

Condition monitoring of train transmission systems based on multimodal fusion improved transformer network

Cun Shi¹, Shutong Zhao¹, Xiyang Chen^{2,3}, Shaoping Wang¹, Di Liu^{1,*}

¹ School of Automation Science and Electrical Engineering, Beihang University, Beijing 100191, China

² Beijing Institute of Precision Electromechanical Control, Beijing 100000, China

³ Innovation Center for Control Actuators, Beijing 100000, China

* Corresponding author: Di Liu, liudi54834@buaa.edu.cn

CITATION

Shi C, Zhao S, Chen X, et al.
Condition monitoring of train transmission systems based on multimodal fusion improved transformer network. *Sound & Vibration*. 2025; 59(2): 2904.
<https://doi.org/10.59400/sv2904>

ARTICLE INFO

Received: 6 March 2025

Accepted: 3 April 2025

Available online: 25 April 2025

COPYRIGHT



Copyright © 2025 by author(s).
Sound & Vibration is published by Academic Publishing Pte. Ltd. This work is licensed under the Creative Commons Attribution (CC BY) license.
<https://creativecommons.org/licenses/by/4.0/>

Abstract: The train transmission system is a critical component of railway operations, playing a pivotal role in ensuring service safety and reliability. However, existing condition monitoring approaches face two major challenges: (1) the coupling of rich multimodal signals, such as vibration, acoustics, current, and rotational speed, is often overlooked, limiting monitoring accuracy; (2) the small data problem in multimodal signals adversely affects the performance of neural networks. To address these issues, this paper proposes a Multimodal Fusion Improved Transformer Network for Condition Monitoring of Train Transmission Systems. The proposed network first explores interdependencies among different modalities of signals and compresses data to reduced dimensions through correlation analysis. It then infers global dependencies through computing self-attention scores based on Q , K , and V matrices. The approach is better than traditional CNN-based models in handling single-modality constraints, with the former demonstrated to be more accurate and trustworthy on publicly available datasets.

Keywords: condition monitoring; multi-modal fusion; transformer; self-attention; MFITN

1. Introduction

Railways have become the premier mode of passenger transportation due to their efficiency, economy, and sustainability [1]. Ensuring train operation safety and reliability remains a key priority, particularly with increased operational speeds and intensities introducing new challenges [2]. As a critical component of the train power transmission system that includes traction motors, driving gearboxes, and axle boxes, the transmission system of the train plays a big role in the process of transferring kinetic energy from the motor to the wheelsets [3]. The health status of these components directly affects the operational safety and stability of railway systems [4]. Failure of any of these key elements can lead to severe accidents, involving economic loss and even loss of lives [5]. Thus, the achievement of real-time condition monitoring of train transmission systems is vital to the guarantee of safe railway operations.

With the development of sensor technologies, large-scale monitoring data from railway systems can now be effectively gathered using different sensing methods [6]. Large-scale monitoring data from railway systems can now be efficiently gathered using different sensing methods [7]. Among them, Condition Monitoring is due to their high sensitivity to mechanical defects and operational anomalies [8]. Vibration signals, traditionally used for fault detection, provide rich information on structural integrity, wear progression, and imbalance of components such as wheelsets, gearboxes, and traction motors [9]. In parallel, sound signals pick up acoustic

emissions generated by friction, impacts, and material degradation and provide additional information that is not always evident from vibration signals alone [10]. By integrating both sound and vibration signals, condition monitoring systems are able to provide increased diagnostic accuracy, fault detection at an earlier stage, and improved reliability in rail transport, making them critical to predictive maintenance as well as real-time health checking.

Existing work has verified the effectiveness of data-driven intelligent condition monitoring methods through their robust end-to-end performance [11]. Zhang et al. [12] utilized time-frequency analysis and symbolic recurrent neural networks to introduce a data-driven prognostic system for fuel cells under dynamic load conditions. Tsunashima et al. [13] utilized convolutional neural networks (CNNs) for unsupervised monitoring of rail track conditions, affirming the potential of deep learning techniques in critical railway infrastructure. Zhang et al. [14] designed a fault diagnosis framework based on the Boosting ensemble learning method, which is based on XGBoost, LightGBM, and CatBoost models. Besides, Shim et al. [15] also proposed a combination of both supervised and semi-supervised learning towards condition monitoring systems in the railroad sector. Zou et al. [16] explored fault diagnosis for traction motor bearings in high-speed trains using deep learning approaches, further extending the application of machine learning to train transmission components. In the domain of self-supervised learning, Fu et al. [17] leveraged TFAI for fault diagnosis, demonstrating the potential of self-supervised frameworks to monitor the health status of machinery under limited annotated data conditions. Wang et al. [18] demonstrated the advantages of supervised learning in the study of a novel method based on comparative learning (CL) and the Grampian angular field (GAF) in this method.

Though deep learning-based techniques such as CNNs, hybrids of STFT-CNNs, and DBNs have shown promising performance in railway condition monitoring, there are still some challenges. Davies [19] describes how most techniques care only about handling a single type of signal, say vibration or acoustic data, which can limit their ability to fully monitor the complex interdependence among different elements of the train transmission system. Furthermore, CNN-based models require a great amount of labeled data, which is difficult to obtain in real railway operations [20]. Semi-supervised and anomaly detection approaches relieve this issue to some extent but still struggle to distinguish normal variations from emerging faults accurately [21]. The bogie system of a train comprises multiple interdependent components, including wheelsets, gearboxes, traction motors, and axle bearings, which collectively generate heterogeneous signal modalities with distinct physical attributes. In addition to the previously referenced vibration and acoustic signals, current and rotational speed signals must be systematically considered. Conventional time-domain or frequency-domain analyses prove insufficient for these signals, necessitating feature extraction methodologies. Primary time-domain features encompass Root Mean Square (RMS), variance, Skewness, Kurtosis, and cross-correlation coefficients [22]. Frequency-domain characterization predominantly employs spectral centroid and spectral energy distribution metrics [23]. The multidimensional nature of these extracted features significantly amplifies the complexity of subsequent data processing and analysis tasks [24]. Such complexity induces cross-component dependencies that cannot be

effectively represented by single-modality approaches [25]. Thus, conventional deep learning models may fail to capture significant correlations among sensor inputs, leading to incomplete or suboptimal fault detection performance.

Transformer, a deep learning architecture based on sequence-to-sequence modeling, does well for time series prediction by effectively capturing long-range dependencies and temporal behaviors [26]. Transformer models have proven to be quite helpful in railway condition monitoring by combining various sensor data and improving the fault detection accuracy. For example, Zhang et al. [27] designed a feature fusion Transformer that combines dynamic sensor signals like vibration, acoustic, and temperature data to provide more accurate diagnostics. Ma et al. [28] introduced a self-attention time-frequency Transformer for rolling bearing fault diagnosis that can enable the detection of fine signal variations. With such self-attention capabilities, Transformers can merge diverse sources of signal, enhancing the analysis of system health. Additionally, Transformers' capacity for self-supervised learning, as shown by Wang et al. [29], enables efficient extraction of features from unlabeled datasets, reducing reliance on large sets of labeled data. Transformers work particularly well in identifying faults at an early stage by learning sequential dependencies among sensor data. Ding et al. [30] applied a Transformer model to rolling bearing fault diagnosis, showing its ability for recognizing minimal signal changes. Additionally, Transformers can be trained under varied operating conditions, e.g., varied loads and train speeds, to ensure monitoring stability in different situations, as Yasuda et al. [31] noted.

The parallel processing ability of Transformer models, demonstrated by Ahmed et al. [32], ensures real-time fault detection, thereby making Transformers an efficient and scalable solution for ensuring the reliability of train bogie systems. While Transformer models have been very effective in handling long-distance dependencies and temporal structures in sequential data [33,34], the reliance on large data is a limitation when they are applied in scenarios where data is scarce [35].

