Sound & Vibration 2025, 59(4), 3645.
doi: https://doi.org/10.59400/sv3645

Article

AS-FCRNet: a lightweight multi-frame acoustic–seismic fusion network for high-precision ground moving target recognition on UGS

Zheyu Liu, Kunsheng Xing, Wei Wang*, Nan Wang

College of Intelligence Science and Technology, National University of Defense Technology, Changsha 410073, China

*** Corresponding author:** Wei Wang, wangwei.4526@163.com

CITATION

Liu Z, Xing K, Wang W, et al. AS-FCRNet: a lightweight multi-frame acoustic–seismic fusion network for high-precision ground moving target recognition on UGS. Sound & Vibration. 2025; 59(4): 3645. https://doi.org/10.59400/sv3645

ARTICLE INFO

Received: 18 June 2025
Revised: 2 August 2025
Accepted: 20 August 2025
Available online: 29 August 2025

Abstract: To address the challenges of high computational complexity and temporal modeling difficulties caused by high-dimensional data in acoustic and seismic signal classification, this paper proposes a multi-stage dimensionality reduction and classification framework based on the integration of Mel-frequency spectrum feature extraction, Convolutional Neural Networks (CNNs), and Long Short-Term Memory (LSTM) networks. The method significantly reduces computational complexity while maintaining competitive classification accuracy through progressive feature compression and acoustic-seismic feature fusion. Specifically, Mel-frequency spectrum feature extraction is first performed on dual-channel input signals (acoustic and seismic) to extract perceptually relevant physical features aligned with human auditory characteristics. Then, a lightweight CNN is designed to perform further feature extraction on log-Mel energy representations; in the fusion stage, we investigate three fusion strategies (information-level, feature-level, and decision-level fusion) for acoustic and seismic signals to identify the optimal approach, before fusing the information and compressing the fused features into a short vector for subsequent temporal modeling. A sequence of compact feature vectors extracted from consecutive frames (e.g., four-frame segments) is fed into an LSTM network to capture temporal dependencies, and the final classification is performed based on the output of the last time step. Experimental results demonstrate that the proposed approach effectively balances inference efficiency and model performance, achieving accurate and reliable classification results with low computational complexity.

Keywords: fusion of acoustic and seismic features; multi-frame temporal modeling; lightweight neural network; deep learning; long short-term memory (LSTM)

1. Introduction

The classification of moving ground targets is a critical task in many military and security missions. Unattended Ground Sensor (UGS) systems are widely used to monitor the activity of ground-moving targets and collect relevant information [1]. These systems can be equipped with various types of sensors, such as seismic, acoustic, and optical sensors [2]. Among these, seismic and acoustic sensors are particularly favored due to their low cost an d high concealability. Both seismic and acoustic signal classification have their respective advantages and limitations. Seismic signals can propagate over long distances and are less affected by environmental noise, such as wind, but their characteristics can vary significantly across different geological conditions [3]. This variability poses a significant challenge for identifying moving targets in diverse ground geological conditions [4]. In contrast, acoustic signals propagate through air and

are unaffected by changes in the propagation medium, but they suffer from a shorter transmission range and are inevitably influenced by ambient noise [5]. To address these limitations, this paper proposes the joint use of seismic and acoustic sensors to capture the signals generated by ground moving targets and to fuse both acoustic and seismic information for improved classification performance.

Classification methods for moving targets using acoustic and seismic signals include threshold-based methods and machine learning approache. Threshold-based methods rely on empirically set fixed or adaptive feature thresholds. When a feature value exceeds the predefined threshold, a target intrusion is considered detected. These methods offer high computational efficiency, but their classification accuracy is susceptible to environmental noise.

Machine learning is currently regarded as the best solution in ground target recognition techniques based on seismic or acoustic signals [6]. According to the developmental trajectory of machine learning, it can be categorized into traditional machine learning (TML) and deep learning (DL). TML are widely used in object detection, dynamic system optimization [7], power load forecasting [8] and related fields.TML methods applied to moving target recognition using ground seismic signals include, but are not limited to, support vector machines (SVM) [9–11], k-nearest neighbors (k-NN) [12–14], Decision tree(DT) [15, 16], ANN [17], naive Bayesian algorithm (NB) [18], boost [19], and Gaussian mixture models (GMM) [20, 21]. Traditional machine learning methods can achieve high computational efficiency and are easily deployable on embedded devices. However, due to the difficulty in leveraging handcrafted features to extract nonlinear relationships, the classification accuracy and generalization capability of traditional machine learning methods are often inadequate [22].

The advancement of deep learning has significantly accelerated progress in the field of computer vision [23–25]. Furthermore, deep learning methods have been extensively applied to other critical domains, including Remaining Useful Life (RUL) prediction [26], wind power forecasting [27], and mechanical equipment lifetime prediction [28]. These advanced techniques have also been increasingly adopted in the area of acoustic and seismic signal recognition.Deep neural networks (DNNs) can extract underlying patterns from large amounts of training data, and many researchers have proposed various deep learning methods that demonstrate strong performance. Existing deep learning-based methods for ground moving target recognition using sound and seismic signals can be categorized into four types.The first directly feeds raw audio signals into neural networks. For instance, Wang et al. (2019) proposed VibCNN, a deep learning architecture for one-dimensional seismic signal classification. The network incorporates a custom 1D SE-Inception module that integrates both the multi-branch design of GoogLeNet and the channel-wise attention mechanism from Squeeze-and-Excitation Networks (SENet). Evaluated on a dedicated seismic dataset containing three types of moving ground targets, the model achieved a classification accuracy of 93.44%. The second type involves preprocessing the raw signals into two-dimensional representations suitable for CNN input, while reducing data dimensionality and computational load. For example, Xing et al. proposed the MFC-TinyNet method, which combines Mel-frequency spectrum

feature extraction and CNN to recognize moving targets across multiple terrains [29]. Bin et al. introduced a compressed-aware edge convolutional neural network (CS-ECNN), an edge-oriented intelligence approach designed for efficient target recognition [30]. Jin et al. employed a convolutional neural network (CNN) for vehicle classification based on seismic signals. They utilized particle swarm optimization (PSO) to tune parameters of the feature extraction method and introduced a log-scaled frequency cepstral coefficient (LFCC) matrix as input to the network. Their approach achieved a classification accuracy of 91.93% on the DARPA SensIT dataset.

The third type combines raw waveform input with time-frequency representations in a multimodal framework. Tran and Tsai [21] introduced a CNN-based ensemble model named SirenNet for emergency vehicle detection via siren sounds. The model employs a dual-path architecture: one branch (WaveNet) processes raw audio waveforms using a 1D-CNN, while the other (MLNet) uses combined MFCC and log-mel spectrogram features with a 2D-CNN. Evaluated on a diverse dataset integrating real-world recordings and public datasets (UrbanSound8K and ESC-50), their method achieved an accuracy of 98.24%. The fourth type integrates CNNs and LSTM networks [31], For example, Mohine et al. (2022) developed a hybrid deep learning model integrating a 1D convolutional neural network with a bidirectional long short-term memory network (1D CNN-BiLSTM) for classifying moving vehicles using acoustic signals. The model was evaluated on an experimentally generated dataset comprising multiple vehicle types, achieving a classification accuracy of 92%, and reached 96% accuracy on the military SITEX02 dataset [32]. Nie et al. [33] proposed a CNN-DBiLSTM model for recognizing moving targets using seismic signals, achieving an accuracy of 97.23% on the JL dataset. Sun et al. [34] introduced a long-term correlation feature network (LTCFN) for classifying vehicles using synchronized acoustic and seismic signals. The model employs an AlexNet-based feature extractor to capture frame-level features, which are then fused and processed by a long short-term memory (LSTM) network to exploit inter-frame temporal correlations. Evaluated on a custom-collected dataset comprising three vehicle types, the method achieved a classification accuracy of 96%.

Current methods for moving target recognition based on acoustic and seismic signals are mostly built on single-frame features, which are unable to capture the temporal dependencies across frames generated during target movement. Sun et al. [34] proposed a multi-frame acoustic–seismic recognition approach; however, their method uses raw waveform data, which typically has high dimensionality, and employs a relatively complex single-frame feature extraction network, making the overall system computationally expensive.Furthermore, current methods for moving target identification using acoustic and seismic signals have yet to explore the optimal fusion strategy for integrating these two modalities.

In view of the advantages and limitations of existing methods, this paper proposes a moving target recognition method called Acoustic–Seismic Frame-based Convolutional Recurrent Network (AS-FCRNet), which balances recognition accuracy and low computational complexity. The method consists of three components: Mel-frequency spectrum feature extraction, a lightweight CNN for feature fusion and compression, and an LSTM for temporal modeling. For each frame, Mel-frequency spectrum features

are extracted separately from the acoustic and seismic signals, rather than using the raw waveforms directly as done in Sun et al.'s method [34]. This preprocessing step reduces the dimensionality of the original signals through a form of down-sampling, while preserving critical frequency and temporal information. Mel-spectrograms are particularly advantageous over raw signals because they transform the data into a time-frequency domain, which is better suited for detecting patterns related to moving targets. The Mel-frequency scale mimics the way humans perceive sound, focusing on perceptually important frequencies that are crucial for distinguishing between different types of movement and environmental noise. Additionally, using Mel-spectrograms helps to reduce high-frequency noise and irrelevant details commonly present in raw signals. This transformation not only makes the input data more compact and informative but also reduces the computational complexity of the model, thereby improving its efficiency. By converting the signals into a 2D representation, the model is able to better capture both frequency and time-domain patterns, which are essential for recognizing the dynamic, time-varying characteristics of moving targets.

