

Article

A Risk-Sensitive Fault Diagnosis Framework for AUVs Using Learnable Dual-Branch Bandpass Filtering and Bayesian Minimum-Risk Decision

Lingyan Dong * and Yan Huo

School of Intelligent Science and Information Engineering, Shenyang University, Shenyang 110040, China

* Correspondence: donglingyan@syu.edu.cn

Abstract

An autonomous underwater vehicle (AUV) is an important operation platform in ocean exploration engineering, and the operation state of its actuator directly determines the navigation safety. However, the marine environment is complex and changeable, and the state signal characteristics of autonomous underwater vehicles are complex and vulnerable to noise interference. The performance of the traditional fault identification model degrades significantly in the noise environment, and the model ignores the risk difference of misjudgment, which may lead to serious misjudgment consequences. In order to solve the above problems, we propose a Risk-Sensitive Dual-Branch Learnable Filtering (RS-DBLF) framework for AUV fault diagnosis. The feature extraction backbone network adopts a dual-branch structure, which can adaptively capture the low-frequency slow degradation fault features and high-frequency impact fault features, and simultaneously complete the noise suppression and multi-scale fault frequency domain feature separation. Secondly, a fault asymmetric cost matrix is established for the cost-sensitive diagnosis scenario, and the expected misjudgment loss corresponding to various faults is calculated by using the posterior probability of the network output. The global minimum-risk criterion is used to replace the traditional cross entropy hard classification reasoning, so as to reduce the high-cost fault misjudgment probability from the decision level. The “Haizhe” open dataset is used to verify the effectiveness of the algorithm. The results show that the cost-sensitive dual-branch learning bandpass network has a stronger robustness in noise environments, and the minimum-risk decision based on Bayesian can effectively reduce the loss of misjudgment.

Keywords: fault diagnosis; AUV; Bayesian minimum risk; cost-sensitive classification; learnable bandpass filtering



Academic Editor: Rui Araújo

Received: 18 August 2026

Revised: 19 September 2026

Accepted: 20 September 2026

Published: 22 September 2026

Copyright: © 2026 by the authors.

Licensee MDPI, Basel, Switzerland.

This article is an open access article distributed under the terms and conditions of the [Creative Commons Attribution \(CC BY\)](https://creativecommons.org/licenses/by/4.0/) license.

1. Introduction

In recent years, autonomous underwater vehicles (AUVs) have undertaken important tasks in the fields of marine resource exploration, environmental monitoring, and military reconnaissance, and more attention has been paid to the safety of the platform [1]. In the process of executing the task, the running state of the actuator is directly related to the success or failure of the task and navigation safety. However, the real marine environment is full of uncertainty [2]. The complex underwater acoustic channel, ocean current, and the electronic noise of the sensor itself make the state monitoring results of AUVs contain some noise. Under the background of different noise intensity, how to accurately identify the

fault symptoms of the actuator from the highly coupled and non-stationary observation data has become an urgent technical problem in the field of AUV fault diagnosis.

In the AUV fault diagnosis task, different types of misjudgments often produce different consequences and bring different losses [3]. Specifically, if the fault state is misjudged as a normal state (missed judgment), it may lead to the continuous accumulation of faults, and eventually lead to the failure of the propulsion system, resulting in the irreversible loss of equipment, and serious misjudgment may even lead to the loss of the aircraft. If the normal state is misjudged as a fault state (false alarm), unnecessary redundancy detection or task interruption will be triggered, resulting in the reduction of operation efficiency, but the harm is relatively controllable. Obviously, the risks arising from different judgment results are quite different. However, the traditional fault diagnosis methods usually take the overall classification accuracy or symmetry loss as the optimization template in the design process, and ignore the significant risk asymmetry between missed and misjudged in the real task scenario.

Some researchers have tried to embed misclassification into the model training process through cost-sensitive learning, such as introducing the cost-sensitive integrated learning method into the failure prediction of submarine combat systems to reduce the risk of missing reports. However, in the field of AUV fault diagnosis, there is less research on the cost-sensitive mechanism for actuator fault diagnosis, and most methods ignore the differentiated cost matrix, resulting in the lack of an effective risk-avoidance decision mechanism in the face of high-risk missing reports.

The marine environmental noise makes the data collected by the sensor show the characteristics of strong nonlinearity and low signal-to-noise ratio. Although the traditional feature extraction methods perform well under ideal conditions, some features are easily submerged by the environmental noise in the strong noise environment, resulting in a significant decline in the signal-to-noise separation ability of the model input signal. At present, the fault diagnosis method based on deep learning is widely used in the field of underwater vehicle fault diagnosis. However, the deep learning method can automatically learn hierarchical features, but it is sensitive to input noise. In the actual deployment of AUV, the traditional model often processes the signals of each channel independently, ignoring the complementarity of multi-sensor data and failing to effectively reduce the negative impact of noise.

In order to solve the problem that the existing AUV fault diagnosis methods based on deep learning ignore the asymmetry of misjudgment cost and the lack of robustness of feature extraction in strong noise environments, this paper proposes a risk-sensitive fault diagnosis framework based on a learning double-branch bandpass filter network and Bayesian minimum-risk decision. The framework uses the dual-branch structure to adaptively extract time-frequency features, combines the display modeling ability of Bayesian decision on asymmetric misjudgment risk, and takes into account the robustness and reliability of fault diagnosis tasks, providing an innovative and practical solution for safety critical underwater applications. The main contributions of this paper are as follows:

- (1) A learning dual-branch bandpass filter network is proposed, which can directly extract high-quality time-frequency representation from strong noise signals by keeping the branch in the time domain to retain the transient impact details of the signal, and learning the optimal bandpass filter parameters in the frequency domain, which overcomes the limitation of the traditional fixed wavelet base lacking data adaptive ability.
- (2) In the reasoning stage, the Bayesian minimum-risk decision module is constructed, and the configurable risk matrix is introduced to impose explicit punishment on high-cost misjudgments. The diagnosis decision objective is transformed from maximizing

the overall accuracy to minimizing the expected risk, which effectively reduces the probability of serious failure omission.

- (3) The cost-sensitive multi-sensor fusion network has been validated on the “Haizhe” AUV benchmark dataset. The results show that compared with the baseline model, it has stronger robustness and low risk.

The remainder of this paper is organized as follows. Section 2 reviews related work. Section 3 presents the proposed methodology. Section 4 describes the experimental setup and results.

2. Related Work

2.1. Feature Extraction Network

The end-to-end feature extraction method driven by deep learning has become the mainstream paradigm in the field of mechanical fault diagnosis. According to the different signal processing methods of network structure, the existing research can be roughly divided into three categories: time-domain feature extraction, multi-scale feature extraction, and time-frequency fusion feature extraction.

In terms of time-domain feature extraction, Ince et al. [4] took the lead in applying a one-dimensional convolutional neural network (1d-cnn) to motor fault diagnosis, directly learning fault features from the original vibration signal, avoiding the dependence of manual feature design on expert experience. The experimental results show that the detection accuracy of this method is significantly better than that of the traditional shallow classifier. Zhang et al. [5] introduced the long-term and short-term memory network (LSTM) into the fault diagnosis of rotating machinery, and effectively captured the long-range time-series dependence in the vibration signal by using its gating mechanism. Chen et al. [6] proposed an end-to-end hybrid model combining gating cycle unit (Gru) and convolutional neural network (CNN) for fault diagnosis of rolling bearings, which effectively captures the long-range time-series dependence in vibration signals by using the gating mechanism of Gru. However, such pure time-domain methods only extract features from the time dimension, lack the explicit modeling ability of signal frequency-domain characteristics, and the discrimination performance is significantly degraded in strong background noise [7].

In order to enhance the richness of feature expression, a multi-scale feature extraction network has attracted extensive attention. Ismail et al. [8] migrated the perception-time architecture to the task of time-series classification. By using convolution kernels of different sizes in the parallel branch, they realized the synchronous extraction of multi granularity time-domain features, and achieved excellent performance on multiple time-series classification benchmark datasets. Zhang et al. [9] combined the Inception module with residual connections and proposed a multi-scale residual convolutional network for rolling bearing fault diagnosis, which alleviates the degradation problem of deep networks through residual learning.

In summary, existing research lacks a lightweight feature extraction structure that can both adaptively learn frequency-domain filter parameters and preserve time-domain transient details.

2.2. Cost-Sensitive Decision

Most of the existing fault diagnosis research uses cross-entropy loss function to train the model, and output the prediction category with argmax rule in the reasoning stage [10,11]. This paradigm aims to maximize the overall classification accuracy, and implicitly assumes that the costs caused by all types of misjudgments are equal. However, in safety critical systems such as AUVs, the consequences of misjudging a serious fault (such

as thruster fracture) as a healthy state are far greater than those in the opposite direction, and the costs of the two types of errors are highly asymmetric [12].

In order to deal with the problem of misjudgment cost asymmetry, a cost-sensitive learning method is introduced into the field of fault diagnosis. Chen et al. [13] proposed a cost-sensitive dense network combining improved DenseNet and adaptive weighted cross-entropy loss to solve the problem of data imbalance in industrial scenarios, and achieved better performance than the standard method in the task of unbalanced fault diagnosis. Xu et al. [14] designed a cost-sensitive learning mechanism that enables the GBDT classifier to effectively learn discriminant boundaries under extremely imbalanced fault data, significantly reducing missed alarm rates and average misclassification costs compared to standard SVM, GBDT, and mainstream resampling methods such as SMOTE and ADASYN. Li et al. [15] proposed a cost-sensitive dense network, which assigns appropriate misclassification costs to each class through an adaptive weighted cross-entropy loss, effectively mitigating the impact of data imbalance on deep learning-based fault diagnosis models. However, the above methods have common problems at the cost modeling level: the category weight of weighted cross entropy usually depends on experience setting and lacks probability interpretation, while the cost matrix of cost-sensitive SVM only acts on the classification boundary, and the posterior probability is not derived from the perspective of decision theory.

