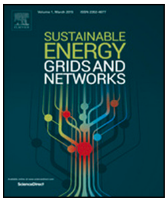
Contents lists available at ScienceDirect

Sustainable Energy, Grids and Networks

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



Leak identification and quantification in gas network using operational data and deep learning framework

Elham Ebrahimi [a,1], Mohammadrahim Kazemzadeh [b,1], Antonio Ficarella [a,*]

[a] *Dept. of Engineering for Innovation, University of Salento, Italy*
[b] *Istituto Italiano di Tecnologia Center for Biomolecular Nanotechnologies, Italy*

ARTICLE INFO

Keywords:
Gas network
Hydrogen network
Green energy
Leak detection
Deep learning
Manifold learning

ABSTRACT

In this study, we introduce an innovative deep learning framework designed to achieve precise detection, localization, and rate estimation of gas distribution pipeline system leakages. Our method surpasses conventional statistical approaches, particularly those based on Bayesian inference, by accommodating the system's intricate behaviors, including variable usage and production from both sources and sinks. Notably, our approach demonstrates remarkable accuracy in localizing leakages even amidst multiple occurrences within the system. Specifically, achieving over 98% accuracy in single-leakage scenarios underscores its effectiveness. Furthermore, through data augmentation involving the introduction of noise into the training dataset, we significantly enhance the model's performance, particularly when tested against real-world-like noisy data. This study not only showcases the efficacy of our proposed deep learning framework but also underscores its adaptability and robustness in addressing complex challenges in gas pipeline systems.

1. Introduction

The continuous expansion of gas networks, spanning thousands of kilometers annually, underscores their crucial role in long-distance transportation and local distribution. Ensuring the integrity and safety of these extensive systems is essential for a reliable gas supply. Gas leaks pose significant risks, including environmental damage, property damage, and personal injury. Effective leak detection mechanisms are indispensable for preventing such failures and preserving the integrity of gas networks.

Advancements in pipeline infrastructure have led to various leak detection techniques. The negative pressure wave technique uses pressure waves generated by leaks to indicate their presence [1]. Acoustic pressure wave detection explores acoustic emissions from leaks [2]. Pressure Point Analysis (PPA) compares current pipeline pressure data to historical records to identify leaks [3]. Wavelet transform technology, combined with an average-weighted localization scheme, is effective in detecting and locating leaks in linear pipelines [4]. Other methods, such as frequency response-based approaches [5,6], mass/volume balance analysis [7], and steady-state or transient models [8], offer additional leak detection and assessment avenues but may struggle with complex pipeline networks.

The interconnected nature of gas distribution networks poses challenges for effective leak detection and localization. Traditional methods, suitable for single-pipe scenarios, are inadequate for these complex networks. Conventional methods, including manual patrolling, are time-consuming and inefficient [9,10]. Recent advancements in sensor technology have led to real-time leak detection systems utilizing distributed temperature sensors [11–13], acoustic sensors [14], and pressure/flow sensors [15]. However, these systems can suffer from false alarms and faulty diagnostics due to sensor noise and environmental interference. Moreover, the need to place sensors at all nodes presents practical challenges.

Deep learning offers a potent solution to the challenges of leak detection and localization in gas networks [16–24], categorized into supervised and unsupervised techniques. Supervised methods leverage extensive datasets encompassing normal and leakage conditions. Recently, some models have integrated artificial intelligence into digital twins of gas networks [25,26], enhancing their capability for leakage detection. A digital twin emulates a physical system using real-time data, simulations, and machine learning to replicate real-world behavior and performance.

However, these digital twin and deep learning-based solutions heavily rely on access to comprehensive datasets containing anomalies

[*] Corresponding author.
E-mail address: antonio.ficarella@unisalento.it (A. Ficarella).
[1] These two author collaborated equally

https://doi.org/10.1016/j.segan.2024.101496
Received 23 April 2024; Received in revised form 20 July 2024; Accepted 11 August 2024
Available online 13 August 2024

indicative of leakage. The scarcity of such anomalous events in operational data makes data acquisition time-consuming and impractical for large-scale networks. Training models on small-scale lab setups poses challenges for real-world applicability due to issues like ambient noise and disturbances.

Unsupervised models train deep learning models to produce output identical to their input, functioning as identity operators. These models rely on the premise that deviations from the trained data represent anomalies [23,24]. While effective in detecting anomalies, these models may misinterpret unusual system operations or natural pipeline aging as anomalies, limiting their real-world applicability.

Researchers have developed mathematical models and probabilistic approaches for leak detection and localization, but these methods often have limitations, such as being restricted to single leaks or requiring sensors at all nodes. The Leak Analytics System (LAS) uses statistical analysis to detect leaks and approximate their location with some accuracy [27]. However, analyzing a small number of potential leak locations limits its applicability, especially in large-scale network failures due to natural hazards.

A critical oversight in both classical and deep learning-based methods is the neglect of the random nature of sinks and sources in real-world gas networks. The variable nature of sinks and sources introduces additional complexity and statistical variations. An unusual increase in sink usage due to extreme weather conditions might be erroneously flagged as a leak by unsupervised techniques. Classical models struggle to consider such factors due to the vast number of simulation scenarios required.

We propose a deep learning digital twin solution tailored for leak detection and localization, accounting for the random nature of sinks and sources in gas networks. Our method operates effectively with only a few sensors in large-scale networks. The training and validation of our method are conducted on simulated scenarios from a complex gas network. This simulation-based approach facilitates the incorporation of rare leakage scenarios and accommodates network changes, such as aging or modifications to the topology.

For comparison with the latest classical models, we chose the gas network featuring 25 sinks, with 6 selected as variable. Our findings showcase significantly more robust performance compared to classical models, even with variable sinks. Our method demonstrates heightened resilience against sensor noise and disturbances, positioning it as a more viable option for real-world applications.

2. Methodology

2.1. Gas network

We investigate by delving into a standardized example network consisting of 38 nodes, 50 pipes, a gas inlet source, and 6 variable sinks, as illustrated in Fig. 1. This particular network has been previously employed in studies focused on optimizing algorithms for gas distribution networks, as referenced in prior works [27]. The pipe diameters within this network span from 4 to 12 inches, and detailed specifications can be found in Supplementary Table 1. Notably, the average distance between adjacent nodes hovers around 100 m, with a minimum distance of approximately 50 m.

The inlet node of this distribution network serves as the gas reduction station or pressure regulator node. This component is responsible for converting high-pressure gas into safe and manageable low-pressure levels. The primary objective of this pressure reduction station is to uphold the outlet pressure at a predefined level, thereby ensuring consistent gas delivery across the entire network. Specifically, our target outlet pressure in this scenario is precisely 5 kPa. Within this network, we explored six potential leak locations, all strategically positioned at the junctions. It is worth noting that junctions and pipe connections emerge as the most probable sites for leakage occurrences, as noted in prior literature [27–30].

