PAPER

Automatic pipeline leak detection based on AST deep learning using a free-swimming detector with acoustic resonance cavity

To cite this article: Li Jian *et al* 2025 *Meas. Sci. Technol.* **36** 086126

IOP Publishing Measurement Science and Technology
Meas. Sci. Technol. **36** (2025) 086126 (14pp) https://doi.org/10.1088/1361-6501/adf90d

Automatic pipeline leak detection based on AST deep learning using a free-swimming detector with acoustic resonance cavity

Li Jian[1], Yin Xiaofeng[1], Zhang Tao[2], Ren Xiaoyu[1], Ma Jinyu[1] and Huang Xinjing[1,3,*]

[1] State Key Laboratory of Precision Measurement Technology and Instruments, Tianjin University, Tianjin, People's Republic of China
[2] Tianjin Institute of Metrological Supervision and Testing, Tianjin, People's Republic of China
[3] Guangxi Key Laboratory of Automatic Detecting Technology and Instruments, Guilin University of Electronic Technology, Guilin, People's Republic of China

E-mail: huangxinjing@tju.edu.cn

Received 15 May 2025, revised 16 July 2025
Accepted for publication 7 August 2025
Published 19 August 2025

Abstract

The free-swimming in-pipe spherical detector (SD) has found practical application in pipeline leak detection due to its proximity to the leak points, low risk of obstruction, and low cost. Acoustic resonance air cavity design enables the SD to capture small leak acoustic signals with high sensitivity, but it simultaneously renders the SD highly vulnerable to collision noise interference, increasing the difficulty of leak identification. For long-time-series acoustic data acquired by the SD, traditional manual analysis methods are inefficient and difficult to ensure accuracy. Additionally, valid leak samples of field pipelines are insufficient, and the actual leak events represent an extremely low proportion, which makes the collection result in a small sample imbalance dataset. To address these issues, this paper proposes an automatic identification method for pipeline leak sounds collected by the SD via incorporating audio spectrogram transformer (AST) deep learning. First, a pipeline leak voice dataset is constructed using the field pipeline acoustic signals captured by the SD, which contained three types of acoustic events: normal rolling, collision, and leak. Second, the problems of category imbalance and sample insufficiency are addressed by combining data augmentation and transfer learning. Finally, the AST model is applied to pipeline leak sound identification and leak point localization. Experiments under different pressure conditions show that this method can accurately identify and locate 1 mm aperture leaks, and the localization errors under 1 MPa and 0.5 MPa conditions are 2.46 m and 3.07 m, with relative errors of 0.37% and 0.47%, respectively. This research provides a new solution for the automation and intelligence of pipeline leak detection and localization, with good engineering application prospects.

Keywords: pipeline leak detection, acoustic event classification, audio spectrogram transformer (AST), transfer learning

* Author to whom any correspondence should be addressed.

1. Introduction

Pipeline leaks can lead to severe environmental damage and economic losses. For instance, oil pipeline leaks may result in substantial oil seepage into soil and water, causing ecosystem destruction and economic repercussions [1, 2]. Water supply pipeline leaks can lead to water resource wastage and contamination of drinking water [3]. However, existing traditional detection methods exhibit significant limitations in detecting small leaks. Mass-volume balance methods [4] are simple to set up but cannot localize the leaks; Negative pressure wave methods [5] can quickly and accurately locate large and sudden leaks but are ineffective for slow, small, or continuous leaks that do not generate pressure changes. Vibration monitoring methods [6] can effectively identify leaks but are costly, and their accuracy is limited by significant signal attenuation. Fiber optic sensing technology [7] can precisely locate leaks by measuring vibrations and thermal changes along the pipeline but requires optical fibers to be embedded during pipeline construction. Acoustic methods are simple to implement, using devices such as listening rods [8, 9], and acoustic emission monitoring [10–14] for leak detection and localization. However, these approaches are typically limited to accessible pipeline sections and cannot achieve comprehensive monitoring of deeply buried pipelines due to access limitations and signal attenuation.

In contrast, free-swimming in-pipe inspection methods have developed significantly in recent years [15, 16]. Among various free-swimming detection solutions, spherical detectors (SDs) have received widespread attention due to their simplicity and practicality. Early research focused on effectively sealing sensors and electronic components within spherical shells designed for high-pressure environments. Traditional designs utilized an inner aluminum shell covered by an outer polyurethane protective layer. While this enhanced pressure resistance also significantly attenuated leak sounds entering the sphere, resulting in reduced leak detection sensitivity. In recent years, Tianjin University [17] has redesigned the SD's structure by removing the polyurethane layer and has proposed integrating electronic components into the spherical shell to create an air cavity, utilizing acoustic resonance to achieve high-sensitivity detection. In a static state, this design achieves detection sensitivity as high as 0.164 l min^{-1} for small continuous leaks. However, the high-sensitivity characteristics cause SD with acoustic resonance air cavity (ARAC) to simultaneously capture substantial background noise, while collision vibrations generated during SD rolling within pipelines further interfere with leak signal identification [18]. Furthermore, during actual inspections, the SD often operates inside pipelines for tens of hours or even days, generating large volumes of long-time-series acoustic data. Relying on manual analysis of this data is time-consuming and can lead to inconsistencies in results, which severely limits detection efficiency and reliability.

The acoustic signals collected by the SD in pipeline leak detection are similar to data acquired by other leak monitoring sensors, such as pressure sensors, vibration sensors, acoustic sensors, and distributed fiber optic sensors [4, 19, 20]. These signals share similar characteristics: they are continuous one-dimensional time series signals and frequently contain complex environmental noise. In particular, the movement of the SD within the pipeline causes the collected acoustic signals to exhibit significant non-stationary characteristics and strong noise features. Currently, widely applied signal analysis methods for these one-dimensional leak detection data include short-time Fourier transform (STFT) [20, 21], wavelet transform [21, 22], and modal decomposition [20]. However, these methods have several limitations. They require manual selection and adjustment of multiple key parameters, such as time-frequency resolution, wavelet basis functions, and decomposition levels. When pipeline operational parameters like pressure and flow velocity change, experts need to readjust the analysis parameters. For signals collected by the SD, the large data volume and continuously changing signal characteristics during sphere movement make manual parameter adjustment approaches severely impact detection efficiency [20, 21].

In recent years, deep learning methods have gained widespread application in pipeline leak detection due to remarkable capabilities in feature extraction and data processing. The main deep learning frameworks used for this purpose are convolutional neural networks (CNNs), recurrent neural networks (RNNs), and models based on the Transformer architecture. CNNs achieve multi-scale analysis of raw data through hierarchical feature extraction structures [23, 24]. Siddique *et al* [25] achieved a detection accuracy of 99.22% by converting one-dimensional acoustic emission signals from pipelines into two-dimensional images and applying CNNs. RNNs and their variants (LSTM, GRU) capture temporal dependencies through memory mechanisms. Zhang *et al* [26] proposed a dual-layer LSTM structure to analyze pipeline pressure signals collected by sensors, achieving a detection accuracy of 99% in real natural gas pipeline networks. Models based on attention mechanisms can directly establish long-range dependencies between signals and support parallel computation. Liu *et al* [27] proposed a time series Transformer model to process acoustic wave signals collected from pipe walls by acoustic sensors, achieving a detection accuracy of 88.78% in actual water supply networks.

Deep learning models typically require large amounts of training data to optimize their parameters, learn features, and ensure generalization. However, in the field of pipeline leak detection, the rarity of leak events and challenges in data collection result in small data problems for deep learning-based detection methods. To tackle this issue, researchers have proposed a variety of solutions. Common approaches include data augmentation [28, 29] and transfer learning [30, 31]. Wang *et al* [32] enhanced acoustic signals through denoising and sample generation to mitigate the effects of insufficient samples and background noise, achieving 93.02% leak detection accuracy in pressure pipelines. Glynis *et al* [33] utilized LSTM neural network models to transfer knowledge from fixed sensor configurations to new sensor scenarios, attaining 98.1% accuracy in detecting water pipeline leaks under data-scarce conditions. Wu *et al* [34] proposed an intelligent diagnostic framework combining hybrid feature selection and

transfer learning techniques. This approach enables knowledge transfer between laboratory and field data, resulting in a 90% accuracy for pipeline defect diagnosis even with small sample sizes.