Bayesian minimum-risk decision provides a more rigorous theoretical framework for the above problems. Based on the estimated posterior probability and the predefined risk matrix, the method selects the decision category that minimizes the expected risk [16]. At present, the framework has been successfully applied in the fields of medical diagnosis and communication signal detection. In the field of fault diagnosis, a few studies began to explore the potential of Bayesian decision making. For the problem of active fault diagnosis of discrete-time stochastic systems, Guo et al. [17] designed auxiliary input signals based on a Bayesian minimum-risk decision framework, and improved the diagnosis performance by minimizing the risk of misdiagnosis. Because the misdiagnosis risk cannot be closed, its upper bound is derived and the input design criteria are established based on Bhattacharyya distance. Tao et al. [18] proposed a multi signal fusion fault diagnosis method. Its core is probabilistic neural network (PNN) based on the optimal parameters, and the minimum Bayesian risk reclassification model is introduced. The prior knowledge of a PNN classifier is used to reduce the risk of misdiagnosis, especially the misclassification of high-risk fault modes. Ren et al. [19] proposed an adaptive cost-sensitive learning strategy. This method can adaptively calculate the misclassification cost for each sample according to the sample number distribution, the convergence trend of the category, and the convergence trend of the sample, so as to solve the problem of industrial data imbalance. The strategy is applied to different deep learning models and loss functions, and the effectiveness is verified on the rolling bearing dataset.

In summary, existing research lacks a Bayesian decision framework that systematically couples high-quality posterior probabilities generated by deep feature extractors with flexibly configurable risk matrices. This paper addresses the problem at both the feature-extraction and decision-mechanism levels by introducing an adaptive dual-branch filtering network and a risk-matrix-driven Bayesian decision framework, enabling fault diagnosis that reconciles accuracy with risk sensitivity.

3. Methodology

In this research, a closed set fault diagnosis model for underwater vehicles is proposed, which is composed of two parts: feature extraction network and decision reasoning module. Firstly, the AUV multi-sensor time-series data are processed in parallel through the double-

branch feature extraction network. The global branch uses the large-scale convolution kernel to capture the system-level fault features, and the local branch uses the small-scale convolution kernel to extract the local correlation features between sensors. Secondly, the learnable bandpass filter module is introduced to enhance the dynamic frequency domain of the fused multimodal features to adaptively suppress noise interference and strengthen the fault-related frequency band components. Then, it is classified into output category posterior probability by full connection. Finally, in the decision-making stage, the model introduces the minimum-risk decision-making mechanism based on the cost matrix in addition to the traditional probability maximization strategy, and realizes the systematic avoidance of high-risk miscalculations by quantifying the differentiated consequences of different miscalculations. Figure 1 shows the overall structure of the proposed frame.

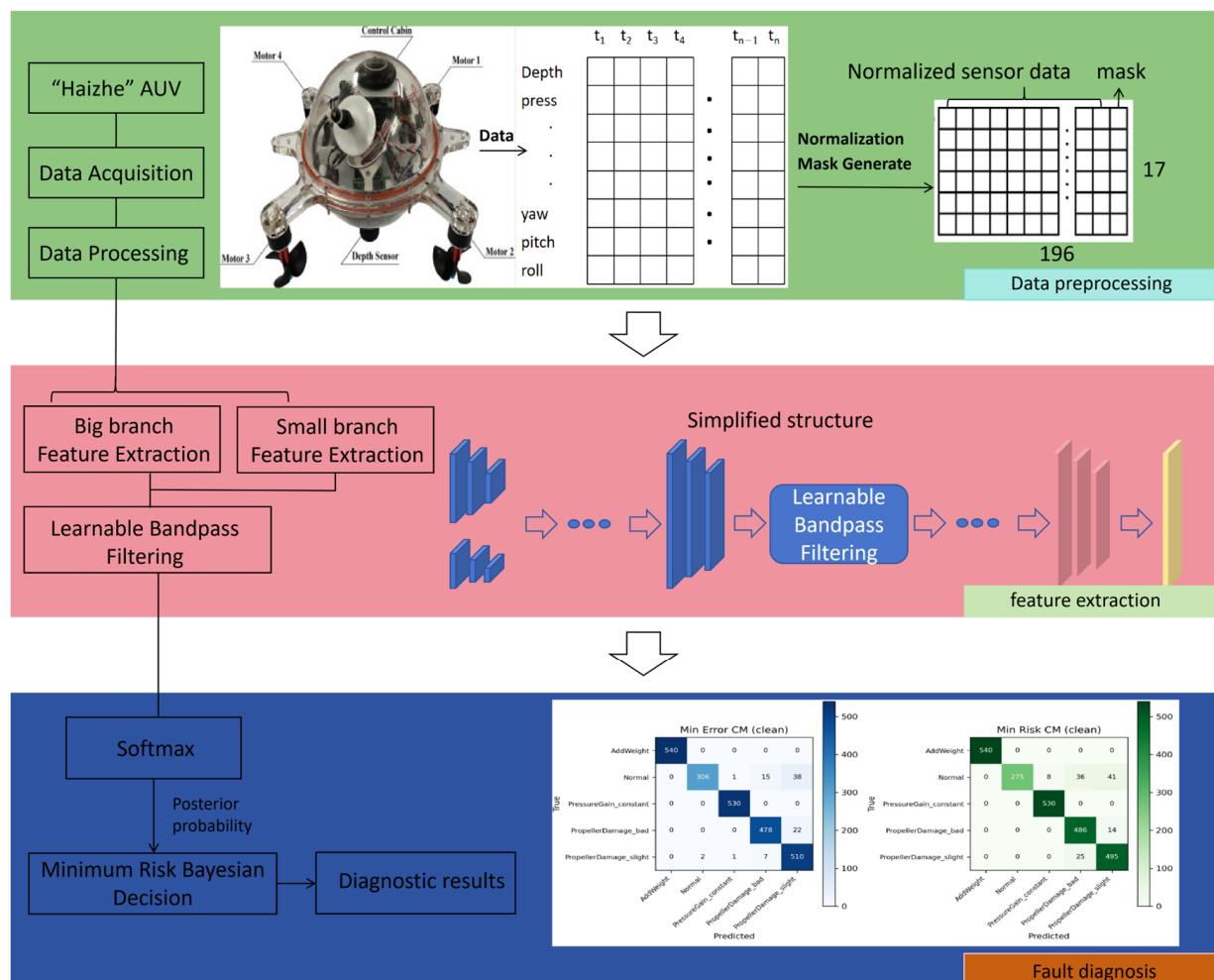


Figure 1. The structure of AUV fault diagnosis using RS-DBLF.

3.1. Dual-Branch Learnable Filtering Network

In order to ensure that the model can simultaneously perceive the long-term dependence and transient anomalies in the AUV operating state, a parallel double-branch convolution feature extraction network is designed in this paper. The network contains two different feature extraction branches of the receptive field, which are responsible for global trend modeling and local detail capture, and realize multi-scale information fusion through a unified feature projection layer. After fusing features, a trainable bandpass enhancement module is introduced, which adaptively learns the optimal frequency band response characteristics through end-to-end back-propagation. The feature extraction network framework is shown in Figure 2.

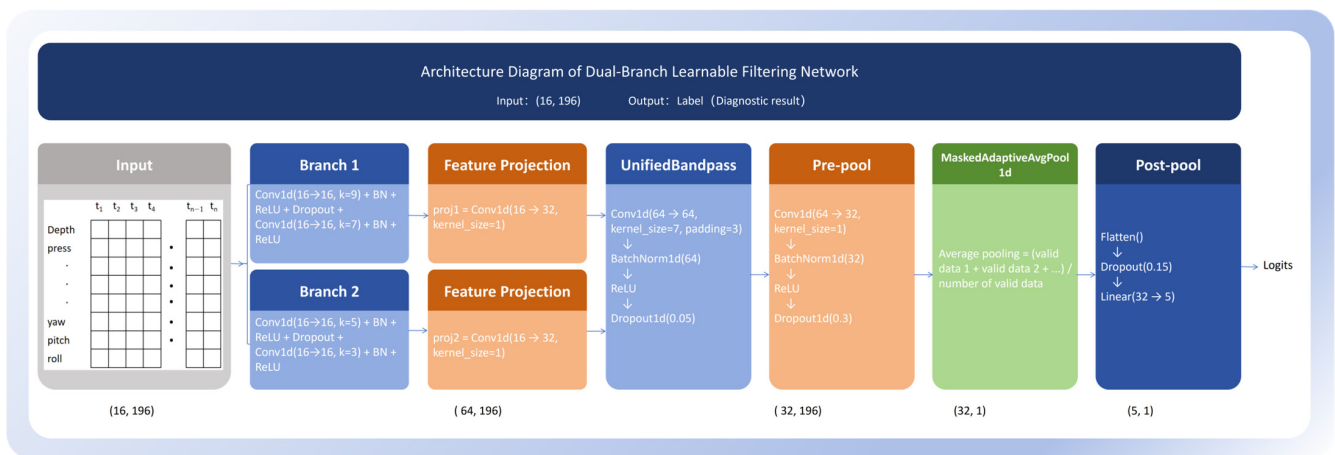


Figure 2. Network architecture of the DBLF model.

