

FROM RAW ENVIRONMENTAL DATA TO RELIABLE MACHINE LEARNING MODELS: A DATA QUALITY-DRIVEN FRAMEWORK WITH APPLICATION IN WASTEWATER TREATMENT

Nikolina Ivanova

University of Mining and Geology “St. Ivan Rilski”, 1700 Sofia; E-mail: nikolina.ivanova@mgu.bg

ABSTRACT. Modern environmental monitoring systems generate large volumes of heterogeneous data from sensor networks, laboratory analyses, operational systems, and external information sources. Data quality determines the reliability, accuracy, and interpretability of machine learning models used for environmental analysis, forecasting, and decision support. This review examines the major challenges of preparing environmental monitoring data for machine learning by analysing their characteristics, key data quality dimensions, and contemporary preprocessing techniques.

The main contribution of this paper is a conceptual framework for data quality-driven data preparation. Unlike conventional approaches, where preprocessing follows a predefined sequence of operations, the proposed framework places data quality assessment at the core of preprocessing strategy selection. The four principal data quality dimensions—completeness, reliability, temporal consistency, and informativeness—are linked to corresponding preprocessing techniques, including imputation, noise and anomaly filtering, temporal synchronisation, and feature selection or dimensionality reduction.

Wastewater treatment plants are used as a representative application illustrating the typical challenges of environmental data analysis. However, the proposed framework is applicable to a wide range of environmental monitoring systems and provides a systematic basis for preparing reliable, interpretable, and practically applicable datasets for machine learning.

Key words: environmental monitoring, data quality assessment, data preprocessing, machine learning, wastewater treatment plants.

ОТ СУРОВИ ЕКОЛОГИЧНИ ДАННИ ДО НАДЕЖДНИ МОДЕЛИ ЗА МАШИННО ОБУЧЕНИЕ: КОНЦЕПТУАЛНА РАМКА, ОСНОВАНА НА ОЦЕНКАТА НА КАЧЕСТВОТО НА ДАННИТЕ, С ПРИЛОЖЕНИЕ ПРИ ПРЕЧИСТВАНЕТО НА ОТПАДЪЧНИ ВОДИ

Николина Иванова

Минно-геоложки университет „Св. Иван Рилски“, 1700 София

РЕЗЮМЕ. Съвременните системи за екологичен мониторинг генерират големи обеми хетерогенни данни от сензорни мрежи, лабораторни анализи, експлоатационни системи и външни информационни източници. Качеството на тези данни определя надеждността, точността и интерпретируемостта на моделите за машинно обучение, използвани за анализ, прогнозиране и подпомагане на вземането на решения. Настоящият обзор разглежда основните предизвикателства при подготовката на екологични данни за машинно обучение, като анализира техните характеристики, ключовите измерения на качеството и съвременните подходи за предварителна обработка.

Основният принос на статията е предложената концептуална рамка за подготовка на данни, управлявана от оценката на тяхното качество. За разлика от традиционните подходи, при които предварителната обработка следва предварително определена последователност, предложената рамка поставя оценката на качеството на данните в основата на избора на подходящи стратегии за предварителна обработка. Четирите основни измерения на качеството на данните – пълнота, надеждност, времева съгласуваност и информативност – са свързани със съответните методи за предварителна обработка, в т.ч. импутация, филтриране на шум и аномалии, синхронизиране на времеви редове и подбор или намаляване на размерността на признаците.

Като представителен пример са използвани пречиствателните станции за отпадъчни води, които илюстрират типичните предизвикателства при анализа на екологични данни. Предложената рамка има и по-широка приложимост и може да се използва при подготовката на данни от различни системи за екологичен мониторинг, създавайки основа за разработването на по-надеждни, интерпретируеми и практически приложими модели за машинно обучение.

Ключови думи: екологичен мониторинг, оценка на качеството на данните, предварителна обработка на данни, машинно обучение, пречиствателни станции за отпадъчни води.

Introduction

Driven by advances in computational methods and environmental sensing technologies, machine learning (ML) is now widely used to analyse, model, and optimise environmental processes. The widespread deployment of sensor technologies, Internet of Things (IoT) devices, SCADA platforms, and automated monitoring networks has resulted in the continuous generation of large and highly heterogeneous environmental datasets. These developments have created new opportunities for prediction, anomaly detection, and decision support in a wide range of applications, including air quality monitoring, water resource management, precision agriculture, ecosystem monitoring, and wastewater treatment.