2.2. Gas network simulation

The simulations conducted in this study were executed using the Pandapipes package in Python. This package uses the Newton–Raphson method, which adeptly solves for the steady-state flow in gas pipe networks. Automation was employed to simulate various scenarios, and the resultant data has been stored for subsequent multivariate analyses. The Pandapipes package accounts for meteorological conditions such as ambient and gas temperature and pressure, as well as the thermal conductivity and thickness of the pipes. In our simulations, we assumed an ambient temperature of 293 degrees Kelvin at one atmospheric pressure for the surrounding environment, appropriate for above-ground pipes at sea level. This value should be adjusted for underground pipes to account for different environmental conditions. Among all the stored data, only 6 pressure points and 6 flow rates (as illustrated in Fig. 1) have been chosen for the multivariate and deep learning training analysis.

Please note that to accurately determine leak locations in a gas network with “m” variable elements and “n” possible leak sites, it is essential to have at least “m + n” independent sensor readings. This requirement arises from the need to solve a set of nonlinear equations that describe the system’s behavior. Each sensor reading provides an independent equation, and having “m + n” readings ensures that we have sufficient information to determine the “m + n” unknown parameters (i.e., the variable sink usages and potential leak locations). These readings can be any combination of pressure or flow sensors, provided they are independent.

2.3. Data pre-processing

All recorded data underwent a comprehensive data normalization step prior to any multivariate and deep learning analysis. This crucial step involved linearly mapping all sensor readings to set the mean of each sensor reading to zero and standardizing the variance to one, a process known as z-score normalization. Specifically, for each sensor reading x_i, the normalized value z_i was calculated using the formula:

$$z_i = \frac{x_i - \mu}{\sigma} \tag{1}$$

where μ represents the mean of the sensor readings and σ represents the standard deviation. This normalization process is essential due to the significant differences in numerical values between flow meters and pressure meters. Without normalization, these discrepancies can lead to poor model performance, as the varying scales can disproportionately influence the learning process in both multivariate and deep learning models. By ensuring that all sensor readings are on a common scale, the normalization step helps to improve the convergence of optimization algorithms and enhances the generalization capabilities of the models. Additionally, it mitigates the risk of numerical instability and ensures that the models treat all input features with equal importance, thereby facilitating more accurate and robust analysis.

2.4. Deep neural network architectures

The primary goal of this study is to develop a deep learning model capable of identifying one or multiple leakages within a gas network. Our proposed approach involves segregating the task of leakage detection (determining whether there is any leakage) from the processes of localization and rate detection. The rationale behind this separation lies in the enhanced robustness achieved by first establishing the presence or absence of a leakage before delving into its precise localization.

Mathematically, this approach is grounded in probability theory. For instance, if the chance of identifying no leakage at a location is represented by u, then on T possible leak locations, the probability of identifying no leakages is u^T. This value would be considerably smaller than $1 - u^T$, which represents the probability of having at least one location with leakage. This substantial difference in probabilities

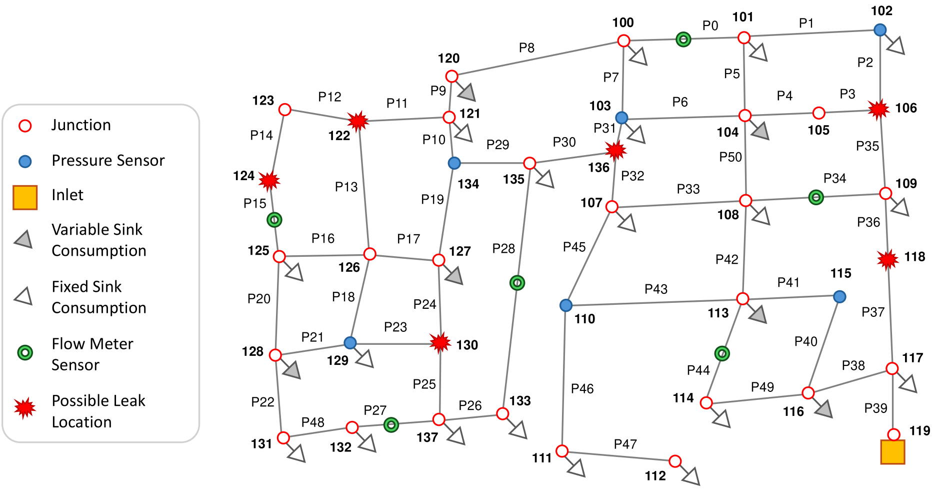


Fig. 1. The schematic of the gas network used in this study. Details such as the diameter and length of each pipe are presented in the supplementary information. The possible leak locations are selected to be the junctions, as these are the most likely sites for leaks [27–30]. The variable sinks are considered to consume gas with rates randomly sampled from a Gaussian distribution with a mean of 0.0075 kg/s and a variance of 0.001 kg/s.