To explore the combination of the two modalities, three fusion strategies are considered: data-level fusion, which directly merges the two inputs for CNN-based feature extraction; feature-level fusion, where each modality is first processed by a CNN and their extracted features are concatenated for classification; and decision-level fusion, where independent classification results are integrated by an expert network. A sequence of compact feature vectors from consecutive frames is then stacked in temporal order and fed into an LSTM network for temporal modeling, followed by fully connected layers for classification.

2. The AS-FCRNet architecture

The proposed classification method consists of two main steps: 1) Single-frame feature extraction: For each frame, Mel-frequency spectrum features are extracted separately from the acoustic and seismic signals. These features are combined into a two-channel representation and fed into a lightweight CNN, which performs further feature extraction and acoustic–seismic fusion, producing a compact feature vector for each frame; 2) Temporal sequence modeling and classification: A sequence of frame-level feature vectors, arranged in temporal order, is fed into an LSTM for temporal modeling, followed by fully connected layers for classification.

2.1. Single-frame feature extraction

2.1.1. Mel-frequency spectrum feature extraction

In the proposed method, the raw acoustic and seismic signals are first segmented into multiple 1-second frames. Each frame is first processed using Mel-frequency spectrum feature extraction to obtain two-dimensional spectrograms of the acoustic and seismic signals. This process includes the following steps: pre-emphasis, framing, windowing, short-time Fourier transform (STFT), Mel filter bank processing, and logarithm.The overall procedure is illustrated in **Figure 1** [35].

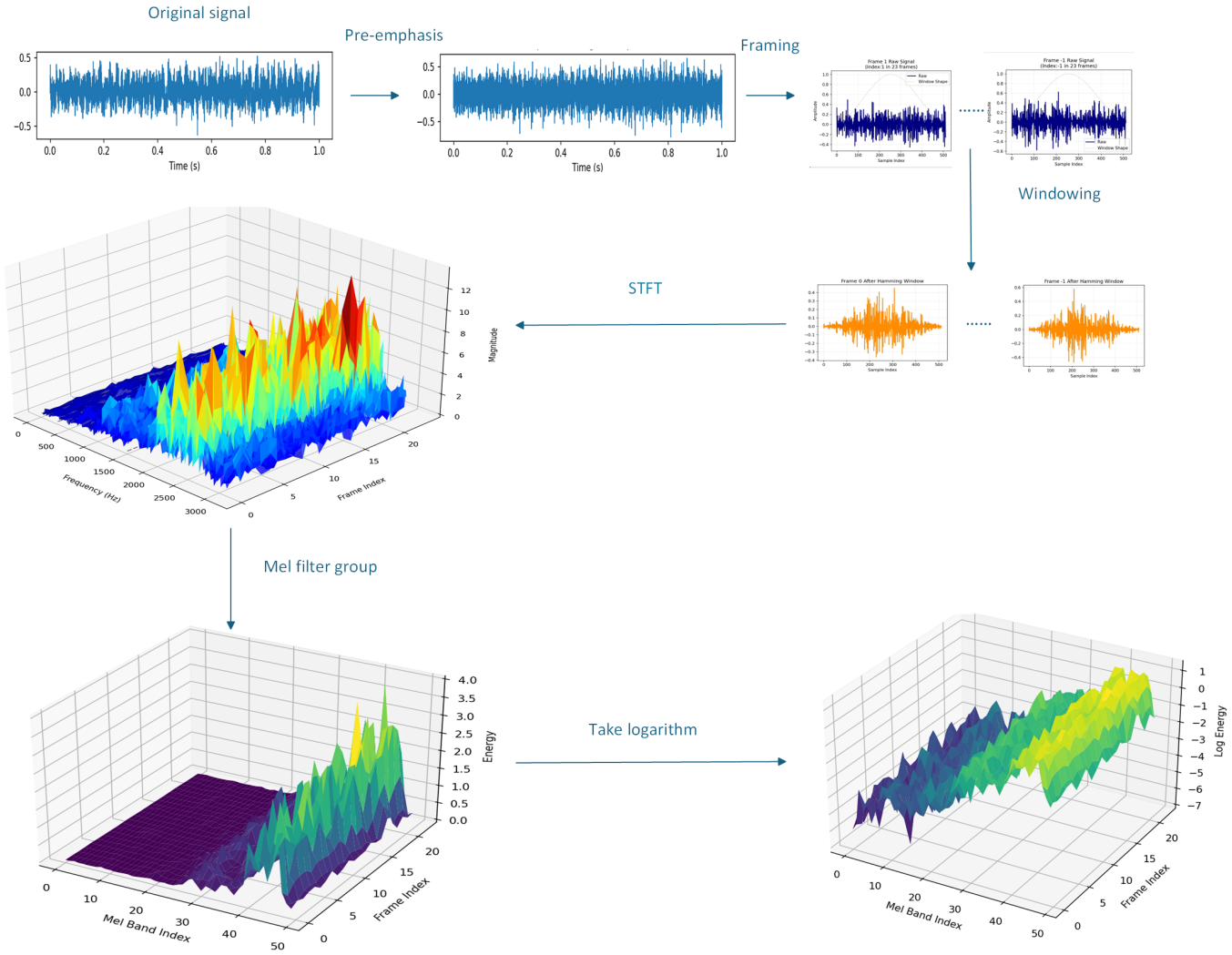


Figure 1. Mel-frequency spectrum extraction process.

(1) Pre-emphasis

Because high-frequency acoustic and seismic components attenuate rapidly during propagation, a pre-emphasis filter is applied to compensate for this loss before spectral analysis. In practice, the signal is passed through a first-order high-pass filter with transfer function shown in (1). This step improves the signal-to-noise ratio of higher frequencies and flattens the spectrum, which benefits subsequent short-time analysis.

$$H(z) = 1 - az^{-1} \tag{1}$$

(2) Framing

Given the non-stationary but short-time stationary nature of the signals, each waveform is divided into partially overlapping frames. To avoid large variations between adjacent frames and to ensure smooth transitions, an overlap is introduced. In this paper, the frame length is set to 82 ms, and the frame shift to 41 ms.

(3) Windowing

To reduce spectral leakage caused by frame-edge truncation (Gibbs phenomenon) and to emphasize the central portion of each frame while suppressing edge noise, every frame is multiplied by a window function. We use the Hamming window in

all experiments.

(4) Short-Time Fourier Transform (STFT)

Since it is often difficult to observe the characteristics of a signal directly in the time domain, the signal is typically transformed into its frequency-domain energy distribution for analysis. Different energy distributions correspond to different speech characteristics. Therefore, after applying the Hamming window, each frame is further processed using the Fast Fourier Transform (FFT) to obtain its energy distribution in the frequency spectrum.

(5) Mel Filter Bank

The human ear is more sensitive to slight variations at low frequencies (e.g., around 500 Hz) but has reduced resolution at higher frequencies (e.g., around 5 kHz). The Mel filter bank is designed to be dense in the low-frequency range and sparse in the high-frequency range, consistent with the critical bandwidth characteristics of human hearing. On the Mel scale, frequencies below 1000 Hz are approximately linear, while those above 1000 Hz grow logarithmically, matching the perceptual property that humans are more sensitive to low frequencies and less discriminative at high frequencies. Therefore, the linear frequency (Hz) obtained from the short-time Fourier transform is converted into the Mel scale according to the transformation formula (2), simulating the human auditory perception of frequency.

$$m = 2595 \times \ln(1 + \frac{f}{700}) \tag{2}$$

2.1.2. CNN architecture for frame-wise modality fusion

Following Mel-frequency spectrum extraction, the resulting acoustic and seismic spectrograms are stacked along the channel dimension to form a two-channel input, which is then fed into a custom-designed convolutional neural network (CNN) for acoustic–seismic feature fusion and further feature extraction. To accommodate the limited computational resources of unattended ground sensors, this paper introduces a Lightweight Hierarchical Depthwise–Pointwise Convolutional Network (HDP-ConvNet) optimized for low computational complexity. HDP-ConvNet consists of three convolutional blocks, each comprising a convolutional layer, a ReLU activation, and a pooling layer. The design choice of using three convolutional blocks is driven by the need for progressively refining the features from low-level to high-level representations. This relatively small number of blocks ensures both sufficient feature extraction and computational efficiency. To support progressive feature extraction, the number of output feature maps is doubled after each block. Doubling the output feature maps after each block increases the model's ability to learn complex features while maintaining a manageable computational cost. The final convolutional block utilizes depthwise separable convolutions to further reduce computational cost and improve generalization. The combination of only three convolutional blocks and the use of depthwise separable convolutions makes HDP-ConvNet a lightweight model, striking a balance between feature extraction capacity and low computational demand. After the three convolutional blocks, the output is flattened and passed through fully connected layers for classification. The overall structure of HDP-ConvNet is shown in **Figure 2**.