Despite these opportunities, developing reliable machine learning models remains challenging because environmental datasets often contain quality-related limitations. Environmental monitoring data originate from multiple heterogeneous sources

and differ in sampling frequency, measurement accuracy, and temporal resolution. As a result, the quality of the available data directly influences the performance, robustness, and interpretability of machine learning models.

Although data preprocessing is traditionally regarded as a technical step within the analytical workflow, preprocessing strategies should be selected according to the characteristics and quality of the available data rather than applied universally. Therefore, data quality assessment should be regarded as the foundation for selecting appropriate preprocessing methods. Despite the rapid growth of machine learning applications in environmental monitoring, considerably less attention has been devoted to systematically linking data quality assessment with preprocessing strategy selection.

This review examines the relationship between environmental monitoring challenges, data quality assessment, and data preprocessing for machine learning. Based on a comprehensive literature review, it proposes a conceptual data

quality-driven framework linking data quality dimensions with appropriate preprocessing strategies. Although wastewater treatment plants (WWTPs) are used as a representative application, the proposed framework is applicable to a broad range of environmental monitoring systems.

Environmental Monitoring Data: Challenges for Machine Learning

Environmental monitoring data differ substantially from the datasets commonly used in many traditional machine learning applications, posing unique challenges for machine learning because of their heterogeneity, dynamic behaviour, and complex environmental interactions (Zhong et al., 2021). They describe dynamic natural and engineered processes, are collected from multiple independent sources, and are characterised by varying sampling frequencies, measurement accuracies, and spatial-temporal structures, presenting unique challenges that directly influence the reliability and performance of machine learning models.

Heterogeneous Data Sources

One of the primary challenges arises from the heterogeneity of data sources. Depending on the application, environmental datasets may integrate measurements obtained from online sensors, laboratory analyses, SCADA systems, meteorological observations, remote sensing platforms, geographic information systems (GIS), and operational records. These sources differ in their measurement principles, data formats, sampling frequencies, and levels of accuracy, making their integration into a unified analytical dataset a challenging task (Ullo & Sinha, 2020). Accordingly, heterogeneous data integration is the first step in preparing environmental datasets for machine learning.

Missing and Unreliable Observations

Real-world environmental datasets frequently contain missing values, noise, outliers, and measurement errors resulting from sensor malfunctions, communication failures, adverse weather conditions, or human intervention (Alferes & Vanrolleghem, 2016). Such imperfections reduce data reliability and may lead to inaccurate predictions, unstable models, and misleading interpretations (Aribisala, 2024). Reliable machine learning therefore requires identifying missing and unreliable observations before model development.

Temporal Heterogeneity

Environmental processes exhibit pronounced temporal dynamics. Measurements may be collected at intervals ranging from seconds to days or even weeks, while the monitored variables are influenced by seasonal patterns, climatic conditions, and the specific characteristics of the observed system.

Differences in sampling frequency and temporal resolution complicate the analysis of relationships among variables, making temporal alignment a prerequisite for reliable machine learning in environmental monitoring (Zhu et al., 2022).

High-Dimensional Environmental Datasets

Modern environmental monitoring systems simultaneously record a large number of physical, chemical, biological, and operational variables. These variables are often characterised by complex linear and non-linear relationships, while some

features contain redundant or only marginally informative information (Bagherzadeh et al., 2021). High data dimensionality increases computational complexity, complicates result interpretation, and may reduce the performance and generalisability of machine learning models (Aribisala, 2024). These challenges make feature selection and dimensionality reduction essential steps in preparing environmental datasets for machine learning.

Wastewater Treatment as an Illustrative Case Study

The application of machine learning in wastewater treatment has expanded rapidly during the last decade and has been comprehensively reviewed from different perspectives, including biological treatment processes, process optimisation, operational control, and AI-driven decision support (Sundui et al., 2021; Rashid et al., 2025; Jaganathan et al., 2025; Zamfir et al., 2025). As a result, WWTPs have become one of the most representative environmental systems for developing and evaluating data-driven models.

WWTPs represent a typical example of environmental systems in which all of the previously discussed challenges occur simultaneously. They integrate continuous measurements from online sensors, laboratory analyses of water quality, operational process parameters, SCADA systems, and external meteorological observations (Newhart et al., 2019). The diversity of data sources, sampling frequencies, and levels of reliability results in complex multidimensional datasets containing missing values, noise, outliers, and temporal inconsistencies (Wang et al., 2021).

