Article:

Habib, M.A., Jurcut, A.D., Ahmed, H. et al. (2026) Cybersecurity in water distribution networks: a systematic review of AI-based detection algorithms. Water, 18 (4). 519. ISSN: 2073-4441

https://doi.org/10.3390/w18040519

water

MDPI

Systematic Review

Cybersecurity in Water Distribution Networks: A Systematic Review of AI-Based Detection Algorithms

Md Arman Habib [1,*], Anca Delia Jurcut [2], Hafiz Ahmed [3,4], Wenhui Wei [1] and Md Salauddin [1]

[1] UCD Dooge Centre for Water Resources Research, UCD School of Civil Engineering, UCD Earth Institute, University College Dublin, 4 Dublin, Ireland; wenhui.wei@ucdconnect.ie (W.W.); md.salauddin@ucd.ie (M.S.)
[2] UCD School of Computer Science, University College Dublin, 4 Dublin, Ireland; anca.jurcut@ucd.ie
[3] School of Electrical & Electronic Engineering, University of Sheffield, Sheffield S1 3JD, UK; hafiz.h.ahmed@ieee.org
[4] Autonex Systems Limited, London WC2H 9JQ, UK
[*] Correspondence: md.habib@ucd.ie; Tel.: +353-17163201

Abstract

Water Distribution Networks (WDNs) are critical infrastructure for delivering clean and safe drinking water. As modern WDNs increasingly integrate cyber technologies, they evolve into complex cyber–physical systems (CPSs). This connectivity, however, introduces new vulnerabilities, including cyberattacks. Cybersecurity protects systems from unauthorized access, attacks, and data breaches. In this systematic review, we adopted the PRISMA 2020 reporting guideline. Predefined keyword strings were designed to extract relevant articles from Scopus and Web of Science during the period of 2014–2025. In total, 32 peer-reviewed studies were included for narrative synthesis following duplication and eligibility screening. The review protocol was not registered. This review provides a unified perspective on how Artificial Intelligence (AI) contributes to WDNs resilience. The literature is evaluated in terms of detection tasks, data modalities, learning paradigms, and model architecture. The results highlight three key findings: (a) data bias, reflected in significant reliance on specific synthetic datasets and limited use of real-world utility network data; (b) performance, with deep learning architecture, such as long-short-term memory models, achieving commendable levels of accuracy in intrusion detection, however, overall comparison with other models remain scenario-dependent; and (c) future directions, synthesized through an AI-centered perspective that emphasizes resilience and identifies research gaps in adaptive online learning, attack prediction, interpretability, federated learning and topology localization. This study concludes with recommendations for the broader integration of AI tools to support resilient WDN operation.

Keywords: cybersecurity; machine learning; resilience; water distribution systems

Check for updates

Academic Editors: Katarzyna Pietrucha-Urbanik and Janusz Rak

Received: 12 January 2026
Revised: 12 February 2026
Accepted: 18 February 2026
Published: 22 February 2026

1. Introduction

According to the European Program for Critical Infrastructure Protection (EPCIP), Water Distribution Networks (WDNs), which ensure the supply of clean and safe water to households, are defined as critical infrastructures. Any disruption of their services could, therefore, have a serious impact on national security, economic well-being, public health or safety, or any combination thereof [1]. Modern WDNs are increasingly integrating advanced technologies and remote monitoring facilities that rely on cyberspace. While this integration improves operational efficiency, it also exposes WDNs to vulnerabilities arising from cyberattacks [2]. Cyberattacks on WDNs can disrupt the supply of clean and safe water,

https://doi.org/10.3390/w18040519

potentially leading to public health crises and affecting overall community well-being [3]. Even a minor breach can propagate consequences across the entire system. The integration of physical and cyber–physical systems (CPSs) has enhanced WDN operations; however, it also raises concerns regarding increased exposure to organized cyberattacks [4]. For example, a *Stuxnet-type* attack can potentially jeopardize critical infrastructure by manipulating operational technologies, resulting in severe physical consequences [5]. Moreover, the frequency and complexity of cyberattacks are expected to continue increasing in the coming decades [6,7]. A recent large scale cyberattacks in WDNs occurred in October 2024, where 14 million people across 14 states and 18 military installations were affected. The utility company detected unauthorized and malicious activity in the online network that resulted in the shutdown of key network components.

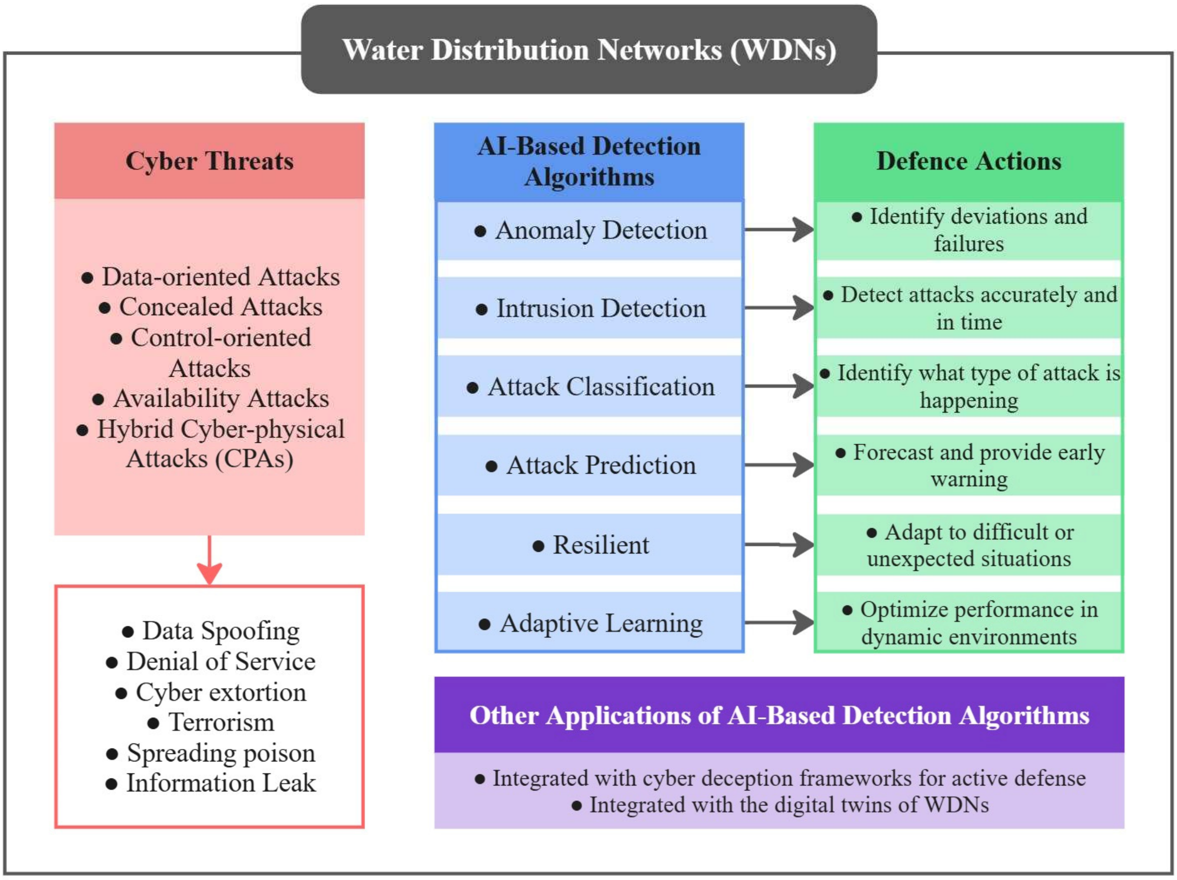
Traditional cybersecurity approaches rely on static rules and heuristics, such as Intrusion Detection Systems (IDS). These methods are limited in their ability to capture the dynamic nature of cyberattacks. To address this limitation, Artificial Intelligence (AI), particularly supervised and unsupervised Machine Learning (ML) algorithms, has been increasingly employed to detect sophisticated cyberattacks with high accuracy and efficiency [8]. Traditional ML algorithms can identify threats and intrusions simultaneously and are capable of handling large volumes of cyberattacks efficiently [9]. More recently, deep learning and reinforcement learning have further contributed to the adaptive capability of security systems, improving resilience against the evolving cyber threat landscape [10]. Supervisory Control and Data Acquisition (SCADA) systems and Programmable Logic Controllers (PLCs) are essential components of WDNs. Together with numerous sensors and actuators, they enable digital control and monitoring of WDNs. However, several vulnerabilities in current and future SCADA systems have been identified, including the lack of built-in security mechanisms, open access networks, assumptions of closed environments, and reliance on legacy systems [11]. Cyberattacks targeting SCADA systems and PLCs may involve false data injection or contamination attacks, potentially compromising water quality and availability. Recent research has showcased the effectiveness and potential of AI-based detection algorithms in WDNs, highlighting their growing importance, as illustrated in Figure 1. Despite these advances, significant challenges continue to emerge alongside the evolution of WDNs.

The current literature includes several review studies addressing the application of ML algorithms for anomaly detection [12,13] and cyberattacks in SCADA systems [14]. More recently, the study of Kanyama et al. [15] reviewed ML-based anomaly detection in smart water metering networks. A review focusing on theoretical frameworks for cybersecurity research aimed at improving water distribution and wastewater collection systems is presented in Tuptuk et al. [16]. Existing reviews have predominantly focused on anomaly detection in water systems and ML-based approaches for cyberattacks on SCADA environments and WDNs [12–16], with an emphasis on algorithmic applications. In contrast, the present study critically analyzes selected articles by mapping the cyber–physical chain to resilience features. A detailed algorithm taxonomy is compiled to highlight the specific categories of algorithms used for cybersecurity tasks. Each article is reviewed to determine whether it addresses only cyber–physical processes, such as intrusion or anomaly detection, or also considers associated impacts and response or recovery strategies. This approach aims to provide a holistic view of cyber–physical security in WDNs. Additionally, a data-centric meta-analysis is conducted to categorize the literature by detection type, data modality, learning paradigm, and model architecture. This review aims to answer the following Research Questions (RQs):

https://doi.org/10.3390/w18040519

RQ1: What are the main task types covered by ML applications?
RQ2: What evidence exists regarding the real-world deployment of the developed algorithms and cross-dataset generalization?
RQ3: To what extent do existing studies link the ML algorithms to resilience outcomes and mitigation enhancement?

Finally, to the best of the authors' knowledge, this review is the first to present a resilience-aware AI perspective that integrates detection performance with potential cyberattack impacts and emphasizes model interpretability and federated learning to support broader adoption of ML-based cybersecurity solutions in WDNs.


