aerospace

MDPI

Article

Temporal-Enhanced GAN-Based Few-Shot Fault Data Augmentation and Intelligent Diagnosis for Liquid Rocket Engines

Hui Hu [1], Rongheng Zhao [2], Chaoyue Xu [2], Shuai Ren [2,*] and Hui Wang [1,*]

[1] Beijing Aerospace Propulsion Institute, Beijing 100076, China; hh13141291049@163.com
[2] School of Automation, Beijing Institute of Technology, Beijing 100081, China; zhaorongheng@bit.edu.cn (R.Z.); xuchaoyue@bit.edu.cn (C.X.)
* Correspondence: renshuai@bit.edu.cn (S.R.); whitneywxl@sina.com (H.W.)

Abstract

(1) Background: The scarcity and imbalance of real fault data significantly limit the development of data-driven fault diagnosis methods for liquid rocket engines (LREs), especially under few-shot conditions. (2) Methods: To address this issue, this study proposes a GAN-based fault data augmentation framework for multivariate LRE time-series signals and a hybrid diagnostic classifier combining convolutional neural networks (CNNs), bidirectional long short-term memory (BiLSTM), and multi-head attention (MHA). The GAN component is introduced to alleviate fault-data scarcity and class imbalance by generating additional fault-like samples, while the classifier is designed to capture local features, long-range temporal dependencies, and diagnostically informative temporal regions. (3) Results: A multidimensional evaluation based on temporal similarity, statistical consistency, and global distribution discrepancy indicates that the generated samples preserve important characteristics of the original signals under the current evaluation protocol. On the augmented LRE dataset, the proposed classifier achieved strong diagnostic performance. In addition, supplementary experiments on the public HIT aero-engine dataset further support the effectiveness of the classifier architecture, its component-wise contribution, and its behavior under imbalanced few-shot settings, while also demonstrating the value of uncertainty-aware prediction. (4) Conclusions: The results provide encouraging evidence that the proposed framework can improve LRE fault diagnosis under data-scarce conditions. However, the present findings should be interpreted within the scope of the available data and evaluation setting. More comprehensive generator-side ablation, broader external validation, and physics-oriented assessment of the generated signals are still needed before stronger conclusions can be made.

Keywords: liquid rocket engine (LRE); generative adversarial network (GAN); few-shot learning; temporal dynamics; multi-head attention

Check for updates

Academic Editors: Stephen Whitmore and Jae Hyun Park

Received: 7 January 2026
Revised: 20 March 2026
Accepted: 23 March 2026
Published: 25 March 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) license.

1. Introduction

With advances in global aerospace technology, reusable launch vehicles (RLVs) have revolutionized space exploration by enabling cost-effective missions. As the core and irreplaceable propulsion systems, liquid rocket engines (LREs) are among the most critical components of RLVs, significantly enhancing system reliability and safety. This technology holds substantial engineering value and has emerged as a key development focus for future aerospace endeavors [1–4]. However, due to the high-pressure conditions and complex

operational environments of its internal components, the rocket engine system (referring to LRE) has become one of the most failure-prone subsystems in launch vehicles [5]. Consequently, real-time anomaly detection and fault diagnosis techniques are essential for ensuring the safe operation of LRE.

In recent years, the rapid progress of artificial intelligence (AI) and deep learning technologies has led to their widespread adoption in fault diagnosis, decision support, anomaly detection, and failure prediction [6–8]. Intelligent fault diagnosis has garnered increasing attention. Traditional LRE fault diagnosis primarily relies on physical or expert models. However, these approaches often struggle to achieve accurate diagnostics under complex and dynamic operating conditions, failing to capture subtle fault signatures. In contrast, deep learning techniques can effectively diagnose and classify fault types by analyzing multidimensional sensor data from LRE, demonstrating strong potential to ensure operational safety. However, due to LRE's operational constraints, each new testing phase prevents the engine from operating until a fault occurs, making fault events extremely rare and difficult to reproduce during flight. Consequently, historical fault data are vastly outnumbered by normal operational data, resulting in severe class imbalance. This data scarcity significantly impedes the application of deep learning in LRE fault diagnosis, particularly in detecting minor or emerging faults, where intelligent diagnostic capabilities remain critically inadequate [9]. Although many methods have been developed to address this issue [10], such as embedding physical knowledge into deep neural networks [11], only data augmentation can more intuitively improve model training.

Therefore, developing appropriate data augmentation techniques is important for alleviating the shortage of LRE fault data and improving fault diagnosis under data-scarce conditions. In this context, this paper proposes a generative adversarial network (GAN)-based framework to generate additional fault-like samples with improved class balance. The generated data are used to support the training of a deep learning-based fault diagnosis model and to enhance classification performance under few-shot conditions. The proposed framework is intended to provide a data-driven basis for improving LRE fault diagnosis, rather than to claim complete fault reproduction or direct engineering deployment.

2. Related Work

In recent years, machine learning and deep learning technologies have demonstrated promising applications in LRE fault diagnosis due to their strong adaptive and nonlinear mapping capabilities. However, due to the complex structure of LRE, harsh operating conditions, scarcity of fault samples, and the high requirements for autonomy and real-time performance in fault identification, existing diagnostic techniques still face numerous challenges. Based on their methodological principles, current research on LRE fault diagnosis can be categorized into three approaches: model-driven, knowledge-driven, and data-driven. Among these, model-based fault diagnosis methods use mathematical and physical models of the engine to estimate relevant parameters [12]. These estimated parameters are then compared with the engine's measured parameters, and a residual analysis is performed to identify faults [13]. For instance, Maru et al. [14] combined autoregressive moving-average models with extended Kalman filters to develop a residual-based method based on the Mahalanobis distance, enabling fault detection in LRE. Such methods offer clear physical interpretability and can provide a theoretical basis for fault localization. However, their diagnostic performance heavily relies on the accuracy of the engine model, and establishing precise physical models is often challenging in practice. Kanso et al. [15] employed extended Kalman filtering to estimate the health status of the combustion chamber, providing support for lifetime prediction. Nevertheless, the generalizability of such models is relatively poor, making them challenging to apply to different systems. Expert

systems represent another category of traditional methods and are typical examples of knowledge-driven fault diagnosis approaches. These systems simulate the decision-making capabilities of human experts using intelligent computer programs to address complex fault-diagnosis problems. For instance, Moonis [16] developed a fault-diagnosis expert system, while Gupta et al. [17] designed the LEADER system for the space shuttle main engine, both aiming to achieve real-time diagnosis through rules and knowledge bases. However, these systems heavily depend on expert experience and historical rules, making it difficult to handle new or unknown fault types so their scalability and adaptability remain limited.

With the advent of the digital, informational, and intelligent era, data-driven intelligent fault diagnosis methods, typified by machine learning, have become the mainstream approach in current fault diagnosis research. These methods do not rely on precise physical models or expert knowledge but instead directly leverage observable data from the engine system. They employ data-processing technologies for fault detection and fault mode identification, demonstrating strong adaptive capabilities. For example, Tsutsumi et al. [18] used principal component analysis and dynamic time warping to achieve LRE fault detection on a two-dimensional phase plane. Park et al. [8] constructed fault datasets using numerical simulations and achieved multi-category fault identification based on deep learning models. Flora et al. [19] designed an artificial neural network-based algorithm for sensor fault isolation and replacement, thereby enhancing system fault tolerance. Although these data-driven methods have made significant progress, fault samples for LREs are extremely scarce in practical engineering applications. This leads to insufficient model training and poor generalization, particularly in small-sample settings where the diagnostic performance of traditional deep learning methods is severely limited. Therefore, generating fault-like data that preserve important temporal and statistical characteristics from limited samples has become an important issue for enhancing LRE fault diagnosis capability.

Generative adversarial networks (GANs), introduced by Goodfellow et al. [20], have emerged as powerful data-generation tools and have been widely applied in recent years in image synthesis and time-series data augmentation. They also offer a novel approach to addressing data scarcity in LRE fault diagnosis. Through adversarial training of the generator and discriminator, GANs can synthesize samples that closely match the distribution of real data, demonstrating excellent generative performance across various domains. For instance, Yoon et al. [21] proposed the TimeGAN framework, which integrates unsupervised adversarial training with autoregressive models to preserve the dynamic characteristics of time-series data effectively. In [22], an auxiliary classifier GAN was introduced, combined with a meta-learning strategy, to enhance the quality of fault data generation under small-sample conditions. Furthermore, in scenarios with severe data imbalance, few-shot GAN architectures have emerged as robust solutions to prevent overfitting when training with limited samples. For example, Ren et al. [23] developed a few-shot learning strategy for GANs that utilizes sample-rich classes to guide the distribution learning of sample-poor fault classes. This approach justifies the use of generative models to expand tiny datasets into balanced sets suitable for deep learning-based diagnosis. Recent research specifically targeting LREs has also demonstrated the effectiveness of integrating temporal modeling with generative frameworks. For instance, Deng et al. [24] proposed an FDD method for large LOX/kerosene rocket engines based on an LSTM-GAN architecture, which uses an LSTM to capture the dynamic characteristics of the engine's steady-state and startup processes. Their work highlights the potential of GANs in enhancing diagnostic reliability by simulating engine operating states. Xu et al. [25] utilized a combination of GAN and LSTM to achieve accurate pipeline leak prediction. Furthermore, GANs have been widely applied in the aerospace field to generate fault samples to augment datasets and improve model

training. Cheng et al. [26] proposed a GAN framework based on Wasserstein distance, enabling high-precision LRE fault detection by simulating the distribution of normal data. However, instability during GAN training remains a key issue that affects the quality of generated data. Enhancing the authenticity and diversity of the generated data remains a significant focus of current research.

While the aforementioned approaches, such as transfer learning from physical models [27], provide valuable pathways for leveraging external knowledge, they often introduce dependency on complex simulations or source-domain data. In contrast, directly generating fault-like samples that preserve key temporal and statistical characteristics from a limited number of real events presents a more agile and data-centric solution. This study therefore focuses on developing a temporal-enhanced GAN framework coupled with a hybrid deep classifier, aiming to address the core challenge of data scarcity in LRE fault diagnosis through end-to-end, few-shot data augmentation, and intelligent pattern recognition.

In summary, while various methods have sought to address the challenges of data insufficiency and model generalization in LRE fault diagnosis, GAN-based data augmentation techniques offer unique advantages due to their intuitive effectiveness and strong scalability. Building on this foundation, this paper proposes an enhanced GAN framework that incorporates temporal modeling mechanisms to generate LRE fault samples that are empirically consistent with important temporal and statistical characteristics, thereby providing data support for subsequent intelligent diagnosis models. In the present study, the role of the GAN component is to alleviate fault-data scarcity and class imbalance by generating additional fault-like samples, rather than to claim full physical equivalence between generated and real fault signals.