Numerous machine learning algorithms, including artificial neural networks, Random Forest, and model tree approaches, have already demonstrated their potential for predicting wastewater treatment performance (Kumar & Deswal, 2020). Machine learning has also been successfully applied to specialised monitoring tasks such as odor classification around wastewater treatment facilities (Cangialosi et al., 2021).

Although these challenges are commonly discussed as independent preprocessing tasks, they rarely occur in isolation. Their interaction highlights the need to view data quality assessment as the basis for selecting appropriate preprocessing strategies, which forms the central concept of the framework proposed in this review.

Data Quality Dimensions for Machine Learning Readiness

Data quality assessment forms the foundation of reliable machine learning because it determines the suitability of the available data for machine learning and guides the selection of appropriate preprocessing strategies. Its importance is particularly evident in environmental monitoring systems, where heterogeneous data sources, long-term observations, and dynamic processes introduce multiple quality-related challenges.

Unlike traditional information systems, where data quality is primarily evaluated in terms of correctness and completeness, machine learning applications require additional characteristics that directly influence model training, generalisation, and robustness. In this review, data quality is organised into four functional dimensions – completeness, reliability, temporal consistency, and informativeness – which encompass the most common challenges encountered when preparing

environmental monitoring data for machine learning. Their relationships, associated challenges, impacts on machine

learning, and corresponding preprocessing strategies are summarised in Table 1.

Table 1. Data quality assessment framework

DATA QUALITY DIMENSIONS			
Dimension	Typical problems	Impact on ML	Recommended methods
 Completeness Extent to which data are complete and not missing	 Missing data ratio  Patterns of missingness  Temporal coverage	 Reduced prediction accuracy  Biased model estimates	 Imputation  Interpolation
 Reliability Degrees to which data are accurate and consistent with reality	 Outlier rate  Sensor/instrument reliability  Measurement error	 Model instability  Overfitting	 Filtering  Outlier detection
 Temporal consistency Degree to which data are temporally aligned and consistent	 Inconsistent sampling frequency  Time lag/ synchronization  Non-stationarity	 Distorted temporal relationships  Reduced forecasting performance	 Aggregation  Resampling  Synchronization
 Informativeness Degree to which data contain useful information for the target task	 Low signal-to-noise ratio  Feature redundancy  Multicollinearity	 Increased model complexity  Reduced interpretability	 Principal component analysis (PCA)  Feature selection  Factor analysis
			
QUALITY ASSESMENT SUMMARY			
Summary of identified strengths, weaknesses, and priority issues requiring preprocessing before machine learning.			

Completeness

Data completeness refers to the extent to which the information required for analysis is available and accessible. In environmental monitoring systems, missing values may result from sensor malfunctions, communication failures, power outages, scheduled equipment maintenance, or human factors associated with laboratory analyses.

Missing data reduce the amount of usable information, distort the statistical properties of the dataset, decrease prediction accuracy, and may limit the applicability of certain machine learning algorithms. Systematic missingness is particularly problematic, as the absence of data is associated with specific processes or monitoring conditions rather than occurring at random (Alejo-Sanchez et al., 2025).

In WWTPs, missing values commonly result from sensor interruptions, laboratory sampling schedules, or differences in sampling frequency (Ivanova et al., 2024).

Reliability

Data reliability reflects the extent to which measurements accurately represent the actual state of the monitored process. Even when a dataset is complete, individual measurements may be affected by noise, sensor errors, temporary equipment malfunctions, improper calibration, or human errors during data recording (Alferes & Vanrolleghem, 2016).

These imperfections may produce outliers or anomalous observations that do not reflect the actual behaviour of the monitored system (Kim et al., 2025). If not identified and

addressed in a timely manner, these observations may significantly affect model training, particularly for machine learning algorithms that are sensitive to data anomalies.

In WWTPs, unreliable measurements may directly affect process control and treatment evaluation.

Temporal Consistency

A distinguishing characteristic of environmental monitoring data is the variation in sampling frequencies across different variables. While some parameters are recorded continuously or at intervals of only a few minutes, others are measured daily, weekly, or only when laboratory analyses are performed.

These differences complicate the integration of data into a unified analytical dataset. Variations in temporal resolution may lead to incorrect alignment, information loss, and misleading relationships among variables.