Water Distribution Networks (WDNs)

Cyber Threats
- Data-oriented Attacks
- Concealed Attacks
- Control-oriented Attacks
- Availability Attacks
- Hybrid Cyber-physical Attacks (CPAs)

- Data Spoofing
- Denial of Service
- Cyber extortion
- Terrorism
- Spreading poison
- Information Leak

AI-Based Detection Algorithms
- Anomaly Detection
- Intrusion Detection
- Attack Classification
- Attack Prediction
- Resilient
- Adaptive Learning

Defence Actions
- Identify deviations and failures
- Detect attacks accurately and in time
- Identify what type of attack is happening
- Forecast and provide early warning
- Adapt to difficult or unexpected situations
- Optimize performance in dynamic environments

Other Applications of AI-Based Detection Algorithms
- Integrated with cyber deception frameworks for active defense
- Integrated with the digital twins of WDNs

Figure 1. The application of AI-based detection algorithms in cybersecurity for water distribution networks.

A comparative discussion of this review work with the existing reviews on relevant topics is tabulated in Table 1.

Table 1. Comparison of the proposed study with contemporary review works on similar topics.

References	Review Topic	Review Focus
[16]	Cybersecurity in water infrastructure (WDNs and Waste Water Network)	Discusses cybersecurity threats; proposes research themes. Highlights skill gaps and recommends set of priorities. No specific focus on ML applications and ML taxonomy.
[12]	PRISMA Systematic Review of ML applications for cyberattack detection in SCADA–WDNs	Algorithm taxonomy and database dominance discussed. Resilience aspects and mitigation strategies not highlighted. Focuses only on intrusion detection.

Table 1. *Cont.*

References	Review Topic	Review Focus
[15]	ML-based anomaly detection in Smart Metering Water Networks	Synthesizes applications of ML frameworks for anomalies (leaks/usage) in SMWNs; little or no emphasis on adversarial/cyberattack context.
[17]	ML applications for intrusion detection in WDNs and treatment networks	Reviews cross-layer security layers and links attacks to operational impacts. Focuses on creating mixed datasets based on review findings.
[18]	AI/ML frameworks for water cybersecurity	Conceptual discussion on ML frameworks; does not discuss technical applications of ML algorithms.
[19]	ML Applications for SCADA Intrusion Detection Systems (IDSs)	Reviews supervised ML-based IDS methods only; not specific to WDNs.
Proposed paper	ML Applications for cybersecurity in WDNs from 3 lenses (intrusion + anomaly + resilience)	Emphasizes algorithm taxonomy, synthesizes three different streams of ML applications and recommends an inclusive framework of ML applications incorporating resilience and mitigation strategies.

2. Materials and Methods

The articles selected for this review were identified based on several criteria, including publication type, publication period (2014–2025), and peer-review status. In addition, the following categories of articles were excluded from this review:

- Book chapters.
- Non-peer-reviewed conference papers.
- Studies with restricted data sources or inaccessible full text.

A set of five keywords was used to retrieve articles from major databases, including Web of Science and Scopus. The retrieved studies were screened for eligibility using the PRISMA methodology [20] (see Figure 2), resulting in the selection of 32 articles for the final review. The keywords used to identify the relevant literature are listed below:

- Machine Learning AND Cyberattack detection AND Water Distribution Networks
- Anomaly Detection AND Water Distribution Networks AND Machine Learning
- Artificial Intelligence AND Cybersecurity AND Water Infrastructure
- Supervised AND Unsupervised Learning AND SCADA AND Water Networks

Each of the 32 selected publications was then categorized to enable a systematic comparison of current research developments. The categorization was based on one or more task types addressed by the ML algorithms (intrusion detection, operational anomaly detection, or resilience-oriented approaches), the primary data modality (e.g., hydraulic, water quality, cyber network traffic, or mixed cyber–physical features), the learning paradigm (supervised, unsupervised, or semi-supervised; offline or online), and the model architecture family (e.g., tree-based, kernel-based, density-based, autoencoder, recurrent, or graph-based models). Furthermore, the classification distinguished between studies using simulated datasets (such as BATADAL/CTown, SWaT, WADI, and EPANET) and those relying on utility-based data. The following Supporting Information can be downloaded at: https://www.mdpi.com/article/10.3390/w18040519/s1: PRISMA 2020 checklist and a PRISMA 2020 checklist has been provided as a Supplementary Material.

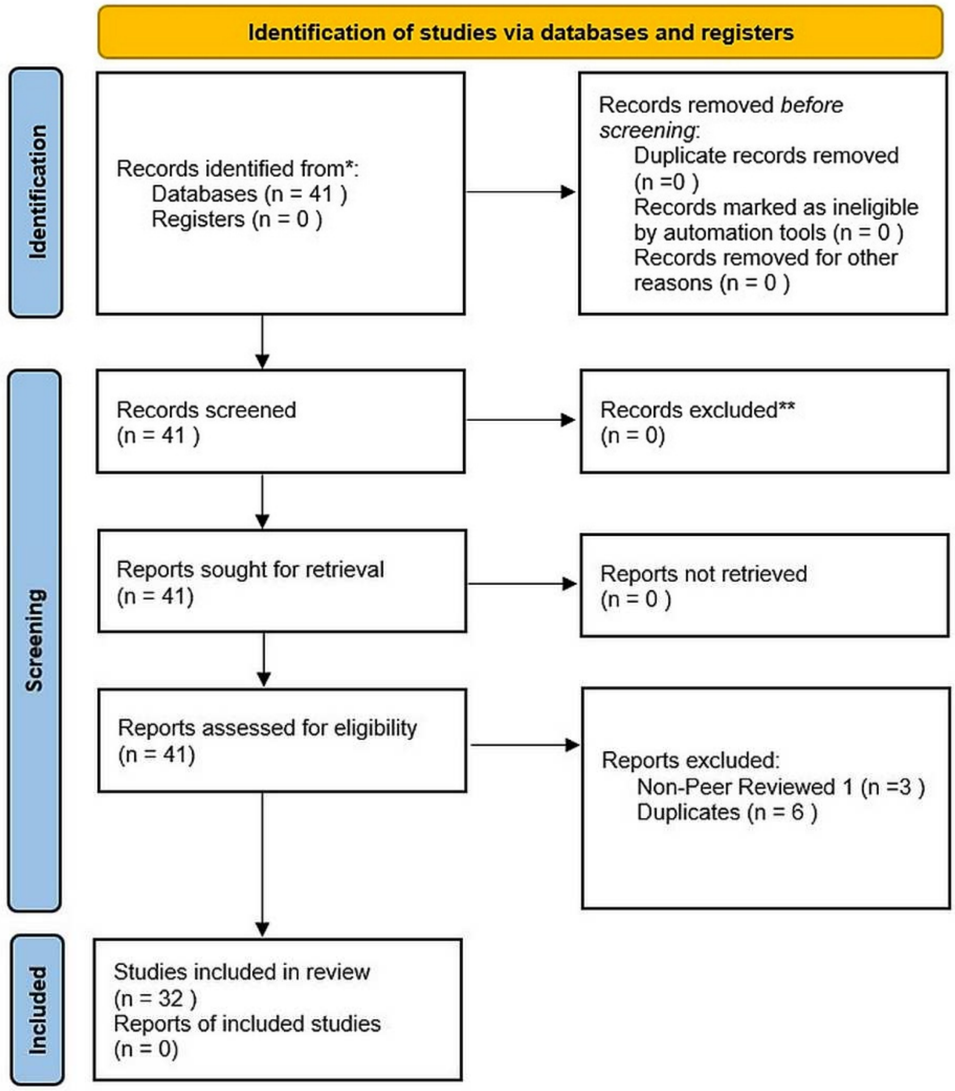

Identification of studies via databases and registers
Identification
Screening
Included
Records identified from*:
Databases (n = 41)
Registers (n = 0)
Records removed *before screening*:
Duplicate records removed (n =0)
Records marked as ineligible by automation tools (n = 0)
Records removed for other reasons (n = 0)
Records screened (n = 41)
Records excluded** (n = 0)
Reports sought for retrieval (n = 41)
Reports not retrieved (n = 0)
Reports assessed for eligibility (n = 41)
Reports excluded:
Non-Peer Reviewed 1 (n =3)
Duplicates (n = 6)
Studies included in review (n = 32)
Reports of included studies (n = 0)

Figure 2. Screening approach followed within this study for the selection from the literature (PRISMA diagram). * Databases used: Scopus, Web of Science. *Google* Scholar, ** Automation Tools used: Python-based *pyscholar*.

The PRISMA [20] guideline was adopted to conduct and report the findings from the systematic review. The rationale and objectives are stated in the Introduction section, and the eligibility criteria, information sources, search strategy, study selection, data extraction, data items, quality appraisal approach, and synthesis methods are reported in the Methodology section. The PRISMA flow diagram (Figure 2) shows the selection results of the relevant articles. The characteristics of the selected studies and findings are summarized in the Results and Discussion section through tables/figures. Additionally, the interpretation, limitations, and implications are also discussed. The review protocol was not registered, and statements on funding/support, conflicts of interest, and data/materials availability are provided at the end of this review paper to complete the PRISMA 2020 reporting requirements.

3. Results and Discussion

3.1. Search Results

The publication trend in Figure 3 indicates a noticeable increase in studies focusing on the application of ML algorithms to cybersecurity after 2016. This growth can largely be attributed to the increased availability of open-access datasets and an increasing awareness of cybersecurity challenges in water infrastructure. Prior to 2016, only one of the 32 selected studies addressed this topic. The observed trend suggests that a substantial and growing body of research is now dedicated to enhancing cybersecurity measures in WDNs using ML-based approaches.

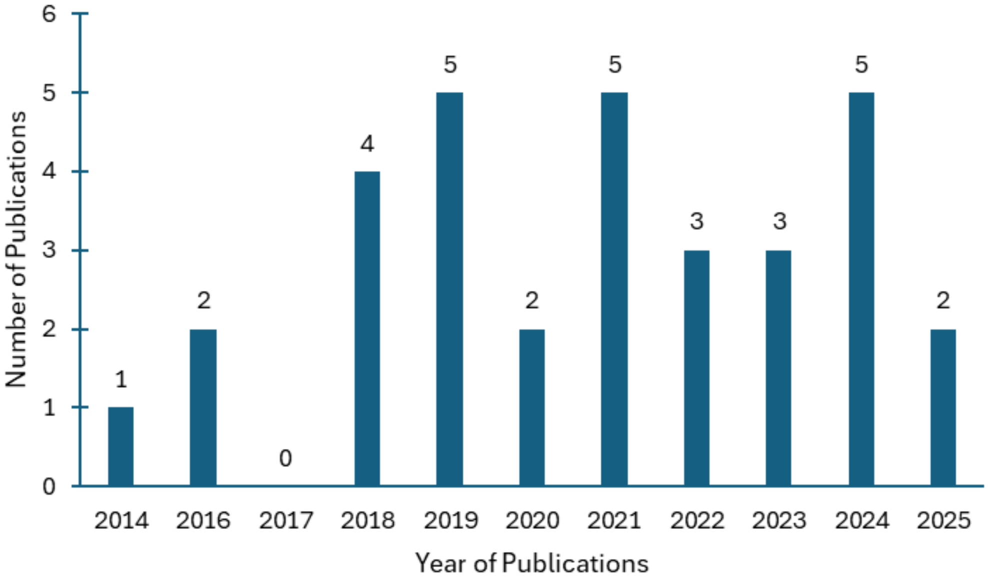


Figure 3. Publication trend of articles related to cybersecurity enhancement using AI-based algorithms.

