



International Journal of Data Mining and Bioinformatics

ISSN online: 1748-5681 - ISSN print: 1748-5673

<https://www.inderscience.com/ijdmb>

Construction of a natural gas pipeline intelligent early warning platform based on the integration of multi-source heterogeneous data management and GAN reinforcement learning

Huichao Guo, Runhua Huang, Yanzhi Huang

DOI: [10.1504/IJDMB.2026.10079311](https://doi.org/10.1504/IJDMB.2026.10079311)

Article History:

Received:	06 January 2026
Last revised:	04 March 2026
Accepted:	01 April 2026
Published online:	13 July 2026

Construction of a natural gas pipeline intelligent early warning platform based on the integration of multi-source heterogeneous data management and GAN reinforcement learning

Huichao Guo*

School of Safety Science and Emergency Management,
Wuhan University of Technology,
Wuhan, 430000, Hubei, China
Email: 334998@whut.edu.cn
and

Zhejiang Business College,
Hangzhou, 310000, Zhejiang, China

*Corresponding author

Runhua Huang

Zhejiang Business College,
Hangzhou, 310000, Zhejiang, China
and
The Chinese University of Hong Kong,
Shenzhen, 518100, Guangdong, China
Email: runhuahuang@link.cuhk.edu.cn

Yanzhi Huang

Zhejiang Business College,
Hangzhou, 310000, Zhejiang, China
Email: huangyanzhide@163.com

Abstract: The platform proposed in this paper is a combination of two technologies (a data governance module and an intelligent early warning model). It features a metadata-driven heterogeneous data fusion model that provides a means of efficiently cleaning, aligning, and building feature sets to integrate multiple sources of data (such as SCADA and GIS) by providing a solution for the challenge of integrating disparate types of datasets. The combination of a conditional generative adversarial network (CGAN) and reinforcement learning enables the generation of high-quality synthetic anomaly sample datasets based on the 'expectation' of future performance, informed by conditional fault characteristics (CFC). The newly generated enhanced data will be utilised to train an early warning agent using a deep Q-QN model. Using this approach to maximise cumulative rewards will provide a means to train the best possible early warning strategy, enabling a proactive and accurate approach to identify and manage risk in pipelines.

Keywords: multi-source heterogeneous data governance; conditional generative adversarial network; deep Q-network; DQN; natural gas pipeline; intelligent early warning; conditional fault characteristics; CFC.

Reference to this paper should be made as follows: Guo, H., Huang, R. and Huang, Y. (2026) ‘Construction of a natural gas pipeline intelligent early warning platform based on the integration of multi-source heterogeneous data management and GAN reinforcement learning’, *Int. J. Data Mining and Bioinformatics*, Vol. 30, No. 7, pp.1–25.

Biographical notes: Huichao Guo received his Bachelor’s degree from the Nankai University. Currently, he is chasing the PhD degree in the School of Safety Science and Emergency Management, Wuhan University of Technology. His research interests include big data analysis, digital management, emergency management, and intelligent management.

Runhua Huang received his BE degree from The Chinese University of Hong Kong, Hong Kong in 2014, MS degree from Columbia University, New York, NY, USA in 2019, and PhD in Data Science from The Chinese University of Hong Kong, Shenzhen. His research interests include smart transportation networks, population migration analysis, and urban mobility systems.

Yanzhi Huang is a Lecturer at Zhejiang Business College and holds a PhD from the University of York, UK. Her research spans entrepreneurial psychology, pedagogical innovations in entrepreneurship education, and AI-driven transformation strategies for SMEs. Her work emphasises actionable insights to enhance organisational agility and foster sustainable growth in dynamic markets.

1 Introduction

As a significant source of clean energy, natural gas (NG) plays a pivotal role in the global transition toward carbon neutrality. Recent research highlights that accurate time-series forecasting of energy consumption and emissions, supported by advanced analytical techniques, is critical for balancing economic growth with environmental protection (Wu, 2024). In the era of Industry 4.0, the maturity of data analytics has become a core determinant of organisational innovation and competitive advantage. Recent empirical research indicates that data analytical skills significantly influence both explorative and exploitative innovation capabilities, especially when complemented by effective leadership (Rizvi and Khalid, 2025). The ability to operate pipeline systems safely directly affects a nation’s energy security, economic stability, and environmental protection. With the increasing age of pipeline systems and their exposure to additional external factors related to the environment; also, as the infrastructure ages, there is a corresponding increase in the likelihood of safety events such as corrosion, leaks and other forms of third-party sabotage, increasing risk and potential consequence to our society and environment (Singh and Poblete, 2025; Jones et al., 2025; Pengchao, 2025). Hence, there is a need for advanced, proactive, and intelligent early warning technology to ensure the early and accurate identification and prediction of pipeline hazards, as well as for converting the traditional response mode to a preventive mode. The integration of

artificial intelligence (AI) into industrial operations is no longer just a technical upgrade but a strategic necessity. Recent studies emphasise that AI capabilities significantly boost organisational performance and innovation by effectively organising development resources and intangible assets (Ali and Simmou, 2025; Zhuang, 2025). There are both theoretical and practical implications for reliability and sustainability of the NG supply chain (Adebisi et al., 2021; Asadi, 2024; Fazle et al., 2023).

At present, the research and practical application of intelligent early warning for NG pipelines are facing severe technical challenges. A core problem lies in the effective management and integration of multi-source heterogeneous data.

The integration of multi-source heterogeneous data is not only a technical requirement but also a key factor in enhancing system effectiveness. According to recent research on multimodal AI acceptance, increased media richness significantly strengthens the perceived usefulness and positive attitudes of users toward advanced technological frameworks (Lu, 2025). By providing a comprehensive data view, our platform facilitates a more intuitive understanding of pipeline risks. Pipeline operation data comes from multiple systems, including SCADA, fibre-optic monitoring, GIS, and environmental sensors. These data have significant differences in acquisition frequency, format structure, and time synchronisation (Enemosah and Chukwunweike, 2022; Nuruzzaman and Rana, 2025; Imam, 2025). Kornilova's (2025) research pointed out that the lack of a unified metadata standard is the main reason for the distortion of the fusion of real-time data such as SCADA and static data such as GIS. Villegas-Ch et al. (2024) found that when aligning the features of high-speed optical fibre vibration signals with low-frequency pressure data, time-series interpolation can easily introduce new errors. Another prominent challenge is the extreme sparsity of pipeline abnormality samples. Real pipeline leakage or early-failure events are rare, making the fault samples used for training models extremely scarce. According to a report by Awuku et al. (2024), the proportion of datasets with faults was extremely low, resulting in a high level of class imbalance. Yutong et al. (2025) highlighted that the severity of the data imbalance made it difficult for deep learning models to adequately represent the subtle anomalous features inherent in these samples, resulting in poor generalisation and low warning accuracy. Ahmadi et al. (2025) further noted that many current techniques are unable to accurately capture and represent the complex temporal dependencies, making it even more difficult to develop a robust warning model.

To address the above challenges, academia and industry have adopted a variety of methods. Abdelhafidh et al. (2023) used Bayesian networks to probabilistically infer pipeline failures. However, Bayesian networks rely on prior knowledge and expert experience and are difficult to adapt to the dynamic changes in pipeline operation status (Abdelhafidh et al., 2023). Mahapatra and Patra (2025) proposed a method based on principal component analysis (PCA). The PCA method is used to process multidimensional features, but its essence remains dimensionality reduction, and it lacks deep semantic associations in heterogeneous data feature fusion. To address sample sparsity, Riaz et al. (2025) used the synthetic minority over-sampling technique (SMOTE) to generate abnormal samples. However, SMOTE only generates samples via linear interpolation, lacks characterisation of the complexity of fundamental fault modes, and often produces low-quality synthetic data. The Toufigh and Ranjbar (2024) team used the isolation forest for anomaly detection. Still, this method is more inclined to unsupervised anomaly point recognition and cannot predict specific fault types and evolution trends. The research of Ullah and Mahmoud (2025) shows that although

traditional deep learning methods, such as recurrent neural networks (RNNs), have advantages in processing pipeline time series data, when faced with highly unbalanced datasets, model training is prone to falling into local optimality, and the recall rate of key faults is poor. The limitations of these methods collectively point to the key problem this paper aims to solve: existing methods lack an effective mechanism for high-quality fusion of heterogeneous multi-source data, and it is challenging to build a high-precision, high-generalisation intelligent early warning model when abnormal samples are incredibly sparse.

The primary contributions of this research are twofold, spanning both data engineering and machine learning domains. At the data governance level, we propose a metadata-driven fusion architecture that systematically resolves the semantic and temporal misalignments of multi-source heterogeneous pipeline data, providing a high-fidelity ‘data soil’. At the intelligent model level, we build a synergistic model using a CGAN coupled with deep reinforcement learning (DRL). The CGAN serves as the ‘sample factory’ that addresses data sparsity, while the DRL agent functions as the ‘decision brain’ to help optimise early-warning strategies. The relationship between the two models is interrelated; data governance will support CGAN sample production through quality of the feature space, which will affect the reliability of the CGI-model’s synthetic samples, while these better samples permit a DRL agent to discover sound policies for high-impact fault modes that are frequently under-represented in raw industrial datasets.