There are some challenges in using the SD with ARAC inside for detecting small leaks in water pipelines:

- The ARAC design enables the SD to have high sensitivity and capture small leak signals, but this design also makes the SD highly susceptible to collision and environmental noise, increasing the difficulty of leak identification.
- Existing datasets are small in scale and limited in scenarios, lacking coverage of complex situations in actual engineering environments.
- Long-sequence acoustic data collected by SD primarily relies on traditional methods (spectral analysis, energy detection), requiring time-consuming expert analysis with inconsistent results. Meanwhile, deep learning approaches also have limitations: CNNs are constrained by local receptive fields and cannot effectively capture long-range acoustic dependencies; LSTMs suffer from gradient vanishing and sequential processing constraints that limit feature extraction efficiency.

To address these issues, this research focuses on the following key tasks:

- Sound signals are collected using the SD with the ARAC inside, leading to the establishment of a pipeline leak voice dataset, which includes data from normal rolling, collisions, and leak events.
- The challenges of data scarcity and class imbalance are addressed through data augmentation and transfer learning techniques.
- Leak acoustic features are extracted and analyzed using the AST model, which overcomes the limitations of traditional approaches including CNN and LSTM through global feature modeling and parallel processing, enabling the effective identification and localization of small pipeline leaks.

The subsequent sections in this paper are arranged as follows. Section 2 presents the proposed methodology, including the AST model architecture and few-shot learning strategy. Section 3 describes the data acquisition process using the SD, including experimental methodology, dataset construction, and model training. Section 4 presents the results and discussion, analyzing the detection and localization performance under various operating conditions. Finally, section 5 provides conclusions.

2. Proposed method

The SD leak detection method based on AST is shown in figure 1. The free-swimming SD is thrown into the pipeline, where it moves forward propelled by the fluid flow, collecting acoustic signals to identify leaks. Due to collection constraints and the rarity of pipeline leak points, the sound data samples acquired by the SD are few, with an extremely low proportion of leak sounds. This leads to two main problems. First, field pipeline acoustic signals collected by the SD exhibit a class imbalance between leak and non-leak sounds, causing models to bias toward predicting the majority class. Therefore, data augmentation methods utilizing signal concatenation and background mixing were implemented to achieve class balance. Secondly, training the model from scratch using only a few samples of acoustic data makes it difficult to learn the fundamental features of sound signals. Hence, transfer learning with AST models pre-trained on large-scale datasets is employed to enhance feature extraction capabilities and model generalization performance under small-sample conditions. The leak detection method proposed in this paper comprises three key steps: data acquisition using the SD with ARAC to collect in-pipeline acoustic signals, data processing through augmentation and transfer learning to address sample scarcity and class imbalance, and AST-based feature extraction and classification for leak detection and localization.

2.1. AST model

This paper adopts the AST model as the acoustic event classifier for several key reasons. First, the AST model introduces the Transformer architecture to the audio domain, effectively capturing the sequential dependencies between audio features through the self-attention mechanism. Second, the model processes spectrograms using patch embedding, which preserves local structural information while also achieving global modeling. This characteristic is particularly suitable for scenarios in this research where both local impact features and continuous leak features need to be considered simultaneously. Lastly, the Transformer's parallel computing capabilities lend the model good inference efficiency when deployed in practical applications, which supports its use in engineering contexts.

The leak localization model based on AST is shown in figure 2. This model comprises four key components: the data preprocessing module, the Patch Embedding module, the Transformer encoder module, and the classification output module.

In the data preprocessing stage, input audio waveforms are transformed into a Mel-spectrogram using a Mel Transform. Each Mel-spectrogram is then normalized using its own mean and standard deviation, which helps to reduce intensity differences between various audio samples. The computation process of Mel-spectrograms sequentially includes STFT and Mel filter bank mapping. The Mel filter bank consists of a series of triangular filters with central frequencies distributed according to the Mel scale. By multiplying the STFT squared magnitude spectrum with the frequency response of each filter and summing, the energy distribution of each time frame across different Mel frequency bands can be obtained, which is the Mel-spectrogram:

$$X\mathrm{mel}\,(m,l) = \sum_{k=0}^{N/2} |XSTFT(k,m)|^2 Hl(k) \tag{1}$$

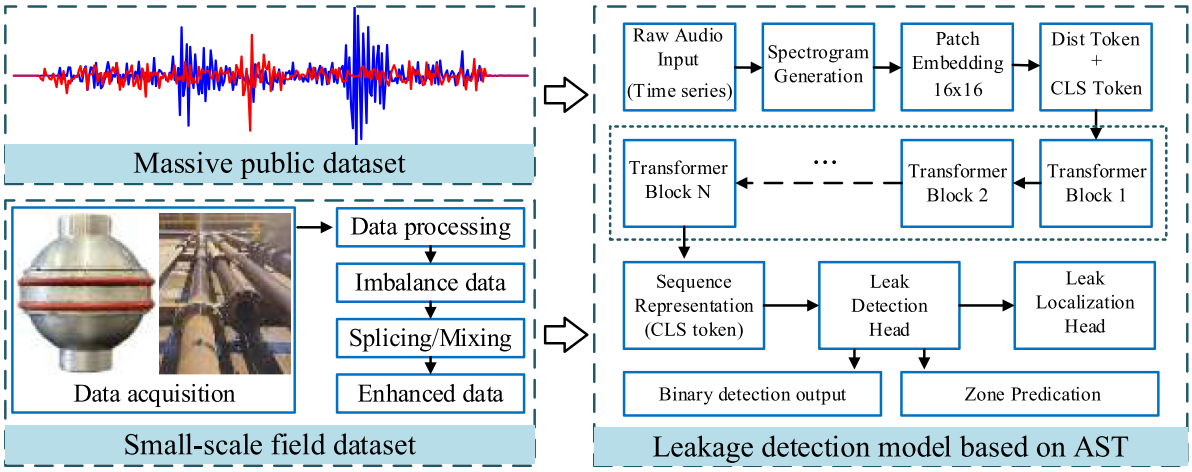


Figure 1. A leak detection approach based on audio spectrogram transformer.

