



Delft University of Technology

Document Version

Final published version

Licence

CC BY

Citation (APA)

Oberascher, M., Brentan, B. M., Menapace, A., Herrera, M., Fu, G., Taormina, R., & Sitzenfrei, R. (2026). Assessing supervised machine learning practice in urban water networks: A critical review of methodological transparency, reproducibility and reporting. *Water Research X*, 32, Article 100569. <https://doi.org/10.1016/j.wroa.2026.100569>

Important note

To cite this publication, please use the final published version (if applicable).
Please check the document version above.

Copyright

In case the licence states "Dutch Copyright Act (Article 25fa)", this publication was made available Green Open Access via the TU Delft Institutional Repository pursuant to Dutch Copyright Act (Article 25fa, the Taverne amendment). This provision does not affect copyright ownership.
Unless copyright is transferred by contract or statute, it remains with the copyright holder.

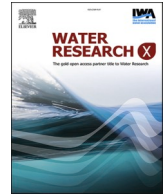
Sharing and reuse

Other than for strictly personal use, it is not permitted to download, forward or distribute the text or part of it, without the consent of the author(s) and/or copyright holder(s), unless the work is under an open content license such as Creative Commons.

Takedown policy

Please contact us and provide details if you believe this document breaches copyrights.
We will remove access to the work immediately and investigate your claim.

This work is downloaded from Delft University of Technology.



Assessing supervised machine learning practice in urban water networks: A critical review of methodological transparency, reproducibility and reporting

Martin Oberascher^a, Bruno M. Brentan^b, Andrea Menapace^c, Manuel Herrera^d, Guangtao Fu^e, Riccardo Taormina^f, Robert Sitzenfrei^{a,*}

^a Unit of Environmental Engineering, Department of Infrastructure Engineering, University of Innsbruck, Technikerstraße 13, 6020 Innsbruck, Austria

^b Hydraulics and Water Resources Department, Universidade Federal de Minas Gerais, Av. Presidente Antonio Carlos, 6947, Belo Horizonte, 31555250, Minas Gerais, Brazil

^c Institute for Renewable Energy, Eurac Research, Viale Druso 1, Bolzano, 39100, Italy

^d School of Engineering, Newcastle University, Newcastle upon Tyne, NE1 7RU, United Kingdom

^e Centre for Water Systems, University of Exeter, Exeter EX4 4QF, United Kingdom

^f Delft University of Technology, Faculty of Civil Engineering and Geosciences, 2600 GA Delft, Delft, P.O. Box 5048, the Netherlands

ARTICLE INFO

Keywords:

Artificial Intelligence
Best practice
Methodology checklist
Urban drainage network
Supervised learning
Water distribution network

ABSTRACT

In recent years, the advantages of machine learning (ML) have been clearly demonstrated in research on urban water infrastructure (UWI) and has been applied in a wide range of applications. This review critically assesses the current quality of ML implementations in UWI by examining >100 articles from the recent literature, with a particular focus on common pitfalls throughout the development process. Most reviewed articles placed strong emphasis on performance benchmarking but provided limited reporting on key ML methodology implementation and deployment. Only around one third of the reviewed articles documented essential tasks such as feature scaling or automatic hyperparameter optimisation, despite their importance for performance and generalisation. Additionally, fewer than 25% reported explainability and uncertainty quantification techniques, making reported performance gains difficult to explain or operationalise. Furthermore, the lack of standardised documentation makes the extraction of relevant information about the methods, workflows, and steps needed for reproducibility difficult. Together, these issues negatively affect the reproducibility and reduce comparison and trust in the developed approaches. This is particularly important, as trust and confidence in ML-based decisions are key requirements for successful transformation of research into practice. To address these issues, a methodological and reporting checklist is provided to guide the design and integration of ML applications in UWI throughout the development process and to highlight common pitfalls. The review and the checklist can support developers, technicians, and operators in future ML applications, helping to raise awareness on reproducible ML implementations in UWI applications.

1. Introduction

Urban water infrastructure (UWI), including urban drainage networks (UDN) and water distribution networks (WDN), is an important part of the national critical infrastructure. It ensures an environmentally sound disposal of wastewater, a safe stormwater discharge, and a reliable supply of drinking water in adequate quantity and quality. UWI is exposed to constant changes, such as population growth, demographic shifts, and climate change (He et al., 2021; Yazdanfar and Sharma,

2015). In recent years, digitalisation is increasingly seen by operators as a dynamic measure to address these challenges (Daniel et al., 2023). In this context, the collection and analysis of operating data as well as the remote control of a wide variety of system components is frequently applied (Creaco et al., 2019; Yuan et al., 2019), with the Internet of Things further enhancing the availability of high-resolution measurement data (Oberascher et al., 2022).

Machine learning (ML), as a subset of artificial intelligence (AI), has emerged as a promising tool for managing large volumes of data,

* Corresponding author.

E-mail address: robert.sitzenfrei@uibk.ac.at (R. Sitzenfrei).

<https://doi.org/10.1016/j.wroa.2026.100569>

Received 29 April 2026; Received in revised form 10 June 2026; Accepted 10 June 2026

Available online 11 June 2026

2589-9147/© 2026 The Authors. Published by Elsevier Ltd. This is an open access article under the CC BY license (<http://creativecommons.org/licenses/by/4.0/>).

supporting decision-making, and enhancing performance in UWI (Doorn, 2021; Rapp et al., 2023). ML, which involves learning complex relationships between variables, has been applied across a wide range of applications, and its advantages for UWI are well documented in research (Rapp et al., 2023). It is referred to recent reviews for more information about the widespread possibilities in the management of UWI such as Ahmed et al. (2024), Fu et al. (2022) Li et al. (2024), and Kanyama et al. (2024) for example.

Despite its promises, similar errors and misunderstandings are often observed in ML implementations as shown in general reviews (Liao et al., 2021; Lones, 2024), as well as in domain-specific analysis in other fields (Roberts et al., 2021; Zhu et al., 2023), reducing overall trustworthiness of the achievable results. Since these ML applications are often developed and applied by domain experts, these reviews have also introduced field-specific best practice frameworks to counteract this trend. To the best of the authors' knowledge, such a discussion tailored to the UWI is still at an early stage. For example, Jun et al. (2025) identified that comparatively limited attention is paid to pre-model setup procedures, such as data processing and model optimisation in UWI, and proposed a checklist to improve reporting on data integrity to avoid common pitfalls. However, this analysis is restricted to articles published in the *Journal of Water Resources and Management* and focused on implementation quality up to the stage of hyperparameter optimisation. In contrast, this review extends this analysis by adopting a cross-journal perspective and providing a critical evaluation of the entire ML development process, ranging from data pre-processing to model application. Furthermore, based on the identified shortcomings, a corrective checklist is introduced to guide ML implementation in the field of UWI. This checklist is broadly applicable across a wide range of ML application, which may differ substantially in terms of data characteristics, model architectures and workflows. In contrast to existing reviews which focus primarily on applications and performance gains, this review adopts an explicitly critical perspective on the implementation. The focus is not which ML methods are used in UWI, but whether their development ensures reliable generalisation, transparency, and reproducibility.

The aim of this review is to:

- Critically assess the documented methodological correctness and implementation quality of supervised ML applications in UWI, identifying systematic weaknesses that undermine reported performance and operational trust.
- Providing a methodological and reporting checklist to guide the implementation of supervised ML in UWI throughout the development process and to highlight common pitfalls

2. General workflow of supervised machine learning

In general, ML can be divided into unsupervised, supervised, and reinforcement learning (Huang et al., 2021). Supervised ML is based on the usage of labelled data, allowing the numerical calculation of a loss function for model training and evaluation (Zhu et al., 2023). With focus on UWI, supervised ML is applied across a wide range of applications in the management of UWI, typically used for real-time operations such as predicting system states, anomaly detection or surrogate model. It is also employed for operation optimisation, including asset management for rehabilitation measures (Fu et al., 2022; Huang et al., 2021; Kanyama et al., 2024). This work focuses on supervised ML, as the initial results of the literature review have shown that this is the most common approach.

To ensure consistent terminology throughout this work and to support a better understanding how the reviewed articles were collected and evaluated, this section provides a short overview of a general workflow for implementing supervised ML, based on literature (Chauhan et al., 2020; Géron, 2022; Salehin et al., 2024; Zhu et al., 2023). The supervised ML workflow is divided into four phases as

outlined in Fig. 1:

- In the first phase '**project setup**', the purpose of the ML application and the corresponding supervised ML task are defined, subdivided into classification and regression. The required data are then collected from relevant sources and may include different data types.
- In the second phase '**data preprocessing**', raw data collected in the first phase are processed and transformed into input features for ML models. This includes splitting the data into training, validation, and test sets for use in the 'model development' phase. As UWI data is normally noisy and contains data gaps, outlier detection and missing data management are required. Feature selection or feature engineering may then be performed, often in interaction with model development, to identify informative and relevant features; however, this step is not always necessary, as modern deep neural networks often can automatically learn hierarchical feature representations from raw data. The last step is feature scaling to ensure similar value ranges.
- The prepared data sets are used in the third phase called '**model development**' to train different ML models and to select the best one. This includes optimising ML model configurations, called hyperparameters, and evaluating model performance using appropriate validation strategies and multiple performance metrics. Different ML models should be compared to identify the best ML model.
- In the last phase, the '**model application**', the selected ML model is prepared for deployment in operational and decision-support contexts. This includes integrating explainability techniques for an understanding of its decision-making and applying uncertainty quantification techniques to address data and model uncertainties. Finally, the full ML development process, including hyperparameters, should be documented and reported, ideally with shared data and code to support reproducibility and transparency.

3. Review methodology

This review focuses on supervised machine learning in UWI, prioritising highly cited articles to evaluate dominant methodological practices. Consequently, recent or less-cited high-quality articles may be underrepresented. The aim of this review is not to provide an up-to-date overview of ML applications in UWI, but rather to assess their documented implementation quality and their implications for trustworthiness and practical deployment.

3.1. Literature search and selection

The literature search and selection process followed the "Preferred Reporting Items for Systematic Reviews and Meta-Analyses (PRISMA) 2020 guidelines" (Page et al., 2021), as shown in Fig. 2. A structured search in "Web of Science" was carried out in June 2025 and updated in April 2026 using the keywords ("*machine learning*" OR "*artificial intelligence*" OR additional ML keywords) AND (type network keywords) within the topic field (title, abstract, keywords).

Additional ML keywords were drawn from widely used ML methods identified in previous UWI-focused reviews, including "*deep learning*", "*neural network*", "*random forest*", "*support vector machine*", "*decision tree*", "*XGBoost*", "*classification and regression trees*", and "*long short-term memory*" (Boo et al., 2024; Li et al., 2025; Niknam et al., 2022; Zhu et al., 2023, 2022). The type of network keywords were ("*urban drainage*" OR "*sewer systems*" OR "*urban wastewater*" OR "*urban stormwater*") for UDN and ("*water supply*" OR "*water distribution*") for WDN, resulting in 388 and 2156 records, respectively. These records were pre-screened by their title, abstract, and keywords, which narrowed down supervised ML applications for UDN and WDN to 295 and 1012 articles, respectively.

The review period spanned 2017 to 2025, reflecting the start of the approximate exponential growth in supervised ML publications and

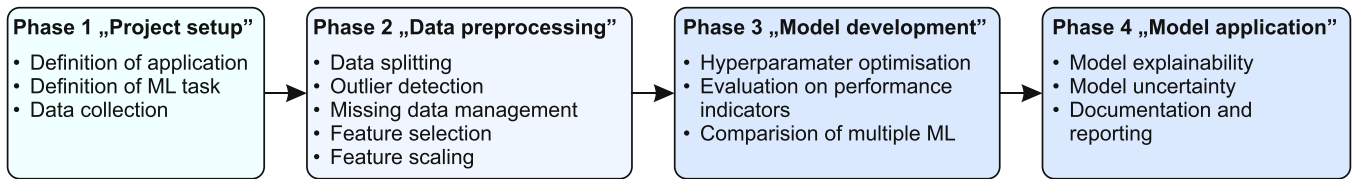


Fig. 1. General workflow supervised machine learning.

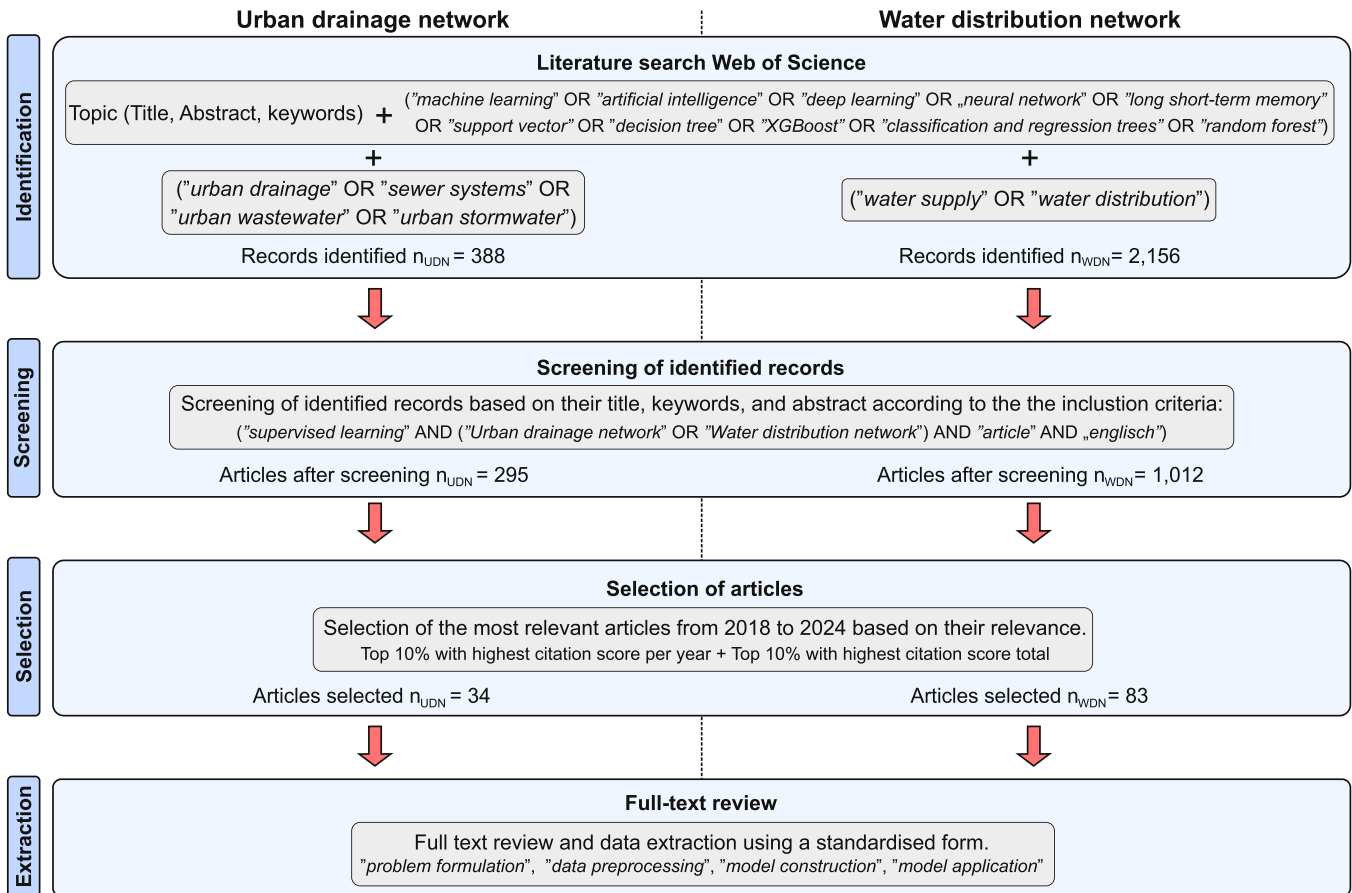


Fig. 2. PRISMA flow diagram.

covering nine years in total. To emphasise influential work, the analysis prioritised highly cited articles, assuming these have greater outreach within the community and are more likely used as inspiration for ML implementations. Articles were therefore selected if they were within the top 10% most cited articles of their publication year (after pre-screening) or within the top 10% of the most cited articles from the entire period between 2017 and 2025. These values were selected to ensure enough articles for the review while keeping the review manageable in scope.