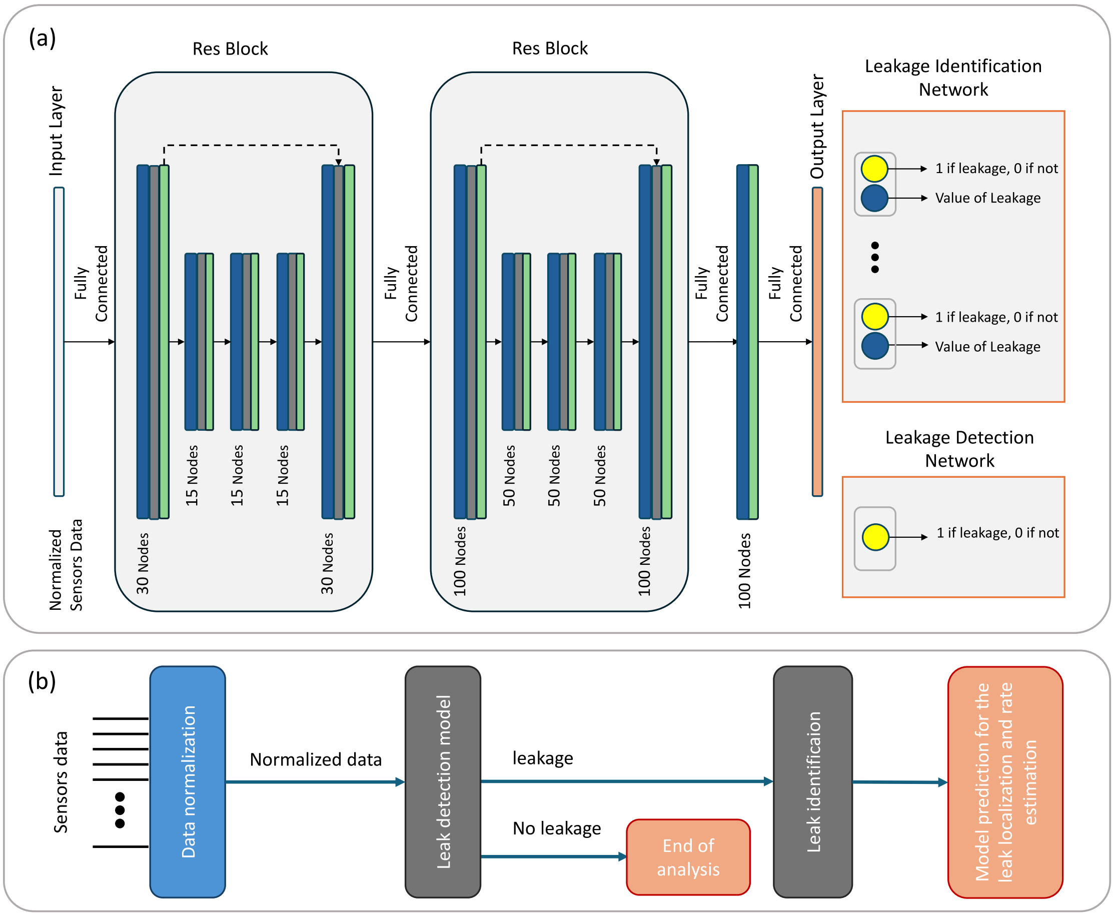


Fig. 2. (a) The neural network schematic features two residual block stacks connected to an input layer, expanding sensor data dimension to 100 nodes. Output nodes include one for binary leak detection, and for leak identification, there are nodes equal to possible leak locations for binary classification, along with nodes for estimating leakage rates. (b) A flowchart illustrating the workflow of the presented deep learning models. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

underscores the challenge of confirming the absence of leakage across all potential leak locations simultaneously. Therefore, by prioritizing the initial investigation of the presence or absence of leakage, we navigate the inherent complexity of verifying a non-leakage scenario across numerous potential locations. This underscores the robustness of our approach, ensuring a more reliable foundation for subsequent localization and rate detection tasks.

The schematic illustrated in Fig. 2 is the architecture of the neural network employed in this study, comprising two stacks of residual blocks connected to the input layer. The input layer serves as the gateway for passing sensor data into the neural network. The output of the residual block is then connected to a layer with 100 nodes, effectively expanding the sensor data dimension to 100 through the nonlinear mapping of the residual blocks. This strategic mapping enhances the network's capability to capture intricate patterns within the sensor data. In This illustration, the blue ribbon shows the nodes while the gray and green ribbons show the batch normalization and Exponential Linear Unit (ELU) activation layer, respectively.

Orange rectangular boxes show the internal structure of the final layer connected to the aforementioned neural network. For leak detection, a single node (yellow) is used for binary classification to determine the presence or absence of a leak in the system. This node is accompanied by a Sigmoid activation function to limit its corresponding output between 0 and 1. Meanwhile, for leakage identification, the network is configured to have output nodes (in yellow color) corresponding to the number of possible leak locations. Each of these nodes is dedicated to binary classification, discerning the existence of leakage at a specific location, similar to the leak detection output, these nodes are followed by a Sigmoid activation function.

Additionally, another set of nodes is incorporated to estimate the leakage rate (blue color), providing a comprehensive understanding of the potential leaks detected by the network. This design ensures the neural network's versatility in simultaneously addressing leak detection, identification, and rate estimation tasks within the gas network.

Please note that although the residual block section shares the same architecture for both the leakage detection and identification networks, they do not share the same weights and biases. Therefore, we are indeed dealing with two independent networks.

In summary, the workflow of the presented method is depicted in the flow chart located at the bottom of Fig. 2. Following the data normalization process outlined in the previous section, the normalized data is directed to the leak detection phase. If no leakage is identified during this initial step, further investigation is unnecessary. However, should a leakage be detected, the normalized data proceeds to the leak identification network. This network not only localizes the leakage but also provides an accurate estimate of its rate. This streamlined process ensures efficient and effective detection and localization of gas distribution system leakages.

2.5. Leak detection

2.5.1. Data

To statistically account for the extensive array of operational conditions in the simulated gas network system, we conducted 20,000 simulation scenarios using the Pandapipes package in Python. This comprehensive dataset comprises 10,000 scenarios representing the normal operational state of the system, devoid of any leakage, and an additional 10,000 scenarios introducing leakages with variable rates ranging from 0.1% to 10% of the total consumption of the network. Specifically, we allocated 50% of the leakage data, equivalent to 5,000 simulation scenarios, to the range of 0.1% to 1%, while the remaining 5,000 scenarios were assigned to the 1% to 10% range of leakage rates. This deliberate distribution emphasizes addressing challenging scenarios where the leakage rate is notably low compared to the overall gas network consumption. Please note that the simulations used in this study are steady-state. This means that all gas outflows, including sinks and leakages, are assumed to have reached equilibrium. Consequently, leakages are modeled as sinks with an unknown flow rate.

For the gas network, we incorporated six variable sinks, with their usages sampled from a Gaussian distribution with a mean of 0.0075 kg/s and a variance of 0.001 kg/s for both normal and leakage datasets. This meticulous approach is pivotal, as variations in sink usages can mimic the behavior of leakages, adding complexity to leakage detection and identification while significantly enhancing the robustness of the detection method.

2.5.2. Leakages statistics

We considered the possibility of all potential locations experiencing leaks in each simulation. The number of leakages for each scenario was randomly selected between one and all possible leak locations. For instance, in the case of our gas network, there is a 1/6 chance of having only one leakage, a 1/6 chance of having two leakages, and so on. This fair selection of the number of leakages aims to create a dataset that encompasses a broad spectrum of leakage scenarios, accounting for various circumstances such as large-scale disasters or aging-related leaks. The probability of each possible leak location being the source of leakage is formulated as follows:

$$p = 1 - \frac{1}{T}\sum_{m}^{T}\frac{{}^{T-1}C_m}{{}^{T}C_m} \qquad p = \frac{T+1}{2T} \tag{2}$$

In the context of our analysis, let T represent the total possible number of leakage points within a given system, and denote m as the actual number of occurrences of leakage. The term ${}^{T}C_m$ signifies the combination of m leakages chosen from the possible T leakage locations. Importantly, ${}^{T-1}C_m$ takes into account combinations where a specific leakage point is excluded from the selection. The division of this term by the total number of combinations (${}^{T}C_m$) provides the probability that, out of the m selected leakage points from the total of T, a particular leakage point is excluded. As the selection of the number of leakage occurrences (m) is a random choice ranging from 1 to T, we can normalize this probability by dividing it by T and summing over all possible values of m. It is crucial to note that, up to this point, we have calculated the probability of excluding one specific location; thus, subtracting this probability from 1 yields the probability of having that specific location as the point of leakage. This equation can be simplified and rewritten, as illustrated at the bottom of Eq. (2). This analytical derivation is used in the next section for dealing with unbalanced data.

