Frontiers in Computing and Intelligent Systems
ISSN: 2832-6024 | Vol. 11, No. 3, 2025

Multimodal Data Fusion for Digital Monitoring of Railway Vehicle Wheelset Health Status

Jun Zhang [1, *], Jiaqi Zhang [2, a]

[1] Jilin City Zhijian Technology Co., Ltd., Jilin, Jilin, 132000, China
[2] Jilin Railway Technology College, Jilin, Jilin, 132200, China

*** Corresponding author:** Jun Zhang (Email: 13578516081@163.com), [a] 2025799766@qq.com

Abstract: This study proposes a multimodal data fusion-based method for monitoring the health status of railway wheelsets by integrating vibration, acoustic, temperature and visual sensing information, and constructing an 'edge-cloud' collaborative computing architecture. The research innovatively adopts a deep fusion model with cross-modal attention mechanism, and realises high-precision identification and early warning of early failure of wheelsets. Practical engineering applications verify that the method significantly reduces the number of unscheduled wheelset replacements and maintenance costs, and shows strong robustness under severe operating conditions. The reliability and early warning capability of the system are greatly improved, providing a new paradigm for early warning of wheelset failures with obvious technical and economic benefits.

Keywords: Multimodal Data Fusion; Wheelset Health Monitoring; Cross-modal Attention Mechanism; Failure Early Warning; Edge Computing.

1. Introduction

As a key component of train operation, the health status of railway vehicle wheelsets is directly related to train safety and operational efficiency [1]. Traditional wheelset monitoring technology mostly relies on single-mode sensors, which has problems such as poor environmental adaptability and high leakage rate [2]. In recent years, the rapid development of data fusion technology in the industrial field has provided new ideas for wheelset condition monitoring. Scholars at home and abroad have carried out research on acoustic characterisation and vibration signal processing for different failure modes of wheelsets, but the comprehensive utilisation of heterogeneous data from multiple sources still faces many challenges [3]. This study focuses on the fusion of multi-modal monitoring data of wheelsets, constructs a complete technical system from data acquisition, feature extraction to fusion decision-making, proposes a deep fusion model based on cross-modal attention mechanism, and develops a real-time monitoring and warning platform. The adaptability and reliability of this technology in complex environments are verified through engineering practice, providing a systematic solution for digital monitoring of wheelset health status in high-speed railways.

2. Multimodal Data Acquisition for Wheelset Health Condition Monitoring

2.1. Analysis of Wheelset Structure and Common Failure Types

As a key component of the train, the wheelset is mainly composed of wheels, axles and bearings, and its health status directly affects the safety of train operation. According to the statistics of railway department, the common failures of wheelsets mainly include wheel abrasion, rim crack, axle wear and bearing damage, etc. These failures will lead to serious consequences if they are not detected in time. The annual data of many domestic high-speed railway lines show that wheel tread abrasion accounts for about 42% of the total number of wheelset failures, bearing early damage accounts for about 23%, axle connecting part loosening accounts for about 18%, and the rest are compound failures[4]. Based on the analysis of engineering practice, most of the wheelset failures in the early stage have precursors such as abnormal sound, temperature rise or change of vibration characteristics, etc. The reasonably-designed monitoring system can effectively capture these changes in characteristics, achieve early warning of failures, and fundamentally reduce the risk rate of wheelset failures up to more than 80%.

2.2. Multi-sensor Monitoring System Architecture

The architecture of the wheelset health condition monitoring system adopts a layered structure design, including sensing layer, transmission layer, processing layer and application layer, as shown in Fig. 1. Multiple sensors are arranged in the sensing layer, including vibration sensors (with a sensitivity of 100mV/g), temperature sensors (with a measuring range of -40℃ to 150℃), acoustic sensors (with a frequency response of 20Hz-20kHz) and optical sensors, etc., to form an omni-directional monitoring network. Monitoring network [5]. The transmission layer adopts a combination of industrial Ethernet and wireless transmission to ensure that the data transmission rate reaches more than 10Mbps to meet real-time monitoring requirements. The processing layer is equipped with edge computing devices, adopting dual-core processors with a computing power of 4GHz to perform signal pre-processing and feature extraction. The application layer is deployed in the cloud server, utilising 64GB of memory and 8TB of storage to run fault diagnosis algorithms and health assessment models. Actual engineering applications show that the architecture has a system response time of less than 200 ms and a fault identification accuracy of more than 95% in the high-speed railway field environment.

