Addressing data limitations in leakage detection of water distribution systems
Data creation, data requirement reduction, and knowledge transfer

Wu, Yipeng; Liu, Shuming; Kapelan, Zoran

DOI
10.1016/j.watres.2024.122471

Publication date
2024

Document Version
Final published version

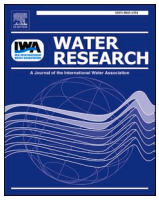
Published in
Water Research

Water Research 267 (2024) 122471

Contents lists available at ScienceDirect

Water Research

journal homepage: www.elsevier.com/locate/watres



Review

Addressing data limitations in leakage detection of water distribution systems: Data creation, data requirement reduction, and knowledge transfer

Yipeng Wu [a,b,*], Shuming Liu [a], Zoran Kapelan [b]

[a] School of Environment, Tsinghua University, 100084, Beijing, China
[b] Faculty of Civil Engineering and Geosciences, Delft University of Technology, 2628 CN Delft, the Netherlands

ARTICLE INFO

Keywords:
Leakage detection
Artificial intelligence
Data augmentation
Transfer learning
Water distribution system

ABSTRACT

Leakage in water distribution systems is a significant problem worldwide, leading to wastage of water resources, compromised water quality and excess energy consumption. Leakage detection is essential to reduce the duration of leaks and data-driven methods are increasingly being used for this purpose. However, these models are data hungry and available observed data, especially leakage data, is limited in most cases. In addition, these data need to be manually processed to label whether leaks occur, which is time-consuming and costly. These are significant obstacles for the development and application of these methods. This article provides a comprehensive review of relevant journal papers, categorizing all data-driven methods into unsupervised anomaly detection, semi-supervised anomaly detection and supervised classification methods based on how the data are utilized for developing these methods. In addition, strategies to address data limitations are summarized from both data and model perspectives, including data creation, reduction of a model's data requirements and knowledge transfer. After detailing these strategies, research gaps are identified. Based on these, future research directions are suggested, highlighting the need for further research in data augmentation, development of semi-supervised classification methods, exploration of multi-classification methods with model updating mechanisms, and development of novel knowledge transfer methods.

1. Introduction

Global water scarcity, exacerbated by climate change and urbanization, is increasingly severe, with over 80 major cities experiencing extreme drought and water shortages in the past two decades (Zhang et al., 2019). Water distribution systems (WDSs) are crucial for urban water management, but issues like aging pipelines, corrosion, external damage, and poor management lead to significant leakage (Bozkurt et al., 2022). A World Bank study reports that developing countries lose about 45 million cubic meters of water daily, causing an annual economic loss exceeding US$3 billion. Moreover, global physical water losses are estimated at 32 billion cubic meters annually, with half in developing countries (Kingdom et al., 2016). Water leakage not only wastes resources and has economic impacts but also poses public health risks through contamination (Fox et al., 2016). Additionally, the energy used in treating and distributing leaked water adds to environmental burdens (Jernigan, 2024). Given these challenges, effective leakage management strategies are urgently needed.

Leakage detection is a crucial component of leakage management strategies, aimed at promptly detecting existing leaks within WDSs, including bursts, visible leaks, and invisible leaks. This allows water utilities to swiftly undertake pipeline repairs to reduce water loss and associated damages. Leakage detection can be achieved through various methods, including hardware-based approaches, hydraulic model-based methods, and data-driven techniques (Wan et al., 2022).

Hardware-based methods for leakage detection, such as ground penetrating radar, fiber optics, and smart balls, are accurate but costly, require skilled personnel, and are often challenging to implement due to site conditions (El-Zahab and Zayed, 2019; Wong and Mccann, 2021). Hydraulic model-based methods use model calibration and other manners to identify leaks. Despite the maturity of hydraulic modeling technology, its adoption is limited by the need for precise pipe data, specialized staff, and financial investment, which many water utilities may lack (Yu et al., 2024). Due to advancements in monitoring,

* Corresponding author at: Faculty of Civil Engineering and Geosciences, Delft University of Technology, 2628 CN Delft, the Netherlands.
E-mail addresses: wu_yp2021@mail.tsinghua.edu.cn, W.Y.P.Wu@tudelft.nl (Y. Wu).

https://doi.org/10.1016/j.watres.2024.122471
Received 24 June 2024; Received in revised form 3 September 2024; Accepted 16 September 2024
Available online 18 September 2024

communication, and artificial intelligence technologies, data-driven approaches have gained significant attention from researchers and have been applied in real-world WDSs. An example is the deployment of an AI system using Artificial Neural Networks (ANNs) and a fuzzy inference system for burst detection in the UK (Mounce et al., 2011).

Leakage detection is poised to become one of the first mature applications among data-driven urban water management technologies (e.g., anomaly detection, system prediction, optimal design and operation) (Fu et al., 2022). To advance this technology towards application, focusing on data availability is crucial (Eggimann et al., 2017; Fu et al., 2022). Although the optimization of sensor placement and the decreasing cost of sensors have made it possible to collect large amounts of data (Islam et al., 2022; Yu et al., 2024), leak events, especially bursts, are infrequent, making leakage data very scarce compared to data under normal operating conditions. Additionally, data labeling is tedious and difficult, often requiring on-site inspections and pipeline excavations to confirm leaks, further reducing the availability of leakage data (Wu et al., 2023). The inherent class imbalance in monitoring data (i.e., leakage data is much less prevalent than normal data) and the difficulty of data labeling have become significant obstacles to the development of data-driven leakage detection methods.

Amid the prevalence of data-driven methods, several review articles have been published. However, these reviews mainly discuss the data types, models used, and performance obtained by data-driven methods (Fan et al., 2022; Kammoun et al., 2022; Wan et al., 2022; Wu and Liu, 2017), without specifically discussing the data limitations faced by these methods and corresponding strategies. Therefore, this review paper focuses on overcoming the data limitations, discussing efforts made by scholars from three aspects of data creation, data requirement reduction, and knowledge transfer, and identifies future research challenges and directions based on existing gaps in research.

The paper is organized as follows. To facilitate the smooth progression of the article, the authors first define the review scope and classify data-driven methods in Section 2. Then strategies for addressing data limitations in current research are discussed in Section 3. Subsequently, Section 4 presents research challenges and directions before conclusions are drawn.

2. Review methodology

2.1. Review scope

As mentioned earlier, data-driven leakage detection methods gained significant attention and research due to their independence from costly specialized equipment and the need to construct and maintain hydraulic models. Therefore, this review focuses on data-driven leakage detection methods. Considering the variability in terminology used by scholars concerning leak events, WDSs, and data-driven approaches, the authors conducted a search on Web of Science using the following keywords: "TS = (leakage detection OR leak detection OR burst detection) AND TS = (water distribution system OR water supply system OR water distribution network) AND TS = (data-driven OR machine learning OR deep learning)". As of June 2024, a total of 197 relevant journal articles were retrieved from the Web of Science Core Collection database.

After careful screening, review articles (Wan et al., 2022; Wang et al., 2022) and those focused on other topics, such as monitoring sensors (Awwad et al., 2023; Okosun et al., 2019), sensor placement (Ayati and Haghighi, 2023; Rayaroth and Sivaradje, 2019), pipe condition assessment (Momeni et al., 2023), and leakage localization (Soldevila et al., 2021), were excluded. Ultimately, this review encompasses 85 papers on data-driven leakage detection methods and 3 papers introducing publicly available datasets.

The data-driven leakage detection methods discussed in this paper cover a broader scope compared to other reviews, utilizing a range of data including flow data, pressure data (collected from both conventional loggers and transient devices), and acoustic signals. Unlike specialized equipment in hardware-based methods, sensors collecting such data enable long-term monitoring over large areas. These sensors are essential daily monitoring facilities within pipeline networks (e.g., flow meters and pressure gauges) or easily installed with relatively low equipment costs (e.g., accelerometers for collecting acoustic signals) (Islam et al., 2022). Therefore, the data-driven approach in this review refers to using long-term monitoring data from WDSs, complemented by statistical or machine learning techniques, for effective leakage detection. Although both acoustic and hydraulic data-driven methods are reviewed together, it should be noted that they are dependent on different characteristics of pipe networks. Methods using acoustic signals are influenced by pipeline characteristics, leak point shapes, and the surrounding environment. In contrast, methods using hydraulic data are affected by network topology, water consumption patterns, and the operation of pumps and valves. As a result, different strategies are employed to address data limitations in these distinct methods, which will be detailed in Section 3.

Poor data quality, such as missing data, extreme outliers and duplicates, can also result in low data availability, impacting the development and performance of data-driven methods (Kirstein et al., 2019). However, improving data quality is not a trivial matter, involving various aspects such as sensors, data transmission, and algorithms. Additionally, the articles reviewed here have limited coverage of data quality issues. Therefore, the data limitations mentioned in this review mainly refer to the two major issues mentioned in the introduction: class imbalance and labeling difficulty, which persist even in the presence of relatively good data quality. Normal data in this context refers to monitoring data generated under the normal operating conditions of WDSs, as opposed to abnormal scenarios like leakage and firefighting. However, normal data may still show periodic or non-periodic fluctuations due to holidays, seasonal changes, or environmental factors. Leakage data can significantly differ from normal data, such as with higher flow rates, sudden pressure drops, and stronger acoustic signal amplitudes (Lee et al., 2016; Liu et al., 2024; Romano et al., 2014). Without expending additional manpower and resources, the data collected from WDSs typically consist of a large amount of normal data mixed with a small amount of anomalous data (including leakage data), resulting in unlabeled and imbalanced datasets.

2.2. Data-driven method categorization

Since the categorization of data-driven methods proposed by Wu and Liu (2017), researchers have generally adopted their classification based on the utilized techniques, gradually forming three major categories: prediction-classification methods, statistical methods, and clustering methods (Wan et al., 2022). However, this classification primarily focuses on studies using flow and pressure data, and with the emergence of various strategies to address data limitations, many new technology-based methods have emerged, rendering this classification somewhat outdated. This paper proposes a new categorization of data-driven leakage detection methods based on data utilization manners, particularly suitable for discussing how to overcome data limitations. Through an analysis of 85 primary literature sources, data-driven methods are categorized into three main classes: unsupervised anomaly detection, semi-supervised anomaly detection, and supervised classification.

In the first two categories, leakage detection is regarded as an anomaly detection task, whereas it is approached as a classification task in the third category. Anomaly detection here refers to developing a data-driven method to identify abnormal patterns in observed data that indicate a leak. Classification, on the other hand, involves developing a method to classify observed data to determine whether they indicate the presence of a leak or not. The term "supervised" refers to using manually labeled data (e.g., indicating whether given pressure/flow/acoustic signals suggest a leak) to develop data-driven leakage detection methods. "Unsupervised" denotes the use of unlabeled data. "Semi-

supervised" refers to not requiring fully labeled data, which will be detailed in Section 2.2.2.

The brief description about each category is presented in Table 1. Supplementary material provides detailed information on the type of observed data used, data sources, techniques and respective categories for the leakage detection methods presented in the 85 analyzed articles.

2.2.1. Unsupervised anomaly detection

This category requires the least data requirements, as it does not rely on labeled data or address class imbalance. The underlying assumption is that normal data points share similar statistical properties or cluster tightly in feature space (i.e., a multi-dimensional space where each dimension represents a different characteristic or measurement used to describe the corresponding monitoring data), whereas anomalous data like leakage data exhibit distinct statistical properties or are dispersed across the feature space. Consequently, statistical process control (SPC) (Jung et al., 2015) or unsupervised techniques such as clustering (Hu et al., 2022) can be employed to identify anomalous data, including leakage data. This category constitutes the smallest proportion among all methods, appearing in only 10 out of 85 literature sources.