3.1.1. Dual-Branch Feature Extraction

The dual-branch architecture extracts multi-scale features via parallel kernels with different receptive fields, enabling complementary representations that improve noise robustness and discriminative power [20,21]. Given the preprocessed multi-sensor input sequence $\mathbf{X} \in \mathbb{R}^{T \times F}$, where $T = 196$ denotes the temporal length and $F = 16$ denotes the number of sensor channels, a mask matrix $\mathbf{M} \in \{0, 1\}^{T \times 1}$ is introduced to identify valid observation positions, ensuring that padded regions do not participate in feature computation. The input data are first transposed to the channel-first format $\mathbf{X}' \in \mathbb{R}^{T \times F}$ and element-wise multiplied with the mask, yielding the masked input:

$$\tilde{\mathbf{X}} = \mathbf{X}' \odot \mathbf{M}' \quad (1)$$

where $\odot$ denotes element-wise multiplication and $\mathbf{M}' \in \mathbb{R}^{1 \times T}$ is the channel-first form of the mask.

The global feature extraction branch employs one-dimensional convolutional layers with kernel sizes $k_1 = [29, 17]$, whose larger receptive fields enable effective capture of low-frequency trend components and long-range temporal dependencies in sensor signals. The forward propagation of this branch can be formally expressed as shown in Equations (2)–(4):

$$\mathbf{Z}_1^{(1)} = \sigma \left(\text{BN} \left(\text{Conv1D}_{k=29} \left(\tilde{\mathbf{X}} \right) \right) \right) \quad (2)$$

$$\mathbf{Z}_1^{(2)} = \sigma \left(\text{BN} \left(\text{Conv1D}_{k=17} \left(\mathbf{Z}_1^{(1)} \right) \right) \right) \quad (3)$$

$$\mathbf{Z}_1 = \text{Dropout} \left(\mathbf{Z}_1^{(2)} \right) \quad (4)$$

The local feature extraction branch adopts one-dimensional convolutional layers with kernel sizes $k_2 = [5, 3]$, whose smaller receptive fields focus on high-frequency transient features and local abrupt patterns within sensor channels, as shown in Equations (5)–(7):

$$\mathbf{Z}_2^{(1)} = \sigma \left(\text{BN} \left(\text{Conv1D}_{k=5} \left(\tilde{\mathbf{X}} \right) \right) \right) \quad (5)$$

$$\mathbf{Z}_2^{(2)} = \sigma \left(\text{BN} \left(\text{Conv1D}_{k=3} \left(\mathbf{Z}_2^{(1)} \right) \right) \right) \quad (6)$$

$$\mathbf{Z}_2 = \text{Dropout} \left(\mathbf{Z}_2^{(2)} \right) \quad (7)$$

where $\sigma(\cdot)$ denotes the ReLU activation function, BN denotes batch normalization, and the dropout rate is set to 0.1. The mask is reapplied after each convolutional layer to ensure that padded positions do not participate in feature learning, which can be formulated as

$$\mathbf{Z} \leftarrow \mathbf{Z} \odot \tilde{\mathbf{M}} \quad (8)$$

where $\tilde{\mathbf{M}}$ is interpolated to match the spatial dimensions of $\mathbf{Z}$.

Subsequently, 1×1 convolutions are applied to project the outputs of the two branches into a unified feature space $\mathbb{R}^{D \times T}$, where $D = 32$:

$$\mathbf{H}_1 = \text{Conv1D}_{1 \times 1}(\mathbf{Z}_1) \quad (9)$$

$$\mathbf{H}_2 = \text{Conv1D}_{1 \times 1}(\mathbf{Z}_2) \quad (10)$$

Finally, the multi-scale fusion representation is obtained through channel-wise concatenation:

$$\mathbf{H} = [\mathbf{H}_1; \mathbf{H}_2] \in \mathbb{R}^{2D \times T} \quad (11)$$

3.1.2. Learnable Bandpass Filtering Module

Unlike traditional fixed-parameter filter designs, this module adaptively learns the optimal frequency band response characteristics through end-to-end back-propagation, enabling the model to automatically focus on the most discriminative frequency intervals according to the distribution characteristics of fault data. The proposed learnable band-pass front-end takes the concatenated multi-branch feature map as input and produces a frequency-selective feature representation. Let $\mathbf{H} \in \mathbb{R}^{C_{in} \times L}$ denote the input feature map, where C_{in} is the number of input channels and L is the temporal length. The module is defined as follows:

$$\mathbf{U} = \mathbf{H} * \mathbf{W} + \mathbf{b}, \mathbf{W} \in \mathbb{R}^{C_{out} \times C_{in} \times K} \quad (12)$$

where $*$ denotes the 1D convolution operation along the temporal dimension, $\mathbf{W}$ is the learnable convolution kernel, $\mathbf{b} \in \mathbb{R}^{C_{out}}$ is the bias vector, C_{out} is the number of output channels, and K is the kernel size. The output $\mathbf{U} \in \mathbb{R}^{C_{out} \times L}$ is the pre-normalization feature map.

Each convolutional kernel with K learnable taps is mathematically equivalent to a finite impulse response (FIR) filter, whose frequency response is given by

$$H_W(f) = \sum_{k=0}^{K-1} W[k] e^{-j2\pi f k} \quad (13)$$

where $W[k]$ is the k -th tap of the kernel and f is the normalized frequency. The magnitude response $|H_W(f)|$ determines the frequency selectivity of the filter.

To ensure a bandpass response, we initialize the kernel with a difference-of-Gaussians (DoG) kernel:

$$g_1[k] = \exp\left(-\frac{\left(k - \frac{K-1}{2}\right)^2}{2\sigma_1^2}\right), g_2[k] = \exp\left(-\frac{\left(k - \frac{K-1}{2}\right)^2}{2\sigma_2^2}\right) \quad (14)$$

$$\omega^{DoG}[k] = \frac{g_1[k]}{\sum_{k'} g_1[k']} - \frac{g_2[k]}{\sum_{k'} g_2[k']} \quad (15)$$

$$\mathbf{W}^{(0)}[k] = \omega^{DoG}[k] + \eta, \eta \sim \mathcal{N}(0, \sigma_{init}^2) \quad (16)$$

where K is the number of kernel taps, and σ_1 and σ_2 are the standard deviations of the two Gaussian kernels. Since $\sigma_1 < \sigma_2$, their difference yields a bandpass-like shape. In Equation (16), σ_{init} is the standard deviation of a small random perturbation added to the DoG kernel, which prevents all output channels from having identical initial weights and thus preserves the expressive capacity of the module.

After initialization, the forward propagation of the module proceeds as follows:

$$\tilde{\mathbf{U}} = \text{BN}(\mathbf{U}) \quad (17)$$

$$\hat{\mathbf{U}} = \sigma(\tilde{\mathbf{U}}) \quad (18)$$

$$\mathbf{H}' = \text{Dropout}(\hat{\mathbf{U}}) \quad (19)$$

3.1.3. Mask-Based Adaptive Pooling and Classification

In the pooling stage, traditional global average pooling incorporates zero values from padded positions into the computation, leading to dilution of effective signals and consequently compromising the accuracy of feature representations. To address this issue, this paper designs a masked adaptive average pooling layer that performs pooling operations exclusively on valid time steps.

First, the bandpass-filtered features are channel-compressed through a 1×1 convolution:

$$\mathbf{V} = \text{Dropout}(\sigma(\text{BN}(\text{Conv1D}_{1 \times 1}(\mathbf{H}')))) \quad (20)$$

where $\mathbf{V} \in \mathbb{R}^{C \times T}$ and $C = 32$ denotes the compressed feature dimension. The masked adaptive average pooling is then expressed as

$$\mathbf{v} = \frac{\sum_{t=1}^T \mathbf{V}_{:,t} \cdot M_t}{\sum_{t=1}^T M_t + \epsilon} \in \mathbb{R}^C \quad (21)$$

where $\epsilon = 10^{-8}$ is a small constant to prevent division by zero. This operation ensures that the pooling result is unaffected by sequence length variations and padding strategies, thereby yielding more robust feature representations.

The compressed feature vector $\mathbf{v}$ is then fed into the classifier, where it is mapped to class logits through a dropout layer and a fully connected layer:

$$\mathbf{o} = \mathbf{W}_c \cdot \text{Dropout}(\mathbf{v}) + \mathbf{b}_c \quad (22)$$

$$\mathbf{p} = \text{Softmax}(\mathbf{o}) = \frac{\exp(o_k)}{\sum_{j=1}^K \exp(o_j)}, \quad k = 1, 2, \dots, K \quad (23)$$

where $\mathbf{W}_c \in \mathbb{R}^{K \times C}$, $\mathbf{b}_c \in \mathbb{R}^K$ are learnable parameters, $K = 5$ denotes the total number of fault categories, $\mathbf{p} \in \Delta^{K-1}$ denotes the posterior probability distribution over all classes, and o_k denotes the logit for the k -th class. This classification head employs a dropout rate of 0.15 to prevent overfitting.

3.2. Cost-Sensitive Decision Strategy

Traditional fault diagnosis methods typically adopt the cross-entropy decision strategy during the inference stage, i.e., selecting the class with the maximum posterior probability as the final prediction [22,23]. However, misclassifying a severe propeller fault (PropellerDamage_bad) as normal operation (Normal) incurs substantially higher costs than misclassifying a minor load change (AddWeight) as a pressure gain anomaly (PressureGain_constant). Thus, in practical AUV engineering applications, different types of

misclassifications entail significantly different safety consequences and economic costs. To address this, the present research introduces a cost-sensitive decision mechanism on top of the traditional probabilistic decision framework, achieving optimized decision making during inference by minimizing the expected conditional risk. The decision process is illustrated in Figure 3.