Among the 32 selected articles, 22 are investigative studies employing defined datasets or experimental setups, while 10 articles provide reviews of the state of the art. A large proportion of first-author affiliations originates from the USA, the UK, and South Africa, (see Figure 4) with a smaller number of studies from other regions. This distribution indicates that research on AI-based cybersecurity enhancement in WDNs remains geographically uneven.

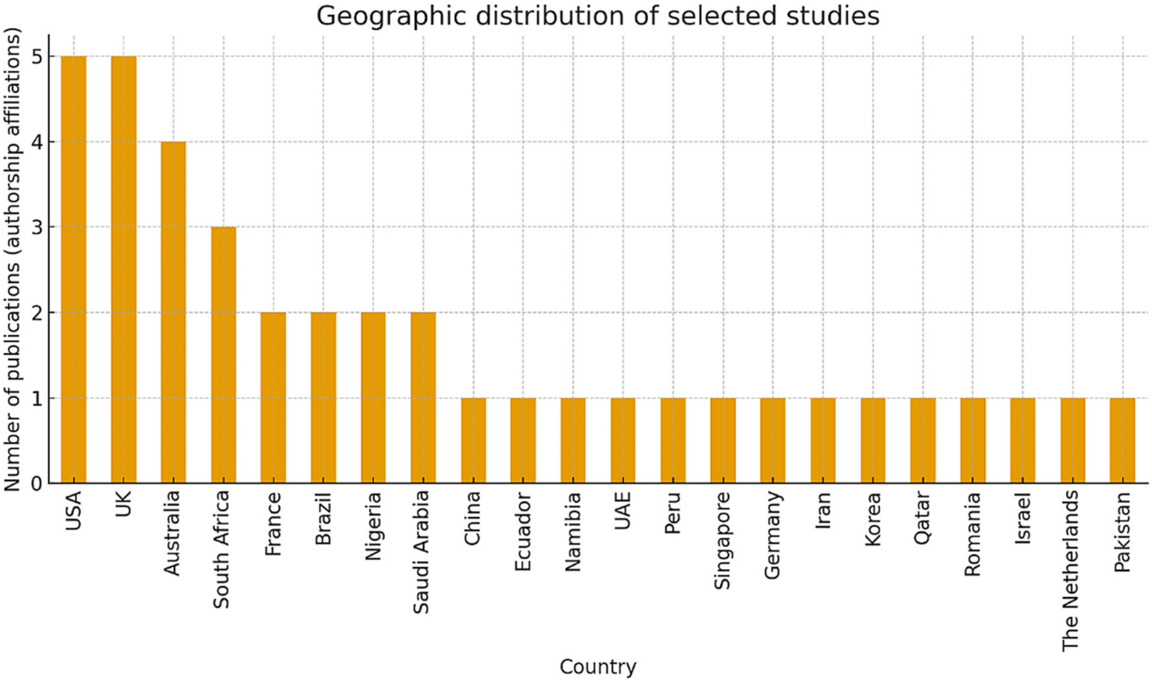


Figure 4. Origin countries of first-author affiliations of the selected studies.

https://doi.org/10.3390/w18040519

A bias towards the evidence base may have been induced by the evidence of first-author affiliations from a small number of high-income countries. The bias may be towards (i) utilities with higher instrumentation density, (ii) SCADA governance, and (iii) labeled historical events. Consequently, this will lead to the diminished transferability of performance metrics and deployment feasibility towards low- and middle-income countries where operational practices and data collection/recording are different from high-income nations. This review therefore recommends considering the results reported in this review according to the context of individual studies.

3.2. Algorithm Taxonomy

A considerable diversity in task type, data modality, and ML paradigm is evident across the reviewed literature. To provide a structured overview, this subsection summarizes the main ML components employed. The algorithms are broadly classified into two groups: classical ML models, including Random Forest, Support Vector Machines (SVM), k-Nearest Neighbors, and Extreme Learning Machines; and deep or hybrid architectures, such as Autoencoders, Convolutional and Recurrent Neural Networks (CNNs and RNNs), Temporal Graph Convolutional Networks (GNs), and hybrid encoder–decoder models. Figure 5 illustrates the algorithm families used in the selected studies.

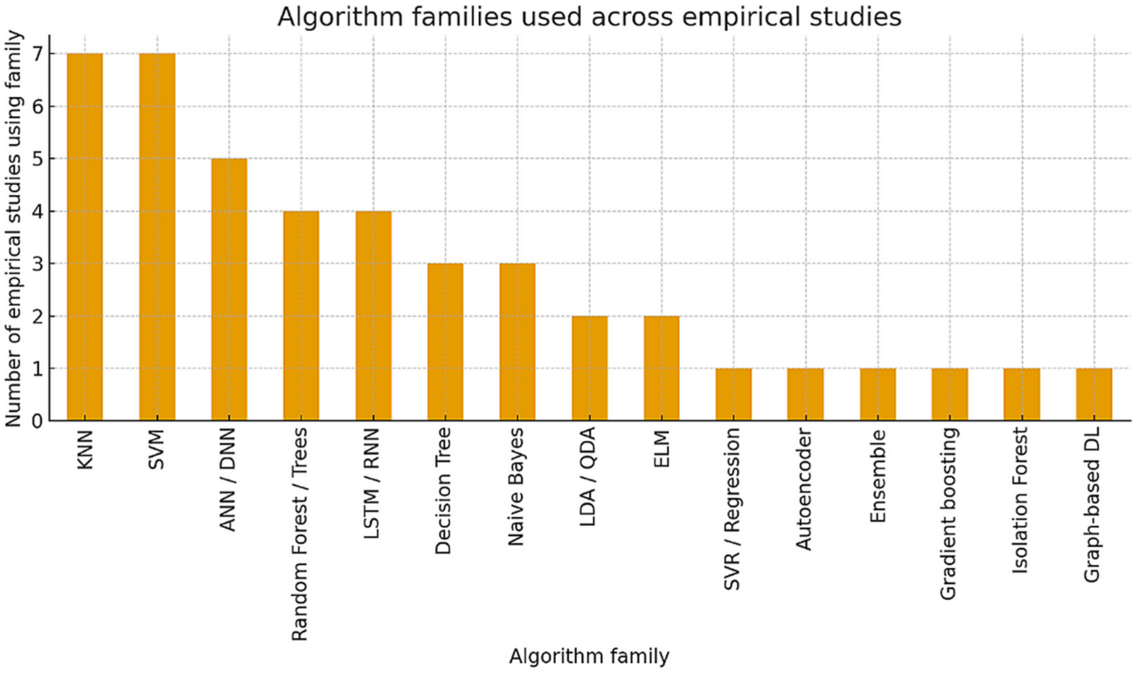


Figure 5. Taxonomy of the algorithms applied in cybersecurity enhancement in WDNs.

3.3. Databases

A strong reliance on the C-Town network from the BATtle of the Attack Detection Algorithms (BATADAL) dataset is observed across the reviewed studies (see Table 2), with approximately one-third of the articles using this dataset. SCADA testbeds also represent a frequently adopted data source. A limited number of studies utilized data from public water utilities, including network topology and operational data, as well as hydraulic simulation data generated using tools such as EPANET [21]. Synthetic network data generation was reported in Abughali et al. [22]. Several studies employed the SWaT and WADI testbeds, which capture water treatment and distribution dynamics [23,24] with applications reported in Abughali et al. [22]. In Teixeira et al. [25], a new dataset was developed by extracting features using Argus and Wireshark from attacks executed on a SCADA system testbed. Private datasets were used in a small number of cases [26,27], and

https://doi.org/10.3390/w18040519

combined data sources were also reported, for example, in Govea et al. [26]. Given the evolving nature of cyberattacks, several studies emphasized the need for periodic updates of benchmark datasets to support the development and training of ML-based cybersecurity systems [25].

Table 2. Datasets adopted across the selected studies.

Dataset Category (Examples)	Tier	Frequency	Topology Availability	Ground-Truth Maturity	Service-Impact Evaluation Feasibility
BATADAL/CTown	T1	8	Yes (simulated)	Attack labels typically available	Yes (via simulation replay)
Virtual SCADA (EPANET)	T1	1	Yes (model-based)	Scenario-defined	Yes (via hydraulics/WQ simulation)
Simulated WDNs model	T1	1	Yes (model-based)	Scenario-defined	Yes (via hydraulics/WQ simulation)
Synthetic MATLAB-generated	T1	1	Yes (model-based)	Scenario-defined	Yes (via simulation replay)
SCADA testbed (incl. SWaT/WADI-type)	T2	4	Often partial/small-scale	Attack labels available; limited diversity	Partly (depends on testbed instrumentation)
Generic IT/ICS IDS dataset (non-water)	T2	1	No (not WDN)	Attack labels available	No (no water-service impact)
Real utility SCADA	T3	3	Variable (often restricted)	Often weak/partial event logs	Difficult (needs counterfactual + ops logs)
Public open utility data	T3	2	Variable	Typically anomaly-/event label-limited	Difficult (context-dependent)
Smart-meter WDNs	T3	1	Implicit (DMA/zone)	Leak/anomaly labels variable	Possible for demand loss proxies
Water-quality dataset (tier not specified)	T2–T3	1	Variable	Variable	Variable

Three dataset-realism tiers (Tier-1 synthetic/simulated, Tier-2 SCADA/ICS testbeds, Tier-3 real-utility datasets) were named to stratify the evidence gathered to support a sensitivity synthesis of findings. Table 2 summarizes tier-wise evaluation feasibility according to the topology/label availability. Given the dominance of Tier-1/Tier-2 benchmarks, the reported detection performances should be interpreted tier-wise due to the dominance of Tier-1/Tier-2 datasets, and therefore, offer limited value to Tier-3 datasets. Operational and service-impact metrics are rarely reported, particularly for Tier-3 evidence, constraining generalizability.

The distribution of datasets across the reviewed studies reveals a strong bias toward synthetic and testbed-based benchmarks, with BATADAL and SCADA datasets collectively used in nearly 54% of the studies. In contrast, datasets derived from real utility networks or smart metering systems remain significantly underrepresented. Moreover, cross-validation across heterogeneous datasets is rarely reported in the current literature.

