International Journal for Housing Science and Its Applications
Publish August 10, 2025. Volume 46, Issue 4 Pages 1804-1820

https://doi.org/10.70517/ijhsa464157

# Research on the Full Life Cycle Management and Intelligent Detection and Early Warning System of PHM Technology for EMU Wheel Sets Based on Big Data Analysis

**Jun Ma[1,*], Xu Xue[2] and Bingzhi Chen[1]**
[1] School of Mechanical Engineering, Dalian Jiaotong University, Dalian, Liaoning, 116028, China
[2] School of Electrical Engineering, Dalian Jiaotong University, Dalian, Liaoning, 116028, China
Corresponding authors: (e-mail: 15904951689@163.com).

**Abstract** To improve the effectiveness of fault prediction and health management for high-speed train wheels, this paper proposes a data processing framework based on Spark Streaming and Kafka to collect, clean, and transform high-speed train PHM source data. Based on the data processing, correlation algorithms were used to identify the influencing factors of wheel set wear. Considering the complexity of changes in high-speed train wheel size data influenced by operational environments and other factors, a wheel size prediction model based on VMD-PSO-MKELM was constructed to achieve accurate prediction of high-speed train wheel set data. In wheel diameter data, the MSE, MAE, and MAPE of the VMD-PSO-MKELM model in this paper are 0.0012, 0.0294, and 0.0004%, respectively, with R2 reaching 0.9968; For flange thickness data, the corresponding values are 0.0081, 0.0741, and 0.0005% for MSE, MAE, and MAPE, respectively, with an $R^2$ of 0.9251. Whether in wheel diameter data or flange thickness data, the MSE, MAE, and MAPE of the VMD-PSO-MKELM model in this paper are lower than those of the compared ELM, L-ELM, P-ELM, and R-ELM models, and the $R^2$ is the highest, demonstrating high prediction accuracy and greater practicality.

**Index Terms** PHM, correlation algorithm, multi-kernel extreme learning machine, PSO algorithm

## I. Introduction

With the continuous opening of new high-speed rail lines, China's high-speed rail operational mileage has also been steadily increasing. By the end of 2024, China's railway operational mileage reached 162,000 kilometers, with 48,000 kilometers of high-speed rail. It is projected that by the end of 2035, high-speed rail mileage will exceed 70,000 kilometers [1]. With breakthroughs in China's high-speed train technology, operational speeds have continued to increase, and operational scale has grown significantly. However, this has also brought about increasingly prominent challenges in ensuring train safety [2], [3]. To ensure train operational safety, China's high-speed trains currently adopt a maintenance model combining scheduled maintenance and fault-based maintenance, which has to some extent ensured wheel operational safety. However, this model does not offer advantages in terms of human and material resource costs for wheel set maintenance or in terms of intelligence [4]-[6]. Unlike conventional trains, high-speed trains are composed of powered locomotives and equipment-carrying trailers arranged in different traction units, with each subsystem having clear responsibilities and functions. Data between systems is interconnected via train-level MVB, car-level WTB, and Ethernet, giving high-speed trains characteristics such as systematization, proceduralization, specialization, and intensification that conventional trains lack [7]-[9]. The wheelsets of EMUs consist of axles, hubs, and rims, bearing 80% of the dynamic loads of rail transit vehicles [10]. Currently, China's high-speed rail operates at speeds as high as 350 km/h. Accidents such as axle breakage or wheel failure could result in catastrophic consequences [11]. According to incomplete statistics, 65% of the total operating costs of high-speed rail are spent on maintenance, parts replacement, and vehicle depreciation. Therefore, regular inspection of wheel sets is a critical safeguard for operational safety, service quality, and economic efficiency.

PHM, short for Predictive Maintenance and Health Management, is a core technology in the industrial field used to monitor equipment operating conditions [12]. This system uses sensors to collect real-time key parameters such as equipment vibration, temperature, and pressure, akin to installing a 24/7 health monitor for machinery. It can detect internal abnormalities invisible to the naked eye and is widely applied in fields such as aerospace, shipbuilding, energy, and power systems [13]-[15]. As early as the late 1990s, fault prediction technology was applied to the JSF project in the U.S. military aviation industry, marking the official birth of PHM technology [16]. When large machine tools in factories experience bearing wear, PHM can issue a warning three weeks in

advance with an accuracy rate of 92%, preventing sudden shutdowns of the entire production line [17]. The accumulation of data has made PHM increasingly intelligent. In the stamping workshop of an automobile factory, the PHM system records 150,000 stamping data points from each press over a 20-year period. When a new piece of equipment exhibits a 0.03-millimeter mold offset, the system immediately retrieves similar case studies from its database, making decisions 17 times faster than human judgment [18].

