https://doi.org/10.37965/jdmd.2026.1796

Empirical Evaluation of Normalized Temperature-Scaled Cross-Entropy Pre-training in a Complete Ensemble Empirical Mode Decomposition with Adaptive Noise–Convolutional Neural Network–Bidirectional Long Short-Term Memory–Cross-Attention Framework for Unmanned Surface Vehicle Diesel Engine Fault Diagnosis

Yibo Ai,[1] Haobo Niu,[2] Hongtao Zhu,[2] Qiuyu Sun,[2] Yulin Guo,[2] Tao Xiong,[2] Hanxiao Xie,[2] Huangxiangyu Han,[2] Hengze Zhang,[2] Anbang Sun[2]*

（1. National Center for Materials Service Safety, University of Science and Technology Beijing, Beijing 100083, China;

2. State Key Laboratory of Electrical Insulation and Power Equipment, School of Electrical Engineering, Xi'an Jiaotong University, Xi'an 710049, China）

Received 3 August , 2026; Revised 15 October,2026; Accepted 26 October, 2026; Published online 3 November , 2026.

Abstract: Reliable diagnosis of unmanned surface vehicle (USV) diesel-engine faults requires representations that preserve multiscale vibration content and crank-angle-dependent context. Therefore, a CEEMDAN-CNN-BiLSTM-Cross-Attention framework is constructed, and NT-Xent pre-training is evaluated without presuming that it provides a performance benefit. CEEMDAN generates eight multiscale input channels, a CNN encodes local vibration transients, and the BiLSTM represents within-cycle context. The Cross-Attention uses CNN features as queries and BiLSTM features as keys and values. Six health states recorded on a laboratory diesel-engine platform at 1800 r/min and 80% load are used for experiments. The acquisition order defines the training, validation, and test subsets containing 2100, 450, and 450 cycles, respectively. Across five paired seeds, the Cross-Attention model with NT-Xent pre-training achieve 99.16 ± 0.40% accuracy and 99.15 ± 0.40% macro-F1 on clean test data. Five-fold acquisition-ordered block evaluation yield 99.07 ± 0.32% for both accuracy and macro-F1. Matched protocols, paired statistics, nested ablations, reduced-data tests, and additive-noise tests are used to separate overall framework performance from incremental component effects. NT-Xent pre-training increase clean-test accuracy by +0.09 percentage points relative to supervised training alone (95% CI; -0.95 to 1.13; p

* Corresponding author: Anbang Sun (e-mail: anbang.sun@xjtu.edu.cn).

https://doi.org/10.37965/jdmd.2026.1796

= 0.825). The Cross-Attention model scores 0.49 percentage points lower than the without-attention model (p = 0.207). Ordinary instance-level NT-Xent also performs worse at training-data fractions of 1%, 5%, and 10%. Accuracy is fell to 17.56 ± 1.14% at 12 dB, which is close to the chance level for six classes. The framework therefore accurately classifies clean test cycles under a fixed-operating condition , but neither the Cross-Attention model nor NT-Xent pre-training produces a reliable incremental benefit. These findings define the validated scope of the framework. Its performance under different operating conditions, across independent sessions and engines, and in real marine environments remains to be tested independently.

Keywords: USV; fault diagnosis; contrastive learning; CEEMDAN; CNN-BiLSTM-Cross-Attention; vibration signal.

1 Introduction

Unmanned surface vessels (USVs) perform waterborne missions under autonomous or remote control [1]. Their diesel engines determine manoeuverability, endurance, and mission availability [2]. Reliable diagnosis of engine health is therefore required to identify abnormal operating states before they compromise vessel operation.

Diesel-engine vibration signals contain local combustion-related transients at characteristic crank-angle positions. A useful representation must therefore preserve both short-duration patterns and their within-cycle context. High clean-data accuracy alone, however, does not establish robustness to limited training data, additive noise, or changing operating conditions [3,4]. Model evaluation must separate fixed-condition classification from broader generalization claims [5].

Data-driven methods are increasingly used to diagnose USV engine faults. Kowalski et al. [6] applied an extreme learning machine to marine diesel-engine diagnosis. Qu et al. [7] combined variational mode decomposition (VMD) with Long Short-Term Memory (LSTM)-based state prediction. Diez et al. [8] reduced sensor false alarms using a kernel support vector machine, whereas Yuan et al. [9] developed a multi-kernel extreme learning machine. Contrastive learning offers an alternative representation strategy when labelled or balanced data are limited. Wang et al. [10] studied nearest-neighbor self-supervision for few-shot marine-machinery diagnosis. Cho et al. [11] used supervised contrastive learning for multi-sensor engine signals. Wu et al. [12] investigated environment-aware diagnosis under class imbalance. These studies motivate rigorous testing of representation quality under disturbances, limited data, and changing conditions.

Vibration measurements provide information about combustion and moving-component states in diesel engines [13]. Ramteke et al. [14] used vibration and acoustic features to identify liner wear. Wang et al. [15] combined Ensemble Empirical Mode Decomposition (EEMD) features with

support vector machines for gearbox diagnosis. Pająk et al. [16] used polynomial-kernel classification of vibration and noise signals. Bi et al. [17] and Kuai et al. [18] combined adaptive signal decomposition with learning-based valve-clearance diagnosis. Peng et al. [19] incorporated data augmentation into an LSTM-GRU cascade. Zhao et al. [20] combined cylinder-head vibration, VMD, and random forests to identify six valve-fault states. Jafarian et al. [21] used multi-sensor vibration to detect misfire and valve-clearance faults. Bai et al. [22] retained diagnostic information after compressing diesel-engine vibration signals. Despite these advances, signal disturbance, operating-condition variation, and uncertain fault patterns continue to limit deployment.

Adaptive time-frequency analysis is widely used to denoise vibration signals and extract fault-related components. Zhu et al. [23] combined Complete Ensemble Empirical Mode Decomposition with Adaptive Noise (CEEMDAN), Detrended Fluctuation Analysis (DFA), Empirical Wavelet Transform (EWT) denoising, and a residual channel-attention classifier for ship-motor fault diagnosis. Ouelaa et al. [24] combined improved CEEMDAN, wavelet denoising, principal component analysis, and order tracking for nonstationary gear diagnosis. Such pipelines can improve signal quality, but preprocessing may also discard or distort discriminative information.

Existing studies motivate a diagnostic framework that connects multiscale signal content, local vibration transients, and within-cycle context. This study makes three contributions. First, it constructs a CEEMDAN-CNN-BiLSTM-Cross-Attention framework in which CNN-derived local features query BiLSTM-derived contextual features. This architecture implements joint representation without presuming that every component improves accuracy. Second, it establishes a paired and nested evaluation design that separates backbone performance from the effects of the BiLSTM branch, Cross-Attention, and NT-Xent pre-training. Third, it tests ordinary instance-level NT-Xent across full-data, reduced-data, and additive-noise settings to identify its empirical boundary in diesel-engine cycle classification.

The remainder of this paper is organized as follows. Section 2 describes CEEMDAN preprocessing, the CNN-BiLSTM-Cross-Attention framework, the two training protocols, and the experimental design. Section 3 reports matched architecture comparisons, block evaluation, nested ablations, reduced-data tests, additive-noise sensitivity, and computational cost. Section 4 summarizes the three principal findings and their applicability boundaries.

2 Methods and Experiments

2.1 Preprocessing of Vibration Signal Data

Cylinder-head vibration signals contain components at multiple characteristic scales.

https://doi.org/10.37965/jdmd.2026.1796

CEEMDAN was therefore used as a deterministic preprocessing stage before model training. This experiment evaluated the resulting representation under one fixed laboratory operating condition. The simulated additive-noise test was not assumed to reproduce the disturbance distribution of marine operations.