In rail networks, particularly for condition monitoring of train bogie, the quantity of labeled data is typically not rich [36], which poses a paradox between the data-hungry nature of Transformers and the scarcity of data in such tasks [37]. Hei et al. [38] proposed a multi-scale transfer learning model with an attention mechanism to address the issue of insufficient Transformer training data by aligning the feature distribution of generated data with that of the target domain. As Transformer models tend to overfit with ease if they are trained on poor data, the traditional approach may not be optimal for fault detection in these small-data cases [39–41]. Multimodal learning has proven to address the small-data challenge by utilizing different sources of information. For instance, Wang et al. [42] presented a feature fusion Transformer that dynamically combines vibration, acoustic, and temperature sensor signals to improve diagnosis accuracy.

Notwithstanding the advances, the issue of integrating Transformer models into small data for train transmission remains, emphasizing the need for tailored approaches that will leverage multimodal data while striving to overcome the limitations that accompany small, labeled datasets. In this context, the present work presents an improved Transformer model based on multimodal fusion, which aims to take advantage of the solutions provided by multimodal data while addressing the

issue of small data through efficient fusion of multimodal data sources and the application of Transformer's self-attention mechanism to enhance the accuracy of train bogie system condition monitoring. This approach tries to bridge the gap between the need for large amounts of data by the Transformer and the limited amount of labeled data present in railway condition monitoring, offering a scalable real-time condition monitoring solution for different operational conditions.

In conclusion, conventional Transformer models can hardly efficiently characterize intricate relationships in multi-modal signals, especially in state monitoring applications. Existing methods easily lose cross-modal dependencies and suffer from a lack of effective redundancy reduction mechanisms. To overcome these limitations, this paper develops an improved Transformer framework with correlation-based multi-modal fusion and adaptive feature integration. In the field of intelligent condition monitoring, obtaining real-world operational data from large-scale physical systems remains a significant challenge [43–46]. Due to the complexity and cost associated with data collection under actual working conditions, most research efforts in neural network-based condition monitoring focus primarily on model construction and validation rather than real-time deployment on actual systems. This limitation necessitates the use of publicly available datasets and simulated environments to evaluate the performance of proposed methods. This paper's thorough experiments on the BJTU-RAO bogie dataset confirm the effectiveness of the proposed approach. The key contributions of this study are as follows:

- 1) **Correlation-Based Multi-Modal Fusion:** A novel correlation-based fusion strategy is proposed to optimize the integration of features extracted from multi-modal signals. By analyzing the interdependencies between channels and features, the framework adaptively weights different modalities to enhance the quality of input data.
- 2) **Transformer Architecture:** The Transformer model is extended with mechanisms to process multi-modal data efficiently. This includes a fusion preprocessing layer that captures both global dependencies through self-attention and local relations through position-wise encoding. Data aren't available due to [ethical/legal/commercial] restrictions.
- 3) **Strong Feature Processing:** Matrix transformations founded upon inter-channel as well as inter-feature relations are applied in order to reduce redundancy and enhance representational strength within the framework. Informative and compact features are assured consequently, facilitating enhanced learning efficiency.
- 4) **Performance Verification:** The proposed framework has state-of-the-art performance, with improved accuracy and robustness. Compared to conventional models, the enhanced Transformer achieves significant improvement in classification accuracy, recall, and F1 scores under different fault conditions and operating conditions.

The structure of this paper is organized as follows: Section 2 details the theoretical foundation of the proposed methods. Section 3 describes the experimental setup, including data preprocessing and parameter tuning. And extensible comparative experiments are conducted to demonstrate the superiority of the proposed framework and analyze the practical application for real railroad systems.

2. Primary methods

2.1. Feature extraction for multimodal signals

Train transmission systems involve complex interactions among various mechanical and electrical components. In this study, we analyze and extract features from the following signals:

- 1) Vibration signals: Three-axis acceleration signals from the motor, gearbox, and train bogies.
- 2) Acoustic signals: Sound signals captured from key components such as the axle box.
- 3) Current signals: Three-phase electrical current from the motor drive.
- 4) Rotational speed signals: Angular velocity of rotating components such as the motor and gearbox.

Time-Domain Feature Extraction:

Time-domain analysis is a fundamental method for characterizing the dynamics of signals in train transmission systems. This study focuses on vibration signals (three-axis accelerations), acoustic signals, current signals, and rotational speed signals. Below are the specific features extracted for each signal type and the associated formulas:

$$\mu = \frac{1}{N} \sum_{i=1}^N a_i \quad (1)$$

$$\text{RMS} = \sqrt{\frac{1}{N} \sum_{i=1}^N a_i^2} \quad (2)$$

$$\text{Skewness} = \frac{\frac{1}{N} \sum_{i=1}^N (a_i - \mu)^3}{\sigma^3} \quad (3)$$

$$\text{Kurtosis} = \frac{\frac{1}{N} \sum_{i=1}^N (a_i - \mu)^4}{\sigma^4} \quad (4)$$

Here, a_i is the i -th sample of the accelerations, acoustic, current and rotational speed signal. μ is the average over this above signal duration. N is the total number of samples. σ is the standard deviation of this above signal. RMS quantifies the magnitude of the signal, highlighting its energy content. Skewness measures signal asymmetry, which is useful for detecting fault-specific characteristics. Kurtosis quantifies the sharpness of signal peaks, distinguishing between normal and abnormal fault patterns.

$$\|a\| = \sqrt{a_x^2 + a_y^2 + a_z^2} \quad (5)$$

where $a_x + a_y + a_z$ are the acceleration components along x, y, z -axes.

Frequency-Domain Feature Extraction:

Using the Fast Fourier Transform (FFT), signals are analyzed in the frequency domain: Spectral Centroid: The frequency ‘‘center of mass’’ of the signal spectrum.

$$f_c = \frac{\sum_{k=1}^M f_k |X_k|}{\sum_{k=1}^M |X_k|} \quad (6)$$

where f_k and X_k denote the frequency and magnitude of the k -th spectral component.

Spectral Energy: Total power of the signal in the frequency domain.

$$E = \sum_{k=1}^M |X_k|^2 \quad (7)$$

Feature Encoding into Matrices:

The extracted features are encoded into normalized matrices for uniform representation across all modalities: Let $F \in \mathbb{R}^{N \times D}$ denote the feature matrix for a single signal, where N is the number of segments, and D is the number of features. Normalization is applied to scale each feature into the range $[0, 1]$:

$$F'_{ij} = \frac{F_{ij} - \min(F_i)}{\max(F_i) - \min(F_i)}, \forall i, j \quad (8)$$

2.2. Fundamental principles of transformer models

Transformer models, as is shown in **Figure 1**, revolutionized sequence modeling by eschewing recurrent structures in favor of self-attention mechanisms [26]. They efficiently capture global dependencies in data and have become the cornerstone of modern deep learning for sequence-to-sequence tasks.

Self-Attention Mechanism:

Here core component of Transformer is the scaled dot-product self-attention. Given a sequence of input vectors $X \in \mathbb{R}^{N \times d}$, the model computes query Q , key K , and value V matrices:

$$Q = XW_Q, \quad K = XW_K, \quad V = XW_V \quad (9)$$

where $W_Q, W_K, W_V \in \mathbb{R}^{d \times d_k}$ are learnable parameter matrices. The attention scores are computed as:

$$\text{Attention}(Q, K, V) = \text{softmax}\left(\frac{QK^T}{\sqrt{d_k}}\right)V \quad (10)$$

where $\sqrt{d_k}$ scales the dot product to stabilize gradients.

Self-Attention Mechanism:

To enhance representational capacity, Transformer employs multi-head attention:

$$\text{MultiHead}(Q, K, V) = \text{Concat}(\text{head}_1, \dots, \text{head}_h)W^O \quad (11)$$

where $\text{head}_i = \text{Attention}(QW_i^Q, KW_i^K, VW_i^V)$, and W^O is a linear projection matrix.

Self-Attention Mechanism: Each layer of the Transformer contains a feed-forward network applied position-wise:

$$\text{FFN}(x) = \text{ReLU}(xW_1 + b_1)W_2 + b_2 \quad (12)$$

This sublayer projects input vectors to a higher-dimensional space, applies a nonlinearity, and maps them back to the original dimension.