To ensure train safety, comprehensive inspection and management of high-speed train wheel sets are essential. Literature [19] introduces a quality management platform for the reliability, availability, and maintainability of high-speed train wheel sets throughout their lifecycle. This platform is based on a lifecycle quality management system for critical components, which helps extend the service life and economic benefits of high-speed train wheels. With the development of intelligent technology, the inspection and warning systems for high-speed train wheel sets have become intelligent. Literature [20] proposes a defect monitoring and identification data collection system for high-speed train wheel sets, which is formed through robotics technology and ultrasonic data collection, processing, and identification technology. It achieves automated monitoring, automated collection of defect data, automatic warning, intelligent control, intelligent management decision-making, data management, and resource sharing. Therefore, utilizing high-speed train wheel set operational data for their full lifecycle management and inspection and warning is of critical importance. Currently, due to the difficulty in unifying data across high-speed rail management groups and the increasing proportion of unstructured data, data governance challenges have arisen, necessitating the introduction of big data analysis. Big data analysis handles an extremely large volume of data, far exceeding the scope of traditional data processing, aiming to help us better understand the underlying meaning of these data and make more informed decisions [21].

In the railway system, Reference [22] developed an integrated PHM platform combining vehicle, communication, and ground data using PHM technology, integrating multi-source heterogeneous data from high-speed EMUs in various scenarios, and introduced artificial intelligence to establish a traction motor fault prediction model. Reference [23] utilized PHM technology and proactive maintenance technology to construct a maintenance framework for the traction power supply system of high-speed railways, which applied a large amount of online sensor data and offline test data, and applied this framework to case studies. Literature [24] combines big data technology, hidden Markov models, deep belief networks, and PHM technology to monitor, diagnose, and predict the status of critical components in high-speed trains, and also proposes maintenance prediction and decision-making technologies.

The wheels of high-speed trains are critical components that play a crucial role in the vehicle's load-bearing, guidance, and traction braking. This paper first combines relevant big data technology to propose a data processing framework based on Spark Streaming and Kafka, establishing a data processing architecture for high-speed train PHM. This framework enables the collection, cleaning, and transformation of high-speed train PHM source data, supports online processing of streaming data, and meets the data requirements for model calculations. Based on the completed data processing, correlation algorithms were further utilized to identify the influencing factors of high-speed train wheel wear, and strongly correlated influencing parameters were extracted as input parameters for the high-speed train wheel dimension prediction model. To address the issue that the single kernel function of KELM struggles to adapt to the diverse data features of high-speed train wheel samples, multiple kernel functions are weighted to form a multi-kernel extreme learning machine (MKELM), establishing a high-speed train wheel size prediction model based on VMD-PSO-MKELM. The original high-speed train wheel size sequence was decomposed into sub-sequences with better regularity using VMD, and the numerous parameters of the mixed kernel function in the model were optimized using the PSO algorithm to obtain the optimal parameters. Using wheel diameter and flange thickness data, and selecting models such as ELM, L-ELM, P-ELM, and R-ELM as comparisons, the effectiveness and practicality of the proposed VMD-PSO-MKELM model were verified. Finally, combining the proposed high-speed train wheel size prediction method, the application of PHM technology in high-speed train wheel intelligent detection and early warning is explored, and a high-speed train wheel intelligent detection and early warning system is constructed.

## II. PHM Data Processing Framework for EMUs

With the rapid development of high-speed rail, EMUs have become the primary mode of transportation for railways, with the number of EMUs in service increasing annually in tandem with growing demand for transport capacity. Fault prediction and health management (PHM) for EMUs are crucial tools for implementing reforms in maintenance schedules and systems. To address the data requirements for the full lifecycle management of wheel PHM technology for EMUs, a data processing framework based on Spark Streaming and Kafka is proposed to enable the collection, cleaning, and transformation of PHM source data for EMUs [25].