2.2.2. Semi-supervised anomaly detection

This category of methods also has relatively low data requirements, leveraging the relatively abundant normal data in monitoring datasets. This entails the need to filter out existing abnormal data from the raw dataset, although this step may not always be explicitly stated in the literature. These anomalies may or may not be labeled and may or may not signify leakage occurrences. Subsequently, various methods, such as similarity analysis (Wu et al., 2016), prediction error analysis (Mounce et al., 2011), and reconstruction error analysis (Kammoun et al., 2023) can be used to discern discrepancies between new monitoring data and the established normal dataset (more details can be found in Section 3.2.1). Significant disparities suggest outliers, potentially indicative of leakage incidents. This category of methods is notably prevalent, found in 34 out of 85 reviewed articles.

It is important to note that in anomaly detection (Song et al., 2017), the term "semi-supervised" differs from its conventional definition in semi-supervised learning. In machine learning, semi-supervised learning typically involves using a small amount of labeled data along with a large amount of unlabeled data to enhance performance (van Engelen and Hoos, 2020). Under this definition, semi-supervised methods are typically applied in supervised learning tasks such as classification. Therefore, semi-supervised classification is used to differentiate it from semi-supervised anomaly detection. However, within the scope of this review, there is currently no existing research on leakage detection methods falling under semi-supervised classification.

Existing research shows a certain level of confusion regarding unsupervised and semi-supervised leakage detection methods. Some researchers classify methods as unsupervised leakage detection simply because they employ machine learning techniques traditionally associated with unsupervised learning. However, they only use normal data for modeling, thus applying unsupervised machine learning techniques in a semi-supervised manner (Quiñones-Grueiro et al., 2018). To avoid confusion, this paper suggests defining whether a method is unsupervised or semi-supervised leakage detection based on how the data is utilized.

Table 1
Categorization of data-driven methods and corresponding descriptions.

Category	Data utilization approach	Typical technical steps	Representative techniques used
Unsupervised anomaly detection	Building models using unlabeled and imbalanced data	1) Outlier detection using statistical or unsupervised techniques (required step); 2) Leakage identification based on prior knowledge to identify outliers that exhibit leakage characteristics, such as higher flow rates, sudden pressure drops, and stronger acoustic signal amplitudes.	CUSUM, *k*-means, isolation forest
Semi-supervised anomaly detection	Building models only using normal data	1) Original outlier removal to construct a normal dataset (required step); 2) New outlier detection based on similarity analysis or prediction / reconstruction error analysis* (required step); 3) Same as step 2 of unsupervised anomaly detection.	Density-based clustering, AE, LSTM
Supervised classification	Building models with labeled data including both leakage and normal instances	Classifier training based on leakage and normal data or extracted relevant features to directly identify leakage events.	SVM, RF, CNN

Note: * New outlier detection methods will be detailed in Section 3.2.1.
CUSUM, cumulative sum; AE, autoencoder; LSTM, long short-term memory; SVM, support vector machine; RF, random forest; CNN, convolutional neural network.

2.2.3. Supervised classification

This category of methods has the highest data requirements. First, the data must be labeled; supervised learning models use labeled datasets to train, attempting to find a function that can accurately classify new input data. During training, the model adjusts its parameters based on the errors between its predictions and the true labels, to minimize classification errors as much as possible. Second, the data across different classes must be balanced, otherwise the results will be biased towards the majority class (Guo et al., 2024). Bykerk and Valls Miro (2022) did not address class imbalance, and as a result, despite an overall identification accuracy of 98 %, the model's detection accuracy for leakage data in the test dataset was 91 % (sensitivity), significantly lower than the 99 % (specificity) for normal data.

Although previously rare, as noted by Wu and Liu (2017), the emergence of strategies to overcome data limitations, which will be discussed later, has led to a significant increase in such methods, with 40 out of the 85 papers reviewed employing this approach. Among the 85 papers, one (McMillan et al., 2023) focuses on predicting leakage and normal data using labeled data but does not provide specific leakage detection methods and results. This paper's data labeling method and prediction approach may offer valuable insights for data-driven leakage detection, so it is included in the review but does not belong to any specific category.

3. Strategies for addressing data limitations

Continuous data collection is considered the most direct and fundamental way to address data limitations. However, due to constraints such as time costs, labor costs, and objective conditions (e.g., the infrequent occurrence of burst events and the hidden nature of leaks), constructing a comprehensive dataset (including a large number of labeled leakage samples) for any given WDS is deemed difficult, if not unrealistic. Consequently, when developing a data-driven leakage detection method, different degrees of data limitations are typically encountered, necessitating diverse solutions. In this paper, these three

strategies are summarized and analyzed from both the data and model perspectives.

From the data perspective, researchers attempt to either artificially create data (or data labels) that are difficult to collect or increase the quantity of existing limited monitoring data for use in the model development. From the model perspective, two strategies are distinguished: reducing the model's data requirements and knowledge transfer. The former pertains to situations where data-driven models are developed from scratch (e.g., an ANN classifier is trained using observed data for that WDS only), while the latter involves utilizing existing data and models from other WDSs or fields (e.g., a well-established classifier in a different field or WDS is further trained using the data from the analyzed WDS). Fig. 1 illustrates the three strategies used to address data limitations in this paper, together with different specific approaches used for each strategy. These individual approaches are analyzed in detail in the following sections.

Fig. 2 shows the relationship between strategies for addressing data limitations and categories of data-driven leakage detection methods. Overall, data creation provides a foundation for the development of all methods. The emergence and application of unsupervised anomaly detection and semi-supervised anomaly detection methods are mainly attributed to the modeling using unlabeled or normal data, while the flourishing of supervised classification methods relies on comprehensive support from most strategies.

3.1. Creating data to increase data availability

3.1.1. Automatic labeling

Automatic labeling is primarily used for creating labeled data by automatically classifying raw, unlabeled data into normal data and anomalous data (including leakage data). This is an essential step in semi-supervised anomaly detection methods. Typically, automatic labeling does not concern itself with whether the labeled anomalous data truly represents leakage events, the objective is merely to construct a sufficiently clean normal dataset (i.e., data points share similar statistical properties or cluster tightly in feature space). Subsequently, new data with significant differences from the normal dataset (e.g., surpassing a pre-defined threshold) can be identified as anomalous through prediction or reconstruction error analysis, similarity analysis, and other methods (as detailed in Section 3.2.1). As a result, most automatic labeling techniques are relatively simple and coarse. For example, Romano et al. (2014) used statistical tests (specifically, the Shewhart control chart to determine if individual measurements or historical data averages exceed control limits) to filter out anomalies in historical pressure and flow data. Wu et al. (2016) and Wu et al. (2018) identified observed data anomalies based on the density and distance of flow data in the feature space, classifying data points with the lowest density and far from others as anomalous.

In recent years, more sophisticated methods have emerged to accurately label anomalous data. These methods typically rely on water utility repair records, making it possible to directly label leakage data. McMillan et al. (2023, 2024) used Isolation Forest (IF) method to identify anomalies in flow monitoring data. The principle behind IF is to isolate observations through random partitioning; anomalies require fewer partitions to be isolated. If the timestamps of anomalous data match the timestamps in repair records and the duration exceeds five hours, the corresponding anomalous data are considered leakage data. Yan and Huang (2023) initially labeled raw monitoring data using repair records to obtain pseudo-labels and then used confident learning to clean the labels, improving accuracy. Specifically, this study constructs a Random Forest (RF) classifier based on pseudo-labels, compares pseudo-labels with predicted labels to obtain label confidence, and removes low-confidence labeled data. Because the labeled data generated by these methods are more accurate, the labeled leakage data are not discarded but can be used to train classification and prediction models (McMillan et al., 2023, 2024).

3.1.2. Experiment-based data generation

Experiment-based data generation methods can increase the data volume, especially the amount of leakage data, and directly produce labeled and balanced datasets. These experiments are divided into two categories here.

The first category involves constructing pipelines or simple pipe networks in a laboratory and then artificially altering various conditions to create with and without leakage scenarios, collecting the corresponding data using sensors. This approach is common in acoustic leakage detection research. For example, Cody et al. (2018) simulated leaks by using valves to release water in an open-air branched polyvinyl chloride (PVC) pipeline network. Cody et al. (2020) upgraded the experiment by establishing a simple 30-m long open-air looped PVC

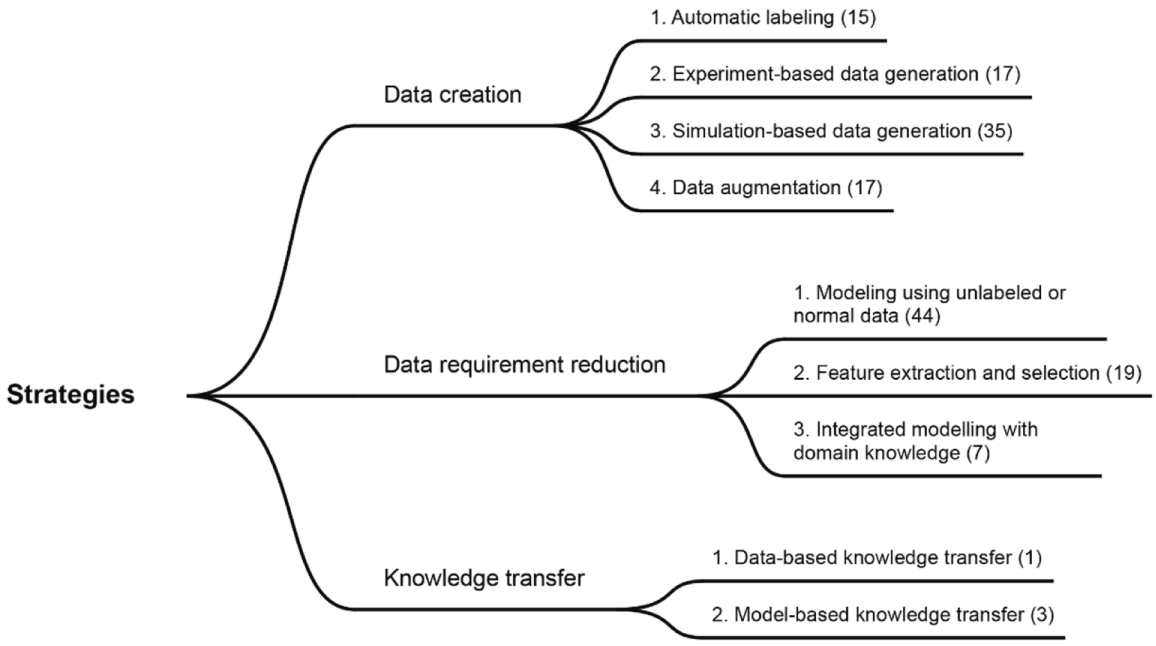


Fig. 1. Strategies for addressing data limitations. Numbers in brackets denote the number of existing publications reviewed here that fall under different approaches.