2 Methods

2.1 Metadata-driven heterogeneous data fusion model

In transforming raw data into training datasets, a step-by-step description of how data extracted from various Pipeline Operations through means such as SCADA systems, geographic information systems (GIS), optical sensors, and environmental monitoring systems, among others, are combined into a comprehensive dataset. Each data source provides unique types of information with different formats and frequency ranges. Hence, it is crucial to incorporate and standardise all the datasets from each Pipeline Operation data source (Laayati et al., 2025; Mijwil, 2025). This article’s focus is on developing a metadata-driven heterogeneous data fusion model that clearly outlines and governs the operation of unified metadata standards; allows for the real-time cleaning of input data, aligning and combining input data to common time intervals, mapping it spatially and extracting input data features (Machireddy, 2023; Sultana and Rozony, 2025). Data cleaning will consist of the removal of outlying data points. Filtering data occurs when the data point is outside a defined physical boundary, (i.e., negative pressure), while statistical outliers will be determined using the Z-score approach. For example, if x is a data point, then it will be marked as an outlier if $|(x - \mu)/\sigma| > 3$, where μ and σ are the rolling average and standard deviation of a dataset’s specified window.

Outlier values identified are fixed utilising the median of nearby values to ensure data continuity. The missing value replacement method will depend on both the characteristics of the data source and the length of time that the data are missing from their original position. For SCADA and fibre-optic sensor data where high-frequency sampling is conducted, linear interpolation will be used to replace short-range missing data (less than

three consecutive values) to maintain local temporal coherence. Conversely, environmental or long-range missing operational data will be repaired through the application of a random forest regression algorithm to predict values based on complex nonlinearity between environmental variables and the historical trends of operational data. The random forest regression algorithm allows for the prediction of values based on the historical relationships of the input variables as opposed to using simple linear transitions to replace missing values. The time alignment module will resample and synchronise heterogeneous time series to ensure strict temporal correspondence among the feature vectors by applying a common 10-minute sampling interval ΔT . To achieve precise integration, the spatial mapping module employs a coordinate-based nearest neighbour (CNN) matching algorithm. Specifically, dynamic streaming data from SCADA sensors (identified by unique Device IDs) is bound to their respective longitude and latitude coordinates in the metadata layer. Static attributes from the GIS database, such as pipe material, age, and burial depth, are indexed using a spatial R-tree. By calculating the Euclidean distance between the SCADA sensor coordinates (x_s, y_s) and the nearest GIS pipeline segment vertices (x_g, y_g), the system establishes a many-to-one mapping. A distance threshold $d_{\max}$ (e.g., 5 metres) is applied to ensure spatial consistency, thereby successfully appending static geological and structural features to the real-time operational feature vector. The metadata-driven multi-source heterogeneous data fusion architecture is shown in Figure 1.

Figure 1 Metadata-driven multi-source heterogeneous data fusion architecture (see online version for colours)

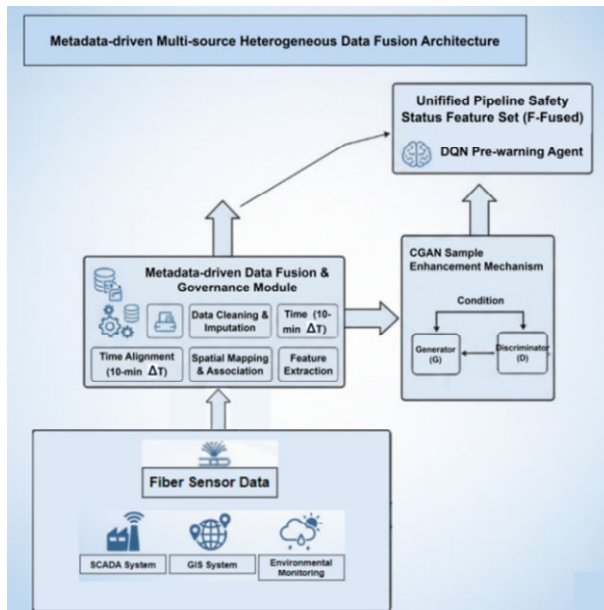


Figure 1 depicts the core architecture of the NG pipeline intelligent early warning platform, illustrating the processing flow from raw multi-source heterogeneous data to a final unified safety status feature set. It integrates a metadata-driven data fusion and governance module responsible for cleaning, temporal alignment, spatial mapping, and

feature extraction of data from diverse sources, including SCADA systems, GIS, fibre-optic sensing, and environmental monitoring. The processed data from this module, combined with the conditional generative adversarial network (CGAN) sample enhancement mechanism, addresses the challenge of sparsity in pipeline anomaly samples. Ultimately, all processed and enhanced information is aggregated into a unified pipeline safety status feature set, which serves as input to the DQN early warning agent's decision-making, fully supporting intelligent identification and early warning of pipeline risks.

The pipeline operation status feature set design and pre-processing parameters are shown in Table 1.

Table 1 Pipeline operation status feature set design and pre-processing parameters

<i>Feature name</i>	<i>Original source</i>	<i>Data type</i>	<i>Normalisation method</i>	<i>Time window ΔT</i>
Pressure fluctuation	SCADA	Numerical	Z-Score	10 min
Flow rate gradient	SCADA	Numerical	Min-max	10 min
Pipeline material code	GIS	Categorical	One-hot encoding	Static
Soil temperature avg.	Environmental	Numerical	Min-max	10 min
Strain rate change	Fibre sensor	Numerical	Z-Score	10 min
Accumulated flow diff.	SCADA	Numerical	Z-score	10 min
Pipeline depth	GIS	Numerical	Min-max	Static
Rainfall amount	Environmental	Numerical	Min-max	10 min
Velocity of sound	SCADA	Numerical	Z-Score	10 min
Density variation	SCADA	Numerical	Z-Score	10 min

Table 1 lists the ten key features and their corresponding pre-processing parameters used to construct a unified pipeline safety status feature set, and demonstrates the specific implementation plan for multi-source heterogeneous data fusion. The features in this system are each identified by their original source of data (this includes real-time, operational status data from the SCADA system; static characteristics for the pipeline from the GIS system; and time series data via environmental monitoring and/or fibre-optic sensing). Numeric features are scaled to the same range by standardising them via either Z-score or min-max to eliminate dimensionality differences and evenly distribute the model training process. The time interval used for all time series features is consistently and uniformly defined as 10 minutes (ΔT) allowing the different types of data to be aligned and synchronised for temporal alignment/coordination. Finally, the static properties of these features contain spatial metadata so all data describing the overall pipeline safety condition will have been through integration of spatially relevant data to produce an overall input feature vector (F_{Fused}).

2.2 CGAN sample enhancement mechanism

Deep learning models typically require extensive datasets for training, and their application to smaller datasets is often hindered by the risk of overfitting (Shu, 2025). To address the problem of extremely sparse samples of abnormal events, such as early failures or minor leaks, this paper uses a CGAN to enhance sample quality

(Woldesellasse and Tesfamariam, 2023; Sauber-Cole and Khoshgoftaar, 2022; Ribas et al., 2025). A CGAN consists of a generator G and a discriminator D , trained via adversarial learning. The goal of the generator G is to generate synthetic abnormal samples $\tilde{x}$ that are sufficiently realistic, while the goal of the discriminator D is to accurately distinguish whether the input sample comes from the real dataset x or $\tilde{x}$ synthesised by the generator G . In this CGAN framework, the conditional variable c is structured as a joint-encoded vector to represent both the fault category and its operational severity. Specifically, c is a concatenated vector $[c_{type}, c_{severity}]$. The component c_{type} is a discrete, one-hot-encoded vector representing four primary fault categories (e.g., $[1, 0, 0, 0]$ for leakage). The component $c_{severity}$ is a continuous scalar value normalised between $[0, 1]$, where 0 represents a nascent anomaly, and 1 represents a critical failure. This hybrid encoding allows the generator G not only to distinguish between different physical failure modes but also to simulate the continuous evolution and intensity of these faults. It is used to guide the sample generation process and ensure that the synthetic samples have specific abnormal properties. The objective function of CGAN is a minimum-maximum game problem, defined as follows:

$$\min_G \max_D E_{x \sim p_{data}(x)} [\log D(x | c)] + E_{z \sim p_z(z)} [\log(1 - D(G(z | c)))] \quad (1)$$

As shown in formula (1), $E_{x \sim p_{data}(x)}$ represents the expectation of the real data distribution $p_{data}(x)$, $E_{z \sim p_z(z)}$ represents the expectation of the random noise distribution $p_z(z)$, x is the real sample, z is the input noise, c is the conditional variable, $D(\cdot | c)$ represents the probability of the discriminator judging the authenticity of the sample under the given condition c , and $G(z | c)$ represents the synthetic sample generated by the generator under the given noise z and condition c . To enhance training stability and mitigate the vanishing or exploding gradient problems common in GANs, spectral normalisation (SN) is implemented in each convolutional and fully connected layer of the discriminator D . Specifically, the weight matrix W of each layer is normalised by its largest singular value (spectral norm) $\sigma(W)$, such that $\bar{W}_{SN} = W / \sigma(W)$. The discriminator's Lipschitz constant has a maximum of 1 with this approach, the discriminator does not have an excessive amount of confidence and as a result, smooth gradient signals can be sent to the Generator G . This quality of smooth gradients in SN helps reduce the occurrence of mode collapse and enables a more stable and converging adversarial training process while performing across multi-source pipelines with significant imbalances in the amount of input data. The output $\tilde{x}$ of the Generator G is a high dimensional feature vector that represents a synthetic abnormal pipeline condition. Its generation process is based on the nonlinear mapping of noise z and condition c , which can be expressed as follows:

$$\tilde{x} = G(z, c; \theta_G) \quad (2)$$

As shown in formula (2), where θ_G represents the parameters of the generator network. This sample enhancement mechanism effectively addresses the issue of sparse abnormal samples and significantly improves the generalisation ability of the subsequent warning model.