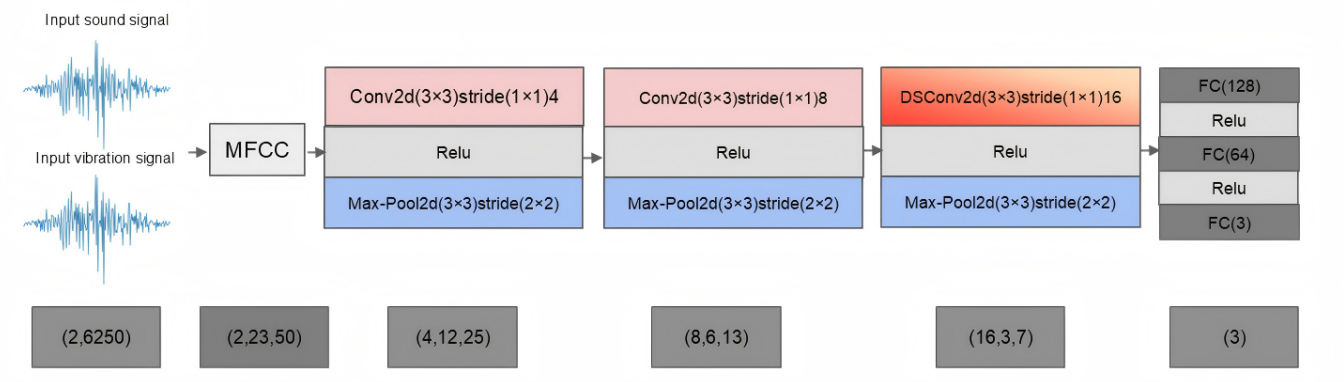


Figure 2. Network structure.

The depthwise separable convolution layer consists of a depthwise convolution followed by a pointwise convolution. The depthwise convolution applies spatial filtering independently to each input channel, while the pointwise convolution uses 1×1 kernels to fuse information across channels. This structure significantly reduces parameters and computation compared to standard convolution. The structure is illustrated in **Figure 3** [36].

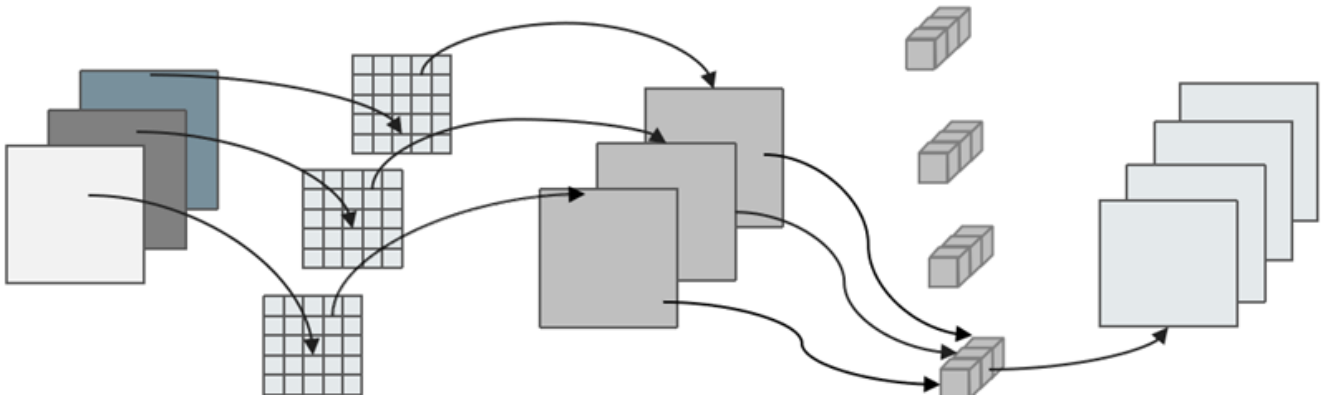


Figure 3. Depthwise separable convolutional layer.

In addition to this baseline data-level fusion with HDP-ConvNet, we further investigate two complementary fusion strategies that share a unified CNN feature-extraction backbone. Specifically, both the acoustic and seismic modalities are processed by the same lightweight three-block CNN extractor—two blocks of 3 ×3 convolution + ReLU + 3×2 max pooling (with dropout), followed by a depthwise–pointwise separable convolution block—to produce compact modality-specific feature maps; the two branches adopt an identical architecture (parameters are not shared). The feature-extraction module is illustrated in **Figure 4a**.

(i) Feature-level fusion [37]. The per-modality feature maps are flattened and concatenated; a small expert module (a shallow MLP with sigmoid gating) produces element-wise weights to adaptively reweight the concatenated feature vector, which is then passed to a linear classifier. The post-extraction processing for feature-level fusion is summarized in **Figure 4b**.

(ii) Decision-level fusion [38]. Each modality's feature map is first fed into a lightweight classifier to produce modality-specific logits; the two logits are concatenated and input to a decision expert (a shallow MLP with softmax weighting) that computes

data-dependent fusion weights, followed by a final linear layer to yield the fused prediction. The post-extraction processing for decision-level fusion is shown in **Figure 4c**.

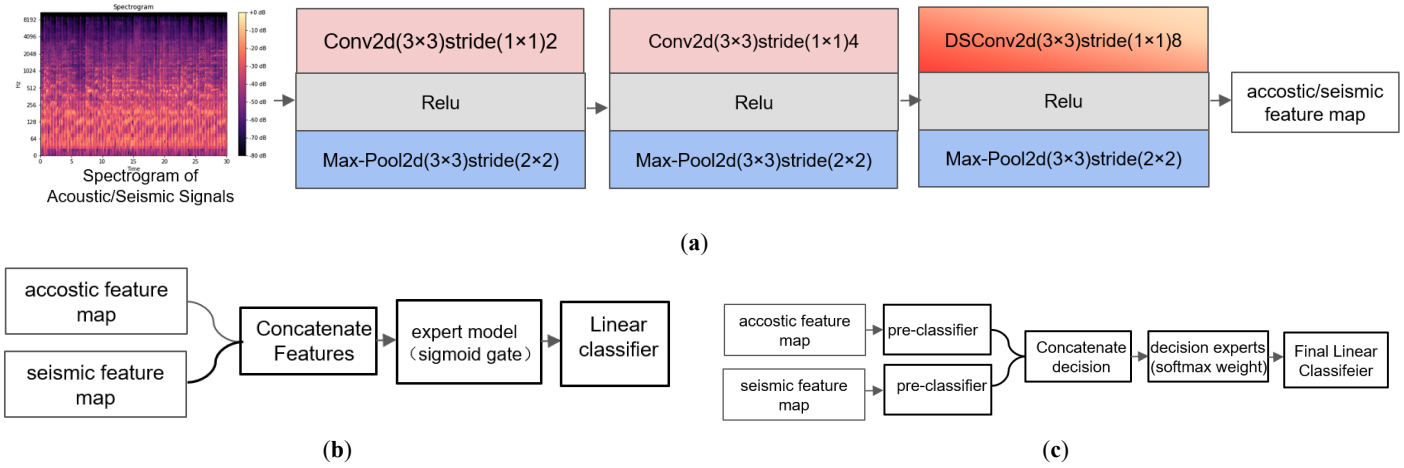


Figure 4. **(a)** Feature extractor for acoustic and seismic spectrograms; **(b)** Feature-level fusion pipeline; **(c)** Decision-level fusion pipeline.

For a fair comparison of fusion strategies, the experimental design ensured that the feature-level and decision-level methods shared an identical CNN feature extraction backbone, All models adhered to the same lightweight design principles to maintain practical feasibility on resource-constrained sensors. Under this unified setting, we systematically investigated the three fusion paradigms to highlight their respective advantages and limitations, and ultimately identified the most effective strategy to serve as the feature extraction backbone for subsequent experiments.

2.2. Temporal sequence modeling and classification

Each frame of the fused acoustic and seismic signals is represented by a compact one-dimensional feature vector. These vectors are then stacked in chronological order to form the input for further processing. For instance, the feature vector of the n-th frame is denoted as F_n, and that of the (n+1)-th frame as F_{n+1}. For a four-frame input, the resulting sequence would be $[F_n, F_{n+1}, F_{n+2}, F_{n+3}]$.

The acoustic and seismic signals, being typical time-series data, exhibit strong temporal dependencies within each frame's feature vector. Therefore, after stacking the feature vectors in chronological order, the next step is to input the multi-frame features into the LSTM layer, which captures deeper-level features by leveraging LSTM's ability to process time-series information.

LSTM (Long Short-Term Memory) is a type of recurrent neural network (RNN) designed to capture long-range dependencies in time-series data. Its core features include gating mechanisms (forget gate, input gate, output gate) and a cell state, which serves as the network's 'memory line,' passing information across time steps. The gates regulate updates to the cell state: the forget gate discards old information, the input gate stores new information, and the output gate generates the current output. The structure of LSTM is shown in **Figure 5**, with computation formulas for the gates and updates provided below [39].