3.4. Accuracy Metrics

The performance of ML algorithms is commonly evaluated using metrics derived from the confusion matrix, including accuracy, precision, recall, and F1-score. Among the reviewed studies, deep learning (DL) algorithms were consistently associated with higher accuracy values. Their effectiveness in handling complex, multivariate, and temporal data was particularly emphasized. In one instance, a DL model achieved an accuracy of 95% and an Area Under Curve (AUC) of approximately 99% [28]. Hybrid approaches that combined DL with metaheuristic techniques, such as *Grey Wolf* Optimization, or signal processing methods further improved accuracy while reducing false alarm rates [29].

3.5. The State-of-the-Art Applications

The reviewed articles are classified into three application streams: (i) intrusion detection, using ML to identify cyberattacks on SCADA and PLC components; (ii) operational anomaly detection, using ML to detect abnormal behavior in hydraulic and water quality components, including leaks, sensor faults, and contamination events; and (iii) resilience-oriented approaches, coupling ML detection with mitigation and response strategies. This classification extends the scope of existing reviews by moving from isolated subsystems toward a unified CPS perspective.

3.5.1. Intrusion Detection

Recent studies frequently investigate multi-level architectures and multi-component frameworks, within which detection classifiers are integrated. This design choice is driven by datasets containing multiple and heterogeneous cyberattack types. The following subsections categorize the literature based on the taxonomy of algorithms used for intrusion detection.

Classical and Ensemble Algorithms

Several studies demonstrate the application of classical Machine Learning and ensemble methods for intrusion detection in WDNs control systems, often using simulated SCADA data or benchmark networks. In Teixeira et al. [25], classifiers such as Random Forest (RF), Decision Tree (DT), and k-Nearest Neighbor (KNN) were trained on a SCADA testbed, with tree-based models exhibiting the best offline and online detection performance. Similarly, Almalawi et al. [30] evaluated clustering-based algorithms and KNN on simulated EPANET data and reported encouraging detection accuracy for KNN, albeit with higher computational costs. Ensemble strategies aimed at improving detection performance were explored in multiple studies. For example, [29] combined bagging, stacking, and boosting on an industrial control system intrusion dataset (UNSW-NB15), concluding that an ensemble Extreme Learning Machine (ELM) achieved the highest accuracy, precision, and recall. An ensemble stacking model combining KNN, DT, and Naïve Bayes was evaluated in Lachure et al. [31], achieving high detection metrics (99% accuracy and 96% precision) on the WADI and BATADAL datasets. However, a recurring observation across these studies is that classical and ensemble approaches often require careful parameter tuning and still struggle to generalize beyond their respective training scenarios.

Autoencoder and Ensemble Algorithms

Several studies adopted unsupervised approaches, particularly AutoEncoders (AEs), for strategic multisite cyberattack detection. This is especially relevant for WDNs, where labeled attack data are scarce and attack signatures evolve rapidly. In Taormina et al. [32], a multisite intrusion detection framework based on AEs was developed and trained using BATADAL data. The results indicated that deeper AE architectures with lower compression factors achieved stronger detection performance. These findings highlight the advantages of representation learning, particularly when attack patterns span multiple sensors and control variables. However, the authors also raised concerns regarding generalizability, noting that while AEs can perform well on datasets containing stealth attacks, broader evaluation across diverse benchmarks is required. Also, the study of Mahmoud et al. [33] proposed a two-stage intrusion detection pipeline combining multiple supervised classifiers (including SVM, KNN, RF, XGBoost, and BOSS) for initial screening, followed by an Isolation Forest for attack differentiation. This approach achieved a normalized accuracy of 94% and strong sensitivity performance. Importantly, the authors emphasized the need to balance detection accuracy with detection timeliness, as delayed alarms in WDNs can lead to significant physical consequences.

The study of Ramotsoela et al. [34] further suggests that while robustness-enhancing ensemble methods remain attractive for intrusion detection, they impose operational constraints. By combining density-based outlier detectors with quadratic discriminant analysis, the authors demonstrated improved robustness through complementary detection mechanisms. However, the approach was noted to be computationally intensive, highlighting a trade-off between robustness and real-time deployability in SCADA and WDN environments. These studies indicate that integrating AEs with ensemble methods can enhance detection accuracy, but challenges related to computational burden and detection latency must be addressed before large-scale deployment.

Deep Learning Algorithms

Recent research has increasingly focused on DL and hybrid approaches for intrusion detection in WDNs. For example, the study of Sikder et al. [35] introduced “Deep H_2O”, a hybrid framework combining a Temporal Graph Convolutional Network (TGCN) with a high-confidence AE to detect subtle attacks in the BATADAL dataset. The study reported improved generalizability and successful identification of poisoned data compared to classical methods. A residual LSTM model was employed in Abughali et al. [22] within a coupled water–power grid test system, achieving detection metrics exceeding 96% across accuracy, precision, recall, F1-score, and AUC. Notably, this LSTM-based detector reduced the impact of stealthy attacks on water services by approximately 88% under real-time operation, highlighting its resilience benefits. In the study by Chandran et al. [36], a custom DL architecture, termed the Rectilinear Hybrid Belief Classifier (RHBC), was evaluated on the BATADAL dataset and achieved a 99% attack detection accuracy. Comparative studies assessing DL against classical approaches are also present in the literature. In [23,25], deep neural networks consistently outperformed simpler classifiers in intrusion detection tasks. Specifically, the study by Govea et al. [26] reported that AI-based models, ranging from SVM to deep neural networks, achieved a 98% threat detection rate while significantly reducing incident response time. In Ali et al. [28], trade-offs between accuracy and precision were observed when comparing classical algorithms (SVM, KNN, and Naïve Bayes) with deep learning models (LSTM, RNN, and CNN). While the LSTM achieved the highest accuracy, the CNN yielded superior precision, recall, and F1-score. Overall, these findings underscore the promise of deep learning for intrusion detection in WDNs. However, similar to classical methods, most high-performing models were predominantly evaluated using synthetic datasets, such as those from BATADAL and the Cybersecurity and Infrastructure Security Agency (CISA), rather than real-world operational data. Recently, features like actionable diagnostics, detection reporting, localization, and severity of attacks have been integrated together in intrusion detection algorithms [37].

Hybrid and Knowledge-Guided Algorithms

Several physics-inspired detection approaches have been proposed to address domain-specific challenges in WDN cybersecurity. In Nader et al. [38], a one-class classifier based on truncated *Mahalanobis* distance was developed and trained using real network data from a water distribution plant. Compared to traditional algorithms, the proposed method achieved faster detection times and was recommended for online intrusion monitoring. In Housh et al. [39], a hydraulic simulation model based on EPANET was combined with physics-aware, threshold-based rules to detect anomalies in the BATADAL dataset. While all simulated cyberattacks were successfully detected, careful calibration of the hydraulic model was required to minimize false alarms.

The study in Abokifa et al. [40] aimed to enhance detection robustness by applying quadratic discriminant analysis to an ensemble of outlier detection algorithms, although high computational demands were identified as a limitation. In Choi et al. [41] and Brentan et al. [42], a range of algorithms, including KNN, SVM, artificial neural networks (ANNs), physics-informed ELM, and standard ELM models, were evaluated on the BATADAL dataset. The results indicated that simpler ELM models offered superior computational efficiency but suffered from lower accuracy, higher false positive rates, and delayed detections compared to more complex methods. These limitations were consistently noted across both studies.

A consolidated discussion of ML-based intrusion detection approaches is provided in Table 3.

Table 3. Summary of the state-of-the-art ML applications for intrusion detection in WDNs.

Algorithm Taxonomy	Dataset	Evidence from Current Literature	Critical Analysis	References
Supervised and Classical ML (RF/DT/KNN/SVM/NB/LR) & Ensemble Learning (bagging/boosting/stacking).	Heavy dependence on SCADA testbeds, EPANET simulations, and benchmark intrusion datasets; limited evidence on real-utility deployment	(i) "Workhorse" techniques for intrusion detection and repeatedly reported as effective in offline/online phases, (ii) ensembles often delivered near-maximum levels of accuracy	(i) Performance depends on meticulous training and benchmark thresholds, (ii) generalization and study of attack evolutions remain under-explored; operational metrics (such as alarm burden, latency, and compute cost) are inconsistently reported.	[25,29,30]
Unsupervised Representation Learning, such as Autoencoders (AE)	Dominated by BATADAL-style benchmark use; generally limited cross-network transfer evaluation.	(i) Application suits to WDN realities where labels are scarce and multisite signatures may emerge, (ii) deep/stacked AEs exhibit strong detection when trained on "normal" operational patterns, (iii) two-stage architectures emulated near real-time conditions.	(i) Generalization concerns due to applications in limited dataset, (ii) computationally expensive, (iii) time-to-detection is flagged as underweighted relative to accuracy.	[31,32,40]

Table 3. *Cont.*

Algorithm Taxonomy	Dataset	Evidence from Current Literature	Critical Analysis	References
Deep temporal and hybrid architectures (e.g., LSTM/RNN/CNN; graph-temporal models; bespoke deep hybrid frameworks)	Frequently benchmark-driven (BATADAL/CISA) with fewer real-utility validations; sometimes mixed sources.	(i) DL is repeatedly favored to be effective for multivariate and temporal complex attack patterns, (ii) some works report very high metrics (accuracy/precision/F1/AUC) and improved poisoned-data handling/regularization.	(i) Benchmark bias: High reported scores often coincide with synthetic/testbed, (ii) model performance tends to degrade under new attacks or configurations; transfer learning is recommended	[22,26,28]
Interpretable/physics or signal-aware detection:	Mixed realism: Includes a real plant dataset exemplar, plus benchmark-driven hydraulics/signal evaluations.	(i) Strong interpretability and operator-actionability potential (clear rules/thresholds; model-based reasoning), (ii) demonstrated feasibility for timely alarming in simulated cyberattack settings; some methods highlight speed.	(i) Hydraulics-assisted thresholds require careful calibration; risk of false alerts if model mismatch exists. (ii) Trade-offs between speed and accuracy: lightweight approaches can be fast yet may miss subtle attacks. (iii) Some approaches still detect certain scenarios late, limiting operational value.	[38,39,41,42]

3.5.2. Anomaly Detection

Timely detection and evaluation of anomaly events, such as leaks, contamination, and unauthorized access, are essential and should be distinguished from cyberattack events.