The generator G was constructed as a deep fully connected neural network in order to guarantee accurate synthesis of the input from all feature vectors originating from multiple sources. The G network has an input layer, three separate hidden layers with

128-256-128 neurons, and one output layer equal in dimension to F_{Fused} . The neurons in each hidden layer utilise the leaky rectified linear unit (Leaky ReLU) activation function, where the negative slope is 0.2, to avoid creating dead neurons. The batch normalisation technique is then applied to all hidden layers to accelerate the convergence process. The output layer of the G network uses the Tanh activation function, which normalises output data consistent with real pipeline flow data. This method effectively allows the G network to extract and represent the more complex cross-correlation between all of the SCADA, GIS and other environmental features present in the multi-source feature vectors.

The CGAN process generates synthetic anomalies, and we have created methods to establish whether they have distributions like those of real anomalies. The Kolmogorov-Smirnov test (KS test) is a statistical test for evaluating the similarity between the CDFs of synthetic anomaly datasets and the CDF of the actual anomaly data. A main assumption of the KS-test is that we can compare two datasets by looking at where the largest vertical difference exists (vertical difference occurs at same point along the X-axis) between the cumulative distributions of both datasets. The KS statistic measured during testing is designated as ‘D’, with a lower D indicating a greater similarity between the two distributions. Thus, sufficient statistical confidence exists in the generation of the synthetic anomaly state via the CGAN process.

2.3 Design of early warning agent based on deep Q -network (DQN)

By employing DQN, the research team has established a foundation for developing an early warning agent. To accomplish this, they treated the pipeline early warning problem as a Markov decision process (MDP). Using these models, the agent acquires an understanding of how best to respond to the ambiguous conditions surrounding pipeline operation (Salehpour et al., 2025; Yilmaz Eroglu, 2025; Ong et al., 2021). The environmental state S_t is defined as the sum of the features that comprise the pipeline safety states F_{Fused} ; it represents the full picture of what it means at time t . The agent’s action space A consists of three possible actions: no warning ($a = 0$), minor warning ($a = 1$), and severe warning ($a = 2$). Ultimately, the agent seeks to find a sequence of actions that yields the maximum cumulative discounted future reward given its current state S_t (Khlifi et al., 2025; Kathiresan, 2025; Heckmann, 2023).

The core of the DQN model is to use a deep neural network to approximate the optimal action-value function $Q^*(S, a)$, which represents the expected maximum future reward obtained by acting a in state S . DQN uses two identical neural networks: a main Q -network (parameterised by θ) and a target Q -network (parameterised by θ_{target}) to improve training stability and convergence. The agent explores the environment using a ϵ -greedy strategy and collects experience $\langle S_t, a_t, R_{t+1}, S_{t+1} \rangle$, which is stored in an experience replay pool. During training, the parameters θ of the main Q -network are updated by minimising the loss function $L(\theta)$, which calculates the mean squared error between the target Q -value Y_t and the current Q -value $Q(S_t, a_t; \theta)$:

$$L(\theta) = E_{S_t, a_t, R_{t+1}, S_{t+1} \sim U(D)} \left[\left(Y_t - Q(S_t, a_t; \theta) \right)^2 \right] \quad (3)$$

As shown in formula (3), where E represents the expectation of uniform sampling from the experience replay pool, and $DQ(S_t, a_t; \theta)$ is the prediction value of the main

Q -network for the current state-action pair. The target Q value Y_t is calculated using the target Q network and is defined as follows:

$$Y_t = R_{t+1} + \gamma \max_{a'} Q(S_{t+1}, a'; \theta_{target}) \quad (4)$$

As shown in formula (4), R_{t+1} is the immediate reward the agent receives after executing action a_t in state S_t ; γ is a discount factor that balances the importance of immediate rewards and future rewards; S_{t+1} is the state at the next moment; and $\max_{a'} Q(S_{t+1}, a'; \theta_{target})$ represents the target Q network θ_{target} 's prediction of the maximum Q value for all possible actions a' in the next state S_{t+1} . The design of the reward function R is crucial. It accounts for the reward $R_{Success}$ for successful warnings, the reward R_{Safe} for correctly identifying safe states, and the costs of false alarm penalties P_{False} and missed alarm penalties P_{Leak} to guide the agent in learning the optimal, risk-minimising warning decision strategy π .

To balance exploration and exploitation, the DQN agent adopts a linearly decaying ϵ -greedy strategy. The value of ϵ is set to 1.0 to ensure maximum exploration of the state space during early training. It is then scheduled to decrease linearly according to the formula: $\epsilon = \max(\epsilon_{min}, \epsilon_{start} - n \cdot \Delta\epsilon)$, where $\epsilon_{min} = 0.05$ is the floor value and n represents the current training step. The decay rate $\Delta\epsilon$ is set such that ϵ reaches its minimum value after approximately 70% of the total training episodes. This scheduling ensures that the agent gradually shifts from broad exploration to stable exploitation of the learned optimal early-warning policy.

Incorporating an experience replay pool extending up to an $N = 50,000$ sample capacity will remove temporal correlations among samples to increase sample efficiency. To provide sufficient variability over time with respect to pipeline operating states and infrequently arising anomaly conditions produced by the CGAN, this experience replay pool will keep the agent from overfitting recent experiences with respect to expectations. When training the model, the size of the sampling batch was established as 128. This value represents a compromise between accuracy in estimating gradients and efficiency in processing; using a batch size of 128 results in stable policy gradient estimates within a multi-source heterogeneous feature set while maximising throughput on the GPU-based hardware accelerator.

The design parameters of DQN warning action space and reward function are shown in Table 2.

In Table 2, we present the discrete action space A available to the DQN early warning agent and also present the design parameters upon which the immediate reward function (R_{t+1}) is based, which drives the learning of the agent. Fundamental to all reinforcement learning (RL) agents is the action space and reward function, as these are the primary components of how agents make decisions through RL. The action space for the agent is made up of three types of decisions, specifically: a decision to give no warning, a minor warning and a severe warning. These decisions correlate to the three states associated with pipeline operational safety: safe; minor risk; and critical risk.

The reward function for the early warning task provides a value for the consequences of the agent performing incorrectly with respect to giving correct, (i.e., successful) warnings ($R_{Success}$), providing accurate assessments of pipeline safety (R_{Safe}), and provides significant penalties for giving false alarms (P_{False}); as well as for not issuing warnings (P_{Leak}).

Table 2 DQN warning action space and reward function design parameters

<i>Action ID</i>	<i>Action a</i>	<i>Action description</i>	<i>Environment outcome (true state)</i>	<i>Immediate reward R_{t+1}</i>
a = 0	No warning	No alarm is issued	Safe (no risk detected)	R_{Safe}
a = 0	No warning	No alarm is issued	Critical risk (failure occurs)	P_{Leak}
a = 1	Slight warning	Minor alert issued for potential risk	Minor risk	$R_{Success}$
a = 1	Slight warning	Minor alert issued for potential risk	Safe (false alarm)	P_{False}
a = 2	Severe warning	Immediate high-level alert	Critical risk (failure occurs)	$R_{Success}$
a = 2	Severe warning	Immediate high-level alert	Safe (false alarm)	P_{False}

By structuring the reward function in this way, the agent’s learning can concentrate on minimising the cost associated with false alarms while at the same time maximising cumulative discounted rewards (Y_t), allowing for a trade off between safety and cost within the overall early warning decision-making process.

As a result, the augmented data is utilised by a DQN to train the early warning agent. Cumulative reward maximisation enables the agent to develop an optimal early-warning strategy to efficiently and effectively make proactive, accurate decisions about risks in the pipeline.

3 Experiment

3.1 Experimental dataset construction and processing

The purpose of this section is to describe the origin and make-up of the dataset employed in the experiment, being used to train and assess the intelligent early warning system (IEWS) for NG pipelines through multi-source heterogeneous data governance (MHGD) combined with generative adversarial networks (GANs) RL. The simulated long-distance gas pipeline operation dataset generated by means of simulation using PIPESIM (Schlumberger/SLB) is used to collect data for this experiment. This dataset provides data on the various operational states and environmental conditions a gas pipeline can experience, thereby enabling repeatable, comparable experimental results. The dataset comprises four major sources of data that are integrated into one comprehensive dataset:

- 1 real-time data from SCADA, such as pressure, flow, temperature, and other operational parameters, obtained by sampling every minute
- 2 GIS provides the static attributes associated with the pipeline, (e.g., pipeline topology, material type and burial depth) that are used as secondary features in the dataset.
- 3 Simulated data generated from the use of sensors that simulated the signals recorded by optical fibre and vibration sensors and at a significantly higher frequency.

- 4 environmental monitoring data that contain parameters that can affect the gas pipeline, (e.g., soil temperature and rainfall), and were sampled at a significantly lower frequency than the SCADA data.