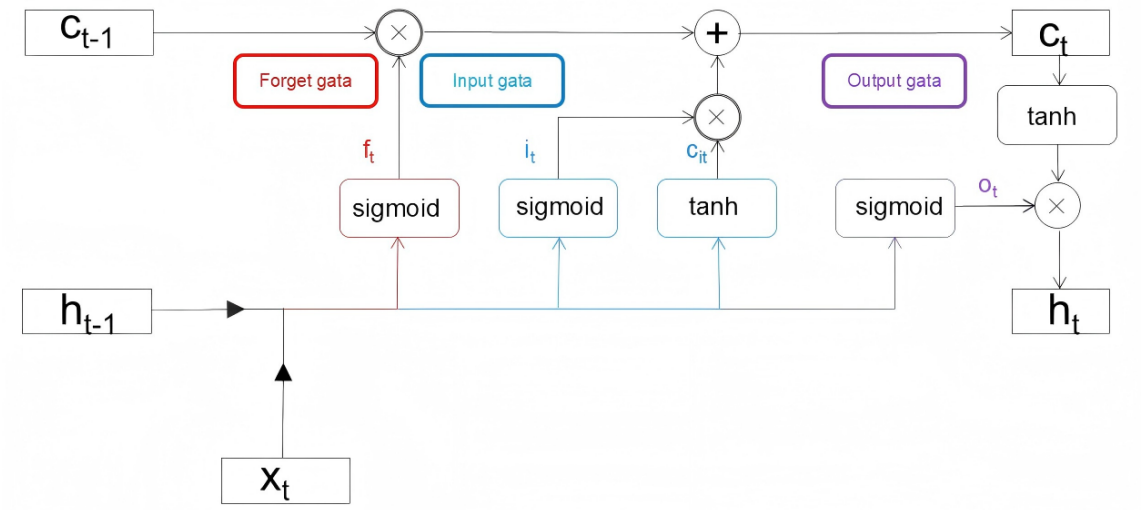


Figure 5. Schematic of the LSTM module.

$$\text{Forget gate:} \quad f_t = \text{sigmoid}\left(w_f\left[h_{t-1}, x_t\right] + b_f\right) \tag{3}$$

$$\text{Input gate:} \quad i_t = \text{sigmoid}\left(w_i\left[h_{t-1}, x_t\right] + b_i\right) \tag{4}$$

$$f_t = \text{sigmoid}(w_f[h_{t-1}, x_t] + b_f) \tag{5}$$

$$\text{Output gate:} \quad o_t = \text{sigmoid}\left(w_o\left[h_{t-1}, x_t\right] + b_o\right) \tag{6}$$

$$c_t = f_i c_{t-1} + i_t c_{ii} \tag{7}$$

In the formulas, x represents the input vector, h is the hidden state, c is the cell state, and w and b are the weight parameters that need to be learned.

The LSTM layer produces multiple outputs. In this method, we use the output of the last time step as the feature vector that integrates information from multiple frames, which is then fed into two fully connected layers for classification.

3. Experiment

3.1. Data acquisition

The dataset used in this study was collected by our team, with data collection sites in Changsha City, Hunan Province, and Zhangjiakou City, Hebei Province. The background noise during the collection of signals from ground-moving targets was approximately 35–60 dB. The sampling frequency was 6250 Hz.

The acoustic and seismic dataset was collected across various geological terrains, including concrete roads, asphalt roads, dirt roads, and grasslands. The dataset is divided into three types: vehicle data, human activity data, and environmental data. Environmental data includes noise when no target is present, and consists of various sounds from the natural environment, such as rain, bird calls, the rustling of leaves caused by wind, and subtle vibrations from wind blowing against the sensor housing. Human activity signals include sounds from human movements, such as speech, breathing, coughing, walking noises, and seismic signals. Vehicle signals consist of sounds generated by wheeled vehicles, including engine noise, horn sounds, and sounds produced during movement. To ensure the purity of collected human, vehicle, and

environmental sound/seismic data, we adopted a two-step process: First, we controlled the collection environment initially—collecting environmental data in quiet areas, human activity data in quiet locations with people but no vehicles, and vehicle data near lanes. Since unavoidable impure data still existed, We then conducted manual inspection after initial collection: for segments containing impure data, we removed and truncated the impure portions, resulting in multiple segments of pure data.Some examples of the data are shown in **Figure 6**. The sensors were installed in two different ways: for soft ground, the probe was directly inserted into the soil, while for hard surfaces, the sensor was placed directly on the surface, as shown in **Figure 7**.

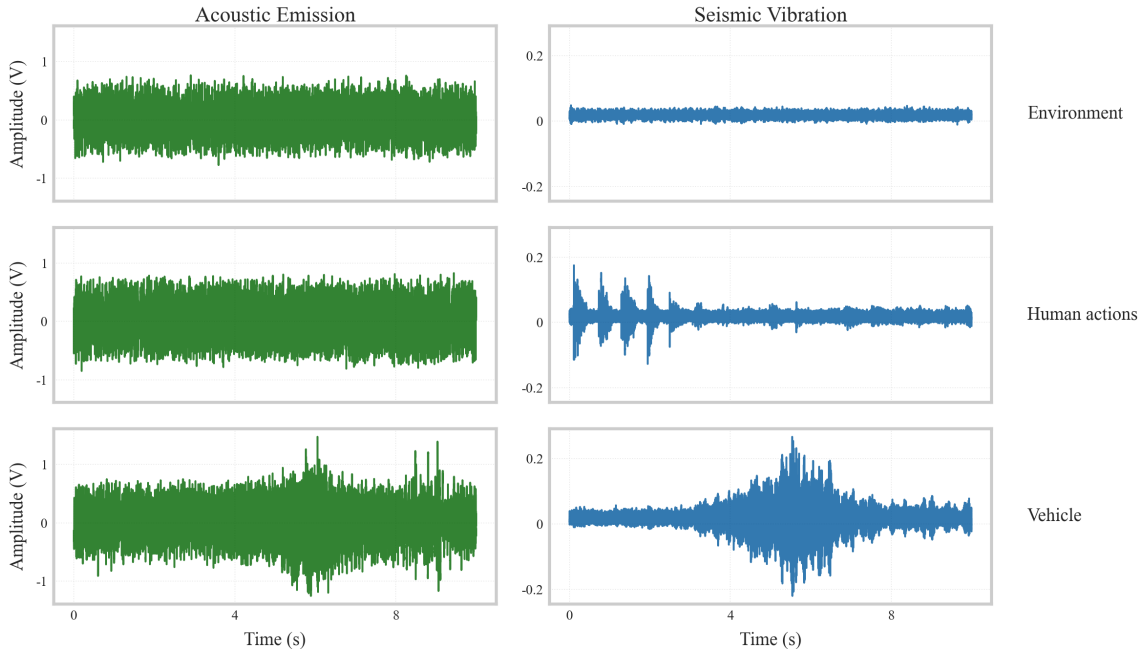


Figure 6. The three types of target sounds and seismic signals.

(a) **(b)**

Figure 7. Installation of the acoustic and seismic sensor for data collection. **(a)** installation in soft ground with the sensor's sharp end directly inserted into the soil; **(b)** Installation on hard ground with a base placed under the sensor.

During dataset collection, the unattended ground sensor is placed and connected to a computer via USB, as shown in **Figure 8**. Acoustic and seismic signals are recorded when a target passes through the alarm zone of the sensor. The alarm zone is a circular area centered on the unattended ground sensor. The diameter of the human alarm zone is 15 meters, while the diameter of the vehicle alarm zone is 25 meters. Humans sometimes pass through the alarm zone while talking, and vehicles sometimes pass through the alarm zone while honking their horns.

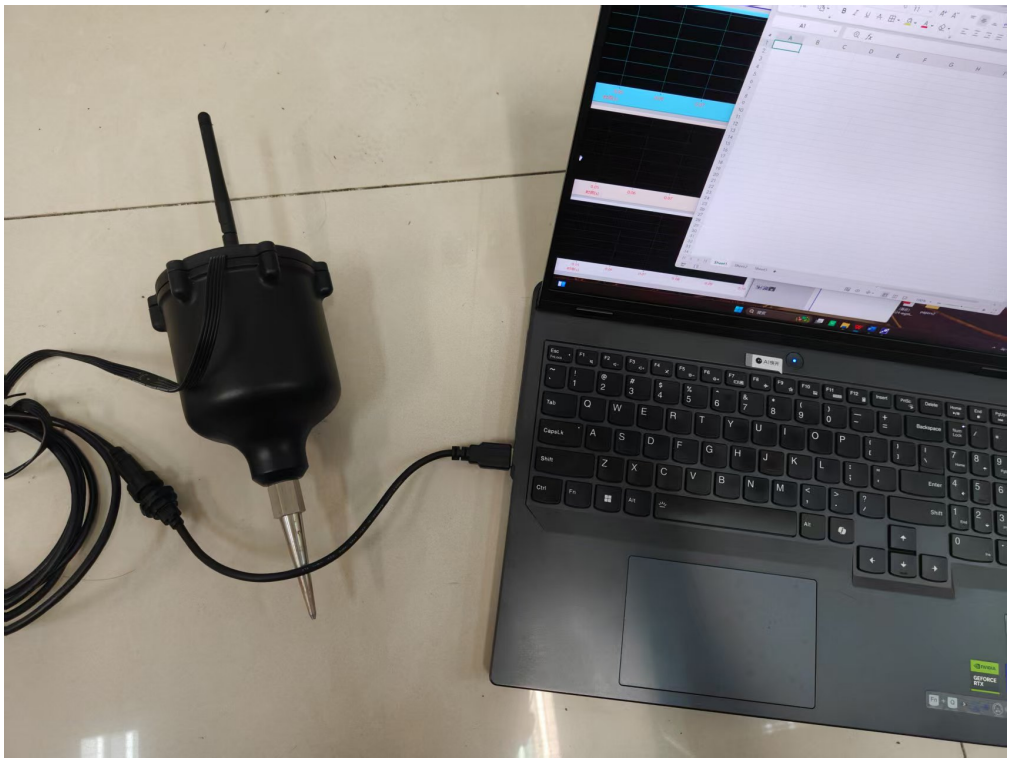


Figure 8. Schematic of the data collection device installation.