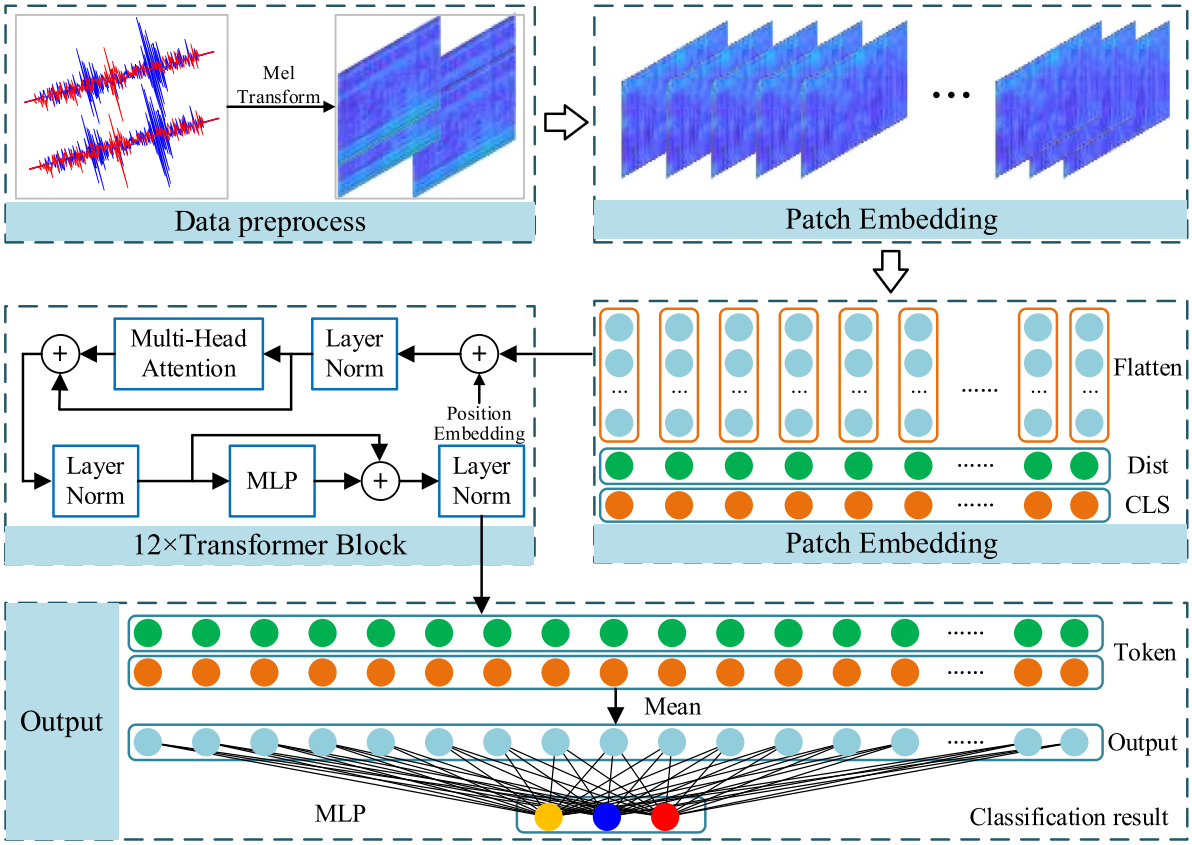


Figure 2. Audio spectrogram transformer model structure.

where, $H_l(k)$ represents the frequency response of the kth Mel filter.

In the Patch Embedding module, convolution operations are used to output the Mel-spectrogram as patches, and this process can be expressed as:

$$Z = \mathrm{Conv}\,(X; k, s, \mathrm{cout}) \tag{2}$$

$$E = \mathrm{Flatten}\,(z) \tag{3}$$

where, X represents the input Mel-spectrogram, k is the convolution kernel size, s is the convolution stride, C_{out} is the output channel. Then, each patch is converted into a one-dimensional vector directly through a Flatten operation, and two special tokens are added at the beginning of the sequence: a classification token (CLS token) and a distillation token (Dist. token). The CLS token is used for the final classification task, while the Dist. token is used for knowledge distillation. The process of adding tokens can be expressed as:

$$E' = [e_{\text{cls}}; e_{\text{dist}}; E]. \tag{4}$$

Finally, learnable position encoding E_{pos}, is added to obtain the final patch embedding representation:

$$x_{\text{p}} = E + E_{\text{pos}}. \tag{5}$$

Through the attention calculation formula, weights are dynamically allocated to measure the proportion of importance of different parts, and the calculation method is:

$$\text{Attention}(Q, K, V) = \text{softmax}(\frac{QK^{\text{T}}}{\sqrt{d_k}})V \tag{6}$$

where, Q, K, V represent the query, key, and value matrices respectively, d_k is the dimension of the key vector. Each Transformer block consists of two sub-layers: a multi-head self-attention mechanism and a feed-forward neural network. Additionally, each sub-layer employs layer normalization and residual connection,

$$x' = \text{LayerNorm}(x + \text{MultiHeadAttention}(x)) \tag{7}$$

$$\text{y}' = \text{LayerNorm}\,(x' + MLP\,(x')). \tag{8}$$

Finally, in the output module, the features are normalized through layer normalization, and then all the tokens are averaged and pooled:

$$h = \frac{1}{N}\sum_{i=1}^{N} xi \tag{9}$$

where, N is the number of tokens, and x_i is the feature vector of the ith token. Finally, classification results are output through a multi-layer perceptron:

$$y = \text{softmax}\,(W2s\,(W1h + b1) + b2) \tag{10}$$

where, W_1, W_2 are weight matrices, b_1, b_2 are bias terms, σ is the activation function.

2.2. Strategy for small sample problem

To address the small sample learning problem in pipeline leak detection, this paper adopts a dual-strategy approach combining data augmentation and transfer learning, as shown in figure 3.

Data augmentation expands training sets by generating synthetic samples, effectively addressing class imbalance, improving minority class recognition, and enhancing model adaptability to complex environments. Methods include two parts: signal concatenation and background mixing.

Transfer learning enables knowledge transfer from a data-rich source domain to a data-scarce target domain [35]. Let $D_{\text{s}} = \{X_{\text{s}}, Y_{\text{s}}\}$ represent the source domain and $D_{\text{t}} = \{X_{\text{t}}, Y_{\text{t}}\}$ represent the target domain (pipeline acoustic events), where X_{s}, X_{t} are the feature spaces and Y_{s}, Y_{t} are the corresponding label spaces. The source and target tasks can be defined as:

$$Y_{\text{s}} = f_{\text{s}}\,(X_{\text{s}}, \theta_{\text{s}}) \tag{11}$$

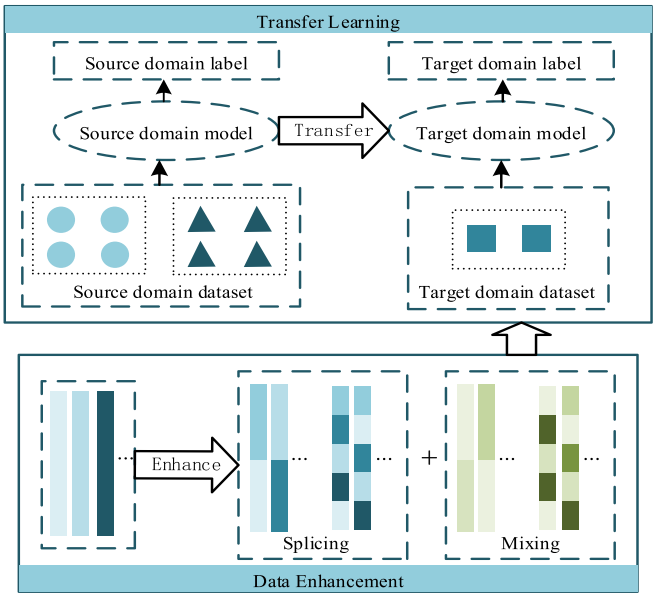


Figure 3. Data enhancement and transfer learning.

$$Y_{\text{t}} = f_{\text{t}}\,(X_{\text{t}}, \theta_{\text{t}}) \tag{12}$$

where in, f_{s} and f_{t} represent the mapping functions for source and target tasks respectively, and θ_{s}, θ_{t} are the corresponding model parameters. The transfer learning objective is to leverage the pre-trained parameters θ_{s} from the source domain to initialize the target model parameters:

$$\theta_{\text{t}} = \theta_{\text{s}} + \Delta\theta \tag{13}$$

where $\Delta\theta$ represents the parameter adaptation for the target task. The model adaptation process involves three key components:

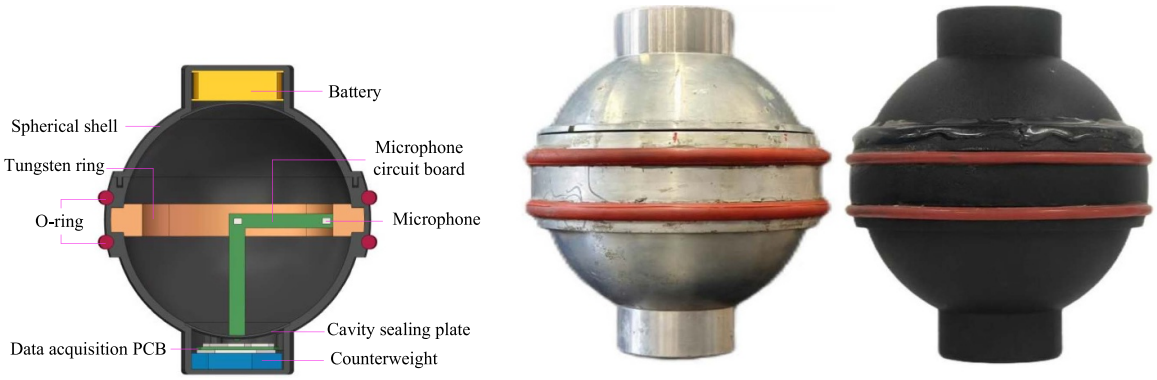
- The pre-trained Transformer encoder layers from the source domain are directly transferred to preserve learned acoustic feature representations.
- The original classification head is replaced with a new head specifically designed for pipeline acoustic classification.
- All parameters are then jointly optimized on the target dataset using cross-entropy loss function to minimize classification error across target samples.