The integrated dataset is partitioned into training, validation, and test sets using a 70/10/20 ratio. This distribution is selected to provide a sufficiently large training volume (70%) for the CGAN and DQN models to converge, while reserving 10% for hyperparameter tuning and 20% for an unbiased final evaluation of the early warning performance. To prevent temporal data leakage, we implemented a chronological splitting strategy instead of random shuffling. Specifically, the data were sorted by timestamp, with the first 70% used for training, the next 10% for validation, and the final 20% for testing. A temporal gap (buffer zone) of 24 hours was inserted between the segments to prevent transient hydraulic states in the training set from statistically biasing the evaluation of the test set, thereby ensuring strict temporal independence.

A summary of the components and scale statistics of the original dataset is reported in Table 3.

Table 3 Composition and size statistics of the original dataset

<i>Data source</i>	<i>Data type</i>	<i>Sampling frequency</i>	<i>Time span</i>	<i>Total records (approx.)</i>	<i>Fault records (%)</i>
SCADA	Time-series	1 min	Three years	1.5×10^6	< 0.05
GIS	Static attributes	N/A	Current	1,000 segments	N/A
Fibre sensor	High-frequency	100 Hz	Three years	1.5×10^8	< 0.05
Environmental	Time-series	1 hour	Three years	2.6×10^4	N/A
Fault log	Event records	Irregular	Three years	750 events	100

In the data processing phase, raw data undergoes a series of cleaning, time alignment, and metadata-driven feature extraction via a fusion model to create a unified feature vector for pipeline safety status. In the cleaning phase, raw data points outside their physical range are corrected, and outlier data points are corrected based on statistical criteria. In time alignment, all heterogeneous time-series datasets are synchronised and aggregated into fixed 10-minute time units. The final feature vector created is then separated into three sets (training, validation, and test) at a ratio of 70% for training, 10% for validation, and 20% for testing. In addition, temporal independence across all three sets will allow us to test our model’s generalisation to unseen data. Lastly, all real-world failure samples (including minor leaks and corrosion) constitute less than 0.000 0.05% of the total data to simulate both the extreme rarity of failure events and the fact that the pipeline’s normal operation is dominated by non-pandemic events. The processed data serves as input to both CGAN sample augmentation and DQN early-warning agents.

3.2 Performance evaluation indicators and comparison baseline setting

Key metrics for evaluating the performance of the IEWS are outlined in this section, serving as a baseline for comparative evaluations. The selection of accuracy, precision, recall, F1 score, false alarm rate (FAR), and missed alarm rate (MAR) to comprehensively evaluate NG pipeline early warning tasks was based on their high safety and timely nature (Aljameel et al., 2022; Al-Khalidi et al., 2024). Accuracy represents the

percentage of correctly classified observations made by the model overall. However, its reference value is limited for highly imbalanced datasets because it produces a high number of false positives relative to actual trends. Precision represents the percentage of true anomalies out of all samples predicted as anomalies, (i.e., FAR), and represents model's ability to help limit or prevent false alarms. Recall represents the percentage of true anomaly samples from all samples identified by the model as anomaly and, therefore, highlights the model's ability to help limit or prevent missed alarms, which is particularly important when dealing with safety early warning. F1 score represents the harmonic mean of precision and recall and can be used to evaluate model performance holistically.

Given the particularities of pipeline early warning tasks, this paper focuses on the FAR and the MAR. The FAR is the proportion of standard samples mistakenly identified as abnormal. A high FAR leads to decreased operator trust and unnecessary inspection downtime. The missed-alarm rate is the proportion of abnormal samples that are mistakenly identified as usual. A high MAR leads to serious safety accidents (Aderubi et al., 2023; Amini et al., 2022). The MAR is calculated as shown in formula (5):

$$MAR = \frac{FN}{TP + FN} \quad (5)$$

In formula (5), FN represents the false negatives, i.e., the number of samples that are truly abnormal but misclassified as usual by the model; TP represents the true positives, i.e., the number of samples that are truly abnormal and correctly classified as abnormal by the model.

This paper sets three comparative baseline models:

- 1 Feature-engineered support vector machine (SVM) model: this model uses a classic radial basis function (RBF) kernel SVM, and the input features are the original sparse feature set that has undergone data management but not CGAN enhancement.
- 2 Based on the Synthetic Minority Oversampling Technique-Long Short-Term Memory (SMOTE-LSTM) model: traditional SMOTE is used to oversample sparse abnormal samples, and then an LSTM network is used for time series prediction and warning classification to compare the effectiveness of CGAN sample enhancement.
- 3 Raw data DQN model: the original sparse feature set without CGAN enhancement is directly input into the DQN agent for training to compare the improvement of CGAN RL on DQN decision-making performance.

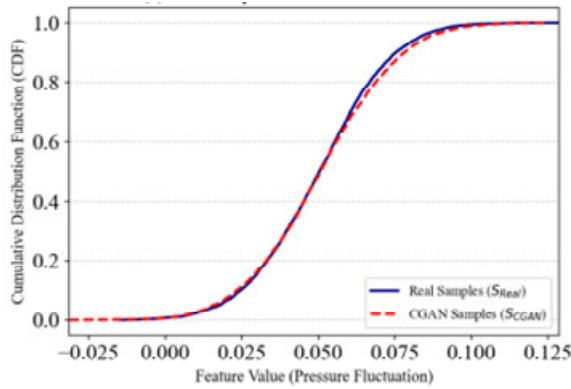
To quantify the actual contribution of the fusion feature set to early warning decision-making, this paper applies Shapley Additive exPlanations (SHAP) as a metric to measure the value of the fused feature set. Based on the Shapley value principle in cooperative game theory, the SHAP method can fairly decompose the complex output of the CGAN-DQN model into the dimensions of each input feature. Calculating the mean SHAP absolute value for each feature allows ranking features by their importance and contribution to the overall model output. This interpretability analysis provides strong quantitative support for validating the necessity of multi-source heterogeneous data fusion.

4 Results analysis

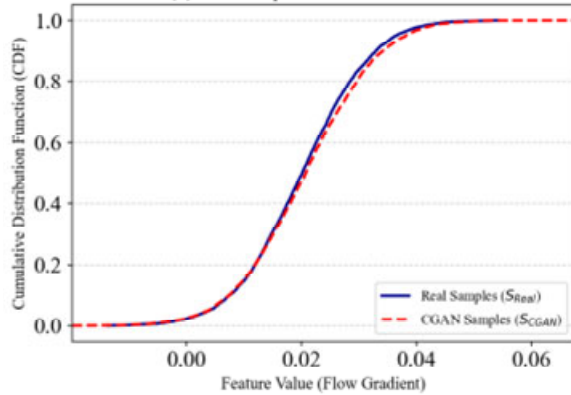
4.1 Effectiveness and quality analysis of the CGAN sample augmentation mechanism

This section aims to verify the effectiveness and quality of the CGAN in enhancing the quality of anomaly samples. To this end, it designed a scheme to compare the distributions of key features between real and synthetic anomaly samples, and a performance comparison scheme for models trained on augmented data versus those trained on original data. The former uses the KS test as a statistical metric to quantify the distributional similarity between synthetic and real samples; the latter uses recall and MAR as core evaluation metrics to quantify the performance of the early warning model. The results are shown in Figures 2 and 3.

Figure 2 Distribution quality and statistical similarity analysis of CGAN-enhanced samples, (a) comparison of cumulative distribution functions of pressure fluctuation characteristics (b) comparison of the cumulative distribution function of flow gradient characteristics (c) KS statistic D value of CGAN samples and real samples (see online version for colours)

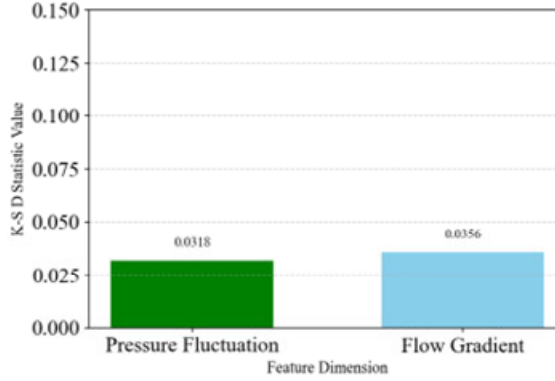


(a)



(b)

Figure 2 Distribution quality and statistical similarity analysis of CGAN-enhanced samples, (a) comparison of cumulative distribution functions of pressure fluctuation characteristics (b) comparison of the cumulative distribution function of flow gradient characteristics (c) KS statistic D value of CGAN samples and real samples (continued) (see online version for colours)



(c)

As shown in Figure 2, for the pressure fluctuation feature, the most sensitive feature in the fused feature set to pipeline anomalies, the KS statistic D value for both the real anomaly sample set and the synthetic sample set is 0.0318. The KS statistic D value for the flow gradient feature is 0.0356. This low D value demonstrates the CGAN mechanism's ability to generate high-fidelity anomalies, ensuring that the augmented training data accurately represents the complex, nonlinear characteristics of real faults. This high similarity in distribution indicates that the adversarial training mechanism forces the generator G to learn the underlying complex distribution of real anomaly data in the fused feature space, enabling synthetic data to accurately represent the physical characteristics of pipeline faults.