2.5.3. Data partitioning strategy and augmentation

To facilitate Neural Network training, we strategically divided the dataset, allocating 80% for training, 10% for validation, and the remaining 10% for testing purposes.

In our pursuit of method robustness and resistance to sensor noise impact, we incorporated a data augmentation step. This involved introducing a Gaussian noise with a variance of .05%, .25%, and .5% percent of the sensor's average reading to both the training and validation datasets, effectively tripling their sizes. The rationale behind this augmentation is twofold. Firstly, during the training phase, we anticipate the neural network learning to effectively ignore noise, thereby enhancing its performance in the presence of such disturbances. Secondly, the expanded dataset resulting from augmentation serves as a powerful tool to counter overfitting, promoting heightened generalization of the network.

2.6. Leak identification

The main difference between the leak identification and leak detection models lies in the composition of their training datasets. The neural network designated for leak identification exclusively operates with leakage data, necessitating 20,000 simulation data points concentrated solely on leakage scenarios. To ensure a fair and uniform distribution of leakage rates for the regression tasks, the leakage rate is uniformly sampled between 0.1% and 10% of the total network consumption. This deliberate selection enhances the model's ability to accurately identify and quantify leakages across a broad spectrum of scenarios.

Please note that the strategies for data augmentation and segmentation are similar between leak detection and identification models. However, the more intricate output layer of the leak identification model necessitates additional attention, as elaborated in the following paragraphs.

2.6.1. Multi-label classification and inherently unbalance data

In the context of the leakage identification network, a critical consideration arises from the potential occurrence of multiple leakages simultaneously, which transforms the problem into a multi-label classification task. This complexity is compounded when ensuring fairness in selecting the number of leakages. As discussed earlier, this inherently leads to unbalanced data for each label, as outlined in Eq. (2). To tackle this challenge within the leak identification model, employing a weighted loss becomes imperative. This approach enables the network to appropriately weigh the impact of different leakage scenarios based on their significance. The weights for the weighted binary cross-entropy loss can be calculated using the probabilities derived in Eq. (2).

On the other hand, it is worth emphasizing that for the leakage detection network, achieving balanced data is more straightforward. This has been accomplished by designing 10,000 scenarios representing normal system operations and an additional 10,000 scenarios introducing leakage. Consequently, a standard binary cross-entropy loss function suffices for training the network in this context.

2.7. Neural network training

We implemented all the presented deep learning models using the TensorFlow open-source package in Python. For training the models, we utilized the Adam optimizer.

During the training process, two callbacks were employed to monitor progress. The first callback monitored the validation loss to save the best model based on validation loss criteria. Meanwhile, the second callback monitored the validation loss. If no improvement occurred after 20 consecutive epochs of training, it triggered a reduction in the learning rate by a factor of 0.1. The initial learning rate was set to be 10^{-3}.

3. Results and discussion

3.1. Leak detection

The learning curve of the leak detection model is depicted in Fig. 3(a1), illustrating the accuracy of both the training and validation data across 100 epochs. Fig. 3(a2) presents the model's accuracy on the testing dataset. Notably, the network demonstrates remarkable accuracy exceeding 99% for leak rates between 1%–10%. Even for leak rates of 0.1%–1%, the network maintains an accuracy of around 95%, showcasing robust performance even with minor leaks. This result is comparable to the findings presented in [27], with the main distinction being that our achieved result accounts for 6 variable sinks with a random consumption pattern.

The confusion matrix for the entire testing dataset, including high and low leakage rates, is displayed in Figures 3(b1)–(b3) respectively. An interesting observation is the minimal false positives observed across all scenarios, with only three instances of normal system states being misclassified as leaks. As previously discussed, this small occurrence may arise from the statistical behavior of variable sinks, which can resemble leak behavior on selected sensors.

For a clearer visualization of the testing dataset's topology, we employed Uniform Manifold Approximation and Projection (UMAP) as a form of unsupervised learning to project the testing data corresponding to low leakage rates and 500 samples of normal system operation into a two-dimensional plane (Fig. 3(c1)). In this projection, normal operation is represented in blue, while leakage is shown in orange. Notably, both normal and leakage states exhibit similar distributions, occupying almost the same area on the projection plane. Variations in the projected points arise from the random nature of variable sinks within the system.

An additional output from the layer connected to the output layer of the model (with 100 nodes in Fig. 2) provides a nonlinear map, mapping input data to a 100-dimensional space. We have also performed the UMAP projection on the result of this nonlinear map, as depicted in Fig. 3(c2). Here, it is evident that the majority of leakage data is distinct from the normal state of the gas network, highlighting the effectiveness of the discovered nonlinear map by the proposed deep learning model in distinguishing leaks from normal system states.

We also investigated the impact of noise on the model's performance by adding varying levels of noise to the testing set. Fig. 3(d) showcases the model's accuracy when evaluated with different levels of noise. For comparison, we conducted the same test without employing the data augmentation strategy discussed in the methods section.

Clearly, data augmentation significantly enhances the model's performance, even in the presence of noisy data, including instances with noise levels four times higher than those present in the augmented training and validation datasets.

3.1.1. Test on the generalization of the model

We further evaluated the model's performance by testing it under normal scenarios (without leakage) where the mean usage of variable sinks deviates significantly from that of the training dataset. Specifically, we considered mean values of 0.0085 kg/s and 0.0065 kg/s, maintaining the variance identical to that of the training data (0.001 kg/s) for high and low sink usage, respectively. Although the sink usage pattern has changed (significantly increased or decreased), these scenarios still represent the normal operation of the system. Therefore, an ideal leakage prediction model should be able to detect them as the normal state of the system.

This test serves to assess the model's ability to discern sink usage patterns from sensor data to detect leaks effectively. Additionally, it evaluates the model's capacity to generalize and make accurate predictions beyond its training dataset.

The statistical distribution of readings from all 12 sensors is illustrated in Fig. 4(a). Here, red denotes higher consumption of variable sinks, while blue and green represent the distribution of training data and lower consumption, respectively. Furthermore, orange depicts the statistical distribution of leakage ranging from 0.1% to 1% in the testing dataset.

An intriguing observation is that the distribution of leakage across these sensors predominantly falls between the consumption levels considered for the training dataset and the higher consumption range. This was expected, as leaks bear closer resemblance to higher sink consumption rather than lower.