3. Materials and Methods

3.1. Data Collection and Preprocessing

Based on the test data from an LRE test bench, this study constructed a multi-channel time-series dataset comprising normal conditions (class0) and four typical fault states: ball bearing failure (class1), fatigue fracture of the turbine rotor disk-shaft joint (class2), oxygen turbopump ablation (class3), and pipeline fatigue fracture (class4). The selection of these four fault types was based on domain expert knowledge and historical failure cases, as they represent typical and frequently encountered failure modes during actual LRE operation. This dataset was used for training and validating the proposed fault diagnosis model. The data were collected from over 20 sensors distributed across key locations, including the combustion chamber and the turbopump, with sampling rates of either 1 kHz or 100 Hz, and a total duration of approximately 500 s. During the construction of the final dataset for training and generation, based on expert knowledge and fault-relevance analysis, we selected 14 signal channels strongly correlated with fault evolution from the original set of over 20 sensors, excluding some channels that monitored redundant environmental parameters or remained static throughout the tests. This feature selection process aims to reduce noise interference and computational complexity while retaining the most representative physical information for fault diagnosis.

During the data preprocessing stage, the raw engine time-series data were first subjected to outlier removal and cleaning, followed by signal smoothing. Specifically, targeted Gaussian smoothing filtering was applied to the multidimensional sensor signals, with different smoothing intensities applied to different parameters. This operation effectively suppressed high-frequency noise and improved signal stability. Next, to ensure consistent sampling lengths across dimensions such as engine temperature and pressure, a time-window-based truncation strategy was employed. Using the engine emergency shutdown

time as the reference, data segments 10 s before shutdown and 5 s after were truncated. This approach not only captured the pre-fault evolution process and post-fault characteristics but also standardized the time length of all sample data. Subsequently, standardization was performed to eliminate dimensional differences among the engine's physical parameters. The standardized data exhibited a uniform distribution, reducing the model's dependence on parameter magnitudes and thereby enhancing training stability. Furthermore, inverse standardization was applied to the data to ensure the generated data could be mapped back to real physical dimensions.

Through the aforementioned preprocessing steps, a high-quality, representative dataset was constructed to support the generation of small-sample fault data and fault classification for LREs. For each engine test, a fixed-length segment centered on the emergency shutdown event was extracted. Non-overlapping sliding windows were employed, ensuring that each sample corresponded to an independent test event. The original dataset comprised 407 samples, including 325 normal samples and 4 fault classes with 14, 27, 8, and 33 samples, respectively, resulting in pronounced class imbalance.

To address training instability and insufficient fault coverage under limited sample conditions, a two-stage data augmentation strategy was adopted. Initially, simulation-augmented data from domain experts at a collaborating institution were used to improve class balance and stabilize generative model training (due to confidentiality agreements, the institution's identity and simulation details are not disclosed). After this stage, the normal samples increased to 1167, and the four fault classes expanded to 47, 51, 39, and 62 samples, respectively. Subsequently, a generative adversarial network (GAN) trained on real data was used to generate additional fault samples, resulting in final fault class sizes of 251, 347, 249, and 250 for model training and validation. To prevent potential information leakage due to temporal autocorrelation, the real dataset was partitioned into training and test sets at the sample (event) level, with a 7:3 split, thereby ensuring temporal independence between the two subsets.

3.2. GAN-Based Fault Data Generation Model

This study adopted a GAN for multidimensional time-series data to generate synthetic LRE fault samples through adversarial training between a generator and a discriminator, thereby alleviating the limited availability of real fault samples. Due to the scarcity of LRE fault data samples—especially for extracting weak features—insufficient data severely compromises diagnostic accuracy. Thus, a GAN was used to generate augmented samples similar to real fault data, supporting dataset augmentation and improving the training conditions for downstream classification under few-shot settings. Figure 1 illustrates the overall architecture.

The generator employs a hierarchical architecture that starts from a low-resolution latent feature representation and progressively upsamples it to reconstruct multivariate time-series signals, thereby balancing computational efficiency and detail recovery capability. The model takes a 100-dimensional latent space vector as input, which is first mapped to a high-dimensional feature representation via a fully connected layer. A temporal downsampling strategy is subsequently adopted, significantly reducing the number of parameters and computational complexity of subsequent convolutional operations by lowering the time-step dimension. This design enables the model to process only small-scale feature maps in the initial stage, thereby achieving efficient resource allocation across both spatial and temporal dimensions.

During the feature reconstruction phase, the downsampled features are progressively upsampled across multiple layers of transposed convolutions. This hierarchical feature refinement strategy decomposes the complex temporal reconstruction task into multiple

scale-specific processing stages. Consequently, the model can focus solely on the feature maps at the current scale in each step, avoiding the high computational cost of directly processing large-scale sequences while maintaining the capacity for detail restoration. To address the strong temporal characteristics of liquid rocket engine fault data, such as the dynamic correlation between sudden pressure changes and temperature rises—which can lead to "temporally chaotic" data when using standard GANs—this study integrated a bidirectional long short-term memory (BiLSTM) network into the generator. This module receives the preliminarily upsampled features from the transposed convolutions and leverages its bidirectional architecture to capture both forward and backward dependencies within the time series. This effectively enhances the temporal coherence of the generated data without substantially increasing the computational burden, preventing the production of synthetic samples with inconsistent temporal dynamics.

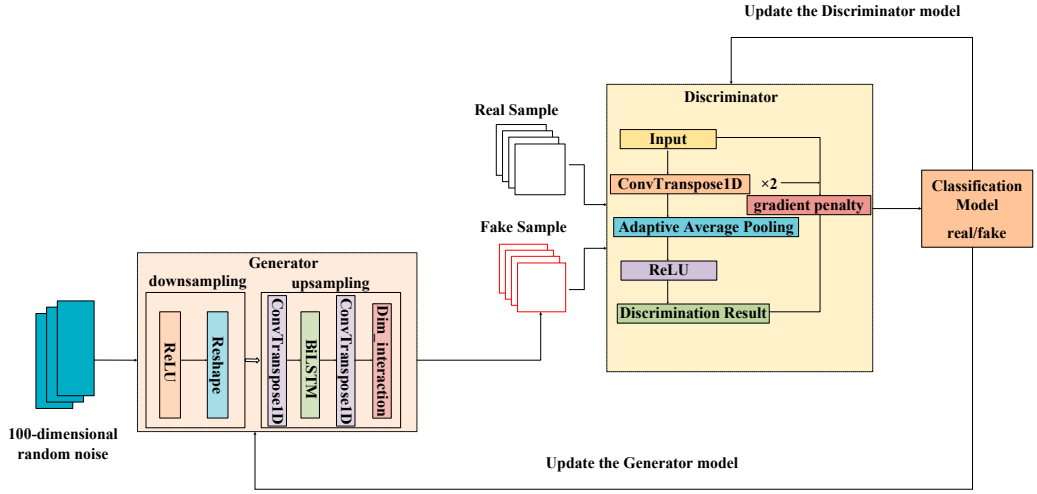


Figure 1. Improved GAN architecture.

To ensure independence and correlation are unified across the various dimensions in multivariate time-series data, the generator incorporates a dimension-interaction module. This module uses 1D convolutional layers to learn implicit relationships across parameter dimensions, generating a dimensional weight matrix. Feature-wise multiplication is then applied to achieve coordinated generation across multiple variables. This mechanism maintains the independent statistical properties of each parameter while ensuring the consistency of the multivariate time series. Finally, the output layer uses the hyperbolic tangent (Tanh) activation function to constrain the generated data to the interval $[-1, 1]$. This design ensures that the generated data have the same value range as the real data, normalized during preprocessing, thereby guaranteeing the rationality of the generated samples and their compatibility with subsequent processing steps.

The core task of the discriminator is to distinguish between real LRE fault data and synthetic samples. To balance feature discrimination capability and model training stability, its architecture employs a spectral normalization convolutional network [28]. The discriminator in the proposed framework functions as a critic rather than as a probabilistic binary classifier. Under the WGAN-GP objective, it outputs an unconstrained scalar score instead of a probability bounded between 0 and 1. Spectral normalization is applied to the convolutional and linear layers to improve training stability, and a gradient penalty is introduced during optimization to enforce the Lipschitz constraint.

During the training of the generator and discriminator, to optimize the model and enhance convergence stability, studies have shown that both WGAN [29] and WGAN-GP [30] can stabilize the optimization process and be better applied to other architectures. There-

fore, this study adopted the WGAN-GP framework for network training. Correspondingly, the generator was optimized using a composite objective consisting of an adversarial loss, a multi-scale temporal consistency loss, a frequency-domain loss, and an adaptive parameter-specific regularization term. The temporal term constrains first-order, second order, and window-level trend consistency between generated and real sequences; the frequency-domain term matches the spectral characteristics of the generated and real signals; and the parameter-specific term places additional emphasis on channels identified as difficult to model during preprocessing. That is, the generator not only needs to "deceive" the discriminator but also ensure that the temporal variation of the generated data is consistent with that of the real data. Through the adversarial learning between the generator and discriminator, 14-dimensional time-series data are generated. During training, the generator combines residual blocks and LSTM modules to capture multi-scale temporal features and dynamically weights its output via a dimension interaction layer. In the current implementation, the discriminator is updated on every other mini-batch, whereas the generator is updated once per mini-batch. Gradient clipping and adaptive learning-rate scheduling are further used to stabilize training.

In the model evaluation phase, to comprehensively assess the consistency between generated data and real fault data across multidimensional characteristics, this study adopted a comprehensive evaluation system based on temporal characteristics, statistical distribution, and overall similarity. This system is composed of three core dimensions: temporal indicator score, statistical indicator score, and Wasserstein distance, and comprehensively measures the quality of the generated data through a weighted comprehensive score.