Positional Encoding:

$$\text{PE} = \sin\left(\frac{\text{pos}}{10000^{\frac{2i}{d}}}\right) \quad \text{PE} = \cos\left(\frac{\text{pos}}{10000^{\frac{2i}{d}}}\right) \quad (13)$$

where pos is the position, i is the dimension index, and d is the embedding dimension.

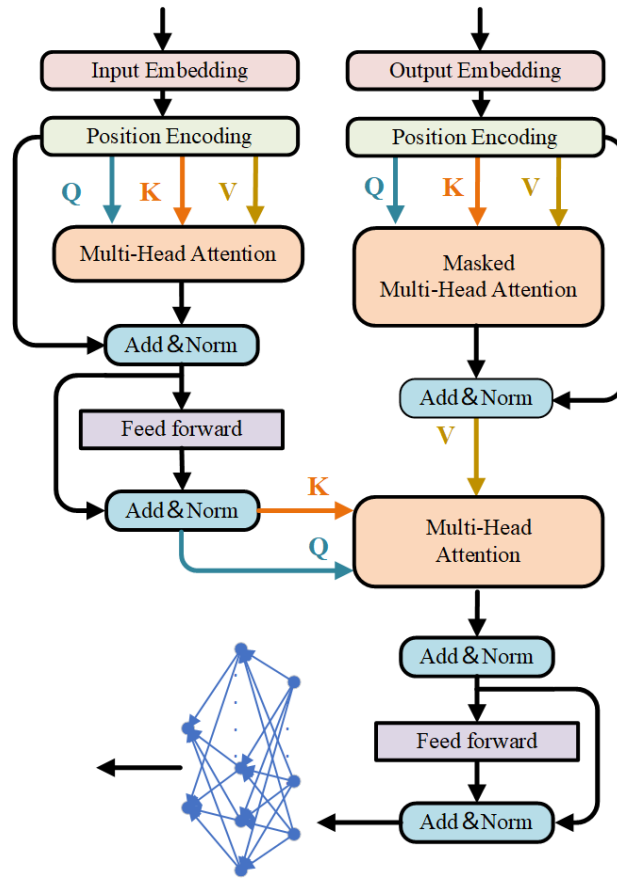


Figure 1. Structure of transformer.

Loss Function for Sequence Learning:

Training is performed using cross-entropy loss, which quantifies the difference between the predicted distribution and the actual distribution.

$$L = -\frac{1}{N} \sum_{i=1}^N \sum_{j=1}^C y_{ij} \log(\hat{y}_{ij}) \quad (14)$$

where y_{ij} and \hat{y}_{ij} represent the true and predicted probabilities for token j in sequence i .

This chapter outlines the methodologies for feature extraction and encoding of multi-modal signals and provides a detailed explanation of the Transformer model's fundamental principles, laying the groundwork for subsequent experiments and applications.

2.3. Multi-modal fusion

The primary novelty of this method is its robust multi-modal data fusion mechanism, which surpasses the shortcomings of typical concatenation or dimensionality reduction techniques. Typical techniques fail to deal with the redundancies and complementary information across modalities, thus their ability to take full advantage of the rich, high-dimensional features contained in multi-modal datasets is not maximized. Conversely, the proposed method uses correlation analysis to quantify and adjust the contribution of a modality based on its correlation with the target variable.

Challenges of multimodal fusion addressed are: Complexity of Multi-Modal Signals: Multidimensional and heterogeneous data require sophisticated techniques to extract useful features without additional noise or redundant information. Inter-modal interactions are generally non-trivial and need to be modeled explicitly in order to best utilize multi-modal information. The method integrates correlation-based feature weighting and transformation as a preprocessing layer, best conditioning the input data for neural network processing. Both univariate and multivariate correlation analysis are used to capture single-modality significance and cross-modality interactions, yielding a complete representation of the underlying data structure.

Univariate Correlation Analysis:

Let X denote the multi-modal input data matrix ($n \times m$, where n is the number of samples and m is the number of modalities or features) and y the target variable ($n \times 1$). The Pearson correlation coefficient $r(X_i, y)$ is computed for each feature X_i as:

$$r(X_i, y) = \frac{\text{Cov}(X_i, y)}{\sigma_{X_i} \times \sigma_y}, \quad i = 1, 2, \dots, m \quad (15)$$

This results in a correlation vector r , which encodes the relevance of each modality to the target.

Multivariate Correlation via Canonical Correlation Analysis (CCA):

To capture the joint influence of multiple modalities on the target variable, canonical correlation analysis (CCA) is employed. The multivariate correlation coefficient R^2 quantifies the combined predictive power of multiple modalities:

$$R^2 = 1 - \frac{\sum_{i=1}^n (y_i - \hat{y}_i)^2}{\sum_{i=1}^n (y_i - \bar{y})^2} \quad (16)$$

where, \hat{y}_i is the prediction from a linear regression model using X , and \bar{y} is the mean of y . R^2 provides a global measure of multi-modality relevance.

Feature Fusion and Dimensionality Reduction:

After determining the significance of each modality, a feature fusion matrix W is constructed, where each weight ω_i is proportional to the corresponding $r(X_i, y)$. The fused data representation is computed as:

$$X_{\text{fused}} = X \times \text{diag}(W) \quad (17)$$

Integration into Neural Network Architectures:

The fused feature representation is then fed into a Transformer-based architecture. By integrating the correlation-based fusion step as a preprocessing block, the Transformer model benefits from high-quality, low-redundancy inputs, which enhances both training stability and predictive performance.

2.4. Framework

The proposed framework is a novel approach to improving the Transformer model with multi-modal data fusion in **Figure 2**. The framework is designed to address challenges in integrating and extracting information from high-dimensional coupled multi-modal signals. The key steps of the framework are:

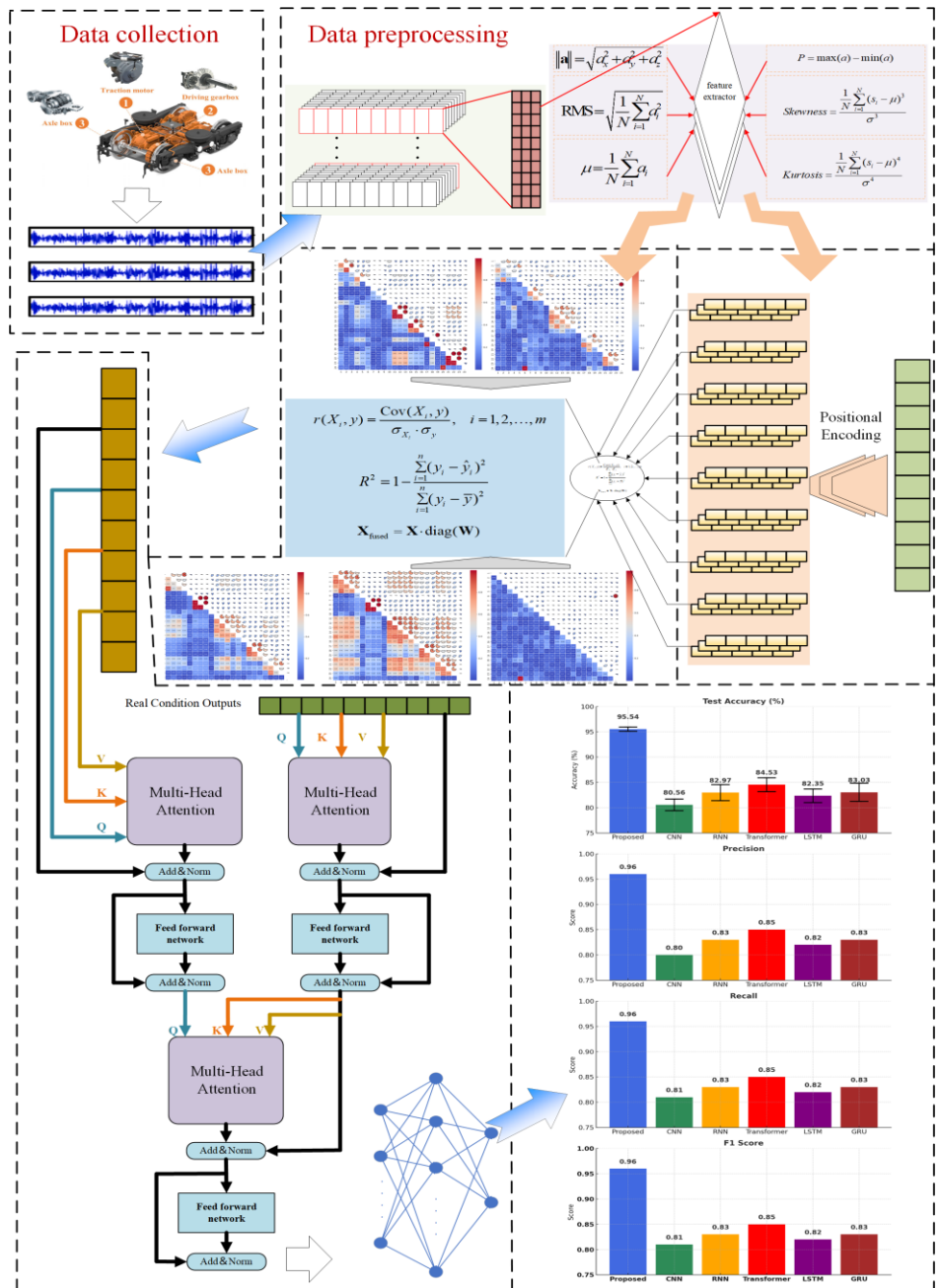


Figure 2. Framework of multimodal fusion improved transformer network.

- 1) **Data Preprocessing:** Multi-channel and multi-modal signals are collected and processed in order to reduce noise and normalize input data. Processing guarantees the raw signals are conditioned for successful feature extraction.
- 2) **Correlation-Based Fusion:** To enhance multi-modal data representation, correlation analysis is employed to evaluate the utility of each modality and their inter-relationships. This involves the analysis of single-variable and multi-variable relationships, which are then used to transform and fuse the feature representations into a unified input matrix.
- 3) **Improved Transformer Model:** The combined representation of data is fed into an improved Transformer model, which involves advanced feature fusion processes to improve learning and forecasting ability.
- 4) **Experiment and Comparison:** Experimental evidence proves the efficiency of the proposed framework against baseline models by highlighting correlation-based fusion policy advantages in producing precise and robust outcomes.

3. Experiments

3.1. Experiment design and data acquisition

The experimental setup utilized a scaled-down (1:2 ratio) metro train bogie transmission system test bench [47], comprising a three-phase asynchronous AC motor, a reduction gearbox (with 16-tooth driving gear and 107-tooth driven gear), and axle boxes. The motor was controlled via a frequency converter, and hydraulic loading simulated operational resistance. Sensor deployment mirrored real-world configurations, including triaxial accelerometers (measuring vibration in g), current sensors (A), tachometers (V), and acoustic sensors (Pa) at critical positions: motor drive/non-drive ends, gearbox input/output shafts, and left/right axle boxes, as is shown in **Table 1**.

The 1:2 scale experimental platform preserved dynamic similarity through proportional motor speeds (20–60 Hz) and lateral loads (0–10 kN), ensuring frequency band distributions (e.g., gear/bearing fault harmonics) aligned with real train systems. Sensor placements mirrored actual metro bogies, capturing equivalent vibration modes (64 kHz sampling resolved scaling-invariant transient features like bearing spalls).

As shown in **Figure 3**, faults were physically induced via component modifications (e.g., seeded gear cracks, motor winding defects) and validated against ISO 26262 severity benchmarks. Fault simulations encompassed 51 health states. (1) Single faults, motor short circuit, gear tooth crack axle bearing inner race failure etc.; (2) Component-level composite faults, concurrent faults within a subsystem; (3) System-level composite faults, cross-component failures. Each fault type was tested under 9 operational conditions to simulate varying speeds and loading scenarios. Dataset reproducibility was ensured through rigorous sensor calibration and cross-verification with physical fault models.

The dataset's controlled fault injection and ISO 26262-aligned severity classifications provide practical utility for algorithm development. These measures confirm that the scaled model's data retains diagnostic fidelity for real-world applications, while providing a cost-effective, risk-free platform for algorithm development.

Table 1. Sensor signal channels.

Channel number	Component	Deployment location	Signal type	Unit
CH 1	Motor	Motor (drive end)	Tri-axial acceleration	g
CH 2				g
CH 3				g
CH 4		Motor (fan end)	Tri-axial acceleration	g
CH 5				g
CH 6				g
CH 7		Motor (cable)	Three-phase current	A
CH 8				A
CH 9				A
CH 10	V			
CH 11	Gearbox	Gearbox (input axle)	Tri-axial acceleration	g
CH 12				g
CH 13				g
CH 14		Gearbox (output axle)	Tri-axial acceleration	g
CH 15				g
CH 16				g
CH 17	Axle box (left)	Axle box (end cover)	Tri-axial acceleration	g
CH 18				g
CH 19		Axle box (adjacency)	Sound	Pa
CH 20				Pa
CH 21	Axle box (right)	Axle box (end cover)	Tri-axial acceleration	g
CH 22				g
CH 23		Axle box (adjacency)	Sound	Pa
CH 24				Pa

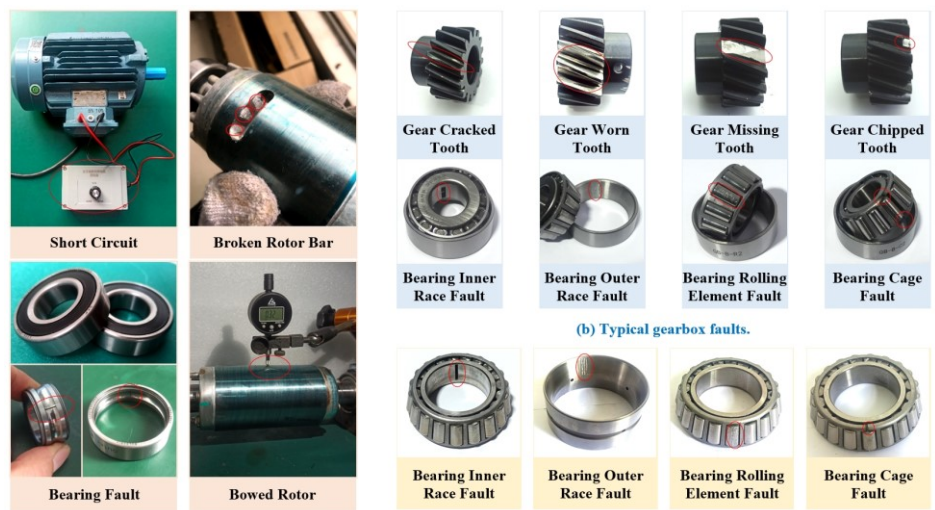


Figure 3. Photographs of fault simulations.