The final selection included 34 UD articles and 83 WDN articles, giving a total of 117 articles in UWI reviewed. While this approach favours influential publications, it does not necessarily reflect methodological quality and high-quality but less-cited or recently published work may therefore be underrepresented. Consequently, the results represent a critical assessment of the overall reported quality of methodological implementation in ML research in UWI, rather than an evaluation of individual implementation errors.

3.2. Data extraction and thematic analysis

In addition to the literature selection, an analytical framework was

developed to assess and evaluate the reported quality of supervised ML applications in UWI. To this end, the following information was extracted from each selected article and structured into the following phases:

‘Project setup’ covers general information about the ML application, including network type (e.g., UDN or WDN), the application purpose, the ML task, and details on the type and source of the data used. **‘Data preprocessing’** records how the data were prepared as input features for the ML model. This includes sample size, data splitting approach, outlier and missing data management, feature selection, and feature scaling. For the later three tasks, it was recorded whether they were reported in the article, and, when applied, which specific method were used. Where feature selection was mentioned, the applied methods were classified into filter, wrapper, and embedded methods. For feature scaling, only scale-sensitive models, such as neural networks or distance-based methods, were considered relevant, whereas tree-based models were not evaluated on this aspect as their performance is largely insensitive to feature scaling. **‘Model development’** describes the workflow used to develop and evaluate the ML model. It includes hyperparameter optimisation, performance evaluation using multiple metrics, comparison with different ML models, and the final ML model

selection. Where hyperparameter optimisation was reported, the applied methods were classified into trial-and-error tuning, grid search, random search, evolutionary algorithms, and Bayesian optimisation. In addition, the types of evaluation metrics and ML models used were recorded, together with the final selected ML model. **‘Model application’** focuses on deploying the trained ML model and documenting it for reproducibility. This includes the usage of model explainability and uncertainty quantification techniques, as well as the availability of code and data. It was recorded of these tasks were reported in the reviewed articles. Model explainability and uncertainty were considered as relevant reporting aspects for all ML applications, although it may be more relevant for critical applications affecting the supply reliability. When documented, the applied methods were grouped into white-box methods, feature relevance explanations, and visual explanations for model explainability, and into ensemble, deterministic, conformal prediction, or Bayesian neural networks approaches for model uncertainty. Code and data availability were assessed based on the information provided in the corresponding data availability statements, as well as in supplementary materials or any external links provided by authors. Where external links were included, their accessibility was checked. In cases where the use of benchmark data was reported, preliminary verifications were carried out to confirm the existence of the referred benchmark data.

Additionally, the articles were screened for potential data leakages, referring to the unintended use of information that would not be available during training, which can artificially increase achievable performance. During the review, data leakage was recorded only when there was clear evidence of data leakage in the reported workflow. Examples include detailed workflow diagrams showing feature scaling before data splitting, or explicit statements in the text indicating that preprocessing was performed on the entire data sets. In contrast, simplified workflow diagrams, unclear methodological descriptions, or unreported tasks were not classified as data leakage, although they may indicate potential data leakage. Data leakages were categorised into three types, used as qualitative criteria to evaluate the reported workflow: (1) data-induced data leakage, where a input feature is directly or

indirectly related or contain information to the target variable, (2) preprocessing-related data leakage, where preprocessing tasks (e.g., scaling, imputation, or feature selection) are applied to the entire dataset before data splitting, and (3) split-related data leakages, where data leakage induce overlapping time-related information between different data sets.

A detailed description of all extracted information is provided in the supplementary materials, together with a full list of the 117 articles included in the evaluation.

4. Common pitfalls and best practices

The evaluation of the reported implementation quality of ML applications in UWI is presented anonymously, based on 34 UDN articles and 83 WDN articles. The individual analyses are enhanced with examples of good implementation practice, supplemented through an additional manual literature search.

4.1. Project setup

Fig. 3 gives an overview of the ML applications. In UDN, ML applications mainly address flood prediction, state prediction, and classifying sewer conditions, while the focus in WDN lies on leakage detection, demand prediction, and state prediction. The ML applications are further subdivided into the ML task. Classification refers to the prediction of discrete values such as classes, whereas with regression continuous values are predicted (Huang et al., 2021). 10 articles are formulated as classification problem and 24 articles fall into regression in UDN, while for WDN, 44 articles are classification and 39 articles are regression. This division influences subsequent model development. Regarding data sources, Fig. 3 shows that the majority of the articles uses real-world data, whereas simulated and test-bed data is usually applied for applications dealing with rare events negatively impacting operations, such as leakages, cyber-physical attacks or floods.

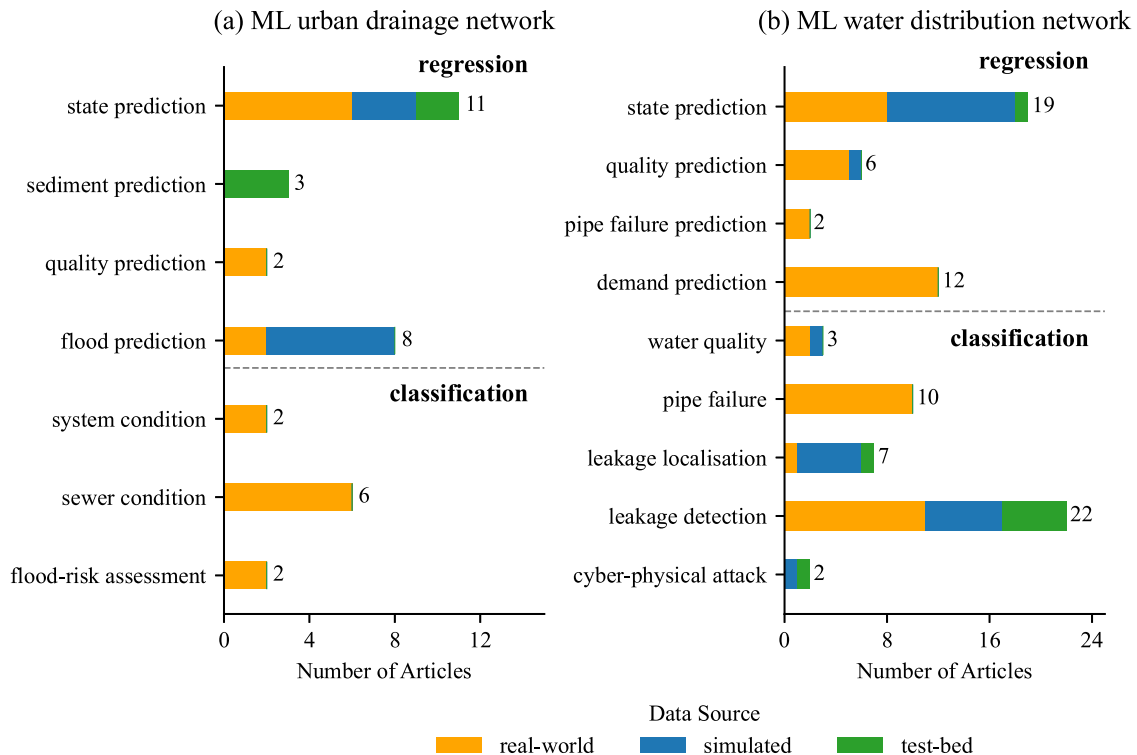


Fig. 3. Overview of the ML applications subdivided into ML task and data source for (a) UDN and (b) WDN.

4.2. Data preprocessing

The performance with supervised ML models depends heavily on the quality and quantity of available data. Consequently, processing tasks such as data splitting, outlier detection, missing data management, and feature selection are important parts of the development process. However, the evaluation of these tasks, summarised in Fig. 4, reveals a systematic neglect in documentation of these essential data processing steps, despite their recognised importance for generalisability and reliability.

In supervised ML, it is considered good practice to divide the dataset into three subsets, used for training, validation, and test (TVT), respectively. The training set is used to learn the relationship between input features and the target; the validation set is employed to prevent underfitting (e.g., inadequate learning of underlying patterns) or overfitting (e.g., excessive learning of data-specific patterns) and to optimise hyperparameters and model architecture; and the independent test set ensures an unbiased performance evaluation (Eelbode et al., 2021). When data availability is limited, k-fold cross-validation (CV) can be used for model selection. In this approach, the training data are split into k folds, and k models are trained, each using $k-1$ folds for training and the remaining fold for validation. Validation results are then averaged to optimise hyperparameters and model architecture. An independent test set, if available, should be used only for final performance assessment (Vabalas et al., 2019; Zhu et al., 2023). For computationally intensive

models such as deep learning architectures, k-fold cross-validation can be challenging in practice and is therefore often replaced by simpler hold-out or reduced-fold strategies. Across the reviewed articles, TVT and k-fold CV with independent test set (CV+T) were applied in 11 and 10 UDN articles and in 28 and 14 WDN articles, respectively. Surprisingly, the second most common approach after TVT was a training and test (TT) split, used in 8 UDN articles and 24 WDN articles. This increases the risk of under- or overfitting of the ML model due to the absence of a validation set. Other splitting approaches, such as training and validation (TV) or k-fold CV without a test set, also limit the ability to obtain an independent performance evaluation. It should be noted, however, that some of these articles used ML model outputs as reference representations of expected system behaviour, against which real measurements were compared, for example in anomaly detection tasks such as leakage identification. In such cases, the primary objective is the detection of deviations rather than unbiased estimation of predictive generalisation performance, making strict independence of test data less critical.

Fig. 4(b) shows the ratio of training, validation, and test data using a ternary plot, where the values of each axis are read according to the angle of corresponding ticks. For example, TT has no validation data, so these points lie on the 0.0 line of the validation axis, which is the axis line for the training data. The ratio used to split data into the different sets is flexible and typically depends on the number of available samples; however, the test set should capture the main characteristics of the data

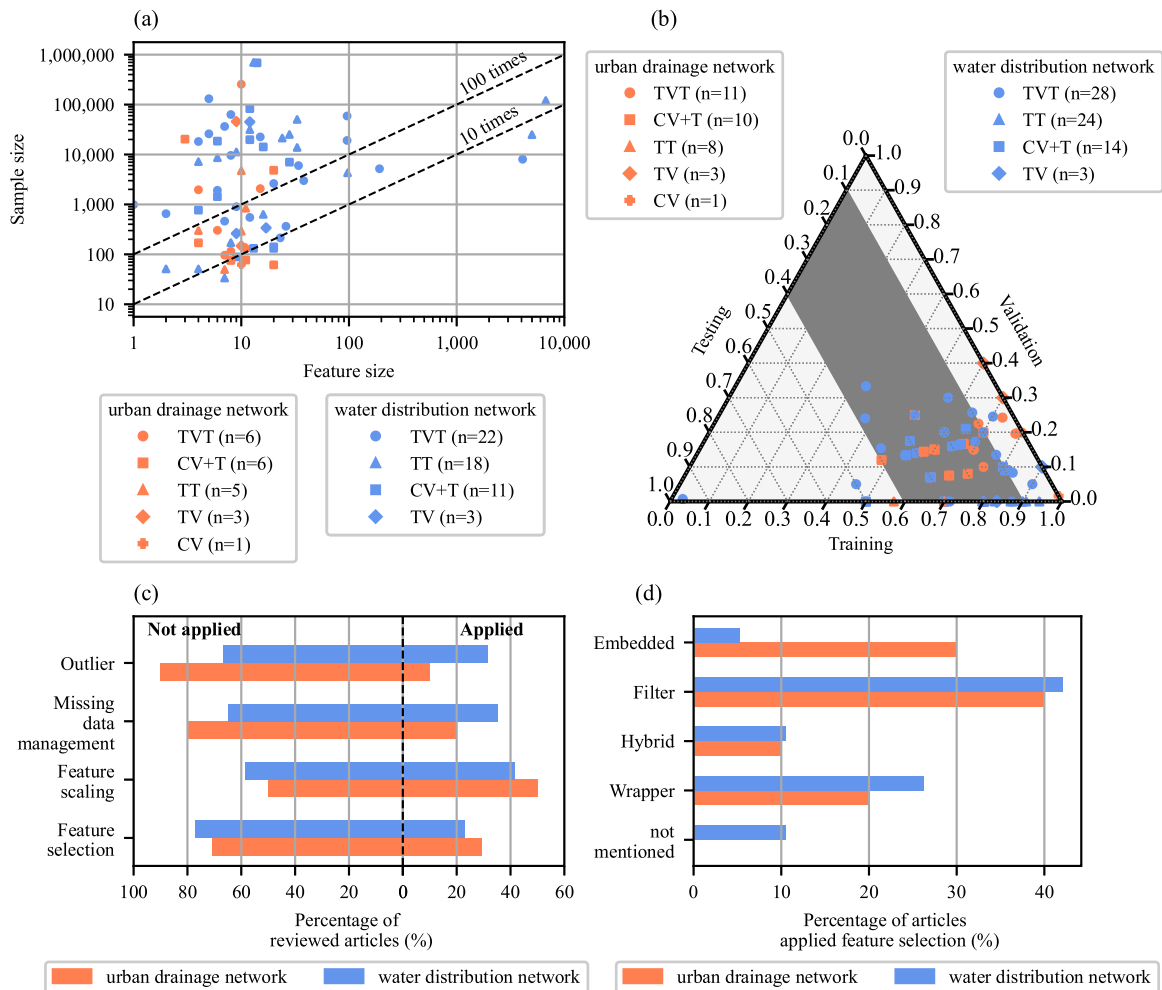


Fig. 4. Evaluation of the implementation quality for data preprocessing subdivided into UDN and WDN: (a) comparison of number of features and sample size for training used in model development, (b) ternary plot for the data splitting, (c) analyse of the applied steps for data preprocessing, and (d) overview of utilised methods if feature selection is applied.

to ensure a fair model evaluation (Eelbode et al., 2021). As a guideline, the literature recommends a value between 0.1 and 0.4 of the total dataset size (Jun et al., 2025; Zhu et al., 2023), whereas for computationally intensive models such as deep learning architectures, this range should be interpreted as a heuristic rather than a strict rule, as test-set design is often constrained by data availability and computational cost. As indicated by the dark shaded area, most articles using CVT or TVT data splitting meet this requirement, being 71% for UDN and 84% for WDN. In UDN articles, 29% applied a lower test set ratio. In WDN articles, 8% used a lower ratio and 8% a higher one. These differences arise partly from CV and TV splitting, with the remainder mainly resulting from TVT data splitting.