The dataset is divided into a training set and a test set for model training and evaluation. The training set is used for parameter estimation, while the test set evaluates generalization performance. To ensure evaluation reliability, most test signals were collected at different times from those in the training set.

The sound and seismic data were all segmented with a 1s sliding step, generating two types of samples for recognition tasks—with the 1s-time-window sample as the minimum data unit. For 1s samples, the window size was set to 1s (matching the 1s sliding step), so each 1s segment is independent and the total count is directly calculated from the cumulative valid data duration; for 4s samples, the window size was expanded to 4s (still using the 1s sliding step), leading to 3s overlap between adjacent 4s samples (only 1s of new data differs). Since the raw data is composed of discrete segments (not a single continuous sequence), each segment yields far fewer non-truncated 4s samples than 1s samples (e.g., a 5s segment produces 5 1s samples but only 2 4s samples). Specific sample sizes are shown in **Table 1**.

Table 1. Sample Size Statistics of Sound and Seismic Data for Recognition Tasks with Different Time Lengths.

Sample size	Category	Environmentent	Human actions	Vehicle
4s	Train set	4905	5840	3544
	Test set	710	657	601
1s	Train set	5241	6220	4437
	Test set	755	708	694

3.2. Experiment design

3.2.1. Fusion-strategy study and CNN backbone selection

We first investigate three acoustic–seismic fusion strategies to select the CNN feature-extraction backbone for subsequent experiments: (i) data-level fusion, which stacks acoustic and seismic spectrograms as a two-channel input and is implemented by the lightweight HDP-ConvNet; (ii) feature-level fusion, where acoustic and seismic spectrograms are processed by two identical lightweight CNN extractors (non-shared weights; extractor structure shown in **Figure 4a**, and the flattened features are concatenated and reweighted by a small expert module before classification (post-extraction pipeline in **Figure 4b**; and (iii) decision-level fusion, where each modality is independently classified by a lightweight CNN-based head and the modality-specific logits are fused by a softmax-weighted expert followed by a final linear layer (pipeline in **Figure 4c**).

After determining the fusion method, we further explored the convolutional network structure. To design an optimal and compact CNN feature extractor for subsequent multi-second (e.g., 4-second) sound-seismic signal recognition tasks, a series of ablation experiments are conducted to explore the impacts of four key factors on the CNN's feature extraction performance and parameter scale. First, different Mel-band numbers (25, 50, 75) are tested during the Mel-spectrogram extraction of sound and seismic signals, to analyze how spectral resolution affects the CNN's ability to capture discriminative features. Second, variations in CNN width are evaluated—with "half-width" (channel count of each layer halved) and "double-width" (channel count doubled) variants constructed based on the base CNN configuration—to balance the network's feature expression capability and model size. Third, the effect of dropping the sound modality is verified by comparing the CNN's performance under dual-modality input (sound + seismic) and single-modality input (seismic only), to assess the contribution of sound signals to feature extraction. Fourth, the role of depthwise separable convolution is examined: the base CNN (with standard convolutions) is compared against a modified CNN adopting depthwise separable convolutions, to verify whether the latter can reduce the network's parameter count while maintaining the CNN's feature extraction accuracy (evaluated via the classification performance of features output by the CNN).

To ensure a fair comparison, The strategy with the best accuracy–efficiency trade-off is selected as the final CNN backbone for the subsequent AS-FCRNet experiments.

During the training process, all experiments use the cross-entropy loss function,

whose formula is shown in (8).

$$L = -\frac{1}{N}\sum_{1}^{N}\sum_{1}^{C} y_{i,c}\log(p_{i,c}) \tag{8}$$

In the formula, L represents the total loss, which quantifies the difference between the predicted probabilities and the true labels. N is the number of samples in the dataset, indicating the total number of data points. C is the number of classes, referring to the total possible categories in the classification task. $y_{i,c}$ is the true label for sample i and class c, where $y_{i,c}$=1 if sample i belongs to class c, and $y_{i,c}$=0 otherwise (one-hot encoding). $p_{i,c}$ is the predicted probability of sample i belonging to class c, which is the output of the model. The formula calculates the cross-entropy loss for each sample and class, averages it over all samples, and provides a measure of how well the model's predictions match the actual labels.

3.2.2. AS-FCRNet experiments

After training the CNN feature extraction network, we proceed to train the AS-FCRNet. Following the procedure described in Section 2, we use 4 frames of signals as input for classification. Adam optimizer is employed with a learning rate of 0.00015 and a batch size of 64. After completing the training, we further conduct 5-fold cross-validation on the AS-FCRNet to demonstrate its benefits, including enhanced generalization capability and stability across different data partitions.

3.2.3. Performance evaluation

In this paper, we evaluate the performance of the model based on accuracy, false alarm rate, underreporting rate, and floating-point operations. The formulas for calculating accuracy(Acc), false alarm rate(FAR), underreporting rate(UR) and floating-point operations(FLOPs) are as follows:

$$ACC = (TP + TN)/(TP + TN + FP + FN) \tag{9}$$

$$FAR = FP/(TP + TN) \tag{10}$$

$$UR = FN/(TP + TN) \tag{11}$$

In the formulas, true positive (TP) represents the correctly classified intrusion events, true negative (TN) represents the correctly classified noise, false positive (FP) represents the misclassified noise, and false negative (FN) represents the missed intrusion events. FLOPs (Floating-Point Operations) denote the total number of floating-point arithmetic operations (e.g., additions and multiplications) performed by a model. The overall FLOPs are obtained by summing the contributions of all layers, each computed with its own standard formula; for conciseness, those layer-specific expressions are not listed here. Accuracy, false alarm rate, and underreporting rate are used to evaluate the model's classification performance, while FLOPs are used to assess the computational complexity of the model [40]. Accordingly, the overall quality of a model is determined jointly by its classification performance and computational efficiency.

3.3. Training environment and implementation details

All deep neural networks were implemented in the PyTorch framework. The experiment used Python 3.12, PyTorch 2.5.1, and CUDA 12.4. It was conducted on a laptop with 32 GB of RAM, an i9-14900H processor, and an RTX 4070 GPU.

4. Result

4.1. Comparative evaluation of fusion strategies

To investigate the effect of different fusion strategies, we evaluate data-level, feature-level, and decision-level fusion methods under a comparable computational budget. The results are summarized in **Table 2**, where classification accuracy and computational complexity (measured in MFLOPs) are reported.

Table 2. Comparative Evaluation of Fusion Strategies.

Model	Acc (%)	FAR(%)	UR(%)	FlOPs (MFlops)
Data_level	93.97	2.86	1.92	0.41
Feature_level	90.86	1.96	5.65	0.41
Decision_level	92.55	1.67	4.95	0.21

From **Table 2**, it can be seen that the data-level fusion (HDP-ConvNet) achieves the highest accuracy (93.97%) and the lowest underreporting rate (UR=1.92%) under a comparable compute budget (≈0.41 MFLOPs). In contrast, decision-level fusion attains the lowest false-alarm rate (FAR=1.67%) and the smallest compute (0.21 MFLOPs), but with a drop in accuracy (92.55%) and a higher UR (4.95%). The feature-level fusion, where acoustic and seismic representations from the shared CNN feature extractor (Section 2) are concatenated before classification, records the lowest accuracy (90.86%) and the highest UR (5.65%), despite having the same compute as the data-level model (0.41 MFLOPs). Considering the accuracy–efficiency trade-off and the need to minimize missed intrusion events, we adopt the data-level fusion (HDP-ConvNet) as the CNN feature extractor for subsequent experiments.

This paper conducts comparative experiments on key design parameters (Mel-band number, convolution type, and CNN width) of the CNN feature extraction network for ground moving target recognition using acoustic-seismic signals. The results are shown in **Table 3**. It shows that depthwise separable convolution (DSconv) significantly reduces computational cost (FLOPs) while maintaining or even improving classification accuracy, thus enhancing generalization; the Mel-band number is not “the more the better” and performs best at 50 in terms of ACC, false alarm rate, and miss rate; although increasing CNN width slightly boosts accuracy, it causes a sharp rise in FLOPs, so the “Normal” width offers a better accuracy-efficiency balance. Therefore, for subsequent experiments, the CNN with depthwise separable convolution, 50 Mel-bands, and normal width is selected as the feature extractor to achieve the optimal trade-off between classification performance and computational efficiency.

Table 3. Performance comparison of acoustic-seismic signal classification under different Mel-band numbers, convolution types, and CNN widths.