In WWTPs, continuous sensor measurements must be synchronised with laboratory analyses and meteorological observations to ensure reliable modelling of environmental processes.

Informativeness

Modern environmental monitoring systems generate a large number of variables describing interconnected physical, chemical, biological, and operational processes. As a result, datasets often contain highly correlated variables, redundant information, or features with limited predictive value (Bagherzadeh et al., 2021).

Using all available variables does not necessarily improve machine learning performance. On the contrary, redundant or irrelevant features may increase model complexity, raise computational costs, and reduce the ability of models to generalise to previously unseen data.

In WWTPs, identifying informative variables improves both predictive accuracy and process interpretation.

The four functional dimensions presented in this section provide a practical framework for organising the most common data quality challenges encountered in environmental monitoring. Together, they establish a direct link between data quality assessment and preprocessing by identifying whether data preparation should focus on missing data handling, data cleaning, temporal harmonisation, or feature selection.

Data Quality-Driven Data Preparation

Data quality assessment does not directly improve environmental monitoring data; rather, it guides the selection of preprocessing strategies according to the characteristics of the available dataset. Consequently, data preparation should be regarded as a decision-driven rather than a fixed sequence of operations.

Environmental data preparation comprises preprocessing techniques that improve data suitability for machine learning. Figure 1 presents the proposed data quality-driven workflow, illustrating how each data quality dimension guides the selection of appropriate preprocessing strategies before model development. The following subsections review the principal methods for handling missing data, improving data reliability, ensuring temporal consistency, and enhancing data informativeness.

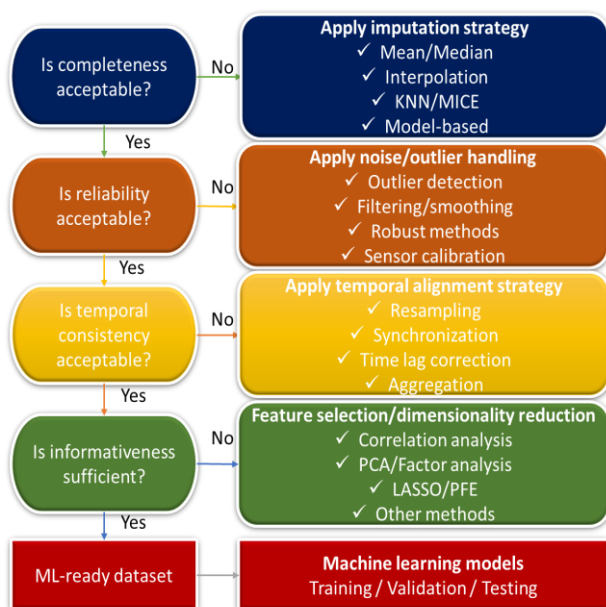


Fig. 1. Data quality-driven preprocessing workflow

Handling Missing Data

Missing values are among the most common challenges in environmental monitoring because they reduce the amount of usable information, distort statistical properties, and negatively affect model performance. They should therefore be addressed

before machine learning model development (Alejo-Sanchez et al., 2025).

Selecting an appropriate imputation technique requires considering the missingness mechanism (MCAR, MAR, or MNAR), the amount of missing data, and the characteristics of the monitored environmental process. Common approaches include interpolation techniques (linear, spline, and PCHIP) for continuously measured variables and model-based approaches such as k-nearest neighbours (kNN), Random Forest, Expectation-Maximisation (EM), and Multiple Imputation by Chained Equations (MICE) for multidimensional datasets.

In WWTP applications, missing observations commonly result from sensor malfunctions, laboratory sampling schedules, or differences in sampling frequency among data sources, making the choice of an imputation strategy a direct outcome of data quality assessment (Ivanova et al., 2024).

Improving Data Reliability

Improving data reliability involves reducing measurement errors and removing unreliable observations. The choice of appropriate techniques depends on the nature and origin of the identified deviations (Alferes & Vanrolleghem, 2016).

Random measurement noise is commonly addressed using smoothing and filtering techniques, such as moving average, median filtering, and the Savitzky–Golay filter, whereas abnormal observations are typically detected using statistical methods (e.g., Z-score and interquartile range) or machine learning algorithms, including Isolation Forest, Local Outlier Factor (LOF), DBSCAN, and One-Class Support Vector Machines (Kim et al., 2025).