To assess the network's ability to detect leaks amidst varying sink consumption levels, we utilized UMAP to project this data before and after network application. The results are depicted in Figures 4 (b1) and (b2) for the original and transformed data, respectively, into 100 dimensions using the proposed model. Notably, Fig. 4(b2) demonstrates that the network successfully generalized the problem, with all normal system states projected closely together on the plane while leakage data was distinctly separated.

For quantitative analysis, we conducted classification over data with lower and higher consumption levels and depicted the resulting accuracy in Fig. 4 (b3). As anticipated, almost all instances of lower consumption were correctly identified as normal system states, with an accuracy rate nearing 97% for instances of higher consumption.

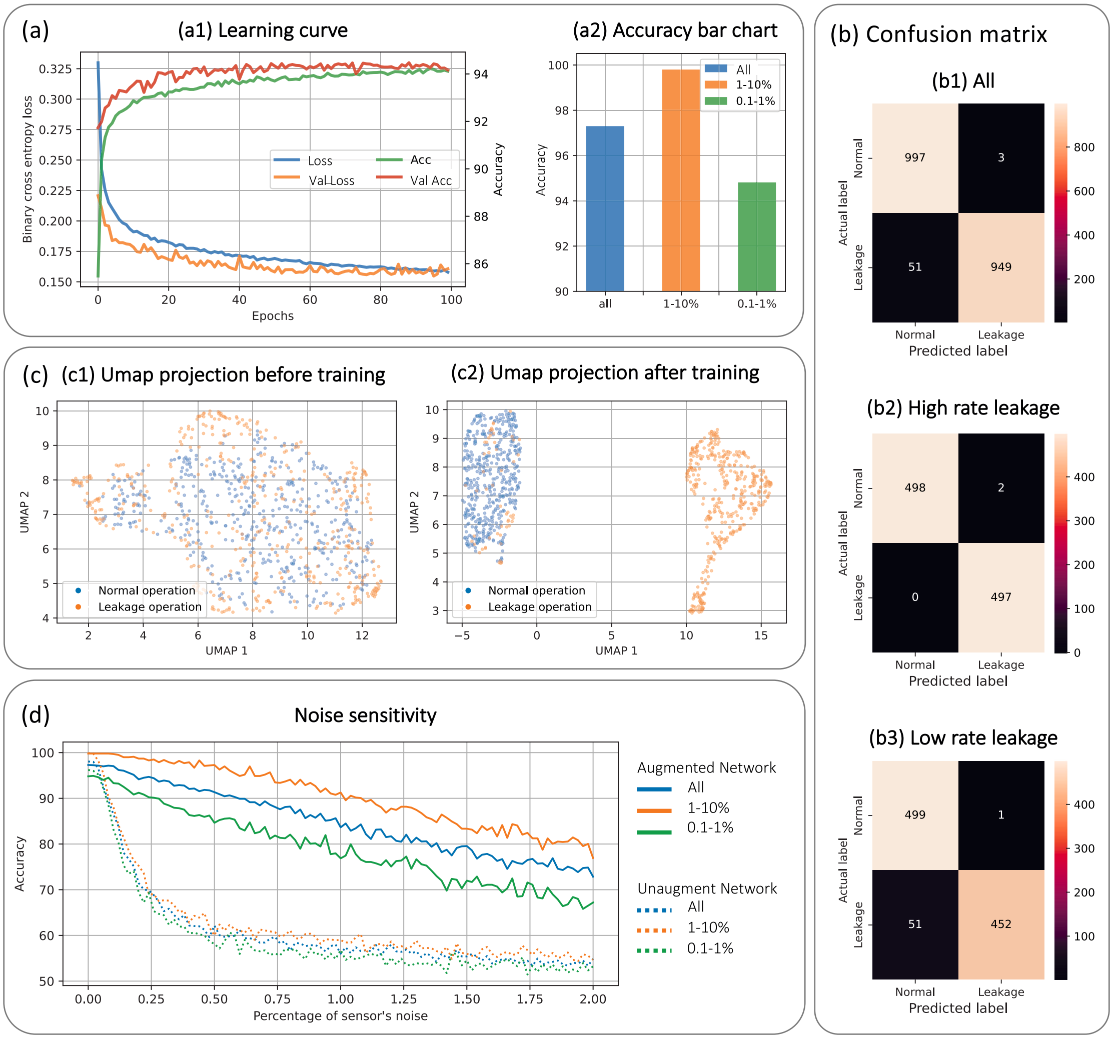


Fig. 3. (a1) Learning curve and model accuracy during training across 100 epochs. (a2) Accuracy over the testing dataset. (b1)–(b3) Confusion matrices for all, high leakage rate, and low leakage rate, respectively. (c1) and (c2) UMAP projections of sensor data before and after model projection, respectively. (d) Sensitivity comparison of the trained model with and without augmented data across various noise levels.

3.2. *Leak identification*

As mentioned earlier, the leak identification model aims to localize leaks and estimate their rates. The learning curve of this model, trained on augmented leakage data over 500 epochs, is depicted in Fig. 5 (a). The sudden improvements observed at certain points in the learning curve result from the reduction of the learning rate facilitated by the employed callback mechanism.

The accuracy of leak localization, whether there are multiple or single leakages presented in the gas network, is depicted in Fig. 5 (b1) and (b2), respectively. These results indicate that the proposed model performs better in identifying leak locations when dealing with only one leakage in the network. Although the model's performance is weaker in scenarios with multiple leakages, it still achieves accuracy beyond 97% for certain leakage sites. This result, akin to the leak detection findings, demonstrates enhanced performance compared to the Bayesian analysis method introduced in [27]. However, a significant difference lies in the neural network's capability to effectively manage 6 variable sinks while still producing accurate results.

We also examined the impact of sensor noise and data augmentation on the performance of the proposed model, as shown in Fig. 5 (c). Similar to the leak detection model, the model trained with augmented data demonstrates significantly improved robustness against noise. This underscores the importance of our proposed data augmentation technique in enhancing the final model performance.

The first 20 examples of the testing set are also presented as heat maps in Fig. 5 (d1)–(d4). Fig. 5 (d1) illustrates the locations of leaks as heat maps for each testing scenario. Each row of this heat map represents one of the leakage scenarios from the testing set. Dark red indicates the presence of a leak at the identified location along the horizontal axis, while light red signifies no leakage. Predictions made by the proposed neural network are also depicted alongside the testing set. It is evident that our method effectively identifies the majority of leakages. To illustrate further, the difference between the predicted leak location and the ground truth of the testing set is shown as a heat map on the right side of the model prediction.

The heat map in Fig. 5 (d2) also portrays the leakage rates for scenarios involving multiple leakages. Here, the model effectively estimates the rates of the leakages. However, upon closer inspection of the difference between predicted and actual rates, it becomes apparent that the model encounters challenges in accurately estimating leakages, particularly at junctions 22 and 24. Notably, these locations exhibit lower accuracy in leak detection compared to other potential leak sites (refer to Fig. 5 (b1)).