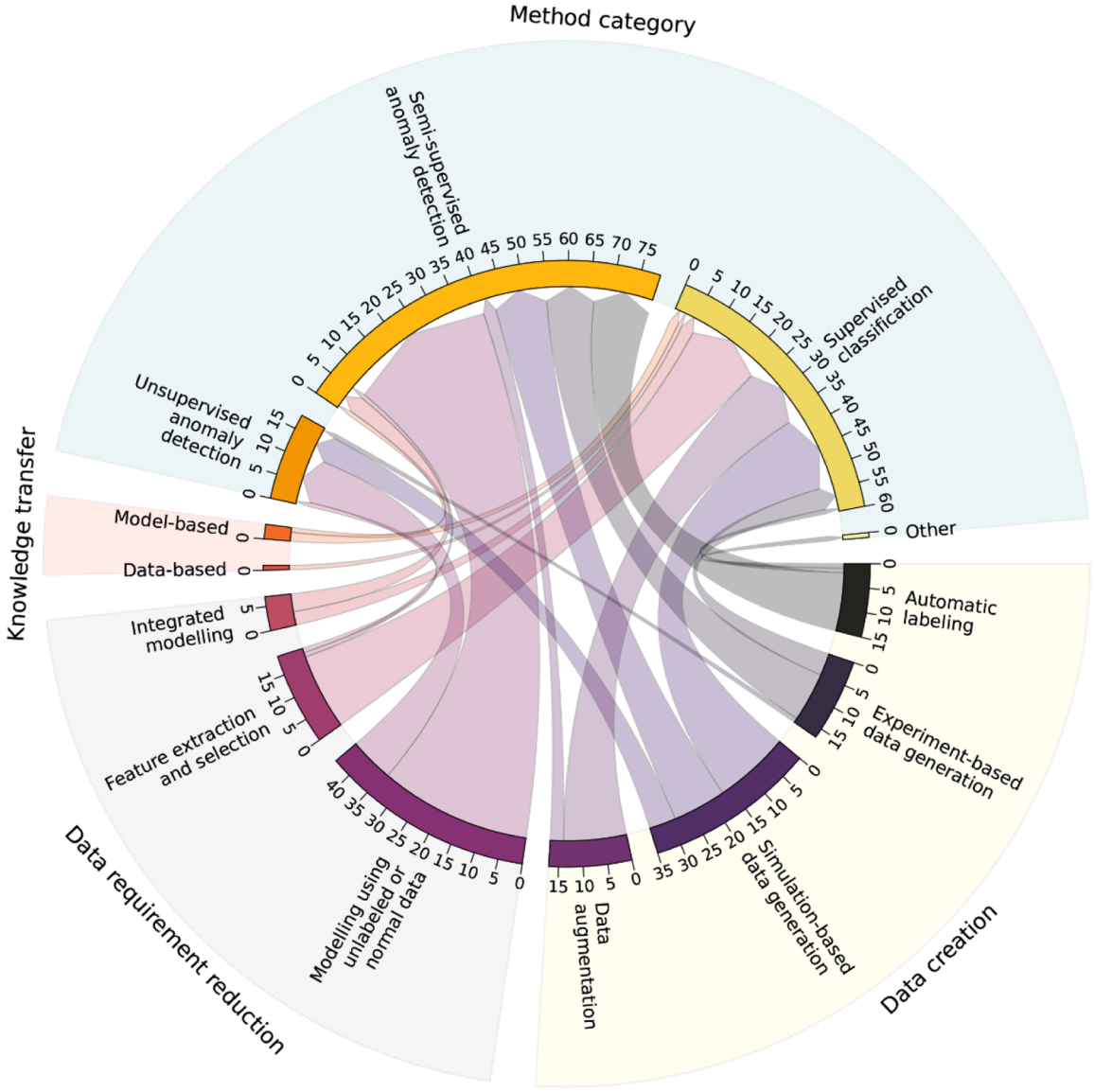


Fig. 2. Relationship between strategies for addressing data limitations and categories of data-driven leakage detection methods. Numbers in the figure denote the number of reviewed papers using the corresponding strategy.

pipeline network. Shukla and Piratla (2020) experimented with a more complex PVC network, incorporating buried pipes at varying depths and diameters. Aghashahi et al. (2023) published a dataset simulating normal and four types of leak signals (orifice, longitudinal, circumferential, and gasket) in open-air branched and looped plastic networks under six different flow rates and noise conditions, generating 140 30-s signals from each network. Despite these efforts, laboratory conditions (e.g., leak point shapes and pressures) remain too simplistic compared to real pipeline networks, and most experiments are conducted in open-air networks. Consequently, the experimental data cannot fully replicate the signals in actual operating WDSs.

The second category involves using fire hydrant tests to simulate leaks in real operational WDSs. This is a widely used approach across various studies, whether focusing on flow and pressure (Glynis et al., 2023; Huang et al., 2018; Weyns et al., 2023; Wu et al., 2016; Ye and Fenner, 2014), acoustic signals (Kang et al., 2018; Ravichandran et al., 2021), or transient pressure (Lee et al., 2016). However, opening fire hydrants repeatedly or for longer periods of time can lead to water waste and pose a risk to the water supply of nearby users in terms of water discoloration. The fire hydrant tests are also time consuming/expensive to conduct and can mimic the real pipe bursts only up to a point, especially for more complex leakage events with dynamically varying flow conditions. Consequently, the number of observed data obtained using this method is rather limited.

3.1.3. *Simulation-based data generation*

Simulation-based data generation methods rely on models constructed based on physical laws. In theory, as long as the physical principles governing the flow in a WDS are understood and translated into mathematical formulas and models, it is possible to generate large amounts of data under various conditions, thereby creating a labeled and balanced dataset. Currently, these methods are primarily used in flow and pressure-based leakage detection studies, with hydraulic models being the main tools for data generation (Fu et al., 2024; Jung et al., 2015; Wan et al., 2022; Zanfei et al., 2022). Among the 58 papers on leakage detection using flow and pressure data, 32 employ hydraulic models to generate data for model development, covering all three categories of data-driven methods.

Calibrated hydraulic models typically generate deterministic and smooth data, which significantly differ from the complex and variable

hydraulic monitoring data encountered in real-world scenarios. To address this issue, Menapace et al. (2020) proposed a stochastic hydraulic time series generator to produce simulated data more akin to real data. This generator first superimposes daily, weekly, seasonal, and random variations to output user water demand patterns, and then uses a hydraulic model to simulate normal water usage (including background leakage) and burst events, generating sufficient data for model development. It is noteworthy that burst events vary as much as possible in terms of flow, location, duration, and type to mimic real-world conditions. In the 2020 Leakage Detection and Isolation Methodologies competition (BattLeDIM), the organizers provided a simulated dataset that accounted for user water demand variations and uncertainties. In addition to burst events, this dataset included concurrent leakages and gradually developing leaks (Vrachimis et al., 2022). The above two articles provide two valuable publicly available datasets for data-driven leakage detection research. Among the 33 papers that used hydraulic models to generate data, 10 utilized this BattLeDIM dataset.

Most of the studies using hydraulic models to generate data do not consider subsequent leakage localization. These models primarily determine whether a leak has occurred in the WDS or a specific district metering area (DMA) based on features like pressure drops and increased flow rates. The advantage of this approach is that these features are universally applicable, meaning that although the model is built using simulated data from a specific network's hydraulic model, it can be easily adapted to other WDSs through techniques like knowledge transfer, as discussed in Section 3.3. Some studies, however, go beyond leakage detection and also consider localization (Cheng et al., 2022; Zhou et al., 2019). These studies begin by optimizing the placement of pressure sensors, taking into account the sensitivity of various nodes to pressure changes within a hydraulic model. They then generate a large number of leakage scenarios for each pipeline and use this leakage-specific data to train a model. This model is designed to predict the probability of a leak occurring in each particular pipeline or DMA, thus achieving localization. However, while these studies accomplish both leakage detection and localization, the sensitivity of pressure monitoring points to pressure changes caused by leaks or bursts will shift if the network topology changes, rendering the localization model ineffective. In other words, this data-driven approach is heavily dependent on a specific hydraulic model, making it difficult to generalize and apply to other WDSs.

In research utilizing transient pressure data, some scholars have also employed simulation methods to generate data. Bohorquez et al. (2020) used the Method of Characteristics, converting two hyperbolic partial differential equations governing unsteady flow behavior into four ordinary differential equations to obtain variations in flow and pressure along the pipeline at different points over time, thereby simulating transient pressure data. Bohorquez et al. (2022) built upon this numerical simulation method by adding white noise to the generated data to simulate the background noise in real pipelines, aiming to enhance the generalization capability of leakage detection models on real data.

Simulated datasets resembling real-world data provide convenience for developing various data-driven methods, ranging from SPC approaches (Ahn and Jung, 2019) to the currently popular deep learning models (Fu et al., 2024). However, most current research based on hydraulic simulation datasets still remains at the stage of developing and comparing models. Whether these models can be equally effective in real-world WDSs requires further investigation. After all, there is still some gap between simulated and real data (Basnet et al., 2023).

3.1.4. Data augmentation

Data augmentation increases the size of a dataset by applying various transformations to it, creating a larger and more diverse dataset to improve model performance and robustness (Wen et al., 2021). Overall, data augmentation can address class imbalance by augmenting labeled leakage data and also increase the total amount of existing data.

In acoustic leakage detection research, the application of data augmentation is relatively common, but it is mainly limited to framing (Guo et al., 2021; Kang et al., 2018; Shukla and Piratla, 2020; Yu et al., 2023; Zhang et al., 2023). Specifically, these studies divide the original acoustic signal into equally sized time frames, treating each frame as a completely new signal. For example, a 5-s signal may be evenly split into five consecutive, non-overlapping 1-s time frames. This way the dataset size can be significantly increased in terms of number of available signals, but not in terms of quantity of observed signals from different scenarios. Acoustic signals tend to be relatively stable over short periods, with minimal changes in pipeline status and surrounding environment. Therefore, it is reasonable to assume high similarity between time frames of the same signal. This, in turn, can lead to data leakage, where frames from the same original signal are distributed across training, validation, and test datasets simultaneously. In supervised classification methods, this can result in classifiers gaining information from unseen test datasets during training, leading to overly optimistic results (Zhu et al., 2023). This issue has been confirmed to exist in the study of acoustic leakage detection by Wu et al. (2024). In the aforementioned studies that used framing to augment data, the accuracy for leakage detection reached unrealistic 95 % or even more than 99 %, which is likely due to aforementioned data leakage.

Besides framing, Tariq et al. (2022) addressed the class imbalance issue using the Synthetic Minority Over-sampling Technique (SMOTE). This method generates synthetic samples by interpolating between existing acoustic leakage signals, effectively creating new signals that fall between known leakage instances and providing new information to data-driven models. This approach is obviously more reasonable than the simple over-sampling (i.e., repeatedly using the same leakage signal) mentioned by Guo et al. (2024). In a recently published paper, Liu et al. (2024) used adversarial training between the generator and discriminator in a generative adversarial network (GAN) for augmenting acoustic signals. The generator is responsible for creating new samples, while the discriminator evaluates their authenticity. The generator continuously improves until its generated samples are indistinguishable from real data, thereby achieving data augmentation. The obtained results indicated that this advanced approach outperformed the SMOTE in enhancing leakage detection performance. In the context of widespread use of simple techniques, this deep learning-based data augmentation method using GAN is an inspiring attempt.

The feasibility of framing stems from the rich information about the vibration status of water-filled pipes contained in high-frequency (typically in the kHz range) acoustic signals, but this does not apply to conventional flow and pressure data with sampling intervals usually ranging from 1 to 15 min (Mounce et al., 2012). For such hydraulic monitoring data, Tornyeviadzi et al. (2023) utilized a pattern mixing method similar to SMOTE, averaging the nighttime flow time series of adjacent days to generate new time series and increase the volume of data. Huang et al. (2018), Wu and Liu (2020) and Jian et al. (2022) employed outlier injection to generate new leakage data. Outliers including peaks and slope-like trends are manually designed based on the characteristics of actual leakage data (e.g., duration and maximum value). Then the outliers are directly added to normal data to simulate burst and slow-developing leakage events. However, these data generation processes did not consider the hydraulic characteristics of the pipe network. If the detection task involves data from multiple adjacent monitoring points, artificially adding outliers to simulate data leakage from multiple hydraulically related monitoring points becomes more challenging.