Temporal characteristics are key features of LRE fault data, and their evaluation includes four critical sub-indicators. The improved dynamic time warping (DTW) distance serves as the core evaluation metric, measuring the alignment degree of temporal morphology between the generated and real sequences. By calculating the matching distance between two time series, DTW overcomes the standard time shifts and nonlinear distortions in engine fault signals. For example, for key parameters such as combustion chamber pressure and turbopump speed, DTW can accurately capture the positions of their mutation points and the trends in amplitude variation. Let the real data time series be $X = (x_1, x_2, \ldots, x_n)$, the generated data time series be $Y = (y_1, y_2, \ldots, y_n)$, and π denote the optimal alignment path; that is, the DTW distance solves the optimal alignment path via dynamic programming:

$$DTW(X,Y) = \frac{\min_{\pi} \sum_{(i,j)\in\pi} (x_i - y_i)^2}{\max(x_n, y_m)} \quad (1)$$

The remaining three indicators are robust autocorrelation similarity (ACF), power spectrum divergence, and trend consistency, respectively. Among them, ACF is mainly used to evaluate the degree of preservation of temporal dependencies, and is particularly suitable for reflecting periodic patterns in vibration signals (e.g., bearing faults) and pressure pulsations. The calculation of this indicator is shown in Equation (2), where ACF denotes the autocorrelation function, and ρ represents the Pearson correlation coefficient.

$$ACF = |\rho(ACF(X), ACF(Y))| \quad (2)$$

For the calculation of the power spectrum divergence indicator, this study adopted the Jensen–Shannon Divergence, as shown in Equation (3):

$$PSD = JS(P_X \parallel P_Y) = \frac{1}{2} D_{KL}(P_X \parallel M) + \frac{1}{2} D_{KL}(P_Y \parallel M) \quad (3)$$

where P_X and P_Y are the normalized power spectra of the real signal and the generated signal, respectively.

The last indicator, trend consistency, is implemented using sliding-window linear regression to verify the accuracy of long-term trends. In the sliding-window linear regression used for trend consistency evaluation, the window length is selected according to the sampling frequency and the characteristic time scale of signal evolution. The chosen window size is sufficiently large to smooth short-term fluctuations while preserving long-term trend information. Empirically, this setting provides a stable estimate of temporal trends, and minor variations in the window length do not lead to significant changes in trend consistency.

The Wasserstein distance measures the difference between the distributions of generated data and real data, and its calculation formula is:

$$D_{\text{Wasserstein}} = \inf_{\gamma \in \Gamma(P_X, P_Y)} E_{(x,y)\sim\gamma}[\|x - y\|] \tag{4}$$

where P_X and P_Y denote the set of all joint distributions of X and Y, respectively.

To comprehensively and quantitatively evaluate the quality of the generated data, this study establishes a hierarchical evaluation framework in which the Temporal Metric Score and Statistical Metric Score are defined as weighted aggregates of multiple underlying metrics. Specifically, the Temporal Metric Score integrates four core sub-metrics: DTW distance, ACF similarity, PSD divergence, and trend consistency, as shown in Equation (5).

$$S_{\text{temp}} = \frac{1}{N}\sum_{\text{j}=1}^{N}\Big(\sum_{i\in\{DTW,ACF,PSD,\ldots\}} \omega_i \cdot Norm(M_{ij})\Big) \tag{5}$$

where N denotes the number of sensor channels, M_{ij} represents the i-th fundamental temporal metric of the j-th channel, Norm(·) signifies the normalization operation, and ω_i refers to the internal weighting coefficient for each respective metric.

The Statistical Metric Score consolidates the relative errors of key statistical moments, including the mean, standard deviation, skewness, and kurtosis, as shown in Equation (6).

$$S_{\text{statistical}} = \frac{1}{N}\sum_{i=1}^{N}\sqrt{\frac{1}{M}\sum_{j=1}^{M}\left(V_{\text{real},i,j} - V_{\text{gen},i,j}\right)^2} \tag{6}$$

where V_{real} and V_{gen} represent the statistical values of the real and generated data, respectively.

The generator and discriminator were trained using the Adam optimizer (β_1 = 0.5, β_2 = 0.9). The model had 5000 training epochs; after every 100 epochs, the model was evaluated. The optimal model was selected by synthesizing statistical similarity, DTW distance, and classifier accuracy. Finally, the generated data were restored to their original dimensions via inverse standardization. In this framework, the GAN component is primarily introduced to improve the training distribution under severe data scarcity and class imbalance. The current evaluation supports temporal and statistical consistency of the generated samples, but does not establish complete physical realism.

3.3. LRE Fault Classification Based on the MHA-CBL Model

Using the generated augmented data samples, this study constructed a hybrid classification model integrating a convolutional neural network (CNN), a bidirectional long short-term memory network (BiLSTM), and a multi-head attention (MHA-CBL) to classify fault types during LRE fault diagnosis. CNN is recognized as an effective technique for extracting spatial features of data for learning [31]; thus, by leveraging CNN to extract

local temporal features, using BiLSTM to capture long-range dependencies, and employing multi-head self-attention to focus on key data features—along with fusing augmented samples generated by GAN and real data for training—this model not only addresses the data scarcity issue but also effectively improves the accuracy of fault classification under small-sample conditions.

First, the input layer of the MHA-CBL model consists of time-series signals, such as temperature and pressure, with a fixed time step per sample. In terms of model architecture, a multi-layer CNN is first used for local feature extraction, with the number of channels (K) gradually increasing from 16 to 128 and the size of convolutional kernels (C) varying across layers. The use of convolutional layers enables the model to extract hierarchical local features from raw samples progressively. After each convolutional operation, a batch normalization layer is applied to accelerate training, followed by an average pooling layer to reduce data dimensionality—this retains key information while effectively reducing computational complexity. Meanwhile, to prevent model overfitting, a dropout layer is added immediately after the pooling layer. After local features are extracted by CNN, a BiLSTM network is adopted to capture long-range temporal dependencies. Owing to its bidirectional structure, the model can capture forward and backward information in time-series data; thus, BiLSTM enhances the model's understanding of time-series data, which is crucial for accurately identifying fault categories. Second, to focus on the critical moments of LRE faults, the model is equipped with multi-head attention. An 8-head attention mechanism dynamically weights each time step, enabling effective identification of key time-step features associated with fault occurrence.

After the hierarchical processing above, the features are fed into a classification head for decision-making. The features are first flattened and then fed into a fully connected layer with 100 neurons for classification. This layer uses the ReLU activation function to introduce nonlinearity; a dropout layer is applied after it for regularization to prevent overfitting. Finally, the model outputs the predicted probabilities for different fault categories via a Softmax activation layer, thereby completing fault classification. The detailed layer structure and parameters of the MHA-CBL model are shown in Table 1.

Table 1. The layer details and parameters of the MHA-CBL model for fault diagnosis type.

Layers	Types	Output Shapes	Params	Activation	Key Hyperparameters/Notes
0	Input	(None, 1500, 14)	0		Input length = 1500, channels = 14
1	Conv1D_1	(None, 1496, 16)	1136	ReLU	filters = 16, kernel_size = 5, stride = 1, padding = valid
2	BatchNormalization	(None, 1496, 16)	64		BN after Conv1D
3	AveragePooling1D	(None, 748, 16)	0		pool_size = 2, stride = 2
4	Dropout (rate = 0.3)	(None, 748, 16)	0		rate = 0.3
5	Conv1D_2	(None, 746, 32)	3104	ReLU	filters = 32, kernel_size = 3, stride = 1, padding = valid
6	BatchNormalization	(None, 746, 32)	128		BN after Conv1D
7	AveragePooling1D	(None, 373, 32)	0		pool_size = 2, stride = 2
8	Dropout (rate = 0.3)	(None, 373, 32)	0		rate = 0.3
9	Conv1D_3	(None, 372, 64)	4160	ReLU	filters = 64, kernel_size = 2, stride = 1, padding = valid
10	BatchNormalization	(None, 372, 64)	256		BN after Conv1D
11	AveragePooling1D	(None, 186, 64)	0		pool_size = 2, stride = 2
12	Dropout (rate = 0.3)	(None, 186, 64)	0		rate = 0.3
13	Conv1D_4	(None, 185, 128)	16,512	ReLU	filters = 128, kernel_size = 2, stride = 1, padding = valid
14	BatchNormalization	(None, 185, 128)	512		BN after Conv1D
15	AveragePooling1D	(None, 92, 128)	0		pool_size = 2, stride = 2
16	Dropout (rate = 0.3)	(None, 92, 128)	0		rate = 0.3

Table 1. *Cont.*

Layers	Types	Output Shapes	Params	Activation	Key Hyperparameters/Notes
17	Bidirectional LSTM	(None, 92, 256)	197,632	tanh/sigmoid (default)	BiLSTM, return_sequences = True (to feed attention); units set per direction to match output dimension (framework default activations)
18	Dropout (rate = 0.3)	(None, 92, 256)	0		rate = 0.3
19	MultiHeadAttention	(None, 92, 256)	131,328	softmax (inside attention)	num_heads = 8, key_dim = 16 (i.e., heads × key_dim = 128); scaled dot-product attention + output projection; optional attention dropout if used
20	Flatten	(None, 23,552)	0		Flatten time and feature dimensions
21	Dense	(None, 100)	2,355,300	ReLU	units = 100
22	Dropout (rate = 0.5)	(None, 100)	0		rate = 0.5
23	Dense	(None, num_classes)	100	Softmax	output layer, units = num_classes

Note: Unless otherwise stated, Conv1D layers use ReLU activation, and the BiLSTM uses the default LSTM activations (tanh for the cell state and sigmoid for gates). The attention layer internally applies softmax to normalized attention scores.

To improve reproducibility, we provide complete architectural specifications of the proposed MHA-CBL model in Table 1. For each layer, we report not only the output shape and parameter count but also the activation function and critical hyperparameters. Specifically, for Conv1D layers, we list the kernel size, stride, padding mode, and the number of filters; for pooling layers we provide the pooling type and window/stride; for the BiLSTM, we report the number of units and whether it is bidirectional and returns sequences; and for the multi-head attention module, we include the model dimension, number of heads, and dropout. We also report the dropout rates and the activation functions used in the fully connected layers (with softmax in the output layer). These details allow the model to be fully reproduced.

During the training process, the input data are normalized to the range [0, 1] in the preprocessing stage to eliminate potential impacts from differences in feature scales during diagnosis, thereby improving the stability of model training. For label processing, fault category labels are converted to numerical labels and transformed into a format suitable for multi-class classification using one-hot encoding. The dataset is divided into a training set and a test set: the training set combines real data and GAN-generated data to enhance the model's robustness, while the test set consists entirely of real data not used in training and is used to evaluate the model's generalization ability objectively. The Adam optimizer was selected for the model, with a learning rate of 0.0001, and the loss function adopted categorical cross-entropy for the multi-classification task. The model used a batch size of 32 and 200 training epochs. To prevent model overfitting, an early stopping strategy was employed, and the model weights at the optimal epoch were saved by monitoring the validation set accuracy. The overall architecture of the proposed MHA-CBL model is illustrated in Figure 2.

To ensure seamless integration of the generated synthetic samples into the fault classifier, a rigorous range-alignment procedure was implemented. Although the generator internally produces normalized outputs in the range $[-1, 1]$ through the Tanh activation, the generated sequences are subsequently inverse-transformed back to the original physical scale using the dataset-level normalization statistics before being saved or evaluated. Therefore, generated and real samples are aligned through the downstream classifier-side preprocessing pipeline rather than through a fixed $X' = (X+ 1)/2$ mapping alone. This step ensures that the synthetic data are numerically aligned with real-world observations while preserving crucial morphological features.

