of-fold predictions.

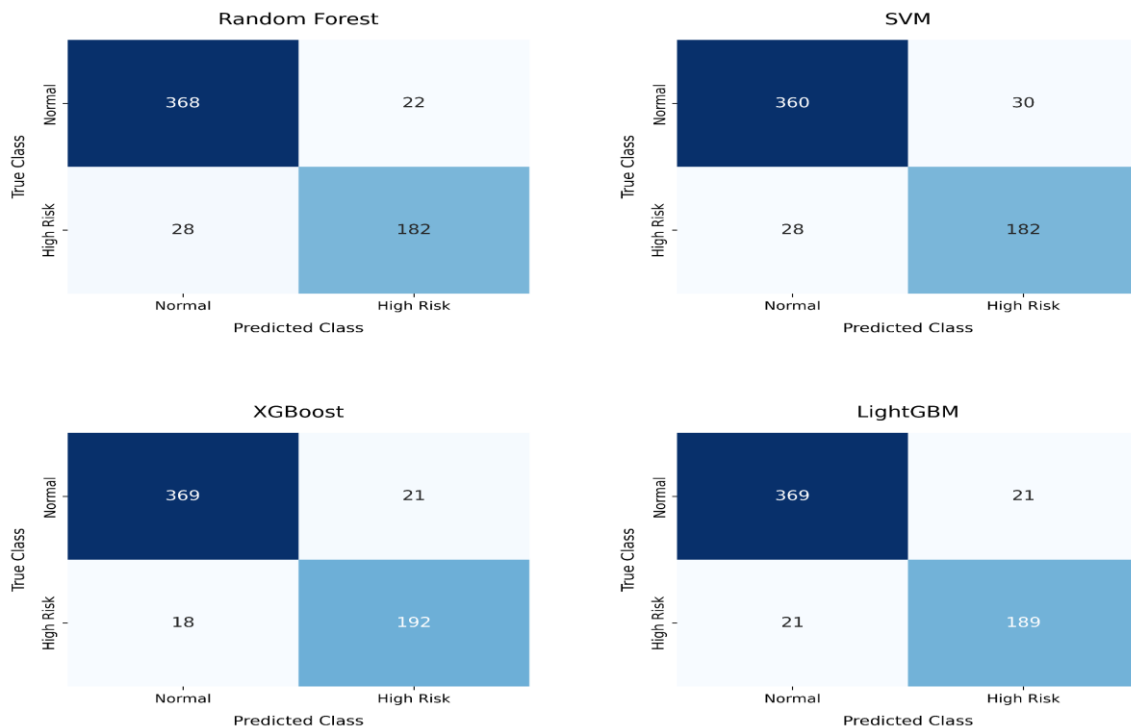


Figure 2. Confusion matrices of the four machine learning models.

The pollution risk labels were generated using a composite index based on selected wastewater parameters. Since COD and BOD5 were included in this index, their contribution to the classification results was expected. Therefore, the model performances represent the classification of the risk levels defined in the dataset and do not directly indicate their ability to detect actual pollution events.

Figure 2 shows the confusion matrices obtained from the combined out-of-fold predictions. XGBoost correctly classified 369 Normal and 192 High Risk observations, while 18 High Risk observations were incorrectly assigned to the Normal class. LightGBM correctly classified 189 High Risk observations, compared with 182 for both Random Forest and SVM. XGBoost had the lowest number of missed High Risk observations among the four models.

The SHAP results for XGBoost are given in Figure 3. Among the 14 input variables, COD had the largest mean absolute SHAP value (4.3872). BOD5 followed with a value of 1.4431. Suspended solids, ammonia nitrogen, and conductivity had much lower values of 0.3019, 0.2725, and 0.2238, respectively. The difference was particularly large between BOD5 and the remaining variables. COD and BOD5 were also among the five variables used to calculate the pollution risk index. Their high importance in the XGBoost model should therefore be considered together with the way the risk classes were defined.