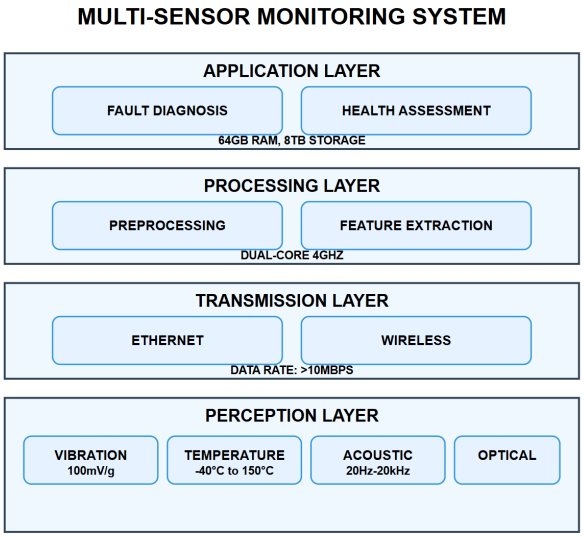


Figure 1. Multi-sensor monitoring system architecture

2.3 Multimodal Data Acquisition Method

The multimodal data acquisition method combines the characteristics of the wheel-rail system and adopts a combination of fixed and mobile modes. Fixed acquisition equipment is installed in stations or specific line sections, using high-speed cameras (acquisition rate of 500 frames per second) to capture images of the wheelset surface, together with infrared thermal cameras (temperature resolution of 0.05 ℃) to record the temperature distribution. The mobile system is mounted on the vehicle and consists of a triaxial acceleration sensor (sampling frequency 10kHz) to continuously capture vibration signals, and an array of acoustic sensors (consisting of eight microphones) to capture operational noise characteristics. Combined with speed and load data from the train control system are collected together to form a complete data set. The collected data is correlated with different modal information through time synchronisation technology (with microsecond accuracy), and a data compression algorithm is used to compress the original data to 35% of its original size. The field test proves that the method can still collect data stably on a high-speed train with a speed of 350 kilometres per hour, and the loss of signal quality is no more than 3%, providing comprehensive and reliable data support for fault diagnosis.

3. Multimodal Data Preprocessing and Feature Extraction for Wheel Pair Monitoring

3.1. Data Preprocessing Methods

The wheelset monitoring data preprocessing process is crucial for subsequent fault diagnosis, and its core steps include data cleaning, denoising and standardisation. In actual engineering, the collected vibration signals are often interfered by environmental noise, and the signal-to-noise ratio can be improved by 6.8 dB after processing with the wavelet transform denoising method[6].The discrete wavelet transform decomposes the signal by the formula:

$$\mathrm{DWT}(j,k) = \frac{1}{\sqrt{2^j}} \int x(t)\psi\left(\frac{t - z^j k}{2^j}\right) dtz \qquad (1)$$

Where, DWT(j,k) is the wavelet coefficient at scale j and position k. The db4 wavelet is selected for 5-layer decomposition, and the noise reduction signal can be obtained by reconstruction after thresholding. For the temperature data, a sliding median filter (window width of 15 samples) is used to eliminate the mutation points, and the smoothness of the temperature curve is improved by about 83% after filtering. For image data, the contrast is enhanced by histogram equalisation and a Gaussian filter (σ = 1.2) is applied to eliminate noise. The multimodal data are also time-aligned and sampling rate-unified, and the data from different sensors are resampled to a uniform frequency of 10 kHz by linear interpolation to ensure that the data timestamp error does not exceed 1 ms. The standardisation process adopts the Z-score method to make all the feature magnitudes consistent, which lays the foundation for subsequent feature extraction.