UDN and WDN datasets typically include outliers and missing values, which must be addressed before model development. Fig. 4(c) shows the proportion of reviewed articles reporting outlier and missing data management, relative to the number of articles using real-world data (51 WDN, 20 UDN). Outlier handling was documented in only 31% of WDN and 10% of UDN articles in WDN and UDN using real-world data. All applied statistical methods such as box plots or z-scores to remove outliers, as for instance applied by Chen et al. (2022). Although statistical methods are easy to implement, their effectiveness depends on the dataset properties and is limited with high dimensional data (Smiti, 2020). Moreover, outliers should not always be removed, because such data points may represent meaningful system behaviour such as peak demands or regular flushing events (Alvisi et al., 2025).

The management of missing data was more frequently reported, 35% of WDN articles and 20% of UDN articles addressed missing values. Two main approaches were used, namely deletion and imputation. Deletion, applied for example by Fan et al. (2022) or Snider and McBean (2020), removes all samples containing missing values, but reduces sample size (Emmanuel et al., 2021) and disrupts temporal continuity in timeseries data. Imputation includes simple imputation (e.g., filling the missing values with mean or median values), interpolation (e.g., filling the missing values as linear function between known neighbour values), Kalman filtering, and ML-based methods such as k-nearest neighbours. The appropriate method depends on the number and length of data gaps. For example, Namdari et al. (2023) filled data gaps using mean values from the same hour on adjacent days, while Zanfei et al. (2022) demonstrated that more complex methods outperform simple imputation or interpolation for a higher proportion of missing data. Combined strategies can also be useful. Daulat et al. (2024), for instance, removed samples missing important features (e.g., pipe age), but used median imputation for less influential features (e.g., diameter) in developing deterioration models.

Feature selection aims to identify the most informative and relevant features while removing redundant information, improving both technical and computational efficiency (Dhal and Azad, 2021; Pudjihartono et al., 2022). Despite its documented benefits, Fig. 4(c) shows that feature selection was reported in only 23% of WDN and 29% of UDN articles. According to Fig. 4(d), the most common approach was filter-based methods (around 40% of the articles), followed by wrapper-based and embedded methods. Each method has different advantages and limitations, and the choice strongly depends on the specific objective (Dhal and Azad, 2021).

In filter-based methods, the dependency between the input features and the target variable is assessed using static measures such as correlation coefficients, and relevant features are selected by applying a predefined threshold (Beig Zali et al., 2024; Zhou et al., 2022). As filter-based methods identify features before model training, they are independent of the ML model itself (Dhal and Azad, 2021); however they cannot consider feature interdependencies or interaction with the ML model (Pudjihartono et al., 2022). In contrast, wrapper-based methods evaluate different combinations of features, iteratively or using grid search, and select the best feature combination based on the performance achieved by the ML model (Ebtehaj et al., 2020; Guo et al., 2018). These methods therefore capture both feature dependence and

model interactions, but at the cost of higher computational effort and dependence on the chosen ML model (Pudjihartono et al., 2022). Embedded methods incorporate feature selection directly within the model training process as is the case for Random Forest, where features are internally selected (Montes et al., 2021; Safari, 2019). They are faster than wrapper-based methods and also account for feature dependencies and model interactions, but their use is limited to tree-based ML models (Pudjihartono et al., 2022). Although a key characteristic of deep learning models is their capacity to automatically learn feature representations from raw inputs, especially in image-related tasks, recent studies suggest that explicit feature selection may be beneficial also for deep learning (Liang et al., 2025). Around 10% of the articles employed hybrid approaches, combining two feature selection methods (Tijani et al., 2022).

Feature scaling is equally important, as UWI data commonly include features with differing ranges. For scale-sensitive models, such as neural networks or distance-based methods, transforming features into comparable value ranges is considered good practice to improve learning performance (Géron, 2022; Zhu et al., 2023). In contrast, tree-based models are largely insensitive to feature scaling, which may partly explain its limited use in some studies. Only 42% of WDN and 50% of UDN articles applying scale-sensitive models documented feature scaling as presented in Fig. 4(c). Normalisation (min-max scaling) and standardisation (z-score normalisation) were the used approaches. For example, Taiwo et al. (2024) reported a performance improvement of 5% in pipe failure predictions after standardising input features with widely varying ranges (e.g., pipe length between 1 and 200 m, diameter between 20 and 3000 mm, and pressure between 0.05 – 21.07 m). The choice of scaling method depends partly on data quality, as normalisation is sensitive to outliers, whereas standardisation is less affected (Géron, 2022).

Finally, determining the appropriate amount of data is required for training remains a key question. A commonly used heuristic indicator is the ratio of training sample size to number of input features, with values greater than 10 for simpler tasks and greater than 100 for more complex tasks (Zhu et al., 2023). This ratio should be interpreted as a coarse guideline, as data requirements also depend on model class, feature redundancy, and temporal or spatial dependence in the data. Fig. 4(a) shows the scatter plot between number of features and training sample size from the 75 articles providing sufficient information. The number of input features ranged from one to roughly 100, whereas training sample size varied between 50 and 100,000, with a maximum of 697,000 samples. Seven articles fell below the minimum recommended ratio, while most (41 articles) exceeded the ratio of 100. Notably, 81% of the 27 articles with ratios between 10 and 100 used a TV, TVT, or TT split, whereas 59% of articles using CVT, which is often adopted to improve robustness under limited or uncertain data conditions, had ratios greater than 100. In practice, data availability in UWI applications is frequently constrained by external factors, such as sensor availability, spatial deployment, measurement frequency, and data quality, rather than solely by modelling choices.

A common pitfall during data preprocessing is data leakage, defined as the unintended use of information during model development that would not normally be available and thus artificially improving the achievable model performance (Apicella et al., 2025; Kaufman et al., 2012). Because data leakages are introduced unintentionally, identifying them in the analysed articles was challenging without access to code or detailed methodological descriptions that follow a standard definition. Consequently, only confirmed cases of data leakage were included in the analyse and were identified in approximately 15% of the reviewed articles based on detailed workflow diagrams or explicit workflow descriptions. Of these, 13 articles included preprocessing-related data leakages, while 7 articles showed split-related data leakages in articles with the application of leakage detection and water demand forecasting. This relatively widespread issue indicates a structural deficiency in the evaluation and validation of

ML models within the field, rather than isolated methodological errors.

Following [Apicella et al. \(2025\)](#), data leakages can be grouped into three categories. Data-induced leakage arises during data collection and include label leakages and synthetic data leakages. Label leakages occur when a feature used for model development is directly or indirectly related to the target variable ([Apicella et al., 2025](#)), such as using temperature measured at time step t to predict water demand at the same time step as illustrated in [Fig. 5\(a\)](#). Synthetic data leakages arise when data augmentation is applied before data splitting, allowing original and synthetic samples to appear in both training and test sets. To avoid this, data augmentation should be restricted to the training set ([Kapoor and Narayanan, 2023](#)) and to use appropriate evaluation metrics for

imbalanced datasets ([Zhu et al., 2023](#)).

Preprocessing-related data leakage occurs when preprocessing tasks are applied to the full dataset prior to splitting, allowing information from the validation or test sets to influence the training process ([Apicella et al., 2025](#)). To avoid these data leakage, all preprocessing tasks should be developed/fitted exclusively on the training set and subsequently applied to the validation and test sets as outlined in [Fig. 5\(b-c\)](#) ([Kapoor and Narayanan, 2023](#); [Zhu et al., 2023](#)). During the review, preprocessing-related data leakages

Split-related data leakage arises from inappropriate dataset splitting methods and includes both similarity (overlapping information) and distributional (violating temporal dependency) issues ([Apicella et al.,](#)

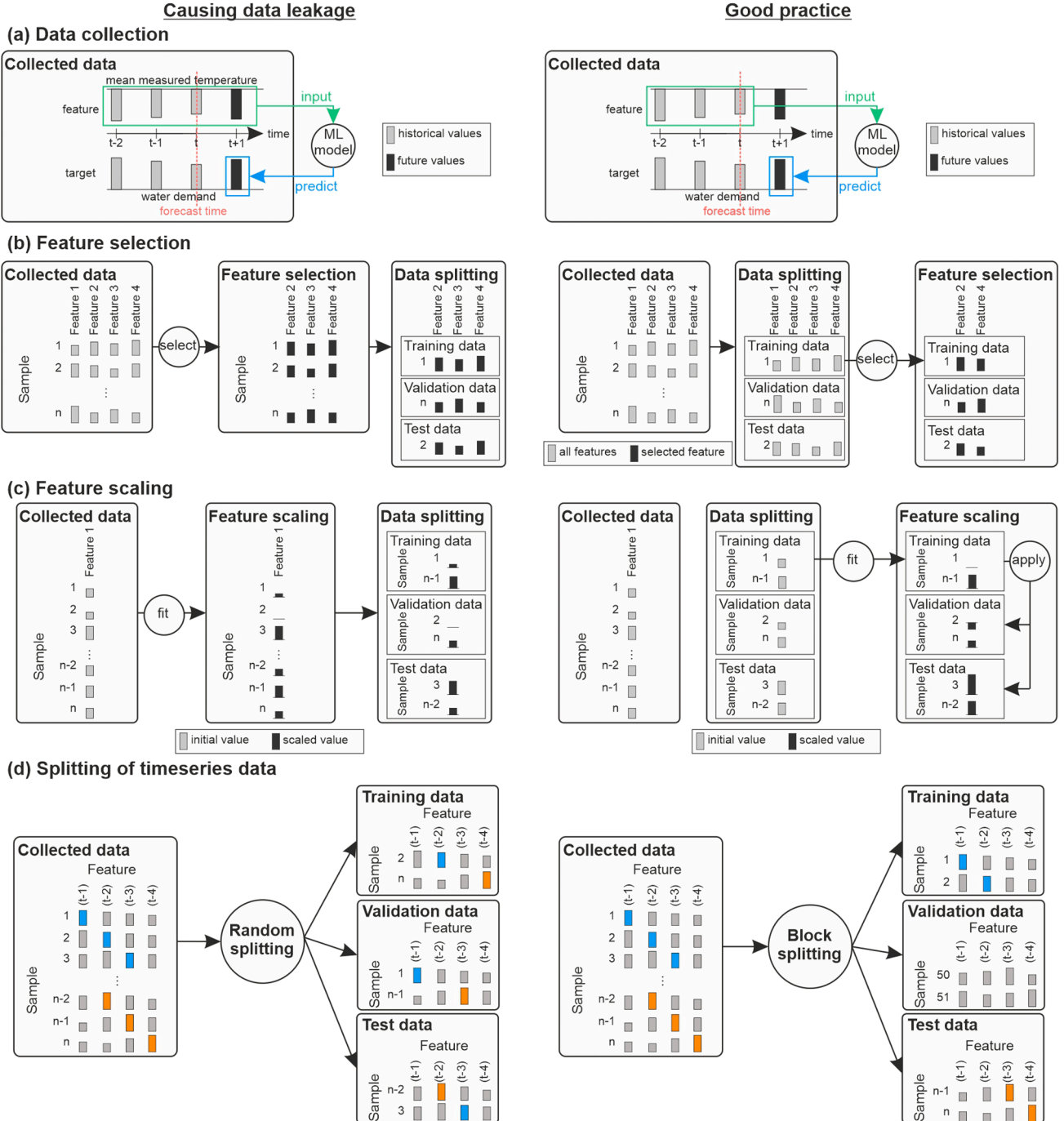


Fig. 5. Approaches causing data leakages (left) in comparison with good practice (right) for (a) data collection, (b) feature selection, (c) feature scaling, and (d) splitting of timeseries data.

2025). This is particularly relevant in timeseries applications common in UWI, where random splitting can lead to overlapping historical input features across training, validation, and test sets and may break temporal dependence. Recommended practices include splitting timeseries into three consecutive and independent blocks for training, validation, and test set as outlined in Fig. 5(d). (Namdari et al., 2023; Zhang et al., 2022), or applying event-based grouping strategies, such as rainfall-based splitting in UDN application (Yang and Chui, 2021).

4.3. Model development

Based on the pre-processed data, ML models are developed. This involves selecting hyperparameters for training, evaluating the trained ML model, and the final model selection. The results of the literature review are shown in Fig. 6.