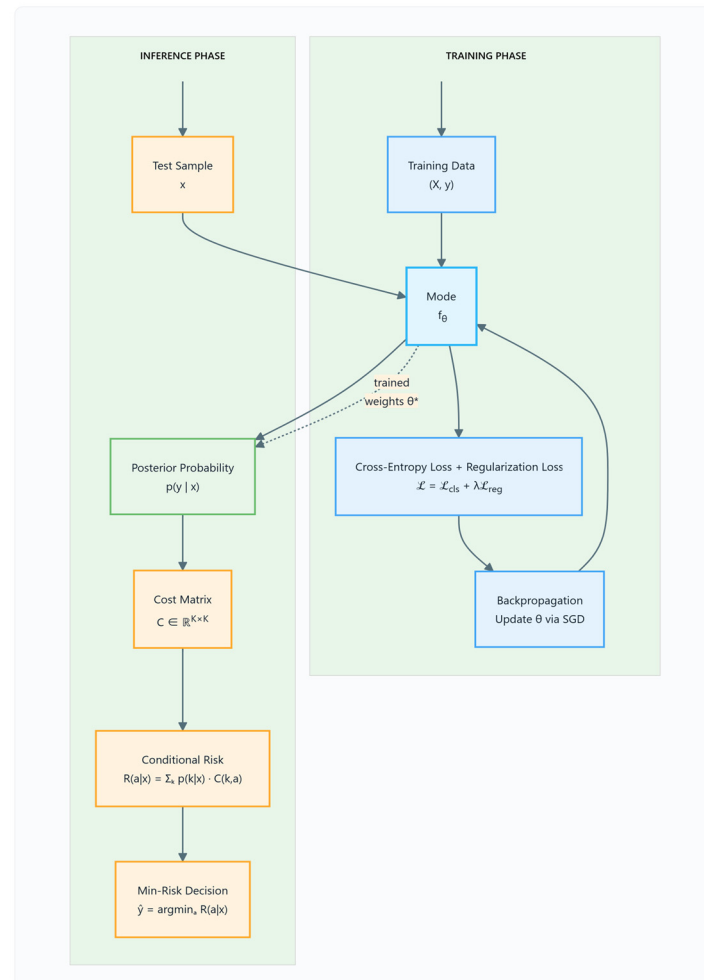


Figure 3. Decision process.

Define the cost matrix $\mathbf{C} \in \mathbb{R}^{K \times K}$, where element C_{ij} denotes the cost of misclassifying true class i as class j . In this matrix, both row and column indices follow the class order: Ad-dWeight, Normal, PressureGain_constant, PropellerDamage_bad, PropellerDamage_slight. The design of this matrix adheres to the following principles:

- (1) The highest costs are assigned to severe fault miss-detections;
- (2) The moderate costs are assigned to false alarms of normal states;
- (3) The low costs are assigned to misclassifications among similar fault types.

For a given posterior probability $\mathbf{p}$, the conditional risk of choosing decision $a \in Y$ is

$$R(a|x) = \sum_{k=1}^K p_k \cdot C_{k,a} \quad (24)$$

This equation can be further expressed in vector inner product form:

$$R(a|x) = \mathbf{p}^T \mathbf{C}_{:,a} \quad (25)$$

where $C_{:,a}$ denotes the a -th column of the cost matrix. The cost-sensitive decision strategy selects the decision that minimizes the risk:

$$\hat{y}_{CS} = \arg \min_{a \in \mathcal{Y}} R(a|x) \quad (26)$$

This strategy achieves a quantitative trade-off between probability estimation and misclassification costs within the framework of Bayesian decision theory. Compared with the traditional CE decision strategy, the cost-sensitive decision approach, through the risk-minimization criterion defined in Equations (25) and (26), effectively reduces the occurrence probability of high-risk misclassification events, making it particularly suitable for safety-critical AUV fault diagnosis scenarios.

3.3. Loss Function

Model training employs a joint objective function consisting of a classification loss and a regularization loss. The total loss function is defined as the weighted sum of the two terms:

$$\mathcal{L} = \mathcal{L}_{cls} + \lambda_{reg} \mathcal{L}_{reg} + \lambda_{freq} \mathcal{L}_{freq} \quad (27)$$

where λ_{reg} and λ_{freq} are the regularization coefficients that balance the three loss terms. $\mathcal{L}_{cls}$ is the classification loss, and $\mathcal{L}_{reg}$ is the regularization loss. $\mathcal{L}_{freq}$ is the frequency-domain regularization loss.

The classification loss adopts the multi-class cross-entropy loss, which measures the discrepancy between the predicted probability distribution and the ground-truth label distribution:

$$\mathcal{L}_{cls} = -\frac{1}{N} \sum_{i=1}^N \sum_{k=1}^K y_{i,k} \log p_{i,k} \quad (28)$$

where N denotes the batch size, $y_{i,k} \in \{0, 1\}$ denotes the one-hot encoded ground-truth label, and $p_{i,k}$ denotes the predicted probability of the i -th sample belonging to the k -th class. For each sample, y_i is a one-hot vector.

To prevent overfitting and encourage weight sparsity, Frobenius norm regularization is imposed on all convolutional layers in the model:

$$\mathcal{L}_{reg} = \frac{1}{N} \sum_{l \in \mathcal{S}} \|W_l\|_F^2 \quad (29)$$

where $\mathcal{S}$ denotes the set of all convolutional layers, and $\|\cdot\|_F$ denotes the Frobenius norm.

While the DoG initialization provides a bandpass-like starting point, the kernel may still drift toward a low-pass solution during training. To prevent this, we add a frequency-domain regularization term to the loss, which penalizes low-frequency and high-frequency energy while encouraging mid-frequency energy:

$$E_{low} = \overline{|H_w(f)|}_{f < 0.05}, E_{mid} = \overline{|H_w(f)|}_{0.10 \leq f \leq 0.35}, E_{high} = \overline{|H_w(f)|}_{f > 0.40} \quad (30)$$

$$\mathcal{L}_{freq} = E_{low} + E_{high} - E_{mid} \quad (31)$$

where $\overline{(\cdot)}$ denotes the mean over the specified frequency band. Under the joint effect of the DoG initialization and the frequency-domain regularization, the learned filter maintains a bandpass shape throughout training, while the exact cutoff frequencies emerge from the K taps rather than being manually specified.

Minimizing this objective function enables the model to accurately fit the training data while maintaining good generalization performance. The cross-entropy loss enables efficient gradient-based optimization, while the Frobenius norm regularization effectively

constrains model complexity, together ensuring both rapid convergence and robust generalization in the proposed diagnostic framework.

4. Experimental Study

4.1. AUV Dataset

The experimental validation of this research is conducted based on a fault diagnosis dataset for AUVs provided by Zhejiang University. The dataset is available at <https://data.mendeley.com/datasets/7rp2pmr6mx/1> (accessed on 1 June 2025), which is a complete open-source dataset [24]. The dataset is collected using a self-developed small-scale AUV named “Haizhe” as the experimental platform, which is equipped with a propulsion motor, a depth sensor, and a control module. The name and description of each column is listed as in Table 1.

Table 1. Sensor data of “Haizhe” AUV dataset.

Name	Description
time	The absolute time for “Haizhe” to record data.
pwm 1–4	The duration (in microseconds) of the high-level signal in a 100 Hz PWM wave. This signal drives Motors 1–4.
depth	The depth value (in meters) measured by depth sensor.
press	The pressure value (in Pa) measured by depth sensor
voltage	The voltage value (in V) of battery.
roll/pitch/yaw	The roll angles/pitch angles/yaw angles (in degrees) measured by nine-axis IMU.
a_x, a_y, a_z	The acceleration (in m/s^2) in the body coordinate frame, measured by nine-axis IMU.
w_row, w_pitch, w_yaw	The angular velocity (in degrees/s) of rotation in the body coordinate frame, measured by nine-axis IMU.

During the experiments, the Haizhe AUV was outfitted with both a depth sensor and a nine-axis IMU for motion and environmental perception. Each trial was designed to simulate a single fault scenario, with each episode spanning 10 to 20 s. The platform was configured to exhibit five distinct operational conditions, namely, normal operation, depth sensor malfunction, additional payload, slight propeller damage, and severe propeller damage. The corresponding fault annotations and the statistical breakdown of dataset samples are summarized in Table 2, while the detailed fault configuration settings are illustrated in Figure 4.

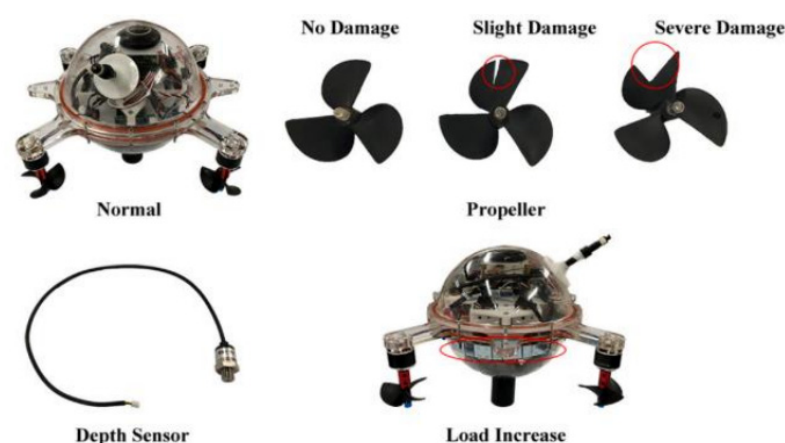


Figure 4. Fault setting types for “Haizhe”. The red circles indicate the fault locations.