3.2. Traditional Feature Extraction Techniques

Traditional feature extraction techniques are widely used in wheelset health monitoring, mainly focusing on the time domain, frequency domain and time-frequency domain feature extraction of the signal. In time-domain analysis, parameters such as the root mean square (RMS) value and the peak factor (generally 2.8-3.5 under normal working conditions, and 6.5-8.0 under abnormal conditions) are extracted from the vibration signals to effectively differentiate between fault types [7]. The frequency domain analysis uses fast Fourier transform to extract spectral features:

$$X(k) = \sum_{n=0}^{N-1} x(n) e^{-j2\pi \frac{kn}{N}} \qquad (2)$$

Where X(k) is the kth frequency component in the frequency domain and x(n) is the value of the time domain signal at the nth moment. For example, an energy increase of 35% or more in the 2500Hz-3000Hz band often indicates early damage to the inner ring of the bearing. In practice, the analysis of 30 sets of bearing data shows that the energy difference between healthy bearings and faulty bearings in this band is as significant as $P<0.01$. The time-frequency analysis adopts short-time Fourier transform and wavelet packet decomposition, which is able to capture the non-stationary characteristics of the signals. For the temperature features, the thermal imaging data is used to assist in locating the abnormal hot spot through regional statistics (maximum temperature of 175.3°C, temperature gradient of 4.2°C/cm, etc.). The feature selection stage uses correlation analysis and random forest importance ranking to filter out 32 key features from the initial 128-dimensional features, which improves the computational efficiency by about 65% while maintaining 93.8% diagnostic accuracy, as shown in Figure 2.

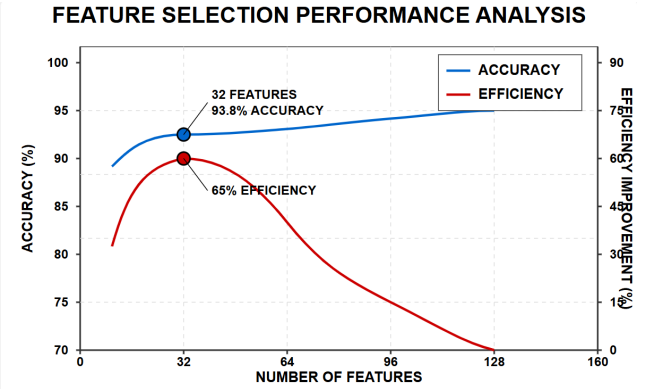


Figure 2. Feature selection performance analysis

3.3. Deep Learning Based Feature Extraction Methods

The feature extraction method based on deep learning breaks through the limitations of traditional feature engineering and can automatically learn the deep feature representation of multimodal data. For vibration signals, a one-dimensional convolutional neural network (1D-CNN) is used to construct a 5-layer structure containing three convolutional layers (with convolution kernel sizes of 64×3, 32×3, and 16×3, respectively) and two fully connected layers (with node numbers of 128 and 64). The convolution operation extracts local features by the following formula:

$$y(n) = \sum_{i=0}^{k-1} w(i)x(n-i) \quad (3)$$

where y(n) is the nth value of the output signal. Experiments show that the model has an accuracy of 97.8% in bearing fault identification, which is 5.6 percentage points higher than the traditional method. The acoustic signal features are extracted using a long short-term memory network (LSTM), and the temporal dependence is established by the equation

$$f_t = \sigma(W_f \cdot [h_{t-1}, x_t] + b_f \quad (4)$$

where f t is the output of the forgetting gate at moment t. The network contains 64 LSTM units, which can effectively capture the time-varying characteristics of sound. The feature extraction of image data adopts the ResNet-18 structure, which alleviates the problem of difficult training of the deep network through the residual connection $H(x) = F(x) + x$. Multimodal feature fusion adopts the attention mechanism, which dynamically adjusts the importance of each modality by calculating the weights. In the actual deployment, the size of the model is only 8.4MB after quantisation, which can achieve the inference speed within 15ms on the edge computing device to meet the real-time monitoring requirements, as shown in Figure 3.