3.1.5. Research gap

Automatic labeling only enriches the information of the existing data (i.e., adding labels) and does not increase data volume. Currently, the primary methods used to increase data volume, especially leakage data, are still experimentation and simulation with hydraulic models, accounting for 61 % in the 85 reviewed papers. However, data generated this way can only partially reflect real-world conditions. As mentioned

earlier, the experimentally analyzed pipe networks are limited, and the factors considered in simulated data are not as comprehensive as those in real operational conditions. An alternative way is to use some data augmentation techniques to increase the observed data volume. When done well the data generated through augmentation can retain the characteristics of real-world conditions and its beneficial effects have been confirmed in the field of speech recognition, which also utilizes time series data (Kong et al., 2020). However, data augmentation in the leakage detection domain has received much less attention so far. Although 17 papers within the review scope involve data augmentation, existing applications are mainly based on simple pattern mixing and framing, which may also pose the risk of data leakage. Therefore, the primary research gap in the strategy of data creation lies in the fact that research on data augmentation for time series is still insufficient and immature.

3.2. Reducing data requirements of a leakage detection model

3.2.1. Modeling using unlabeled or normal data

Unsupervised anomaly detection and semi-supervised anomaly detection methods effectively reduce a model's data requirements because they do not necessitate precise and complete data labels nor address the class imbalance issue. Unsupervised anomaly detection directly utilizes unlabeled and imbalanced data, employing statistical or unsupervised techniques to identify anomalous data that may represent leakage events. However, these methods face two challenges. Firstly, parameters in these methods (e.g., parameters to determine control limits in SPC methods, number of neighbors in density-based unsupervised algorithms) are often difficult to determine, leading to models that are overly sensitive or insensitive to anomalies (Ahn and Jung, 2019; Muniz Do Nascimento and Gomes, 2023). Therefore, some studies have attempted to find more robust statistics to avoid misjudgment of anomalous data (Wan et al., 2022), while others have combined multiple techniques to provide a comprehensive score for anomaly detection (Hu et al., 2022). Secondly, the leakage data and normal data may overlap in feature space, which is particularly common when leakage flows are small. This phenomenon can easily degrade distance-based and density-based unsupervised leakage detection methods. In comparisons with supervised classifiers like ANN and RF, Mashhadi et al. (2021) found this issue. Due to these influences, coupled with the increase in strategies to address data limitations, these methods are not mainstream.

Semi-supervised anomaly detection only require some simple steps to filter out anomalies (i.e., automatic labelling) to obtain normal data for modeling. Among these methods, three popular technical approaches exist. The first approach analyzes prediction errors, known as the prediction-classification method, primarily applied to flow and pressure data in leakage detection. It builds a prediction model using normal data and uses statistical methods to analyze residuals between predicted and measured values, identifying larger differences indicative of leakage events (Wu and Liu, 2017). Recent studies have increasingly utilized deep learning models like Long Short-Term Memory (LSTM) neural networks for precise forecasting to improve leakage detection performance (Fu et al., 2024; Lee and Yoo, 2021; Wang et al., 2020). Some studies have also broken away from the paradigm of single-step prediction, using a multi-input multi-output strategy for multi-step prediction to better identify gradually developing leakage events beyond burst detection (Wan et al., 2023).

The second approach involves analyzing reconstruction errors, applied in studies utilizing both flow/pressure and acoustic data. Autoencoders (AEs) have gained popularity in recent years within this approach. AEs consist of symmetric encoder and decoder structures, where the encoder compresses the original data into a low-dimensional feature space, and the decoder reconstructs the compressed feature back to the original space. After training the AE with a normal dataset, inputting new data yields reconstruction errors, which can then be statistically evaluated to determine whether the error is significantly larger than the reconstruction error of normal data. A large error may indicate the occurrence of a leak (Cody et al., 2020; Fan and Yu, 2022; Tornyeviadzi and Seidu, 2023). AEs can adopt various forms of neural network structures, and research has shown that AEs using LSTM achieve better detection performance than Convolutional Neural Networks (CNNs), primarily because LSTM is more suitable for extracting features from time series data (Kammoun et al., 2023; Tornyeviadzi et al., 2023). Besides AEs, clustering (Zhao et al., 2024) and principal component analysis (Quiñones-Grueiro et al., 2018) can also be used for data reconstruction, although they are less commonly employed.

The third approach involves similarity analysis, which compares the numerical or shape similarity between new data and a normal dataset using various distance measures (e.g., Euclidean distance and cosine distance). Lower similarity indicates anomalous data (Leite et al., 2024; Wu and Liu, 2020; Wu et al., 2018; Wu et al., 2018; Wu et al., 2016). This approach is primarily applied to flow and pressure data.

In addition to the three main approaches, researchers also utilized common unsupervised models such as one-class Support Vector Machine (SVM), Gaussian Mixture Model (GMM), and IF to identify anomalous data (Cody et al., 2018; Cody et al., 2020; Yan and Huang, 2023; Zhang et al., 2023). Blázquez-García et al. (2021) adopt a rather unique self-supervised approach. Firstly, linear transformations are applied to normal nighttime flow rates to generate different levels of flow rate data, where higher flow rates correspond to higher levels. These levels are then used as pseudo-labels to train a classifier. When new data is input into the classifier, obtaining a high-level output indicates a potential leakage.

As mentioned above, various methods using unlabeled or partially labeled data have been developed for leakage detection, treated as an anomaly detection task. Noticeably, not all detected anomalies are leakages, leading to false alarms if not further processed. Scholars have tried to filter leakage data from anomalies based on criteria like anomaly duration (Wang et al., 2020) and the coordinated changes in data from surrounding monitoring points (Wu et al., 2016). However, these are filtering rules artificially set based on researchers' observations of limited data. If a new leakage pattern emerges that does not adhere to the preset rules, there may be missed detections. For example, gradually developing small leaks may pose a challenge for models initially designed to detect bursts.

3.2.2. Feature extraction and selection

In machine learning, feature extraction and selection are two crucial steps that reduce the dimensionality of raw data and eliminate redundant information (Guyon et al., 2008). As the selected features simplify the data and better reflect its essence, models are simplified, and the required data volume is consequently reduced (Wu et al., 2023). This means that a small amount of explicitly labeled leakage data may suffice for model development, particularly facilitating the development of supervised classification methods. Of the 19 articles elaborately discussing feature extraction or selection, 17 fall under supervised classification methods.

Researchers have proposed various methods for feature extraction. In studies utilizing flow and pressure data, common features include time series segments, differences between the current time point and the previous time point, and differences between the current time point and the same time point of the previous day (Hu et al., 2022; Sen et al., 2024; Sun et al., 2022). Xu et al. (2020) applied Fourier transform to pressure data to filter out slow-changing components like daily variations, focusing on rapid changes indicative of pipe bursts. This is an insightful attempt to use signal processing methods for feature extraction in non-high-frequency pressure data. Jian et al. (2022) and Zhang et al. (2022) directly used the differences between predicted model output values and measured values as input features for the classifier. Kim et al. (2022) transformed the preliminary results of six SPC methods, which determine whether there is a pipe burst under different parameter

settings, into grayscale heatmaps. These preliminary results were used as input features for a CNN classifier. This approach eliminates the need for precise determination of the parameters of SPC methods while providing strong detection evidence for the CNN classifier.

In leakage detection research utilizing acoustic signals, statistics in both the time and frequency domains, such as energy, skewness, and kurtosis, are commonly used features, as elaborated in the review by Fan et al. (2022). Differing from conventional statistical features, Tariq et al. (2022) introduced a feature where the ratio of the standard deviation of signals at multiple time instances over a longer duration (e.g., 1 h) to the standard deviation of known normal signals is calculated. This feature typically exhibits higher values during leakage events. While not providing definitive results, the authors indicate its potential superiority in improving leakage detection performance over various other time and frequency domain features. In recent years, scholars have also drawn from research findings in signal processing and speech recognition domains, assuming that acoustic leakage signals result from the convolution of acoustic vibrations at leakage points and the resonant response of water-filled pipes. Building upon this, linear prediction techniques can be employed to extract resonance characteristics. Results obtained suggest that, when using a RF model, features based on linear prediction outperform traditional time and frequency statistical features (Cody et al., 2020; Guo et al., 2020).

Feature selection primarily encompasses three methods: filtering, wrapping, and embedding, all of which are applied in the field of data-driven leakage detection. The filtering method evaluates the importance of individual features based on statistical tests. For instance, Yu et al. (2023) used analysis of variance to validate the significance of nine time and frequency domain features they extracted. Cheng and Shen (2022) found that the three important frequency domain features they identified conformed to different data distributions in normal and leakage categories. Regarding the wrapping methods, Tijani et al. (2022) employed the method viewing the feature selection as a search problem, training classifiers with different feature subsets and evaluating the importance of features based on corresponding performance. Finally, the embedding method integrates feature selection as part of the model training process, with the importance of features determined by the algorithm itself. Tree-based classifiers, such as decision trees (DTs) and RFs, fall into this category (Fares et al., 2023; Guo et al., 2020; Shen and Cheng, 2022).

In situations with limited data, combining carefully extracted and selected features with simple models is an effective strategy. Additionally, using manually extracted features gives water utility staff the opportunity to understand which factors are effective for detecting leaks (i.e., enhancing model interpretability). To give staff a clear understanding of how each feature influences a model's output, Xu et al. (2024) used the SHapley Additive exPlanations (SHAP) method to calculate the attribution value of each feature to classifiers' leakage detection results. This value represents the average contribution of the corresponding feature across all possible feature subsets. Since SHAP is model-agnostic, it is recommended to use this method with any machine learning model based on manually extracted features. However, feature extraction and selection also face practical challenges. Extracting valuable features for leakage detection requires researchers to have a solid interdisciplinary foundation in fields such as acoustics, hydraulics, and signal processing. For example, acoustic leakage detection has been studied and applied in WDSs for a long time, but significant features like linear prediction features have only recently gained attention from researchers (Cody et al., 2020; Guo et al., 2020). The difficulty in achieving a deep understanding and comprehension of the data can be addressed to some extent by integrated modeling, which will be introduced in the next section.

3.2.3. Integrated modeling with domain knowledge

In the field of water engineering, relying solely on physics-based models for anomaly detection and optimization may encounter challenges such as unrealistic simulation results due to assumptions and simplifications, as well as high computational costs. On the other hand, solely relying on data-driven models for leakage detection may suffer from data limitations (Fu et al., 2024). Therefore, integrating domain knowledge such as empirical expertise and physical laws into data-driven models (i.e., integrated modeling), can be a more effective approach, reducing the need for large amounts of labeled data. In the domain of data-driven leakage detection, the emerging studies based on integrated modeling can be divided into two distinctive approaches.