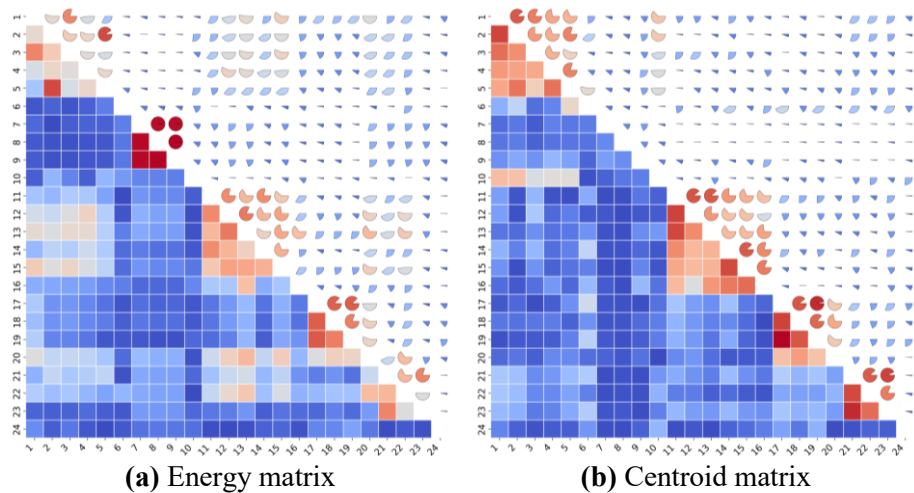
The experimental data employed here are 24-channel measurements captured at a high sampling rate of 64 kHz. These sensors record key information at several

measurement points such as vibration, acoustics, and rotation signals. The experimental setup included 9 working conditions, simulating variations in motor speed and lateral load. The lateral load differences mimic straight-line travel and curved track conditions, while the vertical load was consistently maintained at 10 kN. The experiments introduced a comprehensive array of 51 distinct health states across key components, such as traction motors, gearboxes, and axle boxes. Each health state corresponds to varying levels of mechanical faults and wear conditions, simulating real-world degradation scenarios. For every condition, 10-s recordings were divided into 90 samples, each containing 64,000 sampling points, ensuring sufficient data granularity for analysis.

3.2. Dataset description and processing

The data preprocessing stage focuses on feature extraction and the analysis of correlations among the 24 signal channels. For each channel, six statistical features—mean, root mean square (RMS), skewness, kurtosis, crest factor, and variance—are extracted to capture both the time-domain and statistical characteristics of the signals.

In this paper, the correlation between 24 channels is obtained by calculating the correlation between 24 channels under different feature extraction. The Pearson correlation coefficients among the 24 channels are computed to assess the degree of linear relationships between different channels in **Figure 4**. The calculation results show that there are some differences in the correlation coefficients between the 24 channels under different feature extraction methods. The analysis reveals both strong and weak correlations, highlighting patterns of redundancy and complementarity across the channels. This information serves as the foundation for optimizing the feature integration process and reducing irrelevant or redundant information. Indicating patterns of redundancy and complementarity across the signal channels. Leveraging these insights, the correlation coefficients were calculated to find the correlation between the channels, proving that dimensionality reduction is necessary for data input. The feature integration process of the Transformer network was optimized to reduce irrelevant or redundant information, enhancing the overall feature representation for condition monitoring tasks.



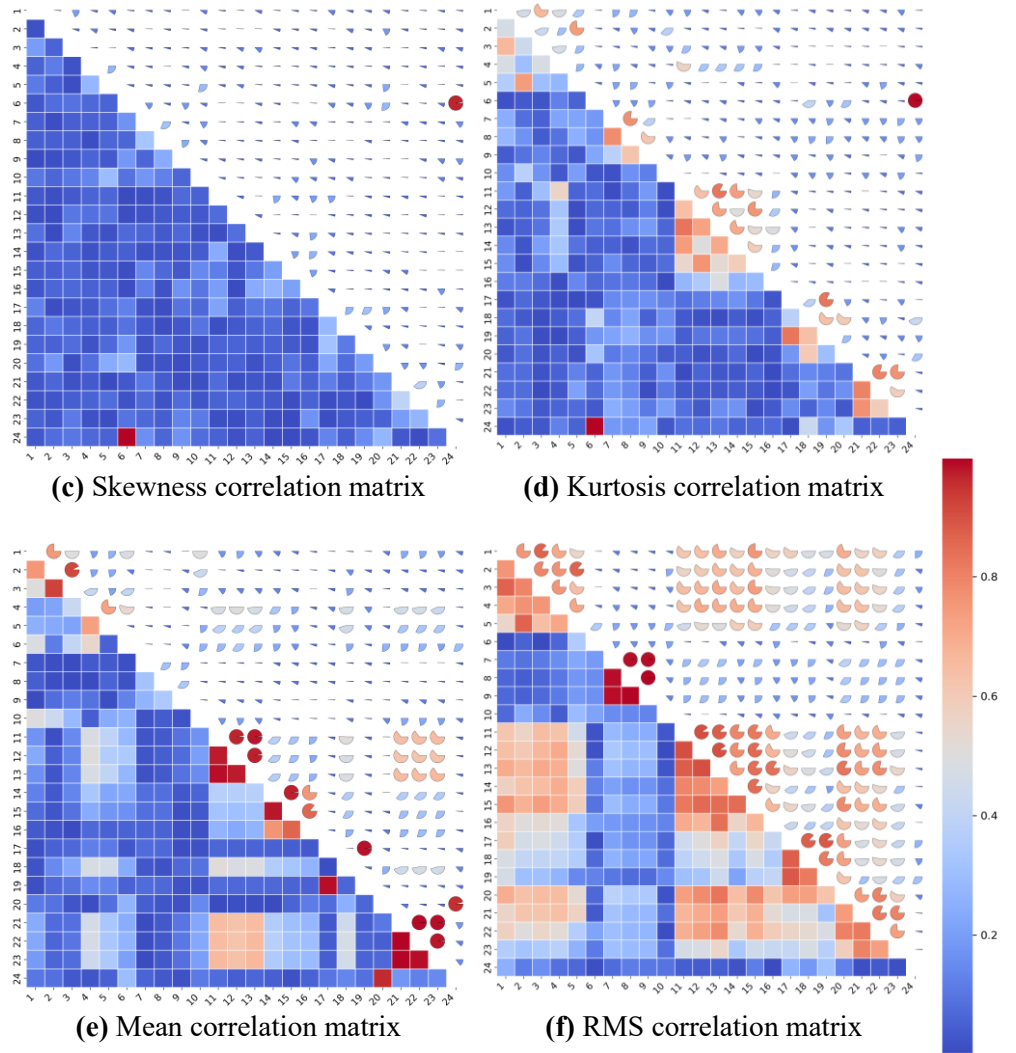


Figure 4. The Pearson correlation coefficients among the 24 channels. **(a)** Pearson correlation of Energy; **(b)** Pearson correlation of Centroid; **(c)** Pearson correlation of Skewness; **(d)** Pearson correlation of Kurtosis; **(e)** Pearson correlation of Mean; **(f)** Pearson correlation of RMS.

An ablation study was performed to assess the effectiveness of the proposed Multimodal Fusion Improved Transformer Network in comparison to the conventional Transformer Network. The evaluation of performance relied on several key metrics. Accuracy (ACC) was utilized to determine the overall correctness of classification results or condition monitoring outcomes. Precision measured the ratio of correctly predicted positive instances among all identified positive cases, highlighting the model’s specificity. Recall, also known as sensitivity, quantified the proportion of actual positive cases that were correctly identified, reflecting the model’s capability in fault detection. The F1 score, as the harmonic mean of precision and recall, provided a comprehensive assessment, particularly beneficial for handling imbalanced datasets.

To evaluate the effectiveness of the proposed Multimodal Fusion Improved Transformer Network (MFITN), we conducted a comparative analysis against three alternative approaches: (1) Manual dimensionality reduction based on multimodal

correlation coefficients (MCC); (2) Unsupervised dimensionality reduction (UDR); and (3) the traditional Transformer model (TTM). The evaluation was conducted using four key metrics: Test Accuracy, Precision, Recall, and F1 Score, with results presented in **Table 2**.

The expected MFITN surpasses all the baselines consistently with the highest Test Accuracy of $95.54\% \pm 0.41\%$, with better Precision (0.96), Recall (0.96), and F1 Score (0.96). The above improvement in performance reveals that the designed method achieves successful modeling and combination of multimodal dependencies, yielding more discriminative feature representations and improved classification performance. In comparison with the base Transformer model which attained a significantly lower Test Accuracy of $84.53\% \pm 1.53\%$, the MFITN enjoys a striking 11.01 percentage points improvement, showcasing the value added through incorporating a multimodal fusion mechanism. Among the baseline methods, manual dimensionality reduction with correlation coefficients achieves a Test Accuracy of $92.16\% \pm 0.61\%$, better than that of unsupervised dimensionality reduction (B) ($89.36\% \pm 1.07\%$) and the base Transformer. This suggests that the use of explicit correlations between modalities is a more informative selection of features than pure data-driven unsupervised methods. Nevertheless, although with a progress, manual selection involves some subjectiveness and does not take advantage of adaptive learning nature captured in the given MFITN. The unsupervised dimension reduction method, although superior to traditional Transformer, lacks significant multimodal interactions and thus is less in Precision (0.89), Recall (0.90), and F1 Score (0.89) compared to MFITN. This suggests data-driven feature compression may miss key multimodal details, limiting it to complex classification tasks. The results of the ablation experiment are shown in **Figure 5**.