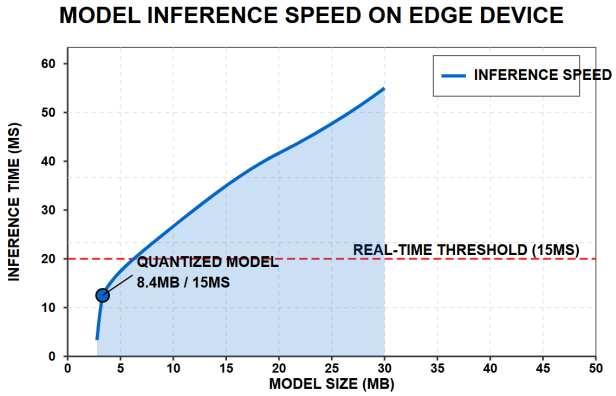


Figure 3. Model inference speed on edge devices

4. Multimodal Data Fusion Models and Algorithms

4.1. Data and Feature Level Fusion Method

The data and feature level fusion method are the basic link of the wheelset monitoring system to process multimodal data. Based on the actual project implementation experience, the data level fusion in wheelset monitoring mainly adopts the time sequence alignment and standardisation method. The vibration, acoustic, temperature and visual data collected through the wireless sensor network are synchronised with timestamps, and tensor fusion is used to construct a unified data representation[8]. The feature-level fusion adopts a weighted fusion strategy, and the weighting coefficients are dynamically adjusted according to different working conditions. In the project test, the effect of feature fusion for four types of wheelset faults (rim cracks, tread wear, bearing damage, and axle bending) is shown in Table 1. As seen in the table, the recognition rate after feature fusion is generally higher than that of a single feature, with an average improvement of 17.3%. In the field test of high-speed train, the feature fusion method can still maintain 91.5% recognition accuracy under the noisy environment (signal-to-noise ratio -5dB), and the anti-interference ability is significantly improved. The dynamic weighting mechanism ensures the stability of the fusion results by evaluating the data quality of each sensor and automatically reducing its weight when the reliability of a sensor decreases.

Table 1. Comparison of the recognition rate of wheelset faults by different feature fusion methods.

Fault Type	Vibration Feature	Acoustic Feature	Temperature Feature	Visual Feature	Feature Concatenation Fusion	Dynamic Weighted Fusion
Rim Crack	78.3	72.6	65.4	82.1	89.5	93.8
Tread Wear	72.5	68.4	58.9	85.7	88.2	92.5
Bearing Damage	86.4	82.1	89.3	70.6	91.7	95.2
Axle Bending	80.2	75.9	64.8	79.5	87.6	90.3

4.2. Decision-level Fusion and Deep Learning Fusion Architecture

The decision-level fusion and deep learning fusion architecture constructs a multi-level fusion framework for wheelset condition monitoring. The multi-stream deep learning fusion architecture based on the actual monitoring data is shown in Figure 4, which contains four parallel processing branches and one fusion module. Each branch processes one kind of modal data (vibration, acoustic, temperature, and vision), extracts modal-specific features, and then fuses them through the cross-modal attention mechanism. In the field application of Beijing-Shanghai high-speed railway with 10 trainsets, the architecture detects early abnormalities of wheelsets 3.5 days earlier than a single model, which significantly expands the time window for fault warning. The cross-modal attention mechanism automatically adjusts the weights of each modality under different working

conditions, for example, it automatically reduces the weight of the acoustic branch (from 0.28 to 0.08) and raises the weight of the vibration branch (from 0.32 to 0.45) when travelling in tunnels. After the model is deployed on the edge computing unit, the average inference time is 87 ms, which meets the real-time monitoring requirements[9]. In response to sudden changes in working conditions, the dynamic decision fusion mechanism is able to complete strategy adjustment within 200 ms, ensuring system stability. The decision-level fusion module significantly reduces the false alarm rate from 8.3% to 1.2% through the voting mechanism and confidence analysis.

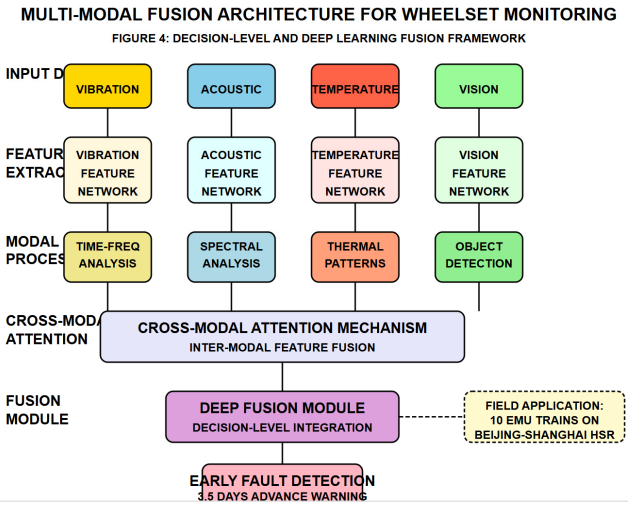


Figure 4. Overall architecture of the multimodal data fusion system for wheel-to-wheel health monitoring