The first approach involves combining easily manually extracted features (e.g., skewness and kurtosis) with relatively simple deep learning models, thereby leveraging both domain knowledge and information automatically extracted by deep learning. Wu et al. (2023) constructed a hybrid model that, in addition to using a single-layer CNN with the spectra (obtained by performing the Fourier transform on original signals) as inputs, employed a single-layer fully connected network (FCN) to transform manually extracted features into a new feature space. The concatenation of the outputs of CNN and FCN entered subsequent FCNs for classification. The results obtained show that with the same amount of data, incorporating common time and frequency domain statistical features into the hybrid model yields superior leakage detection performance when compared to a complex CNN classifier and a traditional extreme gradient boosting (XGBoost) classifier that uses 70 elaborately extracted features. Zhang et al. (2023) followed a similar approach, building two CNNs with identical structures but non-shared parameters. Mel-frequency cepstral coefficients (MFCCs) and commonly used time and frequency domain statistical features served as inputs for one CNN, while the raw acoustic signal served as input for the other CNN. Concatenating the features extracted by both CNNs as input to the classifier also achieved better detection results than other CNN classifiers without using manually extracted features. Therefore, when labeled data is relatively limited, and the intrinsic features (e.g., linear prediction features) of the data cannot be fully understood, it could be beneficial to establish such hybrid models.

The second approach involves directly integrating the topological structure or hydraulic characteristics of the analyzed WDS into model construction. In the prediction-classification framework, Fu et al. (2024) employed Graph Neural Networks (GNN) and attention mechanisms to assess the importance of surrounding monitoring points for the target prediction point based on the topological relationships between monitoring points (i.e., computing weights). Scaled data generated based on these weights replaced the original monitoring data as input to the LSTM. The results obtained demonstrated that, with the same amount of data, the flow prediction performance of this method surpasses the corresponding performances of the LSTM and ANN methods. Moreover, the study found that adding more monitoring points did not necessarily lead to better predictions as excessive monitoring point data could contain redundant information. Weyns et al. (2023) employed a neural network structure combining LSTM with Graph Convolutional Networks (GCN) in an AE. This approach considered both the temporal dependencies of monitoring data using LSTM and the spatial correlations between monitoring points due to topological relationships using GCN, enabling more accurate data reconstruction by the AE. This is a representative and instructive example in leakage detection based on reconstruction error analysis. Leveraging energy conservation in the pipeline network, Daniel et al. (2022) constructed a system of linear equations between any two pressure monitoring points using the Bernoulli equation. These equations included a residual term caused by irregular flows (e.g., pump on/off, leaks) between two monitoring points. By constructing the linear equations based on normal data and using least squares, the coefficients of these linear equations could be obtained. When data entered this trained linear prediction model, the cumulative sum (CUSUM) algorithm could determine if the residual significantly exceeded 0, indicating a leak. This method significantly reduces the data volume requirement. Fereidooni et al. (2021), based on hydraulic principles of the pipeline network, computed features beyond flow, such

as head loss and flow velocity. Results showed that combining features such as flow, velocity and head loss effectively enhanced the leakage detection performance. Although it is an attempt to integrate physical laws into machine learning, this method implicitly requires flow meters to be installed at each water usage node, which is evidently impractical and instead increases the data requirements.

3.2.4. Research gaps

In the real world, the amount of manually labeled data is often scarce compared to the entirety of monitoring data. This labeled data reflects the specific differences between normal and leak data, providing valuable deterministic knowledge. Although unlabeled data cannot offer deterministic knowledge, it contains a wealth of information (van Engelen and Hoos, 2020). However, current strategies aimed at reducing data requirements from a modeling perspective often suffer from a one-size-fits-all problem, resulting in wasted data resources. When using unlabeled or normal data directly for anomaly detection, a considerable amount of deterministic knowledge contained in a small portion of labeled data is discarded. Similarly, when developing supervised classification models using feature extraction and selection and integrated modeling, only a small portion of labeled data is utilized, leading to the neglect of vast amounts of information in unlabeled data. Therefore, a significant research gap is the lack of development of semi-supervised classification methods that can fully use all data resources from both labeled and unlabeled data.

Furthermore, there is also a research gap in the lack of effective strategies to address potential new anomaly patterns that leakage detection models may encounter in practice (e.g., gradually developing small leaks for a burst detection model, leakage signals in plastic pipes for a leakage detection model mainly trained using signals from metal pipes). In fact, the emergence of new anomalies is a derivative issue stemming from class imbalance and labeling difficulty. Identifying a wide range of anomaly categories preemptively is challenging due to their infrequent occurrence and the difficulty in accurately labeling them. Initially, when establishing a leakage detection model, certain anomaly categories may be scarce or may not even be recognized. Consequently, the model may only perform well in detecting a few types of anomalies, such as bursts and leakage signals in mental pipes. This research gap not only exists in unsupervised anomaly detection and semi-supervised anomaly detection methods as described in Section 3.2.1, but also in supervised classification methods. Hashim et al. (2020) observed a significant decrease in SVM classifier's leakage detection accuracy when new flow data differed significantly from training samples.

3.3. Dealing with data limitations by transferring knowledge from other domains

As mentioned earlier, collecting large amounts of training data for each WDS is often costly, time-consuming, and, in many cases, unrealistic. Inspired by humans' ability to transfer knowledge across domains, transfer learning aims to leverage knowledge from related domains (referred to as source domains) to improve learning performance in the target domain, minimizing the amount of labeled data required in the target domain (Zhuang et al., 2021). Knowledge transfer can be divided into data-based and model-based depending on the object of transfer. Although still limited, both have found applications in the field of leakage detection.

3.3.1. Data-based knowledge transfer

Data-based knowledge transfer migrates knowledge by adjusting and transforming data, primarily aiming to reduce distribution differences between source and target domain data. Although hydraulic models can generate large amounts of data, especially leakage samples, there are still differences in characteristics between these simulated data and real monitoring data (Basnet et al., 2023). Moreover, collecting a large number of leakage samples in real networks is impractical. Considering this, Wang et al. (2023) used a projection matrix to project the source domain (hydraulic model simulated data) into the feature space of the target domain (real monitoring data). They utilized maximum mean discrepancy to minimize distribution differences between the two domains, ultimately solving for the optimal projection matrix. Using the transformed data for classifier training, the detection accuracy of the optimal SVM classifier increased by more than 40 % on average compared to only using real monitoring data, effectively addressing the class imbalance issue. Although this method appears promising, it is currently supported by only one study and requires further exploration and validation.

3.3.2. Model-based knowledge transfer

The main goal of model-based knowledge transfer is to use a model trained in the source domain to make accurate predictions or classifications in the target domain. The model's parameters reflect the knowledge it has learned, allowing knowledge transfer at the parameter level (Zhuang et al., 2021). In the field of data-driven leakage detection, one well-known method is parameter sharing. More specifically, after obtaining a pre-trained model in the source domain, the neural network layers used for feature extraction can be frozen (i.e., sharing the parameters optimized in the source domain), and only the layers used for classification or prediction are optimized using the limited data from the target domain. To further enhance the performance, the entire new model can be fine-tuned with a very small learning rate (Chollet, 2023). This approach enables the utilization of the automatic feature extraction capabilities of complex deep learning models under limited data conditions.

In acoustic leakage detection, CNN classifiers are commonly used, with two-dimensional spectrograms (obtained through techniques like short-time Fourier transform) as inputs. Zhang et al. (2022) transferred the CNN model VGGish (Hershey et al., 2017) from the audio classification field to acoustic leakage detection. In this case, both the source and target domain tasks involve time-frequency spectrogram classification, making the knowledge transfer rational. Although the leakage detection accuracy reached approximately 90 %, VGGish exhibited overfitting (i.e., the accuracy on the training dataset was significantly higher than on the validation and test datasets), likely due to the small number of training samples (fewer than 1000) and the lack of fine-tuning. Some researchers have attempted to directly use well-established CNN architectures from the field of computer vision, such as AlexNet (Krizhevsky et al., 2012) and ResNet (He et al., 2016). However, considering the significant differences between image recognition and leakage detection tasks, these studies did not employ knowledge transfer but instead trained the models from scratch using their own collected data (Peng et al., 2024; Shukla and Piratla, 2020).

In addition to transferring pre-trained models from related domains, researchers have also constructed deep learning models using relatively abundant real monitoring data from WDSs, then transferred them to new WDSs or new monitoring points with limited data. Guo et al. (2021) trained a time-frequency CNN (TFCNN) using over 30,000 acoustic signals (including augmented data), with the model inputs being three spectrograms of different time-frequency resolutions. When the TFCNN was applied to the WDS of another city through transfer learning, a detection accuracy similar to that of the original WDS achieved using only a few hundred acoustic signals. To address the issue of quickly building new models for newly added monitoring points in the prediction-classification method, Glynis et al. (2023) established LSTM models for existing monitoring points, training them with several years of extensive monitoring data. When new monitoring points were added, the existing LSTM models were directly transferred and fine-tuned using a small amount of new data collected over one month, significantly reducing the data requirements.

3.3.3. Research gap

Currently, compared to the significant success of knowledge transfer in fields like computer vision, related research in the field of data-driven leakage detection is very limited, with only 4 out of 85 reviewed articles addressing the topic. Existing methods developed are relatively simplistic, mainly using parameter sharing, and have not adopted more advanced techniques from the transfer learning field. Additionally, there is a lack of focus on negative transfer (i.e., poor knowledge transfer performance due to limited similarity between domains). The cause of negative transfer can be attributed to significant differences between the source and target domain data (Wang et al., 2019). Although there is no current research discussing negative transfer in the leakage detection field, this issue has been observed in a similar task of water level anomaly detection. Nicholaus et al. (2022) attempted to transfer the complex ResNet model from the computer vision field, but its detection performance was worse than that of a simple four-layer CNN trained from scratch using limited water level data. Overall, a current research gap is the lack of research and application of advanced knowledge transfer techniques with the consideration of negative transfer issue.

4. Future research recommendations

Anomaly detection, such as leakage detection, is a challenge faced across various fields, including contamination detection in WDSs, industrial defect identification and financial fraud detection. Researchers working on leakage detection should maintain an open attitude toward anomaly detection methods from different domains and develop more effective approaches tailored to the characteristics of leakage events. For example, while unsupervised methods for leakage detection primarily rely on traditional techniques like SPC and clustering, researchers in the field of WDS contamination detection have developed unsupervised methods based on GANs (Li et al., 2023). Additionally, while leakage detection models often use single machine learning models, Li et al. (2022) found that using a meta-model to integrate the results of multiple base models can improve the accuracy and reliability of contamination event detection. Due to space limitations, this review cannot cover all anomaly detection methods and their respective advantages and disadvantages across different fields. Instead, it will primarily address existing research gaps in leakage detection and provide recommendations for future research.

4.1. Further exploration of data augmentation

This section addresses the research gap of the fact that research on data augmentation for time series is still insufficient and immature. When it is not possible to obtain a comprehensive new dataset and experimental/simulated data cannot accurately replicate the characteristics of real monitoring data, one possible solution is to use data augmentation for time series data such as flow, pressure and acoustic signals. This way the amount of observed data can be increased whilst preserving the characteristics of original data (i.e., preserving the same data labels with original data).

Applying data augmentation to time series data, as in leakage detection, poses unique challenges. The dynamic nature of these time series datasets, combined with sensitivity to environmental interference and diversity of monitored targets, requires more nuanced approaches. Not all random transformations effective for images (e.g., adding noise, cropping, and rotating) are applicable to time series data, as they may inadvertently alter the fundamental characteristics of the observed signal, leading to misclassification or information loss (Iwana and Uchida, 2021). In recent studies, researchers have found that adding noise to acoustic signals as a data augmentation method can sometimes even reduce the effectiveness of leakage detection (Liu et al., 2024; Wu et al., 2024). Therefore, it is urgent to investigate which random transformation methods are effective for time series data (both hydraulic data and acoustic signals) in leakage detection. More importantly, these studies need to provide the rationale for the effectiveness of these random transformations, ensuring interpretability (i.e., explaining why the diversity of the data is increased and why the new data's labels remain consistent with original data). Additionally, since data augmentation techniques based on random transformations typically include some manually set parameters (such as a scaling factor to change the signal amplitude), future research can refer to Hataya et al. (2020) and use reinforcement learning techniques to achieve automatic selection of these parameters.