ML models are governed by the choice of non-learned model configuration parameters, called hyperparameters, that influence the learning process and must therefore be carefully selected, as they can significantly affect achievable performance (Bischl et al., 2023; Karl

et al., 2023). For neural networks, for instance, key hyperparameter include the number of hidden layers and the number of neurons per layer (Yu and Zhu, 2020). Menapace et al. (2021) demonstrated that appropriate hyperparameter optimisation for water demand forecasting can increase performance by up to 75% relative to a minimally tuned baseline architecture. Despite this importance, only 41% of UDN articles and 41% of WDN articles reported automatic hyperparameter optimisation, as depicted in Fig. 6(a), increasing to 56% and 54% respectively when non-systematic trial-and-error approaches are included. According to Fig. 6(b), grid-search is the most common optimisation approach, used in >60% of the articles. It discretises each hyperparameter range and evaluates every possible combination. While grid-search can provide good results with sufficiently fine discretisation, and is easy to interpret, computation time increases exponentially with the number of parameters and their discretisation steps. It is therefore typically practical only for a small number of hyperparameters (Bischl et al., 2023). Random search offers a more efficient alternative, as used by Chen et al. (2022) to optimise six hyperparameters. Through a random sampling from predefined distributions, it examines a wider range of

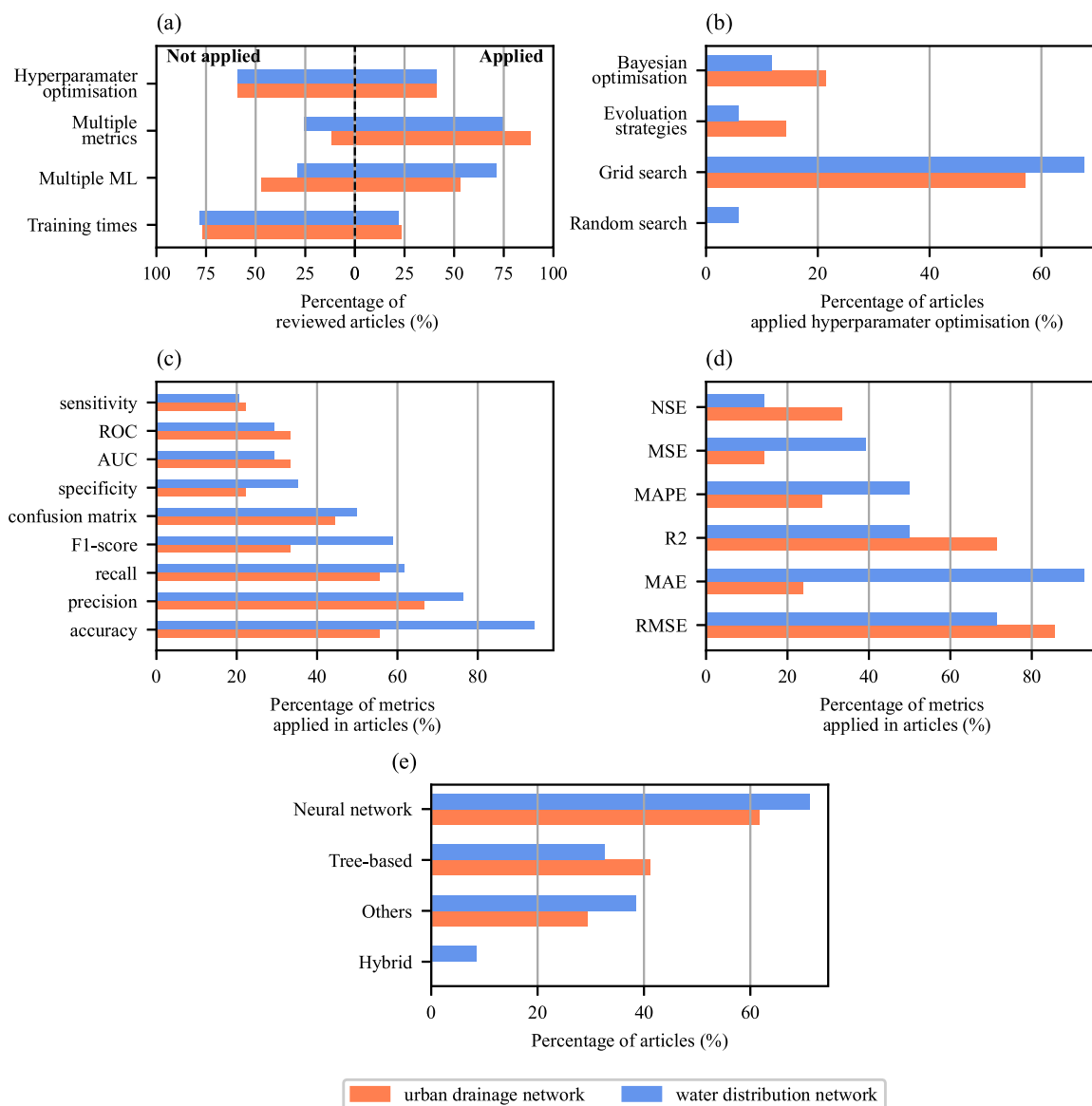


Fig. 6. Evaluation of the implementation quality for model development subdivided into UDN and WDN: (a) analyse of the applied steps for model development, (b) overview of utilised methods if hyperparameter optimisation is applied, (c) utilised performance indicators for classification, (d) utilised performance indicators for regression, (e) selected ML model categories.

configurations and increases the chance of identifying an optimal configuration (Bischl et al., 2023). For highly complex optimisation problems, involving many hyperparameters, fine discretisation, or heterogeneous hyperparameter influence, advanced methods such as evaluation strategies including genetic algorithm (Aderyani et al., 2022; Candelieri et al., 2019) or Bayesian optimisation (Kim et al., 2024; Piadeh et al., 2023; Zhou et al., 2024) should be used, as they refine the search area iteratively based of previous evaluations.

After final training with optimised hyperparameter, the predictive performance of the ML model is evaluated. The literature recommends using multiple metrics and comparing multiple ML models, as each has different characteristics (Lones, 2024; Zhu et al., 2023). As shown in Fig. 6(a), this practice is widely adopted in UWI, roughly 75% of reviewed articles reported at least two metrics and compared at least two ML models. For example, Kavya et al. (2023) evaluated nine ML models, including simple ones such as linear regression, using four metrics to select the best model for short-term water demand forecasting. Similarly, Piadeh et al. (2023) compared seven ML models using seven metrics, concluding that no ML model is optimal, an insight that itself can support improved ML development.

For classification tasks, accuracy is most reported, followed by precision, recall, and F1-score calculated based on the confusion matrix as illustrated in Fig. 6(c). However, accuracy should be used cautiously with imbalanced data sets, a frequent situation in UWI, because this metric gives greater weight to the majority class. Metrics derived from the confusion matrix, that explicitly account for class-specific performance, are therefore preferred (Wang et al., 2021). An alternative is to rebalance the dataset, as applied by Beig Zali et al. (2024) and Dang et al. (2021), who adjusted class distribution for pipe defect classification in WDN and UDN. However, such approaches should be applied carefully as they may affect the representativeness of the original data distribution.

For regression tasks, Fig. 6(d) shows that root mean squared error (RMSE) is the most widely used metric, followed by mean absolute error (MAE), coefficient of determination (R^2) and mean absolute percentage error (MAPE). However, R^2 and the Nash-Sutcliffe efficiency (NSE) should be interpreted with caution: R^2 describes primarily reflects the proportion of variance explained relative to a baseline rather than absolute predictive error (Plevris et al., 2022; Zhu et al., 2023), while NSE introduced for evaluating continuous time-series simulations is highly sensitive to data variability, limiting comparison across case studies (Williams, 2025).

Reporting computational requirements, such as training time and prediction time, is another important information, especially for real-time applications such as automatic anomaly detection. Nevertheless, as presented in Fig. 6(a), only 24% of UDN articles and 22% of WDN articles documented training time, despite its relevance for reproducibility and comparison.

The implemented ML model architecture categories are shown in Fig. 6(e) for UDN and WDN. Neural network methods represent the majority, used in around 65% of the articles in both fields. Convolutional neural networks (CNN) are most common, followed by artificial neural networks (ANN) with one hidden layer, long-short term memory (LSTM) networks, and multilayer perceptron (MLP). Tree-based methods account for 41% of UDN models and 33% for WDN models, with random forest (RF) and XGBoost the most widely used. Other methods, such as support vector machines (SVM) and logistic regression (Log R.), are utilised in 29% of UDN articles and 38% of WDN articles.

Hybrid approaches combining different ML model architectures (such as combining CNN and LSTM architectures) are also relatively frequent, especially for WDNs. While such designs aim to capture complementary dynamics, their added complexity is not always supported by clear justification or systematic ablation. It is plausible that comparable performance could be achieved through better hyperparameter optimisation of simpler architectures, calling into question whether the added complexity is always warranted and the quality of

the benchmarking. Increasing model complexity without sufficient justification undermines both explainability and reproducibility, while offering uncertain practical benefit.

4.4. Model application

After selecting the final ML model, the next step is to prepare it for deployment. This includes integrating model explainability, applying uncertainty quantification techniques, and ensuring a proper documentation, including the sharing of data and code. The results of the literature review are shown in Fig. 7(a).

A key requirement for trustworthy AI is that the decision-making process within the ML model can be understood and traced by humans (Ali et al., 2023). ML models can be categorised in white-box or black-box models (Došilović et al., 2018). White-box models, such as linear regression or decision trees, have an integrated interpretability through their structure. In contrast, black-box models have complex model structures without being directly comprehensible and thus require post-hoc interpretability methods (Belle and Papantonis, 2021). As illustrated in Fig. 7(a), explainability was documented in 26% of UDN articles and 14% of WDN articles. Among these, feature relevance explanations were most common; Fig. 7(b) shows that methods such as SHapley Additive exPlanations (SHAP) (Kumar et al., 2025; Riyadh and Peleato, 2025; Yang and Chui, 2021) and permutation importance (Daulat et al., 2024) were applied in over 77% of UDN articles and 75% of WDN articles. SHAP provides a local explanation of the contribution of each feature to the target variable, whereas permutation importance measures the impact of randomly reshuffling each feature on model performance (Dwivedi et al., 2023). Visual explanations, including partial dependence plots (PDP) (Baziar et al., 2025) and Class Activation Mapping (CAM) (Dang et al., 2021; Xu et al., 2025), were used in 44% of UDN articles and 33% of WDN articles. PDPs show the mean change in model predictions when one feature is varied (Dwivedi et al., 2023), while CAMs highlight the regions of an image that are most influential in a CNN model (Bhati et al., 2024). It should be noted, however, that post-hoc explainability methods attempt to explain the decision-making process of ML models in different approaches and may provide incorrect explanations (Huang and Marques-Silva, 2024). Consequently, it is advisable to use different explanation techniques to address different questions about the ML model (Belle and Papantonis, 2021) to gain deeper insights into the decision-making process (Yu et al., 2024; Zhou et al., 2022). One article used a white-box model, logistic regression, as its final model (Robles-Velasco et al., 2020), which, due to its model structure, allows the direct explanation of the influence of individual features on the result.

The data used in UWI are subject to uncertainties, including missing values, data scarcity, noise, and labelling errors. Furthermore, methodological decisions during data preprocessing and model development, including data splitting, feature selection, hyperparameter optimisation, and final ML model selection, introduce model uncertainties. These uncertainties influence achievable performance, and good practice recommends the use of uncertainty quantification techniques to increase confidence in ML outputs (Fakour et al., 2024). However, fewer than 21% of UDN articles and 13% of WDN articles documented uncertainty quantification techniques as shown in Fig. 7(a). Fig. 7(c) summarises the approaches used. Ensemble models are most common, applied in 71% of UDN articles and 82% of WDN articles applied uncertainty quantification techniques. These either repeat model training multiple times (Chen et al., 2022; Chu et al., 2024), or combine outputs from different ML models (Jun and Jung, 2025; Kang et al., 2018; Li et al., 2022; Zounemat-Kermani et al., 2020). Deterministic approaches, primarily quantile loss functions (Iwakin and Moazeni, 2024), were used in 14% of UDN articles and 9% of WDN articles. Bayesian neural networks, with Monte-Carlo dropout as approximation (Fan et al., 2023), or conformal prediction (Reshadi et al., 2025) were applied in one article each.

The limited documentation of explainability and uncertainty

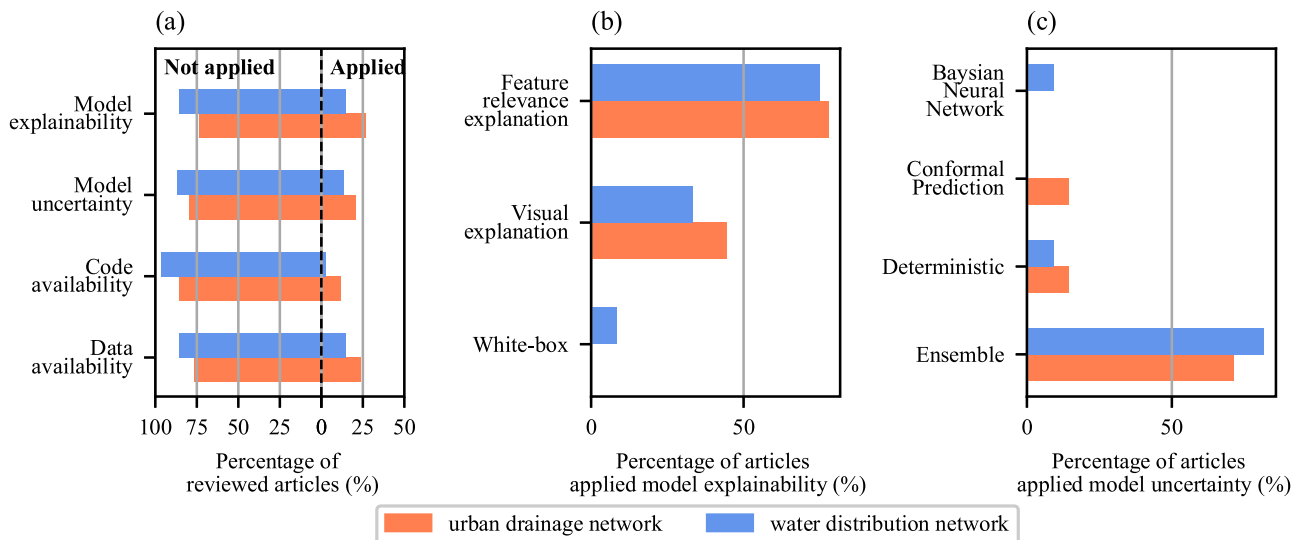


Fig. 7. Evaluation of the implementation quality for model application: (a) analyse of the applied steps for model application, (b) overview of utilised methods if model explainability is applied, and (c) overview of utilised methods if model uncertainty is applied.

quantification techniques is particularly problematic as many UWI applications are classified as high-risk, where failures may negatively affect both urban population and infrastructure itself. Without these techniques, reported performance gains remain difficult to trust, explain, or operationalise, thereby increasing the gap between academic development and real-world deployment.