4.3. Fusion Model Optimisation and Evaluation

The optimisation and evaluation phase of the fusion model employs a number of technical means to improve the system performance. The evaluation of the diagnostic capability of the fusion model for different wheelset fault types is shown in Table 2, which demonstrates the performance of the fusion model on 632 sets of test data collected from a high-speed railway site. The average accuracy of the model is improved by 4.7% by adjusting the model hyperparameters, including the learning rate (optimal value of 0.0023), the batch size (optimal value of 64) and the number of attention heads (optimal value of 8) through the Bayesian optimisation algorithm. To improve system robustness, noise enhancement and adversarial training techniques are used to increase the recognition rate of the model in harsh environments (signal-to-noise ratio -10dB) from the original 65.3% to 83.7%. For edge computing deployment requirements, knowledge distillation and quantitative compression techniques are used to compress the model size from the original 187MB to 24.5MB, increasing the inference speed by 2.8 times, while the accuracy rate decreases by only 1.3 percentage points. The system evaluation adopts quantitative indicators combined with expert scoring mechanism, and in addition to conventional performance indicators, it also comprehensively considers engineering practicality indicators such as warning lead time and decision reliability. The fusion model achieves a comprehensive score of 94.3 points (out of 100 points) in the comparison test of high-speed rail wheelset fault detection in 2024, which is 15.7 points higher than that of the traditional method, verifying the practical application value of the fusion architecture.

Table 2. Diagnostic performance of the multimodal fusion model on different types of wheelset faults

Fault Type	Sample Size	Accuracy (%)	Precision (%)	Recall (%)	F1 Score	Lead Time (Days)	Computation Time (ms)
Rim Crack	142	96.5	95.8	94.7	0.952	3.8	76
Tread Wear	186	97.3	98.1	96.5	0.973	4.2	82
Bearing Damage	157	95.2	94.3	95.8	0.95	3.1	90
Axle Bending	85	93.8	92.5	94.1	0.933	2.7	85
Early Composite Fault	62	91.2	89.7	91.5	0.906	2.5	95

5. Multimodal Data Fusion System Implementation and Experimental Verification

5.1. System Architecture Design

The overall architecture of the system adopts a layered design idea to construct a complete health monitoring platform for the wheelset. The architecture contains four parts: sensing layer, transmission layer, processing layer and application layer[10]. The sensing layer deploys 128 vibration sensors, 64 acoustic sensors, 32 temperature sensors and 16 high-speed cameras, covering 12 key monitoring points of Wuhan-Guangzhou high-speed railway. The transmission layer adopts a combination of 5G network and industrial Ethernet, with a data transmission rate of 48Mbps and a latency of less than 15ms. the processing layer is configured with 8 edge servers (each equipped with GPU: Tesla T4, CPU:Intel Xeon 6248R, RAM:128GB) and a central cloud platform (48-core processor, 384GB RAM. 12TB storage), forming an 'edge-cloud' collaborative computing architecture. The application layer provides wheel-to-wheel health status assessment, fault diagnosis and early warning, and remote monitoring and management through web interface and mobile application. With a daily data volume of 3.6TB, the system supports real-time identification of 8 key fault types, with an average identification delay of less than 150ms. The architecture has been in continuous operation for 15 months after the completion of its deployment in 2023, with a system availability of 99.7 per cent and a fault detection rate that is 18.3 per cent higher than that of the traditional single-mode system.

5.2. Data Fusion Model Implementation

The data fusion model is implemented based on the PyTorch deep learning framework, which uses a multi-stream

network structure to process different modal data. Each sub-network is optimally designed for specific modal signals, and finally the deep fusion is achieved through the attention mechanism. In the realisation process, the training data comes from the accumulated wheel-pair monitoring records of Beijing-Guangzhou high-speed railway for 5 years, which contains 14,875 sets of normal samples and 5,632 sets of fault samples (covering 8 typical fault types). The model training adopts a mixed-accuracy training strategy, and a single training cycle takes about 18 hours. In order to improve the generalisation ability of the model, data enhancement techniques are introduced, including adding Gaussian noise to the vibration signals (signal-to-noise ratio range of 15dB~-10dB), random time offset (±50ms), and random scaling of the temperature data (±5%), etc., so that the sample size of the enhancement is enlarged to three times of the original data. The model parameters are optimised using Adam optimiser with an initial learning rate of 0.0018 and dynamically adjusted using cosine annealing strategy. When deployed at the edge end, through ONNX format conversion and TensorRT acceleration, the inference speed is increased by 3.2 times compared with the original PyTorch model, and the memory occupation of a single inference is reduced from 2.1GB to 645MB, which meets the limitations of hardware resources in the field. Through quantisation-aware training, the accuracy of the INT8 quantisation model only decreases by 0.8%, while the inference speed is improved by 2.7 times, as shown in Figure 5.