The above adaptation process allows the model to leverage source domain knowledge while specializing for pipeline acoustic events.

3. Data acquisition of the SD in leak detection

3.1. Experiment method and process

The SDs [18] employed in the experiment are illustrated in figure 4, with a shell composed of 6063 aluminum alloy and PA12 nylon plastic materials. The SD rolls forward under the propulsion of fluid within the pipeline. The SD rolls around



(a) Internal structure of the SD (b) Aluminum alloy SD (c) PA12 SD

Figure 4. Structure of the SD.

an axis with greater rotational inertia. A tungsten ring counterweight is installed at the equator of the SD to achieve a rotational inertia ratio of 1.25 between the rolling axis (the tungsten ring axis) and the other two axes. This configuration ensures that the SD rolls stably around the tungsten ring axis, keeping the contact points between the SD and the pipe wall always positioned on both sides of the equator [36]. To improve the stability of SD rolling and reduce collision noise, flexible silicone O-rings are installed on both sides of the equator. Notably, the surface of the SD is not covered with additional vibration-damping materials to prevent any attenuation of the leak sound.

The ratio of the SD outer diameter to the pipeline inner diameter is between 0.6 and 0.8 to ensure the SD obtains the appropriate thrust generated by water flow in the pipeline [37]. In this experiment, the inner diameter of the pipeline being tested measures 193 mm, while the outer diameter of the SD is 133 mm. The SD contains a spherical cavity that is designed to resonate acoustically when stimulated by external leak sounds, which enhances its sensitivity for leak detection. The spherical cavity diameter inside the SD is designed to be 116 mm, so the first three resonance frequencies of the SD, fr1, fr2, and fr3, are 2095 Hz, 3384.6 Hz, and 4279.9 Hz, respectively. These frequencies fall within the low-frequency band, where broadband leak sounds typically exhibit higher energy.

When the SD approaches the leak point from a distance, the leak sound source is positioned perpendicular to the SD rotation axis, with leak sounds impinging on the SD approximately perpendicular to its rolling axis [38]. Based on this characteristic, the microphone is arranged near the inner wall of the spherical cavity on the equatorial plane to improve the sensitivity of leak sound detection. The data acquisition and storage system inside the SD integrates an MCU, a triaxial accelerometer, a TF storage card, and a PDM signal demodulation module, achieving 200 Hz sampling for acceleration and 8 kHz sampling for acoustic signals.

The experiment for detecting pipeline leaks is illustrated in figure 5. The leak detection experimental conditions include pressures of 0.5MPa and 1 MPa and flow velocities of 0.4 m s^{-1} and 1 m s^{-1}. The pipeline is constructed from low-carbon steel and has a total length of 657.46 m, with an inner diameter of 193 mm and an outer diameter of 219 mm. The experimental procedure is as follows: First, the SD is placed in the launching tube, and the launching cylinder is closed to ensure the pipeline is sealed. Next, the pressure is set to the desired value (either 0.5 MPa or 1 MPa) using a pressure adjustment device, while the water pump frequency is adjusted to maintain the flow velocity at the predetermined value (either 0.4 m s^{-1}or 1 m s^{-1}). After this, the launching valve is opened, allowing the SD to roll forward through the pipeline driven by the water flow, while it collects acoustic signals. Finally, once the SD has inspected the entire length of the pipeline and returned to the receiving cylinder, the receiving valve is closed, and the receiving cylinder is depressurized. The SD is then removed to download the collected data for analysis.

3.2. Dataset construction and augmentation

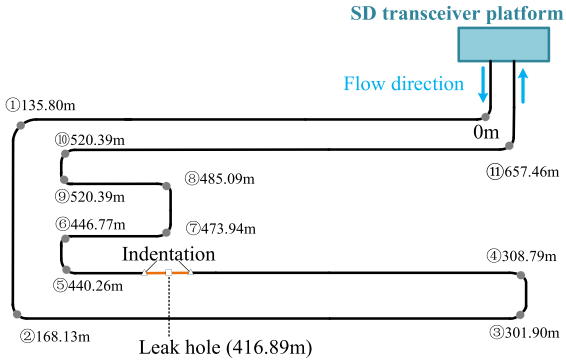
The leak point is located at 416.89 m of the pipeline, with a leak hole diameter of 1 mm. The experiment utilized intelligent spheres made of aluminum alloy and PA12. Multiple comparative experiments were conducted under varying pipeline pressures, flow velocities, and conditions with and without leaks. This approach was taken to comprehensively evaluate the detection performance and adaptability of the SD across different working scenario. The specific experimental parameters and the number of trials are detailed in table 1.

During pipeline operation, the SD primarily encounters three types of acoustic events: normal rolling, collision, and leak. In straight pipe sections without leaks, the sound produced by the SD rolling stably is characterized by a steady and quiet signal. When the SD encounters bends or wall indentations, it generates collision sounds that last from 0.1 to 0.5 s. In some instances, due to instability, these collision sounds may continue for 3–5 s. When the SD passes through a leak point, it continuously collects sound signals indicating the presence of a leak. The duration of these leak sound signals is influenced by both the rolling speed of the SD and the distance that the

(a) Photograph of field pipeline

(b) Photograph of leak point

(c) SD transceiver platform



(d) Pipeline layout

Figure 5. Pipeline leak detection experiment.

Table 1. Experimental parameters.

SD material	Pipeline pressure	Pipeline flow velocity	Leak condition	Number of tests
Aluminum alloy	1 MPa	1 m s^{-1}	No leak	1
Aluminum alloy	1 MPa	1 m s^{-1}	Leak	2
Aluminum alloy	1 MPa	0.4 m s^{-1}	Leak	2
Aluminum alloy	0.5 MPa	1 m s^{-1}	Leak	2
PA12	0.5 MPa	1 m s^{-1}	No leak	1
PA12	0.5 MPa	1 m s^{-1}	Leak	2
PA12	0.5 MPa	0.5 m s^{-1}	Leak	1
PA12	0.5 MPa	0.4 m s^{-1}	Leak	1

sound travels, with previous experiments showing a minimum duration of 20 s. Considering that collision sounds and leak sounds may co-occur near the leak point in actual situations, such composite acoustic events are uniformly labeled as the leak category. Considering the characteristics of sound events in practical applications, a 5 s window is employed to segment continuous audio for sample generation. Therefore, during the data processing phase, the continuous pipeline acoustic signals collected by the SD were segmented into a series of 5 s windows with a 2.5 s step size, generating 2359 original samples. The audio samples were annotated, resulting in an initial dataset comprising 1499 normal rolling events, 612 collision events, and 248 leak events. To ensure the independence of the validation set, acoustic data from one complete experimental run was separately selected and processed using the same segmentation method to form the validation set. This approach guarantees that the validation set is temporally and experimentally independent from the training data, preventing data leakage and ensuring reliable model evaluation. To address the class imbalance issue, two primary data augmentation approaches were implemented: signal concatenation and background mixing.

The signal concatenation approach, targeting the continuity characteristics of leak and normal rolling events, involved dividing the original audio samples into smaller segments and subsequently combining them to form new samples. Two specific strategies were employed: (1) dual-segment concatenation, where original audio clips were subdivided into 2.5 s

Table 2. Statistics of data augmentation methods and sample quantities.