Finally, the ML development process, including the selected hyperparameters, should be fully documented including data and code sharing to improve traceability and transparency. The data extraction process showed that reporting across the reviewed articles lack standardisation, making the extraction of relevant information difficult and increasing the risk of errors for repeating the performed experiments. This challenge compound by the minimal availability of shared code or data as illustrated in Fig. 7(a). Only 11% of UDN articles and 2% of WDN articles provided freely accessible code. Data availability is slightly better at 23% for UDN and 14% for WDN, partly due to the emergence of open-source benchmarking data. Examples include datasets for water demand forecasting (Alvisi et al., 2025), leakage management (Vrachimis et al., 2022), and cyber-physical attacks (Faramondi et al., 2021; Taormina et al., 2018) in WDN, or sewer defect classification dataset (Haurum and Moeslund, 2021), long-term observations (Blumensaat et al., 2025), and hydraulic modelling (Nedergaard Pedersen et al., 2021) in UDN. Additionally, three reviewed articles contained links no longer valid, highlighting the importance of using permanent repositories for long-term availability of code and data.

5. Future challenges and recommended methodological checklist

5.1. Future challenges

The evaluation of the reviewed articles revealed two major issues in the implementation quality of supervised ML applications in the field of UWI. First, the reporting of correct implementation of ML methodology and its practical applicability remain insufficiently documented as shown by the reviewed articles. Second, the lack of uniform and standardised documentation makes it challenging to extract essential information and limits reproducibility. Together, these issues negatively affect reported performance gains and ultimately reduce trust in the developed approaches. This is particularly important, as trust and confidence in ML-based decisions are key requirements for successful transformation of research in practice (Rapp et al., 2023; Sela et al.,

2024).

As discussed by Rapp et al. (2023) and Richards et al. (2023), ML is despite the growing research rarely implemented in real-world applications in UWI and where it is, applications typically remain at an early stage. From the perspective of the operators, lack of confidence in the ML application has been recognised as one of the main barriers (Rapp et al., 2023). Consequently, trust in ML-based applications emerges as a key criterion for successful integration (Sela et al., 2024). This issue is particularly critical because UWI failures or incorrect decisions based on ML recommendations could negatively impact both urban population and urban environment. Hence, ensuring trust, requires a state-of-the-art implementation throughout the entire development process, from data preprocessing to model application, to guarantee traceability and transparency for operators. Moreover, this aspect is increasingly reflected in legal frameworks world-wide (Hilliard et al., 2026; Lund et al., 2025). For example, the European “Artificial Intelligence Act” (European Parliament and Council, 2024) as one of the most comprehensive and binding frameworks, classify WDN applications as high-risk systems when an error could affect the physical integrity of the infrastructure, or health and safety of the population, and ML applications must ensure transparency and technical robustness (Mausner et al., 2025).

At the same time, implementing ML applications has become relatively easy today. Open-source libraries in programming languages such as Python (Raschka et al., 2020) and R (Bischl et al., 2016) support a fast development of ML applications in recent years. In addition, generative AI tools, such as ChatGPT, Claude, and DeepSeek, can assist in preparing and executing ML applications on demand (Palma et al., 2025; Ray, 2023; Wang et al., 2025). However, this increased accessibility also introduces challenges. The reduced requirement of detailed background knowledge may lead to a lack of coding and ML skills in general among developers, technicians, and operators, who are usually traditional hydraulic engineers (Sela et al., 2024). This can negatively affect the quality of ML implementation throughout the development process, particularly as the reliability of generative AI outputs may vary (Cong-Lem et al., 2024; Thelwall et al., 2025). Furthermore, these tools are trained on diverse datasets including scientific articles (Hu et al., 2025), meaning that existing methodological shortcomings may propagate into generated outputs. Consequently, the need of standardised and methodologically correct implementation becomes even more essential.

5.2. Methodological and reporting checklist

To address these shortcomings, a methodological and reporting checklist for supervised ML studies is introduced to help avoid common pitfalls, as depicted in Fig. 8. The checklist is intended to support developers, technicians, and operators in future ML implementation approaches. It is important to emphasise that this checklist provides general guidance towards a correct ML implementation rather than a fixed recipe, and case specific adjustments are necessary. For more detailed explanations of the individual steps, it is referred to the cited literature, which provides an overview of established methods for this task.

The starting point for this checklist assumes that data collection from relevant sources has been completed. Consequently, the initial step following the data collection is to divide the data into independent training, validation, and test sets (TVT), each used for different tasks during model development. For timeseries data, a block-based splitting approach is required to avoid data leakage, whereas random splitting can be used for other data types. As a general recommendation, the test set should comprise between 10 and 40% of the total dataset size to ensure a fair evaluation of the ML model and should be representative for the selected application.

Following data splitting, all data preprocessing strategies are defined and implemented exclusively on the training set to avoid information leakage. This includes outlier detection and removal (Boukerche et al., 2020; Smiti, 2020), and missing data management (Emmanuel et al., 2021; Joel et al., 2022). Considering that UWI data typically consists of parameters with different ranges, feature scaling is applied to improve performance (Ahsan et al., 2021; Géron, 2022; Pinheiro et al., 2025). (Dhal and Azad, 2021; Pudjihartono et al., 2022). The learned preprocessing rules are then applied consistently to the validation and test sets. Feature selection is then performed to identify the most informative features. After preprocessing, the ratio of training samples to features should ideally be greater than or equal to 100 to ensure sufficient learning. If this is not fulfilled, CVT should be used for data splitting, or improvements in data collection should be considered.

The subsequent step involves hyperparameter optimisation (Bischl et al., 2023; Karl et al., 2023; Yu and Zhu, 2020), followed by model training using the optimised hyperparameter and architecture. In both steps with TVT splitting, the training set is used for learning, while the validation set is used to avoid under- or overfitting. Model performance is then evaluated on the test set using appropriate performance metrics, distinct for classification (Naidu et al., 2023) and regression (Botchkarev, 2019; Plevris et al., 2022). Moreover, several ML models should be compared to select the most suitable one. As a minimum requirement, at least two performance metrics should be used, and performance should ideally be compared against an appropriate baseline model to contextualise the added value of the selected ML approach (Lones, 2024; Zhu et al., 2023). Such baselines may include simple or interpretable models (e.g., linear regression or decision trees), as well as established methods reported in the literature. Where available, comparisons against published benchmarks on open-source datasets, or against existing implementations with accessible code, are particularly valuable for ensuring fair evaluation and meaningful performance claims.

Once the best-performing ML model has been selected, it is prepared for deployment. This includes the integration of post-hoc explainability techniques to make the model's decision-making process understandable to humans (Ali et al., 2023; Belle and Papantonis, 2021). Because model performance is affected by uncertainties arising from data preprocessing and model development decisions, uncertainty quantification techniques should be incorporated (Fakour et al., 2024; Hüllermeier and Waegeman, 2021). The final step consists of a proper documentation of the ML development process, including a sharing of data and code to enhance traceability, transparency, and reproducibility.

As an alternative for applications with limited data, the dataset can

be divided into a training set with cross-validation and a separate test set (CV+T). The training set is further divided into n folds. In each iteration, $n-1$ folds are used to define and fit all preprocessing steps, which are then applied to the remaining fold to avoid information leakage. For feature-based models, a commonly used guideline is that the ratio between the number of training samples in the $n-1$ folds and the number of input features should be sufficiently large (e.g., ≥ 10), although this depends on model complexity. Optimal hyperparameters are selected based on mean validation performance, after which the final model is trained on the full training set and evaluated on the independent test set.

This methodological and reporting checklist is provided as an editable Word figure in the supplementary materials, enabling its use as a standardised documentation template.

5.3. Further recommendations and future research topics

The introduced checklist provides methodological and reporting guidelines for supervised ML in UWI. However, several emerging developments highlight the need for increasingly rigorous and transparent ML implementations. One such trend is the integration of domain knowledge throughout the ML pipeline, often referred to as knowledge fusion, to improve efficiency and robustness. During feature selection, engineering knowledge and hydraulic information can be embedded as physically meaningful features to better represent the underlying system (Hu et al., 2026, 2018; Tian et al., 2025). During model development, physical laws can be included into the loss function to ensure mass and energy conservation for state estimations in UWI networks (Garzon et al., 2026; Xing and Sela, 2022). During model evaluation, application specific evaluation metrics can be developed to better reflect operational objectives. Furthermore, the increasing availability of open benchmark data enables a comparison of developed ML applications throughout the articles, further enhancing the need for a common and transparent ML workflow.

Another important development is the increasing application of explainability techniques to help humans understand the model's decisions (Infant et al., 2025). This can be further extended using causal techniques to derive causal relationships between features and the target variable (Chen et al., 2025; Zhang et al., 2026).

Additionally, the long-term maintenance of developed ML models is rarely addressed in the literature. In practice, however, deployment requires continuous monitoring of ML model performance as it is highly susceptible to sensor drift and changing system conditions, which can reduce performance over time. Therefore, an adaptation of SCADA systems is required to enable continuous re-training, adaptation, and validation of the implemented ML pipeline for a long-term use in UWI.

These developments are, however, beyond the scope of this review, which presents promising future research directions to support the real-world deployment of the developed ML applications in the field of UWI.

6. Conclusions

In this work, the current implementation quality of machine learning (ML) applications in the field of urban water infrastructure (UWI) was assessed across all stages of the ML development process. A total of 117 highly cited articles published between 2018 and 2025, selected under the assumption that they have the greatest influence on future research, were systematically reviewed and critically evaluated. Particularly attention was paid to common pitfalls throughout the development process. Based on the results of this review, the following conclusions can be drawn:

- Most reviewed articles place strong emphasis on performance benchmarking. The use of multiple performance metrics and comparison of different ML models is widely adopted, providing a state-of-the-art evaluation of the technical performance.

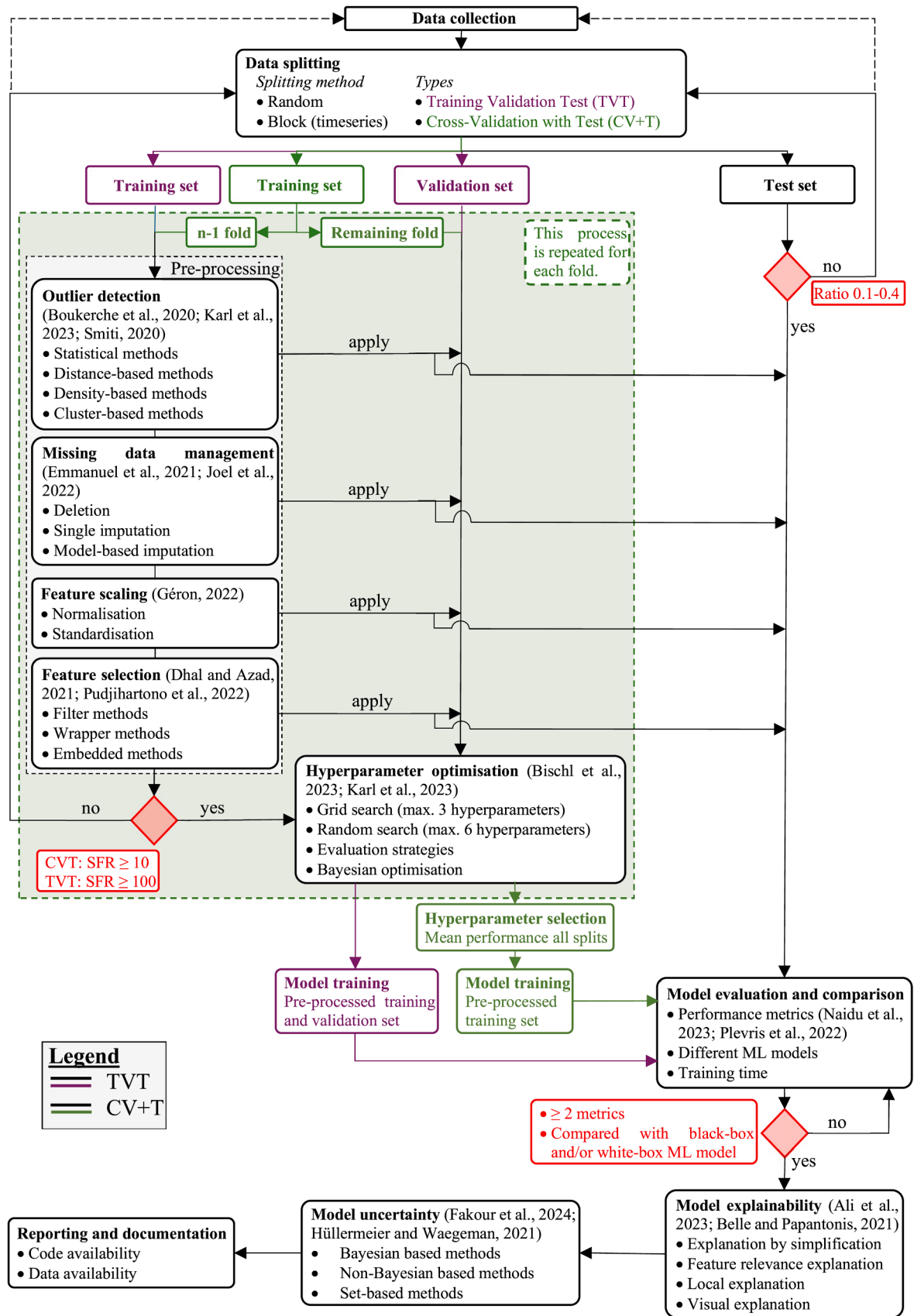


Fig. 8. Methodological and reporting checklist for supervised machine learning in urban water infrastructure.