### *II. A. Analysis of PHM source data for EMUs*

#### II. A. 1) Characteristics of source data

The PHM data for EMUs is sourced from multiple EMU maintenance systems, primarily the EMU Management Information System (EMIS) and the Train Safety Monitoring System. including data on new vehicle and component manufacturing from original equipment manufacturers (OEMs) and parts suppliers, EMU allocation information, operational information, maintenance information, fault information, wireless transmission of EMU onboard information regarding train operational status and fault information, fault monitoring information from the EMU operational fault image monitoring system, as well as data from signaling, civil engineering, weather, and GIS systems. These data packages are characterized by large volumes, multi-source heterogeneity, and the coexistence of bounded and unbounded data.

#### II. A. 2) Source Data Classification

High-speed train PHM data sources are diverse in origin, type, and volume. Proper data classification is crucial for managing and processing high-speed train PHM source data. The following classification of high-speed train PHM data sources is based on an analysis of high-speed train PHM source data.

1) Classification by data type: High-speed train PHM source data can be broadly categorized into three types based on data type: text-based, numerical, and time-based.

2) Classification by data structure: Data can be classified into three categories based on data structure: structured data, semi-structured data, and unstructured data.

3) Classification by data granularity: Data can be classified into two categories based on data granularity: detailed data and summary data.

4) Classification by real-time nature: Data can be classified into two categories based on real-time nature: batch data and real-time data.

### *II. B. PHM Data Processing Framework for EMUs*

This paper applies big data technology to construct a PHM data processing framework for high-speed trains to effectively solve core issues in PHM data processing, such as streaming data processing and parallel computing.

#### II. B. 1) Data Processing Architecture

The data processing architecture diagram for the EMU PHM system is shown in Figure 1, which includes the data source layer, data collection layer, data processing layer, data storage layer, and data analysis layer.

1) Data source layer. This layer primarily includes the source data systems and operational equipment required to support the implementation of EMU PHM functions.

2) Data Collection Layer. This layer defines the specific business data required by the high-speed train PHM system and classifies the data based on its structure following data analysis. The classified data is then subjected to ETL extraction strategies based on its update cycle.

3) Data Storage Layer. This layer employs a distributed multi-type database to accommodate the large-scale storage and application requirements of various data types.

4) Data processing layer. This layer provides batch processing methods for offline data processing and real-time processing methods for streaming data processing. Batch processing methods are mainly used for data cleaning, format conversion, and data aggregation from different dimensions. Real-time data processing methods use big data-related streaming data processing tools to achieve real-time computing and query processing functions.

5) Data Analysis Layer. This layer utilizes data analysis methods such as analytical algorithms, deep learning, and neural networks to perform data mining and analysis. It also combines the failure mechanisms of high-speed train components to construct fault models for these components, and continuously optimizes these models through iterative training to improve the accuracy of the computational results.