Category	Data augmentation method	Training set quantity		Validation set quantity	
		Quantity	Subtotal	Quantity	Subtotal
Normal rolling	Original samples	1324	2000	175	80
	Dual-segment concatenation	338		30	
	Five-segment concatenation	338		45	
Collision	Original samples	563	2000	49	330
	Dual-collision mixing	717		16	
	Triple-collision mixing	720		15	
Leak	Original samples	217	2000	31	250
	Dual-segment concatenation	445		70	
	Five-segment concatenation	446		75	
	Rolling background mixing	446		80	
	Collision sound mixing	446		73	

segments, with two segments from different time periods being combined; (2) five-segment concatenation, where original audio clips were subdivided into 1-second segments, with five randomly selected segments from different time periods being combined.

To enhance the model's adaptability to complex scenarios, two background mixing strategies were designed specifically for the leak category: (1) background noise mixing, where normal rolling sounds were superimposed onto leak sounds as background noise, simulating actual detection environments; (2) collision sound mixing, where collision sounds were overlaid onto leak sounds, enhancing the model's recognition capability for composite acoustic events. For the collision category, data augmentation was achieved by randomly inserting multiple collision sounds into normal operating condition audio recordings.

Through these data augmentation methods, the dataset was expanded from the original 2359 samples to 6660 samples. The expanded dataset comprised 6000 training samples and 660 validation samples. The detailed statistics for each category are presented in table 2, demonstrating that data augmentation significantly mitigated the sample imbalance issue between categories.

3.3. Model transfer learning and training

A GPU-accelerated training approach was implemented. A personal computing terminal equipped with an NVIDIA GeForce RTX 4070 Ti SUPER 16 G graphics card served as the training platform for the AST model. The model training program was developed using the PyTorch 2.1.2 deep learning framework. The model training program was executed in a Windows 11 operating system environment.

To avoid training the deep network from scratch, a transfer learning approach was employed to accelerate the training process, with an AST model pre-trained on the large-scale AudioSet dataset selected as the foundation model [39]. AudioSet contains 632 classes of audio events, encompasses over 5,800 h of total duration, and features 10-second sample lengths. This pre-trained model possesses excellent general feature extraction capabilities for audio signals, significantly enhancing performance in small-sample dataset scenarios.

Table 3. Audio spectrogram transformer model training parameters.

AST model training parameters	Parameter value
Input Mel-spectrogram dimensions	(128, 512)
Convolutional	Conv(16 × 16);Stride(10, 10)
Patch shape	(50, 12)
Number of transformer encoder layers	12
Number of multi-head attention heads	12
Hidden layer dimension	768
Batch size	24
Virtual batch size	48
Number of training epochs	50
Learning rate scheduling strategy	Cosine annealing
Initial learning rate	$1e^{-5}$
Minimum learning rate	$1e^{-6}$
Loss function	Cross-entropy loss
Optimizer	Adam

During the specific training process, the primary parameter settings are presented in table 3. Considering that the AST model had already been pre-trained on the large-scale AudioSet dataset, this paper adopted a transfer learning strategy, fine-tuning the pre-trained model to adapt it to the pipeline acoustic event classification task. Due to the model's established feature extraction capabilities, setting the training for 50 epochs was sufficient to achieve satisfactory results.

4. Results and discussion

4.1. Model evaluation method

To evaluate the performance of the AST model in pipeline acoustic event classification tasks, this paper employs four evaluation metrics: *Accuracy*, *Precision, Recall, F*1 score.

Accuracy represents the proportion of correctly classified samples relative to the total number of samples, reflecting the overall classification performance:

$$\text{Accuracy} = \frac{\text{TP} + \text{TN}}{\text{TP} + \text{TN} + \text{FP} + \text{FN}} \tag{14}$$

where, true positive (TP) denotes the number of samples correctly predicted as positive, true negative (TN) denotes the number of samples correctly predicted as negative, false positive (FP) denotes the number of samples incorrectly predicted as positive, false negative (FN) denotes the number of samples incorrectly predicted as negative.

Precision represents the proportion of truly positive samples among those predicted as positive, reflecting the predictive accuracy of the mode:

$$\text{Precision} = \frac{\text{TP}}{\text{TP} + \text{FP}}. \tag{15}$$

Recall represents the proportion of correctly predicted positive samples among all actual positive samples, reflecting the model's capability to capture positive instances:

$$\text{Recall} = \frac{\text{TP}}{\text{TP} + \text{TN}}. \tag{16}$$

The $F1$ *score* is the harmonic mean of *precision* and *recall*, comprehensively reflecting the classification performance:

$$F1 = \frac{2 \times \text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}}. \tag{17}$$

These metrics collectively evaluate the model's performance: accuracy reflects overall correctness, Precision measures leak prediction accuracy, recall ensures comprehensive leak coverage, and $F1$ *score* provides balanced assessment.

4.2. Model training performance

To assess the effectiveness of feature learning, t-SNE dimensionality reduction is applied to visualize the high-dimensional representations learned by the AST model. Figure 6 shows clear separation between normal rolling, collision, and leak events in the 2D feature space, confirming discriminative feature extraction, while consistent clustering patterns across both PA12 and aluminum alloy SDs indicate minimal impact of detector material on classification performance.

The AST model training process using the original unaugmented dataset is illustrated in figure 7. The model exhibits typical overfitting characteristics: in the early training stage (0–5 epochs), both training and validation accuracy increase rapidly; subsequently, training accuracy continues to rise to 99.66%, while validation accuracy plateaus below 90%. The validation loss increases rather than decreases after the 20th epoch, stabilizing at approximately 0.5, indicating that the model overfits on the training data.

The AST model training process using the augmented dataset is illustrated in figure 8. The accuracy gap between training and validation sets remains around 5%, with validation accuracy steadily increasing to 95%, indicating effective overfitting mitigation. The validation loss stabilizes after the 20th epoch at approximately 0.2, with significantly reduced fluctuation amplitude, confirming the positive effect of data augmentation.

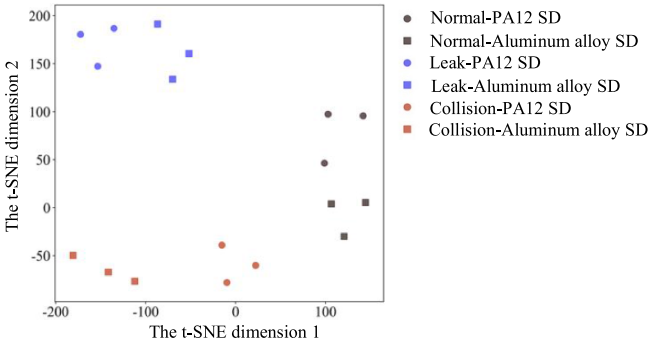


Figure 6. 2D feature distribution via t-SNE.

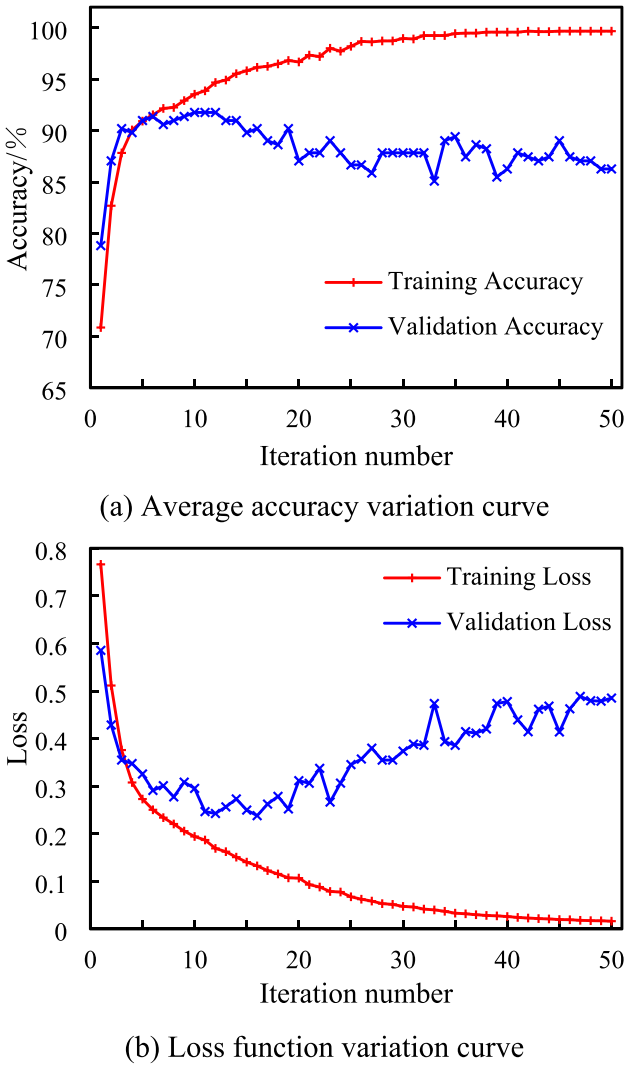


Figure 7. Audio spectrogram transformer model training process without data enhancement.