Table 2. Details of “Haizhe” dataset.

Condition	Normal	PressureGain_Constant	PropellerDamage_Bad	PropellerDamage_Slight	AddWeight
Dataset	182	268	266	260	249
Train	146	214	213	208	199
Test	36	54	53	52	50
Label	0	1	2	3	4

The Haizhe dataset consists of short fault episodes, in which each AUV experimental run is stored as a single CSV file containing one complete fault sequence. In our experiments, each CSV file is treated as one sample, so the number of samples equals the number of episodes. We perform the training/validation/test split at the episode level, with corresponding split ratios of 68%, 12%, and 20%. Each CSV file is assigned to only one of the three subsets, and no single experimental run appears in more than one subset.

This study implements a systematic preprocessing pipeline for raw time-series data. First, 16-channel sensor readings are extracted from CSV files and uniformly truncated to the first 196 time steps to ensure consistent sequence length; short sequences with fewer than 196 steps are padded by replicating the last valid time step, and a binary mask matrix is simultaneously generated to distinguish genuine observations from padded positions. Second, during feature standardization, the mean and standard deviation of each channel are computed over all valid time steps in the training set, and Z-score normalization is applied to the entire dataset. The normalized values are subsequently clipped to the range $[-3, 3]$ to mitigate the influence of extreme outliers. Finally, the standardized features and the mask are concatenated along the last dimension to form the complete input tensor for the model. This preprocessing strategy allows variable-length time series to be fed into deep learning models in a unified format, while the mask mechanism prevents padded positions from interfering with feature learning and loss computation.

4.2. Experimental Setup

The frameworks and versions used in our experiments are PyTorch 2.0.1, torchvision 0.15.2, NumPy 1.24.3, scikit-learn 1.3.0, SciPy 1.11.1, pandas 2.0.3, and Matplotlib 3.7.2. All experiments were run on the same machine, and all sources of randomness were controlled by a single global seed. The seeds for the five independent runs are 142, 242, 342, 442, and 542.

All models share the same input shape and the same five-class output head. A coarse grid search was performed on the validation set for each model family to determine appropriate hyperparameters. To guarantee a fair and reproducible comparison, all methods were trained and evaluated with a consistent set of hyperparameters across the Haizhe dataset. Table 3 provides a comprehensive summary of these settings, which encompass the learning rate, mini-batch size, and total training epochs.

Table 3. Hyperparameter settings for model training.

Hyperparameters	lr	Optimizer	Batch	Epochs
Set	0.0005	Adam	64	200

4.3. Experiment Process

To validate the effectiveness of the proposed algorithm, we conducted multiple sets of comparative experiments. Each experiment was independently repeated five times. The reported metrics are the arithmetic mean of the experimental results, with standard deviation provided to characterize result fluctuations. Evaluation metrics are used to

measure the performance of classification models. The four metrics used in this experiment are explained as follows:

- (1) Accuracy: The proportion of correctly predicted samples to the total number of samples. A higher value indicates stronger overall classification capability of the model.
- (2) F1-Macro: The F1-score is first computed for each class individually, and then the unweighted average is taken across all classes. Values closer to 1 indicate better overall model performance.
- (3) Recall-Macro: The ratio of correctly predicted positive samples of a given class to the actual number of samples belonging to that class. Values closer to 1 suggest a stronger ability of the model to identify positive instances.
- (4) G-Mean: The geometric mean of recall scores across all classes. It is particularly sensitive to disparities in recall among classes. Values closer to 1 denote superior comprehensive performance in balancing both positive and negative classes.

4.3.1. Comparative Study on Feature Extractor

This experiment aims to verify the feature extraction capability of the proposed two-branch learnable bandpass network on clean data. Comprehensive comparisons with seven baseline models are conducted on five-class fault diagnosis tasks. All models adopt consistent CE decision rules, and are implemented with identical standardized preprocessing, unified Adam optimizer configurations, as well as statistical validation based on five independent runs. The experimental results are presented below.

In Table 4, NoBandpass denotes the model with the bandpass module removed, while FixedBandpass represents the model that replaces the learnable filtering module with fixed bandpass filtering. As shown in the table, the proposed DBLF and CNN-LSTM achieve the highest accuracy of 0.9682, and also obtain competitive results in terms of F1-Macro, Recall-Macro, and G-Mean, which verifies the effectiveness of DBLF for this task. The comparative models including SVM, 1D-CNN, and Transformer deliver slightly inferior performance, while LSTM shows relatively poorer results. The ablation experiments reveal that removing the bandpass module (NoBandpass) leads to a decline in all metrics, and the FixedBandpass setting causes a substantial performance drop. This demonstrates that the adaptive bandpass filtering mechanism can effectively extract features and plays a vital role in improving recognition performance. Figure 5 shows the frequency response curve of the learnable bandpass front-end.

Table 4. Results of the comparative study on feature extraction.

Model	Accuracy	F1-Macro	Recall-Macro	G-Mean
SVM	0.9649 ± 0.0055	0.9631 ± 0.0058	0.9629 ± 0.0054	0.9625 ± 0.0054
1D-CNN	0.9620 ± 0.0076	0.9602 ± 0.0088	0.9602 ± 0.0091	0.9608 ± 0.0096
LSTM	0.9535 ± 0.0152	0.9492 ± 0.0169	0.9484 ± 0.0177	0.9463 ± 0.0185
CNN-LSTM	0.9682 ± 0.0070	0.9656 ± 0.0077	0.9652 ± 0.0082	0.9644 ± 0.0086
Transformer	0.9600 ± 0.0142	0.9578 ± 0.0149	0.9574 ± 0.0151	0.9566 ± 0.0155
DBLF	0.9682 ± 0.0094	0.9652 ± 0.0111	0.9628 ± 0.0121	0.9610 ± 0.0135
NoBandpass	0.9608 ± 0.0128	0.9573 ± 0.0142	0.9543 ± 0.0146	0.9526 ± 0.0155
FixedBandpass	0.8506 ± 0.0184	0.8224 ± 0.0298	0.8267 ± 0.0214	0.7719 ± 0.0744

As shown in Figure 5, the mean magnitude response (red curve) exhibits a clear band-pass shape: the low-frequency components (0–0.1) are attenuated, the mid-frequency band (0.15–0.35) is enhanced with a peak of approximately 0.44 at 0.22, and the high-frequency components (>0.35) are attenuated to approximately 0.34 at the Nyquist frequency. The small standard deviation across output channels indicates that this behavior is consis-

tent across the module rather than being dominated by a few channels. The right half of Figure 5 further shows six representative output channels along with their 3 dB bandwidths. All selected channels peak within the target mid-frequency band, and their lower 3 dB cutoff frequencies fall in the range from 0.06 to 0.09, confirming clear suppression of low-frequency components. These results provide direct empirical evidence that the proposed module learns a bandpass response rather than a low-pass or all-pass one.

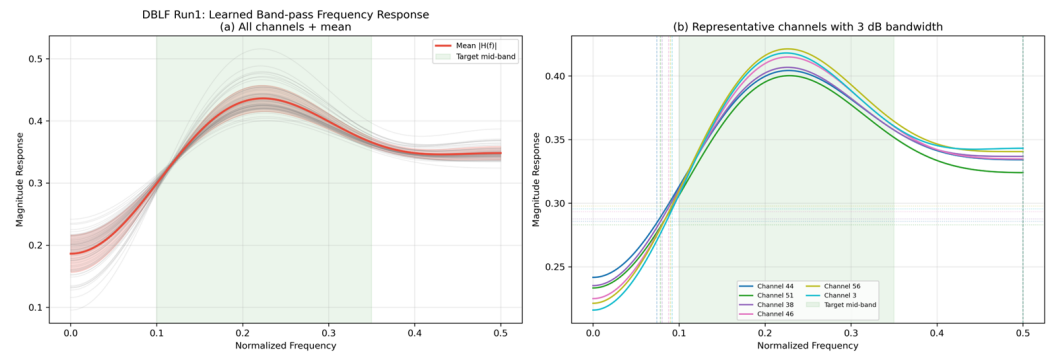


Figure 5. Learned frequency response of the proposed bandpass front-end. (a) Grey lines: All individual channels; Red line: Mean response; Red shading: Variance. (b) Dashed lines: 3 dB thresholds; Vertical lines: Cutoff frequencies.

Since the Bayes minimum-risk decision operates directly on the posterior probabilities, we evaluate the calibration of the model using ECE, NLL, and the multi-class Brier score. The results are shown in Table 5.

Table 5. Comparison of calibration and uncertainty quantification metrics for the DBLF model.

Model	Set	ECE (Raw)	ECE (Cal)	NLL (Raw)	NLL (Cal)	Brier (Raw)	Brier (Cal)	T
DBLF	clean	0.021	0.028	0.103	0.110	0.058	0.056	1.35
DBLF	0 dB	0.026	0.032	0.140	0.151	0.080	0.078	1.35

As shown in Table 5, the raw probabilities are already well calibrated, with ECE = 0.021 and NLL = 0.103 on clean data, and ECE = 0.026 and NLL = 0.140 on 0 dB data. We further apply temperature scaling, which is the standard post hoc calibration method. The fitted temperature is $T = 1.35$. We observe mixed-calibration performance after temperature scaling. Specifically, both expected calibration error (ECE) and negative log-likelihood (NLL) increase slightly under all experimental conditions. This suggests that post hoc temperature scaling brings no consistent calibration improvement for our model in the tested scenarios. Therefore, we adopt the original model probabilities for minimum-risk decision making in this work. Notably, the temperature parameter is exclusively fitted on the training/validation subset rather than the test set to avoid data leakage and ensure rigorous calibration evaluation.