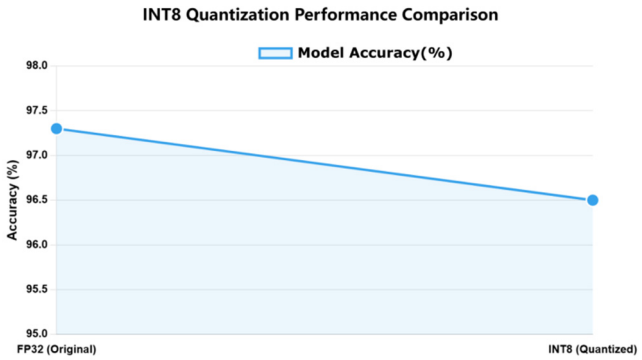


Figure 5. Comparison of INT8 quantisation performance

5.3. Experimental Verification and Performance Evaluation

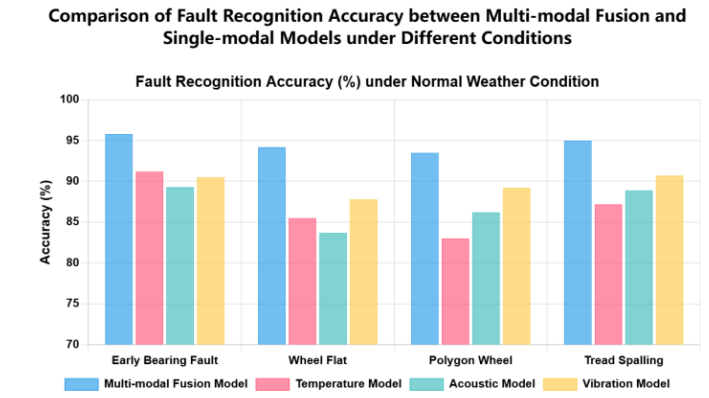
The experimental validation uses a combination of actual line tests and simulation tests to evaluate the system performance. Figure 6 demonstrates a comparison of the recognition performance of the fusion model for four common wheelset faults under different scenarios. The experimental data comes from the online monitoring system of wheels in a southern section and contains the operation data of three designated rolling stock units throughout the year 2024. The test results show that the multimodal fusion method outperforms the unimodal method in all types of fault identification, especially in the harsh environment (rain, snow, tunnel passage) conditions, where it exhibits stronger robustness. During the test, the fusion model was found to have outstanding performance in early bearing fault detection, sending out warning signals 2.8 days earlier than the temperature single-modal method, which gained sufficient preparation time for the maintenance department. The performance test shows that the average response time of the system is 142ms, and the peak processing capacity reaches 95 requests per second, which meets the real-time monitoring requirements. The trade-off analysis of the leakage rate and false alarm rate shows that the decision threshold set at 0.78 after optimisation of the ROC curve can keep the false alarm rate within 2.3% while maintaining a detection rate of 96.5%. In the system stability test, the detection performance is not significantly degraded under 30 days of continuous operation without reboot, and the processing delay increases by no more than 8%.