In addition to random transformations, deep learning-based data augmentation techniques are also worth studying. As mentioned earlier, Liu et al. (2024) have already made an attempt and improved leakage detection performance. However, future research needs to consider the following two points. First, deep learning methods like GANs may require a relatively large dataset to avoid the issue of mode collapse, i.e., generating data that is very similar to the original samples but lacks diversity (Saxena and Cao, 2021). For acoustic leakage detection, many studies only have a few hundred available original signals. Perhaps accumulating more data would better leverage the potential of these deep learning methods. Second, methods based on deep learning typically have relatively poorer interpretability compared to methods based on random transformations, which necessitates further exploration and improvement in the future.

4.2. Development of semi-supervised classification methods

This section addresses the research gap of the lack of development of semi-supervised classification methods that can fully use all data resources from both labeled and unlabeled data. Semi-supervised classification methods can effectively address data limitations because they require only a small amount of labeled data. Before method development, specific areas can be selected to deploy sensors (e.g., pressure gauges and accelerometers) and collect unlabeled data containing relatively rich leakage information without the need for corresponding personnel to excavate pipelines for verification. These areas may include DMAs with poor pipeline quality, long service life, or high night flow rates. Such unlabeled datasets and existing labeled data will provide ample support for the development of semi-supervised classification models.

Semi-supervised classification can be broadly categorized into two approaches: inductive and transductive (van Engelen and Hoos, 2020). The inductive approach learns the differences between leakage data and normal data from all available data resources and generalizes the leakage detection capability to new data. This approach encompasses various methods, such as using one or multiple base classifiers and iteratively retraining these learners with original labeled data and pseudo-labeled data (derived from previous iterations of classifiers). For instance, an ANN classifier initially trained using labeled data can output probabilities indicating whether unlabeled data belongs to normal or leakage categories. Data with high probabilities (e.g., above 0.8) are assigned pseudo-labels and incorporated into the labeled dataset for retraining the ANN. This iterative process continues and could improve leakage detection performance. Some methods directly incorporate unlabeled data into the objective function or optimization process. For instance, Zhao et al. (2019) incorporated both labeled and unlabeled data loss terms into the loss function of a SVM model when predicting the susceptibility to urban flooding.

The transductive approach focuses on predicting labels for given unlabeled data and cannot be applied to new data. A typical representative of this approach is the graph-based label propagation method. This method first constructs a graph where nodes represent all available data, and edges between nodes signify relationships based on data similarity (measured using metrics like Euclidean distance). Then, based on the smoothness assumption of the graph, it propagates label information (e.g., normal or leakage) to unlabeled nodes (data), ensuring that similar nodes (data) have the same label. Since it is only applicable to

existing data, the transductive approach can also be used as a method to generate accurate pseudo-labels. Subsequently, classifiers can be trained using both the original labeled data and pseudo-labeled data to detect leaks.

4.3. Development of multi-classification methods with model updating mechanisms

This section addresses the research gap of the lack of effective strategies to address potential new anomaly patterns that leakage detection models may encounter in practice. Although scholars have developed numerous methods for leakage detection, treating it as either anomaly detection or classification tasks, as discussed earlier, these methods may still face challenges in dealing with new anomaly patterns. This issue may be more severe in acoustic leakage detection, where various factors like the state of leaks, pipe properties, sensor installation distance, and environmental interference, can cause changes in acoustic signals, leading to the deterioration of existing model detection performance (Bakhtawar and Zayed, 2021).

To address the aforementioned issue, future research should focus more on multi-classification models. For instance, acoustic signals could be classified into several classes including background noise, metal pipe leaks, plastic pipe leaks and environmental interference. After implementing multi-classification, a similarity evaluation module could be established to assist in model updating. This module would compute the similarity between newly identified signal belonging to a certain class and the labeled signals used to construct the model. Similarity can be measured using the distance between original signals or the distance between features extracted from the original signals. If the new signal's similarity with each class is low, a potential new type of anomaly may occur. Upon manual verification, if it is confirmed that the new signal represents a new type of anomaly, a new class can be established. The multi-classification model can then be updated as more data belonging to the new anomaly type becomes available. This approach enables the continuous improvement of leakage detection models, preventing them from being overwhelmed by new anomaly types. Zhang et al. (2022) established a shared-parameter Siamese CNN network to compare the similarity between two input samples. Each network independently processes one input and generates a feature vector. Then, the Euclidean distance between the feature vectors is used to assess the similarity between the new data and the known labeled data. This provides insights into constructing a similarity evaluation module.

4.4. Exploration of new ways to transfer knowledge

This section addresses the research gap of the lack of research and application of advanced knowledge transfer techniques. This review does not advocate blindly pursuing the use of transfer learning techniques for leakage detection research. After all, there are significant differences between data from different domains, such as the substantial gap between audio signals in media and acoustic signals in WDSs. This introduces the risk of negative transfer as mentioned earlier. Therefore, this paper offers the following two recommendations.

Firstly, it is advisable to select source domain models that are closely related to the leakage detection task for transfer. For example, building a model similar to AlexNet and ResNet in the field of data-driven leakage detection, supported by abundant data and techniques like data augmentation, would serve as an optimal choice for all water utilities, ensuring a sufficiently high lower bound on knowledge transfer effectiveness.

Secondly, a combination of data-driven and model-driven knowledge transfer approaches can be attempted. Zhuang et al. (2017) proposed a generic deep learning-based transfer learning framework employing two identically structured AEs with shared parameters to process data from both the source and target domains. These AEs encode the source and target domain data and reconstruct them through decoders. In addition to the reconstruction error of AEs and classification error of source domain data, the framework's loss function incorporates a crucial component. This component is the discrepancy between the encoded source and target domain data (measured by Kullback–Leibler divergence), which reflects the data-driven knowledge transfer. The encoder trained in the AE can be directly transferred and applied to classify target domain data, which reflects the model-based knowledge transfer through parameter sharing. Data-driven knowledge transfer helps align the data distribution of the source domain data (e.g., abundant acoustic signals from a particular WDS or abundant hydraulic simulated flow data) with that of the target domain data (e.g., limited acoustic signals from a new WDS or limited flow data from a real-world WDS), thereby mitigating the occurrence of negative transfer in the model-based knowledge transfer.

In the current research state, where much emphasis is placed on developing various models using simulated or experimental data without adequate study of the models' application effectiveness in real-world WDSs, the development of knowledge transfer techniques can contribute to a more comprehensive research framework. Researchers can develop models using simulated or experimental data and then employ knowledge transfer techniques to transfer these models to a WDS with limited data, thereby validating their practical application effectiveness. It is believed that with such a research framework emphasizing both model development and practical application, the data-driven leakage detection technology will mature more rapidly.

5. Conclusion

To address the data limitations in WDS leakage detection, i.e., class imbalance and labeling difficulty, this paper reviews 85 journal articles that focus on data-driven leakage detection methods, along with 3 articles introducing publicly available leakage detection datasets. The review categorizes the methods based on their data utilization into three main categories: unsupervised anomaly detection, semi-supervised anomaly detection, and supervised classification. The following conclusions are drawn:

1) Strategies to address data limitations in leakage detection include data creation, data requirement reduction, and knowledge transfer. These strategies create conditions for developing supervised classification methods that require labeled and balanced datasets, making such methods mainstream. Unsupervised anomaly detection and semi-supervised anomaly detection methods are mainly attributed to the modeling using unlabeled or normal data, with the latter being favored by many researchers due to its relatively low data requirements and relatively good performance.
2) In the data creation strategy, experiment and simulation-based data generation methods remain predominant due to their ability to easily create labeled and balanced datasets with adequate leakage and normal data. However, continued research on data augmentation techniques is necessary to ensure the generated data better reflect real-world conditions.
3) In the data requirement reduction strategy, leakage detection can be treated as an anomaly detection task, directly using unlabeled data or ignoring labeled leakage samples. Alternatively, it can be viewed as a classification task, reducing the total need for labeled data through feature extraction and selection and integrated modeling. However, semi-supervised classification methods that can utilize both a small amount of labeled data and a large amount of unlabeled data have not yet received attention and warrant further research.
4) The application of the knowledge transfer strategy in leakage detection is still relatively rare, primarily involving model-based techniques such as parameter sharing and fine-tuning. To avoid negative transfer, this paper suggests focusing on advanced techniques that combine data-based and model-based knowledge

transfer. Additionally, establishing a general pre-trained model for leakage detection, usable by various water utilities, is recommended.
5) Future research should place greater emphasis on handling new anomalies that a data-driven model may encounter in practice (e.g., gradually developing small leaks for a burst detection model). Development of multi-classification models with a similarity evaluation module is recommended to confirm new anomaly occurrences, facilitating ongoing model updates and ensuring high leakage detection performance.

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

During the preparation of this work the author(s) used ChatGPT 4o in order to improve language and readability. After using this tool/service, the author(s) reviewed and edited the content as needed and take(s) full responsibility for the content of the publication.

CRediT authorship contribution statement

Yipeng Wu: Writing – review & editing, Writing – original draft, Visualization, Methodology, Investigation, Funding acquisition, Formal analysis, Conceptualization. **Shuming Liu:** Writing – review & editing, Supervision, Resources, Conceptualization. **Zoran Kapelan:** Writing – review & editing, Supervision, Methodology.

Declaration of competing interest

The authors declare the following financial interests/personal relationships which may be considered as potential competing interests:

Yipeng Wu reports financial support was provided by National Natural Science Foundation of China. If there are other authors, they declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Data availability

No data was used for the research described in the article.

Acknowledgments

This work was financially supported by National Natural Science Foundation of China (Grant no. 52300119).

Supplementary materials

Supplementary material associated with this article can be found, in the online version, at doi:10.1016/j.watres.2024.122471.