- In contrast, there is a noticeable gap in the reporting of correct ML methodology implementation and practical applicability. Only around one third of the reviewed articles reported essential pre-processing tasks such as feature selection, feature scaling, or automatic hyperparameter optimisation, despite their clear influence on performance and generalisation. Likewise, aspects related to model explainability and uncertainty were seldom documented.
- Additionally, there is a lack of standardised documentation of the developed ML applications and often insufficiently detailed, making it difficult to extract essential methodological information for reproducibility. Data and code sharing is also extremely limited.
- Based solely on workflow illustrations and textual descriptions, evidence of data leakage was found in around 15% of the articles. Data leakage refers to the unintended use of information during model training that would not be available in practice, leading to artificially increased performance.

By showing that even frequently cited articles exhibit gaps in reporting the correct methodological implementation, this review highlights the need for more verifiable, reproducible and transparent ML applications in UWI. To address this, a methodological and reporting checklist was provided to guide the design and integration of ML applications in UWI throughout the development process and to help avoid common pitfalls. This checklist is intended for use by developers, technicians, and operators, supporting future ML implementations.

Funding

This publication is part of the FOUND project that has received funding from the Federal Ministry of Agriculture and Forestry, Climate and Environmental Protection, Regions and Water Management (BMLUK) (Austria) under project number C300198 and of the DECIRE-WATER project that has received funding from the European Union's Horizon Europe research and innovation programme under grant agreement No. 101-182-154. Views and opinions expressed are however those of the authors only and do not necessarily reflect those of the funding agencies. Neither the funding agencies nor the granting authorities can be held responsible for them.

Declaration of AI and AI-assisted technologies in the writing process

During the preparation of this work, the authors used GPT 5 to improve readability and language. After using this tool, the authors reviewed and edited the content as needed and take full responsibility for the content of the publication.

CRedit authorship contribution statement

Martin Oberascher: Writing – original draft, Visualization, Methodology, Investigation, Formal analysis, Conceptualization. **Bruno M. Brentan:** Writing – review & editing, Validation, Methodology, Conceptualization. **Andrea Menapace:** Writing – review & editing, Validation, Methodology, Conceptualization. **Manuel Herrera:** Writing – review & editing, Validation, Methodology, Conceptualization. **Guangtao Fu:** Writing – review & editing, Validation, Methodology, Conceptualization. **Riccardo Taormina:** Writing – review & editing, Validation, Methodology, Conceptualization. **Robert Sitzenfrei:** Writing – review & editing, Validation, Methodology, Conceptualization.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Supplementary materials

Supplementary material associated with this article can be found, in the online version, at [doi:10.1016/j.wroa.2026.100569](https://doi.org/10.1016/j.wroa.2026.100569).

Data availability

The data that has been used is confidential

References

- Aderiyani, F.R., Jamshid Mousavi, S., Jafari, F., 2022. Short-term rainfall forecasting using machine learning-based approaches of PSO-SVR, LSTM and CNN. *J. Hydrol.* 614, 128463. <https://doi.org/10.1016/j.jhydrol.2022.128463>.
- Ahmed, A.A., Sayed, S., Abdoulhalik, A., Moutari, S., Oyedele, L., 2024. Applications of machine learning to water resources management: a review of present status and future opportunities. *J. Clean. Prod.* 441, 140715. <https://doi.org/10.1016/j.jclepro.2024.140715>.
- Ahsan, M., Mahmud, M., Saha, P., Gupta, K., Siddique, Z., 2021. Effect of data scaling methods on machine learning algorithms and model performance. *Technologies* 9 (3), 52. <https://doi.org/10.3390/technologies9030052>.
- Ali, S., Abuhmed, T., El-Sappagh, S., Muhammad, K., Alonso-Moral, J.M., Confalonieri, R., Guidotti, R., Del Ser, J., Díaz-Rodríguez, N., Herrera, F., 2023. Explainable artificial intelligence (XAI): what we know and what is left to attain trustworthy artificial intelligence. *Inform. Fusion.* 99, 101805. <https://doi.org/10.1016/j.inffus.2023.101805>.
- Alvisi, S., Franchini, M., Marsili, V., Mazzoni, F., Salomons, E., Housh, M., Abokifa, A., Arsova, K., Ayyash, F., Bae, H., Barreira, R., Basto, L., Bayer, S., Berglund, E.Z., Biondi, D., Boloukasli Ahmadgourabi, F., Brentan, B., Caetano, J., Campos, F., Cao, H., Cardona, S., Carreño Alvarado, E.P., Carriço, N., Chatzistefanou, G.A., Coy, Y., Creaco, E., Cuomo, S., de Klerk, A., Di Nardo, A., DiCarlo, M., Dittmer, U., Dziedzic, R., Ebrahim Bakhshpour, A., Eliades, D., Farmani, R., Ferreira, B., Gabriele, A., Gamboa-Medina, M.M., Gao, F., Gao, J., Gargano, R., Geranmehr, M., Giudicianni, C., Glynis, K., Gómez, S., González, L., Groß, M., Guo, H., Habibi, M.N., Haghighi, A., Hammer, B., Hans, L., Hayslep, M., He, Y., Hermes, L., Herrera, M., Hinder, F., Hou, B., Iglesias-Rey, A., Iglesias-Rey, P.L., Jang, I.S., Izquierdo, J., Jahangir, M.S., Jara-Arriagada, C., Jenks, B., Johnen, G., Pesantez, J., Polycarpou, M., Kalumba, M., Kang, M.S., Khashei Varnamkhashi, M., Kim, K.J., Kley-Holsteg, J., Ko, T., Koochali, A., Kossieris, P., Koundouri, P., Kühnert, C., Kulaczkowski, A., Lee, J., Li, K., Li, Y., Liu, H., Liu, Y., López-Hojas, C.A., Maier, A., Makropoulos, C., Martínez-Solano, F.J., Marzouny, N.H., Menapace, A., Michalopoulos, C., Moraitis, G., Mousa, H., Namdari, H., Nikolopoulos, D., Oberascher, M., Ostfeld, A., Pagano, M., Pasha, F., Perafán, J., Perelman, G., Pesantez, J., Polycarpou, M., Quarta, M.G., Que, Q., Quilty, J., Quintiliani, C., Ramachandran, A., Reynoso Meza, G., Rodriguez, V., Romano, Y., Saldarriaga, J., Salem, A.K., Samartzis, P., Santonastaso, G.F., Savic, D., Schiano Di Cola, V., Schol, D., Seyoum, A.G., Shen, R., Simukonda, K., Sinske, A., Sitzenfrei, R., Sonnenschein, B., Stoianov, I., Tabares, A., Todini, E., Tsiame, L., Tsoukalas, I., Ulusoy, A.J., Vamvakieridou-Lyroudia, L., van Heerden, A., Vaquet, J., Vaquet, V., Wallner, S., Walraad, M., Wang, D., Wu, S., Wu, W., Wunsch, A., Yao, Y., Yu, J., Zanfei, A., Zanutto, D., Zhang, H., Ziebarth, M., Ziel, F., Zou, J., 2025. Battle of water demand forecasting. *J. Water Resour. Plan. Manag.* 151 (10). <https://doi.org/10.1061/jwrm5.Wreng-6887>.
- Apicella, A., Isgrò, F., Prevete, R., 2025. Don't push the button! exploring data leakage risks in machine learning and transfer learning. *Artif. Intell. Rev.* 58 (11). <https://doi.org/10.1007/s10462-025-11326-3>.
- Baziar, M., Behnami, A., Jafari, N., Mohammadi, A., Abdolhahnejad, A., 2025. Machine learning-based Monte Carlo hyperparameter optimization for THMs prediction in urban water distribution networks. *J. Water Process Eng.* 73, 107683. <https://doi.org/10.1016/j.jwpe.2025.107683>.
- Beig Zali, R., Latifi, M., Javadi, A.A., Farmani, R., 2024. Semisupervised clustering approach for pipe failure prediction with imbalanced data set. *J. Water Resour. Plan. Manag.* 150 (2). <https://doi.org/10.1061/jwrm5.Wreng-6263>.
- Belle, V., Papantonis, I., 2021. Principles and practice of explainable machine learning. *Front. Big. Data* 4, 688969. <https://doi.org/10.3389/fdata.2021.688969>.
- Bhati, D., Neha, F., Amiruzzaman, M., 2024. A survey on explainable artificial intelligence (XAI) techniques for visualizing deep learning models in medical imaging. *J. Imaging* 10 (10), 239. <https://doi.org/10.3390/jimaging10100239>.
- Bischl, B., Binder, M., Lang, M., Pielok, T., Richter, J., Coors, S., Thomas, J., Ullmann, T., Becker, M., Boulesteix, A.L., Deng, D., Lindauer, M., 2023. Hyperparameter optimization: foundations, algorithms, best practices, and open challenges. *WIREs Data. Min. Knowl. Discov.* 13 (2). <https://doi.org/10.1002/widm.1484>.
- Bischl, B., Lang, M., Kotthoff, L., Schiffer, J., Richter, J., Studerus, E., Casalicchio, G., Jones, Z.M., 2016. mlr: machine learning in R. *J. Mach. Learn. Res.* 17 (170), 1–5.
- Blumensaat, F., Bloem, S., Ebi, C., Disch, A., Förster, C., Maurer, M., Rodriguez, M., Rieckermann, J., 2025. The UWO dataset – long-term observations from a full-scale field laboratory to better understand urban hydrology at small spatio-temporal scales. *Earth Syst. Sci. Data Discuss.* 2025, 1–31. <https://doi.org/10.5194/essd-2024-47>.
- Boo, K.B.W., El-Shafie, A., Othman, F., Khan, M.M.H., Birima, A.H., Ahmed, A.N., 2024. Groundwater level forecasting with machine learning models: a review. *Water. Res.* 252, 121249. <https://doi.org/10.1016/j.watres.2024.121249>.

