Probabilistic modelling for methane leak detection in gas distribution networks

Rubab Atwal
Guidehouse Europe
ratwal@guidehouse.com

Katie Green
Guidehouse Europe
kmgreen@guidehouse.com

Abstract

Methane leaks from gas distribution pipelines in the UK contribute significantly to the country's total greenhouse gas emissions. Machine learning methodologies can be employed to improve timely detection of leaks, allowing them to be fixed sooner, therefore reducing emissions. Here we present a probabilistic machine learning framework, based on a Wasserstein autoencoder and Bayesian inference, which has been developed to detect, localise, and quantify leaks within a UK-based gas distribution system with limited data availability.

1 Introduction

Gas distribution networks (GDNs) consist of pipelines carrying natural gas from transmission network offtakes (or sources) to supply gas to consumer's homes. Leaks from these networks, which can be caused by accidental damage, corrosion, aging infrastructure, or incorrect installation, emit methane into the atmosphere. This is a significant environmental challenge due to the potency of methane as a greenhouse gas (GHG) [IPCC, 2021] and a safety hazard due to its flammability. GDNs in the UK suffer from a large proportion of aging infrastructure and methane emissions from these pipelines account for an estimated 1% of total UK GHG emissions [National Physical Laboratory, 2017]. Therefore, reducing these emissions through more accurate and timely detection is key to decarbonising the energy sector and meeting targets set out in the Paris Agreement [Hureau, Geoffroy et al., 2025, Shirizadeh, Behrang et al., 2023].

Machine learning solutions have the potential to provide continuous monitoring of GDNs and proactively detect, localise, and quantify methane leaks in real time allowing for swift response and reduced emissions. Various methodologies have been explored for models to learn the operational dynamics of networks (pressure and flow profiles) and how they correspond to leaks. Previous work in this area has involved probabilistic models using operational pressure and flow sensor data across the network [Gupta et al., 2018] with attention mechanisms developed to weight probabilities [Zhang et al., 2023a]. Both attention mechanisms [Zhang et al., 2023b] and data simulation [Ebrahimi et al., 2024] have been explored to handle the lack of available training data representing anomalous leaking scenarios.

Our challenge was developing these methodologies and deploying a machine learning based leak detection model to meet the business needs of a UK-based GDN. We built on the work of Mücke et al. [2023] to develop a Wasserstein Autoencoder (WAE), trained on simulated data, in combination with a Bayesian inference solver to detect, localise, and quantify leaks from sparse operational pressure data.

Tackling Climate Change with Machine Learning workshop at NeurIPS 2025.

2 Method

2.1 Training data generation

Obtaining sufficient training data to develop leak detection models is a significant challenge due to limited pressure sensors in place across the network and flow data only available at the inlets to the highest pressure tier. This problem was overcome through simulating representative training data, as done by Ebrahimi et al. [2024]. Initially, we transformed GDN data to construct a virtual graph-based representation of the network structure using the networkx package [Hagberg et al., 2008]. We then combined this network representation with the pandapipes simulation package [Lohmeier et al., 2020] to generate our training data. Initial conditions are sampled from distributions of the historical sensor data and corresponding pressures and flows throughout the network are generated. As suggested by Fan et al. [2021] capturing training data representing both leaking (anomalous) and non-leaking (normal) scenarios is important for effective training. To obtain a sufficient amount of training data we generated 1 million non-leaking pressure and flow samples per network segment and 10,000 samples for each leaking scenario. A leaking scenario is each combination of leak location (every 50m along each pipeline) and leak size (0-40% flow out ratio).

2.2 Model architecture

Wasserstein Autoencoders (WAE) are effective for solving a Bayesian inference problem in a gas network due to their ability to accurately perform dimensionality reduction and approximate relevant probability distributions.

We trained the WAE using the operational pressure and flow values generated for the network segment. It then acts as a surrogate model to approximate the likelihood function required for Bayesian inference.