In environmental monitoring, extreme values may represent genuine environmental events rather than measurement errors. In WWTPs, for example, sensor fouling, calibration drift, communication failures, and laboratory sampling errors are common sources of unreliable measurements, highlighting the importance of selecting appropriate filtering and outlier detection techniques.

Temporal Alignment and Synchronisation

Environmental monitoring systems typically integrate data from multiple sources that differ in sampling frequency, temporal resolution, monitoring period, and recording time. Without appropriate temporal synchronisation, these differences may lead to incorrect variable matching, misleading relationships, and reduced machine learning performance.

Common preprocessing techniques include aggregation, resampling, interpolation, synchronisation, and time lag correction. In more complex applications, cross-correlation analysis, lag analysis, and Dynamic Time Warping (DTW) can be used to identify temporal dependencies between variables. However, excessive aggregation or inappropriate interpolation may obscure the natural dynamics of the monitored processes and reduce the information available for machine learning.

WWTPs provide a representative example of these challenges, where continuous sensor measurements must be integrated with laboratory analyses, operational records, and meteorological observations collected at different temporal resolutions.

Feature Selection and Dimensionality Reduction

Environmental monitoring systems often generate large numbers of interrelated physical, chemical, biological, and operational variables. Many of these variables are redundant or

highly correlated, making dimensionality reduction an important step in preparing datasets for machine learning.

Feature selection methods are commonly categorised into filter, wrapper, and embedded approaches, whereas dimensionality reduction techniques include Principal Component Analysis (PCA), Factor Analysis (FA), and non-linear methods, such as t-SNE, UMAP, and autoencoders. Feature selection is generally preferred when preserving the physical interpretation of variables is important, whereas dimensionality reduction is more suitable for reducing data complexity or extracting compact representations.

WWTPs are typical high-dimensional systems in which operational, water quality, and meteorological variables are monitored simultaneously, making feature selection, dimensionality reduction, or their combination essential for efficient machine learning.

The preprocessing strategies reviewed in this section illustrate how different data quality challenges can be addressed through appropriate data preparation techniques. Rather than representing a fixed sequence of operations, these techniques should be selected according to the specific quality characteristics of the available dataset. These relationships are integrated into the data quality-driven conceptual framework presented in the following section.

A Conceptual Framework for Data Quality-Driven Machine Learning in Environmental Monitoring

The literature reviewed in this study indicates that data preprocessing is commonly treated as a sequence of technical operations performed prior to machine learning model development. In many cases, preprocessing techniques are selected based on practical considerations or the characteristics of a particular algorithm, without sufficient consideration of the actual condition of the available data. Such an approach may lead to suboptimal preprocessing decisions, ultimately reducing the reliability and interpretability of machine learning models.

Based on the findings of this review, a conceptual data quality-driven framework is proposed in which systematic data quality assessment guides the selection of preprocessing strategies according to the identified data quality issues rather than following a predefined preprocessing workflow.

As illustrated in Figure 2, the framework begins with the integration of heterogeneous environmental data, followed by the assessment of four key data quality dimensions: completeness, reliability, temporal consistency, and informativeness. This assessment determines the preprocessing strategies required before machine learning model development.

Rather than prescribing a fixed preprocessing pipeline, the proposed framework establishes explicit links between identified data quality issues and the corresponding preprocessing strategies. In this way, data preparation becomes an adaptive process in which preprocessing decisions are driven by the assessed condition of the available data, providing a transparent basis for reliable machine learning in environmental monitoring.

Once these stages have been completed, the dataset is better prepared for machine learning model development. By explicitly linking preprocessing decisions to data quality assessment, the proposed framework reduces the risk of inappropriate preprocessing, improves model reliability and

interpretability, and supports the development of more robust and generalisable machine learning models.

Although illustrated using wastewater treatment plants, the proposed framework is applicable to a broad range of environmental monitoring systems, including air, water, soil, climate, ecosystem, and smart sensing applications (Li et al., 2026).