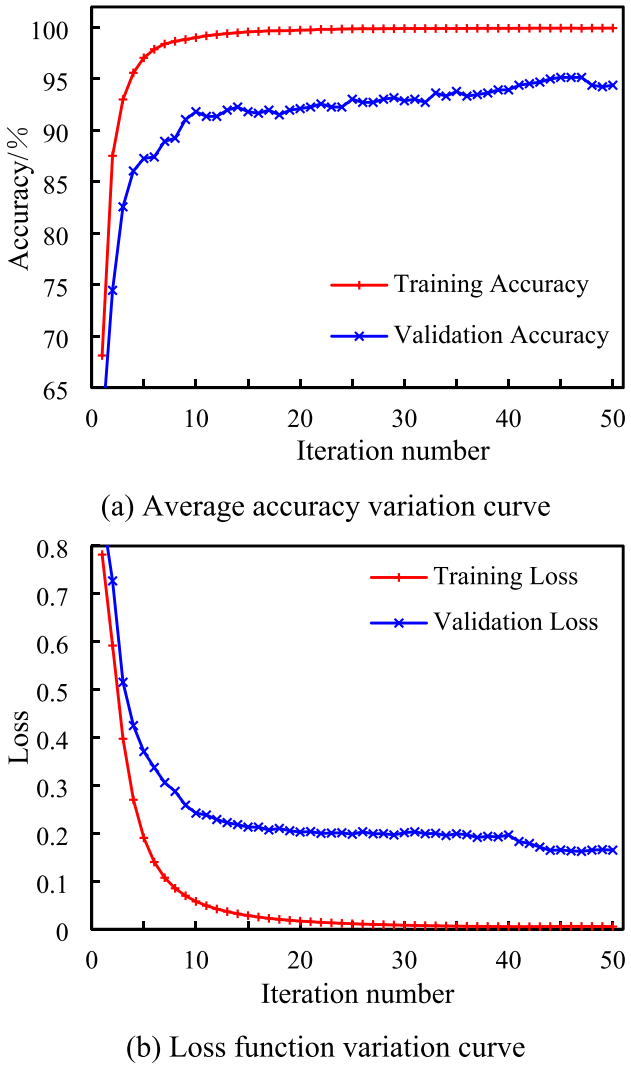


Figure 8. Audio spectrogram transformer model training process after data enhancement.

To further evaluate the classification performance of the leak detection model, the detailed evaluation metrics of the model with data augmentation on the validation set are presented in table 4. As shown in the table, the model performs best in recognizing the normal rolling category, achieving a precision of 96.44% and a recall of 97.60%. It also demonstrates good performance in identifying the leak category, with a precision of 94.24% and a recall reaching 99.09%. The recognition performance for the collision category is relatively lower, but still attains a precision of 95.00% and a recall of 71.25%. From the weighted average metrics, the overall performance of the model shows balanced improvement, with precision, recall, and $F1$ score all exceeding 94%, and the performance disparities between different categories are significantly reduced. This indicates that incorporating data augmentation strategies on the basis of transfer learning not only enhances the overall classification accuracy of the model but also improves its balance across different categories, establishing a solid foundation for the model's deployment in practical engineering applications.

4.3. Leak detection accuracy

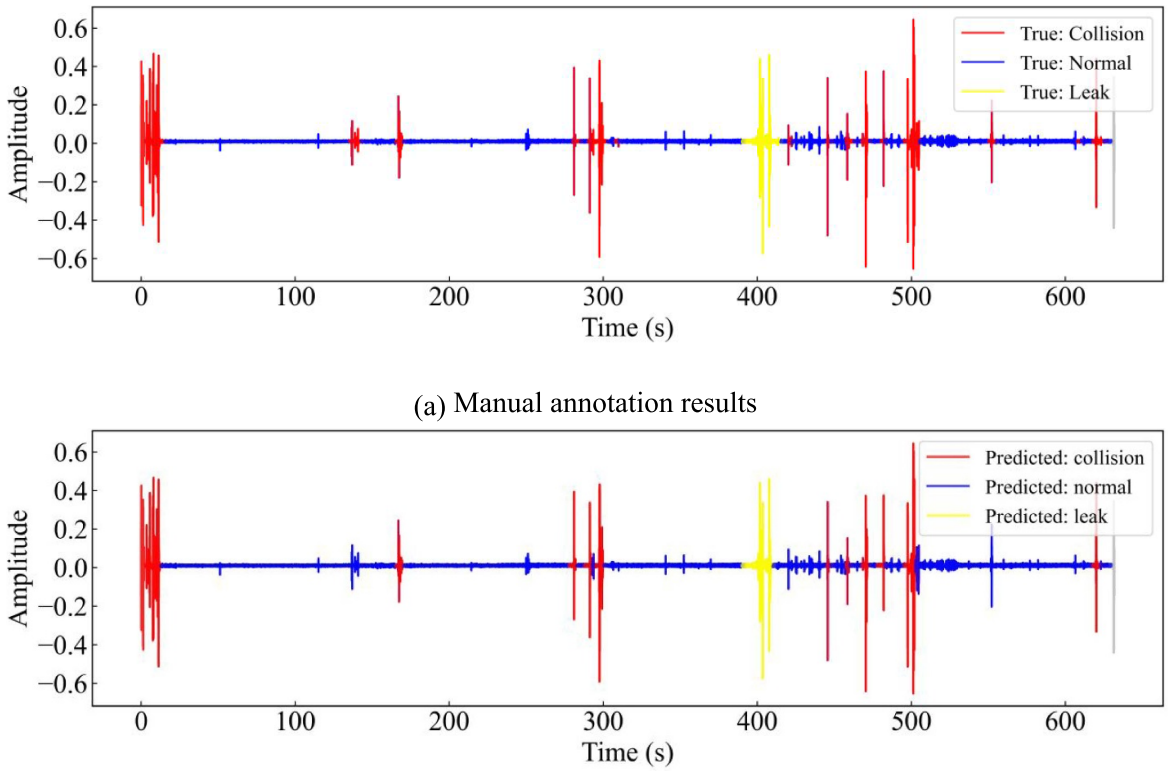
To verify the detection capability of the AST-based leak detection model, an engineering validation experiment independent of the training data was designed. The validation experiment employed newly collected data that had not appeared during the model training process, ensuring the objectivity and reliability of the evaluation results. The experiments were conducted under two pressure conditions: 1 MPa and 0.5 MPa, with flow velocity maintained at 1 m s^{-1}, to comprehensively assess the model's adaptability in different engineering environments.

Under the 0.5 MPa pressure condition, the comparison between model inference results and manual annotations is illustrated in figure 9. As shown in figure 9(a), the manually annotated leak events are primarily distributed within the time period around 400 s, forming a distinct continuous interval. From the time domain perspective, these leak acoustic signals exhibit typical short-duration signal characteristics. As observed in figure 9(b), the model accurately identifies these short-duration leak events, successfully detecting the leak status around 400 s, and the duration of the output leak events highly corresponds with the actual signals. Comparing the manual annotations and model predictions, good consistency is demonstrated in the starting points, ending points, and durations of leak events, indicating that the model possesses strong acoustic event recognition capabilities. This precise recognition capability for short-duration leak signals facilitates successful localization of small leaks in low-pressure pipelines in field conditions.