- Botchkarev, A., 2019. A new typology design of performance metrics to measure errors in machine learning regression algorithms. *Interdiscip. J. Inf.* 14, 045–076. <https://doi.org/10.28945/4184>.
- Boukerche, A., Zheng, L., Alfandi, O., 2021. Outlier detection. *ACM. Comput. Surv.* 53 (3), 1–37. <https://doi.org/10.1145/3381028>.
- Candelieri, A., Giordani, I., Archetti, F., Barkalov, K., Meyerov, I., Polovinkin, A., Sysoyev, A., Zolotykh, N., 2019. Tuning hyperparameters of a SVM-based water demand forecasting system through parallel global optimization. *Comput. Oper. Res.* 106, 202–209. <https://doi.org/10.1016/j.cor.2018.01.013>.
- Chauhan, K., Jani, S., Thakkar, D., Dave, R., Bhatia, J., Tanwar, S., Obaidat, M.S., 2020. Automated machine Learning: the new wave of machine learning. In: 2020 2nd International Conference on Innovative Mechanisms for Industry Applications (ICIMIA). <https://doi.org/10.1109/ICIMIA48430.2020.9074859>.
- Chen, J.-j., Wang, H.-C., Yin, W.-X., Wang, Y.-Q., Lv, J.-q., Wang, A.-J., 2026. From meteorological perturbation mechanisms to sewer methane forecasting: a causal and machine learning approach. *ACS ES&T Eng* 6 (1), 325–339. <https://doi.org/10.1021/acsestengg.5c00822>.
- Chen, L., Yan, H., Yan, J., Wang, J., Tao, T., Xin, K., Li, S., Pu, Z., Qiu, J., 2022. Short-term water demand forecast based on automatic feature extraction by one-dimensional convolution. *J. Hydrol.* 606, 127440. <https://doi.org/10.1016/j.jhydrol.2022.127440>.
- Chu, W., Zhang, C., Li, H., Zhang, L., Shen, D., Li, R., 2024. SHAP-powered insights into spatiotemporal effects: unlocking explainable bayesian-neural-network urban flood forecasting. *Int. J. Appl. Earth. Obs* 131, 103972. <https://doi.org/10.1016/j.jag.2024.103972>.
- Cong-Lem, N., Soyoof, A., Tsering, D., 2024. A systematic review of the limitations and associated opportunities of ChatGPT. *Int. J. Human-Comput. Inter.* 1–16. <https://doi.org/10.1080/10447318.2024.2344142>.
- Creaco, E., Campisano, A., Fontana, N., Marini, G., Page, P.R., Walski, T., 2019. Real time control of water distribution networks: a state-of-the-art review. *Water. Res.* 161, 517–530. <https://doi.org/10.1016/j.watres.2019.06.025>.
- Dang, L.M., Kyeong, S., Li, Y., Wang, H., Nguyen, T.N., Moon, H., 2021. Deep learning-based sewer defect classification for highly imbalanced dataset. *Comput. Ind. Eng.* 161, 107630. <https://doi.org/10.1016/j.cie.2021.107630>.
- Daniel, I., Ajami, N.K., Castelletti, A., Savic, D., Stewart, R.A., Cominola, A., 2023. A survey of water utilities' digital transformation: drivers, impacts, and enabling technologies. *NPJ. Clean. Water* 6 (1). <https://doi.org/10.1038/s41545-023-00265-7>.
- Daulat, S., Rokstad, M.M., Bruaset, S., Langeveld, J., Tschekner-Grat, F., 2024. Evaluating the generalizability and transferability of water distribution deterioration models. *Reliab. Eng. Syst. Saf.* 241, 109611. <https://doi.org/10.1016/j.res.2023.109611>.
- Dhal, P., Azad, C., 2022. A comprehensive survey on feature selection in the various fields of machine learning. *Appl. Intell.* 52 (4), 4543–4581. <https://doi.org/10.1007/s10489-021-02550-9>.
- Doorn, N., 2021. Artificial intelligence in the water domain: opportunities for responsible use. *Sci. Total. Environ.* 755 (Pt 1), 142561. <https://doi.org/10.1016/j.scitotenv.2020.142561>.
- Došilović, F.K., Bricić, M., Hlupić, N., 2018. Explainable artificial intelligence: a survey. In: 2018 41st International Convention on Information and Communication Technology, Electronics and Microelectronics (MIPRO). <https://doi.org/10.23919/MIPRO.2018.8400040>.
- Dwivedi, R., Dave, D., Naik, H., Singhal, S., Omer, R., Patel, P., Qian, B., Wen, Z., Shah, T., Morgan, G., Ranjan, R., 2023. Explainable AI (XAI): core ideas, techniques, and solutions. *ACM Comput. Surv.* 55 (9), 1–33. <https://doi.org/10.1145/3561048>.
- Ebtehaj, I., Bonakdari, H., Safari, M.J.S., Gharabaghi, B., Zaji, A.H., Riahi Madavar, H., Sheikh Khozani, Z., Es-haghi, M.S., Shishegaran, A., Danandeh Mehr, A., 2020. Combination of sensitivity and uncertainty analyses for sediment transport modeling in sewer pipes. *Int. J. Sediment Res.* 35 (2), 157–170. <https://doi.org/10.1016/j.ijsrsc.2019.08.005>.
- Eelbode, T., Sinonquel, P., Maes, F., Bisschops, R., 2021. Pitfalls in training and validation of deep learning systems. *Best Pract. Res. Clin. Gastroenterol* 52–53, 101712. <https://doi.org/10.1016/j.bpg.2020.101712>.
- Emmanuel, T., Maupong, T., Mpoeleng, D., Semong, T., Mphago, B., Tabona, O., 2021. A survey on missing data in machine learning. *J. Big. Data* 8 (1), 140. <https://doi.org/10.1186/s40537-021-00516-9>.
- European Parliament and Council, 2024. Regulation of the European Parliament and of the Council on laying down harmonised rules on artificial intelligence and amending regulations (Artificial Intelligence Act), OJ 2024 L 2024/1689.
- Fakour, F., Mosleh, A., Ramezani, R., 2024. A structured review of literature on uncertainty in machine learning & deep learning. *arXiv preprint*. <https://doi.org/10.48550/arXiv.2406.00332>.
- Fan, X., Wang, X., Zhang, X., Yu, X., 2022. Machine learning based water pipe failure prediction: the effects of engineering, geology, climate and socio-economic factors. *Reliab. Eng. Syst. Saf.* 219, 108185. <https://doi.org/10.1016/j.res.2021.108185>.
- Fan, X., Zhang, X., Yu, X.B., 2023. Uncertainty quantification of a deep learning model for failure rate prediction of water distribution networks. *Reliab. Eng. Syst. Saf.* 236, 109088. <https://doi.org/10.1016/j.res.2023.109088>.
- Faramondi, L., Flammini, F., Guarino, S., Setola, R., 2021. A hardware-in-the-loop water distribution testbed dataset for cyber-physical security testing. *IEEE Access.* 9, 122385–122396. <https://doi.org/10.1109/access.2021.3109465>.
- Fu, G., Jin, Y., Sun, S., Yuan, Z., Butler, D., 2022. The role of deep learning in urban water management: a critical review. *Water. Res.* 223, 118973. <https://doi.org/10.1016/j.watres.2022.118973>.
- Garzon, A., Kapelan, Z., Langeveld, J., Taormina, R., 2026. Evaluation of graph neural networks for urban drainage metamodeling: key components and transferability analysis. *Water. Res.* 290, 125079. <https://doi.org/10.1016/j.watres.2025.125079>.
- Géron, A., 2022. *Hands-on Machine Learning with Scikit-Learn, Keras, and TensorFlow*. O'Reilly Media, Inc.
- Guo, G., Liu, S., Wu, Y., Li, J., Zhou, R., Zhu, X., 2018. Short-term water demand forecast based on deep learning method. *J. Water Resour. Plan. Manag.* 144 (12). [https://doi.org/10.1061/\(asce\)wr.1943-5452.0000992](https://doi.org/10.1061/(asce)wr.1943-5452.0000992).
- Haurum, J.B., Moeslund, T.B., 2021. Sewer-ML: a multi-label sewer defect classification dataset and benchmark. In: *Proceedings of the IEEE/cvf conference on computer vision and pattern recognition*.
- He, C., Liu, Z., Wu, J., Pan, X., Fang, Z., Li, J., Bryan, B.A., 2021. Future global urban water scarcity and potential solutions. *Nat. Commun.* 12 (1), 4667. <https://doi.org/10.1038/s41467-021-25026-3>.
- Hilliard, A., Gulley, A., Kazim, E., Koshiyama, A.S., 2026. Artificial intelligence policy worldwide: a comparative analysis. *R. Soc. Open. Sci.* 13 (2). <https://doi.org/10.1098/rsos.242234>.
- Hu, M., Ma, C., Li, W., Xu, W., Wu, J., Hu, J., Li, T., Zhuang, G., Liu, J., Lu, Y., 2025. A survey of scientific large language models: from data foundations to agent frontiers. *arXiv preprint arXiv:2508.21148*.
- Hu, Q., Zhang, Q., Che, D., Wang, F., Zhang, Z., Zhou, J., 2026. Knowledge-data fusion for water supply pipe failure prediction: a hybrid physics-informed and data-driven method. *Reliab. Eng. Syst. Safety* 271, 112263. <https://doi.org/10.1016/j.res.2026.112263>.
- Hu, Y., Scavia, D., Kerkez, B., 2018. Are all data useful? Inferring causality to predict flows across sewer and drainage systems using directed information and boosted regression trees. *Water. Res.* 145, 697–706. <https://doi.org/10.1016/j.watres.2018.09.009>.
- Huang, R., Ma, C., Ma, J., Huangfu, X., He, Q., 2021. Machine learning in natural and engineered water systems. *Water. Res.* 205, 117666. <https://doi.org/10.1016/j.watres.2021.117666>.
- Huang, X., Marques-Silva, J., 2024. On the failings of Shapley values for explainability. *Int. J. Approx. Reason.* 171, 109112. <https://doi.org/10.1016/j.ijar.2023.109112>.
- Hüllermeier, E., Waegeman, W., 2021. Aleatoric and epistemic uncertainty in machine learning: an introduction to concepts and methods. *Mach. Learn.* 110 (3), 457–506. <https://doi.org/10.1007/s10994-021-05946-3>.
- Infant, S.S., Vickram, S., Saravanan, A., Mathan Muthu, C.M., Yuarajan, D., 2025. Explainable artificial intelligence for sustainable urban water systems engineering. *Results Eng.* 25, 104349. <https://doi.org/10.1016/j.rineng.2025.104349>.
- Iwakiri, O., Moazeni, F., 2024. Improving urban water demand and forecast using conformal prediction-based hybrid machine learning models. *J. Water Process Eng.* 58, 104721. <https://doi.org/10.1016/j.jwpe.2023.104721>.
- Joel, L.O., Doorsamy, W., Paul, B.S., 2022. A review of missing data handling techniques for machine learning. *Int. J. Innov. Technol. Interdiscip. Sci.* 5 (3), 971–1005. <https://doi.org/10.1515/IJTIS.2022.5.3.971-1005>.
- Jun, S., Jung, D., 2025. Exploration of deep learning leak detection model across multiple smart water distribution systems: detectable leak sizes with AMI meters. *Water Research X* 29, 100332. <https://doi.org/10.1016/j.wroa.2025.100332>.
- Jun, S., Yoo, D.G., Jung, D., 2025. Toward transparent and reproducible machine learning—Based research in Urban water networks. *J. Water Resour. Plann. Manag.* 151 (10). <https://doi.org/10.1061/jwrmd5.Wreng-6868>.
- Kang, J., Park, Y.-J., Lee, J., Wang, S.-H., Eom, D.-S., 2018. Novel leakage detection by ensemble CNN-SVM and graph-based localization in water distribution systems. *IEEE Trans. Ind. Electron.* 65 (5), 4279–4289. <https://doi.org/10.1109/tie.2017.2764861>.
- Kanyama, M.N., Bhunu Shava, F., Gamundani, A.M., Hartmann, A., 2024. Machine learning applications for anomaly detection in Smart Water Metering Networks: a systematic review. *Phys. Chem. Earth.* 134, 103558. <https://doi.org/10.1016/j.pce.2024.103558>.
- Kapoor, S., Narayanan, A., 2023. Leakage and the reproducibility crisis in machine-learning-based science. *Patterns. (N. Y.)* 4 (9), 100804. <https://doi.org/10.1016/j.patter.2023.100804>.
- Karl, F., Pielok, T., Moosbauer, J., Pfisterer, F., Coors, S., Binder, M., Schneider, L., Thomas, J., Richter, J., Lang, M., Garrido-Merchán, E.C., Branke, J., Bischl, B., 2023. Multi-objective hyperparameter optimization in machine learning—An overview. *ACM T Evol Learn Optim* 3 (4), 1–50. <https://doi.org/10.1145/3610536>.
- Kaufman, S., Rosset, S., Perlich, C., Stitelman, O., 2012. Leakage in data mining. *ACM T Knowl Discov D* 6 (4), 1–21. <https://doi.org/10.1145/2382577.2382579>.
- Kavya, M., Mathew, A., Shekar, P.R., P, S., 2023. Short term water demand forecast modelling using artificial intelligence for smart water management. *Sustain. Cities Soc.* 95, 104610. <https://doi.org/10.1016/j.scs.2023.104610>.
- Kim, T., Kim, K., Hyung, J., Park, H., Oh, Y., Koo, J., 2024. An interpretable machine learning-based pitting corrosion depth prediction model for steel drinking water pipelines. *Process Safety Environ. Protection* 190, 571–585. <https://doi.org/10.1016/j.psep.2024.08.038>.
- Kumar, S., Weesakul, U., Kumar, D.R., Thangavel, P., Wipulanusat, W., Sunkpho, J., 2025. A machine learning approach for corrosion rate modeling in Patna water distribution network of Bihar. *Sci. Rep.* 15 (1), 11678. <https://doi.org/10.1038/s41598-025-96044-0>.
- Li, Z., Ma, W., Zhong, D., Ma, J., Zhang, Q., Yuan, Y., Liu, X., Wang, X., Zou, K., 2024. Applications of machine learning in drinking water quality management: a critical review on water distribution system. *J. Clean. Prod.* 481, 144171. <https://doi.org/10.1016/j.jclepro.2024.144171>.
- 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. *Sci. Total. Environ.* 828, 154284. <https://doi.org/10.1016/j.scitotenv.2022.154284>.
- Li, Z., Zhou, Z., Wang, H., Li, X., Shi, X., Xiao, J., Yang, Z., Sun, M., Li, X., Jia, H., 2025. Artificial intelligence-incorporated prediction for urban flooding processes in the past 20 years: a critical review. *Environ. Model. Softw.* 192, 106525. <https://doi.org/10.1016/j.envsoft.2025.106525>.
- Liang, S., Zhang, Y., Zheng, K., Bai, Y., 2025. FeatureX: an explainable feature selection for deep learning. *Expert Syst. Appl.* 282, 127675. <https://doi.org/10.1016/j.eswa.2025.127675>.
- Liao, T., Taori, R., Raji, L.D., Schmidt, L., 2021. Are we learning yet? A meta review of evaluation failures across machine learning. In: *Thirty-fifth Conference on Neural Information Processing Systems Datasets and Benchmarks Track (Round 2)*.
- Lones, M.A., 2024. Avoiding common machine learning pitfalls. *Patterns*. (N. Y) 5 (10), 101046. <https://doi.org/10.1016/j.patter.2024.101046>.
- Lund, B., Orhan, Z., Mannuru, N.R., Bevara, R.V.K., Porter, B., Vinai, M.K., Bhaskara, P., 2025. Standards, frameworks, and legislation for artificial intelligence (AI) transparency. *AI and Ethics* 5 (4), 3639–3655. <https://doi.org/10.1007/s43681-025-00661-4>.
- Maussner, C., Oberascher, M., Autengruber, A., Kahl, A., Sitzenfrei, R., 2025. Explainable artificial intelligence for reliable water demand forecasting to increase trust in predictions. *Water. Res.* 268 (Pt B), 122779. <https://doi.org/10.1016/j.watres.2024.122779>.
- Menapace, A., Zanfei, A., Righetti, M., 2021. Tuning ANN hyperparameters for forecasting drinking water demand. *Appl. Sci.* 11 (9), 4290. <https://doi.org/10.3390/app11094290>.
- Montes, C., Kapelan, Z., Saldarriaga, J., 2021. Predicting non-deposition sediment transport in sewer pipes using random forest. *Water. Res.* 189, 116639. <https://doi.org/10.1016/j.watres.2020.116639>.
- Naidu, G., Zuva, T., Sibanda, E.M., 2023. A review of evaluation metrics in machine learning algorithms. *Cham. https://doi.org/10.1007/978-3-031-35314-7-2*.
- Namdar, H., Haghighi, A., Ashrafi, S.M., 2025. Short-term urban water demand forecasting; application of 1D convolutional neural network (1D CNN) in comparison with different deep learning schemes. *Stoch. Environ. Res. Risk Assess.* 39, 4211–4226. <https://doi.org/10.1007/s00477-023-02565-3>.
- Nedergaard Pedersen, A., Wied Pedersen, J., Viguera-Rodríguez, A., Brink-Kjær, A., Borup, M., Steen Mikkelsen, P., 2021. The Bellingde data set: open data and models for community-wide urban drainage systems research. *Earth System Sci. Data* 13 (10), 4779–4798. <https://doi.org/10.5194/essd-13-4779-2021>.
- Niknam, A., Zare, H.K., Hosseini-Nasab, H., Mostafaeipour, A., Herrera, M., 2022. A critical review of short-term water demand forecasting tools—What method should I use? *Sustainability* 14 (9), 5412. <https://doi.org/10.3390/su14095412>.
- Oberascher, M., Rauch, W., Sitzenfrei, R., 2022. Towards a smart water city: a comprehensive review of applications, data requirements, and communication technologies for integrated management. *Sustain. Cities Soc.* 76, 103442. <https://doi.org/10.1016/j.scs.2021.103442>.
- Page, M.J., McKenzie, J.E., Bossuyt, P.M., Boutron, I., Hoffmann, T.C., Mulrow, C.D., Shamseer, L., Tetzlaff, J.M., Akl, E.A., Brennan, S.E., Chou, R., Glanville, J., Grimshaw, J.M., Hrobjartsson, A., Lalu, M.A., Li, T., Loder, E.W., Mayo-Wilson, E., McDonald, S., McGuinness, L.A., Stewart, L.A., Thomas, J., Tricco, A.C., Welch, V.A., Whiting, P., Moher, D., 2021. The PRISMA 2020 statement: an updated guideline for reporting systematic reviews. *BMJ* 372, n71. <https://doi.org/10.1136/bmj.n71>.
- Palma, L., Creaco, E., Iervolino, M., Marocco, D., Santonastaso, G.F., Di Nardo, A., 2025. Could ChatGPT automate water network clustering? A performance assessment across algorithms. *Water* 17 (20). <https://doi.org/10.1016/j.watres.2025.17202995>.
- Piadeh, F., Behzadian, K., Chen, A.S., Kapelan, Z., Rizzuto, J.P., Campos, L.C., 2023. Enhancing urban flood forecasting in drainage systems using dynamic ensemble-based data mining. *Water. Res.* 247, 120791. <https://doi.org/10.1016/j.watres.2023.120791>.
- Pinho, J.M.H., de Oliveira, S.V.B., Silva, T.H.S., Saraiva, P.A.R., de Souza, E.F., Godoy, R.V., Ambrosio, L.A., Becker, M., 2025. The impact of feature scaling in machine learning: effects on regression and classification tasks. *arXiv preprint. https://doi.org/10.48550/arXiv.2506.08274*.
- Plevris, V., Solorzano, G., Bakas, N., Ben Seghier, M., 2022. Investigation of performance metrics in regression analysis and machine learning-based prediction models. In: *8th European Congress on Computational Methods in Applied Sciences and Engineering*. <https://doi.org/10.23967/ecomas.2022.155>.
- Pudjihartono, N., Fadason, T., Kempa-Liehr, A.W., O'Sullivan, J.M., 2022. A review of feature selection methods for machine learning-based disease risk prediction. *Front Bioinform.* 2, 927312. <https://doi.org/10.3389/fbinf.2022.927312>.
- Rapp, A.H., Capener, A.M., Sowby, R.B., 2023. Adoption of artificial intelligence in drinking water operations: a survey of progress in the United States. *J. Water Resour. Plan. Manag.* 149 (7). <https://doi.org/10.1061/jwrmd5.Wreg-5870>.
- Raschka, S., Patterson, J., Nolet, C., 2020. Machine learning in Python: main developments and technology trends in data science, Machine learning, and artificial intelligence. *Information* 11 (4), 193. <https://doi.org/10.3390/info11040193>.
- Ray, P.P., 2023. Leveraging deep learning and language models in revolutionizing water resource management, research, and policy making: a case for ChatGPT. *ACS ES&T Water* 3 (8), 1984–1986. <https://doi.org/10.1021/acsestwater.3c00264>.
- Reshadi, M.A.M., Rezanezhad, F., Shahvaran, A.R., Ghajari, A., Kaykhosravi, S., Slowski, S., Van Cappellen, P., 2025. Assessment of environmental and socioeconomic drivers of urban stormwater microplastics using machine learning. *Sci. Rep.* 15 (1), 6299. <https://doi.org/10.1038/s41598-025-90612-0>.
- Richards, C.E., Tzachor, A., Avin, S., Fenner, R., 2023. Rewards, risks and responsible deployment of artificial intelligence in water systems. *Nat. Water* 1 (5), 422–432. <https://doi.org/10.1038/s44221-023-00069-6>.
- Riyadh, A., Peleato, N.M., 2025. Exploring spatial and temporal importance of input features and the explainability of machine learning-based modelling of water distribution systems. *Digital Chem. Eng.* 14, 100202. <https://doi.org/10.1016/j.dche.2024.100202>.
- Roberts, M., Driggs, D., Thorpe, M., Gilbey, J., Yeung, M., Ursprung, S., Aviles-Rivero, A. I., Etman, C., McCague, C., Beer, L., Weir-McCall, J.R., Teng, Z., Gkrania-Klotsas, E., Ruggiero, A., Korhonen, A., Jefferson, E., Ako, E., Langs, G., Gozaliasl, G., Yang, G., Prosch, H., Preller, J., Stanczuk, J., Tang, J., Hofmanninger, J., Babar, J., Sánchez, L. E., Thillai, M., Gonzalez, P.M., Teare, P., Zhu, X., Patel, M., Cafolla, C., Azadbakht, H., Jacob, J., Lowe, J., Zhang, K., Bradley, K., Wassin, M., Holzer, M., Ji, K., Ortet, M.D., Ai, T., Walton, N., Lio, P., Stranks, S., Shadbahr, T., Lin, W., Zha, Y., Niu, Z., Rudd, J.H.F., Sala, E., Schönlieb, C.-B., 2021. Common pitfalls and recommendations for using machine learning to detect and prognosticate for COVID-19 using chest radiographs and CT scans. *Nat. Mach. Intell.* 3 (3), 199–217. <https://doi.org/10.1038/s42256-021-00307-0>.
- Robles-Velasco, A., Cortés, P., Muñozuri, J., Onieva, L., 2020. Prediction of pipe failures in water supply networks using logistic regression and support vector classification. *Reliab. Eng. Syst. Saf.* 196, 106754. <https://doi.org/10.1016/j.res.2019.106754>.
- Safari, M.J.S., 2019. Decision tree (DT), generalized regression neural network (GR) and multivariate adaptive regression splines (MARS) models for sediment transport in sewer pipes. *Water. Sci. Technol.* 79 (6), 1113–1122. <https://doi.org/10.2166/wst.2019.106>.
- Salehin, I., Islam, M.S., Saha, P., Noman, S.M., Tun, A., Hasan, M.M., Baten, M.A., 2024. AutoML: a systematic review on automated machine learning with neural architecture search. *J. Inf. Intell.* 2 (1), 52–81. <https://doi.org/10.1016/j.jiixd.2023.10.002>.
- Sela, L., Sowby, R.B., Salomons, E., Housh, M., 2025. Making waves: the potential of generative AI in water utility operations. *Water. Res.* 272, 122935. <https://doi.org/10.1016/j.watres.2024.122935>.
- Smits, A., 2020. A critical overview of outlier detection methods. *Comput. Sci. Rev.* 38, 100306. <https://doi.org/10.1016/j.cosrev.2020.100306>.
- Snider, B., McBean, E.A., 2020. Improving urban water security through pipe-break prediction models: machine learning or survival analysis. *J. Environ. Eng.* 146 (3). [https://doi.org/10.1061/\(asce\)ee.1943-7870.0001657](https://doi.org/10.1061/(asce)ee.1943-7870.0001657).
- Taiwo, R., Zayed, T., Ben Seghier, M.E.A., 2024. Integrated intelligent models for predicting water pipe failure probability. *Alex. Eng. J.* 86, 243–257. <https://doi.org/10.1016/j.aej.2023.11.047>.
- Taormina, R., Galelli, S., Tippenhauer, N.O., Salomons, E., Ostfeld, A., Eliades, D.G., Aghashahi, M., Sundararajan, R., Pourahmadi, M., Banks, M.K., Brentan, B.M., Campbell, E., Lima, G., Manzi, D., Ayala-Cabrera, D., Herrera, M., Montalvo, I., Izquierdo, J., Luvizotto, E., Chandy, S.E., Rasekh, A., Barker, Z.A., Campbell, B., Shafiee, M.E., Giacomoni, M., Gatsis, N., Taha, A., Abokifa, A.A., Haddad, K., Lo, C. S., Biswas, P., Pasha, M.F.K., Kc, B., Somasundaram, S.L., Housh, M., Ohar, Z., 2018. Battle of the attack detection algorithms: disclosing cyber attacks on water distribution networks. *J. Water Resour. Plan. Manag.* 144 (8), 04018048. [https://doi.org/10.1061/\(ASCE\)WR.1943-5452.0000969](https://doi.org/10.1061/(ASCE)WR.1943-5452.0000969).
- Thelwall, M., Lehtisaari, M., Katsireia, I., Holmberg, K., Zheng, E.T., 2025. Does ChatGPT ignore article retractions and other reliability concerns? *Learned Publishing* 38 (4). <https://doi.org/10.1002/leap.2018>.
- Tian, J., Chen, Y., Yang, L., Li, D., Liu, L., Li, J., Tang, X., 2025. Enhancing urban flood susceptibility assessment by capturing the features of the urban environment. *Remote Sens. (Basel)* 17 (8), 1347. <https://doi.org/10.3390/rs17081347>.
- Tijani, I.A., Abdelmageed, S., Fares, A., Fan, K.H., Hu, Z.Y., Zayed, T., 2022. Improving the leak detection efficiency in water distribution networks using noise loggers. *Sci. Total. Environ.* 821, 153530. <https://doi.org/10.1016/j.scitotenv.2022.153530>.
- Vabalas, A., Gowen, E., Poliakov, E., Casson, A.J., 2019. Machine learning algorithm validation with a limited sample size. *PLoS One* 14 (11), e0224365. <https://doi.org/10.1371/journal.pone.0224365>.
- Vrachimis, S., Eliades, D., Taormina, R., Kapelan, Z., Ostfeld, A., Liu, S., Kyriakou, M., Pavlou, P., Qiu, M., Polycarpou, M., 2022. Battle of the leakage detection and isolation methods. *J. Water Resour. Plan. Manag.* 148, 04022068. [https://doi.org/10.1061/\(ASCE\)WR.1943-5452.0001601](https://doi.org/10.1061/(ASCE)WR.1943-5452.0001601).
- Wang, J., Fu, G., Savic, D., 2026. Leveraging large language models for automating water distribution network optimization. *Water. Res.* 288 (Pt A), 124536. <https://doi.org/10.1016/j.watres.2025.124536>.
- Wang, L., Han, M., Li, X., Zhang, N., Cheng, H., 2021. Review of classification methods on unbalanced data sets. *IEEE Access*. 9, 64606–64628. <https://doi.org/10.1109/access.2021.3074243>.
- Williams, G.P., 2025. Friends don't let friends use Nash-Sutcliffe Efficiency (NSE) or KGE for hydrologic model accuracy evaluation: a rant with data and suggestions for better practice. *Environ. Model. Softw.* 194, 106665. <https://doi.org/10.1016/j.envsoft.2025.106665>.
- Xing, L., Sela, L., 2022. Graph Neural networks for State estimation in Water Distribution systems: application of supervised and semisupervised learning. *Journal of Water Resources Planning and Management* 148. [https://doi.org/10.1061/\(ASCE\)WR.1943-5452.0001550](https://doi.org/10.1061/(ASCE)WR.1943-5452.0001550).
- Xu, Z., Liu, H., Fu, G., Zheng, R., Zayed, T., Liu, S., 2025. Interpretable deep learning for acoustic leak detection in water distribution systems. *Water. Res.* 273, 123076. <https://doi.org/10.1016/j.watres.2024.123076>.
- Yang, Y., Chui, T.F.M., 2021. Modeling and interpreting hydrological responses of sustainable urban drainage systems with explainable machine learning methods. *Hydrol. Earth. Syst. Sc.* 25 (11), 5839–5858. <https://doi.org/10.5194/hess-25-5839-2021>.
- Yazdanfar, Z., Sharma, A., 2015. Urban drainage system planning and design - challenges with climate change and urbanization: a review. *Water. Sci. Technol.* 72 (2), 165–179. <https://doi.org/10.2166/wst.2015.207>.