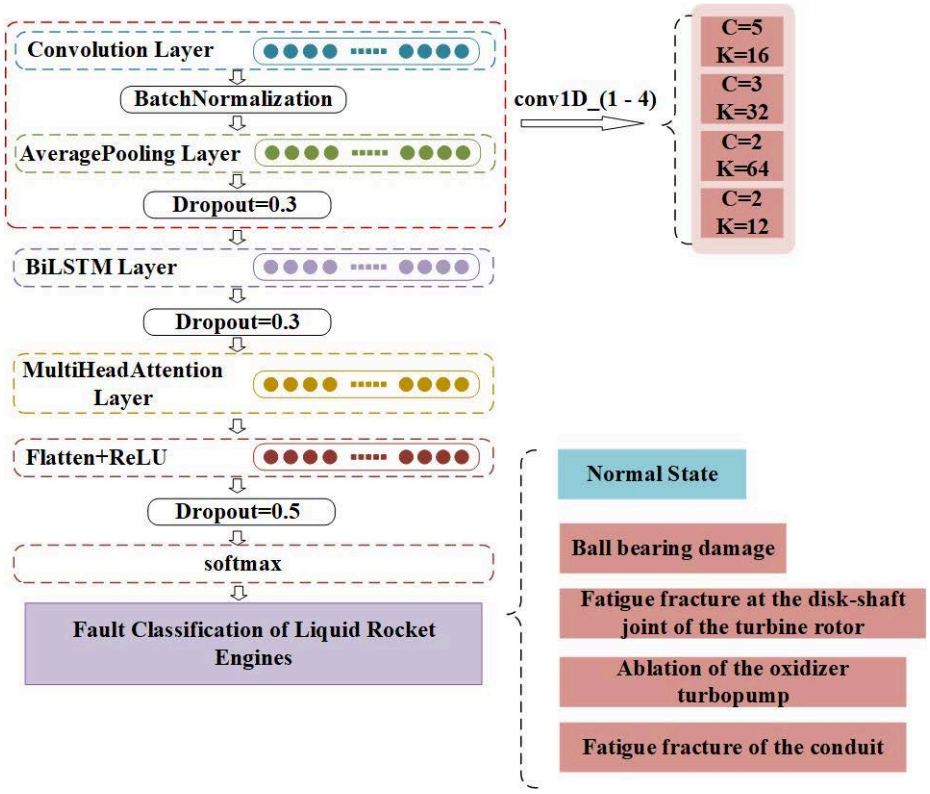


Figure 2. MHA-CBL model architecture.

After training, multidimensional metrics are used to evaluate the model's performance, including classification accuracy, precision, recall, and F1-score. Classification accuracy reflects the model's overall performance on the test set; precision and recall measure the model's prediction accuracy and coverage for each category, respectively; and the F1-score, a weighted average of the two, is suitable for evaluating classification performance on imbalanced datasets. In addition, a confusion matrix is used to visualize the distribution of predictions for each LRE fault category, intuitively showing the model's classification biases across categories. This helps adjust the model parameters to enable high-precision fault detection even with a small number of real samples.

4. Results

4.1. Performance of the GAN Model

In this study, the designed GAN model was used to generate fault data of LRE. After inputting different types of fault data into the GAN model for training, the generator produced corresponding waveform curves. To quantitatively evaluate the similarity between generated and real data, a comprehensive quality assessment was conducted on the generated time-series data for the four fault types. Among the assessment metrics, the comprehensive score was calculated by weighting three indicators—temporal characteristics, statistical distribution, and overall similarity—with weights of 0.6, 0.3, and 0.1, respectively. As shown in Table 2, in terms of overall performance, the comprehensive scores for all categories remained at a low level (<0.14), with the temporal metrics notably outperforming the statistical metrics. This result demonstrates the efficacy of the proposed progressive upsampling structure and bidirectional LSTM modules in preserving temporal feature coherence. It is noteworthy that the evaluation results varied across different fault categories. Fault 3 achieved the lowest comprehensive score (0.071), indicating the highest data quality in this category. In contrast, Fault 2 exhibited a significantly higher Wasserstein distance (112.93),

suggesting a more complex distributional characteristic for this specific fault pattern. This observation provides a clear direction for subsequent model optimization.

Table 2. Comprehensive evaluation results of the four categories of generated data. Note: The Temporal Metric Score and Statistical Metric Score are composite scores derived from multiple underlying metrics (DTW, ACF, PSD, trend consistency, and statistical moments). Please refer to Section 3.2 for detailed definitions and formulas.

Class	Overall Score	Temporal Metric Score	Statistical Metric Score	Wasserstein Distance
1	0.113	0.064	0.199	15.012
2	0.138	0.050	0.024	112.928
3	0.071	0.079	0.026	15.547
4	0.074	0.061	0.071	15.751

To systematically evaluate the quality of the generated data, this study conducted a multidimensional line-plot comparison between the raw and generated data for different fault categories, as shown in Figure 3. Given the large number of sample parameters and the confidentiality constraints associated with part of the data, a set of representative parameters was selected for each fault type, normalized, and then visualized. The comparison results indicate that across various operating conditions, the generated data exhibited substantial temporal consistency with the raw data, with the shapes, magnitudes, and evolution trends of key feature points remaining broadly similar. These observations suggest that the generative model captures important characteristics of the data distribution across different fault modes under the current evaluation setting. Based on this visualization analysis, the generated fault data provide preliminary support for subsequent classifier training and model refinement, particularly for improving the generation of more complex fault patterns.

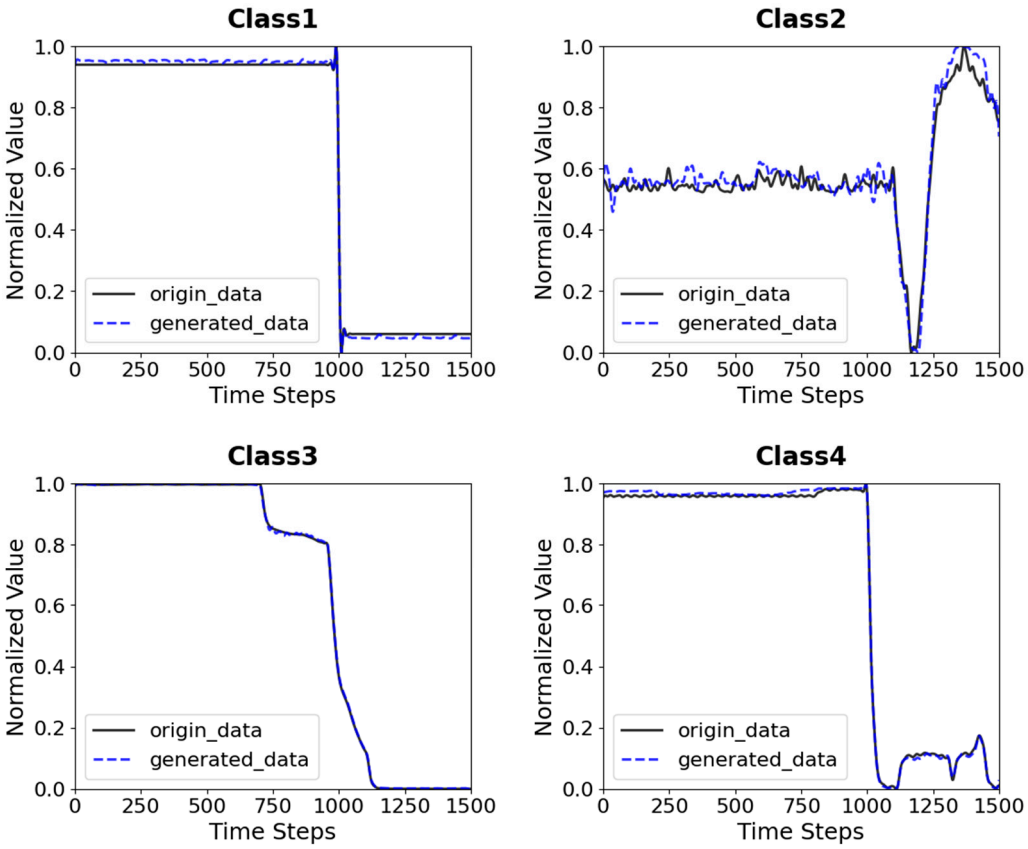


Figure 3. Comparative analysis of raw and generated data.

Although Figure 3 primarily offers a qualitative comparison between real and generated signals, the characteristics of the generated data are intrinsically linked to the parameterization of the generative model. Specifically, the smoothness or fluctuation of the synthesized signals is governed by several key factors: the strength of regularization and smoothing constraints in the objective function, which directly affects temporal continuity—stronger constraints suppress high-frequency components to produce smoother waveforms, whereas relaxed constraints preserve more dynamic variations; the noise injection strategy, including the distribution and scaling of noise, where greater variance enhances signal diversity and fluctuation, while reduced amplitude yields more stable outputs; and model capacity and training stability, particularly the balance between generator and discriminator or between reconstruction fidelity and regularization, which determines whether fine-grained fault dynamics are captured or results become overly smoothed. Through coordinated tuning of these factors, the proposed framework can be adapted to generate signals with tailored smoothness or fluctuation profiles to meet specific application requirements.

By quantitatively analyzing the autocorrelation function (ACF) of each parameter's temporal characteristics, the proposed model was further evaluated for its ability to preserve temporal dependency structures. The results indicate that the model exhibits varying performance in maintaining temporal correlations across different parameters. As shown in Figure 4, based on the ACF similarity metrics, the ACF curves of the model-generated sequences closely matched those of the original data for some key parameters. In particular, the decay trends across lags, periodic patterns, and peak positions remained highly consistent, demonstrating excellent temporal coherence and strong capability in reconstructing underlying dynamical features. For the remaining parameters, although the overall correlation trends generally remained consistent with the original data, noticeable deviations arose in the magnitudes of the correlation coefficients and in specific local details, indicating that the model achieves only a moderate level of temporal modeling performance in these cases. It should be noted that due to confidentiality constraints, the 14 parameters involved in the analysis are anonymized and denoted p1, p2, . . ., p14, without disclosing their specific physical meanings.The mapping between these anonymized features and their general physical measurement categories is provided in Appendix A.1.

Notably, parameters p14, p8, and p11 achieved exceptional ACF correlation coefficients of 0.99997, 0.99498, and 0.98170, respectively, along with impressively low mean absolute errors (MAEs) of less than 0.01. These results suggest near-perfect replication of temporal dynamics for these parameters, which is particularly important for monitoring the operation of the liquid rocket engine. Additionally, p10 demonstrates strong temporal consistency with an ACF correlation of 0.89642, further confirming the model's ability to capture essential dynamic features.