CEEMDAN [25] decomposed each engine-cycle signal into intrinsic mode functions (IMFs). As summarized in Fig. 1, the procedure added adaptive white-noise realizations, averaged the resulting IMFs, and updated the residual recursively. Let x denote the original signal, c_k denote the *k*-th IMF, and r_k denote the residual after k components. The initial residual is defined as:

$$r_0(t) = x(t) \tag{1}$$

The recursive relationship of residual extraction can be expressed as:

$$r_k(t) = r_{k-1}(t) - IMF_k(t) \tag{2}$$

where k represents the index of the extracted IMF component. During the decomposition process, each IMF component is obtained by averaging multiple realizations with adaptive white noise, which can be expressed as:

$$IMF_k(t) = \frac{1}{I}\sum_{i=1}^{I} IMF_{k,i}(t) \tag{3}$$

where I denotes the number of noise-assisted realizations, and $IMF_{k,i}(t)$ represents the *k*-th IMF component obtained from the *i*-th realization.

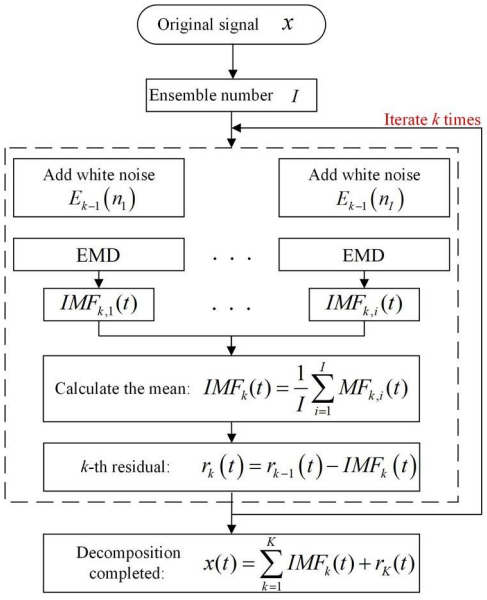


Fig. 1 Flowchart of the CEEMDAN Algorithm

https://doi.org/10.37965/jdmd.2026.1796

The decomposition process terminates when the residual signal no longer contains sufficient extrema for further decomposition. The original vibration signal can then be reconstructed as:

$$x(t) = \sum_{k=1}^{K} IMF_k(t) + r_K(t) \tag{4}$$

where K represents the total number of extracted IMF components and $r_K(t)$ denotes the final residual component.

CEEMDAN is designed to improve decomposition consistency and reduce mode mixing relative to EMD and EEMD. The retained IMF count was selected using validation macro-F1 alone. Candidate counts were 4, 6, 8, and 10. Eight IMFs yielded the highest validation mean and were fixed before formal model comparison. Test labels were not used in this selection.

2.2 CEEMDAN-CNN-BiLSTM-Cross-Attention Framework and Training Protocols

The framework in Fig. 2 contains CEEMDAN preprocessing, a CNN-BiLSTM-Cross-Attention encoder, an optional projection head, and a classifier. The projection head was used only for NT-Xent pre-training. The classifier was used during supervised fine-tuning and supervised-only training.

https://doi.org/10.37965/jdmd.2026.1796

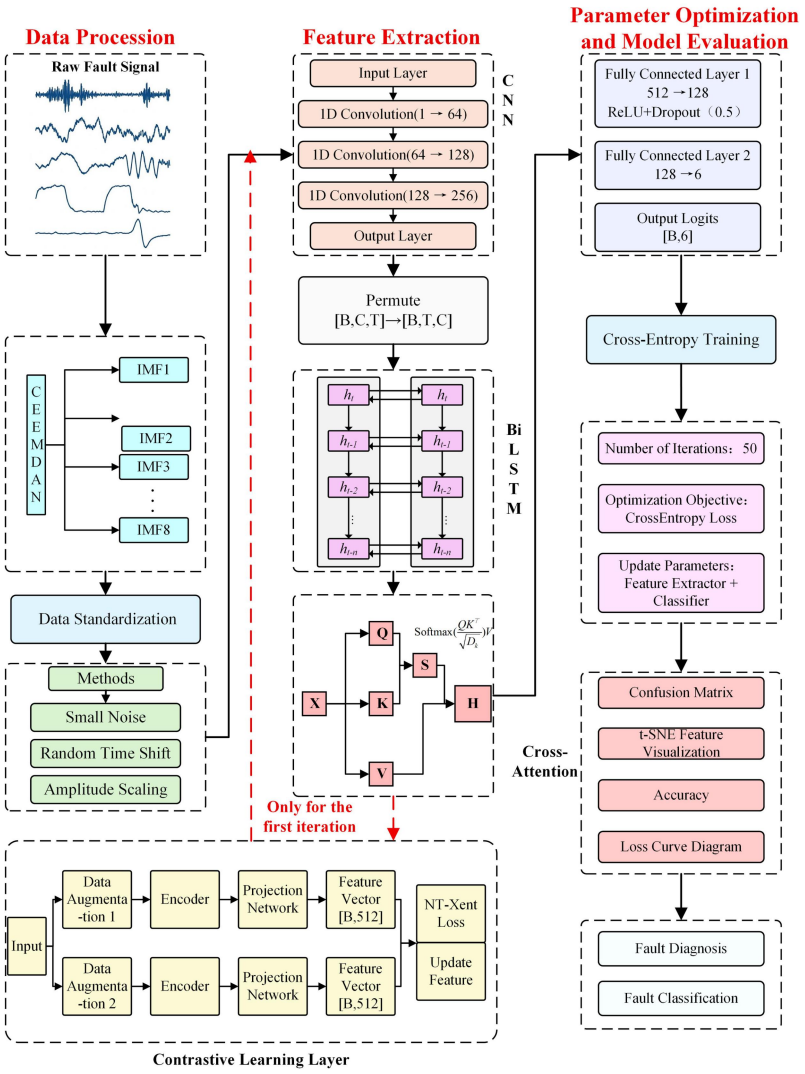


Fig. 2 Overall architecture and workflow of the proposed fault diagnosis framework

Acquisition order defined the training, validation, and test subsets before preprocessing. CEEMDAN was applied independently to each cycle, and StandardScaler statistics were fitted on training data only. For $X \in \mathbb{R}^{[B,8,7200]}$, the CNN produced $F_c \in \mathbb{R}^{[B\times900\times256]}$, and the one-layer BiLSTM produced F_t with the same shape. Cross-Attention used F_c as queries and F_t as keys and values. The attention output was added to F_c and normalized. Separate global pooling of F_c and the enhanced representation produced two 256-dimensional vectors, which were concatenated into a 512-dimensional encoding. Validation macro-F1 selected checkpoints and fixed configurations before test evaluation.

2.2.1 One-Dimensional Convolutional Neural Network (1D-CNN)

CEEMDAN transformed each vibration cycle into eight multiscale IMF channels. The CNN

https://doi.org/10.37965/jdmd.2026.1796

[26] then encoded local transients and short-duration patterns across these channels. This stage was intended to represent fault-related impacts without assigning them a longer-range interpretation.

Three stride-2 convolutional stages transformed the eight IMF channels into 900 local feature vectors, as shown in Fig. 3. These vectors entered the BiLSTM and also served as the query and residual sources in Cross-Attention. Nested ablations evaluated the incremental contribution of this design. CNN parameters were optimized jointly with the downstream modules during each applicable training stage.

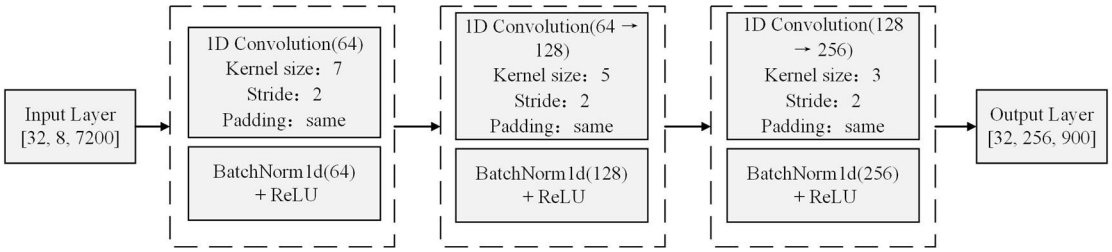


Fig. 3 Architecture of the CNN model

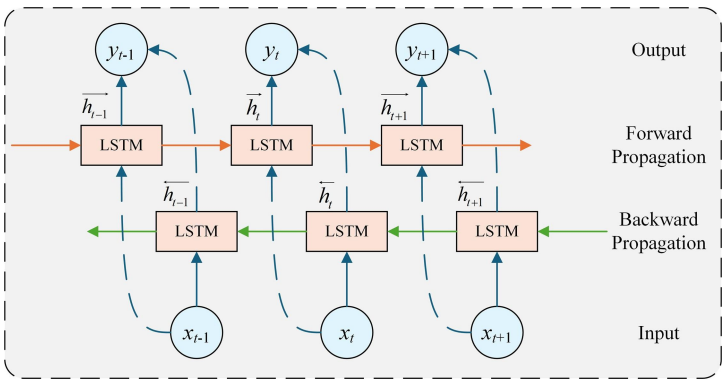
2.2.2 Temporal Feature Extraction Using BiLSTM

The CNN encoded local transients, but its output retained an ordered sequence of crank-angle positions within each 720-degree cycle. A bidirectional long short-term memory network (BiLSTM) [27] therefore represented preceding and following within-cycle context. This sequence did not describe degradation across multiple engine cycles.