Mel-band number	Convolution categories	CNN width	Acc (%)	FAR(%)	UR(%)	FlOPs (MFlops)
25	DSConv	Normal	92.17	3.72	2.36	0.21
		Half	87.90	7.22	4.69	0.08
		Twice	93.42	2.78	3.02	0.63
	Conv	Normal	91.38	3.70	2.59	0.29
		Half	91.24	3.35	4.07	0.10
		Twice	93.23	2.91	1.33	0.96
50	DSConv	Normal	93.97	2.86	1.92	0.41
		Half	93.32	3.28	1.64	0.16
		Twice	93.37	3.23	2.09	1.21
	Conv	Normal	93.18	4.02	1.94	0.56
		Half	92.26	4.77	1.56	0.19
		Twice	94.30	2.21	2.85	1.83
75	DSConv	Normal	92.20	6.08	1.04	0.62
		Half	89.11	4.01	2.24	0.24
		Twice	93.97	3.65	0.84	1.82
	Conv	Normal	92.54	3.11	2.00	0.84
		Half	91.84	4.80	2.22	0.29
		Twice	94.30	2.16	2.21	2.72

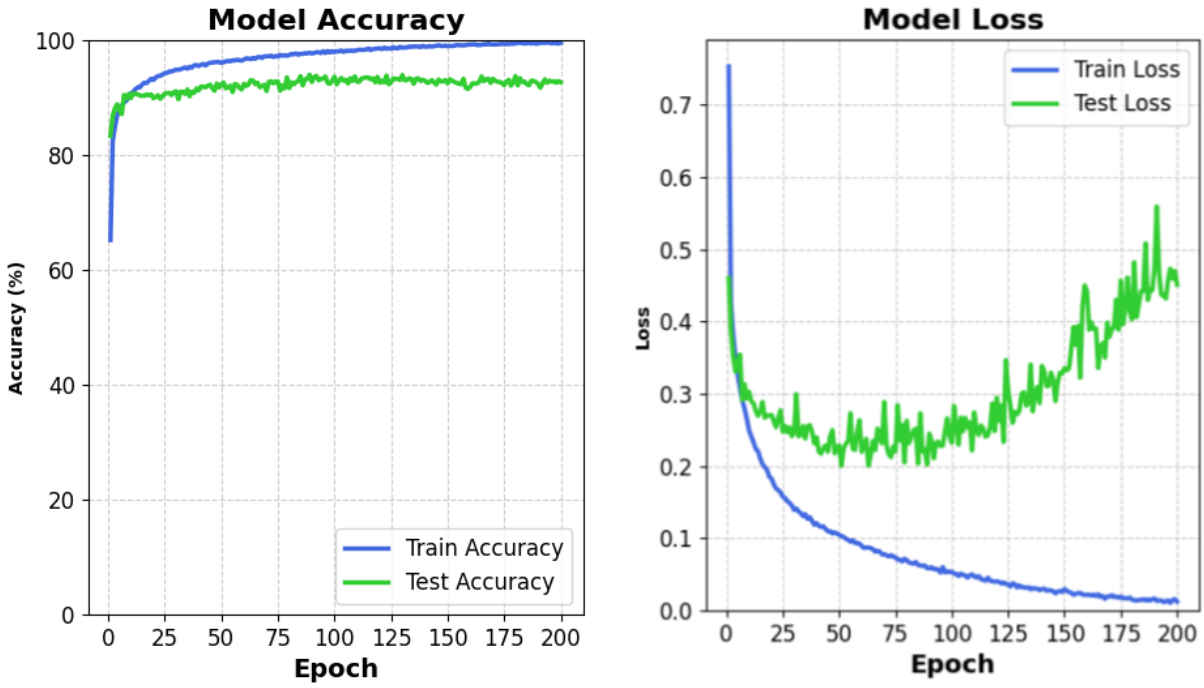
We also conducted an ablation experiment where acoustic signals were removed for recognition. Additionally, in this paper, we constructed a 1DCNNnet for recognition by mimicking the 1D CNN layer components in Mohine 's paper [18]. The results, as shown in **Table 4**, indicate that removing acoustic signals led to a significant drop in accuracy. Meanwhile, the 1DCNN network not only has high FLOPs but also achieves lower accuracy than HDP-ConvNet.

Table 4. Performance Comparison: HDP-ConvNet, Acoustic-Signal-Removed Network, and 1DCNNnet.

Model	Acc (%)	FAR (%)	UR (%)	FlOPs (MFlops)
HDP-ConvNet	93.97	2.86	1.92	0.21
HDP-ConvNet(without sound)	85.07	7.79	5.83	0.33
1D CNN net	87.90	7.91	2.64	675.14

First, the HDP-ConvNet is trained as a single-frame feature extractor for 200 epochs, and its classification performance is evaluated on the test dataset. **Figure 9** shows the loss and accuracy curves during training, while **Figure 10** presents the confusion matrix on the test set, In the confusion matrix, the class labels are abbreviated as follows: Env denotes Environment, HC denotes Human Actions, and Vh denotes Vehicle.

The trained HDP-ConvNet is then frozen and used as the per-frame feature extractor for subsequent training of AS-FCRNet, which is likewise trained for 200 epochs. **Figure 11** displays the loss and accuracy trajectories of AS-FCRNet, and **Figure 12** presents its confusion matrix on the same test set for direct comparison.



(**a**) (**b**)

Figure 9. Accuracy and loss curves during the training process of HDP-ConvNet: **(a)** Accuracy; **(b)** Loss.

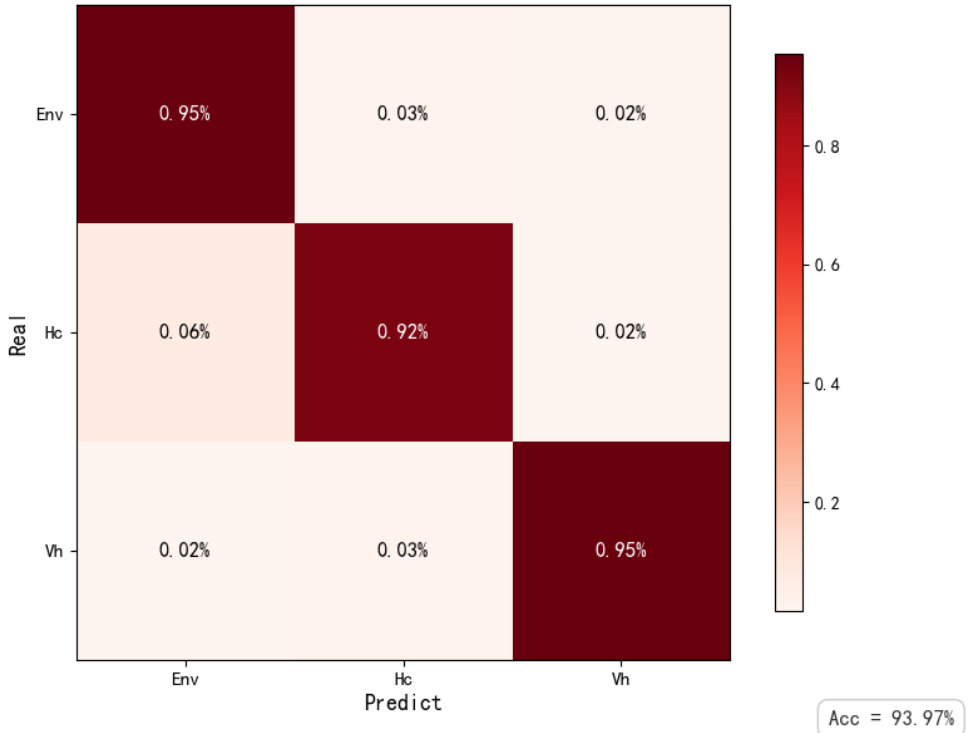
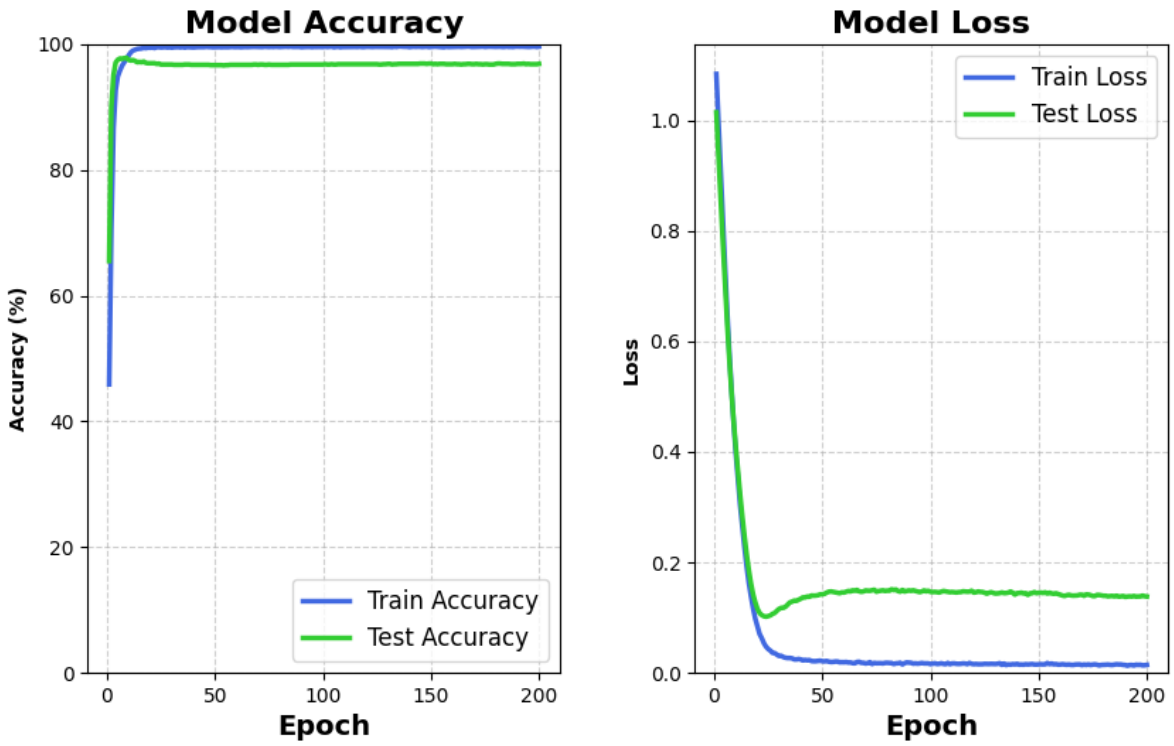


Figure 10. Confusion matrix of the HDP-ConvNet in test dataset.