Figures 5 (d3) and (d4) yield similar outcomes for scenarios where only one leakage occurs in the system. These figures highlight the significantly improved performance of the proposed model, not just in localizing leaks but also in estimating their rates when dealing with singular leakages. It is worth noting that the majority of leakages

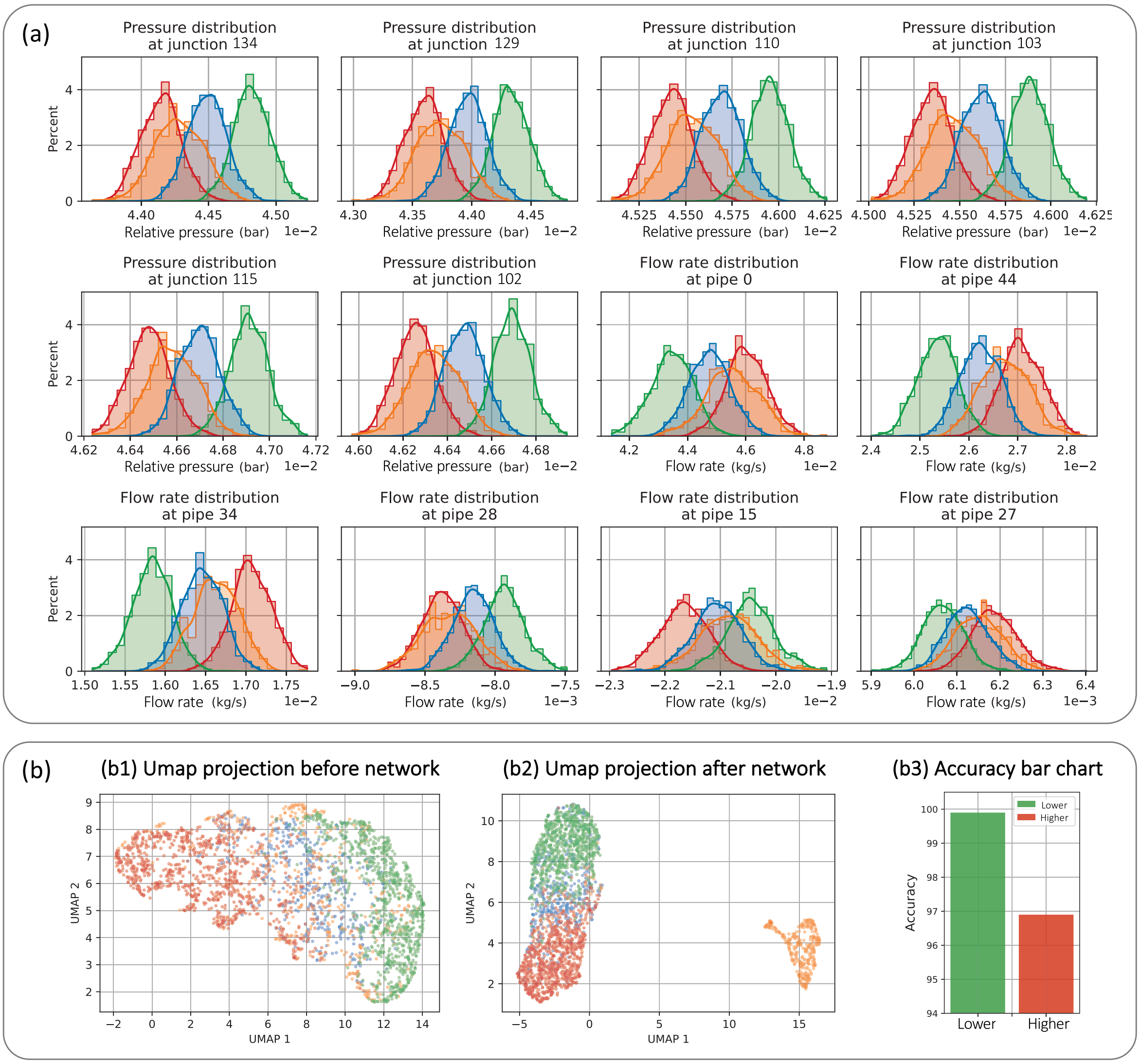


Fig. 4. (a) Distribution of sensor readings in the presence of leakage (orange), the normal state of the system used in training (blue), higher consumption of variable sinks (red), and lower consumption of variable sinks (green). (b1) UMAP projection of sensor data for the normal state (training dataset, lower variable sink consumption, and higher one) and leakage state of the system before application of the developed model. (b2) UMAP projection of the mapped sensor data to 100 dimensions by the proposed model. (b3) Accuracy of the classification of higher and lower variable sink consumption. Please note that The flow rates are defined from one junction to another, and this directional definition results in some flow rates being negative. The negative values indicate the flow direction relative to the defined positive direction between junctions. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

typically occur at a single location at any given time, with occurrences of multiple leakages being rare and usually associated with large-scale catastrophes such as earthquakes.

Similar to the leak detection case, we applied UMAP to the data where only one leakage occurred. This unsupervised projection was applied to both the sensor data and their projection into a 100-dimensional space determined by the last layer of the model before the output layer. Interestingly, UMAP not only identified the locations of the leakages but also revealed a correlation between the resulting projection of the sensor data and the leakage rates (see Fig. 5 (e1) and (e2)). However, this unsupervised projection has its limitations, as data associated with leak locations at junctions 6 and 18 are projected into almost the same location. A similar issue arises for leak locations at junctions 22 and 24.

On the other hand, our method proves to be useful in improving the performance of this unsupervised technique. As depicted in Fig. 5 (e3) and (e4), the projected data shows a better separation between the data points for each leak location. For instance, the previously mentioned issue with leak locations at junctions 6 and 18 was completely resolved, while a significant improvement is observed in the classification of leak locations at junctions 22 and 24.

In addition to the discussions on pressure, temperature, and sensor placement, factors such as meteorological changes affecting precipitation and additional noise in sensor readings are crucial considerations for our model. These variables can potentially impact the accuracy and reliability of simulations. Specifically, variations in meteorological conditions like precipitation can introduce dynamic changes in pressure within the gas network, influencing system behavior unpredictably. Furthermore, sensor noise can distort readings, affecting the quality of data used for training and inference. To mitigate these challenges, our study employs data augmentation techniques tailored to enhance the robustness of the model against such variations. By augmenting simulated data with variations that mirror real-world conditions, our approach ensures that the model is adequately trained to handle these environmental and sensor-related factors effectively. Thus, while some factors require careful consideration in simulation setups, our methodology leverages data augmentation to manage and adapt to these challenges within the digital twin framework for gas network analysis.