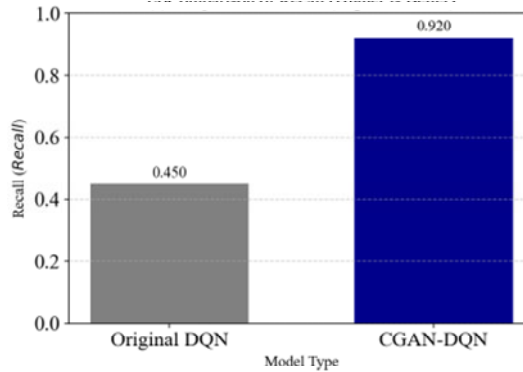
The results, shown in Figure 3, show that the CGAN-DQN model trained with CGAN-augmented data significantly improves Recall and significantly reduces MAR compared to the original DQN model trained on the original sparse data. The original DQN model achieved a recall of only 0.450, corresponding to a FAR of 0.550. The CGAN-DQN model trained with CGAN-augmented data improved recall to 0.920 and reduced MAR to 0.080. This demonstrates that the CGAN mechanism, by providing the DQN agent with sufficient, high-quality experience of abnormal states, addresses decision bias caused by small-sample learning and improves the Model's ability to detect and warn of critical pipeline failures.

4.2 Analysis of the contribution of fusion feature sets to the accuracy of early warning models

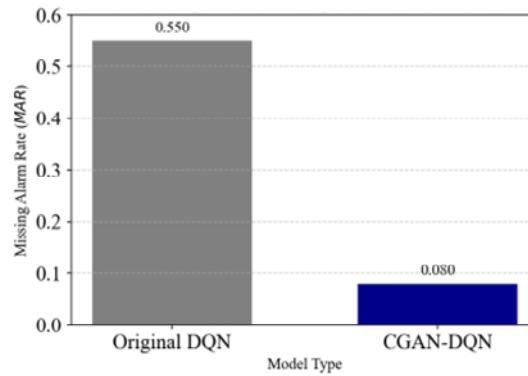
This section quantitatively analyses the contribution of multi-source heterogeneous data fusion to improving the accuracy of the early warning model through feature importance ranking. To this end, a feature importance ranking scheme was designed, employing the SHAP method to measure feature contributions. Specifically, the SHAP algorithm was applied to a trained CGAN-DQN model to calculate the contribution of each input feature in the fused feature set to the model's output (the Q-value of the warning action).

Features were then sorted in descending order based on the absolute value of their contribution to form a feature importance ranking. The results are shown in Figure 4.

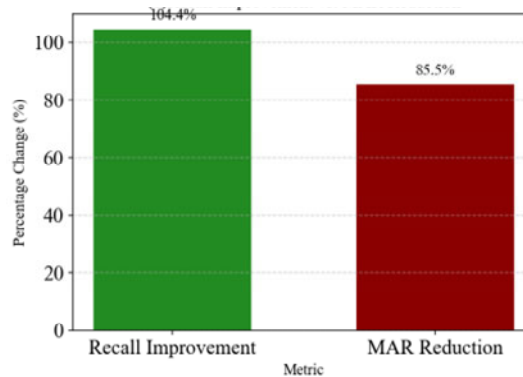
Figure 3 Comparison of the effect of CGAN sample enhancement on warning performance, (a) comparison of the recall rate of the early warning model for critical faults (b) comparison of MAR of early warning models (c) analysis of recall improvement and MAR reduction percentages (see online version for colours)



(a)

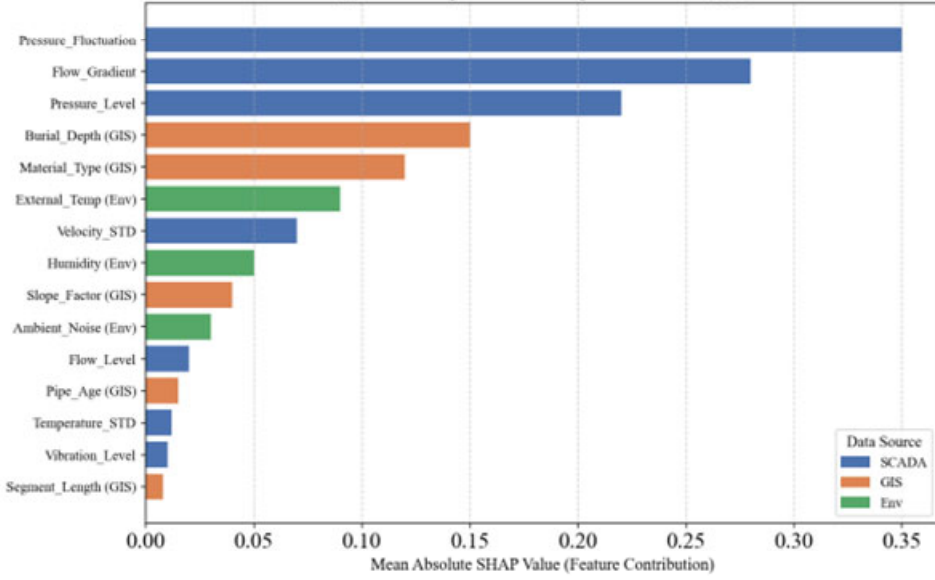


(b)

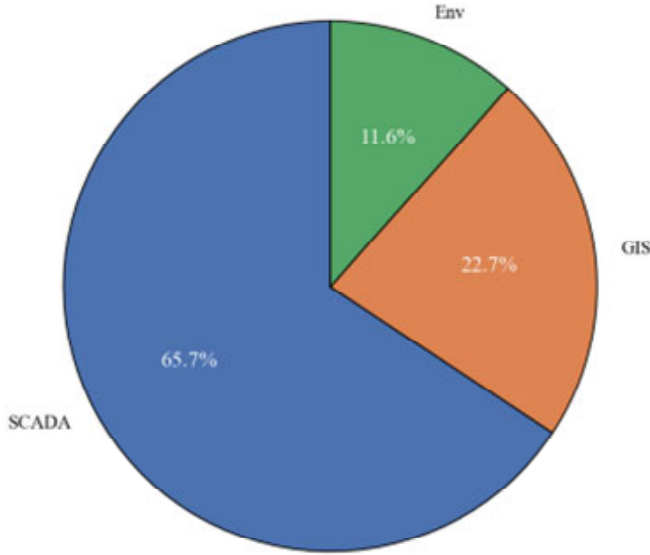


(c)

Figure 4 Analysis of the value contribution of multi-source heterogeneous features to early warning decision-making, (a) feature importance ranking (SHAP value) of the fused feature set (F_{Fused}) (b) contribution of different data source features (SCADA, GIS, environment) (see online version for colours)



(a)



(b)

The core metric of the experiment is the SHAP value, which quantifies each feature's contribution to the early warning decision. The results, as shown in Figure 4, indicate that the multi-source heterogeneous data features in the fused feature set jointly support high-precision early warning by the CGAN-DQN model. The real-time SCADA

monitoring data source has the highest SHAP contribution, with a total SHAP value of 0.962, accounting for approximately 65.7%. Meanwhile, static GIS data sources contribute 0.333, accounting for approximately 22.7%, and environmental monitoring data sources contribute 0.17, accounting for approximately 11.6%. According to the SHAP analysis, features derived from SCADA sensors show a strong and consistent ranking of features based on their importance; they include ‘pressure fluctuation’ (0.35) and ‘flow rate change’ (0.28), both of which are associated with time-related and pear in the same time frame as the time series data which represents a negative change in the behaviour of pipes. However, GIS-derived features of a static nature like ‘burial depth’ (0.15), ‘material type’ (0.12); and environmental factors like ‘outside air temperature’ (0.09) contributed meaningfully to the early warning Q. In the context of hydraulic process physics, both dynamic features can be viewed as ‘real-time’ representations of the hydraulic condition; a drop in hydraulic gradient is a leading indicator of localised loss of containment or an abrupt build-up of hydraulic pressure. With dynamic features related to SCADA being the leading indicators of how kinetic and potential energy are continuously changing within a pipeline, integration of static GIS features provides insight into how ongoing operational pressure impacts long-term structural vulnerabilities. Despite these findings, SCADA features remain the primary ones triggering the shift of the early warning agent from a normal to an abnormal state.

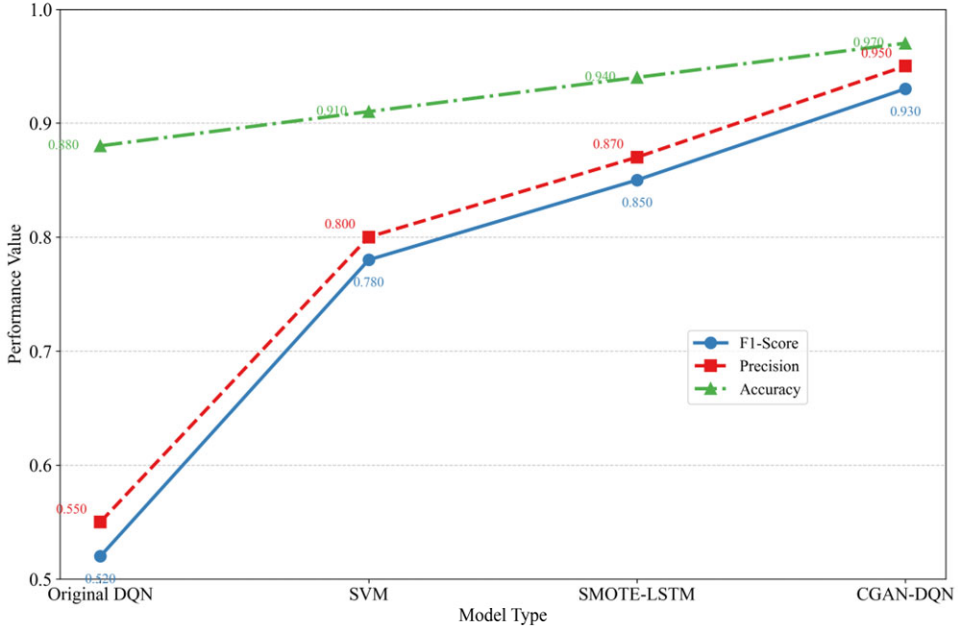
4.3 *Comprehensive performance comparison of CGAN-DQN and other baseline models*

To validate the comprehensive performance of the CGAN-DQN model, this paper designed a quantitative comparison scheme against all baseline models, using F1-score, precision, and accuracy as key performance indicators. Specifically, the baseline models (SVM, SMOTE-LSTM, and original DQN) and the CGAN-DQN model in this paper were evaluated on the same test set. The feature inputs of all models were uniformly normalised using a fused feature set. The results are shown in Figure 5.