(**a**) (**b**)

Figure 11. Accuracy and loss curves during the training process of AS-FCRNet: **(a)** Accuracy; **(b)** Loss.

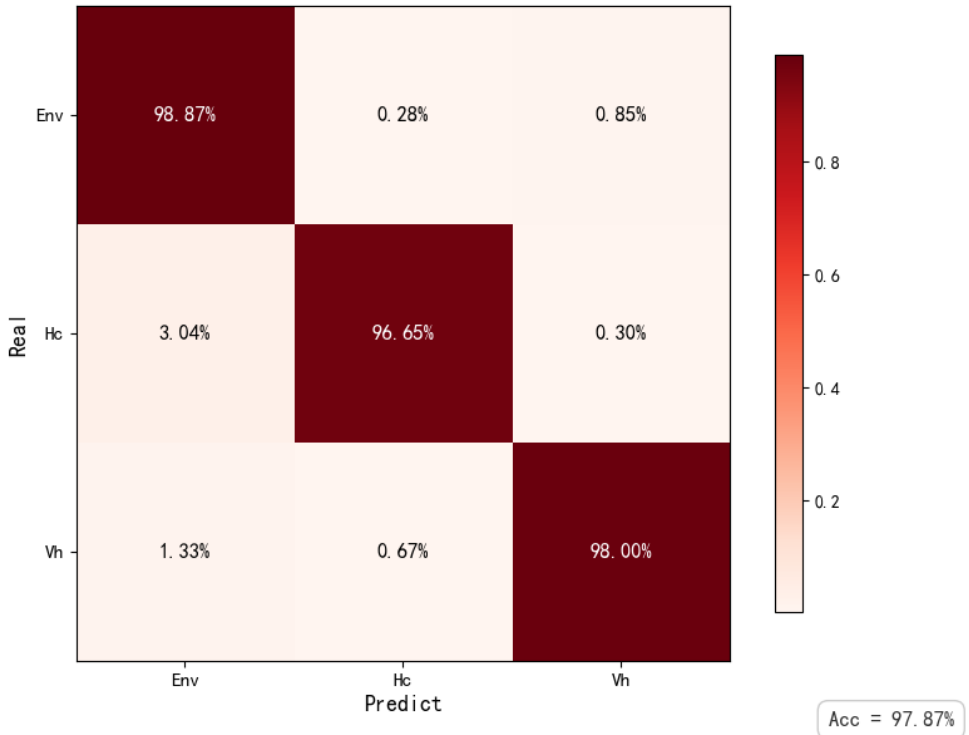


Figure 12. Confusion matrix of the AS-FCRNet in test dataset.

4.2. Performance comparison with existing methods

We replicated the LFCC-CNN model proposed by Jin et al. and the 1D CNN-BiLSTM model proposed by Tong et al. for comparison [13, 19]. Additionally, we developed a traditional machine learning classification algorithm using Mel-Frequency Cepstral Coefficients(MFCC) and SVM for comparison. We also developed two recognition models: the HDP-ConvNet, which takes 4-second signals as input, and the AS-FCRNet, which takes 2-second signals as input.The final comparison results are shown in **Table 5**. From **Table 5**, it can be seen that since AS-FCRNet does not take raw signals as input, but first extracts features and reduces dimensions using Mel-frequency spectrum feature extraction, followed by a lightweight CNN for acoustic-seismic feature fusion, and then employs LSTM to extract temporal information for multi-frame classification, AS-FCRNet significantly reduces computational complexity compared to other deep learning methods while maintaining high accuracy. Although the floating-point operations of the MFCC-SVM approach are slightly lower than those of AS-FCRNet, its other performance metrics are not as good as those of AS-FCRNet, making AS-FCRNet the best-performing model overall.

Table 5. Performance comparison between the proposed method and the benchmark methods.

Model	Acc (%)	FAR (%)	UR (%)	FlOPs (MFlops)
AS-FCRNet(4s)	97.87	1.53	0.43	1.65
LFCC-CNN	89.88	7.18	2.55	4.12
CNN-BiLSTM	92.73	2.16	5.25	651.37
MFCC-SVM	83.33	6.52	3.51	1.51
HDP-ConvNet(4s)	96.60	2.63	0.63	3.18
AS-FCRNet(2s)	95.89	2.71	0.62	0.83

Table 6 presents the 5-fold cross-validation results of the AS-FCRNet. It achieves an average validation accuracy of 99.69% with a standard deviation of merely 0.06%, which indicates the model possesses both high classification performance and excellent stability across various data subsets.

Table 6. 5-Fold Cross-Validation Results of the AS-FCRNet (4s).

Fold	Val Acc(%)	Val Acc mean	Val Acc std
1	99.62		
2	99.69		
3	99.65	99.69	0.06
4	99.79		
5	99.68		

5. Conclusion

This paper presents a ground-moving target recognition method that integrates Mel-frequency spectrum feature extraction with CNN–LSTM modeling, designed for distinguishing among environmental noise (non-target), human intrusion, and vehicle intrusion. Acoustic and seismic signals of moving targets were collected by an autonomously developed unattended ground sensor, providing reliable multi-modal data for the proposed framework.

In addition to establishing the basic AS-FCRNet model, this work further investigated three different fusion strategies—data-level fusion, feature-level fusion, and decision-level fusion—to identify the most effective way to combine acoustic and seismic modalities. Among them, the data-level fusion approach, implemented with the lightweight HDP-ConvNet, consistently achieved the best trade-off between accuracy, false alarm rate, underreporting rate, and computational complexity. The feature-level and decision-level fusion methods also demonstrated competitive performance; however, they showed relatively higher underreporting rates or reduced classification accuracy compared with the data-level approach. These results highlight the importance of carefully selecting the fusion strategy when designing acoustic–seismic recognition systems.

Overall, the proposed method achieves a classification accuracy of 97.87%, outperforming comparable approaches while maintaining low FLOPs and lightweight model complexity, making it suitable for deployment on resource-constrained unattended ground sensors. The exploration of different fusion mechanisms not only validates the robustness of the proposed network design but also provides valuable insights into how acoustic and seismic modalities can be most effectively integrated. In conclusion, the data-level fusion strategy was adopted for the final model used in subsequent experiments, as it offered the most balanced and reliable performance across all evaluation metrics.

While the proposed sound-seismic fusion-based ground moving target recognition method achieves effective performance using datasets collected by self-developed sensors, it has limitations that guide future research directions to enhance practicality and robustness. First, the current dataset lacks geographic diversity, as it focuses on specific local geological conditions—seismic signals are inherently sensitive to geological property changes, so the model's generalization to untested substrates (e.g., snow, sand, or urban pavements) remains unvalidated. To improve accuracy and generalization, future work should prioritize collecting large-scale labeled datasets across diverse geological types, ensuring the model learns substrate-specific signal patterns. Second, snow and sand scenarios pose dual challenges: stable installation of self-developed sensors in loose/soft substrates (e.g., deep snow or shifting sand) is technically difficult, and variations in substrate density (e.g., compact vs. loose sand) directly alter seismic signal characteristics—this may distort feature representations and reduce recognition reliability. Addressing these issues will require developing adaptive sensor mounting mechanisms (e.g., anchor-based or weight-adjusted designs) and optimizing the model to adapt to density-induced signal variations. Additionally, high-traffic urban environments introduce complex background noise (e.g., vehicle honking, construction vibrations), which can mask target-related sound and seismic signals, leading to misclassification. Future research could explore advanced noise suppression strategies (e.g., adaptive filtering, multi-modal attention fusion) to isolate target signals from urban interference. Resolving these limitations will enable the method to achieve broader applicability and higher robustness in real-world diverse scenarios.

Author contributions: Conceptualization, Z.L. and K.X.; methodology, K.X.; software, Z.L.; validation, Z.L.; formal analysis, Z.L.; investigation, Z.L.; data curation, K.X., W.W. and N.W.; writing—original draft preparation, Z.L; writing—review and editing, Z.L. and W.W.; project administration, W.W.; Z.L. is the first author and K.X. is the co-first author. All authors have read and agreed to the published version of the manuscript.

Funding: This work was co-funded by the Natural Science Foundation of Hunan Province (2022JJ40554) and National Natural Science Foundation of China (62305386).

Conflict of interest: The authors declare no conflict of interest.