A BiLSTM is composed of a forward LSTM and a backward LSTM, as shown in Fig. 4. Given the input feature x_t at time step t, the forward hidden state $\overrightarrow{h}_t$ encodes the historical information, while the backward hidden state $\overleftarrow{h}_t$ encodes the future information. The two hidden states are concatenated to obtain the final output y_t at time step t:

$$y_t=[\overrightarrow{h}_t \oplus \overleftarrow{h}_t] \tag{5}$$

where $\oplus$ denotes the feature concatenation operation.



https://doi.org/10.37965/jdmd.2026.1796

Fig. 4 Architecture of the BiLSTM model

The bidirectional representation supplied each crank-angle position with context from both directions within the same cycle. The nested comparison between the without-attention and CNN-only configurations evaluated its incremental contribution.

2.2.3 Cross-Attention Mechanism

The CNN and BiLSTM produced aligned sequence representations. F_t encoded local vibration patterns, whereas F_t encoded their within-cycle context. Both had shape $\mathbb{R}^{B\times 900\times 256}$, permitting scaled dot-product attention and a residual connection without additional temporal alignment.

$$F_c = CNN(X) \tag{6}$$

$$F_t = BiLSTM(F_c) \tag{7}$$

where X denotes the eight-channel IMF input, F_c denotes the CNN local-feature sequence, and F_t denotes the BiLSTM contextual-feature sequence.

A unidirectional Cross-Attention operation [28] conditioned local CNN features on BiLSTM context, as shown in Fig. 5. Its query and key sources, as well as its value source, differed from those of self-attention. The operation estimated which contextual positions in F_t were relevant to each local position in F_c. It was neither bidirectional nor mutual attention.

$$\mathbf{Q} = F_c W_Q \tag{8}$$

$$\mathbf{K} = F_t W_K \tag{9}$$

$$\mathbf{V} = F_t W_V \tag{10}$$

where W_Q, W_K and W_V are learnable projection matrices.

https://doi.org/10.37965/jdmd.2026.1796

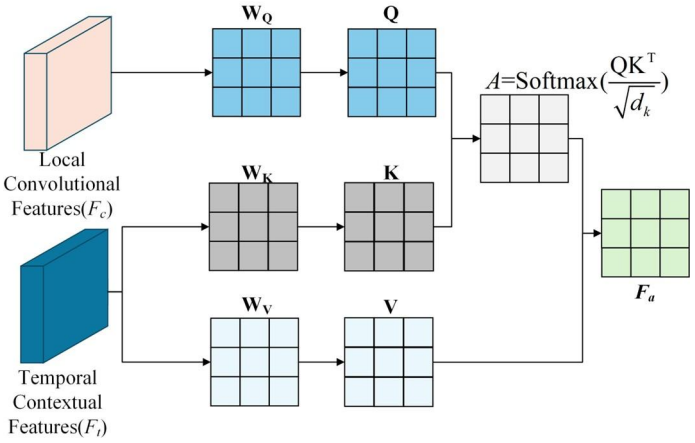


Fig. 5 Cross-attention mechanism algorithm flowchart

The attention weight matrix is calculated using the scaled dot-product attention mechanism:

$$A = \text{Softmax}\left(\frac{\mathbf{QK}^T}{\sqrt{d_k}}\right) \tag{11}$$

Let A denote the scaled dot-product attention output. The enhanced sequence was F_e = LayerNorm ($A + F_c$), with F_c as the residual source. Global average pooling was applied separately to F_c and F_e. Their pooled vectors were concatenated as h = Pool(F_c) ‖ Pool(F_e) in $\mathbb{R}^{512}$ and passed to either the projection head or classifier.

$$F_a = AV \tag{12}$$

2.2.4 Contrastive Learning Method

Fig. 6 illustrates the contrastive-learning branch of the framework. NT-Xent pre-training was evaluated as an optional initialization strategy for the encoder. Two independently augmented views of each training cycle formed a positive pair. Other views in the batch entered the denominator as negatives. The reduced-data and additive-noise experiments tested this instance-level objective without presuming improved sample efficiency or robustness [29].

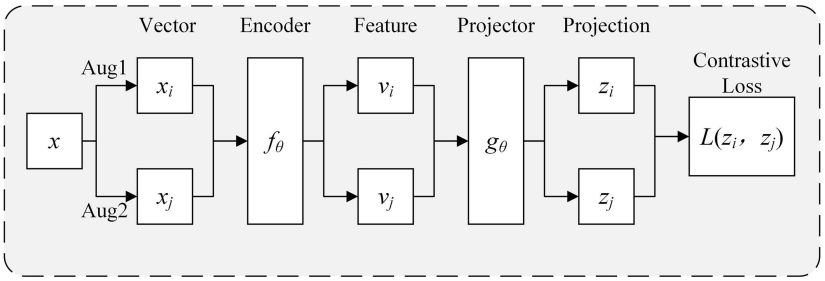


Fig. 6 Architecture of the proposed contrastive learning framework

https://doi.org/10.37965/jdmd.2026.1796

For each training signal x_i, Gaussian perturbation, temporal shifting, and amplitude scaling generated two independent views. A shared encoder *f* mapped both views to features. A 512-256-128 projection head g then produced the embeddings used by NT-Xent at $\tau = 0.5$:

$$\ell_{\text{NT-Xent}} = -\log \frac{\exp(sim(z_i, z_j) / \tau)}{\sum_{k=1, k \neq i}^{2N} \exp(sim(z_i, z_k) / \tau)} \tag{13}$$

where $sim(,)$ denotes the cosine similarity function, defined as:

$$\text{sim}(u, v) = \frac{u^{\tau} v}{\|u\| \cdot \|v\|} \tag{14}$$

where τ denotes the temperature, which was set to 0.5 in this study.

The two views of the same cycle formed the positive pair. With batch size 32, the remaining views were negatives. Cycles from the same health state could therefore enter the negative set and create class-level false negatives. Section 3 evaluates this mechanism against the reduced-data results. Matched supervised-only training provided the control required to determine whether NT-Xent produced a stable difference across paired random seeds.

2.3 Experimental Validation and Parameter Settings

2.3.1 Experimental Setup

Tab. 1 lists the valve-clearance settings for the six health states, and Fig. 7 shows the laboratory diesel-engine fault-simulation platform used to acquire the vibration data. The six states comprise four valve-clearance faults, one fuel-system misfire fault, and normal operation. Valve-clearance faults were created by adjusting the intake and exhaust clearances while maintaining fuel supply to all cylinders. For cylinder deactivation, the clearances remained at their normal values (0.35 and 0.55 mm), but fuel supply to one cylinder was cut off to simulate misfire.

Tab. 1 Experimental conditions and valve-clearance parameters for six health states

Health state	Intake valve clearance (mm)	Exhaust valve clearance (mm)
Excessively small	0.25	0.45
Slightly small	0.30	0.50
Normal	0.35	0.55
Cylinder deactivation	0.35	0.55
Slightly large	0.40	0.60
Excessively large	0.50	0.70

Vibration was acquired at 54 kHz. Each 720-degree four-stroke cycle was resampled on a 0.1-

https://doi.org/10.37965/jdmd.2026.1796

degree crank-angle grid from -360.0 to 359.9 degrees. This representation yielded 7200 points per model sample. The acquisition rate and angle-domain length describe different processing stages and do not imply a 3600-point model input [30,31].