References

Aghashahi, M., Sela, L., Banks, M.K., 2023. Benchmarking dataset for leak detection and localization in water distribution systems. Data Brief 48, 109148.
Ahn, J., Jung, D., 2019. Hybrid statistical process control method for water distribution pipe burst detection. J. Water Resour. Plan. Manag. - ASCE 145 (9), 6019008.
Awwad, A., Albasha, L., Mir, H.S., Mortula, M.M., 2023. Employing robotics and deep learning in underground leak detection. IEEE Sens. J. 23 (8), 8169–8177.
Ayati, A.H., Haghighi, A., 2023. Multiobjective wrapper sampling design for leak detection of pipe networks based on machine learning and transient methods. J. Water Resour. Plan. Manag. - ASCE 149 (2).
Bakhtawar, B., Zayed, T., 2021. Review of water leak detection and localization methods through hydrophone technology. J. Pipel. Syst. Eng. Pract. 12 (4), 3121002.
Basnet, L., Brill, D., Ranjithan, R., Mahinthakumar, K., 2023. Supervised machine learning approaches for leak localization in water distribution systems: impact of complexities of leak characteristics. J. Water Resour. Plan. Manag. - ASCE 149 (8), 4023032.
Blázquez-García, A., Conde, A., Mori, U., Lozano, J.A., 2021. Water leak detection using self-supervised time series classification. Inf. Sci. 574, 528–541.
Bohorquez, J., Lambert, M.F., Alexander, B., Simpson, A.R., Abbott, D., 2022. Stochastic resonance enhancement for leak detection in pipelines using fluid transients and convolutional neural networks. J. Water Resour. Plan. Manag. - ASCE 148 (3), 4022001.
Bohorquez, J., Simpson, A.R., Lambert, M.F., Alexander, B., 2020. Merging fluid transient waves and artificial neural networks for burst detection and identification in pipelines. J. Water Resour. Plan. Manag. - ASCE 147 (1), 4020097.
Bozkurt, C., Firat, M., Ates, A., 2022. Development of a new comprehensive framework for the evaluation of leak management components and practices. AQUA 71 (5), 642–663.
Bykerk, L., Valls Miro, J., 2022. Detection of water leaks in suburban distribution mains with lift and shift vibro-acoustic sensors. Vibration 5 (2), 370–382.
Cheng, W., Shen, Y., 2022. Frequency characteristic analysis of acoustic emission signals of pipeline leakage. Water 14 (24), 3992.
Chollet, F., 2023. Complete guide to transfer learning & fine-tuning in Keras.
Cody, R., Harmouche, J., Narasimhan, S., 2018. Leak detection in water distribution pipes using singular spectrum analysis. Urban Water J. 15 (7), 636–644.
Cody, R.A., Dey, P., Narasimhan, S., 2020. Linear prediction for leak detection in water distribution networks. J. Pipel. Syst. Eng. Pract. 11 (1), 4019043.
Cody, R.A., Tolson, B.A., Orchard, J., 2020. Detecting leaks in water distribution pipes using a deep autoencoder and hydroacoustic spectrograms. J. Comput. Civ. Eng. 34 (2), 4020001.
Daniel, I., Pesantez, J., Letzgus, S., Khaksar, F.M.A., Alghamdi, F., Berglund, E., Mahinthakumar, G., Cominola, A., 2022. A sequential pressure-based algorithm for data-driven leakage identification and model-based localization in water distribution networks. J. Water Resour. Plan. Manag. - ASCE 148 (6), 4022025.
Eggimann, S., Mutzner, L., Wani, O., Schneider, M.Y., Spuhler, D., Moy De Vitry, M., Beutler, P., Maurer, M., 2017. The potential of knowing more: a review of data-driven urban water management. Environ. Sci. Technol. 51 (5), 2538–2553.
El-Zahab, S., Zayed, T., 2019. Leak detection in water distribution networks: an introductory overview. Smart Water 4 (1), 5.
Fan, H., Tariq, S., Zayed, T., 2022. Acoustic leak detection approaches for water pipelines. Autom. Constr. 138, 104226.
Fan, X., Yu, X.B., 2022. An innovative machine learning based framework for water distribution network leakage detection and localization. Struct. Health Monit. 21 (4), 1626–1644.
Fares, A., Tijani, I.A., Rui, Z., Zayed, T., 2023. Leak detection in real water distribution networks based on acoustic emission and machine learning. Environ. Technol. 44 (25), 3850–3866.
Fereidooni, Z., Tahayori, H., Bahadori-Jahromi, A., 2021. A hybrid model-based method for leak detection in large scale water distribution networks. J. Ambient Intell. Humaniz. Comput. 12 (2), 1613–1629.
Fox, S., Shepherd, W., Collins, R., Boxall, J., 2016. Experimental quantification of contaminant ingress into a buried leaking pipe during transient events. J. Hydraul. Eng. 142 (1), 4015036.
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.
Fu, G., Savic, D., Butler, D., 2024. Making waves: towards data-centric water engineering. Water Res. 256, 121585.
Fu, M., Zhang, Q., Rong, K., Yaseen, Z.M., Zheng, L., Zheng, J., 2024. Integrated dynamic multi-threshold pattern recognition with graph attention long short-term neural memory network for water distribution network losses prediction: an automated expert system. Eng. Appl. Artif. Intell. 127, 107277.
Glynis, K., Kapelan, Z., Bakker, M., Taormina, R., 2023. Leveraging transfer learning in LSTM neural networks for data-efficient burst detection in water distribution systems. Water Resour. Manag. 37 (15), 5953–5972.
Guo, G., Yu, X., Liu, S., Ma, Z., Wu, Y., Xu, X., Wang, X., Smith, K., Wu, X., 2021. Leakage detection in water distribution systems based on time–frequency convolutional neural network. J. Water Resour. Plan. Manag. - ASCE 147 (2), 4020101.
Guo, G., Yu, X., Liu, S., Xu, X., Ma, Z., Wang, X., Huang, Y., Smith, K., 2020. Novel leakage detection and localization method based on line spectrum pair and cubic interpolation search. Water Resour. Manag. 34 (12), 3895–3911.
Guo, X., Song, H., Zeng, Y., Chen, H., Hu, W., Liu, G., 2024. An intelligent water supply pipeline leakage detection method based on SV-WTBSVM. Meas. Sci. Technol. 35 (4), 46125.
Guyon, I., Gunn, S., Nikravesh, M., Zadeh, L.A., 2008. Feature Extraction: Foundations and Applications. Springer.
Hashim, H., Ryan, P., Clifford, E., 2020. A statistically based fault detection and diagnosis approach for non-residential building water distribution systems. Adv. Eng. Inform. 46, 101187.
Hataya, R., Zdenek, J., Yoshizoe, K., Nakayama, H., 2020. Faster AutoAugment: learning augmentation strategies using backpropagation. Vedaldi, A., Bischof, H., Brox, T. and Frahm, J. (Eds.), pp. 1–16, Springer International Publishing, Cham.
He, K., Zhang, X., Ren, S., Sun, J., 2016. Deep residual learning for image recognition, pp. 770–778.
Hershey, S., Chaudhuri, S., Ellis, D.P., Gemmeke, J.F., Jansen, A., Moore, R.C., Plakal, M., Platt, D., Saurous, R.A., Seybold, B., 2017. CNN Architectures for Large-Scale Audio Classification. IEEE, pp. 131–135.
Hu, Z., Chen, W., Wang, H., Tian, P., Shen, D., 2022. Integrated data-driven framework for anomaly detection and early warning in water distribution system. J. Clean. Prod. 373, 133977.
Huang, P., Zhu, N., Hou, D., Chen, J., Xiao, Y., Yu, J., Zhang, G., Zhang, H., 2018. Real-time burst detection in district metering areas in water distribution system based on patterns of water demand with supervised learning. Water 10 (12), 1765.