References

1. William PE, Hoffman MW. Classification of military ground vehicles using time domain harmonics' amplitudes. IEEE Transactions on Instrumentation and Measurement. 2011; 60(11): 3720–3731. doi: 10.1109/TIM.2011.2135110
2. Prado G, Johnson R. Changing requirements and solutions for unattended ground sensors. In: Carapezza EM (editor). Unmanned/Unattended Sensors and Sensor Networks IV, Proceedings of the Optics/Photonics in Security and Defence; 5 October 2007; Florence, Italy. SPIE. 2007. p. 67360X. doi: 10.1117/12.748638
3. Tian Y, Qi H, Wang X. Target detection and classification using seismic signal processing in unattended ground sensor systems. In: Proceedings of the 2002 International Conference on Acoustics Speech and Signal Processing; 13–17 May 2002; Orlando, FL, USA. p. IV-4172-IV–4172. doi: 10.1109/ICASSP.2002.5745620
4. Bin K, Jiang Y, Fu R, et al. Multimodal attention transformer encoder for acoustic-seismic fusion target recognition. arXiv preprint. 2025. doi: 10.2139/ssrn.5254396
5. Weisser A, Miles K, Richardson MJ, et al. Conversational distance adaptation in noise and its effect on signal-to-noise ratio in realistic listening environments. The Journal of the Acoustical Society of America. 2021; 149(4): 2896–2907. doi: 10.1121/10.0004774
6. Ekpezu AO, Wiafe I, Katsriku F, et al. Using deep learning for acoustic event classification: The case of natural disasters. The Journal of the Acoustical Society of America. 2021; 149(4): 2926–2935. doi: 10.1121/10.0004771
7. Yuan Y, Shen Q, Xi W, et al. Multidisciplinary design optimization of dynamic positioning system for semi-submersible platform. Ocean Engineering. 2023; 285: 115426. doi: 10.1016/j.oceaneng.2023.115426
8. Yuan Y, Yang Q, Ren J, et al. Short-term power load forecasting based on SKDR hybrid model. Electrical Engineering. 2025; 107(5): 5769–5785. doi: 10.1007/s00202-024-02821-x
9. George J, Mary L, Riyas K. Vehicle detection and classification from acoustic signal using ANN and KNN. In: Proceedings of the 2013 International Conference on Control Communication and Computing (ICCC); 13–15 December 2013; Thiruvananthapuram, India. pp. 436–439. doi: 10.1109/ICCC.2013.6731694
10. Jin X, Sarkar S, Ray A, et al. Target detection and classification using seismic and PIR sensors. IEEE Sensors Journal. 2012; 12(6): 1709–1718. doi: 10.1109/JSEN.2011.2177257
11. Ozkaya SG, Baygin M, Dogan S, et al. Machine learning-based equipment sound classification for advanced construction management and site supervision. World Journal of Advanced Research and Reviews. 2025; 26(3): 317–329. doi: 10.30574/wjarr.2025.26.3.2178
12. Bin K, Lin J, Tong X, et al. Moving target recognition with seismic sensing: A review. Measurement. 2021; 181: 109584. doi: 10.1016/j.measurement.2021.109584
13. Dibazar AA, Yousefi A, Park HO, et al. Intelligent acoustic and vibration recognition/alert systems for security breaching detection, close proximity danger identification, and perimeter protection. In: Proceedings of the 2010 IEEE International Conference on Technologies for HomelandSecurity (HST); 8–11 November 2010; Waltham, MA, USA. pp. 351–356. doi: 10.1109/THS.2010.5654931
14. Cunningham P, Delany SJ. K-Nearest neighbour classifiers—a tutorial. ACM Computing Surveys. 2022; 54(6): 1–25. doi: 10.1145/3459665
15. Kalra M, Kumar S, Das B. Analysis of instantaneous amplitude and frequency of EWT modes for automatic target classification. In: Proceedings of the 2021 IEEE Bombay Section Signature Conference (IBSSC); 18 November 2021;

Gwalior, India. pp. 1–6. doi: 10.1109/IBSSC53889.2021.9673272
16. Narayanaswami R, Gandhe A, Tyurina A, et al. Sensor fusion and feature-based human/animal classification for unattended ground sensors. In: 2010 IEEE International Conference on Technologies for Homeland Security (HST); 8–10 November 2010; Waltham, MA, USA. pp. 344–350. doi: 10.1109/THS.2010.5655025
17. Cyriac S, Harsha BM, Woon Kim Y. Seismic activity-based human intrusion detection using deep neural networks. In: 2022 13th International Conference on Information and Communication Technology Convergence (ICTC); 19 October 2022; Jeju Island, Republic of Korea. pp. 130–135. doi: 10.1109/ICTC55196.2022.9952913
18. Damarla T, Mehmood A, Sabatier J. Detection of people and animals using non-imaging sensors. In: Proceedings of the 14th International Conference on Information Fusion; 5 July 2011; Chicago, IL, USA. pp. 1–8. Available online: https://ieeexplore.ieee.org/abstract/document/5977674
19. Park HO, Dibazar AA, Berger TW. Cadence analysis of temporal gait patterns for seismic discrimination between human and quadruped footsteps. In: Proceedings of the 2009 IEEE International Conference on Acoustics, Speech and Signal Processing; 19–24 April 2009; Taipei, Taiwan. pp. 1749–1752. doi: 10.1109/ICASSP.2009.4959942
20. Wang Y, Cheng X, Zhou P, et al. Convolutional neural network-based moving ground target classification using raw seismic waveforms as input. IEEE Sensors Journal. 2019; 19(14): 5751–5759. doi: 10.1109/JSEN.2019.2907051
21. Jin G, Ye B, Wu Y, et al. Vehicle classification based on seismic signatures using convolutional neural network. IEEE Geoscience and Remote Sensing Letters. 2019; 16(4): 628–632. doi: 10.1109/LGRS.2018.2879687
22. Tran VT, Tsai WH. Acoustic-based emergency vehicle detection using convolutional neural networks. IEEE Access. 2020; 8: 75702–75713. doi: 10.1109/ACCESS.2020.2988986
23. Yu Y, Rashidi M, Samali B, et al. Crack detection of concrete structures using deep convolutional neural networks optimized by enhanced chicken swarm algorithm. Structural Health Monitoring. 2022; 21(5): 2244–2263. doi: 10.1177/14759217211053546
24. Zhao X, Wang L, Zhang Y, et al. A review of convolutional neural networks in computer vision. Artificial Intelligence Review. 2024; 57(4): 99. doi: 10.1007/s10462-024-10721-6
25. Li J, Han L, Li X, et al. An evaluation of deep neural network models for music classification using spectrograms. Multimedia Tools and Applications. 2022; 81(4): 4621–4647. doi: 10.1007/s11042-020-10465-9
26. Wang Z, Ma Y, Gao J, et al. Remaining useful life prediction for solid-state lithium batteries based on spatial–temporal relations and neuronal ODE-assisted KAN. Reliability Engineering & System Safety. 2025; 260: 111003. doi: 10.1016/j.ress.2025.111003
27. Yuan Y, Yang Q, Ren J, et al. Short-term wind power prediction based on IBOA-AdaBoost-RVM. Journal of King Saud University - Science. 2024; 36(11): 103550. doi: 10.1016/j.jksus.2024.103550
28. Yuan Y, Yang Q, Wang G, et al. Combined improved tuna swarm optimization with graph convolutional neural network for remaining useful life of engine. Quality and Reliability Engineering International. 2025; 41(1): 174–91. doi: 10.1002/qre.3651
29. Xing K, Wang N, Wang W, et al. CNN-based multiterrain moving target recognition model for unattended ground sensor systems. Journal of Sensors. 2022; 2022: 1–10. doi: 10.1155/2022/7542114
30. Bin K, Lin J, Tong X. Edge intelligence-based moving target classification using compressed seismic measurements and convolutional neural networks. IEEE Geoscience and Remote Sensing Letters. 2022; 19: 1–5. doi: 10.1109/LGRS.2021.3055795
31. Akter R, Islam MdR, Debnath SK, et al. A hybrid CNN-LSTM model for environmental sound classification: Leveraging feature engineering and transfer learning. Digital Signal Processing. 2025; 163: 105234. doi: 10.1016/j.dsp.2025.105234
32. Mohine S, Bansod BS, Bhalla R, et al. Acoustic modality based hybrid deep 1D CNN-BiLSTM algorithm for moving vehicle classification. IEEE Transactions on Intelligent Transportation Systems. 2022; 23(9): 16206–16216. doi: 10.1109/TITS.2022.3148783
33. Nie T, Wang S, Wang Y, et al. An effective recognition of moving target seismic anomaly for security region based on deep bidirectional LSTM combined CNN. Multimedia Tools and Applications. 2023; 83(22): 61645–61658. doi: 10.1007/s11042-023-14382-5
34. Sun L, Zhang Z, Tang H, et al. Vehicle acoustic and seismic synchronization signal classification using long-term features. IEEE Sensors Journal. 2023; 23(10): 10871–10878. doi: 10.1109/JSEN.2023.3263572
35. Abdul ZKh, Al-Talabani AK. Mel frequency cepstral coefficient and its applications: A review. IEEE Access. 2022; 10: 122136–122158. doi: 10.1109/ACCESS.2022.3223444
36. Chollet F. Xception: deep learning with depthwise separable convolutions. arXiv preprint. 2016. doi: 10.48550/ARXIV.1610.02357

37. Dong S, Chen Z. A multi-level feature fusion network for remote sensing image segmentation. Sensors. 2021; 21(4): 1267. doi: 10.3390/s21041267
38. Oh SI, Kang HB. Object detection and classification by decision-level fusion for intelligent vehicle systems. Sensors. 2017; 17(1): 207. doi: 10.3390/s17010207
39. Hochreiter S, Schmidhuber J. Long short-term memory. Neural Computation. 1997; 9(8): 1735–1780. doi: 10.1162/neco.1997.9.8.1735
40. Liu S, Jiang W, Wu L, et al. Real-time classification of rubber wood boards using an SSR-based CNN. IEEE Transactions on Instrumentation and Measurement. 2020; 69(11): 8725–8734. doi: 10.1109/TIM.2020.3001370