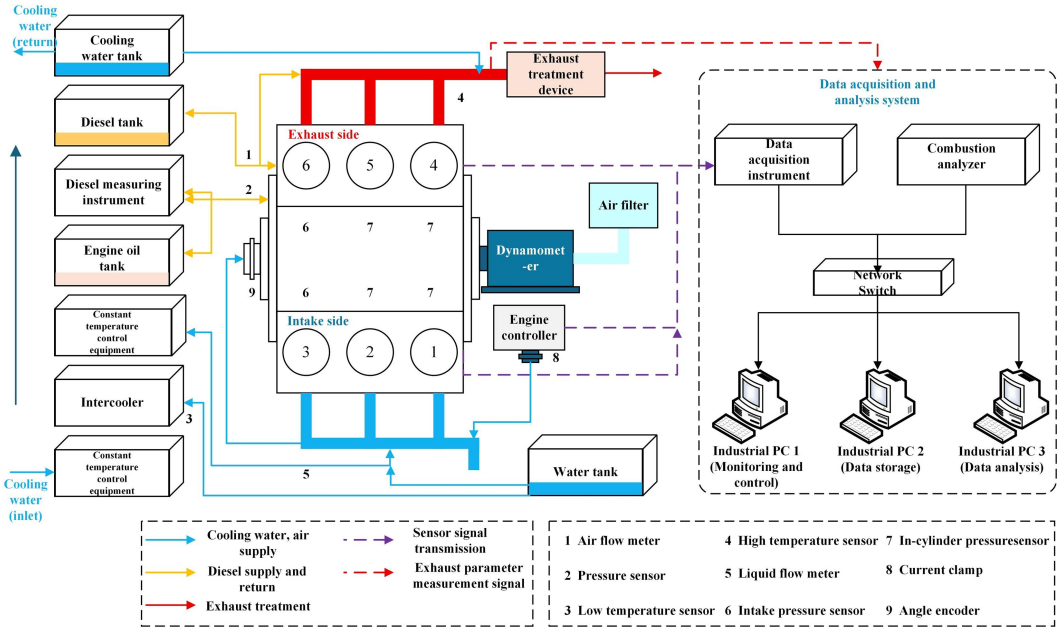


Fig. 7 Schematic diagram of the diesel engine fault simulation platform

Each health state contained 500 consecutive cycles, yielding 3000 samples in total. A cylinder-head sensor measured vibration along the Z direction. The first eight CEEMDAN IMFs formed an 8×7200 input tensor for each cycle. Acquisition order defined the data split before model selection. Within each health state, cycles 1-350 formed the training subset, cycles 351-425 the validation subset, and cycles 426-500 the test subset. The resulting sizes were 2100, 450, and 450 samples. StandardScaler statistics were fitted on the relevant training subset only. Test labels did not guide checkpoint or hyperparameter selection. Section 3.3 describes a separate five-fold acquisition-ordered block analysis.

2.3.2 Model Parameter Settings

Models were implemented in PyTorch and trained on a CUDA-enabled GPU. Random seeds-controlled initialization, dropout, DataLoader shuffling, stochastic augmentation, and reduced-data subset selection. The fixed acquisition-ordered split did not depend on a random seed. Main comparisons used paired seeds 42, 52, 62, 72, and 82. Reduced-data comparisons used ten paired seeds.

Two matched protocols were evaluated. The NT-Xent protocol used 10 pre-training epochs and 50 supervised fine-tuning epochs. The supervised-only protocol used the same encoder and

https://doi.org/10.37965/jdmd.2026.1796

classifier for 60 epochs. Both used batch size 32, AdamW, a 3×10^{-4} learning rate, 1×10^{-4} weight decay, 0.1 label smoothing, and validation macro-F1 for checkpoint selection. Enabled augmentations were applied to training signals only. Each protocol executed 3960 optimizer updates. NT-Xent processed 147000 encoder input views, whereas supervised-only training processed 126000 samples. Their measured training times were 198.3 ± 14.7 s and 188.2 ± 6.3 s, respectively. Peak GPU memory was 1.08 and 0.64 GiB on an NVIDIA GeForce RTX 5070 Ti Laptop GPU. Both models contained 803782 trainable parameters.

Tab. 2 specifies the operation, input tensor, and output tensor of each module in the final CEEMDAN-CNN-BiLSTM-Cross-Attention model. It makes the dimensional transformations from the eight-channel IMF input to the classifier and projection head explicit.

Tab. 2 Parameter settings of the modules in the proposed model

Module	Operation	Input	Output
CNN stage 1	Conv1d(8,64,*k*=7,*s*=2,*p*=3)	[B,8,7200]	[B,64,3600]
CNN stage 2	Conv1d(64,128,*k*=5,*s*=2,*p*=2)	[B,64,3600]	[B,128,1800]
CNN stage 3	Conv1d(128,256,*k*=3,*s*=2,*p*=1)	[B,128,1800]	[B,256,900]
Permutation	[B,C,T] to [B,T,C]	[B,256,900]	[B,900,256]
BiLSTM	1 layer, 128 units per direction	[B,900,256]	[B,900,256]
Cross-Attention	Q: CNN; K/V: BiLSTM	2×[B,900,256]	[B,900,256]
Residual + norm	LayerNorm(A + F_c)	[B,900,256]	[B,900,256]
Two-branch pooling	Pool F_c and enhanced features separately	2×[B,900,256]	2×[B,256]
Concatenation	Feature concatenation	2×[B,256]	[B,512]
Classifier	512 -> 128 -> 6	[B,512]	[B,6]
Projection head	512 -> 256 -> 128 (pre-training only)	[B,512]	[B,128]

3 Results and Discussion

3.1 Experimental Setup and Fair Comparison Protocol

The Cross-Attention model, BiLSTM, 1D-ResNet50, and 1D-Transformer used the same 8×7200 IMF input and acquisition-ordered split. Training-only standardization, augmentation operators, optimizer settings, learning rate, seed set, and checkpoint rule were also matched. Each

https://doi.org/10.37965/jdmd.2026.1796

encoder connected either to a 512-256-128 projection head for NT-Xent pre-training or to its supervised classifier. All NT-Xent runs used $\tau = 0.5$.

Each architecture was evaluated under both protocols with the same five paired seeds. NT-Xent used 10 pre-training and 50 fine-tuning epochs. Supervised-only training used 60 epochs. Hyperparameters were fixed before test evaluation, and validation macro-F1 selected the best checkpoint within each run. No random 10% validation holdout was introduced.

This design controlled the data, preprocessing, augmentation, projection dimension, optimization settings, total epoch count, paired seeds, and checkpoint criterion. Parameter counts and measured compute were reported separately because the architectures were not strictly matched for size or FLOPs. Test labels were used only after each configuration had been fixed.

3.2 Analysis of Experimental Results

Fig. 8 reports the mean optimization curves across the five Cross-Attention NT-Xent runs. Pre-training NT-Xent loss decreased over 10 epochs, while the training and validation cross-entropy losses decreased during the subsequent 50-epoch supervised fine-tuning stage. The shaded regions denote one sample standard deviation across seeds; the two stages are shown on separate axes because their objectives and numerical scales differ.

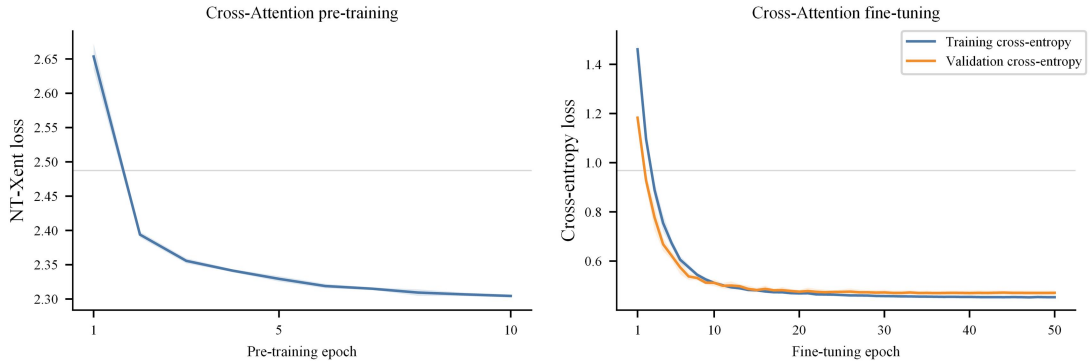


Fig. 8 Optimization curves for the Cross-Attention model with NT-Xent pre-training.