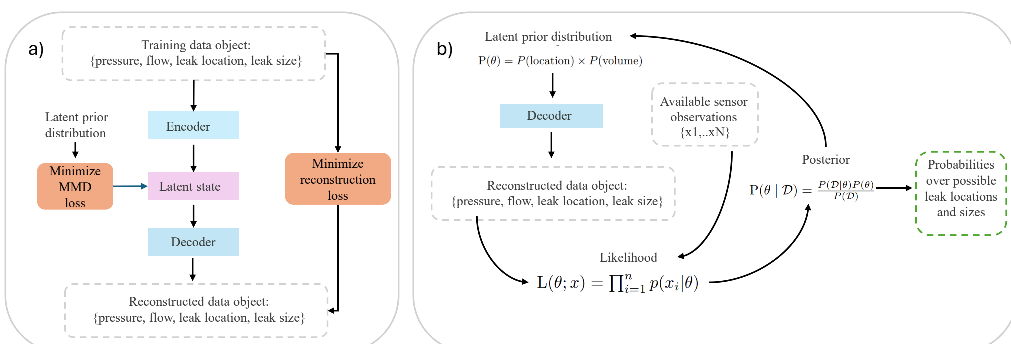

Figure 1: (a) Architecture diagram for the process of training the WAE to encode the network dynamics in the latent space and reconstruct the data object. (b) Schematic for the inference process for probabilistic leak detection through iterative calculation of the posterior from available sensor data.

To enable leak detection and quantification, we configured the prior distribution and likelihood over both the leak location and leak volume. This combined prior can be written as $P(\theta) = P(\text{location}) \times P(\text{volume})$. The likelihood function represents the probability of observing data. For a set of observations, $x = x_1, x_2, x_3, .., x_n$, $L(\theta; x) = \prod_{i=1}^{n} p(x_i|\theta)$.

The Bayesian inference solver then uses the trained decoder to combine the prior and likelihood to estimate the posterior distribution. The posterior distribution for both leak location and leak volume is computed as $P(\theta \mid \mathcal{D}) = \frac{P(\mathcal{D}|\theta)P(\theta)}{P(\mathcal{D})}$.

By using the WAE, the Bayesian inference process becomes computationally feasible, as the autoencoder allows for fast and accurate evaluation of the likelihood function. This framework can efficiently handle the high-dimensional and complex nature of gas network data, enabling real-time leak detection and probabilistic assessment of leak locations and volumes.

2.3 Experiments

To test the applicability of our model in practice, we selected a network segment covering 42km of gas distribution pipelines in the UK. We constructed the graph-based data structure, generated leaking and non-leaking scenarios and trained two autoencoders: one for predicting leak presence and one for predicting leak absence. Positive leak detections are triggered if the probability of leak presence is significantly higher than the probability of leak absence according to a paired t-test.

To validate that our models effectively learned the dynamics of the network, we first tested on a set of 100 synthetically generated leaking scenarios. This test dataset was constructed through sampling the training dataset and reserving unseen samples for validation. Since there are 12 pressure sensors in place across the 42km of network, we assume pressure data is available from all 12 in this test set.

To test performance on real incidents we back-tested the model on operational pressure data from 3 time periods where leaks were reported and compared the resulting detections to historical records. Additionally, we ran the model on pressure data from a 6-month period when no leaks were recorded to test performance under standard operating conditions.

3 Results and discussion

Our models demonstrated effective learning of the dynamics of the gas network as shown both by the comparison of operational pressure to latent dimensions learned by the autoencoder (Figure 2) and through the results on the synthetically generated test data set (Table 1).

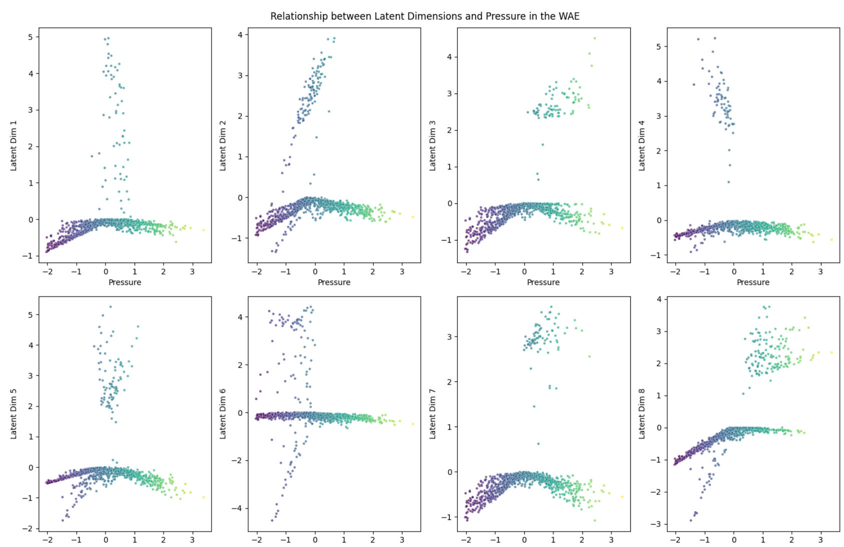

Figure 2: We examined the relationship between latent dimensions and pressure in the WAE to determine whether the WAE captures physically meaningful features.