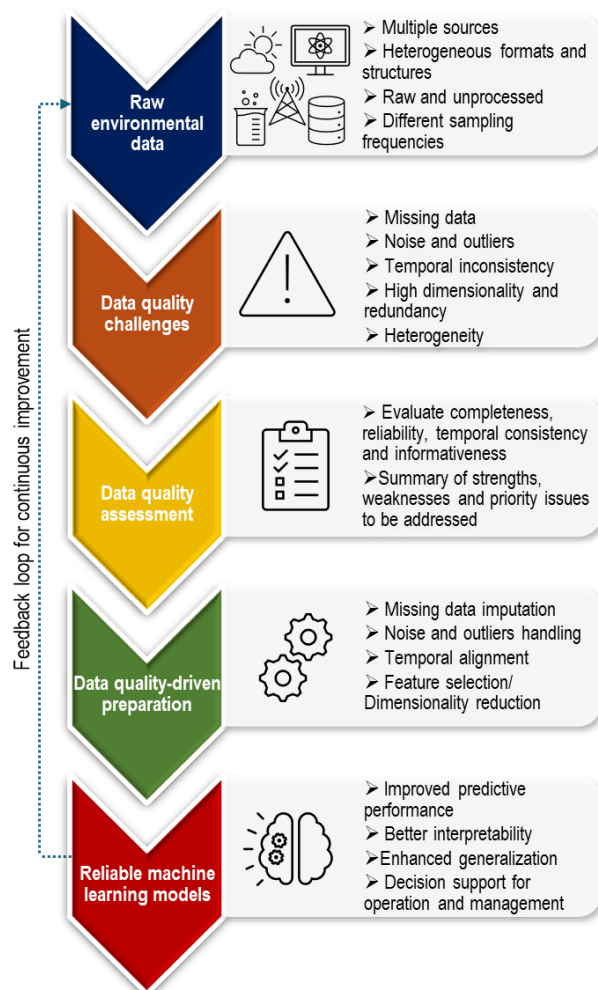


Fig. 2. Conceptual data quality-driven framework

Although the proposed framework is derived from a comprehensive review of the literature, it remains conceptual in nature and has not yet been validated through a unified experimental study. Future research should validate the proposed framework across different environmental monitoring systems and compare its performance with conventional preprocessing workflows.

Conclusion

The growing application of machine learning in environmental monitoring places increasing demands on the quality and preparation of the underlying data. Despite continuous advances in machine learning algorithms, the reliability, robustness, and interpretability of predictive models remain strongly dependent on data quality and the appropriateness of the applied preprocessing strategies.

This review examined the principal challenges associated with preparing environmental monitoring data for machine learning, focusing on the characteristics of environmental datasets, the key dimensions of data quality, and the most widely used preprocessing techniques. Wastewater treatment plants served as the representative application, illustrating challenges that are common across many environmental monitoring systems.

The principal contribution of this review is a conceptual data quality-driven framework linking data quality assessment with preprocessing strategy selection. Though developed for environmental monitoring, its principles are applicable to other domains characterised by heterogeneous, multidimensional, and dynamic datasets.

Future research should focus on validating the proposed framework using diverse real-world environmental datasets and developing intelligent decision-support systems for data quality assessment capable of automatically recommending appropriate preprocessing strategies according to dataset characteristics. Such developments could further improve the reliability of machine learning models and support more effective environmental monitoring and decision-making. Ultimately, reliable machine learning begins with systematic data quality assessment and appropriate data preparation.