Table 2. Test accuracy, precision, recall, F1 Score in the ablation experiment.

	Test accuracy	Precision	Recall	F1 Score
MFITN	$95.54\% \pm 0.41\%$	0.96	0.96	0.96
MCC	$89.36\% \pm 1.07\%$	0.89	0.90	0.89
UDR	$92.16\% \pm 0.61\%$	0.92	0.92	0.92
TTM	$84.53\% \pm 1.53\%$	0.85	0.85	0.85

This experimental result considerably validates the excellence of the Multimodal Fusion Improved Transformer Network, performing best in all of the measures. Multimodal fusion processes implemented in the network not only help in enhanced feature representation but also avoid loss of information typically associated with dimension reduction processes. The results exhibit the strengths of our method in discovering cross-modal relations, increasing classification reliability, and achieving more robust performance in multimodal learning problems.

- 1) Improved Feature Integration: The multimodal fusion method beautifully represented the cross-modal dependency and reduced redundancy to improve feature representation.

- 2) Correlation-Based Dimensionality Reduction: Removing irrelevant and redundant information, the model targeted better and made the feature learning efficient.
- 3) Improved Attention Mechanism: Transformer’s self-attention mechanism was improved using multimodal information and produced enhanced global dependency extraction.

Experimental findings validate the effectiveness of the proposed framework, identifying its capacity to improve condition monitoring of complex train transmission systems.

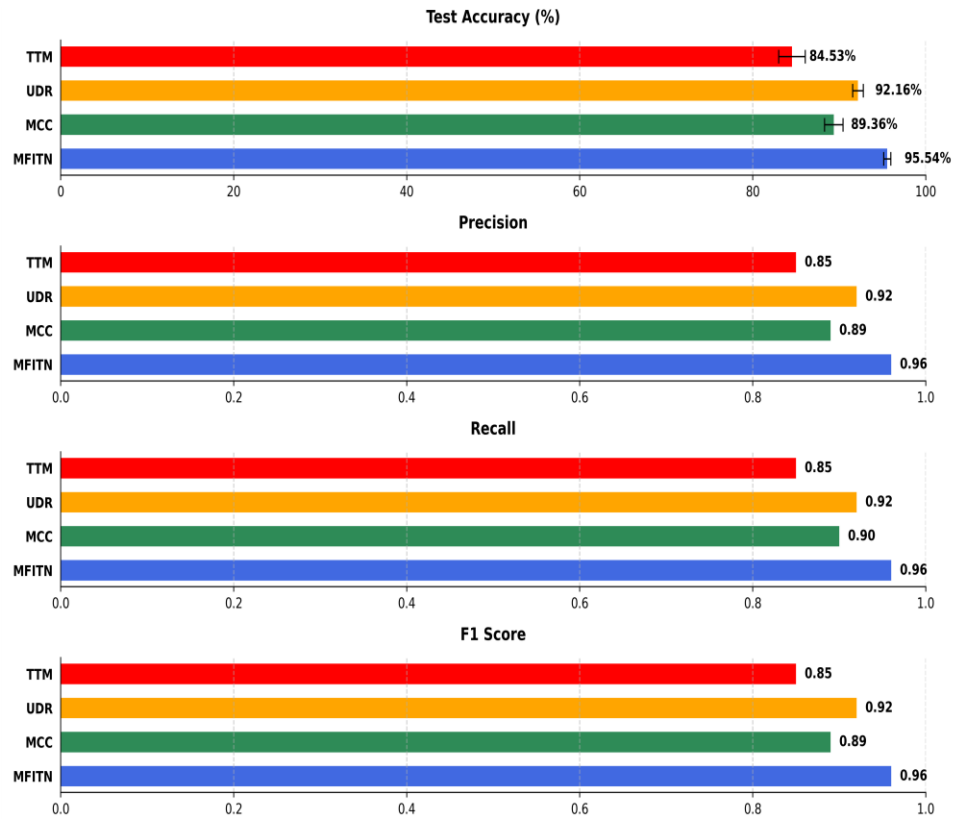


Figure 5. Comparison of different methods.

In this paper, a comprehensive comparison has been performed between the proposed Multimodal Fusion Improved Transformer Network (MFITN) and several baseline models, including Convolutional Neural Networks (CNN), Recurrent Neural Networks (RNN), Transformer, Long Short-Term Memory (LSTM), and Gated Recurrent Units (GRU).

The evaluation metrics are Precision, Recall, F1 Score, and Test Accuracy, with standard deviations considered to quantify result consistency **Table 3**. Experimental results indicate that MFITN performs better than all baseline models on all metrics consistently. Specifically, it achieves a Precision, Recall, and F1 Score of 0.96, which is significantly better than the Transformer (0.85), RNN (0.83), LSTM (0.82), GRU (0.83), and CNN (0.80). Besides, MFITN has a Test Accuracy of 96.16%, which is significantly higher than that of Transformer (84.53%), RNN (82.97%), LSTM (82.35%), GRU (83.03%), and CNN (80.56%). The results of the Comparison experiment are shown in **Figure 6**.

Table 3. The result of the comparative experiment.

	Test Accuracy	Precision	Recall	F1 Score
MFITN	95.54% ± 0.410%	0.96	0.96	0.96
CNN	80.56% ± 1.154%	0.80	0.80	0.80
RNN	82.97% ± 1.581%	0.83	0.83	0.83
TTM	84.53% ± 1.531%	0.85	0.85	0.85
LSTM	82.35% ± 1.339%	0.82	0.82	0.82
GRU	83.03% ± 1.793%	0.83	0.83	0.83

In practical railway operations, real-time train bogie condition monitoring requires the system to meet stringent performance criteria to ensure safety and reliability [48]. Each condition monitoring cycle must be completed within a few seconds to allow timely fault detection and maintenance scheduling. The monitoring system must achieve above 90% accuracy to minimize false positives and false negatives, ensuring effective predictive maintenance. The system must handle continuous real-time monitoring with the ability to process large-scale sensor data without latency issues.

The proposed MFITN model was tested on two configurations: 1) Workstation equipped with an Intel Core i9-13900HX CPU, NVIDIA RTX 4060 GPU, and 32 GB of RAM; 2) PC: equipped with an AMD Core R5-2600 CPU, AMD RX580 GPU, and 16 GB of RAM. The computational time of this paper includes both data processing time and modeling time, data processing can be performed while receipts are collected. Under the two configurations, the data processing time each 10-s sensor data (64 kHz) batches in approximately 10 s and 40 s, and the modeling time of both is less than 1 s. The workstation (i9-13900HX, RTX 4060, 32 GB of RAM) ensuring practical applicability for real-world deployment, as is shown in **Table 4**. In a continuous monitoring scenario, using the workstation configuration, condition evaluations can be performed at fixed intervals, for example, every half minute, allowing for near real-time diagnostics.

If deployed on higher-end hardware configurations, such as an RTX 4090 GPU, Intel Xeon processors, or NVIDIA A100 AI accelerators, computational efficiency would significantly improve. With increased parallelism and optimized tensor processing, the system could achieve faster inference times, making it even more suitable for high-speed railway operations and large-scale fleet monitoring.

Table 4. Comparisons of two configurations.