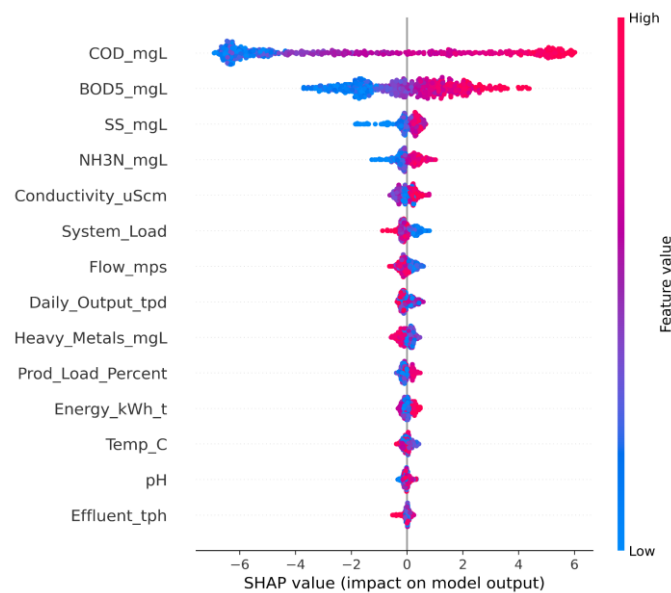


Figure 3. SHAP summary plot showing the contribution of input variables to XGBoost predictions.

Conclusion

This study investigated the classification of pollution risk conditions in industrial wastewater treatment using four machine learning algorithms: Random Forest, SVM, XGBoost and LightGBM. The Melbourne Eastern Industrial Wastewater Plant dataset was used, containing 600 observations and 14 input variables. Model performance was evaluated using stratified five-fold cross-validation based on accuracy, precision, recall, F1-score and ROC-AUC.

The results of the experiments showed that all four models achieved high classification performance. XGBoost performed the best overall, achieving an accuracy of 93.5%, precision of 90.32%, recall of 91.43%, an F1 score of 90.72% and a ROC-AUC of 0.9803. LightGBM achieved comparable results with an accuracy of 93%, followed by random forest and SVM with 91.67% and 90.33%, respectively. XGBoost correctly classified 192 out of 210 high-risk observations. SHAP analysis showed that COD and BOD5 had the greatest influence on XGBoost predictions. Both variables were also used to calculate the composite index from which the pollution risk labels were generated. Therefore, their importance in the model predictions should be interpreted in relation to the labeling method.

The four models were able to classify the predefined pollution risk categories with high accuracy. However, the dataset did not contain independently observed pollution events, so the reported results are limited to the risk categories defined by the composite index. Further studies could use larger wastewater datasets and independent

test data to examine how well these models perform on new observations. Different feature selection methods could also be evaluated to determine their effects on classification performance.

Recommendations

Future studies should investigate larger and independent wastewater datasets to evaluate the generalization performance of machine learning models. Different feature selection and hyperparameter optimization methods could be applied to improve classification performance. In addition, the integration of machine learning models into wastewater monitoring systems could be investigated to support early pollution risk detection.

Scientific Ethics Declaration

* The author declares that the scientific ethical and legal responsibility of this article published in EPSTEM Journal belongs to the author.

Conflict of Interest

* The author declares that there is no conflict of interest.

Funding

* This study was supported by COST Action CA23104 – Mainstreaming Water Reuse into the Circular Economy Paradigm (Water4Reuse).

Acknowledgements or Notes

* This article was presented as an oral presentation at the International Conference on Basic Sciences, Engineering and Technology (www.icbaset.net) held in Istanbul/Türkiye on September 17-18, 2026.