As shown in Figure 5, the comprehensive performance comparison results of different early warning models show that the CGAN-DQN model shows the best performance in all indicators, with its F1-Score, precision and accuracy reaching 0.93, 0.95 and 0.97 respectively. This excellent performance is due to its integration of metadata-driven data governance modules and CGAN RL mechanism, which effectively solves the bottleneck problems of complex integration of multi-source heterogeneous data of NG pipelines and sparsity of abnormal samples; the high-quality synthetic abnormal data generated by CGAN balances the training set and significantly enhances the model’s ability to recognise rare fault modes. At the same time, the DQN early warning agent learns the optimal early warning strategy by maximising the cumulative reward, enabling the model to achieve more accurate risk identification and decision-making in a complex pipeline operation environment; in comparison, baseline models such as the original DQN model’s F1-score are only 0.52, clearly insufficient. While the SMOTE-LSTM model uses data augmentation, its F1 score is 0.85, its precision is 0.87, and its accuracy is 0.94, slightly inferior to the proposed Model. This indicates that relying solely on traditional sampling augmentation methods and LSTM time-series feature extraction remains inferior to the CGAN-DQN model, which combines a CGAN with RL strategies, in terms of robustness and recognition accuracy for complex anomaly patterns. Therefore, the

CGAN-DQN model provides the most reliable and high-precision solution for intelligent early warning of NG pipelines.

Figure 5 Key comprehensive performance comparison of CGAN-DQN and other early warning models (see online version for colours)



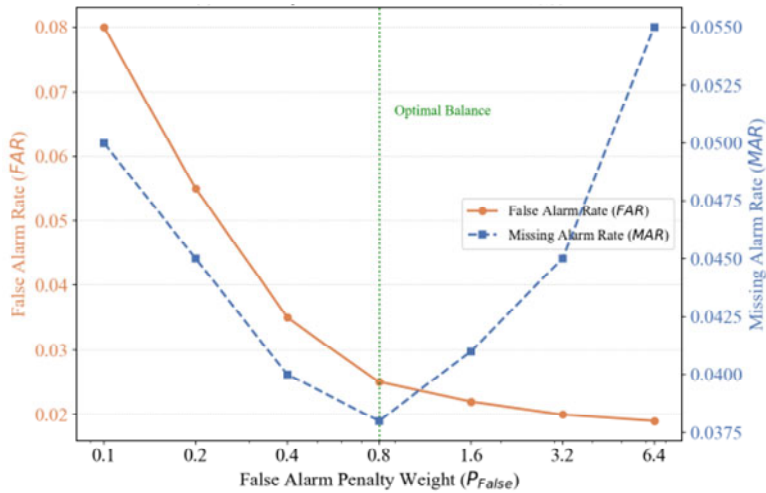
4.4 Risk and cost trade-off analysis of early warning decision-making strategies

This section aims to verify that our model can achieve a trade-off between economy and safety while ensuring high safety. To this end, it designed a sensitivity analysis scheme for the penalty weights in the reward function, using changes in the FAR and MAR as quantitative indicators. Specifically, while maintaining a high, constant value for missed-alarm penalties to ensure safety, it systematically adjusted the weights for false-alarm penalties and trained multiple CGAN-DQN models. These models were applied to the test set, and the results are shown in Figure 6.

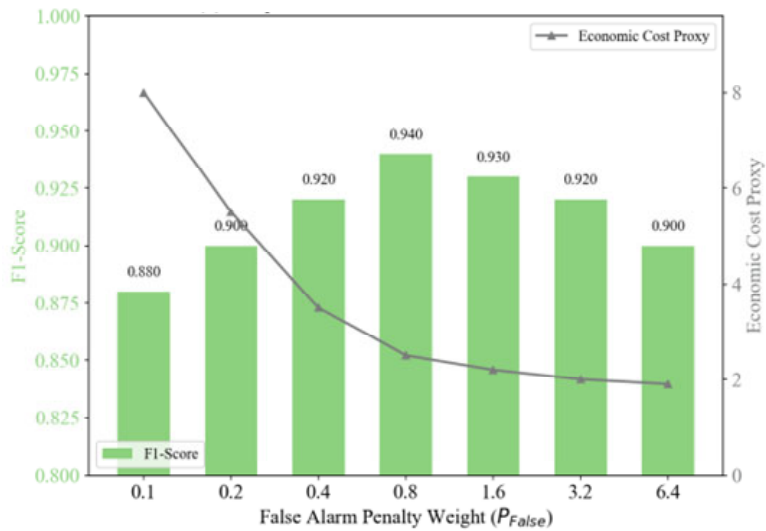
The trends in system safety and economic costs changed with changes in the false alarm penalty weight as shown in Figure 6. The FAR decreases as P_{False} increases from 0.1 to 6.4; the FAR is at its lowest of 0.019 when P_{False} is at its maximum of 0.8. The MAR is lowest at 0.038 when P_{False} is at 0.8, whereas it has increased slightly to 0.055 when $P_{\text{False}} = 6.4$. The FAR decreases, and the MAR initially decreases after the penalty weight for false alarms increases, because the DQN agent, due to the increased penalty weight, encourages a more conservative warning strategy, thereby lowering the FAR and raising the MAR. Furthermore, as P_{False} increases from 0.1 to 0, the FAR decreases 8. From 8.0, the economic cost index of the system has decreased dramatically to 2.5; therefore, it is believed that there is an initial optimal equilibrium between economic and safety. In addition, the F1-score is at its highest value of 0.94 when $P_{\text{False}} = 0.8$; therefore, the overall system performance has been maximised for that penalty weight. Although

BER decreased slightly after 0.8, MAE had increased somewhat due to the overly conservative strategy, so the F1-score dropped to 0.9 and economic cost dropped to 1.9. Therefore, the overall performance for false alarm punishment was suboptimal because of some loss of missed alarm performance; however, at $P_{\text{False}} = 0.8$, this system was the best trade-off for economy versus overall performance and safety.

Figure 6 Trade-off between economy and safety in early warning decision-making strategies, (a) sensitivity curve of false alarm penalty P_{False} to FAR and MAR (b) comprehensive performance vs. economic cost (see online version for colours)



(a)



(b)

Furthermore, we investigated the sensitivity of the DQN agent’s policy to the false alarm penalty weight P_{False} . Experimental results indicate a clear trade-off: as P_{False} increases from 1.0 to 5.0, the agent develops a more ‘conservative’ triggering policy. Specifically, the optimal action threshold for issuing an early warning shifts towards higher confidence levels, effectively reducing the false discovery rate. However, excessively high P_{False} values, (e.g., > 10.0) lead to a significant increase in the missed detection rate (false negatives), as the agent becomes reluctant to act on subtle anomaly signals to avoid high penalties. This behaviour confirms that P_{False} serves as a critical ‘sensitivity dial,’ allowing pipeline operators to balance the operational cost of unnecessary inspections against the safety risk of undetected failures.

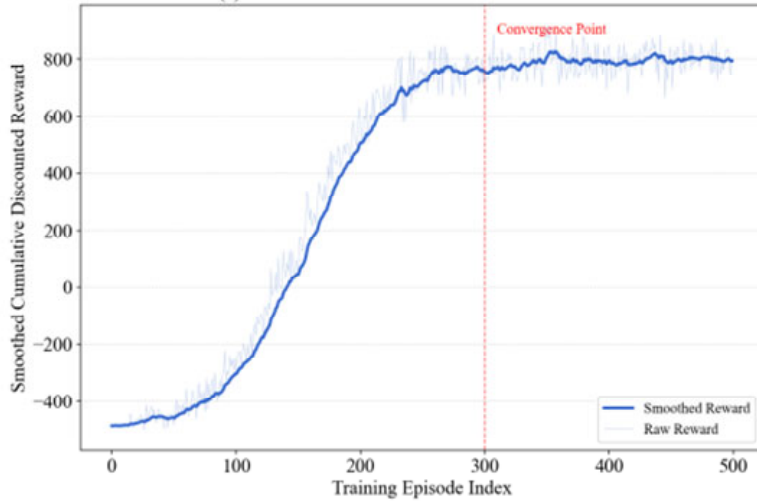
4.5 *Analysis of convergence and timeliness of DQN early warning decision*

This section analyses the dynamic performance of the CGAN-DQN agent from the perspectives of training convergence and the timeliness of warning decisions. To this end, it designed a curve analysis of the cumulative discounted reward as a function of iteration number during training, and a comparison of the average warning lead time of the CGAN-DQN agent with a baseline prediction model (SVM/SMOTE-LSTM). The former uses the cumulative discounted reward sliding average as a metric, while the latter uses the average warning lead time (AAT). The results are shown in Figures 7 and 8.