4.3.2. Comparative Study on Decision Strategies

Next, we performed a comparative study of decision strategies. For an individual sample, when the predicted class is a , the risk associated with that prediction is

$$Risk(a) = C[y_{true}, a] \quad (32)$$

where C denotes the cost matrix and y_{true} is the ground-truth class. Thus, the total risk over the entire test set is computed as

$$\text{Total_Risk} = \sum_{i=1}^N C[y_i, \hat{y}_i] \quad (33)$$

where N is the total number of samples in the test set, y_i is the true class label of the i -th sample, $\hat{y}_i$ is the predicted class label of the i -th sample.

The average risk is defined as the total risk divided by the number of samples:

$$\text{Avg_Risk} = \frac{1}{N} \sum_{i=1}^N C[y_i, \hat{y}_i] \quad (34)$$

The distinction between the two decision strategies lies in how the predicted class $\hat{y}$ is selected. Cross-entropy (CE) decision selects the class with the highest posterior probability, whereas minimum-risk decision selects the class that minimizes the expected risk, which inherently accounts for the penalties imposed by the cost matrix on high-cost misclassifications. The corresponding calculation formulas are given in Equations (35) and (36):

$$\hat{y} = \arg \max P(y|x) \quad (35)$$

$$\hat{y} = \arg \min_a \sum_c P(c|x) \cdot C[c, a] \quad (36)$$

where $P(c|x)$ denotes the posterior probability estimated by the model, $C[c, a]$ denotes the cost incurred when the true class c is misclassified as a .

In this experiment, we conduct a comparative study of two decision strategies on clean datasets, utilizing the cost matrix presented in Equation (37). To assess the model's effectiveness in reducing risk, we employ the average risk reduction metric, which is formally defined in Equation (38).

$$C = \begin{bmatrix} 0 & 5 & 3 & 2 & 2 \\ 2 & 0 & 2 & 8 & 5 \\ 3 & 6 & 0 & 4 & 3 \\ 4 & 50 & 6 & 0 & 3 \\ 3 & 20 & 4 & 2 & 0 \end{bmatrix} \quad (37)$$

$$\text{Risk_Reduction} = \frac{\text{Avg_Risk}_{CE} - \text{Avg_Risk}_{CS}}{\text{Avg_Risk}_{CE}} \times 100\% \quad (38)$$

where Avg_Risk_{CE} denotes average risk achieved by the CE decision strategy, Avg_Risk_{CS} denotes average risk achieved by the cost-sensitive (CS) decision strategy. The experimental results are shown in Table 6.

Table 6. Experimental results of decision strategy comparison.

Model	Accuracy		Average Risk		Average Risk Reduction
	CE	CS	CE	CS	
SVM	0.9649 ± 0.0055	0.9233 ± 0.0111	0.3339 ± 0.1896	0.2106 ± 0.0348	36.9%
1D-CNN	0.9788 ± 0.0060	0.9633 ± 0.0052	0.2408 ± 0.1632	0.1094 ± 0.0213	54.6%
LSTM	0.9461 ± 0.0140	0.9306 ± 0.0213	0.4841 ± 0.2664	0.3355 ± 0.1411	30.7%
CNN-LSTM	0.9633 ± 0.0093	0.9584 ± 0.0060	0.3233 ± 0.1261	0.2653 ± 0.0823	17.9%
Transformer	0.9657 ± 0.0133	0.9510 ± 0.0184	0.2694 ± 0.1562	0.2033 ± 0.0865	24.5%
RS-DBLF	0.9714 ± 0.0089	0.9461 ± 0.0083	0.1886 ± 0.0912	0.1608 ± 0.0402	14.7%

The experimental results demonstrate that the minimum-risk Bayes decision entails a slight degradation in accuracy, while yielding a marked reduction in the average risk.

4.3.3. Robustness Analysis Under Noise Interference

To demonstrate the robustness of our proposed algorithm, we compared its performance under varying noise levels. The training set was kept noise-free, and Gaussian white noise was added to the test set to simulate different SNR conditions. To reduce experimental runtime, we selectively reduced the number of test models. The results are summarized in Tables 7–9.

Table 7. Accuracy analysis under noise interference.

Model	Clean	20 dB	15 dB	10 dB	5 dB	0 dB
SVM (mean)	0.9592	0.1469	0.1469	0.1469	0.1469	0.1469
SVM (std)	0.0310	0.0000	0.0000	0.0000	0.0000	0.0000
1D-CNN (mean)	0.9755	0.9755	0.9796	0.9796	0.9469	0.7959
1D-CNN (std)	0.0125	0.0123	0.0125	0.0128	0.0100	0.0419
LSTM (mean)	0.9306	0.9265	0.9306	0.9347	0.9224	0.8776
LSTM (std)	0.0152	0.0065	0.0076	0.0061	0.0048	0.0106
Transformer (mean)	0.9796	0.9796	0.9755	0.9673	0.9714	0.9551
Transformer (std)	0.0080	0.0085	0.0076	0.0079	0.0089	0.0190
CNN-LSTM (mean)	0.9714	0.9714	0.9714	0.9714	0.9388	0.8122
CNN-LSTM (std)	0.0095	0.0098	0.0125	0.0090	0.0110	0.0129
RS-DBLF (mean)	0.9714	0.9714	0.9755	0.9673	0.9755	0.9347
RS-DBLF (std)	0.0084	0.0095	0.0085	0.0104	0.0085	0.0090

Table 8. Average risk analysis under noise interference.

Model	Clean	20 dB	15 dB	10 dB	5 dB	0 dB
SVM (mean)	0.1510	2.3959	2.3959	2.3959	2.3959	2.3959
SVM (std)	0.0439	0.0000	0.0000	0.0000	0.0000	0.0000
1D-CNN (mean)	0.0939	0.1020	0.0980	0.1143	0.1918	0.6367
1D-CNN (std)	0.0742	0.0778	0.0775	0.0669	0.0830	0.3141
LSTM (mean)	0.6163	0.6163	0.4898	0.4571	0.6490	0.5673
LSTM (std)	0.0687	0.0743	0.0712	0.0547	0.0684	0.1600
Transformer (mean)	0.1714	0.1714	0.1796	0.1959	0.1224	0.2408
Transformer (std)	0.0440	0.0385	0.0578	0.0591	0.0769	0.0954
CNN-LSTM (mean)	0.2408	0.2490	0.2490	0.2449	0.3388	0.5551
CNN-LSTM (std)	0.0820	0.0605	0.0647	0.0744	0.0806	0.0793
RS-DBLF (mean)	0.1510	0.1673	0.1633	0.1837	0.1837	0.2980
RS-DBLF (std)	0.0343	0.0396	0.0673	0.0505	0.0571	0.0868

Table 9. Comparison of model efficiency.

Model	Params (K)	Inf. Time (ms)	Batch Time (ms)	Throughput (Samples/s)
1D-CNN	39.717	0.293	1.951	32,802.1
LSTM	149.893	2.483	15.649	4089.7
Transformer	103.749	4.176	98.494	649.8
CNN-LSTM	188.581	1.3	5.866	10,873.9
RS-DBLF	57.814	0.835	7.85	8152.8

Tables 7 and 8 present the accuracy and average risk of each model under different SNR conditions, respectively. Experimental results demonstrate that the proposed DBLF method achieves superior comprehensive performance in both noise robustness and risk control.

In terms of accuracy, DBLF attains 93.47% under extreme 0 dB noise with a degradation rate of only 3.78%, outperforming LSTM at 87.76%, ranking closely behind Transformer at 95.51%. Transformers achieve high pattern recognition accuracy [25]. However, when

considering computational efficiency, DBLF achieves anti-noise capability, significantly more efficient than Transformer with its 103.7 K parameters and 4.18 ms inference latency.

In terms of risk control, DBLF maintains the average risk within an extremely narrow range from 0.151 to 0.184 from clean data to 5 dB, with a fluctuation of only 0.034, the smallest risk variation among all models. This validates the stability of the Bayesian minimum-risk decision framework under noisy environments. Even under strong noise interference, the model maintains the cost-sensitive optimal decision boundary, effectively avoiding missed detections of high-risk fault modes.

Notably, LSTM exhibits an anomalous risk decrease at 0 dB, revealing the inconsistency between its cross-entropy training objective and the minimum-risk objective. In contrast, DBLF fundamentally resolves this issue through explicit cost-sensitive decision making.

Comprehensively considering accuracy, risk control, and computational efficiency, DBLF provides a favorable balance among these competing objectives. Its lightweight design and strong anti-noise capability make it a promising candidate for resource-constrained AUV-embedded platforms, offering a potential solution for real-time underwater fault diagnosis.

4.3.4. Comparative Experiment on Cost Matrices

Finally, we investigate how the structure of the cost matrix influences the performance discrepancy between the CE decision and the minimum-risk Bayes decision, with the aim of verifying whether the minimum-risk decision can consistently reduce risk under different cost settings, and how the “degree of extremity” of the cost matrix affects the comparative outcomes of the two strategies.

The training set remains noise-free, while the test set is evaluated under two conditions: clean (noise-free) and with 0 dB Gaussian white noise. Three types of cost matrices are adopted—uniform, moderate, and extreme—as detailed below.