However, the analysis also identified areas for improvement in parameters p1, p2, and p12, which exhibited moderate ACF correlations ranging from 0.377 to 0.413 and higher MAE values exceeding 0.82. This variation in performance across parameters indicates that while the temporal-enhanced GAN architecture effectively captures global temporal patterns, the specific characteristics of each parameter may require further architectural adjustments or tailored training strategies.

To further validate the realism of the generated data beyond autocorrelation analysis, additional qualitative and distributional evaluations were conducted. First, temporal morphology comparisons between real and generated signals were performed through direct time-domain visualization, showing that the generated signals preserve the characteristic waveform patterns and dynamic behaviors of real engine data.

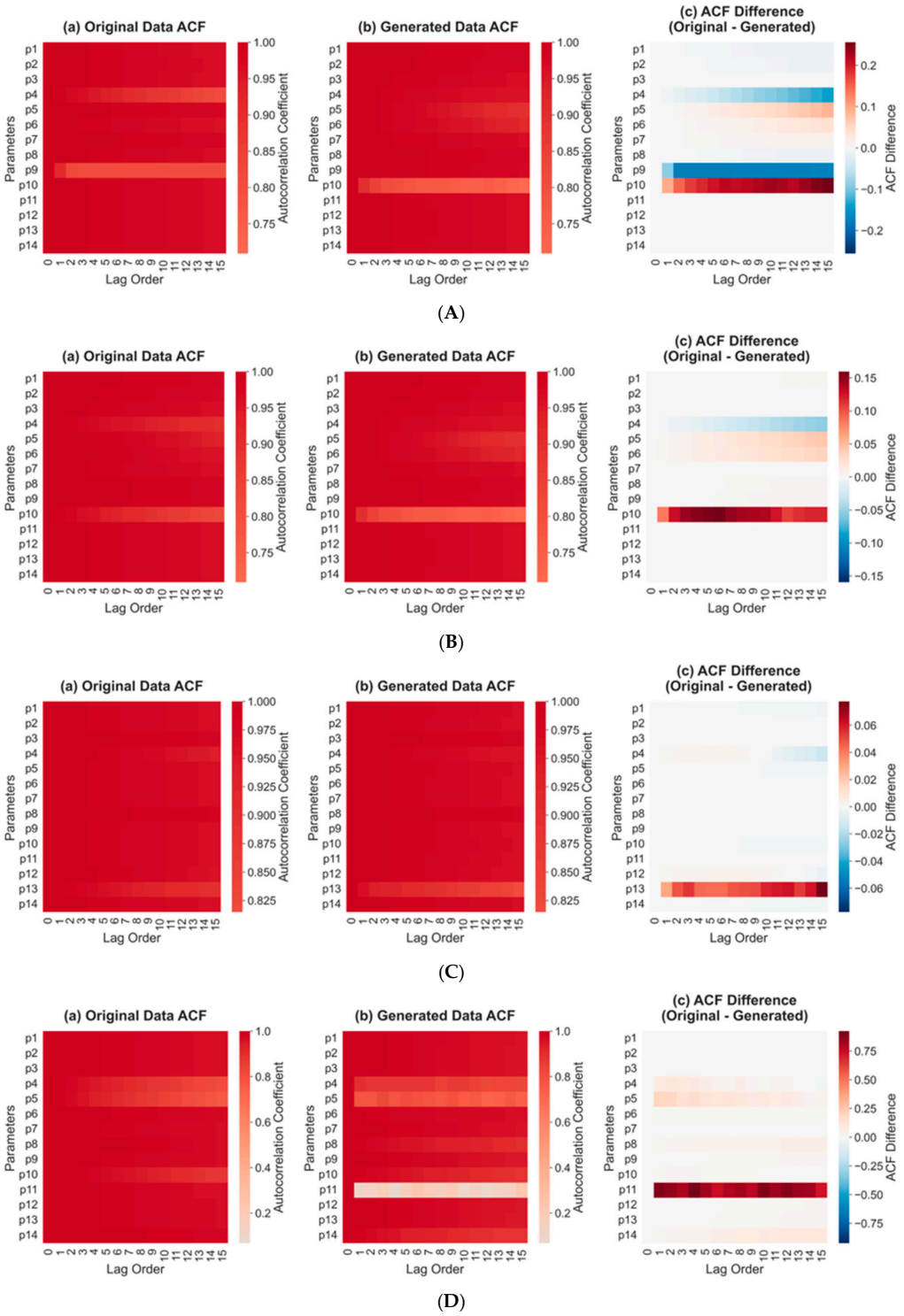


Figure 4. (**A**) Class1: ACF Evaluation; (**B**) class2: ACF Evaluation; (**C**) class3: ACF Evaluation; (**D**) class4: ACF Evaluation.

Furthermore, a t-SNE visualization was employed to compare real and generated samples in a unified low-dimensional embedding space. In this analysis, both real and generated multi-channel samples were jointly embedded using the exact feature representation. As shown in Figure 5, the generated data exhibited substantial overlap with the real data distribution, indicating that the proposed generative model captures the underlying data structure rather than merely reproducing limited statistical characteristics.

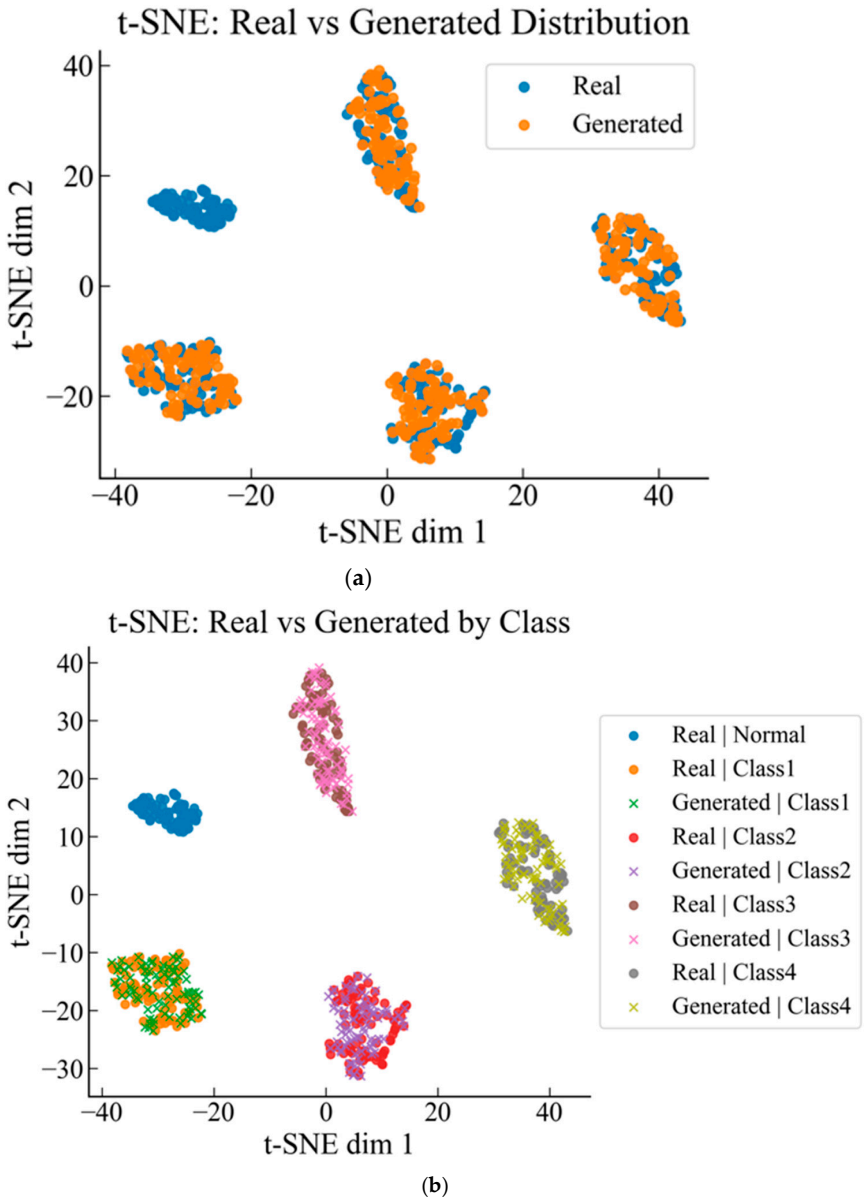


Figure 5. (**a**) t-SNE visualization of real and generated data; (**b**) t-SNE visualization of real and generated data by class.

Overall, the generator-side results provide empirical evidence that the generated samples preserve important temporal and statistical characteristics of the original signals under the current evaluation setting. However, these results should be interpreted as evidence of temporal and statistical consistency rather than as a complete validation of physical equivalence. Further physics-oriented validation of inter-channel coupling, phase consistency, and fault-propagation-related properties remains an important direction for future work.

4.2. Classification Results of the MHA-CBL Model

As shown in Figure 5, the dataset exhibited clear class separation in the t-SNE embedding space, suggesting favorable conditions for subsequent classification. This separation indicates the potential for effective classification. The balanced distribution of samples across the five classes further enhances model training and mitigates issues related to class imbalance. Notably no significant overlap was observed in the t-SNE plot. The separation

between the classes was very distinct and provided favorable conditions for subsequent classification tasks.

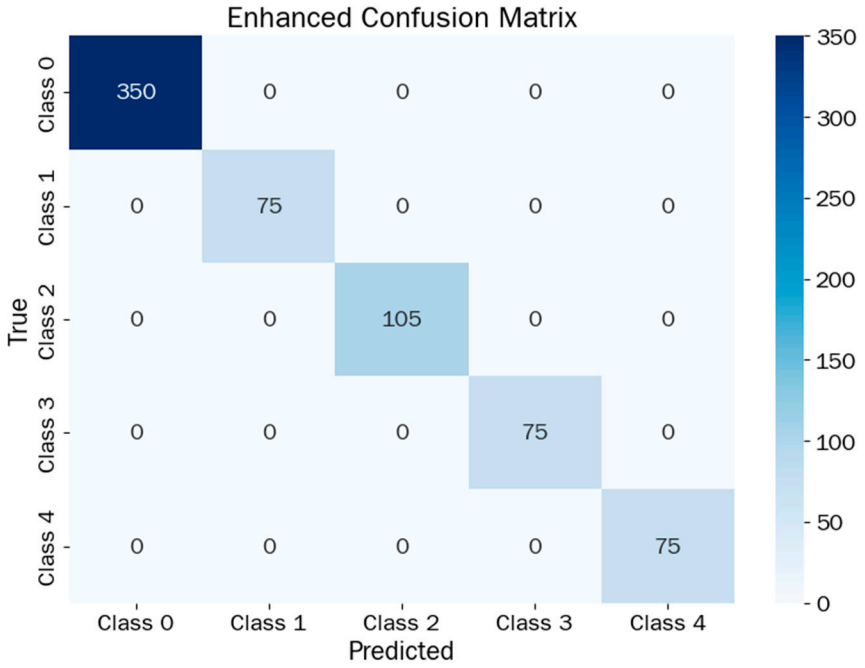
During classification model training, we designed differentiated data processing strategies based on the distribution characteristics of samples within each category. Specifically, Label 0 denotes the normal operating condition of the liquid rocket engine (LRE). At the same time, Labels 1, 2, 3, and 4 correspond to four typical fault types occurring during LRE operation. These faults include ball bearing damage, fatigue fracture at the rotor disk–shaft junction of the turbopump, oxygen turbopump ablation, and conduit fatigue fracture. To effectively mitigate overfitting, we incorporated an early stopping strategy into LSTM network training. This strategy automatically saves the optimal weights by monitoring validation-set accuracy and ensures satisfactory generalization performance under complex operating conditions.