- Yu, T., Zhu, H., 2020. Hyper-parameter optimization: a review of algorithms and applications. arXiv preprint arXiv:2003.05689. <https://doi.org/10.48550/arXiv.2003.05689>.
- Yu, Y., Hossain, M.M., Sikder, R., Qi, Z., Huo, L., Chen, R., Dou, W., Shi, B., Ye, T., 2024. Exploring the potential of machine learning to understand the occurrence and health risks of haloacetic acids in a drinking water distribution system. *Sci. Total. Environ.* 951, 175573. <https://doi.org/10.1016/j.scitotenv.2024.175573>.
- Yuan, Z., Olsson, G., Cardell-Oliver, R., van Schagen, K., Marchi, A., Deletic, A., Urich, C., Rauch, W., Liu, Y., Jiang, G., 2019. Sweating the assets - the role of instrumentation, control and automation in urban water systems. *Water. Res.* 155, 381–402. <https://doi.org/10.1016/j.watres.2019.02.034>.
- Zanfei, A., Menapace, A., Brentan, B.M., Righetti, M., 2022. How does missing data imputation affect the forecasting of urban water demand? *J. Water Resour. Plan. Manag.* 148 (11). [https://doi.org/10.1061/\(asce\)wr.1943-5452.0001624](https://doi.org/10.1061/(asce)wr.1943-5452.0001624).
- Zhang, H., Li, Y., Li, M., Wei, S., Zhang, H., 2026. Explainable and causal machine learning to investigate the spatiotemporal dynamics patterns of coastal water quality in Hong Kong. *Water. Res.* 290, 125026. <https://doi.org/10.1016/j.watres.2025.125026>.
- Zhang, Y., Li, C., Jiang, Y., Sun, L., Zhao, R., Yan, K., Wang, W., 2022. Accurate prediction of water quality in urban drainage network with integrated EMD-LSTM model. *J. Clean. Prod.* 354, 131724. <https://doi.org/10.1016/j.jclepro.2022.131724>.
- Zhou, S., Liu, Z., Wang, M., Gan, W., Zhao, Z., Wu, Z., 2022. Impacts of building configurations on urban stormwater management at a block scale using. XGBoost. *Sustain. Cities Soc.* 87, 104235. <https://doi.org/10.1016/j.scs.2022.104235>.
- Zhou, S., Zhang, D., Wang, M., Liu, Z., Gan, W., Zhao, Z., Xue, S., Müller, B., Zhou, M., Ni, X., Wu, Z., 2024. Risk-driven composition decoupling analysis for urban flooding prediction in high-density urban areas using Bayesian-Optimized LightGBM. *J. Clean. Prod.* 457, 142286. <https://doi.org/10.1016/j.jclepro.2024.142286>.
- Zhu, J.J., Yang, M., Ren, Z.J., 2023. Machine learning in environmental research: common pitfalls and Best practices. *Environ. Sci. Technol.* 57 (46), 17671–17689. <https://doi.org/10.1021/acs.est.3c00026>.
- 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>.
- Zounemat-Kermani, M., Stephan, D., Barjenbruch, M., Hinkelmann, R., 2020. Ensemble data mining modeling in corrosion of concrete sewer: a comparative study of network-based (MLPNN & RBFNN) and tree-based (RF, CHAID, & CART) models. *Adv. Eng. Inform* 43, 101030. <https://doi.org/10.1016/j.aei.2019.101030>.