The uniform cost matrix is defined as follows:

$$C = \begin{bmatrix} 0 & 1 & 1 & 1 & 1 \\ 1 & 0 & 1 & 1 & 1 \\ 1 & 1 & 0 & 1 & 1 \\ 1 & 1 & 1 & 0 & 1 \\ 1 & 1 & 1 & 1 & 0 \end{bmatrix} \quad (39)$$

A uniform cost matrix has zeros on the diagonal, and all misclassification costs are equal. This matrix serves as a baseline control group, representing the traditional classification scenario in which fault severity is not distinguished and all misclassifications are penalized equally. It is used to compare and verify the risk changes brought about by the introduction of asymmetric costs.

The moderate cost matrix is defined as follows:

$$C = \begin{bmatrix} 0 & 5 & 3 & 2 & 2 \\ 2 & 0 & 2 & 8 & 5 \\ 3 & 6 & 0 & 4 & 3 \\ 4 & 50 & 6 & 0 & 3 \\ 3 & 20 & 4 & 2 & 0 \end{bmatrix} \quad (40)$$

The medium cost matrix assigns asymmetric misclassification costs based on the hazard severity levels of AUV faults. For example, PropellerDamage is a high-risk fault that directly leads to loss of navigation control, mission interruption, or even destruction of the vehicle; the cost of misclassifying propeller damage as normal ($3 \rightarrow 0$) is set to 50, significantly higher than that of other types of misclassification. The misclassification costs

among the remaining faults are set relatively according to their hazard severity. This matrix group simulates a conventional offshore AUV operation scenario and represents engineers' moderate-level risk estimation of the consequences of various misclassifications.

The extreme cost matrix is defined as follows:

$$C = \begin{bmatrix} 0 & 10 & 5 & 3 & 3 \\ 3 & 0 & 3 & 15 & 8 \\ 5 & 12 & 0 & 6 & 5 \\ 6 & 100 & 10 & 0 & 4 \\ 5 & 80 & 6 & 3 & 0 \end{bmatrix} \quad (41)$$

The extreme cost matrix further amplifies the penalties for misclassifying high-risk faults on the basis of the medium cost matrix, simulating deep-sea AUV missions characterized by long distance and the absence of real-time human intervention. The above values do not come from precise quantified losses of real AUV accidents (the losses of actual maritime accidents are difficult to quantify directly), but are determined based on the relative risk ranking of underwater fault consequences. These matrices serve as illustrative, scenario-oriented cost structures for sensitivity analysis, rather than real empirical values derived from expert elicitation. Our focus is on the relative changing trends of model risk under different cost matrices, rather than the absolute risk values. Following the experimental conditions outlined above, the results are summarized in Tables 10–12.

Table 10. Experimental results with the uniform cost matrix.

Model	Clean					0 dB				
	Accuracy		Average Risk		Risk_Reduction	Accuracy		Average Risk		Risk_Reduction
	CE	CS	CE	CS		CE	CS	CE	CS	
SVM (mean)	0.9592	0.9592	0.0408	0.0408	0%	0.1469	0.1469	0.8531	0.8531	0%
SVM (std)	0.0060	0.0045	0.0439	0.0400	-	0.0000	0.0000	0.0000	0.0000	-
1D-CNN (mean)	0.9755	0.9755	0.0245	0.0245	0%	0.7959	0.7959	0.2041	0.2041	0%
1D-CNN (std)	0.0078	0.0073	0.0568	0.0498	-	0.0095	0.0087	0.0621	0.0618	-
LSTM (mean)	0.9673	0.9673	0.0327	0.0327	0%	0.9306	0.9306	0.0694	0.0694	0%
LSTM (std)	0.0091	0.0095	0.0552	0.0620	-	0.0085	0.0079	0.0586	0.0577	-
Transformer (mean)	0.9673	0.9673	0.0327	0.0327	0%	0.9633	0.9633	0.0367	0.0367	0%
Transformer (std)	0.0098	0.0089	0.0495	0.0440	-	0.0084	0.0085	0.0402	0.0358	-
CNN-LSTM (mean)	0.9592	0.9592	0.0408	0.0408	0%	0.7918	0.7918	0.2082	0.2082	0%
CNN-LSTM (std)	0.0075	0.0098	0.0587	0.0555	-	0.0067	0.0058	0.0438	0.0411	-
RS-DBLF (mean)	0.9796	0.9796	0.0204	0.0204	0%	0.9469	0.9469	0.0531	0.0531	0%
RS-DBLF (std)	0.0070	0.0061	0.0425	0.0408	-	0.0068	0.0066	0.0410	0.0351	-

Under the uniform cost matrix, all misclassifications incur equal cost, and the CS decision degenerates to the CE decision; consequently, both strategies yield identical accuracy and average risk for all models. This confirms that, under uniform costs, the cost-sensitive decision is consistent with the Bayes-optimal decision. The accuracy results indicate that RS-DBLF and Transformer are significantly more robust to strong noise than the other baselines. Under the moderate cost matrix, the CE and CS decisions begin to diverge. In the clean condition, CS reduces the average risk for most models. Notably, RS-DBLF exhibits a slight negative reduction (−3.03%). This reflects that the empirical risk realized by the CS rule on the finite test sample is marginally higher than that of the CE decision, rather than implying that the CE decision approaches the Bayes-optimal decision. Factors contributing to this discrepancy may include calibration error in the predicted probabilities, sampling variability in the finite test set, and the specific cost structure employed. Under the extreme cost matrix, because an extremely high penalty is imposed on misclassifying Propeller-

Damage as Normal, the CS decision yields substantial risk reductions for all models under both conditions. The considerable risk reduction achieved by RS-DBLF suggests that the cost-sensitive decision effectively mitigates noise-induced errors.

Table 11. Experimental results with the moderate cost matrix.

Model	Clean					0 dB				
	Accuracy		Average Risk		Risk_Reduction	Accuracy		Average Risk		Risk_Reduction
	CE	CS	CE	CS		CE	CS	CE	CS	
SVM (mean)	0.9592	0.9429	0.2122	0.1510	28.85%	0.1469	0.2204	16.8490	2.3959	85.78%
SVM (std)	0.0058	0.0050	0.0425	0.0390	-	0.0000	0.0000	0.0000	0.0000	-
1D-CNN (mean)	0.9755	0.9755	0.6122	0.0939	84.67%	0.7959	0.7796	0.9469	0.6367	32.76%
1D-CNN (std)	0.0085	0.0080	0.0474	0.0476	-	0.0089	0.0091	0.0307	0.0420	-
LSTM (mean)	0.9673	0.9592	0.1959	0.1429	27.08%	0.9306	0.9265	0.7837	0.3837	51.04%
LSTM (std)	0.0081	0.0059	0.0432	0.0329	-	0.0098	0.0069	0.0453	0.0328	-
Transformer (mean)	0.9673	0.9673	0.3714	0.1837	50.55%	0.9633	0.9592	0.3265	0.1673	48.75%
Transformer (std)	0.0075	0.0069	0.0325	0.0322	-	0.0069	0.0054	0.0328	0.0321	-
CNN-LSTM (mean)	0.9592	0.9592	0.2449	0.1429	41.67%	0.7918	0.7796	0.7959	0.6490	18.46%
CNN-LSTM (std)	0.0089	0.0075	0.0254	0.0298	-	0.0085	0.0067	0.0325	0.0325	-
RS-DBLF (mean)	0.9796	0.9429	0.1347	0.1388	−3.03%	0.9469	0.8980	0.7755	0.3061	60.53%
RS-DBLF (std)	0.0054	0.0049	0.0297	0.0255	-	0.0074	0.0065	0.0267	0.0198	-

Table 12. Experimental results with the extreme cost matrix.

Model	Clean					0 dB				
	Accuracy		Average Risk		Risk_Reduction	Accuracy		Average Risk		Risk_Reduction
	CE	CS	CE	CS		CE	CS	CE	CS	
SVM (mean)	0.9592	0.9184	0.5429	0.3020	44.36%	0.1469	0.2204	42.1878	3.8082	90.97%
SVM (std)	0.0054	0.0049	0.0521	0.0500	-	0.0000	0.0000	0.0000	0.0000	-
1D-CNN (mean)	0.9755	0.9592	1.5347	0.1959	87.23%	0.7959	0.7959	2.6816	1.1592	56.77%
1D-CNN (std)	0.0076	0.0069	0.0598	0.0533	-	0.0098	0.0075	0.0432	0.0333	-
LSTM (mean)	0.9673	0.9469	0.5102	0.2653	48.00%	0.9306	0.9224	2.3224	0.7102	69.42%
LSTM (std)	0.0085	0.0079	0.0526	0.0495	-	0.0076	0.0059	0.0433	0.0398	-
Transformer (mean)	0.9673	0.9592	0.8776	0.2122	75.81%	0.9633	0.9510	1.1265	0.4980	55.80%
Transformer (std)	0.0067	0.0065	0.0473	0.0399	-	0.0078	0.0069	0.0359	0.0287	-
CNN-LSTM (mean)	0.9592	0.9510	0.5878	0.2694	54.17%	0.7918	0.7714	1.8571	1.2327	33.63%
CNN-LSTM (std)	0.0092	0.0090	0.0432	0.0395	-	0.0099	0.0085	0.0521	0.0439	-
RS-DBLF (mean)	0.9796	0.9347	0.4082	0.2367	42.00%	0.9469	0.8816	1.9878	0.5143	74.13%
RS-DBLF (std)	0.0066	0.0061	0.0359	0.0359	-	0.0075	0.0071	0.0431	0.0354	-

5. Conclusions

To address the challenges of strong noise interference and high-cost misjudgment risks in AUV fault diagnosis, this paper proposes a cost-sensitive multi-sensor fusion fault diagnosis method. The method first designs a learning-based dual-branch bandpass filtering network, which employs a large-kernel branch to capture low-frequency trends and transient impact details, and a small-kernel branch to capture high-frequency components; the two branches are then fused by a learnable bandpass front-end, which is initialized with a difference-of-Gaussians kernel and constrained by a frequency-domain regularization loss, thereby enabling the direct extraction of discriminative features from heavily noisy signals. Subsequently, a Bayesian minimum-risk decision module is constructed at the inference stage, incorporating a configurable cost matrix to transform the diagnostic decision objective from maximizing overall accuracy to minimizing expected risk, which effectively reduces the probability of critical fault omission.