After completing the training and validation phases, we conducted a more detailed assessment of the model's ability to identify different fault patterns. To evaluate classification performance under different data conditions, we analyzed the model on two subsets: a real-event test subset and an auxiliary augmented-pattern test subset constructed under the current framework. The confusion matrices for these two subsets are shown in Figure 6a,b, respectively, and Table 3 summarizes the corresponding precision, recall, and F1-score values. Following the 7:3 split described in Section 3.1, the final augmented dataset comprised 1167 normal samples, together with 251, 347, 249, and 250 samples for the four fault classes, including both real and generated samples. Within the test partition, the numbers of real samples were 98, 4, 8, 2, and 10 for the five classes, respectively, while the remaining fault samples were generated under the current framework. Accordingly, the real-event subset was used to assess feasibility on scarce real fault cases, whereas the augmented-pattern subset served only as an auxiliary evaluation of the learned fault patterns under the present framework.



(**a**)

Figure 6. *Cont.*



(**b**)

Figure 6. Classification confusion matrices of the MHA-CBL model on (**a**) original data and (**b**) synthetic data.

Table 3. Classification metrics of the MHA-CBL model on original and synthetic data.

Class	Precision	Recall	F1-Score	Support
Original data				
0	1.00	1.00	1.00	98
1	1.00	1.00	1.00	4
2	1.00	1.00	1.00	8
3	1.00	1.00	1.00	2
4	1.00	1.00	1.00	10
Accuracy		1.00	1.00	122
Weighted avg	1.00	1.00	1.00	122
Macro avg	1.00	1.00	1.00	122
Synthetic data				
0	1.00	1.00	1.00	350
1	1.00	1.00	1.00	75
2	1.00	1.00	1.00	105
3	1.00	1.00	1.00	75
4	1.00	1.00	1.00	75
Accuracy		1.00	1.00	680
Weighted avg	1.00	1.00	1.00	680
Macro avg	1.00	1.00	1.00	680

As shown in Figure 6b and the "synthetic data" section of Table 3, the proposed method achieved perfect classification results on the augmented-pattern test subset. This result indicates that the MHA-CBL model trained on the augmented dataset can effectively recognize the fault patterns represented in the generated samples under the current framework. As shown in Figure 6a and the "original data" section of Table 3, the model also achieved perfect accuracy on the extremely limited real-event test subset. However, these results should be interpreted cautiously because the number of real fault samples

remains very small. Taken together, the findings provide preliminary evidence that the introduction of generated data helps alleviate the small-sample problem and supports the learning of more balanced feature representations across classes. They also suggest that the MHA-CBL model can effectively discriminate complex fault patterns under the present dataset and evaluation protocol, rather than establishing universal superiority or broad engineering generalization.

4.3. Verification of Accuracy and Data Integrity

Given the exceptionally high diagnostic accuracy achieved under small-sample conditions, we conducted a rigorous verification process to rule out overfitting and potential data leakage.

Firstly, it is essential to clarify the experimental protocol. We trained the proposed MHA-CBL model on an augmented dataset comprising synthetic samples generated by the improved GAN together with a portion of the original data. For evaluation, we considered both an extremely limited real-event test subset and an auxiliary augmented-pattern test subset under the current framework. The reported 100% accuracy on the real-event subset reflects performance on a very small number of real fault cases. Due to the inherent scarcity of actual LRE fault events, this subset remains sparse, and the resulting performance should therefore be interpreted with caution.

To ensure the integrity and reliability of the reported results, we implemented several rigorous measures. Firstly, we strictly maintained temporal independence by segmenting the data based on independent engine firing events rather than random time-point sampling. This approach ensures that no overlapping fragments occurred between the training and testing phases and effectively prevents data leakage. Additionally, we acknowledge the statistical limitations of the current study. Although the proposed model demonstrated superior performance on the available dataset, the results should be interpreted with caution due to the inherent scarcity of real-world fault samples. Consequently, future research will focus on validating the model's robustness across a broader range of engine operating conditions and previously unseen fault modes to further verify its generalizability.

4.4. Performance Validation on the Aero-Engine Inter-Shaft Bearing Dataset

To further evaluate the diagnostic robustness and generalizability of the proposed MHA-CBL model in a complex aerospace environment we additionally conducted benchmarking experiments on the official inter-shaft bearing fault diagnosis dataset from an aero-engine system (HIT dataset). This dataset was acquired from a real dual-rotor test rig and is characterized by strong background noise and pronounced non-stationarity. These characteristics make it a useful public benchmark for supplementary validation [32]. We first compared the proposed MHA-CBL model with two classical machine learning baselines including random forest and support vector machine (SVM). As shown in Table 4, the proposed model achieved 97.97% test accuracy and a macro-F1 score of 97.77%. In comparison random forest and SVM achieved 71.56%/69.21% and 70.90%/70.25%, respectively. These results provide additional public-benchmark evidence that the proposed classifier can extract informative spatiotemporal features under high-noise and long-sequence conditions. However, rather than attributing the entire gain to the attention mechanism alone, we provide the following supplementary analyses. These analyses further examine component-wise contribution, behavior under imbalanced few-shot conditions, and uncertainty-aware prediction.

Table 4. Comparison of experimental results on the HIT-dataset.

Model	Test Accuracy (%)	Macro-F1 Score (%)
Random Forest	71.56	69.21
SVM	70.90	70.25
MHA-CBL	97.97	97.77

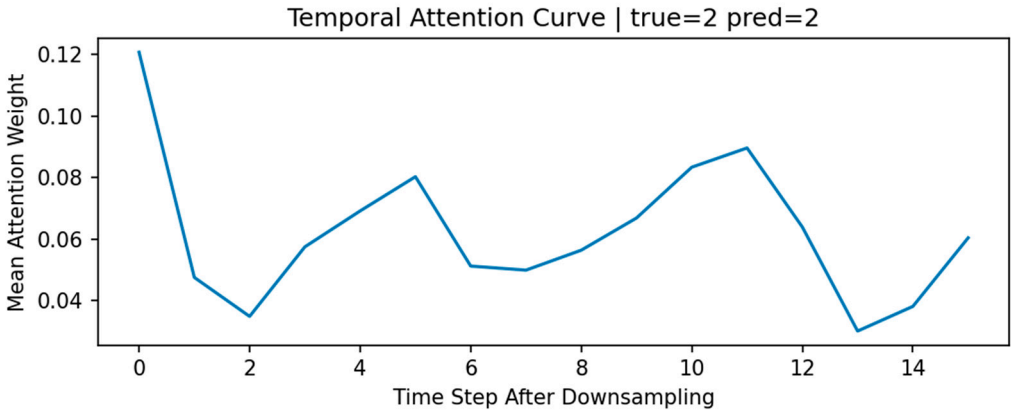
It should be noted that the following supplementary analyses were conducted under controlled auxiliary settings designed for specific purposes. These purposes include component-wise ablation, extreme imbalanced few-shot evaluation, and uncertainty-aware prediction. Therefore, their numerical results are not intended to be compared directly with the main benchmark result in Table 4.

To better clarify which architectural components are responsible for the observed performance gain, we further conducted a classifier-side ablation study on the public HIT benchmark. The results show clear progression from CNN to CNN + BiLSTM and then to CNN + BiLSTM + MHA. Specifically, CNN achieved an accuracy of 82.75% and a macro-F1 score of 82.43%. Then, CNN + BiLSTM improved to 93.53% and 93.00% while CNN + BiLSTM + MHA further improved to 93.91% and 93.54%, respectively, as summarized in Table 5. These findings indicate that temporal sequence modeling is the primary contributor to performance improvement. Meanwhile, the MHA module provides an additional but relatively smaller and consistent gain. To further support the interpretability of the attention-enhanced classifier, we additionally visualized representative attention heatmaps and temporal attention curves, as shown in Figure 7. As illustrated in Figure 7, the attention weights were strongly concentrated around a limited number of temporal positions rather than being uniformly distributed over the sequence. This behavior supports the interpretation that the MHA module helps the classifier focus on fault-relevant temporal regions.

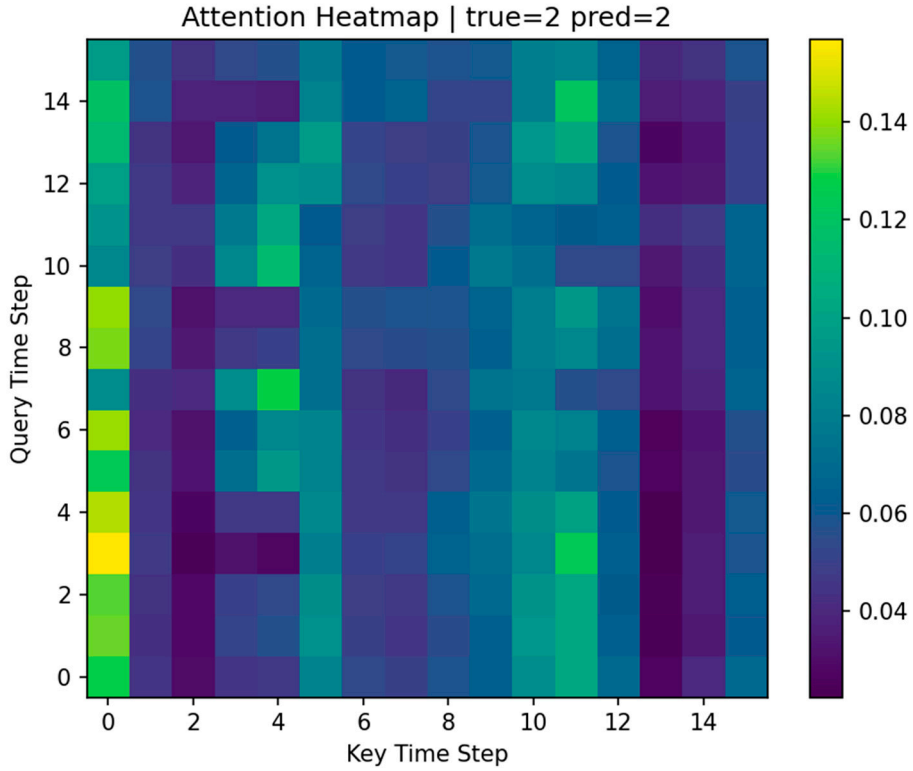
Table 5. Classifier-side ablation results on the public HIT dataset.