Under the 1MPa pressure condition, the comparison between model inference results and manual annotations is illustrated in figure 10. As shown in figure 10(a), with increased pressure, the greater pressure differential at the leak point generates stronger acoustic leak signals, enabling the detection sphere to capture characteristic signals from greater distances. This characteristic manifests in the time-domain distribution of acoustic signals: in the leak event around 400s, the signal duration captured by the detection sphere is significantly longer than under the 0.5 MPa condition, indicating increased signal propagation distance. From the model recognition performance shown in figure 10(b), the AST model demonstrates excellent detection capabilities, not only accurately identifying leak events but also precisely determining signal durations, with prediction results highly consistent with actual annotations. Notably, multiple collision events in the 0-100s and 300-500s intervals are also accurately captured, indicating that increased pressure does not affect the model's

Table 4. Model performance metrics on the validation set.

Category	Precision (%)	Recall (%)	$F1$	Sample quantity
Normal rolling	96.44	97.60	0.9702	250
Leak	94.24	99.09	0.9660	330
Collision	95.00	71.25	0.8143	80
Weighted average	95.16	95.15	0.9492	660



(a) Manual annotation results

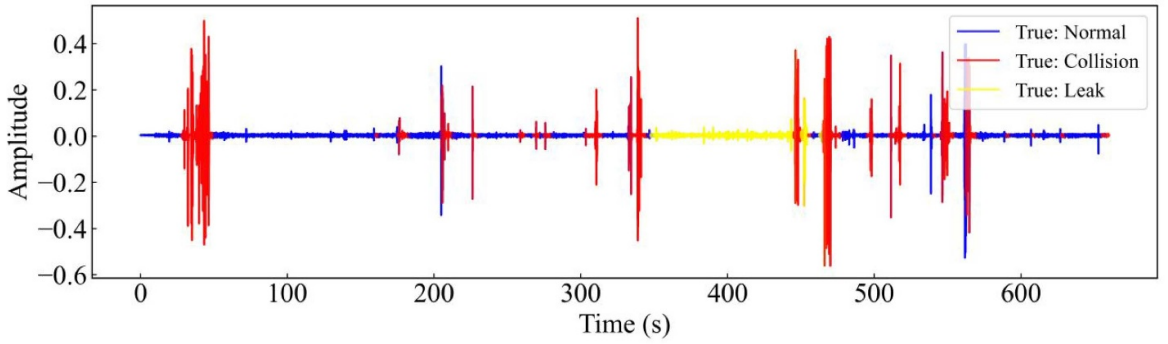
(b) Model inference results

Figure 9. Model inference performance of the SD under 0.5 MPa operating condition.

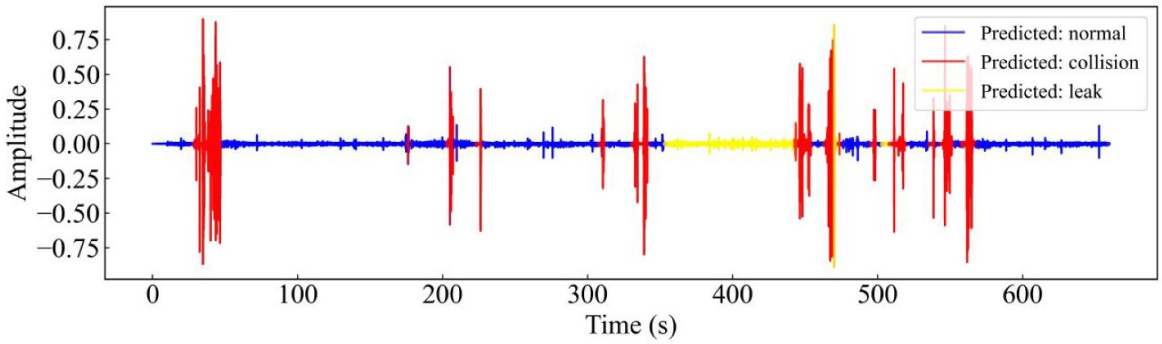
ability to distinguish different types of events, but rather enhances its sensitivity to leak characteristics.

Based on the analysis of both experimental results, the AST model effectively captures the time periods when anomalous events such as leaks and collisions occur. To achieve leak localization, this paper utilizes the built-in acceleration odometer of the SD to convert the time information identified by the model into spatial positions within the pipeline. In principle, for each complete forward axial rotation of the intelligent sphere, the X and Y axis outputs alternate once. By measuring the rotation period and combining it with the geometric dimensions of the sphere, the rolling travel distance of the intelligent sphere within the pipeline can be calculated [40], thereby enabling time-to-space mapping.

After converting the aforementioned time-domain leak detection results to the spatial domain, specific leak location data are obtained. Under the 1MPa and 0.5 MPa conditions, the leak areas identified by the SD are distributed with center points at 406.5 m and 401.6 m, respectively. After correcting the cumulative distance error using pipeline elbow ④ shown in figure 5(d) as a marker point, the final localization results are presented in table 5. The localization errors under both conditions are controlled within 3.1 m, with relative errors less than 0.5%, meeting the requirements for engineering applications. The experimental data indicate that the AST-based leak detection model proposed in this study not only identifies leak events more accurately but also determines leak locations more precisely. The localization precision with relative errors controlled within 0.5% fully satisfies the requirements for actual pipeline inspection and repair, significantly reducing excavation range and lowering maintenance costs.



(a) Manual annotation results



(b) Model inference results

Figure 10. Model inference performance of the SD under 1 MPa operating condition.

Table 5. Leak localization results under various operating conditions.

Operating parameters	Corrected localization result	Corrected localization result	Actual leak position	Localization error	Relative error
0.5 MPa	406.50 m	406.50 m	416.89 m	3.07 m	0.47%
1 MPa	401.60 m	414.43 m	416.89 m	2.46 m	0.37%

5. Conclusion

This study presents an automatic leak sound recognition method based on the AST model for pipeline leak detection using the SD with internal ARAC. The experiments and results analyses yield the following main conclusions:

(1) The study utilizes SD with ARAC to collect pipeline acoustic signals, building a dataset containing 1499 normal rolling, 612 collision, and 248 leak samples (2359 total). This three-class approach reduces interference from non-leak sounds in leak identification.
(2) The approach combining data augmentation and transfer learning successfully addresses class imbalance and small sample limitations. Through signal concatenation, background mixing, and AST model pre-trained on AudioSet, the model achieves weighted average precision of 95.16% and recall of 95.15%, providing a reliable foundation for practical applications.
(3) The AST-based leak detection method achieves accurate identification and localization of 1 mm aperture leaks with localization errors of 2.46 m and 3.07 m under 1 MPa and 0.5 MPa conditions respectively, with relative errors of 0.37% and 0.47%, significantly reducing excavation scope for pipeline maintenance. Compared to traditional methods like CNN and LSTM, the AST approach demonstrates superior global feature capture and processing efficiency, requiring only 2.58 s to process an 11 min audio sample (300 times faster than real-time) to meet offline processing requirements.

Data availability statement

The data cannot be made publicly available upon publication because they contain commercially sensitive information. The data that support the findings of this study are available upon reasonable request from the authors.

Acknowledgments

This work is supported by National Natural Science Foundation of China (No. 62473279), Natural Science Foundation of Tianjin (No. 24JCZDJC01070), and Guangxi Key Laboratory of Automatic Detecting Technology and Instruments (No. YQ24203).