References

- Alejo-Sanchez, L. E., Márquez-Grajales, A., Salas-Martínez, F., Franco-Arcega, A., López-Morales, V., Acevedo-Sandoval, O. A. & González-Ramírez, C. A. (2025). Missing data imputation of climate time series: A review. *MethodsX*, 15, 103455. <https://doi.org/10.1016/j.mex.2025.103455>
- Alferes, J. & Vanrolleghem, P. A. (2016). Efficient automated quality assessment: Dealing with faulty on-line water quality sensors. *AI Communications: The European Journal on Artificial Intelligence*, 29(6), 701-709. <https://doi.org/10.3233/AIC-160713>
- Aribisala, O. E. (2024). Data Quality Assessment and Preprocessing Techniques for Enhancing Machine Learning Model Performance. *International Journal of Innovative Research in Engineering & Multidisciplinary Physical Sciences*, 12. <https://doi.org/10.37082/IJIRMP.v12.i3.232799>
- Bagherzadeh, F., Mehrani, M.-J., Basirifard, M. & Roostaei, J. (2021). Comparative study on total nitrogen prediction in wastewater treatment plant and effect of various feature selection methods on machine learning algorithms performance. *Journal of Water Process Engineering*, 41, 102033. doi: 10.1016/j.jwpe.2021.102033
- Cangialosi, F., Bruno, E. & De Santis, G. (2021). Application of machine learning for fenceline monitoring of odour classes and concentrations at a wastewater treatment plant. *Sensors*, 21(14), 4716. <https://doi.org/10.3390/s21144716>
- Jaganathan, J. S., Ali, J. M. & Abdullah, S. R. S. (2025). Unlocking the Potential of Artificial Intelligence in Wastewater Treatment: Innovations, Opportunities, and Challenges. *Journal of Environmental Chemical Engineering*, 13(6). <https://doi.org/10.1016/j.jece.2025.119671>
- Ivanova, N., Yanev, N. & Vasileva, S. (2024). Utilisation of Basic Methods for Filling Missing Data from Wastewater Treatment Plants. – 2024 5th International Conference on Communications, Information, Electronic and Energy Systems (CIEES), 1-6. doi: 10.1109/CIEES62939.2024.10811430
- Kim, C.-S., Kim, C.-R. & Kok, K.-H. (2025). Outlier Detection in Hydrological Data Using Machine Learning: A Case Study in Lao PDR. *Water*, 17(21), 3120. <https://doi.org/10.3390/w17213120>
- Kumar, S. & Deswal, S. (2020). Estimation of Phosphorus Reduction from Wastewater by Artificial Neural Network, Random Forest and M5P Model Tree Approaches. *Pollution*, 6(2), 417-428. <https://doi.org/10.22059/POLL.2020.293086.717>
- Li, Y., Han, F. & Zheng, Y. (2026). Artificial Intelligence in Surface Water Quality Research and Management: Recent Progress and Future Directions. *Ecosystem Health and Sustainability*, 12, 0474. <https://doi.org/10.34133/ehs.0474>
- Newhart, K. B., Holloway, R. W., Hering, A. S. & Cath, T. Y. (2019). Data-Driven Performance Analyses of Wastewater Treatment Plants: A Review. *Water Research*, 157, 498-513. <https://doi.org/10.1016/j.watres.2019.03.030>
- Rashid, A., Ramya, P., Kumar, K. S., Manohar, Ch. S., Pratap, Ch. S. & Suresh, K. (2025). Optimisation of Sewage Treatment Plant Using Machine Learning Algorithms: A Comprehensive Literature Review. *International Journal of Research Publication and Reviews*, 6(10), 5292-5293. <https://doi.org/10.55248/gengpi.06.1025.3765>
- Sundui, B., Ramirez Calderon, O. A., Abdeldayem, O. M., Lázaro-Gil, J., Rene, E. R. & Sambuu, U. (2021). Applications of machine learning algorithms for biological wastewater treatment: Updates and perspectives. *Clean Technologies and Environmental Policy*, 23, 127-143. <https://doi.org/10.1007/s10098-020-01993-x>
- Ullo, S. L. & Sinha, G. R. (2020). Advances in Smart Environment Monitoring Systems Using IoT and Sensors. *Sensors*, 20(11), 3113. <https://doi.org/10.3390/s20113113>
- Wang, D., Thunell, S., Lindberg, U., Jiang, L., Trygg, J., Tysklind, M. & Souihi, N. (2021). A machine learning framework to improve effluent quality control in wastewater treatment plants. *Science of The Total Environment*, 784, 147138. <https://doi.org/10.1016/j.scitotenv.2021.147138>
- Zamfir, F.-S., Madalina, C. & Mihalanche, S. (2025). Application of Machine Learning Models in Optimising Wastewater Treatment Processes: A Review. *Applied Sciences*, 15(15), 8360. <https://doi.org/10.3390/app15158360>
- Zhong, S., Zhang, K., Bagheri, M., Burken, J. G., Gu, A., Li, B., Ma, X., Marrone, B. L., Ren, Z. J., Schrier, J., Shi, W., Tan, H., Wang, T., Wang, X., Wong, B. M., Xiao, X., Yu, X., Zhu, J.-J. & Zhang, H. (2021). Machine Learning: New Ideas and Tools in Environmental Science and Engineering. *Environmental Science & Technology*, 55(19), 12741-12754. <https://doi.org/10.1021/acs.est.1c01339>
- Zhu, M., Wang, J., Yang, X., Zhang, Y., Zhang, L., Ren, H., Wu, B. & Ye, L. (2022). A review of the application of machine learning in water quality evaluation. *Eco Environ Health*, 1(2), 107-116. <https://doi.org/10.1016/j.eehl.2022.06.001>