Model	Accuracy(%)	Macro-F1 Score (%)
CNN	82.75	82.43
CNN + BiLSTM	93.53	93.00
CNN + BiLSTM + MHA	93.91	93.54

To provide additional context for the few-shot imbalanced diagnosis setting, we further constructed an intentionally challenging protocol on the public HIT dataset. In this protocol, the three training classes were reduced to 20, 10, and 20 samples, respectively, while the test set remained unchanged. Under this setting, we compared four commonly used imbalance-mitigation strategies: standard cross-entropy (CE), class-weighted cross-entropy (weighted CE), focal loss, and random oversampling. The results show that imbalance handling remains highly challenging under such extreme sample scarcity. Specifically, CE achieved accuracy/macro-F1/macro-recall of 41.29%/29.76%/39.03% and weighted CE achieved 56.30%/50.43%/50.70%. Similarly, focal loss achieved 58.13%/42.80%/47.90%, and random oversampling achieved 39.55%/18.89%/33.33%. Among these baselines, weighted CE yielded the most balanced overall performance. In contrast, plain CE and random oversampling showed severe majority-biased collapse. These results provide stronger context for the intrinsic difficulty of the imbalanced few-shot diagnosis problem and help position the proposed method in a more realistic comparative setting. The corresponding results are summarized in Table 6.



(**a**)



(**b**)

Figure 7. (**a**) Temporal attention curve for a representative correctly classified sample. (**b**) Attention heatmap for the same sample.

Table 6. Comparison of conventional imbalance-mitigation strategies under an intentionally imbalanced few-shot protocol on the public HIT dataset.

Method	Accuracy (%)	Macro-F1(%)	Macro-Recall (%)
CE	41.29	29.76	39.03
Weighted CE	56.30	50.43	50.70
Focal Loss	58.13	42.80	47.90
Random Oversampling	39.55	18.89	33.33

Because fault diagnosis for aerospace systems is inherently safety-critical, we additionally examined whether low-confidence predictions could be handled more conservatively. To this end, we performed an uncertainty-aware evaluation on the public HIT dataset

using MC-dropout during inference and incorporated a simple confidence-threshold-based reject option. The base model achieved an accuracy of 92.41%, a macro-F1 of 92.14%, and an expected calibration error (ECE) of 0.0596. The corresponding calibration behavior is illustrated in Figure 8. We then evaluated selective prediction under different confidence thresholds. The results show a clear trade-off between coverage and reliability. Specifically, as low-confidence samples are rejected, the accuracy of the predictions retained increases substantially. At coverage levels of 97.51%, 81.59%, 72.01%, 58.37%, and 47.55% the corresponding selective accuracies were 93.58%, 97.51%, 98.62%, 99.50%, and 99.91%, respectively. This trend is visualized more clearly in Figure 9. This behavior is desirable in safety-critical diagnosis scenarios where conservative decision-making may be preferable to forced misclassification. It should be noted, however, that this experiment should be interpreted as an initial uncertainty-aware extension toward safer deployment rather than a complete open-set diagnosis framework.

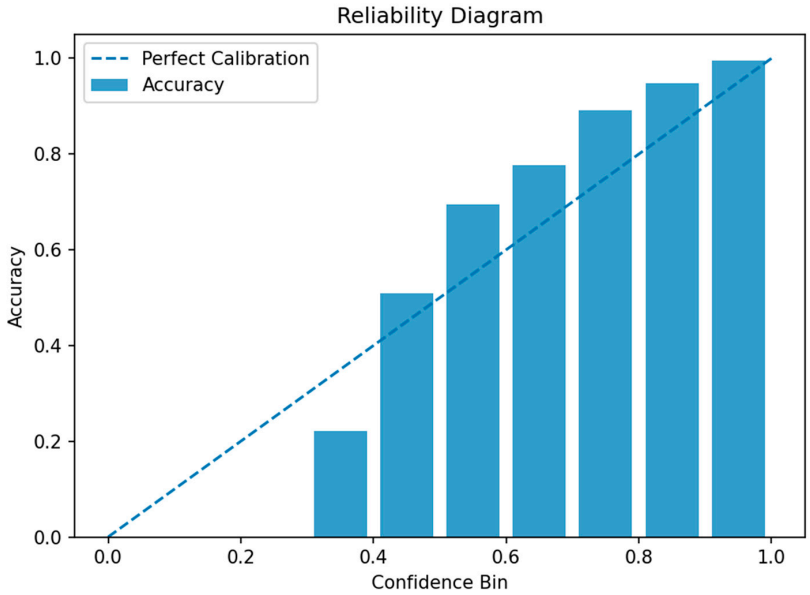


Figure 8. Reliability diagram of the uncertainty-aware MHA-CBL classifier on the public HIT dataset.

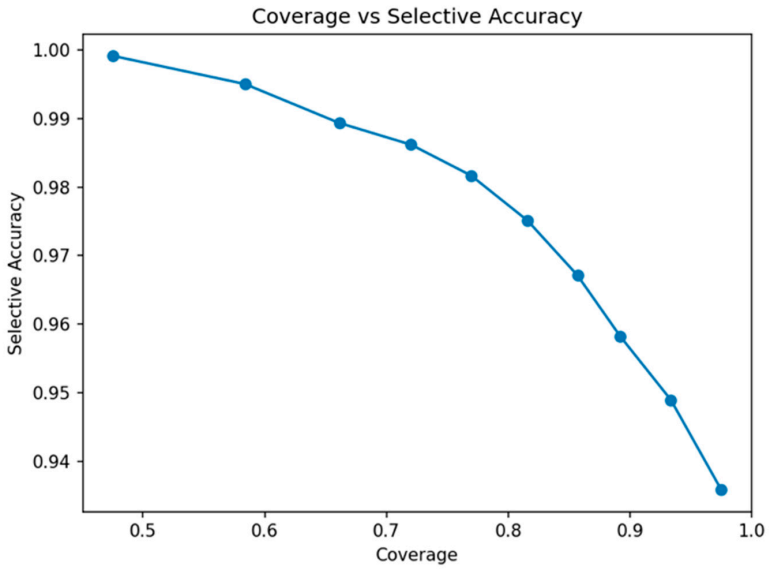


Figure 9. Coverage-selective-accuracy trade-off under confidence-based rejection on the public HIT dataset.

Taken together, the supplementary experiments on the public HIT dataset provide classifier-side evidence from three complementary perspectives. The ablation results clarify that BiLSTM-based temporal modeling is the primary contributor and that MHA provides an additional gain. Furthermore, the imbalanced few-shot comparison demonstrates the intrinsic difficulty of the target problem. Finally, the uncertainty-aware analysis shows that the model exhibits a meaningful confidence–reliability relationship relevant to safety-oriented deployment. These supplementary results do not replace the generator-side evidence on the confidential LRE dataset. Instead, they strengthen the overall argument of the manuscript by providing a more transparent and better-grounded evaluation of the proposed diagnosis framework.

5. Discussion

This study addresses fault diagnosis for LREs under severe data scarcity and class imbalance. To cope with this challenge, a GAN-based augmentation framework and a hybrid CNN-BiLSTM-MHA classifier were developed and evaluated. The results provide encouraging evidence that the proposed framework can improve diagnostic performance when only limited real fault samples are available.

In the present framework, the GAN component is intended to alleviate fault-data scarcity and class imbalance rather than to fully reproduce the physical mechanism of fault evolution. Under the current evaluation protocol, the generated samples exhibited temporal and statistical consistency with the real data, suggesting that they can provide useful support for downstream model training. At the classifier level, the CNN extracts local features, the BiLSTM captures long-range temporal dependencies, and the MHA module helps emphasize diagnostically informative temporal regions. In addition, supplementary experiments on the public HIT aero-engine dataset further support the effectiveness of the classifier architecture, including its component-wise contribution and robustness under imbalanced few-shot settings.

Several limitations of this study should be acknowledged. First, the current generator-side evidence mainly supports temporal and statistical consistency, but does not yet establish full physical plausibility of the generated signals. Second, the real-fault test set remains limited, so the reported performance should be viewed as evidence of feasibility under the current dataset rather than proof of universal superiority or direct engineering deployability. Third, more complete generator-side ablation, especially for the temporal-consistency-related design choices, has not yet been fully conducted on the confidential LRE dataset. Future work will therefore focus on broader external validation, more rigorous generator-side ablation, and physics-oriented assessment of the generated signals.

6. Conclusions

This study proposed a GAN-based fault data augmentation framework and a hybrid CNN-BiLSTM-MHA classifier for liquid rocket engine fault diagnosis under few-shot and imbalanced data conditions. The results show that under the present evaluation protocol, the generated samples preserved important temporal and statistical characteristics of the original data, and the proposed classifier achieved strong diagnostic performance on the augmented dataset. Supplementary experiments on the public HIT aero-engine benchmark further support the effectiveness of the classifier architecture, including its component-wise contribution, interpretability, and behavior under imbalanced few-shot settings.

Overall, these findings provide encouraging evidence that the proposed framework can improve LRE fault diagnosis when real fault data are scarce. At the same time, the conclusions of this study should be interpreted within the scope of the available data and current validation setting. More comprehensive generator-side ablation, broader

external testing, and physics-oriented assessment of the generated signals are still required before stronger claims can be made regarding physical realism, generalization, and engineering deployment.

Author Contributions: Conceptualization, H.H. and S.R.; Methodology, H.H. and R.Z.; Validation, H.H. and R.Z.; Formal analysis, R.Z. and S.R.; Investigation, H.H. and S.R.; Resources, H.H., S.R. and H.W.; Data curation, H.H. and R.Z.; Writing—original draft, R.Z. and C.X.; Writing—review & editing, S.R., H.H. and H.W.; Supervision, S.R.; Project administration, H.H. and S.R.; Funding acquisition, S.R. All authors have read and agreed to the published version of the manuscript.

Funding: This research was funded by the Beijing Institute of Technology Research Fund Program for Young Scholars, grant number 202207001.

Data Availability Statement: The data supporting the findings of this study are confidential and cannot be shared publicly due to proprietary and security constraints.