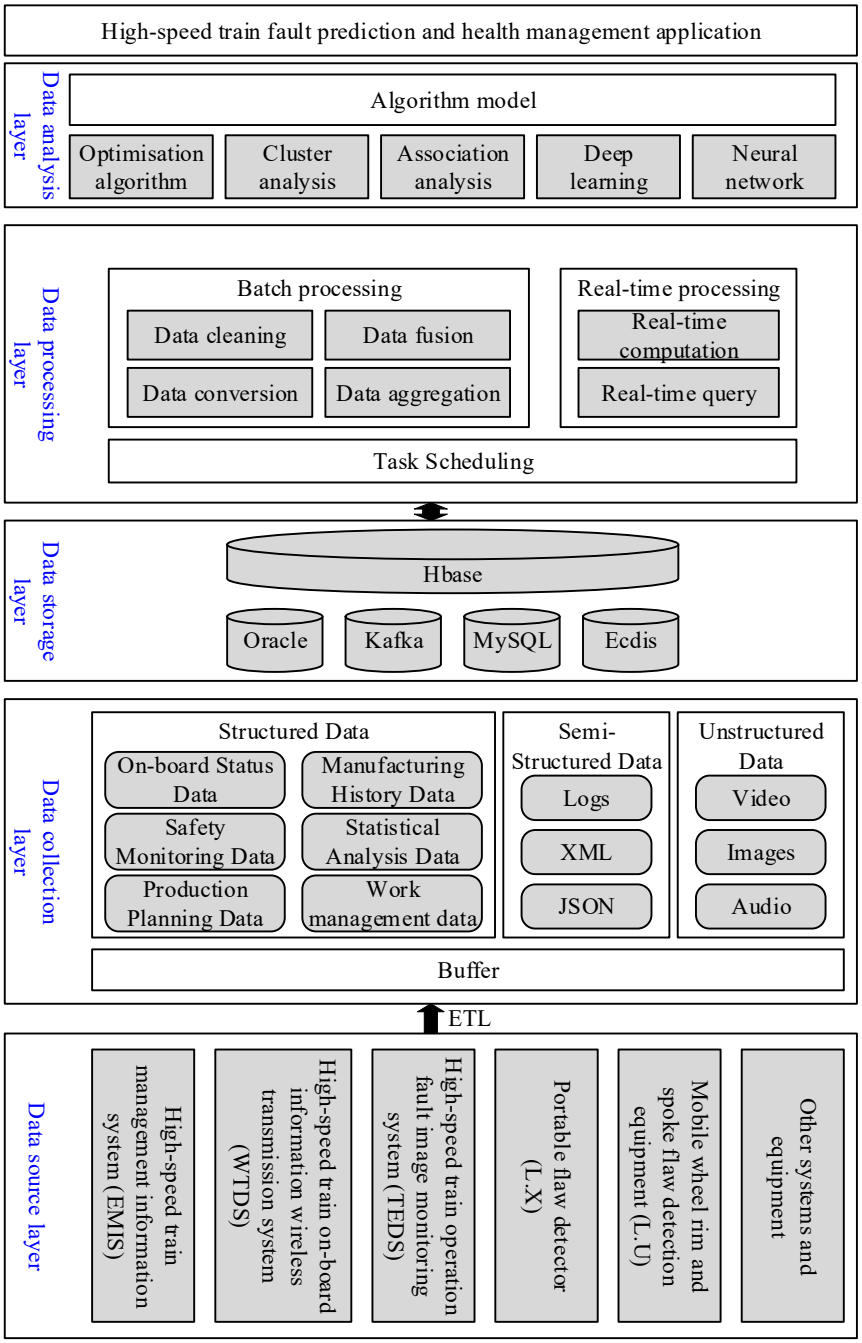



Figure 1: Data processing architecture of EMU PHM

**II. B. 2) Streaming Computing Components**

To meet the requirements of real-time processing of streaming data and online operation of PHM application models, this paper adopts Kafka and Spark Streaming as the core components of the high-speed train PHM data processing framework.

After onboard sensors collect relevant sensor data from components, the WTDS system pushes the collected data via ActiveMQ to the Kafka server in the high-speed train PHM system, which is responsible for receiving external network data. Since the WTDS system encrypts the transmitted data, the high-speed train PHM system cannot use the data normally. Therefore, the data must be decrypted before use. Spark Streaming acts as the consumer of WTDS data in the Kafka server, decrypting the WTDS data retrieved from the broker according to the decryption rules. After processing, the data is split into a data header and a data body according to certain rules. On one hand, Streamsets categorizes and persists the complete WTDS data in HBase based on the information in the data header. On the other hand, to achieve fast response times for querying new data, the split data header

is stored in Redis. On the other hand, the parsed data can be cleaned and transformed using Spark Streaming according to business needs. The processed data is divided into a new topic and stored in the Kafka server to provide data services for the high-speed train PHM system business and models.

#### II. B. 3) Data processing methods

The data processing section provides methods for collecting, cleaning, and storing PHM data from high-speed trains.

1) Data collection

Data collection is divided into offline data collection and streaming data collection based on the real-time nature of the data. The following sections will introduce these two collection methods separately.

(1) Offline Data Collection

In this paper, Sqoop is selected as the extraction tool. Based on the target data update mechanism and the principle of minimizing impact on the business system, the following data extraction methods are established: For tables or views being extracted for the first time, a full extraction is performed without affecting the normal operation of the business system. For small-sized tables or views with data updates, the incremental data is achieved by first deleting and then performing a full extraction. For large-sized tables or views with data updates, an appropriate incremental extraction method is selected based on the specific field design of the table or view, the business system, and the database used.

(2) Real-time data collection

Real-time data is pushed to the Kafka server used for data reception by calling system interfaces or as a service. The newly received data is persisted by the collection tool Streamsets using different storage mechanisms based on the user's data usage requirements. If the user has processing requirements for real-time data, Spark Streaming can be used to parse or compute the data.

2) Data Cleaning

This paper proposes solutions for issues such as data duplication, missing data, and noise in the source data.