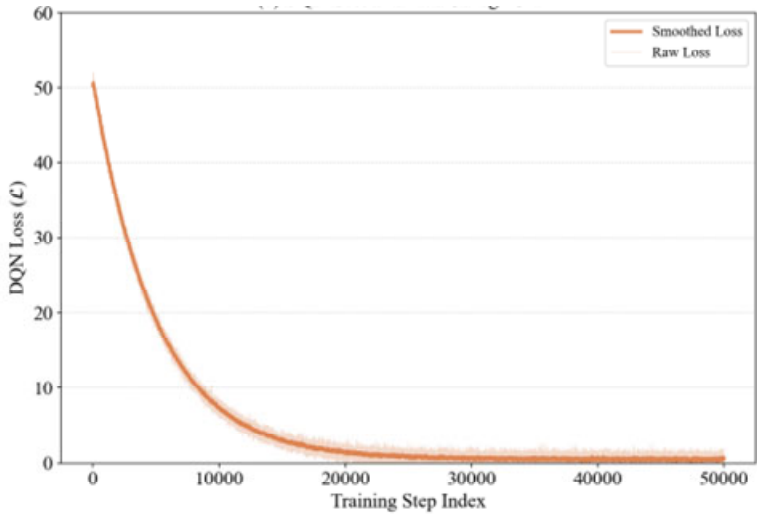
The results, shown in Figure 7, indicate rapid convergence and stability during the CGAN-DQN agent’s training. The sliding average of the cumulative discounted reward rapidly rises from a starting value of -485.72 at the beginning of training. It converges to a stable high of around 800 after approximately 300 episodes. This rapid and stable convergence trend demonstrates that the practical design of the reward function and penalty term guides the agent to rapidly explore the state space and lock in the optimal early warning decision-making strategy. Simultaneously, the DQN loss function curve rapidly decreases from a high of 50.26 at the beginning of training, then converges and stabilises around 0.5. This rapid convergence of the loss function indicates that the agent’s Q network has effectively and stably fitted the optimal Q value, demonstrating the CGAN-DQN mechanism’s high training efficiency and stable strategy.

The results, shown in Figure 8, indicate that the CGAN-DQN agent demonstrates significant advantages in proactive warning timeliness, with an average warning lead time (AAT) of 15.5 minutes, far exceeding those of the baseline models (5.2 minutes for the SVM model and 7.8 minutes for the SMOTE-LSTM model). The enhanced accuracy of proactive predictions can be attributed to the application of the CGAN-DQN system, which uses DRL to optimise long-term cumulative discounted rewards (Q-values). This enables it to identify weak risk signals earlier than traditional baseline models, which classify or predict based on a single moment in time. The CGAN-DQN lead time distribution is positively skewed with density located within the upper tail of the distribution near the average AAT of 15.5 minutes. Thus, the CGAN-DQN system does not only produce larger overall average lead times for warnings but also provides larger overall average lead times for nearly all abnormal events, thereby demonstrating the value of using RL techniques such as those used by CGAN-DQN to increase the timeliness of proactive pipeline risk warnings.

Figure 7 Convergence and stability analysis of the CGAN-DQN agent training process, (a) sliding average curve of cumulative discount rewards (b) DQN loss function change curve (see online version for colours)

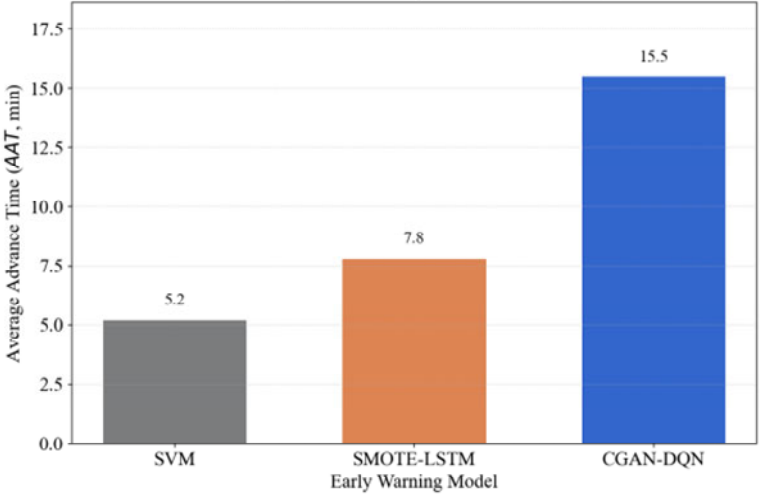


(a)

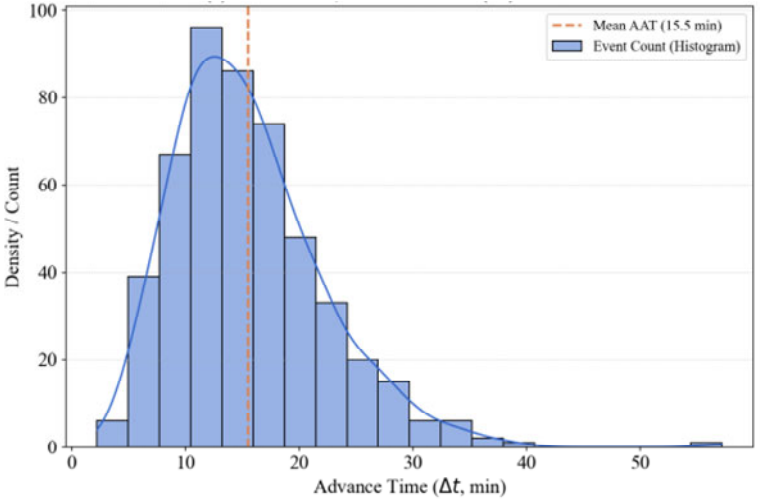


(b)

Figure 8 The timeliness advantage of the RL method in active early warning of pipeline risks, (a) comparison of average warning advance time (AAT) (b) CGAN-DQN warning lead time distribution (see online version for colours)



(a)



(b)

5 Conclusions

To solve the two major bottlenecks faced by NG pipeline intelligent early warning, namely the difficulty in integrating multi-source heterogeneous data and the sparsity of abnormal samples, this paper successfully built an intelligent early warning platform based on the integration of multi-source heterogeneous data governance and CGAN RL. Through the metadata-driven data governance module, the platform efficiently performs cleaning, alignment, and feature set construction for multi-source heterogeneous data,

such as SCADA and GIS, effectively overcoming the data integration problem. To address the scarcity of abnormal samples, the platform uses a CGAN to generate high-quality synthetic abnormal data, balance the training set, and significantly enhance the model's ability to identify rare faults. Finally, the enhanced data is used to train the DQN for early warning intelligence. By maximising cumulative reward, an optimal early-warning decision-making strategy was learned, enabling active and accurate identification of pipeline risks. This study's methodology was validated using experimental analyses and subsequently found to outperform all others' approaches to providing reliable decision making in NG pipelines through improved risk prevention/control and improved intelligence in pipelines, resulting in significant market opportunity. The CGAN-DQN model provided the best overall performance of all models tested with a mean F1 score of 0.93, an average alarm time (AAT) of 15.5 minutes, and a MAR of 0.08. These findings indicate that this approach is highly accurate, has very few missed alarms, and is timely.

In addition, due to the efficiency in computation of the proposed platform this solution can be scaled for industrial use. The offline training of the CGAN and DQN models require considerable GPU resources to execute, however online inference is extremely lightweight. The DQN agent and the data fusion mappings produced require very little memory and processing power, making them highly appropriate for deploying in edge computing environments using either station-based industrial PCs or IoT-based gateways. The architecture supports localised and millisecond-latency early warning responses without the need for continuous, high-bandwidth cloud-based communications, thus offering a robust solution for the geographically distributed NG pipeline networks.

Declarations

All authors declare that they have no conflicts of interest.

References

- Abdelhafidh, M., Fourati, M. and Chaari, L. (2023) 'Dynamic Bayesian network-based operational risk assessment for industrial water pipeline leakage', *Computers & Industrial Engineering*, Vol. 183, No. 1, pp.109466–109478.
- Adebisi, B., Aigbedion, E., Ayorinde, O.B. and Onukwulu, E.C. (2021) 'A conceptual model for predictive asset integrity management using data analytics to enhance maintenance and reliability in oil & gas operations', *International Journal of Multidisciplinary Research and Growth Evaluation*, Vol. 2, No. 1, pp.534–542.
- Aderubi, S., Appah, D., Okafo, E. and Aimike, V. (2023) 'A review of leak detection systems for natural gas pipelines and facilities', *International Journal of Energy Technology and Policy*, Vol. 13, No. 2, pp.9–19.
- Ahmadi, M., Aly, H. and Khashei, M. (2025) 'Enhancing power grid stability with a hybrid framework for wind power forecasting: integrating Kalman filtering, deep residual learning, and bidirectional LSTM', *Energy*, Vol. 2025, No. 1, pp.137752–137765.
- Ali, N.W. and Simmou, W. (2025) 'Framework for measuring the impact of a firm's artificial intelligence capability on creativity within the organization and performance of the firm', *Journal of Intelligence Technology and Innovation*, Vol. 3, No. 1, pp.74–92.
- Aljameel, S.S., Alomari, D.M., Alismail, S., Khawaher, F., Alkhudhair, A.A. and Aljubran, F. et al. (2022) 'An anomaly detection model for oil and gas pipelines using machine learning', *Computation*, Vol. 10, No. 8, pp.138–153.