In Fig. 8, lines show five-seed means and shaded regions indicate one sample standard deviation. These curves describe optimization stability within the present acquisition-ordered split. Checkpoint selection used validation macro-F1; the curves are not evidence of cross-condition generalization.

Tab. 3 reports five-seed means and sample standard deviations under the matched protocols. The Cross-Attention model achieved the highest mean clean-test accuracy among NT-Xent configurations (99.16 ± 0.40%). It also exceeded the other supervised architectures under supervised-only training (99.07 ± 0.60%). These rankings describe complete architectures and do not isolate attention or initialization effects. Paired comparisons and nested ablations address those

questions. The baseline results were retained because they did not depend on the corrected attention implementation. The following analysis uses only the formal Cross-Attention reruns.

Tab. 3 Five-seed performance comparison among four architectures under contrastive and supervised protocols

Architecture	Protocol	Accuracy (%)	Precision (%)	Recall (%)	Macro-F1 (%)
BiLSTM	Contrastive	97.47 ± 0.58	97.46 ± 0.60	97.47 ± 0.58	97.44 ± 0.59
BiLSTM	Supervised	97.64 ± 0.49	97.64 ± 0.50	97.64 ± 0.49	97.63 ± 0.49
Proposed Cross	Contrastive	99.16 ± 0.40	99.17 ± 0.40	99.16 ± 0.40	99.15 ± 0.40
Proposed Cross	Supervised	99.07 ± 0.60	99.09 ± 0.57	99.07 ± 0.60	99.07 ± 0.59
ResNet50	Contrastive	97.73 ± 1.36	97.77 ± 1.36	97.73 ± 1.36	97.73 ± 1.36
ResNet50	Supervised	98.71 ± 0.69	98.74 ± 0.68	98.71 ± 0.69	98.71 ± 0.70
Transformer	Contrastive	98.49 ± 0.43	98.50 ± 0.44	98.49 ± 0.43	98.48 ± 0.44
Transformer	Supervised	98.67 ± 0.63	98.68 ± 0.63	98.67 ± 0.63	98.66 ± 0.63

With NT-Xent pre-training, the Cross-Attention model achieved 99.16 ± 0.40% accuracy and 99.15 ± 0.40% macro-F1. Supervised-only training achieved 99.07 ± 0.60% accuracy and 99.07 ± 0.59% macro-F1. The paired accuracy difference was +0.09 percentage points (95% CI, -0.95 to 1.13; paired t-test p = 0.825). The Wilcoxon p-value was 1.000, Cohen's dz was 0.106, and wins/ties/losses were 3/0/2. The interval included zero, and the effect size was small. Ordinary NT-Xent therefore provided no stable clean-data advantage.

Fig. 9 uses the predefined seed-42 runs for all four architectures. The Cross-Attention confusion matrix was strongly diagonal, and its t-SNE projection contained visually separated clusters. These selected-run diagnostics are qualitative and do not replace repeated paired statistics. The t-SNE projections used the same test samples and seed-specific encoders as the corresponding confusion matrices. They visualized feature geometry but could not establish statistical superiority, causal component effects, or generalization.

https://doi.org/10.37965/jdmd.2026.1796

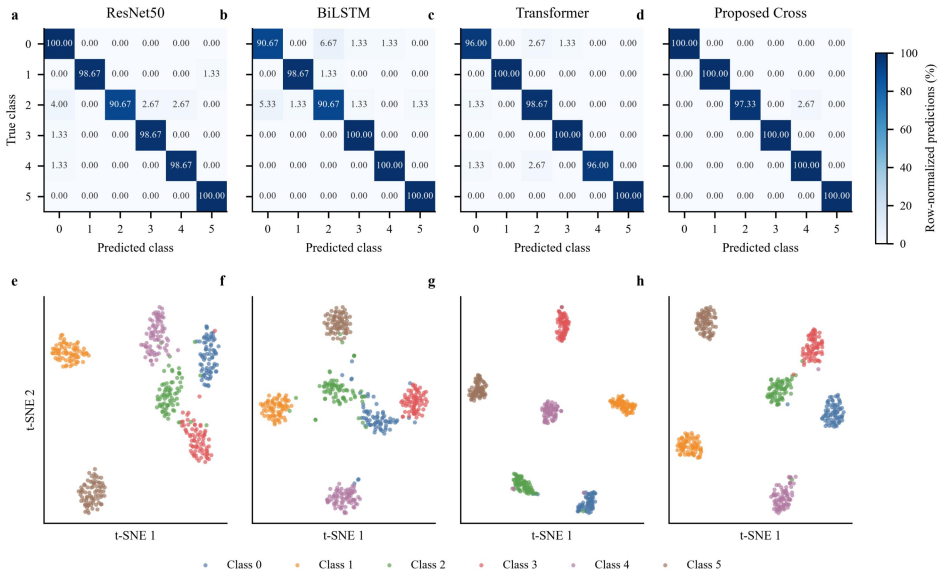


Fig. 9 Fixed-seed qualitative diagnostics for the four NT-Xent-pre-trained architectures.

3.3 Statistical Reliability and Data Independence Analysis

Adjacent cycles from one continuous acquisition may share local patterns. Five acquisition-ordered block folds were therefore defined before final analysis. This evaluation measured variation across held-out blocks from the same run. It was not Group Cross-Validation across independent engines, conditions, sensors, or sessions.

(1) Acquisition-ordered block assignment.

A sample-level random split could distribute adjacent cycles across subsets and yield an optimistic estimate. The block analysis therefore preserves acquisition order and uses contiguous held-out regions.

Within each health state, samples were divided into five contiguous blocks by acquisition order. For each fold, one block was used for testing, the next block for validation, and the remaining three blocks for training. Scaling statistics were fitted only on the training blocks.

Across five held-out blocks, the Cross-Attention NT-Xent model achieved 99.07 ± 0.32% accuracy and 99.07 ± 0.32% macro-F1. Macro-precision was 99.08 ± 0.32%, and macro-recall was 99.07 ± 0.32%. The 95% confidence interval for mean accuracy was 98.66-99.47%. Tab. 4 summarizes accuracy, precision, recall, and macro-F1. Here, within-run consistency is defined as the stability and repeatability of diagnostic performance across different contiguous cycle segments partitioned from a single continuous recording, as distinct from generalization across separate acquisition runs or operating domains.

Tab. 4 Five-fold within-run block-level consistency of the final Cross-Attention model with NT-Xent pre-training

https://doi.org/10.37965/jdmd.2026.1796

Validation scheme	Accuracy (%)	Precision (%)	Recall (%)	Macro-F1 (%)
Five-fold within-run blocks	99.07 ± 0.32	99.08 ± 0.32	99.07 ± 0.32	99.07 ± 0.32

(2) Interpretation.

The small fold-to-fold dispersion indicated consistent classification across contiguous blocks within this acquisition run. The training folds overlapped, and all blocks came from one engine, session, load, and speed. Consequently, the block-evaluation results confirm repeatable classification within the same recording run, but do not establish cross-run or independent-domain generalization.

The reported block results should not be interpreted as evidence of generalization to new loads, speeds, vessels, sensors, or acquisition sessions, which require independent-domain data.

3.4 Ablation Study

Tab. 5 uses nested ablations to separate three effects. Full Cross-Attention with NT-Xent was compared with matched supervised-only training to evaluate initialization. Full Cross-Attention was compared with the without-attention model to evaluate attention. The without-attention model was then compared with the CNN-only model to evaluate the added BiLSTM branch and its pooled representation. This sequence avoids attributing the simultaneous removal of BiLSTM and attention to BiLSTM alone.