References

- Boser, B. E., Guyon, I. M., & Vapnik, V. N. (1992). A training algorithm for optimal margin classifiers. In *Proceedings of the Fifth Annual Workshop on Computational Learning Theory* (pp. 144–152). Association for Computing Machinery.
- Breida, M., Younssi, S. A., Ouammou, M., Bouhria, M., & Hafsi, M. (2019). Pollution of water sources from agricultural and industrial effluents: Special attention to NO₃⁻, Cr(VI), and Cu(II). In *Water chemistry*. IntechOpen.
- Breiman, L. (2001). Random forests. *Machine Learning*, 45(1), 5–32.
- Chen, T., & Guestrin, C. (2016). XGBoost: A scalable tree boosting system. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* (pp. 785–794). Association for Computing Machinery.
- Elsayed, A., Siam, A., & El-Dakhkhni, W. (2022). Machine learning classification algorithms for inadequate wastewater treatment risk mitigation. *Process Safety and Environmental Protection*, 159, 1224–1235.
- Hammond, P., Suttie, M., Lewis, V. T., Smith, A. P., & Singer, A. C. (2021). Detection of untreated sewage discharges to watercourses using machine learning. *npj Clean Water*, 4(1), 18.
- Kato, S., & Kansha, Y. (2024). Comprehensive review of industrial wastewater treatment techniques. *Environmental Science and Pollution Research*, 31(39), 51064–51097.
- Ke, G., Meng, Q., Finley, T., Wang, T., Chen, W., Ma, W., Ye, Q., & Liu, T.-Y. (2017). LightGBM: A highly efficient gradient boosting decision tree. In *Advances in Neural Information Processing Systems* (Vol. 30, pp. 3146–3154).
- Lundberg, S. M., & Lee, S.-I. (2017). A unified approach to interpreting model predictions. In *Advances in Neural Information Processing Systems* (Vol. 30, pp. 4765–4774).

- Melbourne Eastern Industrial Wastewater Plant. (n.d.). *Melbourne Eastern Industrial Wastewater Plant* [Data set]. Kaggle. Retrieved August 15, 2026.
- Nguyen, D. V., Park, J., Lee, H., Han, T., & Wu, D. (2024). Assessing industrial wastewater effluent toxicity using boosting algorithms in machine learning: A case study on ecotoxicity prediction and control strategy development. *Environmental Pollution*, 341, 123017.
- Shu, Y., Kong, F., He, Y., Chen, L., Liu, H., Zan, F., Lu, X., Wu, T., Si, D., Mao, J., & Wu, X. (2025). Machine learning-assisted source tracing in domestic-industrial wastewater: A fluorescence information-based approach. *Water Research*, 268, 122618.
- Singh, G., & Mishra, S. (2023). Industrial pollution management approach. In *Heavy metal toxicity: Environmental concerns, remediation and opportunities* (pp. 363–388). Springer Nature Singapore.
- Trubetskaya, A., Horan, W., Conheady, P., Stockil, K., Merritt, S., & Moore, S. (2021). A methodology for assessing and monitoring risk in the industrial wastewater sector. *Water Resources and Industry*, 25, 100146.
- Trubetskaya, A., Hennessy, A., Ryan, I., Stockil, K., & Gaskin, C. (2025). Industrial assessment tools of risks associated with wastewater and water technologies. *ACS ES&T Water*, 5(10), 6131–6151.
- Zamfir, F. S., Carbureanu, M., & Mihalache, S. F. (2025). Application of machine learning models in optimizing wastewater treatment processes: A review. *Applied Sciences*, 15(15), 8360.
- Zhang, X., & Ren, H. (Eds.). (2019). *High-risk pollutants in wastewater*. Elsevier.

Author(s) Information

Mustafa Ozan Ozkok

Sakarya University

Sakarya, Türkiye

ORCID iD: <https://orcid.org/0000-0001-8109-3197>

*Contact e-mail: mustafaozanzkok@gmail.com

*Corresponding author(s) contact email

To cite this article:

Ozkok, M.O. (2026). Machine learning-based pollution risk classification in industrial wastewater treatment. *The Eurasia Proceedings of Science, Technology, Engineering and Mathematics (EPSTEM)*, 42, 184–190.