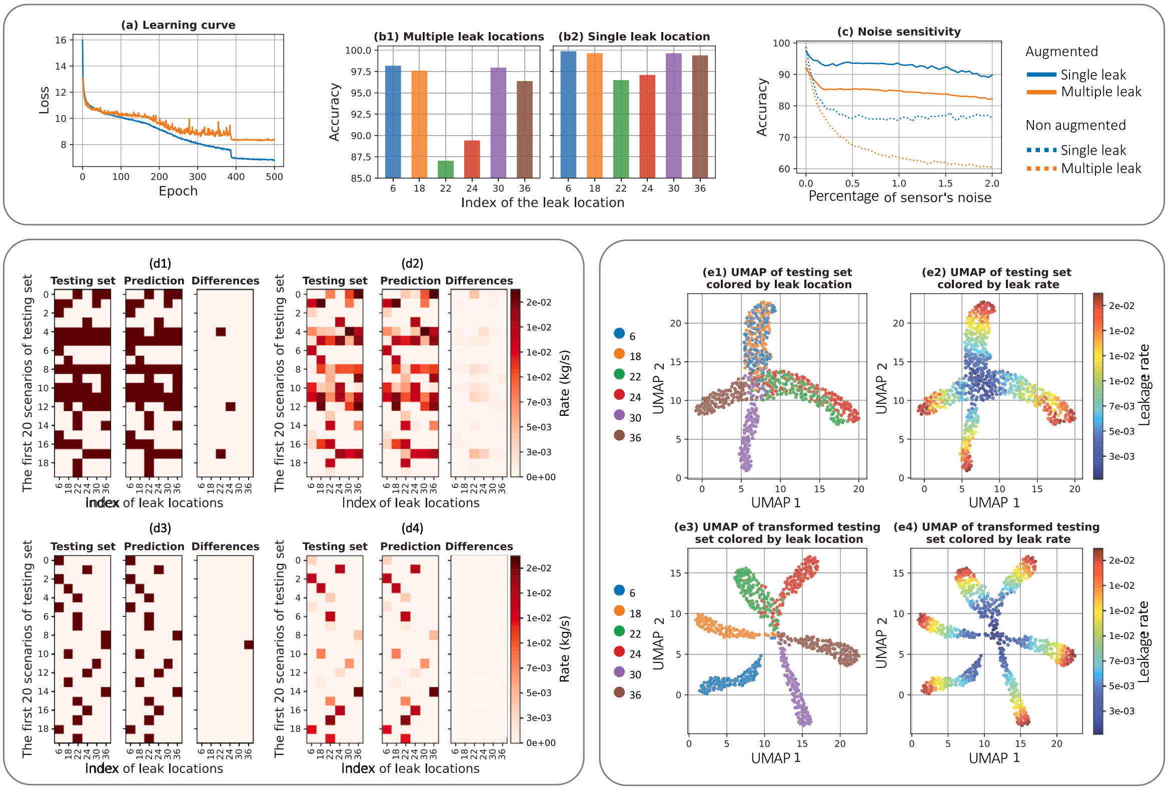


Fig. 5. (a) Learning curve. (b1) and (b2) Accuracy of the testing set for scenarios with multiple and single leakages in the gas network, respectively. (c) Noise sensitivity comparison of the model trained with and without augmented dataset. (d1) and (d2) Examples of the testing set with multiple leakages and the corresponding predictions from the proposed deep neural network. (d3) and (d4) Examples of the testing set with single leakage and the corresponding predictions from the proposed deep neural network. (e1) and (e2) Projection of the single leakage data into a two-dimensional plane using UMAP, with (e1) colored based on the location of the leakage and (e2) each projected point colored based on the leakage rate. (e3) and (e4) Similar projection using the transformed data of the last layer of the proposed model before the output layer. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

4. Discussion and conclusion

In this study, we propose a novel deep neural network framework for investigating leakage occurrences within gas distribution systems, using sparse sensor data collected from operational sensors. The presented method addresses critical issues such as handling inherently unbalanced data, the need for a multilabel classifier to simultaneously detect multiple leakages, and implementing a two-stage neural network model for leak detection and identification. This two-stage approach is strategically developed to reduce false positive leak detections. The developed model has been tested and validated using simulation data from a standard gas network with the Pandapipes package in Python, but it can also be applied to results from other steady-state solvers or real-world data.

Our method effectively manages complex interferences, including unpredictable consumption patterns and noisy sensor readings, exhibiting high accuracy in detecting leakage presence, pinpointing locations, and estimating leakage rates. It shows resilience in handling multiple leakage scenarios, such as those occurring during floods or earthquakes, demonstrating its practical utility in diverse and challenging real-world conditions. The workflow involves an initial deep neural network for detecting leakages, followed by a separate neural network for identification, facilitating precise localization and rate estimation of detected leaks. In addition to the factors discussed, such as pressure, temperature, sensor placement, meteorological changes, and sensor noise, it is crucial to acknowledge their potential impact on model performance in gas network simulations and digital twin development. These variables introduce complexities that must be carefully considered during simulation to ensure accurate representation of real-world conditions. The integration of data augmentation techniques has proven effective in mitigating the effects of variability and noise encountered in sensor readings. By enhancing the robustness of the deep learning models presented in this study, these techniques contribute to their reliability and applicability in practical settings where environmental and operational conditions may vary significantly.

CRediT authorship contribution statement

Elham Ebrahimi: Writing – original draft, Visualization, Validation, Software, Resources, Methodology, Investigation, Formal analysis, Data curation, Conceptualization. **Mohammadrahim Kazemzadeh:** Methodology, Conceptualization. **Antonio Ficarella:** Supervision, Project administration, Funding acquisition, Conceptualization.

Declaration of competing interest

The authors declare the following financial interests/personal relationships which may be considered as potential competing interests: Antonio Ficarella reports financial support was provided by Italian Ministry of University and Research, funded by the European Union - NextGeneratonEU. 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 has been supported under the National Recovery and Resilience Plan (NRRP), Mission 4 Component 2 Investment 1.4 - Call for tender No. 3138 of December 16, 2021 of the Italian Ministry of University and Research, funded by the European Union - NextGenerationEU [Award Number: CNMS named MOST, Concession Decree No. 1033 of June 17, 2022, adopted by the Italian Ministry of University and Research, CUP: F83C22000720001, Spoke 14 "Hydrogen and New Fuels"].

Appendix A. Supplementary data

Supplementary material related to this article can be found online at https://doi.org/10.1016/j.segan.2024.101496.

References

[1] C. Ma, S. Yu, J. Huo, Negative pressure wave-flow testing gas pipeline leak based on wavelet transform, in: 2010 International Conference on Computer, Mechatronics, Control and Electronic Engineering, vol. 5, IEEE, 2010, pp. 306–308.