As shown in Figure 2, several latent dimensions exhibit clear monotonic or nonlinear relationships with pressure, while others appear less correlated. This suggests that the model organises the latent space in a manner that reflects underlying physical dependencies even without explicit supervision. Each component of the model was evaluated using tailored metrics: the latent space structure and generalisation capability of the WAE were assessed through scatter plots and Frechet Inception Distance, while the model's ability to detect leaks was tested on both synthetic and historical datasets using accuracy metrics for correct pipeline identification, leak localisation, and leak volume estimation.

When tested on our generated test set, our model achieved an accuracy of 82% for detecting leaks, 80% for localisation within 50m and 76% for quantification within the correct volume bin. Considering the sensor placement across the pipeline, with a ratio of 1 pressure sensor to 3.5km of pipeline, these are exceptional results and provide promising business implications for enabling accurate leak detection and quantification. The majority of leaks that were undetected had relatively low flow rates indicating the model has a lower detection limit. This occurs as smaller leaks do not significantly impact pressure dynamics in comparison to normal operational fluctuations making them challenging for detection via pressure dynamics.

Table 1: Model performance on generated leak test set

Leak detection accuracy	Localisation accuracy	Quantification accuracy	Working sensors
82%	80%	76%	100%

When tested on the pressure data for the standard 6 month operating period, our model produced no false positives indicating strong network understanding. However, testing on historical operational data over recorded leak periods highlighted the importance of sensor placement and data availability. There were 3 leaks historically recorded on the network section modeled of which 2 were correctly detected at least a week before they were reported, demonstrating the potential of our model to enable timely response and emissions reduction. For the third recorded leak the probability signal output from the model was not significant enough to trigger a leak detection. Across the network section 12 pressure sensors are installed but across the time periods tested only 5-7 of these were collecting data. The undetected leak was recorded in a time period where only 5 pressure sensors were available for the segment and at a location far away from available pressure data. This combination led to the model's inability to detect this leak and highlights the importance of data availability for unlocking the practical benefits of these models.

Table 2: Model performance on historically recorded leaks

Leak	Pipeline detected	Predicted distance along pipeline (m)	Predicted volume (kscm/hr)	Working sensors
1	Correct	400	0.021	58.3%
2	Correct	250	0.286	58.3%
3	Not detected	-	-	38.4%

Our next steps involve continuing to build and deploy these models to cover more of the UK gas distribution network. As our models are deployed, their real-time performance will be monitored and leak detections will be validated. Further optimisation of model performance will include recommendations for pressure sensor placement and settings to increase data availability. Further work on developing our model architecture could include weighting the prior distribution to account for sensor location in a similar manner to Zhang et al. [2023b] and exploring the possibilities of replacing the data generation model with an emulation model to reduce compute costs.

4 Conclusion

Here we present a probabilistic machine learning framework, based on a WAE and Bayesian inference, which has been developed to detect, localise, and quantify leaks within a gas distribution system. This model has been successfully tested on a section of the distribution network within the UK and is in the process of being deployed in production. This will provide continuous monitoring for gas pipelines offering an early warning system for leak detection, increasing safety, and reducing emissions. The ability of the model to accurately quantify detected leaks will also allow distribution networks to improve their overall emissions estimates, tracking over time, and regulatory reporting.

Acknowledgments and Disclosure of Funding

The authors acknowledge and thank the Guidehouse 'Digital Platform for Leakage Analytics' (DPLA) project team, in particular Dale Kruchten, Ribil Gopalakrishnan, and Vinod Ramachandran, for their support, invaluable contributions and leadership throughout the development of this work. The authors also thank the Cadent DPLA team for their guidance and ongoing data provision to make this work possible. Finally, thanks to UKRI and Innovate UK for funding this work through their Stategic Innovation Fund Beta framework.

References

IPCC. *Climate Change 2021: The Physical Science Basis. Contribution of Working Group I to the Sixth Assessment Report of the Intergovernmental Panel on Climate Change*, volume In Press. Cambridge University Press, Cambridge, United Kingdom and New York, NY, USA, 2021. doi: 10.1017/9781009157896.

National Physical Laboratory. Understanding the uk greenhouse gas inventory. Technical report, Committee on Climate Change, 2017. URL `https://www.theccc.org.uk/wp-content/uploads/2017/06/Understanding-the-UK-Greenhouse-Gas-Inventory.pdf`. Commissioned to support the CCC's 2017 report on Quantifying Greenhouse Gas Emissions.