Anomalies in Hydraulics and Consumption Data

Beyond cyber-intrusion detection, several studies have applied ML techniques to identify anomalies arising during normal WDN operation, such as leaks, illegal withdrawals, and sensor faults. Data-driven approaches are often combined with domain-specific knowledge to enhance anomaly identification and localization. For example, the study of Predescu et al. [27] applied a hybrid strategy to smart water meter data from a district in Milan by integrating unsupervised clustering with fault sensitivity analysis. This approach was shown to be effective in detecting distribution anomalies and significantly improved localization accuracy within the network. Similarly, Qian et al. [43] addressed a related problem by combining an ensemble of classical learners (e.g., Random Forests and gradient boosting), data balancing techniques, and sequence-to-point LSTM prediction. The integrated framework achieved a higher F1-score than any individual model, highlighting the benefits of model combination and data balancing for anomaly detection in WDNs. In another study, Vries et al. [44] evaluated three algorithms, namely Support Vector Regression (SVR), Gaussian mixture models, and the SPIRIT time-series pattern discovery method, using data from a public water utility in the Netherlands. While the algorithms successfully detected anomaly events with reasonable accuracy, their evaluation was constrained by limited diversity in anomaly types within the dataset. The authors noted that the small number and variability of historical anomalies hindered a comprehensive assessment of algorithm robustness. These studies demonstrate that integrating ML-based techniques with domain knowledge can effectively identify anomalous conditions, particularly leaks.

However, they also underscore the persistent challenge of data scarcity, as existing datasets often lack sufficient variability to fully evaluate algorithm performance.

Anomalies in SCADA and Sensor Data

Several studies have focused on developing general ML classification frameworks for detecting anomalies in water infrastructure, with particular emphasis on SCADA sensor data. For example, Robles-Durazno et al. [45] evaluated standard classifiers, including KNN, SVM, and RF, using data from a water treatment control rig. The study found that RF achieved the highest classification accuracy in distinguishing between normal and anomalous conditions. The authors recommended enhancing the testbed and incorporating more diverse attack scenarios to better reflect real-world conditions. In Muharemi et al. [46], a comparative study involving multiple algorithms, such as SVM, logistic regression, Linear Discriminant Analysis (LDA), multilayer perceptron neural networks, RNNs, and LSTMs, was conducted using operational data from a German utility company. The results showed that all tested algorithms were susceptible to certain cyber-induced anomalies or noise, and no single model was fully robust. Classical ML models, including SVM, ANN, and logistic regression, demonstrated relatively better robustness to adversarial noise. However, the study concluded that data-related challenges, such as imbalance and noise, often outweigh the impact of algorithm choice. The authors emphasized the need for unified strategies to address dataset imbalance rather than relying exclusively on algorithm tuning. Although these classification-based approaches achieved high detection accuracy on the available datasets, they also revealed a common limitation in terms of generalizability, as model performance varied across different network types.

Anomalies Across Multiple Datasets

In the last year, some studies have emphasized domain shift and dataset realism effects in industrial-control and water testbeds: Studies involving GAN–LSTM and deep learning frameworks continue to report high accuracy on curated SWaT-/WADI-style databases [47,48], while newly released large-scale labeled ICS traffic datasets enable complementary benchmarking for network-centric detectors [49].

To better assess the generalizability of ML-based anomaly detection methods, some studies validated their frameworks across multiple datasets or combinations of datasets. In [50], a regular flow anomaly detection framework was developed and evaluated using two advanced water ICS testbeds (SWaT and WADI), along with a custom industrial dataset. The framework employed combinations of algorithms such as KNN, SVM, RF, and ANN, achieving high accuracy, precision, and recall in detecting abnormal flow behavior. RF performed best when training and testing were conducted on the same dataset, whereas KNN and SVM demonstrated stronger generalization when trained on one dataset and tested on another. This observation suggests that certain algorithms may possess superior transfer learning capabilities, enabling them to adapt learned patterns to new environments. The study further noted that detection performance under diverse conditions could be enhanced through more extensive exploratory data analysis and the development of more advanced model architectures, including deep learning approaches. A related contribution by Moubayed et al. [51] applied a range of ML algorithms to sensor data obtained from a real network provided by an industrial partner, achieving performance levels comparable to those reported in Robles-Durazno et al. [50]. The results reinforced the suitability of RF for supervised anomaly detection while also highlighting that no single model consistently outperformed others across all evaluation criteria. Instead, algorithm selection was shown to depend on dataset characteristics and feature representations. A common theme across studies involving multiple datasets is the emphasis on online detection and adaptability. In particular, Robles-Durazno et al. [50] highlighted the need for online detection systems capable of monitoring anomalies in real time. This underscores the importance of developing adaptive models that continuously learn from evolving data streams rather than relying exclusively on static, offline-trained models. Overall, this body of work indicates that while high detection performance can be achieved in controlled settings, ensuring consistent performance under real-world, evolving conditions remains a key challenge.

A critical synthesis of the discussion covered in Section 3.5.2 is presented in Table 4.

Table 4. Summary of the state-of-the-art ML applications for anomaly detection in WDNs.

Anomaly Type Detected	Algorithm Taxonomy	Current Evidence	Critical Analysis	References
Hydraulic and Demand Anomalies	Hybrid, Classical and Deep Learning Algorithms	Evidence supports research beyond raising alarms, such as efforts made towards pipeline-aware detection and localization of anomalous events.	Additional detection attributes can be investigated, such as alarms/day, detection delay, drift stress tests and cross-dataset validation for localization tests.	[27,43].
Real-Time Anomaly Detection in Utility Network	Hybrid, Classical and Deep Learning Algorithms	Demonstrates the effectiveness of anomaly/event discovery in real operations where labels are weak and events are rare.	Adopt progressive validation: Operator inclusivity, iterative event-set curation, and uncertainty reporting.	[44].
Anomalies in SCADA and Sensor Data	Classical Algorithms	Application in controlled testbeds supports benchmarking; SCADA applications indicate that robustness can be achieved by tackling imbalance/noise rather than model complexity.	Quantification of missingness/noise stress tests and drift-awareness evaluation; prioritization of stability over accuracy.	[45,46].
Anomalies Across Multi-source Datasets	Classical Algorithms	Evidence supports that the “best model” is dependent on the context; generalization requires explicit cross-dataset training, testing and operational reporting.	Standardize transfer reporting across multiple datasets (domain shift, calibration transfer) and require a deployment checklist (runtime, drift, and interpretable outputs).	[50,51].

3.5.3. Resilience Awareness

A recent research focus has emerged on the application of ML techniques to enhance the resilience of WDNs against cyberattacks. Beyond anomaly or attack detection, some studies have proposed mitigation measures and mechanisms to accelerate system recovery. However, this research area remains comparatively less explored than intrusion and anomaly detection (see Figure 6). A large proportion of the reviewed studies focus exclusively on the detection stage, with very few addressing operational consequences or recovery processes. From a practical perspective, water utilities are more concerned with service disruptions, contamination propagation, and the time required to restore unmet demand. Consequently, cyber-detection tools should increasingly incorporate mitigation strategies and physical consequences within a resilience-oriented framework. A recent work was devoted to quantifying cyber risk in operational terms by analyzing WDN vulnerability to false-data injection. The same study also highlighted how attack feasibility is dependent on sensing and control coverage [52]. Practical guidance for monitoring layouts that supports both detection and localization was investigated in Parajuli et al. [53].

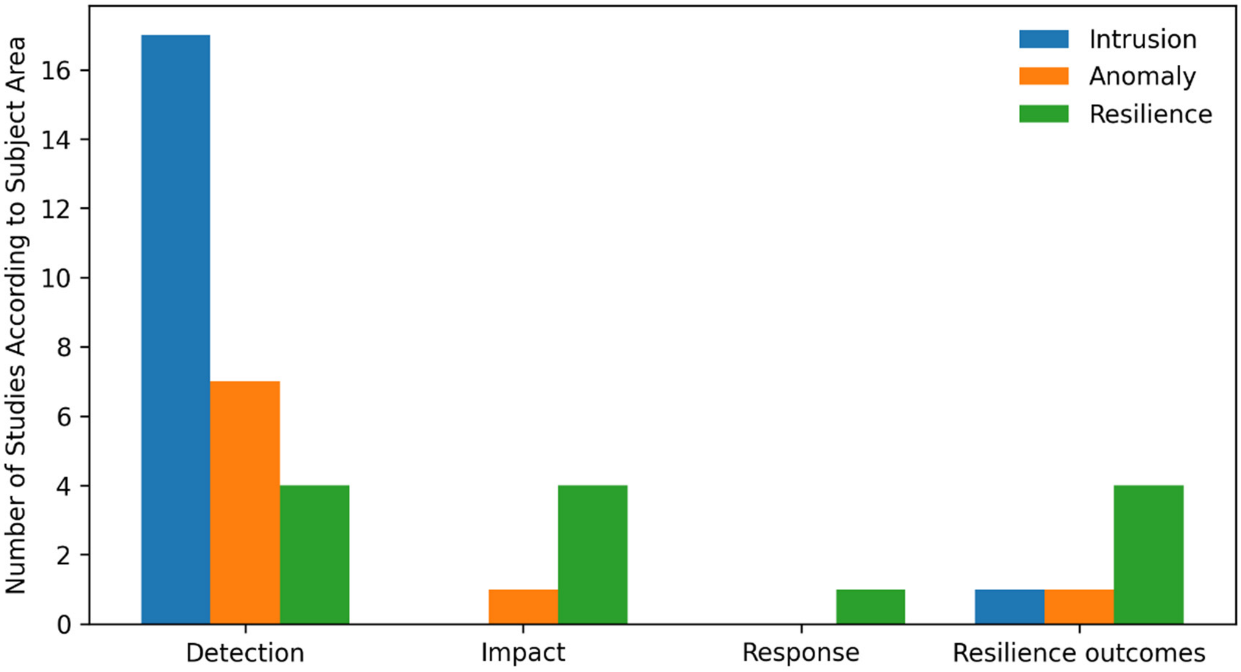


Figure 6. Different categories of applications of ML algorithms in the reviewed literature.

Resilience-Oriented Applications

One of the notable efforts integrating resilience into cyberattack detection is presented in Abughali et al. [22], where detection techniques are explicitly linked with mitigation strategies. The authors developed a residual LSTM-based detection model and evaluated it on a coupled 37-bus power grid and 35-node water distribution system. The model monitored coordinated attacks across both infrastructures and successfully identified and responded to disruptions targeting the water supply. For the water network, precision, recall, F1-score, and AUC values exceeded 95%. Following timely detection, the model reconstructed manipulated measurements as part of the mitigation process, achieving a recovery rate of 87.6%. Additionally, the approach reduced operational costs induced by attacks by approximately 2.19%. This study demonstrates that integrating AI-based detection with mitigation strategies and consequence quantification can significantly enhance resilience by preventing water outages and pressure losses. The authors emphasized the importance of flexibility and transfer learning to improve generalization across different WDNs, noting that static models may become brittle under dynamic conditions.