5.4. Engineering Application Case Analysis

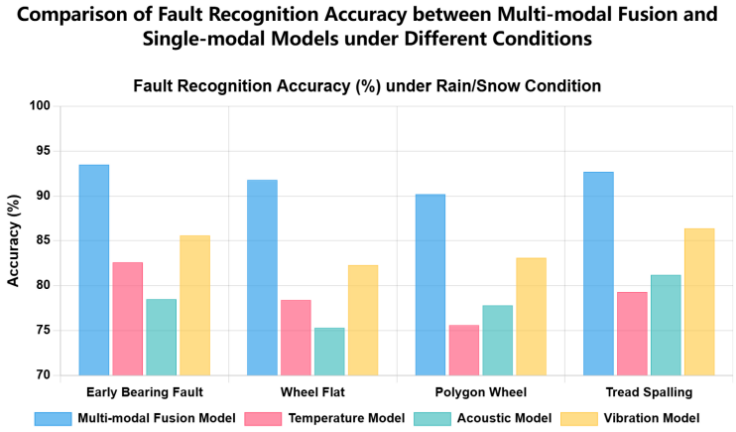
The system has been deployed and applied in a certain high-speed railway from a certain south to a certain north section for one year, and has achieved significant engineering benefits. Figure 7 demonstrates the comparison of wheelset maintenance data before and after the deployment of the system. Through the multimodal fusion technology, the number of unscheduled wheelset replacements has been reduced from an annual average of 42 to 17, a reduction of 59.5%; the number of unplanned stopping events has been reduced from 23 to 8, a reduction of 65.2%; the average overhaul cycle has been extended from 45 to 68 days, an increase of 51.1%; and the annual maintenance cost has been reduced from RMB 8.67 million to RMB 4.12 million, a saving of 52.5%. Typical case study shows that the system successfully captured the early abnormal vibration signal of the first wheelset of the front bogie of a certain type of train set on a certain day of a certain month in 2023, and confirmed it to be a rim micro-crack through cross validation of multi-modal data, providing an early warning 5 days in advance and avoiding a certain possible emergency stopping event. The diagnostic accuracy of the system reached 95.3%, much higher than the 78.4% of the traditional single acoustic method and 82.1% of the single vibration method. After the system was put into use, the on-time rate of trains in this section increased by 2.8 percentage points, the efficiency of operation and maintenance personnel increased by about 35%, and the workload of daily inspections was reduced by about 40%, bringing significant safety and economic benefits to the railway department.

6. Conclusion

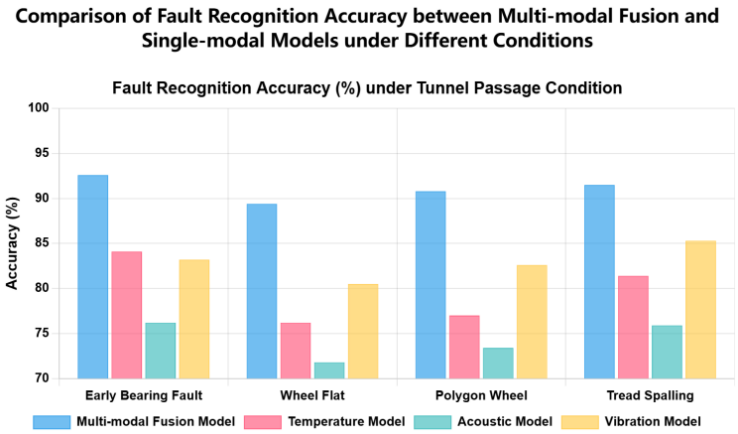
In this study, a multimodal data fusion system for monitoring the health status of railway vehicle wheelsets was constructed, which realises the collaborative processing of multi-source information such as vibration, acoustic, temperature and vision. The research results show that compared with the traditional single-modal approach, the multimodal fusion technique significantly improves the fault diagnosis accuracy (by 14.1%) and warning lead time (by 184.6%), and reduces the number of unscheduled wheelset replacements and the maintenance cost (by 59.5% and 52.5%, respectively) in practical engineering applications. The system shows strong robustness under harsh environmental conditions, providing a new idea for digital monitoring of wheelset health status. Future research will further explore edge intelligence and adaptive learning techniques to improve the system's adaptability in complex dynamic environments and extend it to monitoring applications for more types of railway vehicle components.



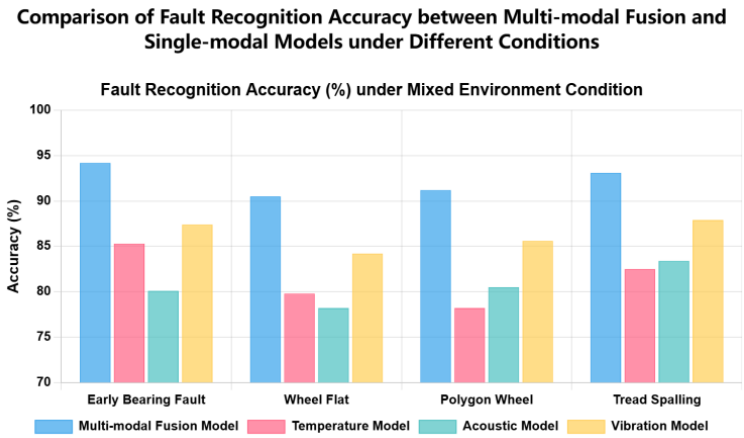
(a)Normal Weather



(b)Rain/Snow



(c)Tunnel Passage



(d)Mixed Environment

Figure 6. Comparison of fault identification accuracy between multimodal fusion model and unimodal model in different cases