(1) Duplicate Data Removal: Based on data analysis, BloomFilter is considered the most suitable method for duplicate data removal in the high-speed train PHM system.

(2) Missing Data Imputation: The EM algorithm is selected as the method for imputing missing data in the high-speed train PHM source data.

(3) Noise Data Processing: The chi-square binning method is chosen as the noise smoothing method for the high-speed train PHM source data.

3) Data Storage

To meet the requirements for rapid response in storing, managing, and querying large volumes of data, the data processing framework adopts HBase, a highly reliable, high-performance, and scalable distributed data storage system, as the primary data storage method. The source data collected from various data source business systems is uniformly stored in HBase after data processing.

## III. Full life cycle management of PHM technology for EMU wheel sets

During the operation of high-speed train wheel sets, various types of wheel set wear, such as radial wear and flange thickness wear, have a significant impact on the smooth and safe operation of trains. These factors must be given priority consideration in the full life cycle management of trains. This chapter will analyze the current status and influencing factors of condition prediction and health management for high-speed trains based on the PHM data processing framework for high-speed trains, using correlation algorithms to determine the influencing factors of wheel set wear [26].

### *III. A. Analysis of factors affecting wheel set wear based on correlation algorithms*

From the start of operation, wheel sets on EMUs undergo wear throughout their entire service life until they are scrapped. Since the wear rate of wheel sets varies greatly under different conditions, determining the parameters that affect wheel set wear is a prerequisite for accurately predicting wheel set wear. Given the uncertainty of the correlation between parameters, linear and nonlinear correlation algorithms are used to calculate correlation coefficients and extract highly correlated influencing parameters to obtain a sample set for the training model.

The formula for calculating the Pearson correlation coefficient $r_p$ is as follows:

$$r_p = \frac{\sum_{i=1}^{n}(y_i - \bar{y})(b_i - \bar{b})}{\sqrt{\sum_{i=1}^{n}(y_i - \bar{y})^2 \sum_{i=1}^{n}(b_i - \bar{b})^2}} \tag{1}$$

where, $y_i$ and $b_i$ are the actual values of the two relevant parameters, $\bar{y}$ and $\bar{b}$ are the sample means of the two relevant parameters, $n$ is the sample size.

The value range of $r_p$ is $[-1,1]$. If $|r_p|$ is closer to 1, it indicates that the linear correlation between wheel set wear and the detection parameter is higher.

The Spearman algorithm is a method for calculating non-linear correlation coefficients. The formula for calculating the Spearman coefficient $r_s$ is as follows:

$$r_s = \frac{\sum_{i=1}^{n}(m_i - \bar{m})(s_i - \bar{s})}{\sqrt{\sum_{i=1}^{n}(m_i - \bar{m})^2 (s_i - \bar{s})^2}} \tag{2}$$

where, $m_i$ and $s_i$ are the actual values of the two relevant parameters, $\bar{m}$ and $\bar{s}$ are the sample means of the two relevant parameters.

Calculate $r_p$ and $r_s$ using the two algorithms, and assign them the corresponding weights $p$ and $q$. In this case, $p$ is set to 0.6 and $q$ is set to 0.4. Then:

$$r_{ab} = pr_p + qr_s \tag{3}$$

In the formula:

$r_{ab}$ ——total correlation coefficient.

### *III. B. Analysis of the correlation between EMU wheel health and performance*

This paper will apply correlation algorithms to historical operational data of a specific model of high-speed train to calculate the correlation coefficients between train operational data, acceleration, mileage, fresh air temperature, track characteristics, and axle box temperature. Through correlation analysis, the key factors influencing the health status of high-speed train wheels will be identified.

The correlation coefficients between gearbox temperature and train operational status parameters are specifically shown in Table 1. As can be seen from the correlation coefficients between gearbox temperature and vehicle operational and track parameters in the figure, the factor most highly correlated with gearbox temperature is fresh air temperature, which has a very high influence weight. The higher the fresh air temperature, the higher the gearbox temperature. Speed and mileage also have certain positive influence weights on gearbox temperature; the higher the speed and mileage, the higher the gearbox temperature.