	CPU	RAM	GPU	Data processing time/s	modeling time/s
Computer 1	i9-13900HX	32G	RTX 4060	10	< 1
Computer 2	R5-2600	16G	RX 580	40	< 1

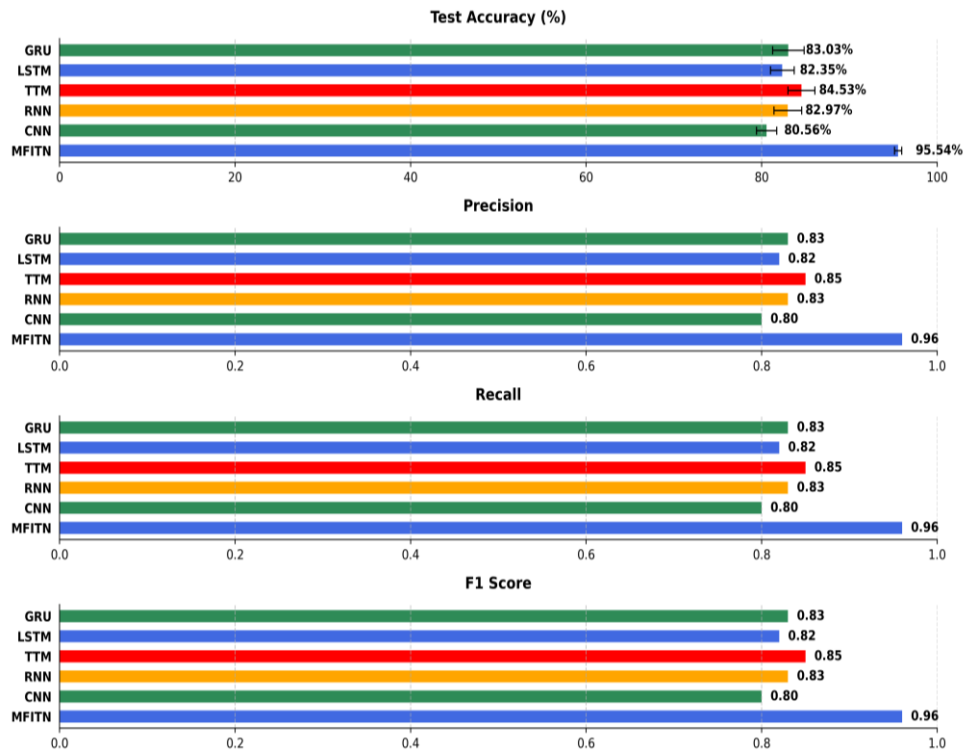


Figure 6. Comparison of accuracy, precision, recall, and F1 score across models.

4. Conclusion

The key design of the Multimodal Fusion Improved Transformer Network (MFITN) is that it includes correlation-based feature selection to reduce redundancy and enhance feature quality, as well as a self-attention mechanism to capture global dependencies across multimodal signals. These advancements enable MFITN to effectively learn complex patterns and interactions among vibration, acoustic, current, and rotational speed signals.

Experimental results demonstrate that MFITN outperforms traditional models, including CNNs, RNNs, LSTMs, GRUs, and even the standard Transformer network, achieving higher accuracy, precision, recall, and F1 score. This confirms its robustness and reliability in real-world railway condition monitoring. The findings highlight the critical role of multimodal fusion in improving predictive maintenance, positioning MFITN as a state-of-the-art solution for future multimodal neural network research.

While the proposed Multimodal Fusion Improved Transformer Network has demonstrated excellent performance on benchmark datasets, its application in real-world engineering scenarios remains to be fully explored. In future work, we will address this limitation by integrating real-world operational data into our framework, enabling a more comprehensive evaluation of the system's performance in actual railway environments.

Author contributions: Writing—original draft, SZ and CS; methodology, CS, SZ and XC; simulation, SZ; validation, CS, SW; supervision and funding acquisition, CS and SW; writing—review and editing, CS and DL. All authors have read and agreed to the published version of the manuscript.

Funding: This study was supported by the Beijing Natural Science Foundation (L221008), National Natural Science Foundation of China (52205046, 52475046), the Aviation Science Foundation (2023M024051001, 2022Z027051001), the Foundation of National Key Laboratory of Aircraft Integrated Flight Control (JSY6142219202402), Ningbo Key R&D Program (2023Z010).

Institutional review board statement: Not applicable.

Informed consent statement: Not applicable.

Availability of data and materials: The data used to support the findings of this study are available from the corresponding author upon request.

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

References

1. World Bank. Bhutan urban policy notes: Urban resilience. Available online: <https://documents1.worldbank.org/curated/en/807961559553043410/pdf/Bhutan-Urban-Policy-Notes-Urban-Resilience.pdf> (accessed on 10 February 2025).
2. Yin J, Tang T, Yang L, et al. Research and development of automatic train operation for railway transportation systems: A survey. *Transportation Research Part C: Emerging Technologies*. 2017; 85: 548–572. doi: 10.1016/j.trc.2017.09.009
3. Jin Y, Chen X. Research on aerodynamic characteristics of high-velocity train bogies. *Journal of Engineering and Applied Science*. 2024; 71(1). doi: 10.1186/s44147-024-00542-3
4. Gavrilovic B, Baboshin VA. Simulations of the operation of the fast light innovative regional train from “Serbian Railways” in traction and electric braking mode. *Mechanical Engineering Advances*. 2023; 2(1): 1214. doi: 10.59400/mea.v2i1.1214
5. Shiao JY, Wang ST. Bogie Stability Control and Management Using Data Driven Analysis Techniques for High-Speed Trains. *Applied Sciences*. 2022; 12(5): 2389. doi: 10.3390/app12052389
6. Bernal E, Spiriyagin M, Cole C. Onboard Condition Monitoring Sensors, Systems and Techniques for Freight Railway Vehicles: A Review. *IEEE Sensors Journal*. 2019; 19(1): 4–24. doi: 10.1109/jsen.2018.2875160
7. He Z, Guo H, Liu H, et al. A Sound Quality Evaluation Method for Vehicle Interior Noise Based on Auditory Loudness Model. *Sound & Vibration*. 2024; 58(1): 47–58. doi: 10.32604/sv.2024.045470
8. Huang W, Xu J. Engineering vibration recognition using CWT-ResNet. *Sound & Vibration*. 2025; 59(1): 2242. doi: 10.59400/sv2242
9. Mousavi SA, Taghipour M. Turbine vibration condition monitoring in region 3. *Mechanical Engineering Advances*. 2023; 1(1). doi: 10.59400/mea.v1i1.219
10. Peng C, Cheng S, Sun M, et al. Prediction of Sound Transmission Loss of Vehicle Floor System Based on 1D-Convolutional Neural Networks. *Sound & Vibration*. 2024; 58(1): 25–46. doi: 10.32604/sv.2024.046940
11. Randall RB. *Vibration-based condition monitoring: Industrial, automotive and aerospace applications*. John Wiley & Sons; 2021.
12. Zhang Y, Tang X, Xu S, et al. Deep Learning-Based State-of-Health Estimation of Proton-Exchange Membrane Fuel Cells under Dynamic Operation Conditions. *Sensors*. 2024; 24(14): 4451. doi: 10.3390/s24144451
13. Tsunashima H, Takikawa M. Monitoring the Condition of Railway Tracks Using a Convolutional Neural Network. *Recent Advances in Wavelet Transforms and Their Applications*; 2022. doi: 10.5772/intechopen.102672
14. Zhang J, Liu M, Deng W, et al. Research on electro-mechanical actuator fault diagnosis based on ensemble learning method. *International Journal of Hydromechanics*. 2024; 7(2): 113–131. doi: 10.1504/ijhm.2024.138231
15. Shim J, Koo J, Park Y. A Methodology of Condition Monitoring System Utilizing Supervised and Semi-Supervised Learning in Railway. *Sensors*. 2023; 23(22): 9075. doi: 10.3390/s23229075
16. Zou Y, Zhang Y, Mao H. Fault diagnosis on the bearing of traction motor in high-speed trains based on deep learning. *Alexandria Engineering Journal*. 2021; 60(1): 1209–1219. doi: 10.1016/j.aej.2020.10.044
17. Fu D, Liu J, Zhong H, et al. A novel self-supervised representation learning framework based on time-frequency alignment and interaction for mechanical fault diagnosis. *Knowledge-Based Systems*. 2024; 295: 111846. doi: 10.1016/j.knosys.2024.111846