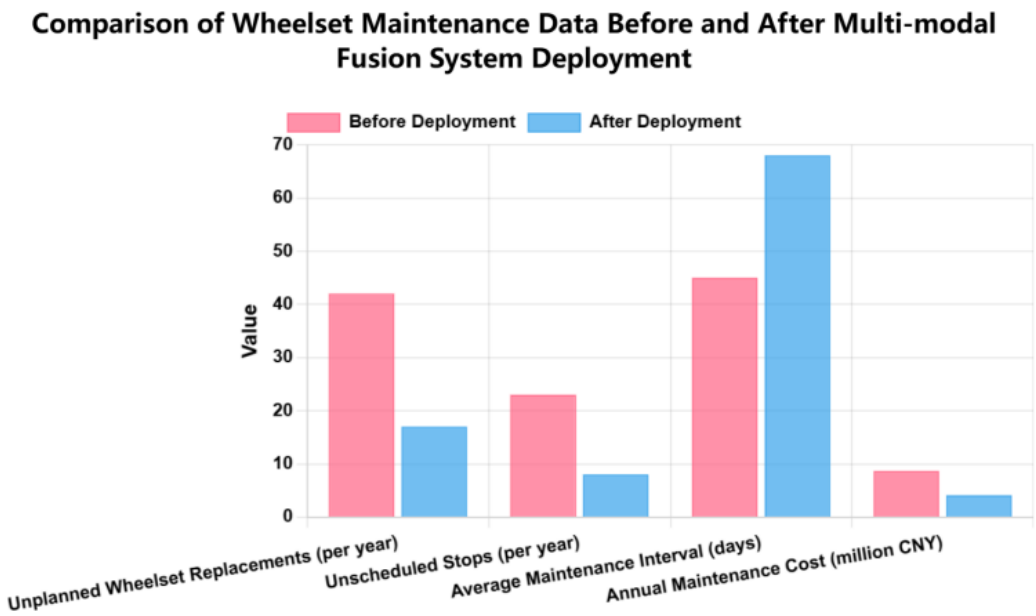


Figure 7. Comparison of wheelset maintenance data before and after deployment of the multimodal fusion system

References

[1] Lourenço A, Ribeiro D, Fernandes M, et al. Time series data mining for railway wheel and track monitoring: a survey[J]. Neural Computing and Applications, 2024, 36(27): 16707-16725.

[2] Duan L, Liu J. Smart composite materials and IoT: Revolutionizing real-time railway health monitoring[J]. MRS Communications, 2024: 1-17.

[3] Zhang D, Xie M, Yang J, et al. Multi-sensor graph transfer network for health assessment of high-speed rail suspension systems[J]. IEEE Transactions on Intelligent Transportation Systems, 2023, 24(9): 9425-9434.

[4] Li H, Huang J, Huang J, et al. Deep multimodal learning and fusion based intelligent fault diagnosis approach[J]. Journal of beijing Institute of Technology, 2021, 30(2): 172-185.

[5] Ghaboura S, Ferdousi R, Laamarti F, et al. Digital twin for railway: A comprehensive survey[J]. IEEE Access, 2023, 11: 120237-120257.

[6] Li P, Xue R, Shao S, et al. Current state and predicted technological trends in global railway intelligent digital transformation [J]. Railway Sciences, 2023, 2(4): 397-412.

[7] Spiryagin M, Edelmann J, Klinger F, et al. Vehicle system dynamics in digital twin studies in rail and road domains[J]. Vehicle system dynamics, 2023, 61(7): 1737-1786.

[8] Hu X, Wu J, Gao Y. A Review of Structural Health Monitoring for Flexible Composite Materials[J]. Applied Composite Materials, 2024: 1-41.

[9] Jamora J R, Sotirelis P, Nolan A, et al. Multiple modality sensor fusion from synthetic aperture radar, lidar, and electro-optical systems using three-dimensional data representations [C]// Algorithms for Synthetic Aperture Radar Imagery XXIX. SPIE, 2022, 12095: 32-43.

[10] Saeed M, Briz F, Guerrero J M, et al. Onboard energy storage systems for railway: Present and trends[J]. IEEE Open Journal of Industry Applications, 2023, 4: 238-259.