Table 1: Correlation coefficient

| Parameters | Correlation coefficient |
|---|---|
| Speed | 0.1451 |
| Acceleration | 0.0243 |
| Mileage | 0.2701 |
| Fresh air temperature | 0.7993 |
| Brake level | -0.0383 |

To conduct an in-depth analysis of the impact of slopes on the bearing temperatures of high-speed trains under different track conditions and operational scenarios, data from the Wuhan-Guangzhou High-Speed Railway between January 2023 and July 2024 was selected. Three bearing component temperatures were chosen: the temperature of the wheel side of the first axle gearbox, the temperature of the first axle box, and the temperature of the stator of the first motor on the first axle. The average temperature values for each component were calculated. The calculation results are shown in Figure 2. Under slope conditions, the bearing temperature is higher than under non-slope conditions, indicating that slope affects bearing temperature.

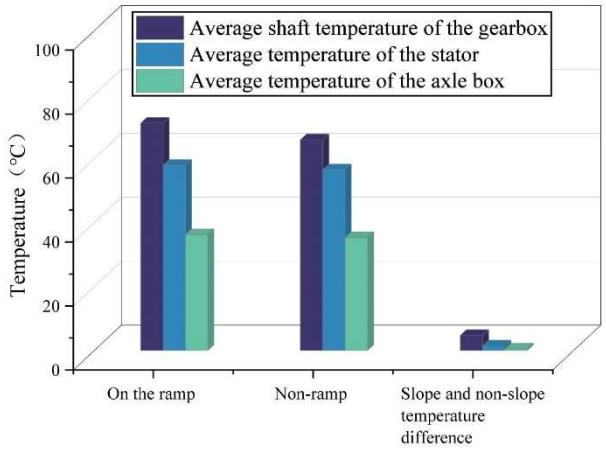
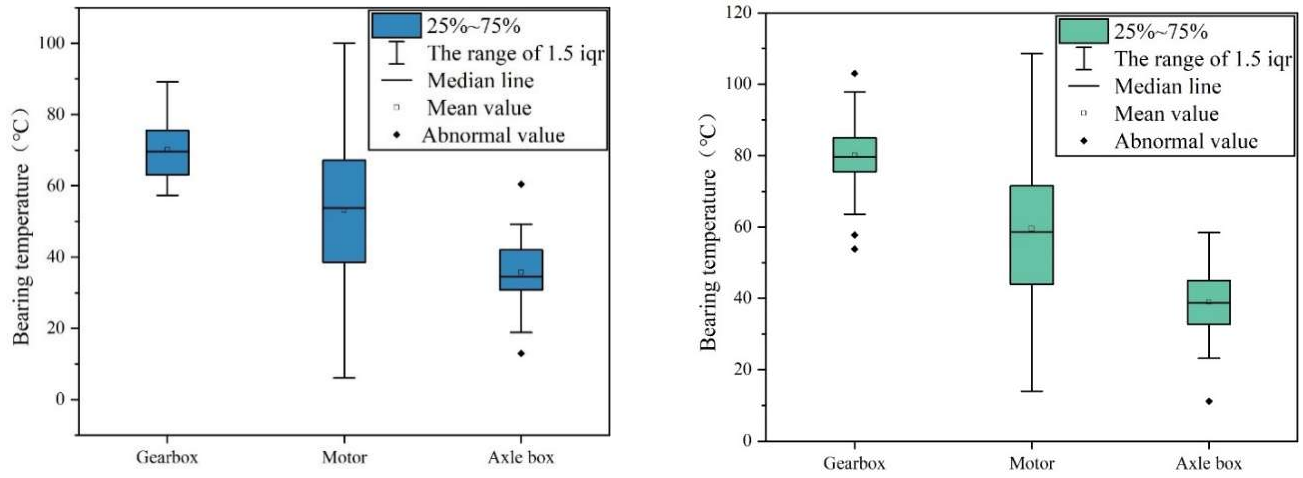



Figure 2: Comparison of bearing temperature under ramp or not

The temperature distribution of the gearbox, motor, and axle box bearings under non-slope and slope conditions is shown in Figure 3, with Figures (a) and (b) corresponding to non-slope and slope conditions, respectively. As can be seen, the average and median bearing temperatures on slopes are higher than those on flat surfaces. It can be concluded that slopes affect bearing temperatures, with temperatures on slopes being higher than those on flat surfaces.




(a) Non-ramp (b) Ramp

Figure 3: Comparison of bearing temperature distribution with or without ramp

**III. B. 1) Comparison and analysis of bearing temperatures under ramp grading**

The distribution of gradients along the Wuhan-Guangzhou Railway Line is shown in Figure 4. The gradients along the Wuhan-Guangzhou Railway Line range from -15 to 15, with a relatively dispersed distribution, indicating that the overall gradient along the Wuhan-Guangzhou Railway Line is relatively high and not primarily centered around 0.