Hureau, Geoffroy, Lecarpentier, Armelle, Serbutoviez, Sylvain, Kaniewicz, Jean, Madden, Mike, Brooks, Chris, Robertson, Aileen, Harrison, Colin, Langston, Chris, and Le Ravalec, Mickaele. Global methane emissions from natural gas transmission and distribution networks. *Sci. Tech. Energ. Transition*, 80:28, 2025. doi: 10.2516/stet/2025007. URL `https://doi.org/10.2516/stet/2025007`.

Shirizadeh, Behrang, Villavicencio, Manuel, Douguet, Sebastien, Trüby, Johannes, Bou Issa, Charbel, Seck, Gondia Sokhna, D'herbemont, Vincent, Hache, Emmanuel, Malbec, Louis-Marie, Sabathier, Jerome, Venugopal, Malavika, Lagrange, Fanny, Saunier, Stephanie, Straus, Julian, and Reigstad, Gunhild A. The impact of methane leakage on the role of natural gas in the european energy transition. *Nature Communications*, 14, 2023. URL `https://doi.org/10.1038/s41467-023-41527-9`.

Payal Gupta, Thaw Zan, Mengmeng Wang, Justin Dauwels, and Abhisek Ukil. Leak detection in low-pressure gas distribution networks by probabilistic methods. *Journal of Natural Gas Science and Engineering*, 58, 08 2018. doi: 10.1016/j.jngse.2018.07.012.

Xinqi Zhang, Jihao Shi, Ming Yang, Xinyan Huang, Asif Sohail Usmani, Guoming Chen, Jianmin Fu, Jiawei Huang, and Junjie Li. Real-time pipeline leak detection and localization using an attention-based lstm approach. *Process Safety and Environmental Protection*, 174:460–472, 2023a. ISSN 0957-5820. doi: https://doi.org/10.1016/j.psep.2023.04.020. URL `https://www.sciencedirect.com/science/article/pii/S0957582023003087`.

Xinqi Zhang, Jihao Shi, Xinyan Huang, Fu Xiao, Ming Yang, Jiawei Huang, Xiaokang Yin, Asif Sohail Usmani, and Guoming Chen. Towards deep probabilistic graph neural network for natural gas leak detection and localization without labeled anomaly data. *Expert Systems with Applications*, 231:120542, 2023b. ISSN 0957-4174. doi: https://doi.org/10.1016/j.eswa.2023.120542. URL `https://www.sciencedirect.com/science/article/pii/S0957417423010448`.

Elham Ebrahimi, Mohammadrahim Kazemzadeh, and Antonio Ficarella. Leak identification and quantification in gas network using operational data and deep learning framework. *Sustainable Energy, Grids and Networks*, 39:101496, 2024. ISSN 2352-4677. doi: https://doi.org/10.1016/j.segan.2024.101496. URL `https://www.sciencedirect.com/science/article/pii/S235246772400225X`.

Nikolaj T. Mücke, Prerna Pandey, Shashi Jain, Sander M. Bohté, and Cornelis W. Oosterlee. A probabilistic digital twin for leak localization in water distribution networks using generative deep learning. *Sensors*, 23 (13), 2023. ISSN 1424-8220. doi: 10.3390/s23136179. URL `https://www.mdpi.com/1424-8220/23/13/6179`.

Aric A. Hagberg, Daniel A. Schult, and Pieter J. Swart. Exploring network structure, dynamics, and function using networkx. In Gaël Varoquaux, Travis Vaught, and Jarrod Millman, editors, *Proceedings of the 7th Python in Science Conference*, pages 11 – 15, Pasadena, CA USA, 2008.

Daniel Lohmeier, Dennis Cronbach, Simon Ruben Drauz, Martin Braun, and Tanja Manuela Kneiske. Pandapipes: An open-source piping grid calculation package for multi-energy grid simulations. *Sustainability*, 12(23): 9899, 2020. doi: 10.3390/su12239899.

Xudong Fan, Xijin Zhang, and Xiong (Bill) Yu. Machine learning model and strategy for fast and accurate detection of leaks in water supply network. *Journal of Infrastructure Preservation and Resilience*, 2(1): 10, 2021. ISSN 2662-2521. doi: 10.1186/s43065-021-00021-6. URL `https://doi.org/10.1186/s43065-021-00021-6`.