Emerging Frameworks

A strategic perspective on incorporating resilience into cybersecurity detection is provided in Sharmeen et al. [54]. The authors explored the feasibility of interpretable AI, federated learning, and quantum computing to enhance resilience in WDNs. Case studies from African and American water systems were used to highlight technical and socio-economic barriers to implementing resilience frameworks. For example, reluctance among utilities to share network data can hinder the adoption of federated learning. The study emphasized that AI tools must be trustworthy for operators to rely on their interpretations and recommendations. The authors concluded that resilience implementation poses not only technical challenges but also policy, privacy, and human-factor issues, highlighting the need for collaboration among researchers, policymakers, and the water sector.

Moving beyond conceptual discussions, ref. [17] attempted to operationalize resilience by designing adaptive security protocols responsive to evolving attack scenarios. The authors proposed a clustering- and deep learning-based detection framework that identified cyberattacks in a testbed dataset with near-perfect accuracy. However, system-level resilience indicators, such as service disruption or recovery time, were not evaluated. Similar to the study of Sharmeen et al. [54], adaptability was emphasized as a core requirement of resilience frameworks, and the use of ensemble classifiers was suggested to support adaptive security strategies.

Based on the findings discussed in this subsection, it can be inferred that resilience-oriented frameworks remain underexplored within the reviewed literature. This gap does not arise from methodological infeasibility, but rather from the difficulty of mapping adaptive detection capabilities to quantified system-level consequences in WDNs.

4. Limitations and Further Research Directions

A critical analysis of the reviewed studies revealed several research gaps and corresponding future directions in AI-based cybersecurity for WDNs. These gaps are organized into four categories: (i) data and benchmarks, (ii) ML paradigms, (iii) evaluation metrics, and (iv) deployment and trustworthiness. The key research gaps identified from the literature are summarized in Table 5.

Overall, in this study, according to the taxonomy of the algorithms, the classical supervised ML algorithms (such as RF, SVM, and KNN) perform better with representative labeled data and when engineered features remain stable under operating conditions. The performance of these algorithms is impacted adversely if class imbalance exists, under unknown operating conditions and unseen attacks.

Unsupervised or semi-supervised algorithms (such as autoencoders) are more appropriate under very limited instances of attacks or anomalies and when there is an abundance of baseline data for rigorous pattern recognition. Higher risk of false alarms arising from limitations related to determining the threshold values can affect the performance of these algorithms negatively.

Deep Learning and Temporal models (such as RNN, CNN and LSTM) are effective for multivariate datasets where there are instances of time-based intrusion patterns. Limitations include high computational demand and data requirement, and the performance also diminishes under stealth attacks, changes in network topology, and sensor reconfiguration, as the algorithms lack the provision of transfer/continual learning.

Physics-aware or hydraulics-assisted models are suitable under well-defined operating conditions of the WDNs. These models also face limitations from topology changes or reconfiguration of demand patterns.

A reliable localization/source attribution in WDN cybersecurity is only feasible under two conditions, when (i) the network topology (node–pipe/DMA structure) is available and (ii) ground-truth source labels are provided at a significant detailing (e.g., asset/node/pipe/DMA). In the reviewed evidence, localization is seldom reported quantitatively, meaning that many studies optimize detection metrics (e.g., F1/AUC) without demonstrating actionable source identification. As a result, topology-agnostic time-series detectors may flag anomalies yet return ambiguous candidate sources, since multiple hydraulically coupled sensors can exhibit correlated deviations under the same event. To improve operational relevance, future studies should explicitly evaluate localization using standardized metrics such as top-k accuracy, distance-to-source error, or candidate-set size, and report how localization performance varies with dataset realism and topology availability.

Table 5. Main research gaps identified from literature.

Topic	Evidence from Current Literature	Research Gap	Suggested Directions
Data & benchmarks	A significant number of studies use BATADAL/CTown and SCADA testbeds; smart-meter and water-quality datasets appear in just 1 study each.	Dataset bias is very strong and lacks real-world and diverse datasets	Development and sharing multi-utility WDN datasets (including water-quality and smart-meter streams), designing new benchmark scenarios beyond CTown and inclusion of multi-attack and noisy conditions.
Data & benchmarks	Cross-dataset validation not observed; models are trained and tested on a single network or testbed dataset	Cross-system generalization is poorly understood.	Establishing protocols where models are trained on one WDN/testbed and tested on another; understand cross-network domain adaptation and transfer learning.
Learning paradigms	Classical ML approaches, such as KNN and SVM, used more frequently in real utility and testbed data; deep/temporal models used mostly on BATADAL/CTown.	Limited exploration of deep/temporal models on real-world utility data	Application of temporal and graph neural networks to real SCADA, smart-meter and water-quality datasets; comparison against classical ML models for both detection and localization.
Learning paradigms	"Attack Prediction" and "Adaptive Learning" are underexplored; most papers explicitly call for future online learning but do not implement it.	Lack of attack prediction and adaptive/online Intrusion Detection Systems (IDS) for WDNs.	Develop forecasting models for attack likelihood and risk indices; design continual-learning IDS that update under drift and evaluate them on long-term, realistic simulations.
Learning paradigms	Only one instance of the application of graph-based deep learning; most models ignore network topology.	Topology-aware, localization-oriented methods are underexplored	Create spatio-temporal graph-based models that exploit WDN topology for attack localization and source attribution, and benchmark localization accuracy vs. topology-agnostic models.
Metrics & evaluation	Nearly all papers report only accuracy/F1/AUC on a single dataset; none systematically link detection to service/resilience metrics.	Over-dependence on static classifier metrics; absence of resilience-aware evaluation.	Introduce metrics such as detection latency, contaminated volume avoided, unmet demand and time to recovery; evaluate detectors within attack–impact–recovery simulations.
Deployment & trustworthiness	Only isolated uses of feature-importance analysis; federated learning appears only as a suggestion.	Limited work on interpretability, privacy and collaborative learning.	Develop explainable IDS interfaces (e.g., SHAP-based maps) for operators; implement federated learning prototypes across multiple utilities; quantify privacy–performance trade-offs.
Deployment & socio-technical	No empirical human-in-the-loop studies, for example, operator interaction is discussed only conceptually.	Human-in-the-loop and socio-technical evaluations are not considered.	Study how operators interpret alarms and explanations; integrate feedback into active learning loops; design and test SOC workflows tailored to WDNs.

Apart from conventional accuracy metrics such as F1/AUC, resilience-relevant outcomes (e.g., unmet demand, contaminated volume avoided, service downtime, recovery time) should be normalized and interpreted for the different tiers of datasets reviewed in this study. Tier-1/Tier-2 settings can compute such impacts via coupled hydraulic/water-quality simulation for testbed datasets, whereas for real utility datasets in Tier 3, this can be

https://doi.org/10.3390/w18040519

computed from factual baselines and ground truth. Table 6 classifies the different tiers of datasets according to the classifier-level metrics and operational metrics reported in the reviewed studies.

Table 6. Datasets adopted across the selected studies.

Dataset Tier	Classifier Metrics (Acc/F1/AUC)	Operational Metrics (Latency/Alarm Burden)	Localization Metrics (Top-k/Distance)	Service-Impact Metrics (Unmet Demand/Contaminated Volume/Downtime)	Resilience Indicators (Recovery/Time-to-Recovery/AUPC)
T1 Synthetic/simulated	Common (Acc/Prec/Rec/F1/AUC)	Rare (latency/alarm burden/runtime usually not quantified)	Rare (standard localization metrics seldom reported)	Rare (unmet demand/contaminated volume/downtime seldom quantified)	Occasional (example: recovery rate/cost reported in one study [25])
T2 Testbed	Common	Rare	Rare–occasional (feasible, but inconsistently reported)	Rare (impact feasible but underreported)	Rare (recovery/impact indicators usually not evaluated)
T3 Real utility	Common–occasional (often constrained by weak labels)	Rare	Very rare (ground truth typically unavailable)	Very rare (counterfactual baselines rarely available)	Very rare

Resilience-Aware Framework for AI Applications

The resilience of infrastructure is commonly expressed through the system's ability to absorb/resist, recover, and adapt under disruptive scenarios, and is often represented by the variation in the overall performance of the system over time [55]. This review proposes a five-step framework that effectively maps action/preparedness (S1) to timely detection/diagnosis (S2) for quantifying cyber–physical performance loss (S3) and proposes recovery through response actions (S4) and checks the overall resilience of WDNs to cyber–physical attacks using specific metrics (S5).

To link AI-based detection models with realistic operational outcomes, a five-step (S1–S5) resilience-aware cyber–physical framework is proposed (see Figure 7).

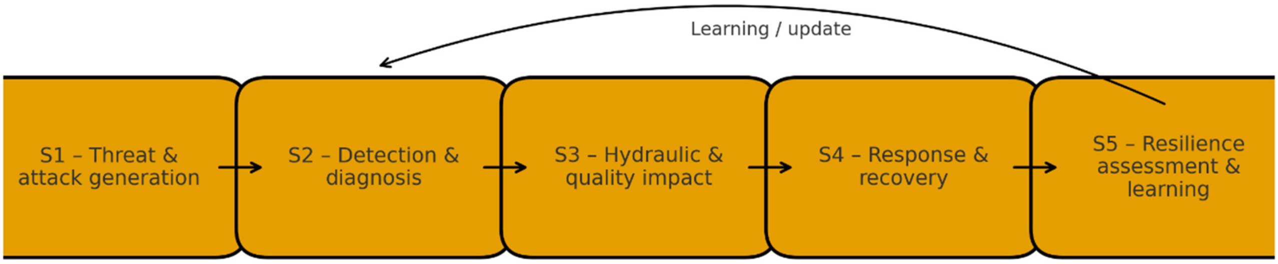

Learning / update
S1 – Threat & attack generation
S2 – Detection & diagnosis
S3 – Hydraulic & quality impact
S4 – Response & recovery
S5 – Resilience assessment & learning

Figure 7. A five-step resilience-aware framework for the widespread use of AI-applications for cyberattack detection in WDNs.

The framework comprises five interconnected components:

- S1—Threat and Attack Simulation: Cyberattacks, A_k, are simulated against WDN components, including SCADA systems, PLCs, and field devices. The target (asset, time window), intensity and attacker capability scenarios are simulated.
- S2—Detection and Diagnosis: AI-based models detect and analyze attacks, generating alarms with associated latency and false-alarm profiles. In this step, detection and diagnostic features such as alarm probability, $p_k(t)$, class detection, $\hat{c}_k$, location of the attack, $\hat{\ell}_k$, and uncertainty quantification are to be reported.
- S3—Impact Analysis: When attacks propagate through the system, the resulting effects on hydraulic and water-quality parameters are quantified. Here, the outputs from S1 and S2 parameterize S3 by enabling cyber–physical simulation under the context of

$\hat{\ell}_k$. An impact vector ΔP is constituted that quantifies impact features such as unmet demand or the contaminated volume.