Islam, M.R., Azam, S., Shanmugam, B., Mathur, D., 2022. A review on current technologies and future direction of water leakage detection in water distribution network. IEEE Access 10, 107177–107201.
Iwana, B.K., Uchida, S., 2021. An empirical survey of data augmentation for time series classification with neural networks. PLoS One 16 (7), e254841.
Jernigan, W., 2024. The hidden link: water leakage, carbon emissions & climate change.
Jian, C., Gao, J., Xu, Y., 2022. Anomaly detection and classification in water distribution networks integrated with hourly nodal water demand forecasting models and feature extraction technique. J. Water Resour. Plan. Manag. - ASCE 148 (11), 4022059.
Jung, D., Kang, D., Liu, J., Lansey, K., 2015. Improving the rapidity of responses to pipe burst in water distribution systems: a comparison of statistical process control methods. J. Hydroinform. 17 (2), 307–328.
Kammoun, M., Kammoun, A., Abid, M., 2022. Leak detection methods in water distribution networks: A comparative survey on artificial intelligence applications. J. Pipel. Syst. Eng. Pract. 13 (3), 4022024.
Kammoun, M., Kammoun, A., Abid, M., 2023. LSTM-AE-WLDL: unsupervised LSTM auto-encoders for leak detection and location in water distribution networks. Water Resour. Manag. 37 (2), 731–746.
Kang, J., Park, Y., Lee, J., Wang, S., Eom, D., 2018. Novel leakage detection by ensemble CNN-SVM and graph-based localization in water distribution systems. IEEE Trans. Ind. Electron. 65 (5), 4279–4289.
Kim, S., Jun, S., Jung, D., 2022. Ensemble CNN model for effective pipe burst detection in water distribution systems. Water Resour. Manag. 36 (13), 5049–5061.
Kingdom, B., Soppe, G., Sy, J., 2016. What is non-revenue water? How can we reduce it for better water service?
Kirstein, J.K., Hogh, K., Rygaard, M., Borup, M., 2019. A semi-automated approach to validation and error diagnostics of water network data. Urban Water J. 16 (1), 1–10.
Kong, Q., Cao, Y., Iqbal, T., Wang, Y., Wang, W., Plumbley, M.D., 2020. PANNs: large-scale pretrained audio neural networks for audio pattern recognition. IEEE/ACM Trans. Audio Speech. Lang. Process. 28, 2880–2894.
Krizhevsky, A., Sutskever, I., Hinton, G.E., 2012. ImageNet classification with deep convolutional neural networks.
Lee, C.W., Yoo, D.G., 2021. Development of leakage detection model and its application for water distribution networks using RNN-LSTM. Sustainability 13 (16), 9262.
Lee, S.J., Lee, G., Suh, J.C., Lee, J.M., 2016. Online burst detection and location of water distribution systems and its practical applications. J. Water Resour. Plan. Manag. - ASCE 142 (1), 4015033.
Leite, R., Amado, C., Azeitona, M., 2024. Online burst detection in water distribution networks based on dynamic shape similarity measure. Expert Syst. Appl. 248, 123379.
Li, Z., Liu, H., Zhang, C., Fu, G., 2023. Generative adversarial networks for detecting contamination events in water distribution systems using multi-parameter, multi-site water quality monitoring. Environ. Sci. Ecotechnol. 14, 100231.
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.
Liu, R., Zayed, T., Xiao, R., 2024. Advanced acoustic leak detection in water distribution networks using integrated generative model. Water Res. 254, 121434.
Mashhadi, N., Shahrour, I., Attoue, N., El Khattabi, J., Aljer, A., 2021. Use of machine learning for leak detection and localization in water distribution systems. Smart Cities 4 (4), 1293–1315.
McMillan, L., Fayaz, J., Varga, L., 2023. Flow forecasting for leakage burst prediction in water distribution systems using long short-term memory neural networks and Kalman filtering. Sustain. Cities Soc. 99, 104934.
McMillan, L., Fayaz, J., Varga, L., 2024. Domain-informed variational neural networks and support vector machines based leakage detection framework to augment self-healing in water distribution networks. Water Res. 249, 120983.
Menapace, A., Zanfei, A., Felicetti, M., Avesani, D., Righetti, M., Gargano, R., 2020. Burst detection in water distribution systems: the issue of dataset collection. Appl. Sci. 10 (22), 8219.
Momeni, A., Piratla, K.R., Anderson, A., Chalil Madathil, K., Li, D., 2023. Stochastic model-based leakage prediction in water mains considering pipe condition uncertainties. Tunn. Undergr. Space Technol. 137, 105130.
Mounce, S.R., Mounce, R.B., Boxall, J.B., 2011. Novelty detection for time series data analysis in water distribution systems using support vector machines. J. Hydroinform. 13 (4), 672–686.
Mounce, S.R., Mounce, R.B., Boxall, J.B., 2012. Identifying sampling interval for event detection in water distribution networks. J. Water Resour. Plan. Manag. - ASCE 138 (2), 187–191.
Muniz Do Nascimento, W., Gomes Jr., L., 2023. Enabling low-cost automatic water leakage detection: a semi-supervised, autoML-based approach. Urban Water J. 20 (10), 1471–1481.
Nicholaus, I.T., Lee, J., Kang, D., 2022. One-class convolutional neural networks for water-level anomaly detection. Sensors 22 (22), 8764.
Okosun, F., Cahill, P., Hazra, B., Pakrashi, V., 2019. Vibration-based leak detection and monitoring of water pipes using output-only piezoelectric sensors. Eur. Phys. J. - Spec. Top. 228 (7), 1659–1675.
Peng, H., Xu, Z., Huang, Q., Qi, L., Wang, H., 2024. Leakage detection in water distribution systems based on logarithmic spectrogram CNN for continuous monitoring. J. Water Resour. Plan. Manag. - ASCE 150 (6), 4024015.
Quiñones-Grueiro, M., Verde, C., Prieto-Moreno, A., Llanes-Santiago, O., 2018. An unsupervised approach to leak detection and location in water distribution networks. Int. J. Appl. Math. Comput. Sci. 28 (2), 283–295.
Ravichandran, T., Gavahi, K., Ponnambalam, K., Burtea, V., Mousavi, S.J., 2021. Ensemble-based machine learning approach for improved leak detection in water mains. J. Hydroinform. 23 (2), 307–323.
Rayaroth, R., Sivaradje, G., 2019. Random bagging classifier and shuffled frog leaping based optimal sensor placement for leakage detection in WDS. Water Resour. Manag. 33 (9), 3111–3125.
Romano, M., Kapelan, Z., Savić, D.A., 2014. Automated detection of pipe bursts and other events in water distribution systems. J. Water Resour. Plan. Manag. - ASCE 140 (4), 457–467.
Saxena, D., Cao, J., 2021. Generative adversarial networks (GANs): challenges, solutions, and future directions. ACM Comput. Surv. 54 (3), 63.
Sen, P., Wang, Y., Xu, F., Wu, Q., 2024. Burst diagnosis multi-stage model for water distribution networks based on deep learning algorithms. Water 16 (9), 1258.
Shen, Y., Cheng, W., 2022. A tree-based machine learning method for pipeline leakage detection. Water 14 (18), 2833.
Shukla, H., Piratla, K., 2020. Leakage detection in water pipelines using supervised classification of acceleration signals. Autom. Constr. 117, 103256.
Soldevila, A., Blesa, J., Jensen, T.N., Tornil-Sin, S., Fernandez-Canti, R.M., Puig, V., 2021. Leak localization method for water-distribution networks using a data-driven model and Dempster–Shafer reasoning. IEEE Trans. Control Syst. Technol. 29 (3), 937–948.
Song, H., Jiang, Z., Men, A., Yang, B., Gutierrez, P.A., 2017. A hybrid semi-supervised anomaly detection model for high-dimensional data. Comput. Intell. Neurosci. 2017, 8501683.
Sun, Q., Zhang, Y., Lu, B., Liu, H., 2022. Flow measurement-based self-adaptive line segment clustering model for leakage detection in water distribution networks. IEEE Trans. Instrum. Meas. 71, 1–13.
Tariq, S., Bakhtawar, B., Zayed, T., 2022. Data-driven application of MEMS-based accelerometers for leak detection in water distribution networks. Sci. Total Environ. 809, 151110.
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.
Tornyeviadzi, H.M., Mohammed, H., Seidu, R., 2023. Robust night flow analysis in water distribution networks: a BiLSTM deep autoencoder approach. Adv. Eng. Inform. 58, 102135.
Tornyeviadzi, H.M., Seidu, R., 2023. Leakage detection in water distribution networks via 1D CNN deep autoencoder for multivariate SCADA data. Eng. Appl. Artif. Intell. 122, 106062.
van Engelen, J.E., Hoos, H.H., 2020. A survey on semi-supervised learning. Mach. Learn. 109 (2), 373–440.
Vrachimis, S.G., Eliades, D.G., Taormina, R., Kapelan, Z., Ostfeld, A., Liu, S., Kyriakou, M., Pavlou, P., Qiu, M., Polycarpou, M.M., 2022. Battle of the leakage detection and isolation methods. J. Water Resour. Plan. Manag. - ASCE 148 (12), 4022068.
Wan, X., Farmani, R., Keedwell, E., 2022. Online leakage detection system based on EWMA-enhanced Tukey method for water distribution systems. J. Hydroinform. 25 (1), 51–69.
Wan, X., Farmani, R., Keedwell, E., 2023. Gradual leak detection in water distribution networks based on multistep forecasting strategy. J. Water Resour. Plan. Manag. - ASCE 149 (8), 4023035.
Wan, X., Kuhanestani, P.K., Farmani, R., Keedwell, E., 2022. Literature review of data analytics for leak detection in water distribution networks: a focus on pressure and flow smart sensors. J. Water Resour. Plan. Manag. - ASCE 148 (10), 3122002.
Wang, C., Xu, Q., Zhou, Y., Qiang, Z., 2022. Research on pipe burst in water distribution systems: knowledge structure and emerging trends. AQUA 71 (12), 1408–1424.
Wang, H., Liu, T., Zhang, L., 2023. Pipeline-burst detection on imbalanced data for water supply networks. Water 15 (9), 1662.
Wang, X., Guo, G., Liu, S., Wu, Y., Xu, X., Smith, K., 2020. Burst detection in district metering areas using deep learning method. J. Water Resour. Plan. Manag. - ASCE 146 (6), 4020031.
Wang, Z., Dai, Z., Póczos, B., Carbonell, J., 2019. Characterizing and avoiding negative transfer, pp. 11293–11302.
Wen, Q., Sun, L., Yang, F., Song, X., Gao, J., Wang, X., Xu, H., 2021. Time series data augmentation for deep learning: a survey. In: International Joint Conferences on Artificial Intelligence Organization.
Weyns, M., Mazaev, G., Vaes, G., Vancoillie, F., De Turck, F., Van Hoecke, S., Ongenae, F., 2023. Leak localization in water distribution networks using GIS-enhanced autoencoders. Urban Water J. 20 (7), 859–881.
Wong, B., Mccann, J.A., 2021. Failure detection methods for pipeline networks: from acoustic sensing to cyber-physical systems. Sensors 21 (15), 4959.
Wu, Y., Liu, S., 2017. A review of data-driven approaches for burst detection in water distribution systems. Urban Water J. 14 (9), 972–983.
Wu, Y., Liu, S., 2020. Burst detection by analyzing shape similarity of time series subsequences in district metering areas. J. Water Resour. Plan. Manag. - ASCE 146 (1), 4019068.
Wu, Y., Liu, S., Smith, K., Wang, X., 2018. Using correlation between data from multiple monitoring sensors to detect bursts in water distribution systems. J. Water Resour. Plan. Manag. - ASCE 144 (2), 4017084.
Wu, Y., Liu, S., Wang, X., 2018. Distance-based burst detection using multiple pressure sensors in district metering areas. J. Water Resour. Plan. Manag. - ASCE 144 (11), 6018009.
Wu, Y., Liu, S., Wu, X., Liu, Y., Guan, Y., 2016. Burst detection in district metering areas using a data driven clustering algorithm. Water Res. 100, 28–37.
Wu, Y., Ma, X., Guo, G., Huang, Y., Liu, M., Liu, S., Zhang, J., Fan, J., 2023. Hybrid method for enhancing acoustic leak detection in water distribution systems: integration of handcrafted features and deep learning approaches. Process Saf. Environ. Protect. 177, 1366–1376.

Wu, Y., Ma, X., Guo, G., Jia, T., Huang, Y., Liu, S., Fan, J., Wu, X., 2024. Advancing deep learning-based acoustic leak detection methods towards application for water distribution systems from a data-centric perspective. Water Res. 261, 121999.
Xu, W., Zhou, X., Xin, K., Boxall, J., Yan, H., Tao, T., 2020. Disturbance extraction for burst detection in water distribution networks using pressure measurements. Water Resour. Res. 56 (5), e2019WR025526.
Xu, Z., Liu, H., Fu, G., Zeng, Y., Li, Y., 2024. Feature selection of acoustic signals for leak detection in water pipelines. Tunn. Undergr. Space Technol. 152, 105945.
Yan, R., Huang, J.J., 2023. Confident learning-based Gaussian mixture model for leakage detection in water distribution networks. Water Res. 247, 120773.
Ye, G., Fenner, R.A., 2014. Weighted least squares with expectation-maximization algorithm for burst detection in U.K. water distribution systems. J. Water Resour. Plan. Manag. - ASCE 140 (4), 417–424.
Yu, T., Chen, X., Yan, W., Xu, Z., Ye, M., 2023. Leak detection in water distribution systems by classifying vibration signals. Mech. Syst. Signal Process. 185, 109810.
Yu, X., Wu, Y., Meng, F., Zhou, X., Liu, S., Huang, Y., Wu, X., 2024. A review of graph and complex network theory in water distribution networks: mathematical foundation, application and prospects. Water Res. 253, 121238.
Zanfei, A., Menapace, A., Brentan, B.M., Righetti, M., Herrera, M., 2022. Novel approach for burst detection in water distribution systems based on graph neural networks. Sustain. Cities Soc. 86, 104090.
Zhang, C., Alexander, B.J., Stephens, M.L., Lambert, M.F., Gong, J., 2022. A convolutional neural network for pipe crack and leak detection in smart water network. Struct. Health Monit. 22 (1), 232–244.
Zhang, P., He, J., Huang, W., Zhang, J., Yuan, Y., Chen, B., Yang, Z., Xiao, Y., Yuan, Y., Wu, C., Cui, H., Zhang, L., 2023. Water pipeline leak detection based on a pseudo-siamese convolutional neural network: integrating handcrafted features and deep representations. Water 15 (6), 1088.
Zhang, X., Chen, N., Sheng, H., Ip, C., Yang, L., Chen, Y., Sang, Z., Tadesse, T., Lim, T.P. Y., Rajabifard, A., Bueti, C., Zeng, L., Wardlow, B., Wang, S., Tang, S., Xiong, Z., Li, D., Niyogi, D., 2019. Urban drought challenge to 2030 sustainable development goals. Sci. Total Environ. 693, 133536.
Zhang, X., Long, Z., Yao, T., Zhou, H., Yu, T., Zhou, Y., 2022. Real-time burst detection based on multiple features of pressure data. Water Sci. Technol. 22 (2), 1474–1491.
Zhang, X., Wu, X., Yuan, Y., Long, Z., Yu, T., 2023. Burst detection based on multi-time monitoring data from multiple pressure sensors in district metering areas. Water Supply 23 (10), 4074–4091.
Zhao, G., Pang, B., Xu, Z.X., Peng, D.Z., Xu, L.Y., 2019. Assessment of urban flood susceptibility using semi-supervised machine learning model. Sci. Total Environ. 659, 940–949.
Zhao, M., Liu, H., Li, G., Zhang, C., 2024. Burst detection in district metering areas using flow subsequences clustering–reconstruction analysis. AQUA—Water infrastructure. Ecosyst. Soc. 73 (5), 853–869.
Zhou, X., Tang, Z., Xu, W., Meng, F., Chu, X., Xin, K., Fu, G., 2019. Deep learning identifies accurate burst locations in water distribution networks. Water Res. 166, 115058.
Zhu, J., Yang, M., Ren, Z.J., 2023. Machine learning in environmental research: common pitfalls and best practices. Environ. Sci. Technol. 57 (46), 17671–17689.
Zhuang, F., Cheng, X., Luo, P., Pan, S.J., He, Q., 2017. Supervised representation learning with double encoding-layer autoencoder for transfer learning. ACM Trans. Intell. Syst. Technol. (Tist) 9 (2), 1–17.
Zhuang, F., Qi, Z., Duan, K., Xi, D., Zhu, Y., Zhu, H., Xiong, H., He, Q., 2021. A comprehensive survey on transfer learning. Proc. IEEE 109 (1), 43–76.