ORCID iDs

Yin Xiaofeng 0009-0006-3116-4640
Ma Jinyu 0000-0003-1509-9421
Huang Xinjing 0000-0002-8964-8502

References

[1] Wang F, Lin W, Liu Z, Wu S and Qiu X 2017 Pipeline leak detection by using time-domain statistical features *IEEE Sens. J.* **17** 6431–42
[2] Li W, Ling W, Liu S, Zhao J, Liu R, Chen Q, Qiang Z and Qui J 2011 Development of systems for detection, early warning, and control of pipeline leakage in drinking water distribution: a case study *J. Environ. Sci.* **23** 1816–22
[3] Xu J, Chai K T, Wu G, Han B, Wai E L, Li W, Yeo J, Nijhof E and Gu Y 2018 Low-cost, tiny-sized MEMS hydrophone sensor for water pipeline leak detection *IEEE Trans. Ind. Electron.* **66** 6374–82
[4] Yuan J, Mao W, Hu C, Zheng J, Zheng D and Yang Y 2023 Leak detection and localization techniques in oil and gas pipeline: a bibliometric and systematic review *Eng. Fail. Anal.* **146** 107060
[5] Zhao L, Cao Z and Deng J 2024 A review of leak detection methods based on pressure waves in gas pipelines *Measurement* **236** 115062
[6] Tran V Q C, Le D V, Yntema D R and Havinga P J M 2022 A review of inspection methods for continuously monitoring PVC drinking water mains *IEEE Internet Things J.* **9** 14336–54
[7] Lalam N, Westbrook P, Naeem K, Lu P, Ohodnicki P, Diemler N, Buric M P and Wright R 2023 Pilot-scale testing of natural gas pipeline monitoring based on phase-OTDR and enhanced scatter optical fiber cable *Sci. Rep.* **13** 14037
[8] Hu Z, Tariq S and Zayed T 2021 A comprehensive review of acoustic based leak localization method in pressurized pipelines *Mech. Syst. Signal Process.* **161** 107994
[9] Yu Y, Cui X, Gao Y, Han X, Song L and Lu F 2025 Acoustic feature processing strategy for leak degree identification in non-metallic pipelines *Appl. Acoust.* **238** 110820
[10] Li Y, Minh H, Cao M, Qian X and Wahab M A 2024 An integrated surrogate model-driven and improved termite life cycle optimizer for damage identification in dams *Mech. Syst. Signal Process.* **208** 110986
[11] YiFei L, MaoSen C, Tran-Ngoc H, Khatir S and Wahab M A 2023 Multi-parameter identification of concrete dam using polynomial chaos expansion and slime mould algorithm *Comput. Struct.* **281** 107018
[12] Ali A, Xinhua W and Razzaq I 2025 Optimizing acoustic signal processing for localization of precise pipeline leakage using acoustic signal decomposition and wavelet analysis *Digit. Signal Process.* **157** 104890
[13] Zhang Z, Huang J, Yu Y, Qin R, Wang J, Zhang S, Su Y, Wen G, Cheng W and Chen X 2025 Microleakage acoustic emission monitoring of pipeline weld cracks under complex noise interference: a feasible framework *J. Sound Vib.* **604** 118980
[14] Huang J, Zhang Z, Qin R, Yu Y, Li Y, Xu Q, Xing J, Wen G, Cheng W and Chen X 2025 Dual channel visible graph convolutional neural network for microleakage monitoring of pipeline weld homalographic cracks *Comput. Ind.* **164** 104193
[15] Wu K 2024 A survey on wireless in-pipe inspection robotics *Int. J. Intell. Robot. Appl.* **8** 648–70
[16] Huang X, Li Z, Li J, Wang X, Feng H, Zhang Y and Rui X 2020 Low-cost, high-sensitivity hydrophone based on resonant air cavity *IEEE Sens. J.* **21** 7348–57
[17] Huang X, Li Z, Li J, Feng H, Zhang Y and Chen S 2021 Acoustic investigation of high-sensitivity spherical leak detector for liquid-filled pipelines *Appl. Acoust.* **174** 107790
[18] Huang X, Ren X, Wang L, Bian X, Li J and Ma J 2024 Design and test of pipeline leak detector with acoustic resonance air cavity *IEEE Sens. J.* **25** 5425–37
[19] Korlapati N V S, Khan F, Noor Q, Mirza S and Vaddiraju S 2022 Review and analysis of pipeline leak detection methods *J. Pipeline Sci. Eng.* **2** 100074
[20] Farah E and Shahrour I 2024 Water leak detection: a comprehensive review of methods, challenges, and future directions *Water* **16** 2975
[21] Yang Y, Peng Z, Zhang W and Meng G 2019 Parameterised time-frequency analysis methods and their engineering applications: a review of recent advances *Mech. Syst. Signal Process.* **119** 182–221
[22] Huang J, Zhang Z, Qin R, Yu Y, Wen G, Cheng W and Chen X 2024 Interpretable real-time monitoring of pipeline weld crack leakage based on wavelet multi-kernel network *J. Manuf. Syst.* **72** 93–103
[23] Satterlee N, Zuo X, Lee C, Park C and Kang J S 2025 Parallel multi-layer sensor fusion for pipe leak detection using multi-sensors and machine learning *Eng. Appl. Artif. Intell.* **153** 110923
[24] YiFei L, Minh H, Khatir S, Sang-To T, Cuong-Le T, MaoSen C and Wahab M A 2023 Structure damage identification in dams using sparse polynomial chaos expansion combined with hybrid K-means clustering optimizer and genetic algorithm *Eng. Struct.* **283** 115891
[25] Siddique M F, Ahmad Z, Ullah N and Kim J 2023 A hybrid deep learning approach: integrating short-time fourier transform and continuous wavelet transform for improved pipeline leak detection *Sensors* **23** 8079
[26] Zhang X, Shi J, Yang M, Huang X, Usmani A S, Chen G, Fu J, Huang J and Li J 2023 Real-time pipeline leak detection and localization using an attention-based LSTM approach *Process Saf. Environ. Prot.* **174** 460–72
[27] Liu R, Zayed T, Xiao R and Hu Q 2024 Time-transformer for acoustic leak detection in water distribution network *J. Civ. Struct. Health Monit.* **15** 1–17
[28] Yang C, Cai B, Wu Q, Wang C, Ge W, Hu Z, Zhu W, Zhang L and Wang L 2023 Digital twin-driven fault diagnosis method for composite faults by combining virtual and real data *J. Ind. Inf. Integr.* **33** 100469
[29] Wang D, Sun Y and Lu J 2025 Pipeline leak detection based on generative adversarial networks under small samples *Flow Meas. Instrum.* **101** 102745
[30] Chen X, Yang R, Xue Y, Huang M, Ferrero R and Wang Z 2023 Deep transfer learning for bearing fault diagnosis: a systematic review since 2016 *IEEE Trans. Instrum. Meas.* **72** 1–21
[31] Weiss K, Khoshgoftaar T M and Wang D 2016 A survey of transfer learning *J. Big Data* **3** 1–40
[32] Wang K, Yang Y and Zhao X 2024 Acoustic signal adversarial augmentation for pressure pipeline leakage detection *Eng. Res. Express* **6** 035538
[33] Glynis K, Kapelan Z, Bakker M and Taormina R 2023 Leveraging transfer learning in LSTM neural networks for

data-efficient burst detection in water distribution systems *Water Resour. Manage.* **37** 5953–72
[34] Wu L, Liang W and Sha D 2023 Cross-domain feature selection and diagnosis of oil and gas pipeline defects based on transfer learning *Eng. Fail. Anal.* **143** 106876
[35] Zhang R, Tao H, Wu L and Guan Y 2017 Transfer learning with neural networks for bearing fault diagnosis in changing working conditions *IEEE Access* **5** 14347–57
[36] Lin G, Zhoumo Z, Xinjing H, Mingze L, Hao F, Jian L and Xiaobo R 2020 Performance enhancements of the spherical detector for pipeline spanning inspection through posture stabilization *Measurement* **165** 108095
[37] Guo S X, Chen S L, Huang X J, Xu T S and Jin S J 2015 Design and application of a leak detector for submarine oil pipelines *Mod. Chem. Ind.* **35** 182–6
[38] Xu T, Chen S, Guo S, Huang X, Li J and Zeng Z 2019 A small leakage detection approach for oil pipeline using an inner spherical ball *Process Saf. Environ. Prot.* **124** 279–89
[39] Gemmeke J F, Ellis D P, Freedman D, Jansen A, Lawrence W, Moore R C, Plakal M and Ritter M 2017 Audio set: an ontology and human-labeled dataset for audio events *2017 IEEE Int. Conf. on Acoustics, Speech and Signal Processing (ICASSP)* (IEEE) pp 776–80
[40] Huang X, Chen S, Guo S, Xu T, Ma Q, Jin S and Chirikjian G S 2016 A 3D localization approach for subsea pipelines using a spherical detector *IEEE Sens. J.* **17** 1828–36