- S4—Response and Recovery: Attack information is communicated to operators, who follow standard operating procedures and adjust control actions to mitigate or isolate affected zones. S4 links the previous steps to propose remediation strategies, such as adjusting pumps/valves at key locations, and to recommend any advisories.
- S5—Resilience Outcome: In the final step, resilience indicators report the performance trajectories of the overall detection and remediation process and compute attributes such as AUPC (Area Under Performance Curve) and time required for recovery.

This framework extends current AI-based detection approaches beyond isolated classification tasks toward a unified cyber–physical evaluation system. Future research should focus on end-to-end detection models integrated with simulations or digital twins that capture attack dynamics, quantify impacts, support operator decision-making, and assess contributions to overall system resilience.

5. Conclusions

This review examined the state of the art in AI-based cybersecurity for water distribution networks (WDNs), with a focus on detection algorithms for cyber and cyber–physical threats. Based on an analysis of 32 selected studies through the PRISMA method, a WDN-centered synthesis was developed across three main application streams: intrusion detection targeting SCADA and PLC components, operational anomaly detection in hydraulic and water-quality domains, and resilience-oriented approaches that integrate detection with mitigation and impact assessment. In addition to a narrative review, a structured, data-centric perspective was adopted, mapping studies across dimensions such as task type, data modality, learning paradigm, model family, deployment maturity, datasets, attack types, and evaluation metrics. This resulted in an extended taxonomy and a quantitative meta-analysis of AI applications in WDN cybersecurity.

The first key finding is that AI-based cybersecurity for WDNs is a relatively recent and geographically uneven research field. Publication trends show a sharp increase only after the mid-2010s, with most contributions concentrated in the past 5–7 years. Research activity is dominated by a small number of high-income countries, while limited representation exists from middle- and low-income regions. Although the topic has gained momentum, the empirical evidence base remains narrow and does not yet reflect the diversity of operational and institutional contexts of WDNs globally.

A second major contribution arises from the dataset-centered analysis. A significant proportion of studies rely on a single benchmark network (BATADAL/CTown)—see, for instance, refs. [56–58] or laboratory SCADA testbeds, with very limited use of real utility SCADA data. Isolated examples employ smart-meter data, water-quality anomaly datasets, or coupled water–energy models. Systematic cross-dataset validation remains rare. This benchmark bias raises concerns regarding the generalizability of reported performance metrics and highlights the need for broader multi-utility data sharing and more rigorous evaluation protocols.

Analysis of algorithm taxonomy shows that classical supervised ML models remain dominant. KNN, SVM, and tree-based ensembles are the most widely used, particularly on SCADA testbeds and real utility data. Deep and temporal models, including ANN, LSTM/RNN, autoencoders, and graph-based networks, are increasingly explored but are primarily validated on synthetic benchmarks. More than half of the reviewed studies rely exclusively on classical ML, while only a small fraction employs advanced architectures. Graph-based and physics-informed approaches that explicitly incorporate WDN topology remain rare. These patterns suggest that the field is in transition, with promising methods identified but not yet widely demonstrated at scale or on real operational data.

A third key outcome concerns underexplored task types. Most studies focus on reactive detection, identifying attacks after they occur. Early warning systems, cyber-risk estimation, and continual-learning IDS that address concept drift and evolving attack strategies are largely absent. Multi-modal data fusion and topology-aware localization remain isolated efforts. These observations indicate that predictive, adaptive, and localization-focused AI applications are still at an early stage.

A novel contribution of this review is the proposal of a resilience-aware cyber–physical framework that links algorithmic performance to system-level outcomes. Mapping existing studies onto this framework reveals that most work is concentrated at the detection stage, reporting classifier-level metrics on single datasets. Few studies quantify hydraulic or water-quality impacts, and even fewer assess resilience indicators such as unmet demand or recovery time. Almost no studies close the loop by adapting detection strategies based on resilience outcomes. This highlights the need for end-to-end evaluation of how AI-based detection influences real-world system performance.

Finally, cross-cutting issues of interpretability, privacy, and deployment are identified as critical barriers to adoption. While some studies explore explainability and federated learning, these approaches remain exceptions. Human-in-the-loop evaluations and socio-technical studies are largely absent. Addressing these challenges will be essential for transitioning AI-based cybersecurity tools from experimental prototypes to trusted, operational systems in water utilities.

Supplementary Materials: The following supporting information can be downloaded at: https://www.mdpi.com/article/10.3390/w18040519/s1: PRISMA 2020 checklist.

Author Contributions: Conceptualization, M.A.H. and M.S.; methodology, M.A.H., W.W. and M.S.; software, M.A.H.; validation, M.A.H.; formal analysis, M.A.H.; investigation, M.A.H. and W.W.; resources, M.S.; data curation, M.A.H.; writing—original draft preparation, M.A.H. and W.W.; writing—review and editing, M.S., H.A. and A.D.J.; visualization, M.A.H. and W.W.; supervision, M.S.; project administration, M.S.; funding acquisition, M.S. All authors have read and agreed to the published version of the manuscript.

Funding: This research and the APC are supported by the Towards Resilient Operation of Critical Infrastructures—Application to Water and Energy Systems (TROCI) Project, through the Taighde Éireann–Research Ireland under Grant CHIST-ERA-22-SPiDDS-07. The UK partner is supported through Engineering and Physical Sciences Research Council (EPSRC) under grant EP/Y036344/1.

Data Availability Statement: No new data were created or analyzed in this study.

Conflicts of Interest: Author H.A. was employed by the company Autonex Systems Limited. The remaining authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

References

1. U.S. Government Accountability Office. *Critical Infrastructure Protection: Cybersecurity Guidance Is Available, but More Can Be Done to Promote Its Use*; U.S. GAO (No. GAO-12-92); U.S. Government Accountability Office: Washington, DC, USA, 2011.
2. Hazell, P.; Novitzky, P.; van den Oord, S. Socio-technical system analysis of responsible data sharing in water systems as critical infrastructure. *Front. Big Data* **2023**, *5*, 1057155. [CrossRef]
3. Malatji, M.; Marnewick, A.L.; von Solms, S. Cybersecurity policy and the legislative context of the water and wastewater sector in South Africa. *Sustainability* **2020**, *13*, 291. [CrossRef]
4. Munteanu, A.; Pasqua, M.; Merro, M. Impact analysis of cyber-physical attacks on a water tank system via statistical model checking. In Proceedings of the 8th International Conference on Formal Methods in Software Engineering, Seoul, Republic of Korea, 13 July 2020. [CrossRef]

5. Khan, S.; Madnick, S. Cybersafety: A system-theoretic approach to identify cyber-vulnerabilities & mitigation requirements in industrial control systems. *IEEE Trans. Depend. Secur. Comput.* **2022**, *19*, 3312–3328. [CrossRef]
6. Dagnas, R.; Barbeau, M.; Boutin, M.; Garcia-Alfaro, J.; Yaich, R. Exploring the quantitative resilience analysis of cyber-physical systems. In Proceedings of the 2023 IFIP Networking Conference (IFIP Networking), Barcelona, Spain, 12–15 June 2023. [CrossRef]
7. Riggs, H.; Tufail, S.; Parvez, I.; Tariq, M.; Khan, M.A.; Amir, A.; Sarwat, A.I. Impact, vulnerabilities, and mitigation strategies for cyber-secure critical infrastructure. *Sensors* **2023**, *23*, 4060. [CrossRef]
8. Alharbi, A.; Seh, A.H.; Alosaimi, W.; Alyami, H.; Agrawal, A.; Kumar, R.; Khan, R.A. Analyzing the impact of cyber security related attributes for intrusion detection systems. *Sustainability* **2021**, *13*, 12337. [CrossRef]
9. Abushark, Y.B.; Khan, A.I.; Alsolami, F.; Almalawi, A.; Alam, M.M.; Agrawal, A.; Khan, R.A. Cyber security analysis and evaluation for intrusion detection systems. *Comput. Mater. Contin.* **2022**, *72*, 1765–1783. [CrossRef]
10. Rahmati, M. Adversarially robust ai for real-time cyber threat detection: A reinforcement learning approach. *Res. Sq.* **2025**. [CrossRef]
11. Skrodelis, H.; Kelle, R.; Romanovs, A. Cybersecurity in SCADA Systems with Advanced AI and ML Techniques. In Proceedings of the 2024 IEEE 65th International Scientific Conference on Information Technology and Management Science of Riga, Riga, Latvia, 3–4 October 2024.
12. Galarza Yallico, A.L.; Santos López, F.M. Detection of Cyberattacks in SCADA Water Distribution Systems Using Machine Learning: A Systematic Review of the Literature. In *Proceedings of the International Conference on Computer Science, Electronics and Industrial Engineering (CSEI 2023, Ambato, Ecuador, 13-17 November 2023)*; Garcia, M.V., Gordón-Gallegos, C., Salazar-Ramírez, A., Nuñez, C., Eds.; Springer: Cham, Switzerland, 2024; Volume 775. [CrossRef]
13. Dogo, E.M.; Nwulu, N.I.; Twala, B.; Aigbavboa, C. A survey of machine learning methods applied to anomaly detection on drinking-water quality data. *Urban Water J.* **2019**, *16*, 235–248. [CrossRef]
14. Sobien, D.; Yardimci, M.O.; Nguyen, M.B.T.; Mao, W.-Y.; Fordham, V.; Rahman, A.; Duncan, S.; Batarseh, F.A. AI for Cyberbiosecurity in Water Systems—A Survey. In *Cyberbiosecurity*; Springer International Publishing: Cham, Switzerland, 2023; pp. 217–263. [CrossRef]
15. Kanyama, M.N.; Bhunu Shava, F.; Gamundani, A.M.; Hartmann, A. Machine learning applications for anomaly detection in Smart Water Metering Networks: A systematic review. *Phys. Chem. Earth* **2024**, *134*, 103558. [CrossRef]
16. Tuptuk, N.; Hazell, P.; Watson, J.; Hailes, S. A Systematic Review of the State of Cyber-Security in Water Systems. *Water* **2021**, *13*, 81. [CrossRef]
17. Gulzar, Q.; Mustafa, K. An analytical survey of cyber-physical systems in water treatment and distribution: Security challenges, intrusion detection, and future directions. *Secur. Priv.* **2024**, *7*, e440. [CrossRef]
18. Adelani, F.A.; Okafor, E.S.; Jacks, B.S.; Ajala, O.A. Theoretical Frameworks for the Role of AI and Machine Learning in Water Cybersecurity: Insights from African and US Applications. *Comput. Sci. IT Res. J.* **2024**, *5*, 681. [CrossRef]
19. Alimi, O.A.; Ouahada, K.; Abu-Mahfouz, A.M.; Rimer, S.; Alimi, K.O.A. A Review of Research Works on Supervised Learning Algorithms for SCADA Intrusion Detection and Classification. *Sustainability* **2021**, *13*, 9597. [CrossRef]
20. 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.; et al. The PRISMA 2020 statement: An updated guideline for reporting systematic reviews. *BMJ* **2021**, *372*, n71. [CrossRef]
21. Ayas, S.; Ayas, M.S.; Cavdar, B.; Sahin, A.K. Detecting cyberattacks based on deep neural network approaches in industrial control systems. *J. Inf. Secur. Appl.* **2025**, *94*, 104206. [CrossRef]
22. Abughali, A.; Alansari, M.; Al-Sumaiti, A.S. Deep Learning Strategies for Detecting and Mitigating Cyber-Attacks Targeting Water-Energy Nexus. *IEEE Access* **2024**, *12*, 129690–129704. [CrossRef]
23. Ahmed, C.M.; Palleti, V.R.; Mathur, A.P. WADI: A water distribution testbed for research in the design of secure cyber physical systems. In *Proceedings of the 3rd International Workshop on Cyber-Physical Systems for Smart Water Networks, CySWATER '17, Pittsburgh, PA, USA, 21 April 2017*; Association for Computing Machinery: New York, NY, USA, 2017; pp. 25–28. [CrossRef]
24. Mathur, A.P.; Tippenhauer, N.O. SWaT: A water treatment testbed for research and training on ICS security. In *Proceedings of the 2016 International Workshop on Cyber-Physical Systems for Smart Water Networks (CySWater), Vienna, Austria, 11 April 2016*; IEEE: Piscataway, NJ, USA; pp. 31–36. [CrossRef]
25. Teixeira, M.; Salman, T.; Zolanvari, M.; Jain, R.; Meskin, N.; Samaka, M. SCADA System Testbed for Cybersecurity Research Using Machine Learning Approach. *Future Internet* **2018**, *10*, 76. [CrossRef]
26. Govea, J.; Gaibor-Naranjo, W.; Villegas-Ch, W. Transforming Cybersecurity into Critical Energy Infrastructure: A Study on the Effectiveness of Artificial Intelligence. *Systems* **2024**, *12*, 165. [CrossRef]
27. Predescu, A.; Mocanu, M.; Lupu, C. A fault sensitivity analysis for anomaly detection in water distribution systems using Machine Learning algorithms. In Proceedings of the 2018 IEEE 14th International Conference on Intelligent Computer Communication and Processing (ICCP), Cluj, Romania, 6–8 September 2018. [CrossRef]