Tab. 5 Five-seed nested ablation results

Configuration	Accuracy (%)	Precision (%)	Recall (%)	Macro-F1 (%)
Full Cross + NT-Xent	99.16 ± 0.40	99.17 ± 0.40	99.16 ± 0.40	99.15 ± 0.40
Full Cross + Supervised	99.07 ± 0.60	99.09 ± 0.57	99.07 ± 0.60	99.07 ± 0.59
Without attention	99.64 ± 0.34	99.65 ± 0.33	99.64 ± 0.34	99.64 ± 0.34
CNN only	99.51 ± 0.19	99.52 ± 0.19	99.51 ± 0.19	99.51 ± 0.19

The without-attention model achieved 99.64 ± 0.34% accuracy, whereas the CNN-only model achieved 99.51 ± 0.19%. The paired difference was +0.13 percentage points (95% CI, -0.24 to 0.50; paired t-test $p = 0.374$). The Wilcoxon p-value was 0.500, Cohen's dz was 0.447, and wins/ties/losses were 2/2/1. The contextual branch was therefore associated with a small numerical difference that was not statistically established.

Full Cross-Attention with NT-Xent achieved 99.16 ± 0.40% accuracy, 0.49 percentage points

below the without-attention model. The paired difference was -0.49 percentage points (95% CI, -1.39 to 0.41; paired t-test $p = 0.207$). The Wilcoxon p-value was 0.313, Cohen's dz was -0.673, and wins/ties/losses were 2/0/3. The present Cross-Attention operation therefore did not establish an accuracy improvement under the clean fixed operating condition.

Within the same Cross-Attention architecture, NT-Xent and supervised-only training differed by +0.09 percentage points in accuracy. The 95% CI ranged from -0.95 to 1.13, and the paired t-test gave $p = 0.825$. The Wilcoxon p-value was 1.000, Cohen's dz was 0.106, and wins/ties/losses were 3/0/2. This isolated comparison did not support a reliable NT-Xent gain.

3.5 Extended Ablation Study and Hyperparameter Sensitivity Analysis

Tab. 6 combines validation-only IMF-count selection with test-set ablations of CEEMDAN and the three augmentation operators. IMF count was selected from validation results. All test ablations were evaluated only after their configurations had been fixed. The full Cross-Attention NT-Xent model achieved 99.16% mean accuracy. Removing CEEMDAN reduced the mean to 98.89 ± 0.27%, a difference of -0.27 percentage points. Removing all random augmentation produced 98.80 ± 0.84%, a difference of -0.36 percentage points. These repeated-run differences describe clean-data rankings and do not establish causal robustness effects.

Tab. 6 IMF sensitivity and extended ablation results

Analysis	Configuration	Metric	Mean ± SD (%)
IMF validation	4 IMFs	Validation macro-F1	91.36 ± 2.96
IMF validation	6 IMFs	Validation macro-F1	97.11 ± 0.23
IMF validation	8 IMFs	Validation macro-F1	99.33 ± 0.39
IMF validation	10 IMFs	Validation macro-F1	97.92 ± 0.13
Test ablation	W/O CEEMDAN	Accuracy	98.89 ± 0.27
Test ablation	W/O random augmentation	Accuracy	98.80 ± 0.84
Test ablation	W/O Gaussian perturbation	Accuracy	98.93 ± 0.67
Test ablation	W/O temporal shift	Accuracy	99.29 ± 0.40
Test ablation	W/O amplitude scaling	Accuracy	98.76 ± 0.68

Removing Gaussian perturbation yielded 98.93 ± 0.67% accuracy. Removing temporal shift yielded 99.29 ± 0.40%, and removing amplitude scaling yielded 98.76 ± 0.68%. The corresponding differences from the full-model mean were -0.23, +0.13, and -0.40 percentage points. Temporal shift was therefore not uniformly beneficial in this experiment. Overlapping dispersions also precluded a firm ranking of augmentation operators.

Validation macro-F1 was 91.36 ± 2.96%, 97.11 ± 0.23%, 99.33 ± 0.39%, and 97.92 ± 0.13%

for 4, 6, 8, and 10 IMFs, respectively. Eight IMFs had the highest validation mean and were retained. These validation results were not compared directly with the locked-test ablations.

The formal Cross-Attention implementation used CNN features as queries and residuals, with BiLSTM features as keys and values. This direction was fixed before rerunning the experiment matrix. Self-attention and multi-head attention results were excluded from formal evidence for the proposed model. Test performance did not select the attention type.

Taken together, the ablations separated framework performance from incremental component effects. The complete framework accurately classified clean cycles under a fixed operating condition. However, the paired comparisons did not establish benefits from Cross-Attention, the BiLSTM contextual branch, or NT-Xent initialization. The pronounced clean-data ceiling limited component attribution. CEEMDAN and augmentation produced modest numerical changes, but their broader value needs to be assessed withvariable-load, variable-speed, cross-session, cross-engine, and field data.

3.6 Reduced-Data Performance and Additive-Noise Sensitivity

Reduced-data tests used 1%, 5%, 10%, and 100% of the 2,100-sample training subset. These fractions corresponded to 21, 105, 210, and 2,100 samples before augmentation. Each of the ten paired seeds used the same subset for both protocols. NT-Xent pre-training and fine-tuning were restricted to that subset. Validation and test sets remained unchanged.

Tab. 7 presents the ten-seed paired results for each training fraction and protocol. It is used to assess whether the effect of NT-Xent initialization changes when labeled training cycles are scarce.

Tab. 7 Ten-seed paired reduced-data results

Training fraction	Training samples	Protocol	Accuracy (%)	Precision (%)	Recall (%)	Macro-F1 (%)
1%	21	Contrastive	39.76 ± 4.72	42.87 ± 6.92	39.76 ± 4.72	35.04 ± 4.88
1%	21	Supervised	43.78 ± 4.35	44.33 ± 6.48	43.78 ± 4.35	39.55 ± 4.98
5%	105	Contrastive	67.29 ± 4.75	67.39 ± 4.75	67.29 ± 4.75	66.16 ± 4.97
5%	105	Supervised	75.47 ± 4.04	75.96 ± 3.88	75.47 ± 4.04	75.27 ± 4.09
10%	210	Contrastive	80.67 ± 4.65	80.95 ± 4.50	80.67 ± 4.65	80.50 ± 4.78

https://doi.org/10.37965/jdmd.2026.1796

Training fraction	Training samples	Protocol	Accuracy (%)	Precision (%)	Recall (%)	Macro-F1 (%)
10%	210	Supervised	85.20 ± 5.29	85.55 ± 5.19	85.20 ± 5.29	85.20 ± 5.32
100%	2100	Contrastive	98.96 ± 0.47	98.97 ± 0.46	98.96 ± 0.47	98.95 ± 0.47
100%	2100	Supervised	99.00 ± 0.59	99.02 ± 0.57	99.00 ± 0.59	99.00 ± 0.58

Tab. 8 reports five-seed results after additive white Gaussian noise was applied to the raw test signals and CEEMDAN was rerun. It distinguishes clean fixed-condition performance from sensitivity to the simulated noise perturbation.