18. Wang H, Sun W, Sun W, et al. A novel tool condition monitoring based on Gramian angular field and comparative learning. *International Journal of Hydromechatronics*. 2023; 6(2): 93. doi: 10.1504/ijhm.2023.130510
19. Davies A. *Handbook of condition monitoring: techniques and methodology*. Springer Science & Business Media; 2012.
20. Xu Y, Wang H, Liu Z, et al. Self-Supervised Defect Representation Learning for Label-Limited Rail Surface Defect Detection. *IEEE Sensors Journal*. 2023; 23(23): 29235–29246. doi: 10.1109/jsen.2023.3324668
21. Zhuang L, Qi H, Wang T, et al. A Deep-Learning-Powered Near-Real-Time Detection of Railway Track Major Components: A Two-Stage Computer-Vision-Based Method. *IEEE Internet of Things Journal*. 2022; 9(19): 18806–18816. doi: 10.1109/jiot.2022.3162295
22. Logan D, Mathew J. Using The Correlation Dimension for Vibration Fault Diagnosis of Rolling Element Bearings—I. Basic Concepts. *Mechanical Systems and Signal Processing*. 1996; 10(3): 241–250. doi: 10.1006/mssp.1996.0018
23. Wang WJ, McFadden PD. Early detection of gear failure by vibration analysis i. calculation of the time-frequency distribution. *Mechanical Systems and Signal Processing*. 1993; 7(3): 193–203. doi: 10.1006/mssp.1993.1008
24. Li W, Zhu Z, Jiang F, et al. Fault diagnosis of rotating machinery with a novel statistical feature extraction and evaluation method. *Mechanical Systems and Signal Processing*. 2015; 50-51: 414–426. doi: 10.1016/j.ymsp.2014.05.034
25. Wang C, Dou M, Li Z, et al. Data-driven prognostics based on time-frequency analysis and symbolic recurrent neural network for fuel cells under dynamic load. *Reliability Engineering & System Safety*. 2023; 233: 109123. doi: 10.1016/j.res.2023.109123
26. Vaswani A, Shazeer N, Parmar N, et al. Attention is all you need. *Advances in Neural Information Processing Systems*; 2017.
27. Zhang X, Wang H, Wang C, et al. Time-segment-wise feature fusion transformer for multi-modal fault diagnosis. *Engineering Applications of Artificial Intelligence*. 2024; 138: 109358. doi: 10.1016/j.engappai.2024.109358
28. Ma Y, Wang L, Chen F, et al. DSAN: An Integrated Bearing Diagnosis Strategy with Dual-Spectral Feature Transform and Adaptive Position Correction Algorithm. *IEEE Transactions on Instrumentation and Measurement*. 2025; 74: 1–11. doi: 10.1109/tim.2024.3502804
29. Wang H, Liu Z, Ge Y, et al. Self-supervised signal representation learning for machinery fault diagnosis under limited annotation data. *Knowledge-Based Systems*. 2022; 239: 107978. doi: 10.1016/j.knosys.2021.107978
30. Ding Y, Jia M, Miao Q, et al. A novel time—frequency Transformer based on self-attention mechanism and its application in fault diagnosis of rolling bearings. *Mechanical Systems and Signal Processing*. 2022; 168: 108616. doi: 10.1016/j.ymsp.2021.108616
31. Yasuda M, Ohishi Y, Saito S, et al. Multi-View and Multi-Modal Event Detection Utilizing Transformer-Based Multi-Sensor Fusion. In: *Proceedings of the ICASSP 2022–2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*; 23–27 May 2022; Singapore. doi: 10.1109/icassp43922.2022.9746006
32. Ahmed HOA, Nandi AK. Convolutional-Transformer Model with Long-Range Temporal Dependencies for Bearing Fault Diagnosis Using Vibration Signals. *Machines*. 2023; 11(7): 746. doi: 10.3390/machines11070746
33. Dou B, Zhu Z, Merkurjev E, et al. Machine Learning Methods for Small Data Challenges in Molecular Science. *Chemical Reviews*. 2023; 123(13): 8736–8780. doi: 10.1021/acs.chemrev.3c00189
34. Li C, Li S, Feng Y, et al. Small data challenges for intelligent prognostics and health management: a review. *Artificial Intelligence Review*. 2024; 57(8). doi: 10.1007/s10462-024-10820-4
35. Zhu Q, Sun B, Zhou Y, et al. Sample Augmentation for Intelligent Milling Tool Wear Condition Monitoring Using Numerical Simulation and Generative Adversarial Network. *IEEE Transactions on Instrumentation and Measurement*. 2021; 70: 1–10. doi: 10.1109/tim.2021.3077995
36. Ding A, Qin Y, Wang B, et al. Brownian Distance Covariance-Based Few-Shot Learning Framework Considering Noisy Labels for Fault Diagnosis of Train Transmission Systems. *IEEE Transactions on Industrial Informatics*. 2025; 21(1): 136–145. doi: 10.1109/tii.2024.3441645
37. Tian X, Jin Y, Tang X. Local-Global Transformer Neural Network for temporal action segmentation. *Multimedia Systems*. 2022; 29(2): 615–626. doi: 10.1007/s00530-022-00998-4
38. Hei Z, Sun W, Yang H, et al. Novel domain-adaptive Wasserstein generative adversarial networks for early bearing fault diagnosis under various conditions. *Reliability Engineering & System Safety*. 2025; 257: 110847. doi: 10.1016/j.res.2025.110847
39. Zhou AY, Barati Farimani A. FaultFormer: Pretraining Transformers for Adaptable Bearing Fault Classification. *IEEE Access*. 2024; 12: 70719–70728. doi: 10.1109/access.2024.3399670

40. Tanha J, Abdi Y, Samadi N, et al. Boosting methods for multi-class imbalanced data classification: an experimental review. *Journal of Big Data*. 2020; 7(1). doi: 10.1186/s40537-020-00349-y
41. Peng L, Jian S, Li M, et al. A unified multimodal classification framework based on deep metric learning. *Neural Networks*. 2025; 181: 106747. doi: 10.1016/j.neunet.2024.106747
42. Wang D, Guo X, Tian Y, et al. TETFN: A text enhanced transformer fusion network for multimodal sentiment analysis. *Pattern Recognition*. 2023; 136: 109259. doi: 10.1016/j.patcog.2022.109259
43. Qi GJ, Luo J. Small Data Challenges in Big Data Era: A Survey of Recent Progress on Unsupervised and Semi-Supervised Methods. *IEEE Transactions on Pattern Analysis and Machine Intelligence*. 2022; 44(4): 2168–2187. doi: 10.1109/tpami.2020.3031898
44. He Z, Wang S, Shi J, et al. Physics-informed neural network supported wiener process for degradation modeling and reliability prediction. *Reliability Engineering & System Safety*. 2025; 258: 110906. doi: 10.1016/j.ress.2025.110906
45. Li X, Wan S, Liu S, et al. Bearing fault diagnosis method based on attention mechanism and multilayer fusion network. *ISA Transactions*. 2022; 128: 550–564. doi: 10.1016/j.isatra.2021.11.020
46. Xu Z, Li C, Yang Y. Fault diagnosis of rolling bearings using an Improved Multi-Scale Convolutional Neural Network with Feature Attention mechanism. *ISA Transactions*. 2021; 110: 379–393. doi: 10.1016/j.isatra.2020.10.054
47. Ding A, Qin Y, Wang B, et al. Evolvable graph neural network for system-level incremental fault diagnosis of train transmission systems. *Mechanical Systems and Signal Processing*. 2024; 210: 111175. doi: 10.1016/j.ymsp.2024.111175
48. Fernandez-Bobadilla HA, Martin U. Modern Tendencies in Vehicle-Based Condition Monitoring of the Railway Track. *IEEE Transactions on Instrumentation and Measurement*. 2023; 72: 1–44. doi: 10.1109/tim.2023.3243673