Acknowledgments: The authors would like to thank the technicians at the Beijing Aerospace Propulsion Institute test facility for their assistance in data acquisition. We also appreciate the constructive discussions with colleagues from the School of Automation, Beijing Institute of Technology.

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

Abbreviations

The following abbreviations are used in this manuscript:

GAN	Generative Adversarial Network
BiLSTM	Bidirectional Long Short-Term Memory
MHA-CBL	Multi-Head Attention-CNN-BiLSTM (model)
WGAN-GP	Wasserstein GAN with Gradient Penalty
CNN	Convolutional Neural Network
LRE	Liquid Rocket Engine
RLV	Reusable Launch Vehicle
MHA	Multi-Head Attention

Appendix A

Appendix A.1

Anonymized Feature Reference Table. This table provides the mapping between the 14 anonymized features (p1–p14) used in this study and their corresponding general physical measurement categories. Specific sensor labels and detailed engineering parameters are withheld due to proprietary and security constraints.

Table A1. Anonymized feature reference table.

Feature	General Physical Category
P1	Pressure
P2	Pressure
P3	Pressure
P4	Pressure
P5	Pressure
P6	Pressure
P7	Pressure
P8	Rotational Speed
P9	Rotational Speed
P10	Flow Rate
P11	Temperature
P12	Pressure

Table A1. *Cont.*

Feature	General Physical Category
P13	Pressure
P14	Temperature

Appendix A.2

To ensure the reproducibility of our methods and results, the complete source code for this study—including data preprocessing, the temporal-enhanced GAN architecture, the MHA-CBL classifier, training scripts, and evaluation metrics—has been made publicly available under an open-source license. The code repository can be accessed at: https://gitcode.com/shuce/MHA-CBL (accessed on 15 March 2026)

References

1. Aiswarya, N.; Suja Priyadharsini, S.; Moni, K.S. An efficient approach for the diagnosis of faults in turbo pump of liquid rocket engine by employing FFT and time-domain features. *Aust. J. Mech. Eng.* **2018**, *16*, 163–172. [CrossRef]
2. Das, D.; Padmanabhan, P.; Kumaresan, V.; Sudhakar, D.P. Case Study on Bearing Fault Diagnosis in Liquid Rocket Engine Using Envelope Detection Technique. In Proceedings of the Advances in Mechanical and Materials Technology, Singapore, 19 December 2022; pp. 357–366.
3. Duyar, A.; Eldem, V. Fault Detection and Diagnosis in Propulsion Systems; A Real Time Identification Approach. *IFAC Proc. Vol.* **1991**, *24*, 473–478. [CrossRef]
4. Lee, K.; Cha, J.; Ko, S.; Park, S.-Y.; Jung, E. Fault detection and diagnosis algorithms for an open-cycle liquid propellant rocket engine using the Kalman filter and fault factor methods. *Acta Astronaut.* **2018**, *150*, 15–27. [CrossRef]
5. Cha, J.; Ha, C.; Suheon, O.; Ko, S. A Survey on Health Monitoring and Management Technology for Liquid Rocket Engines. *J. Korean Soc. Propuls. Eng.* **2014**, *18*, 50–58. [CrossRef]
6. Komlev, G.V.; Mitrofanova, A.S. To the question of forecasting the technical condition of low-thrust liquid rocket engines. *Sib. Aerosp. J.* **2020**, *21*, 78–84. [CrossRef]
7. Li, N.; Xue, W.; Guo, X.; Xu, L.; Wu, Y.; Yao, Y. Fault Detection in Liquid-Propellant Rocket Engines Based on Improved PSO-BP Neural Network. *J. Softw.* **2019**, *14*, 380–387. [CrossRef]
8. Park, S.-Y.; Ahn, J. Deep neural network approach for fault detection and diagnosis during startup transient of liquid-propellant rocket engine. *Acta Astronaut.* **2020**, *177*, 714–730. [CrossRef]
9. He, A.; Jin, X. Failure Detection and Remaining Life Estimation for Ion Mill Etching Process Through Deep-Learning Based Multimodal Data Fusion. *J. Manuf. Sci. Eng. Trans. ASME* **2019**, *141*, 101008. [CrossRef]
10. JDMD Editorial Office; Gebraeel, N.; Lei, Y.; Li, N.; Si, X.; Zio, E. Prognostics and Remaining Useful Life Prediction of Machinery: Advances, Opportunities and Challenges. *J. Dyn. Monit. Diagn.* **2023**, *2*, 1–12. [CrossRef]
11. Chen, X.; Ma, M.; Zhao, Z.; Zhai, Z.; Mao, Z. Physics-Informed Deep Neural Network for Bearing Prognosis with Multisensory Signals. *J. Dyn. Monit. Diagn.* **2022**, *1*, 200–207. [CrossRef]
12. Aswal, N.; Sen, S.; Mevel, L. Switching Kalman filter for damage estimation in the presence of sensor faults. *Mech. Syst. Signal Process.* **2022**, *175*, 109116. [CrossRef]
13. Cha, J.; Ko, S.; Park, S.-Y.; Jeong, E. Fault detection and diagnosis algorithms for transient state of an open-cycle liquid rocket engine using nonlinear Kalman filter methods. *Acta Astronaut.* **2019**, *163*, 147–156. [CrossRef]
14. Maru, Y.; Mori, H.; Ogai, T.; Mizukoshi, N.; Takeuchi, S.; Yamamoto, T.; Yagishita, T.; Nonaka, S. Anomaly Detection Configured as a Combination of State Observer and Mahalanobis-Taguchi Method for a Rocket Engine. *Trans. Jpn. Soc. Aeronaut. Space Sci. Aerosp. Technol. Jpn.* **2018**, *16*, 195–201. [CrossRef]
15. Kanso, S.; Jha, M.S.; Galeotta, M.; Theilliol, D. Remaining Useful Life Prediction with Uncertainty Quantification of Liquid Propulsion Rocket Engine Combustion Chamber. *IFAC-PapersOnLine* **2022**, *55*, 96–101. [CrossRef]
16. ALI, M.; GUPTA, U. An Expert System for Fault Diagnosis in a Space Shuttle Main Engine. In *26th Joint Propulsion Conference*; American Institute of Aeronautics and Astronautics: Reston, VA, USA, 1990. [CrossRef]
17. Gupta, U.K.; Ali, M. LEADER-an integrated engine behavior and design analyses based real-time fault diagnostic expert system for space shuttle main engine (SSME). In *Proceedings of the 2nd International Conference on Industrial and Engineering Applications of Artificial Intelligence and Expert—Systems Volume 1, Tullahoma, TN, USA, 6–9 June 1989*; ACM: New York, NY, USA, 1989; pp. 135–145.
18. Tsutsumi, S.; Hirabayashi, M.; Sato, D.; Kawatsu, K.; Sato, M.; Kimura, T.; Hashimoto, T.; Abe, M. Data-driven fault detection in a reusable rocket engine using bivariate time-series analysis. *Acta Astronaut.* **2021**, *179*, 685–694. [CrossRef]

19. Flora, J.J.; Auxillia, D.J. Sensor Failure Management in Liquid Rocket Engine using Artificial Neural Network. *J. Sci. Ind. Res.* **2020**, *79*, 1024–1027. [CrossRef]
20. Goodfellow, I.; Pouget-Abadie, J.; Mirza, M.; Xu, B.; Warde-Farley, D.; Ozair, S.; Courville, A.; Bengio, Y. *Generative Adversarial Nets*; MIT Press: Cambridge, MA, USA, 2014.
21. Yoon, J.; Jarrett, D.; Schaar, M.v.d. Time-series Generative Adversarial Networks. In Proceedings of the Neural Information Processing Systems, Vancouver, BC, Canada, 8–14 December 2019.
22. Dixit, S.; Verma, N.K.; Ghosh, A.K. Intelligent Fault Diagnosis of Rotary Machines: Conditional Auxiliary Classifier GAN Coupled With Meta Learning Using Limited Data. *IEEE Trans. Instrum. Meas.* **2021**, *70*, 3517811. [CrossRef]
23. Ren, Z.; Zhu, Y.; Liu, Z.; Feng, K. Few-Shot GAN: Improving the Performance of Intelligent Fault Diagnosis in Severe Data Imbalance. *IEEE Trans. Instrum. Meas.* **2023**, *72*, 3516814. [CrossRef]
24. Deng, L.; Cheng, Y.; Shi, Y. Fault Detection and Diagnosis for Liquid Rocket Engines Based on Long Short-Term Memory and Generative Adversarial Networks. *Aerospace* **2022**, *9*, 399. [CrossRef]
25. Xu, P.; Du, R.; Zhang, Z. Predicting pipeline leakage in petrochemical system through GAN and LSTM. *Knowl.-Based Syst.* **2019**, *175*, 50–61. [CrossRef]
26. Cheng, Y.; Deng, L. Application of Wasserstein distance in fault detection for liquid-propellant rocket engines. *J. Natl. Univ. Def. Technol.* **2023**, *45*, 20–27. [CrossRef]
27. Wang, C.; Zhang, Y.; Zhao, Z.; Chen, X.; Hu, J. Dynamic model-assisted transferable network for liquid rocket engine fault diagnosis using limited fault samples. *Reliab. Eng. Syst. Saf.* **2024**, *243*, 109837. [CrossRef]
28. Miyato, T.; Kataoka, T.; Koyama, M.; Yoshida, Y. Spectral normalization for generative adversarial networks. *arXiv* **2018**, arXiv:1802.05957. [CrossRef]
29. Arjovsky, M.; Chintala, S.; Bottou, L. Wasserstein Generative Adversarial Networks. In *Proceedings of the 34th International Conference on Machine Learning*; Proceedings of Machine Learning Research: Cambridge, MA, USA, 2017; pp. 214–223.
30. Gulrajani, I.; Ahmed, F.; Arjovsky, M.; Dumoulin, V.; Courville, A. Improved Training of Wasserstein GANs. In Proceedings of the 31st Annual Conference on Neural Information Processing Systems (NIPS 2017), Long Beach, CA, USA, 4–9 December 2017.
31. Krizhevsky, A.; Sutskever, I.; Hinton, G. ImageNet Classification with Deep Convolutional Neural Networks. In Proceedings of the 26th Annual Conference on Neural Information Processing Systems (NIPS 2012), Lake Tahoe, NV, USA, 3–6 December 2012.
32. Hou, L.; Yi, H.; Yuhong, J.; Gui, M.; Sui, L.; Zhang, J.; Chen, Y. Inter-shaft Bearing Fault Diagnosis Based on Aero-engine System: A Benchmarking Dataset Study. *J. Dyn. Monit. Diagn.* **2023**, *2*, 228–242. [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.