Experimental validation on the “Haizhe” AUV benchmark dataset demonstrates that, compared with baseline models, RS-DBLF achieves a favorable trade-off among

accuracy, average risk, noise robustness, and efficiency, confirming that RS-DBLF is a promising solution for cost-sensitive fault diagnosis under noisy conditions. Future work will focus on improving the calibration of probabilistic outputs so that the CS decision can more fully exploit cost information. These findings should be interpreted within the scope of the experiments. All results were obtained on a single benchmark dataset with additive Gaussian noise, and no embedded-platform or real underwater deployment was demonstrated. Therefore, the reported robustness reflects behavior under the simulated setting, and statements about practical applicability remain preliminary. Future work includes validation on additional datasets or transfer to real platforms.

Author Contributions: Conceptualization, L.D.; methodology, L.D.; software, L.D.; validation, L.D.; formal analysis, L.D.; investigation, Y.H.; resources, L.D.; data curation, Y.H.; writing—original draft preparation, L.D. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: The raw data supporting the conclusions of this article will be made available by the corresponding author on request.

Acknowledgments: During the preparation of this manuscript, the authors used DeepSeek V4.1 Flash for language polishing and grammar checking. The authors take full responsibility for the content of the publication.

Conflicts of Interest: The authors declare no conflicts of interest.

References

1. Wynn, R.B.; Huvenne, V.A.I.; Le Bas, T.P.; Murton, B.J.; Connelly, D.P.; Bett, B.J.; Ruhl, H.A.; Morris, K.J.; Peakall, J.; Parsons, D.R.; et al. Autonomous Underwater Vehicles (AUVs): Their past, present and future contributions to the advancement of marine geoscience. *Mar. Geol.* **2014**, *352*, 451–468. [\[CrossRef\]](#)
2. Ji, S.; Thai, B.S.; Yoo, S.J.; Youn, W. A novel fuzzy adaptive finite-time extended state observer based robust control for an autonomous underwater vehicle subject to external disturbances and measurement noises. *Ocean Eng.* **2025**, *318*, 120141. [\[CrossRef\]](#)
3. Wang, Y.; Zhu, K.; Sun, M.; Deng, Y. An Ensemble Learning Approach for Fault Diagnosis in Self-Organizing Heterogeneous Networks. *IEEE Access* **2019**, *7*, 125662–125675. [\[CrossRef\]](#)
4. Ince, T.; Kiranyaz, S.; Eren, L.; Askar, M.; Gabbouj, M. Real-Time Motor Fault Detection by 1-D Convolutional Neural Networks. *IEEE Trans. Ind. Electron.* **2016**, *63*, 7067–7075. [\[CrossRef\]](#)
5. Zhang, J.; Wang, P.; Yan, R.; Gao, R.X. Long short-term memory for machine remaining life prediction. *J. Manuf. Syst.* **2018**, *48*, 78–86. [\[CrossRef\]](#)
6. Chen, J.; Jing, H.; Chang, Y.; Liu, Q. Gated Recurrent Unit Based Recurrent Neural Network for Remaining Useful Life Prediction of Rolling Bearings. *Reliab. Eng. Syst. Saf.* **2019**, *185*, 372–382. [\[CrossRef\]](#)
7. Zhang, W.; Peng, G.; Li, C.; Chen, Y.; Zhang, Z. A New Deep Learning Model for Fault Diagnosis with Good Anti-Noise and Domain Adaptation Ability on Raw Vibration Signals. *Sensors* **2017**, *17*, 425. [\[CrossRef\]](#) [\[PubMed\]](#)
8. Ismail Fawaz, H.; Lucas, B.; Forestier, G.; Pelletier, C.; Schmidt, D.F.; Weber, J.; Webb, G.I.; Idoumghar, L.; Muller, P.-A.; Petitjean, F. InceptionTime: Finding AlexNet for Time Series Classification. *Data Min. Knowl. Discov.* **2020**, *34*, 1936–1962. [\[CrossRef\]](#)
9. Zhang, W.; Li, X.; Ding, Q. Deep Residual Learning-Based Fault Diagnosis Method for Rotating Machinery. *ISA Trans.* **2019**, *95*, 295–305. [\[CrossRef\]](#) [\[PubMed\]](#)
10. Wu, Y.; Wang, A.; Zhou, Y.; Zhu, Z.; Zeng, Q. Fault Diagnosis of Autonomous Underwater Vehicle with Missing Data Based on Multi-Channel Full Convolutional Neural Network. *Machines* **2023**, *11*, 960. [\[CrossRef\]](#)
11. Chen, Y.M.; Wang, Y.Z.; Yu, Y.; Wang, J.R.; Gao, J. A Fault Diagnosis Method for the Autonomous Underwater Vehicle via Meta-Self-Attention Multi-Scale CNN. *J. Mar. Sci. Eng.* **2023**, *11*, 1121. [\[CrossRef\]](#)
12. Dong, L.; Huo, Y. Research on Fault Diagnosis Method for Autonomous Underwater Vehicles Based on Improved LSTM Under Data Missing Conditions. *Appl. Sci.* **2025**, *15*, 11570. [\[CrossRef\]](#)

13. Chen, J.; Zhang, W.; Li, X.; Ding, Q. A cost-sensitive dense network for imbalanced fault diagnosis. In Proceedings of the 2022 International Conference on Sensing, Measurement & Data Analytics (ICSMD), Harbin, China, 30 November–2 December 2022; pp. 1–6. [\[CrossRef\]](#)
14. Xu, Q.; Lu, S.; Jia, W.; Jiang, C. Imbalanced fault diagnosis of rotating machinery via multi-domain feature extraction and cost-sensitive learning. *J. Intell. Manuf.* **2020**, *31*, 1467–1481. [\[CrossRef\]](#)
15. Li, Y.; Zhang, X.; Liu, Z. Lightweight Network with Variable Asymmetric Rebalancing Strategy for Small and Imbalanced Fault Diagnosis. *Machines* **2022**, *10*, 879. [\[CrossRef\]](#)
16. Angelotti, G.; Drougard, N.; Chaneil, C.P.C. An Offline Risk-Aware Policy Selection Method for Bayesian Markov Decision Processes. *Artif. Intell.* **2026**, *354*, 104519. [\[CrossRef\]](#)
17. Guo, Y.; Liu, Q.; He, X. Active Fault Diagnosis for Stochastic Systems Subject to Non-Convex Input Constraints. *Automatica* **2024**, *164*, 111631. [\[CrossRef\]](#)
18. Tao, L.; Sun, L.; Wu, Y.; Lu, C.; Ma, J.; Cheng, Y.; Suo, M. Multi-signal fusion diagnosis of gearbox based on minimum Bayesian risk reclassification and adaptive weighting. *Measurement* **2022**, *187*, 110358. [\[CrossRef\]](#)
19. Ren, Z.; Zhu, Y.; Kang, W.; Fu, H.; Niu, Q.; Gao, D.; Yan, K.; Hong, J. Adaptive cost-sensitive learning: Improving the convergence of intelligent diagnosis models under imbalanced data. *Knowl.-Based Syst.* **2022**, *241*, 108296. [\[CrossRef\]](#)
20. Deng, L.; Zhao, C.; Qiu, R.; Wang, G.; Ling, Y. DWTNet: A novel dual-branch interactive network for fault feature extraction of rotating machinery. *Digit. Signal Process.* **2026**, *169*, 105682. [\[CrossRef\]](#)
21. Liu, S.; Cao, L.; Zhang, Y.; Li, Y. DBFE-Net: A dual-branch feature extraction network for Gaussian blind denoising. *Displays* **2025**, *90*, 103108. [\[CrossRef\]](#)
22. Dong, L.; Huo, Y. A New Open-Set Recognition Method for Fault Diagnosis of AUV. *Appl. Sci.* **2026**, *16*, 3526. [\[CrossRef\]](#)
23. Li, Y.; Ye, Y.; Zhang, Z.; Wen, L. A New Incremental Learning Method Based on Rainbow Memory for Fault Diagnosis of AUV. *Sensors* **2025**, *25*, 4539. [\[CrossRef\]](#) [\[PubMed\]](#)
24. Ji, D.; Yao, X.; Li, S.; Tang, Y.; Tian, Y. Autonomous underwater vehicle fault diagnosis dataset. *Data Brief.* **2021**, *39*, 107477. [\[CrossRef\]](#) [\[PubMed\]](#)
25. Huo, Y.; Gang, S.; Xiao, X.; Fu, B.; Sun, X. FATSNet: Transformer-based skip network with frequency attention for remote sensing image super-resolution. *Complex Intell. Syst.* **2026**, *12*, 131. [\[CrossRef\]](#)

Disclaimer/Publisher’s Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.