Tab. 8 Five-seed additive white Gaussian noise results for the proposed model

Test condition	Protocol	Accuracy (%)	Precision (%)	Recall (%)	Macro-F1 (%)
Clean	Contrastive	99.16 ± 0.40	99.17 ± 0.40	99.16 ± 0.40	99.15 ± 0.40
Clean	Supervised	99.07 ± 0.60	99.09 ± 0.57	99.07 ± 0.60	99.07 ± 0.59
12 dB	Contrastive	17.56 ± 1.14	7.74 ± 6.90	17.56 ± 1.14	6.13 ± 1.68
12 dB	Supervised	17.87 ± 2.68	4.46 ± 3.76	17.87 ± 2.68	6.38 ± 3.61
8 dB	Contrastive	17.16 ± 1.09	7.28 ± 10.05	17.16 ± 1.09	5.69 ± 2.05
8 dB	Supervised	17.16 ± 1.09	3.44 ± 1.49	17.16 ± 1.09	5.66 ± 2.00
4 dB	Contrastive	16.71 ± 0.24	4.10 ± 1.86	16.71 ± 0.24	5.69 ± 1.49
4 dB	Supervised	16.67 ± 0.00	2.78 ± 0.00	16.67 ± 0.00	4.76 ± 0.00
0 dB	Contrastive	17.11 ± 0.99	3.48 ± 1.57	17.11 ± 0.99	5.67 ± 2.04
0 dB	Supervised	16.67 ± 0.00	2.78 ± 0.00	16.67 ± 0.00	4.76 ± 0.00
-4 dB	Contrastive	16.62 ± 0.10	3.33 ± 1.24	16.62 ± 0.10	5.46 ± 1.56
-4 dB	Supervised	16.67 ± 0.00	2.78 ± 0.00	16.67 ± 0.00	4.76 ± 0.00
-8 dB	Contrastive	16.31 ± 0.80	3.15 ± 0.84	16.31 ± 0.80	5.18 ± 0.93
-8 dB	Supervised	16.71 ± 0.10	3.13 ± 0.79	16.71 ± 0.10	4.93 ± 0.37

Ordinary NT-Xent did not improve sample efficiency. Relative to supervised-only training, accuracy was 4.02 percentage points lower at 1% (95% CI, -5.79 to -2.25; $p = 0.000617$). It was 8.18 percentage points lower at 5% (95% CI, -11.00 to -5.36; $p = 0.000104$). At 10%, it was 4.53 percentage points lower (95% CI, -8.20 to -0.86; $p = 0.0209$). At 100%, the difference was -0.04 percentage points (95% CI, -0.52 to 0.43; $p = 0.836$). Wilcoxon tests and effect sizes supported the same interpretation. Cycles from the same health state could be treated as negatives by the instance-

https://doi.org/10.37965/jdmd.2026.1796

level objective. These class-level false negatives may be especially damaging when few labeled cycles are available. Class-aware masking, supervised contrastive learning, and debiased objectives should be evaluated prospectively in future work.

For noise evaluation, additive white Gaussian noise was applied to raw test signals before CEEMDAN was rerun. Noise power followed $P_{\text{noise}} = P_{signal} / 10^{SNR/10}$. Five model seeds were paired one-to-one with five noise seeds, and all training data remained clean. For NT-Xent, accuracy/macro-F1 fell from 99.16/99.15% on clean data to 17.56/6.13% at 12 dB. Values remained near chance at all lower SNRs.

Both protocols approached chance-level accuracy at 12 dB, indicating sensitivity rather than robustness to additive noise. Macro-F1 remained substantially lower because predictions concentrated in a subset of classes. Gaussian noise does not reproduce wave impact, hull vibration, propeller excitation, speed variation, sensor remounting, or domain shift. This test therefore identified a failure boundary but did not establish real-marine performance.

Fig. 10 compares accuracy and macro-F1 across the four training fractions for the matched NT-Xent and supervised-only protocols. The error bars represent variation across paired seeds and show whether the difference depends on data availability.

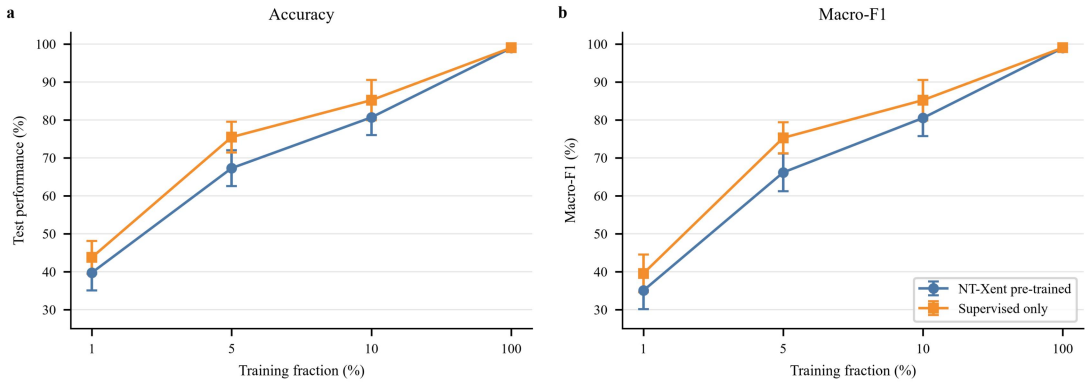


Fig. 10 Reduced-data comparison of NT-Xent pre-training

Fig. 11 reports accuracy and macro-F1 when additive white Gaussian noise is applied to the raw test signals. It shows whether the clean-trained model retains discriminative performance as the signal-to-noise ratio decreases.

https://doi.org/10.37965/jdmd.2026.1796

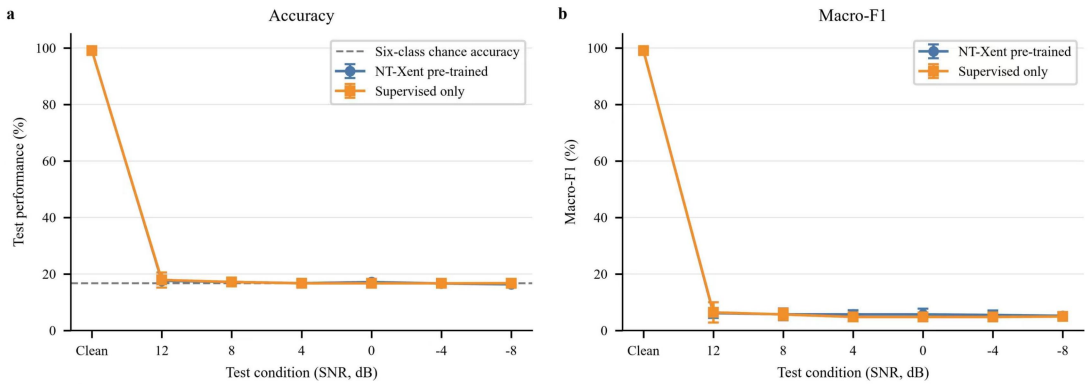


Fig. 11 Additive-noise sensitivity of the proposed Cross-Attention architecture trained on clean data.

3.7 Training Process and Pre-training Stage Description

The NT-Xent protocol contained a self-supervised pre-training stage followed by supervised fine-tuning.

(1) Self-supervised pre-training stage

During the first 10 epochs, the encoder and projection head were optimized with NT-Xent using augmented training views. The classifier was not used during this stage.

(2) Supervised fine-tuning stage

After pre-training, the projection head was removed. The encoder and classifier were then optimized jointly with cross-entropy for 50 epochs. Validation macro-F1 selected the best checkpoint.

The two-stage protocol provided an alternative initialization to 60-epoch supervised-only training. Both protocols executed 3,960 optimizer updates and used the same 803,782-parameter encoder/classifier. NT-Xent processed 147,000 encoder input views, compared with 126,000 supervised samples. It required 198.3 ± 14.7 s and 1.08 GiB peak memory, compared with 188.2 ± 6.3 s and 0.64 GiB. Equal epoch and update counts therefore did not imply equal computation. The additional cost was not accompanied by a stable accuracy benefit.

4 Conclusions

This study constructed a CEEMDAN-CNN-BiLSTM-Cross-Attention framework and evaluated ordinary instance-level NT-Xent for six diesel-engine health states under one fixed operating condition. Three main conclusions are listed as follows.

(1) The framework integrated multiscale CEEMDAN components, CNN-derived local transients, and BiLSTM-derived within-cycle context through unidirectional Cross-Attention. The

configuration with eight retained IMFs achieved the highest validation macro-F1 (99.33 ± 0.39%). The complete framework with NT-Xent pre-training achieved 99.16 ± 0.40% accuracy and 99.15 ± 0.40% macro-F1 on clean test data. A five-fold block evaluation yielded 99.07 ± 0.32% for both metrics, supporting consistency within the acquisition run.

(2) Paired seeds and nested ablations were used to separate overall framework performance from incremental component effects. The Cross-Attention model scored 0.49 percentage points lower than the without-attention model (p = 0.207). The without-attention model scored 0.13 percentage points higher than the CNN-only model (p = 0.374). Neither comparison established a reliable component benefit under the clean fixed operating condition.