[2] M. Rocha, Acoustic monitoring of pipeline leaks, in: ISA Calgary 1989 Symposium-Paper, 1989, pp. 283–290.

[3] A. bin Md Akib, N. bin Saad, V. Asirvadam, Pressure point analysis for early detection system, in: 2011 IEEE 7th International Colloquium on Signal Processing and Its Applications, IEEE, 2011, pp. 103–107.

[4] J. Wan, Y. Yu, Y. Wu, R. Feng, N. Yu, Hierarchical leak detection and localization method in natural gas pipeline monitoring sensor networks, Sensors 12 (1) (2011) 189–214.

[5] W. Mpesha, S.L. Gassman, M.H. Chaudhry, Leak detection in pipes by frequency response method, J. Hydraul. Eng. 127 (2) (2001) 134–147.

[6] S. Askari, N. Montazerin, M.F. Zarandi, High-frequency modeling of natural gas networks from low-frequency nodal meter readings using time-series disaggregation, IEEE Trans. Ind. Inform. 12 (1) (2015) 136–147.

[7] P.L. dos Santos, T.-P. Azevedo-Perdicoúlis, J.A. Ramos, J.M. de Carvalho, G. Jank, J. Milhinhos, An LPV modeling and identification approach to leakage detection in high pressure natural gas transportation networks, IEEE Trans. Control Syst. Technol. 19 (1) (2010) 77–92.

[8] E. Hauge, O.M. Aamo, J.-M. Godhavn, Model based pipeline monitoring with leak detection, IFAC Proc. Vol. 40 (12) (2007) 318–323.

[9] P.-S. Murvay, I. Silea, A survey on gas leak detection and localization techniques, J. Loss Prev. Process Ind. 25 (6) (2012) 966–973.

[10] J. Liu, J. Yao, M. Gallaher, J. Coburn, R. Fernandez, Study on methane emission reduction potential in chinas oil and natural gas industry, Tech. Rep., 2008.

[11] A. Ukil, H. Braendle, P. Krippner, Distributed temperature sensing: Review of technology and applications, IEEE Sens. J. 12 (5) (2011) 885–892.

[12] C.E. Campanella, G. Ai, A. Ukil, Distributed fiber optics techniques for gas network monitoring, in: 2016 IEEE International Conference on Industrial Technology, ICIT, IEEE, 2016, pp. 646–651.

[13] F. Tanimola, D. Hill, Distributed fibre optic sensors for pipeline protection, J. Nat. Gas Sci. Eng. 1 (4–5) (2009) 134–143.

[14] P. Karkulali, H. Mishra, A. Ukil, J. Dauwels, Leak detection in gas distribution pipelines using acoustic impact monitoring, in: IECON 2016-42nd Annual Conference of the IEEE Industrial Electronics Society, IEEE, 2016, pp. 412–416.

[15] R.S. Reddy, G. Payal, P. Karkulali, M. Himanshu, A. Ukil, J. Dauwels, Pressure and flow variation in gas distribution pipeline for leak detection, in: 2016 IEEE International Conference on Industrial Technology, ICIT, IEEE, 2016, pp. 679–683.

[16] B. Lindemann, B. Maschler, N. Sahlab, M. Weyrich, A survey on anomaly detection for technical systems using LSTM networks, Comput. Ind. 131 (2021) 103498.

[17] H. Esen, F. Ozgen, M. Esen, A. Sengur, Artificial neural network and wavelet neural network approaches for modelling of a solar air heater, Expert Syst. Appl. 36 (8) (2009) 11240–11248.

[18] X. Zhang, S. He, V. Stojanovic, X. Luan, F. Liu, Finite-time asynchronous dissipative filtering of conic-type nonlinear Markov jump systems, Sci. China Inf. Sci. 64 (5) (2021) 152206.

[19] X. Song, P. Sun, S. Song, V. Stojanovic, Event-driven NN adaptive fixed-time control for nonlinear systems with guaranteed performance, J. Franklin Inst. 359 (9) (2022) 4138–4159.

[20] J. Zheng, C. Wang, Y. Liang, Q. Liao, Z. Li, B. Wang, Deeppipe: A deep-learning method for anomaly detection of multi-product pipelines, Energy 259 (2022) 125025.

[21] M. Zhou, Y. Yang, Y. Xu, Y. Hu, Y. Cai, J. Lin, H. Pan, A pipeline leak detection and localization approach based on ensemble TL1DCNN, IEEE Access 9 (2021) 47565–47578.

[22] N.V.S. Korlapati, F. Khan, Q. Noor, S. Mirza, S. Vaddiraju, Review and analysis of pipeline leak detection methods, J. Pipeline Sci. Eng. (2022) 100074.

[23] C. Spandonidis, P. Theodoropoulos, F. Giannopoulos, N. Galiatsatos, A. Petsa, Evaluation of deep learning approaches for oil & gas pipeline leak detection using wireless sensor networks, Eng. Appl. Artif. Intell. 113 (2022) 104890.

[24] X. Zhang, J. Shi, X. Huang, F. Xiao, M. Yang, J. Huang, X. Yin, A.S. Usmani, G. Chen, Towards deep probabilistic graph neural network for natural gas leak detection and localization without labeled anomaly data, Expert Syst. Appl. (2023) 120542.

[25] E. Priyanka, S. Thangavel, X.-Z. Gao, N. Sivakumar, Digital twin for oil pipeline risk estimation using prognostic and machine learning techniques, J. Ind. Inf. Integr. 26 (2022) 100272.

[26] J. Liang, L. Ma, S. Liang, H. Zhang, Z. Zuo, J. Dai, Data-driven digital twin method for leak detection in natural gas pipelines, Comput. Electr. Eng. 110 (2023) 108833.

[27] P. Gupta, T.T.T. Zan, M. Wang, J. Dauwels, A. Ukil, Leak detection in low-pressure gas distribution networks by probabilistic methods, J. Nat. Gas Sci. Eng. 58 (2018) 69–79.

[28] P. Gupta, A. Goyal, J. Dauwels, A. Ukil, Bayesian detection of leaks in gas distribution networks, in: IECON 2016-42nd Annual Conference of the IEEE Industrial Electronics Society, IEEE, 2016, pp. 855–860.

[29] A. Preis, A.J. Whittle, A. Ostfeld, L. Perelman, Efficient hydraulic state estimation technique using reduced models of urban water networks, J. Water Resour. Plan. Manag. 137 (4) (2011) 343–351.

[30] M.V. Casillas Ponce, L.E. Garza Castanon, V.P. Cayuela, Model-based leak detection and location in water distribution networks considering an extended-horizon analysis of pressure sensitivities, J. Hydroinform. 16 (3) (2014) 649–670.