- Al-Khalidi, M.R., Abdulsadda, A.T. and Al Zuhry, M.S. (2024) 'Oil pipeline leak detection in Iraqi oil fields based on 1DCNN', *International Journal of Robotics and Control Systems*, Vol. 4, No. 1, pp.312–333.
- Amini, I., Jing, Y. and Chen, T. (2022) 'Adaptive naive bayes classifier-based filter using kernel density estimation for pipeline leakage detection', *IEEE Transactions on Control Systems Technology*, Vol. 31, No. 1, pp.426–433.
- Asadi, Y. (2024) 'Employing machine learning in water infrastructure management: predicting pipeline failures for improved maintenance and sustainable operations', *Industrial Artificial Intelligence*, Vol. 2, No. 1, pp.8–22.
- Awuku, B., Huang, Y., Yodo, N. and Asa, E. (2024) 'Interpretable machine learning models for failure cause prediction in imbalanced oil pipeline data', *Measurement Science and Technology*, Vol. 35, No. 7, pp.076006–076021.
- Enemosah, A. and Chukwunweike, J. (2022) 'Next-generation SCADA architectures for enhanced field automation and real-time remote control in oil and gas fields', *Int. J. Comput. Appl. Technol. Res.*, Vol. 11, No. 12, pp.514–529.
- Fazle, A.B., Prodhan, R.K. and Islam, M.M. (2023) 'AI-powered predictive failure analysis in pressure vessels using real-time sensor fusion: enhancing industrial safety and infrastructure reliability', *American Journal of Scholarly Research and Innovation*, Vol. 2, No. 2, pp.102–134.
- Heckmann, R. (2023) 'Optimal control of a hybrid microgrid for hydrogen-based heat supply using deep reinforcement learning', *Clean Energy*, Vol. 7, No. 5, pp.940–951.
- Imam, M.H. (2025) 'Digital twin technology in utility infrastructure management: opportunities and challenges', *Journal of Sustainable Development and Policy*, Vol. 1, No. 1, pp.315–345.
- Jones, J.F., Raj, V., Abas, P.E., Petra, M.I., Sivan, D. and Kumar, K.S. et al. (2025) 'Application of artificial intelligence for asset integrity management of offshore oil and gas pipelines', *Life Cycle Reliability and Safety Engineering*, Vol. 2025, No. 1, pp.1–28.
- Kathiresan, G. (2025) 'Cybersecurity risk modeling in CI/CD pipelines using reinforcement learning for test optimization', *International Journal of Innovative Science and Research Technology*, Vol. 2025, No. 1, pp.15–25.
- Khlifi, A., Othmani, M. and Kherallah, M. (2025) 'A novel approach to autonomous driving using double deep Q-network-based deep reinforcement learning', *World Electric Vehicle Journal*, Vol. 16, No. 3, pp.138–156.
- Kornilova, Z. (2025) 'Digital twin roadways: integrating BIM, IoT and AI for intelligent infrastructure management', *European Science*, Vol. 2025, No. sge41-03, pp.1–144.
- Laayati, O., El Hadraoui, H., El Maghraoui, A., Guennouni, N., Mekhfouf, M. and Chebak, A. (2025) 'Metaheuristic-optimized forecasting in a smart edge-fog-cloud energy management framework: an industrial mining case study', *Results in Engineering*, Vol. 2025, No. 1, pp.107303–107334.
- Lu, K.T. (2025) 'User acceptance behavior analysis of multimodal generative AI', *Journal of Information and Computing*, Vol. 3, No. 2, pp.23–40.
- Machireddy, J.R. (2023) 'Data quality management and performance optimization for enterprise-scale ETL pipelines in modern analytical ecosystems', *Journal of Data Science, Predictive Analytics, and Big Data Applications*, Vol. 8, No. 7, pp.1–26.
- Mahapatra, A. and Patra, P.K. (2025) 'A unified and scalable machine learning framework for feature fusion in object classification using weighted PCA with adaptive concatenation and dynamic scaling', *Discover Computing*, Vol. 28, No. 1, pp.114–127.
- Mijwil, M.M. (2025) 'Post-quantum secure blockchain-based federated learning framework for enhancing smart grid security', *Iraqi Journal for Computers and Informatics*, Vol. 51, No. 2, pp.157–224.
- Nuruzzaman, M. and Rana, S. (2025) 'IoT-enabled condition monitoring in power distribution systems: a review of scada-based automation, real-time data analytics, and cyber-physical security challenges', *Journal of Sustainable Development and Policy*, Vol. 1, No. 1, pp.25–43.

- Ong, K.S.H., Wang, W., Niyato, D. and Friedrichs, T. (2021) 'Deep-reinforcement-learning-based predictive maintenance model for effective resource management in industrial IoT', *IEEE Internet of Things Journal*, Vol. 9, No. 7, pp.5173–5188.
- Pengchao, C. (2025) 'Advancements and future outlook of safety monitoring, inspection and assessment technologies for oil and gas pipeline networks', *Journal of Pipeline Science and Engineering*, Vol. 2025, No. 1, pp.100267–100281.
- Riaz, R., Han, G., Shaukat, K., Khan, N.U. and Zhu, H. (2025) 'A robust anomaly detector for imbalanced industrial internet of things data', *Journal of Computational Design and Engineering*, Vol. 12, No. 9, pp.46–60.
- Ribas, L.C., Casaca, W. and Fares, R.T. (2025) 'Conditional generative adversarial networks and deep learning data augmentation: a multi-perspective data-driven survey across multiple application fields and classification architectures', *AI*, Vol. 6, No. 2, pp.32–54.
- Rizvi, R.A. and Khalid, W. (2025) 'Unlocking innovation in manufacturing: the impact of data analytics maturity and knowledge-oriented leadership', *Journal of Management Science and Operations*, Vol. 3, No. 2, pp.17–43, DOI: <https://doi.org/10.30210/JMSO.202503.002>.
- Salehpour, S., Eskandari, A., Nedaei, A. et al. (2025) 'Two-stage deep Q-network reinforcement learning based ultra-efficient fault diagnosis and severity assessment scheme for photovoltaic protection', *Energy and AI*, Vol. 20, No. 1, pp.100512–100534.
- Sauber-Cole, R. and Khoshgoftaar, T.M. (2022) 'The use of generative adversarial networks to alleviate class imbalance in tabular data: a survey', *Journal of Big Data*, Vol. 9, No. 1, pp.98–113.
- Shu, M. (2025) 'Utilizing transfer learning for deep learning based image classification', *Journal of Intelligence Technology and Innovation*, Vol. 3, No. 1, pp.58–73.
- Singh, B. and Poblete, B. (2025) 'Offshore pipeline risk, corrosion, and integrity management with lessons learned', *Oil and Gas Pipelines: Integrity, Safety, and Security Handbook*, Vol. 2, No. 1, pp.1224–1256.
- Sultana, R. and Rozony, F.Z. (2025) 'A meta-analysis of artificial intelligence-driven data engineering: evaluating the effectiveness of cloud-based integration models', *Asrc Procedia: Global Perspectives in Science and Scholarship*, Vol. 1, No. 1, pp.193–214.
- Toufigh, V. and Ranjbar, I. (2024) 'Unsupervised deep learning framework for ultrasonic-based distributed damage detection in concrete: integration of a deep auto-encoder and isolation forest for anomaly detection', *Structural Health Monitoring*, Vol. 23, No. 3, pp.1313–1333.
- Ullah, I. and Mahmoud, Q.H. (2022) 'Design and development of RNN anomaly detection model for IoT networks', *IEEE Access*, Vol. 10, No. 1, pp.62722–62750.
- Villegas-Ch, W., Gaibor-Naranjo, W. and Sanchez-Viteri, S. (2024) 'Application of deep learning techniques for the optimization of industrial processes through the fusion of sensory data', *International Journal of Computational Intelligence Systems*, Vol. 17, No. 1, pp.187–199.
- Woldesellasse, H. and Tesfamariam, S. (2023) 'Data augmentation using conditional generative adversarial network (cGAN): application for prediction of corrosion pit depth and testing using neural network', *Journal of Pipeline Science and Engineering*, Vol. 3, No. 1, pp.100091–100125.
- Wu, C-H. (2024) 'Eco-economic predictions: applying QPSO-BiLSTM and attention mechanisms for accurate renewable energy forecasting', *Journal of Management Science and Operations*, Vol. 2, No. 4, pp.28–47, DOI: <https://doi.org/10.30210/JMSO.202402.011>.
- Yilmaz Eroglu, D. (2025) 'Tariff-sensitive global supply chains: semi-Markov decision approach with reinforcement learning', *Systems*, Vol. 13, No. 8, pp.645–663.
- Yutong, L., Ukwuoma, C.C. and Kemayou, L.N.M. (2025) 'Enhancing machine-learning-based anomaly detection in transmission lines models using robust features from neural networks', *Progress in Engineering Science*, Vol. 2025, No. 1, pp.100151–100164.
- Zhuang, H. (2025) 'The application of generative AI among photographers: formation and influencing factors of behavioral intention', *International Journal of Management and Organization*, Vol. 3, No. 3, pp.30–44.