(3) Ordinary NT-Xent increased clean-test accuracy by +0.09 percentage points relative to supervised training alone (p = 0.825). It performed significantly worse at training-data fractions of 1%, 5%, and 10%, and showed no detectable difference at 100%. Class-level false negatives provide a plausible explanation for this reduced-data failure. The accuracy of the clean-trained model also fell to 17.56% at 12 dB and remained near chance at lower SNRs.

Overall, the scope of this work is limited to feasibility validation under fixed laboratory operating conditions, as all experimental data were collected from one engine and a single acquisition session at a constant speed and load. Confusion matrices and t-SNE projections remain qualitative clean-data diagnostics. Generalization to different operating conditions, independent sessions, other engines, and real marine environments remains to be validated in future studies.

ACKNOWLEDGMENTS

The authors sincerely acknowledge the support of the Opening Project Fund of Materials Service Safety Assessment Facilities (Grant No. MSAF-2024-113).

CONFLICT OF INTEREST STATEMENT

The authors declare that they have no conflicts of interest related to this work.

REFERENCES

[1]. BARRERA C, PADRON I, LUIS F S, et al. Trends and challenges in unmanned surface vehicles (USV): From survey to shipping[J]. TransNav: International Journal on Marine Navigation and Safety of Sea Transportation, 2021, 15(1): 123-130.

[2]. VARBANETS R, KARIANSKIY S. Analyse of marine diesel engine performance[J]. Journal of Polish CIMAC, 2012, 7(1): 269-276.

[3]. WRÓBEL K, MONTEWKA J, KUJALA P. Towards the assessment of potential impact of unmanned vessels on maritime transportation safety[J]. Reliability Engineering & System Safety, 2017, 165: 155-169.

[4]. MISHRA C S, ALI F, ADAM S, et al. Study of the causes of engine failure in ship and maintenance required to prevent engine failure[J]. International Journal of Contemporary Research and Review, 2017, 8(7): 20218-20241.

[5]. OU S, YU Y, YANG J. Identification and reconstruction of anomalous sensing data for combustion analysis of marine diesel engines[J]. Measurement, 2022, 193: 110960.

[6]. KOWALSKI J, KRAWCZYK B, WOŹNIAK M. Fault diagnosis of marine 4-stroke diesel engines using a one-vs-one extreme learning ensemble[J]. Engineering Applications of Artificial Intelligence, 2017, 57: 134-141.

[7]. QU C, ZHOU Z, LIU Z, et al. State prediction for marine diesel engine based on variational modal decomposition and long short-term memory[J]. Energy Reports, 2021, 7: 880-886.

[8]. DIEZ-OLIVAN A, PAGAN J A, KHOA N L D, et al. Kernel-based support vector machines for automated health status assessment in monitoring sensor data[J]. The International Journal of Advanced Manufacturing Technology, 2018, 95(1): 327-340.

[9]. YUAN Y, YAN X, WANG K, et al. A new remote intelligent diagnosis system for marine diesel engines based on an improved multi-kernel algorithm[J]. Proceedings of the Institution of Mechanical Engineers, Part O: Journal of Risk and Reliability, 2015, 229(6): 604-611.

[10]. WANG R, CHEN H, GUAN C. A self-supervised contrastive learning framework with the nearest neighbors matching for the fault diagnosis of marine machinery[J]. Ocean Engineering, 2023, 270: 113437.

[11]. CHO Y S, KIM S B. Supervised contrastive learning for multisensor signals classification in automobile engine manufacturing[J]. IEEE Transactions on Industrial Informatics, 2024, 20(5): 7764-7776.

[12]. WU H, ZHONG S, ZHAO M, et al. Continual contrastive reinforcement learning: Towards stronger agent for environment-aware fault diagnosis of aero-engines through long-term optimization under highly imbalance scenarios[J]. Advanced Engineering Informatics, 2025, 65: 103297.

[13]. YOUSSEF A, NOURA H, EL AMRANI A, et al. A survey on data-driven fault diagnostic techniques for marine diesel engines[J]. IFAC-PapersOnLine, 2024, 58(4): 55-60.

[14]. RAMTEKE S M, CHELLADURAI H, AMARNATH M. Diagnosis of liner scuffing fault of a diesel engine via vibration and acoustic emission analysis[J]. Journal of Vibration Engineering & Technologies, 2020, 8(6): 815-833.

[15]. WANG A, LI Y, DU X, et al. Diesel engine gearbox fault diagnosis based on multi-features extracted from vibration signals[J]. IFAC-PapersOnLine, 2021, 54(10): 33-38.

[16]. PAJĄK M, KLUCZYK M, MUŚLEWSKI Ł, et al. Ship diesel engine fault diagnosis using data science and machine learning[J]. Electronics, 2023, 12(18): 3860.

[17]. BI X, CAO S, ZHANG D. Diesel engine valve clearance fault diagnosis based on improved variational mode decomposition and bispectrum[J]. Energies, 2019, 12(4): 661.

[18]. KUAI Z, HUANG G. Fault diagnosis of diesel engine valve clearance based on wavelet packet decomposition and neural networks[J]. Electronics, 2023, 12(2): 353.

[19]. PENG Z, ZHANG W, LIU R. Research on diesel engine fault prediction method based on deep

learning and vibration information[J]. Journal of the Brazilian Society of Mechanical Sciences and Engineering, 2025, 47(10): 526.

[20]. ZHAO N, MAO Z, WEI D, et al. Fault diagnosis of diesel engine valve clearance based on variational mode decomposition and random forest[J]. Applied Sciences, 2020, 10(3): 1124.

[21]. JAFARIAN K, MOBIN M, JAFARI-MARANDI R, et al. Misfire and valve clearance faults detection in the combustion engines based on a multi-sensor vibration signal monitoring[J]. Measurement, 2018, 128: 527-536.

[22]. BAI H, WEN L, MA Y, et al. Compression reconstruction and fault diagnosis of diesel engine vibration signal based on optimizing block sparse Bayesian learning[J]. Sensors, 2022, 22(10): 3884.

[23]. ZHU R, SONG E, YAO C, et al. Marine motor fault diagnosis based on CEEMDAN and BRECAN under strong noise conditions[J]. Chinese Journal of Ship Research, 2025, 20(2): 20-29.

[24]. OUELAA Z, DJEBALA A, YOUNES R, et al. Advanced gear fault diagnosis in non-stationary conditions with an improved CEEMDAN-wavelet denoising technique[J]. Advances in Mechanical Engineering, 2025, 17(7): 16878132251356546.

[25]. TORRES M E, COLOMINAS M A, SCHLOTTHAUER G, et al. A complete ensemble empirical mode decomposition with adaptive noise[C]//2011 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). Prague: IEEE, 2011: 4144-4147.

[26]. Wang, D., Guo, Q., Song, Y., Gao, S., & Li, Y. (2019). Application of multiscale learning neural network based on CNN in bearing fault diagnosis. Journal of Signal Processing Systems, 91(10), 1205-1217.

[27]. Bai, J., Zhu, W., Liu, S., Ye, C., Zheng, P., & Wang, X. (2025). A temporal convolutional network–bidirectional long short-term memory (TCN-BiLSTM) prediction model for temporal faults in industrial equipment. Applied Sciences, 15(4), 1702.

[28]. Jin, Z., Wu, T., Chen, T., Pan, D., Wang, X., Xie, J., ... & Lyu, Q. (2023). CAPLA: improved prediction of protein–ligand binding affinity by a deep learning approach based on a cross-attention mechanism. Bioinformatics, 39(2), btad049.

[29]. Chen, Li, Lingying Chen, and Xiaoqiang Xie. "A fault detection method for power waste material dismantling production line based on SimCLR algorithm." International Conference on Power Electronics Technology and Grid Systems (PETGS 2026). Vol. 14189. SPIE, 2026.

[30]. Ricci, L., Perinelli, A., & Prevedelli, M. (2024). Nyquist–Shannon sampling theorem. In The Physics Behind Electronics (pp. 199-212). Cham: Springer Nature Switzerland.

[31]. Randall, R. B. (2011). Vibration-based condition monitoring: Industrial, aerospace and automotive applications (pp. 13-20). John Wiley & Sons.