28. Ali, S.M.; Razzaque, A.; Yousaf, M.; Ali, S.S. A Novel AI-Based Integrated Cybersecurity Risk Assessment Framework and Resilience of National Critical Infrastructure. *IEEE Access* **2025**, *13*, 12427–12446. [CrossRef]
29. Kayode Saheed, Y.; Harazeem Abdulganiyu, O.; Ait Tchakoucht, T. A novel hybrid ensemble learning for anomaly detection in industrial sensor networks and SCADA systems for smart city infrastructures. *J. King Saud Univ. Comput. Inf. Sci.* **2023**, *35*, 101532. [CrossRef]
30. Almalawi, A.; Yu, X.; Tari, Z.; Fahad, A.; Khalil, I. An unsupervised anomaly-based detection approach for integrity attacks on SCADA systems. *Comput. Secur.* **2014**, *46*, 94–110. [CrossRef]
31. Lachure, J.; Doriya, R. ESML: A Hyperparameter-Tuned Stacking Machine Learning Approach for Anomaly Attack Detection in Water Distribution Systems. *SN Comput. Sci.* **2025**, *6*, 505. [CrossRef]
32. Taormina, R.; Galelli, S. Deep-Learning Approach to the Detection and Localization of Cyber-Physical Attacks on Water Distribution Systems. *J. Water Resour. Plan. Manag.* **2018**, *144*, 04018065. [CrossRef]
33. Mahmoud, H.; Wu, W.; Gaber, M.M. A Time-Series Self-Supervised Learning Approach to Detection of Cyber-physical Attacks in Water Distribution Systems. *Energies* **2022**, *15*, 914. [CrossRef]
34. Ramotsoela, D.; Hancke, G.P.; Abu-Mahfouz, A.M. Attack Detection in Water Distribution Systems Using Machine Learning. *Hum.-Centric Comput. Inf. Sci.* **2019**, *9*, 13. [CrossRef]
35. Sikder, M.N.K.; Nguyen, M.B.T.; Elliott, E.D.; Batarseh, F.A. Deep H_2O: Cyber attacks detection in water distribution systems using deep learning. *J. Water Process Eng.* **2023**, *52*, 103568. [CrossRef]
36. Chandran, P.; Sunil, K.S. Safeguarding water distribution systems: Cyber-physical attack detection using rectilinear hybrid belief classifier network. *J. Water Process Eng.* **2025**, *76*, 108131. [CrossRef]
37. Raza, N.; Moazeni, F. Optimal cybersecurity framework for smart water system: Detection, localization and severity assessment. *Water Res.* **2025**, *281*, 123517. [CrossRef]
38. Nader, P.; Honeine, P.; Beauseroy, P. Detection of cyberattacks in a water distribution system using machine learning techniques. In Proceedings of the 2016 Sixth International Conference on Digital Information Processing and Communications (ICDIPC), Beirut, Lebanon, 21–23 April 2016. [CrossRef]
39. Housh, M.; Ohar, Z. Model-based approach for cyber-physical attack detection in water distribution systems. *Water Res.* **2018**, *139*, 132–143. [CrossRef]
40. Abokifa, A.A.; Haddad, K.; Lo, C.; Biswas, P. Real-Time Identification of Cyber-Physical Attacks on Water Distribution Systems via Machine Learning–Based Anomaly Detection Techniques. *J. Water Resour. Plan. Manag.* **2019**, *145*, 04018089. [CrossRef]
41. Choi, Y.H.; Sadollah, A.; Kim, J.H. Improvement of Cyber-Attack Detection Accuracy from Urban Water Systems Using Extreme Learning Machine. *Appl. Sci.* **2020**, *10*, 8179. [CrossRef]
42. Brentan, B.; Rezende, P.; Barros, D.; Meirelles, G.; Luvizotto, E.; Izquierdo, J. Cyber-Attack Detection in Water Distribution Systems Based on Blind Sources Separation Technique. *Water* **2021**, *13*, 795. [CrossRef]
43. Qian, K.; Jiang, J.; Ding, Y.; Yang, S. Deep Learning Based Anomaly Detection in Water Distribution Systems. In Proceedings of the 2020 IEEE International Conference on Networking, Sensing and Control (ICNSC), Nanjing, China, 30 October–2 November 2020. [CrossRef]
44. Vries, D.; van den Akker, B.; Vonk, E.; de Jong, W.; van Summeren, J. Application of machine learning techniques to predict anomalies in water supply networks. *Water Supply* **2016**, *16*, 1528–1535. [CrossRef]
45. Robles-Durazno, A.; Moradpoor, N.; McWhinnie, J.; Russell, G. A supervised energy monitoring-based machine learning approach for anomaly detection in a clean water supply system. In Proceedings of the 2018 International Conference on Cyber Security and Protection of Digital Services (Cyber Security), Scotland, UK, 11–12 June 2018. [CrossRef]
46. Muharemi, F.; Logofătu, D.; Leon, F. Machine learning approaches for anomaly detection of water quality on a real-world data set. *J. Inf. Telecommun.* **2019**, *3*, 294–307. [CrossRef]
47. Siddique, M.S.; Khan, M.A.R.; Ahammad, I.; Nath, N.; Das, J.R.; Rahman, F. An intelligent intrusion detection system for cyber-physical systems using GAN-LSTM networks. *Frankl. Open* **2025**, *11*, 100281. [CrossRef]
48. Gulzar, Q.; Mustafa, K. Interdisciplinary framework for cyber-attacks and anomaly detection in industrial control systems using deep learning. *Sci. Rep.* **2025**, *15*, 26575. [CrossRef]
49. Zhou, X.; Cheng, Z.; Wang, C.; Wang, S.; Tao, C.; Zhou, Z.; Chen, X.; Luo, J.; Wang, D.; Zhou, H. A dataset collected in real-world industrial control systems for network attack detection. *Sci. Data* **2026**. [CrossRef] [PubMed]
50. Robles-Durazno, A.; Moradpoor, N.; McWhinnie, J.; Russell, G.; Tan, Z. Newly engineered energy-based features for supervised anomaly detection in a physical model of a water supply system. *Ad Hoc Netw.* **2021**, *120*, 102590. [CrossRef]
51. Moubayed, A. Flow anomaly detection in harsh industrial environments: A data analytics & machine learning approach. *Measurement* **2026**, *258*, 119043. [CrossRef]
52. Raza, N.; Moazeni, F. Assessing Water Distribution Systems' Vulnerability to False Data Injection Attacks. *J. Am. Water Work. Assoc.* **2025**, *117*, 30–39. [CrossRef]

53. Parajuli, U.; Magar, B.A.; Ghimire, A.B.; Shin, S. Sensor Placement for the Classification of Multiple Failure Types in Urban Water Distribution Networks. *Urban Sci.* **2025**, *9*, 413. [CrossRef]
54. Sharmeen, S.; Huda, S.; Abawajy, J.; Ahmed, C.M.; Hassan, M.M.; Fortino, G. An Advanced Boundary Protection Control for the Smart Water Network Using Semisupervised and Deep Learning Approaches. *IEEE Internet Things J.* **2022**, *9*, 7298–7310. [CrossRef]
55. Shin, S.; Lee, S.; Judi, D.; Parvania, M.; Goharian, E.; McPherson, T.; Burian, S. A Systematic Review of Quantitative Resilience Measures for Water Infrastructure Systems. *Water* **2018**, *10*, 164. [CrossRef]
56. Rustam, F.; Jurcut, A.D.; Salauddin, M. GSPO-LAD: A Graph Neural Networks based Sensor Placements and Anomaly Detection in Water Distribution Systems. *IEEE Access* **2025**, *13*, 149462–149477. [CrossRef]
57. Rustam, F.; Salauddin, M.; Saeed, U.; Jurcut, A.D. Dual-Approach Machine Learning for Robust Cyber-Attack Detection in Water Distribution System. In Proceedings of the 14th International Conference on the Internet of Things, Oulu, Finland, 19–22 November 2024.
58. Rustam, F.; Saeed, U.; Jurcut, A.D.; Freni, G.; Sambito, M.; Salauddin, M. AI-Based Sensor Optimization and Anomaly Detection in Water Distribution Networks. In *Innovative Perspectives on Computational Intelligence and Data Science*; Chira, C., Matei, O., Pop, F., Pop-Sitar, P., Eds.; Springer: Cham, Switzerland, 2026; Volume 2794. [CrossRef]

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.

https://doi.org/10.3390/w18040519