The average temperatures of the gearbox wheel side under four conditions—steep downhill, gentle downhill, gentle uphill, and steep uphill—are shown in Table 2. As shown in the table, the gearbox wheel side temperature is approximately half a degree Celsius higher on steep uphill and steep downhill slopes compared to gentle downhill and gentle uphill slopes. The average temperature of the axle box on steep downhill slopes is higher than that on gentle downhill slopes, and the average temperature on steep uphill slopes is also higher than that on gentle uphill slopes. From the average bearing temperature, it can be seen that the average temperature on steep slopes is higher than on gentle slopes, and the steeper the slope, the higher the bearing temperature. The slope of the road affects the bearing temperature.

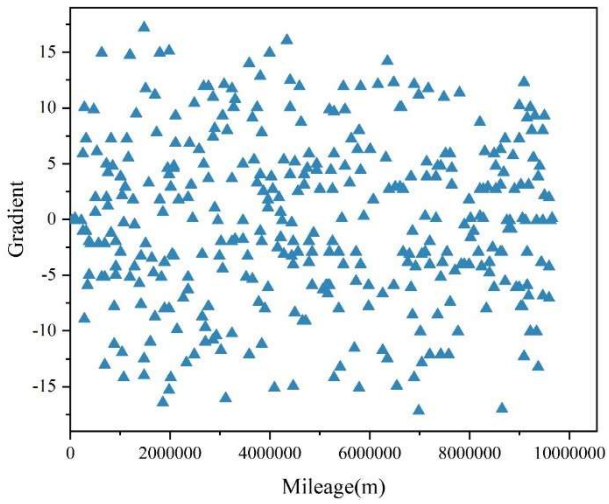



Figure 4: Scatter plot of slope distribution

Table 2: Average temperature of train gearbox under gradient classification

| Ramp level | Average temperature (°C) |
|---|---|
| Steep downhill | 72.48 |
| Slow downhill | 71.61 |
| Slow uphill | 71.78 |
| Steep uphill | 72.55 |

Randomly select a single driving record and compare the bearing temperature performance under four different slope conditions during the single driving process. The scatter plot of gearbox temperature under slope grades is shown in Figure 5, where the green line represents the average bearing temperature under that slope grade. As can be seen from the figure, the bearing temperature on non-slope sections is lower than that on gentle uphill or downhill slopes, and lower than that on steep uphill or downhill slopes. The slope gradient affects bearing temperature, with steeper slopes resulting in higher bearing temperatures.

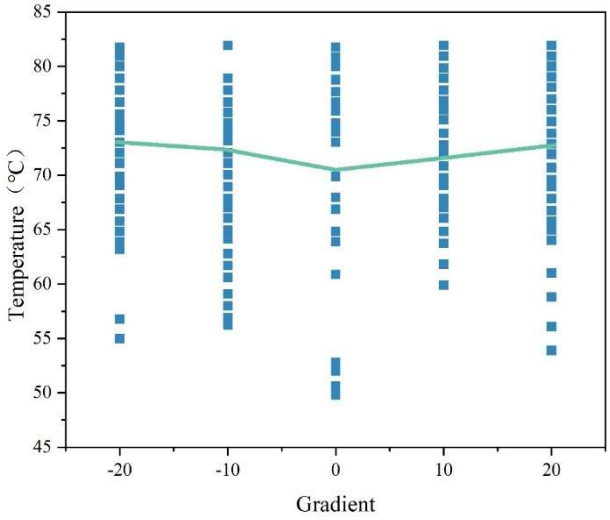



Figure 5: Scatter plot of gearbox temperature under slope grade

**III. B. 2) Comparison and analysis of bearing temperatures on different slopes**

The sample data used in this analysis consists of approximately 1 million multi-vehicle trip records, with 350,000 records at low speeds and 580,000 records at high speeds.

1) Comparison of bearing temperatures between slopes under high-speed conditions

Under high-speed conditions, the average and median values of bearing temperature changes with slope are shown in Figure 6. Under high-speed conditions, bearing temperature is lowest on non-sloped sections of track, and increases as the slope increases. Bearing temperature increases relatively slowly when descending slopes. Bearing temperature increases more significantly when ascending slopes.

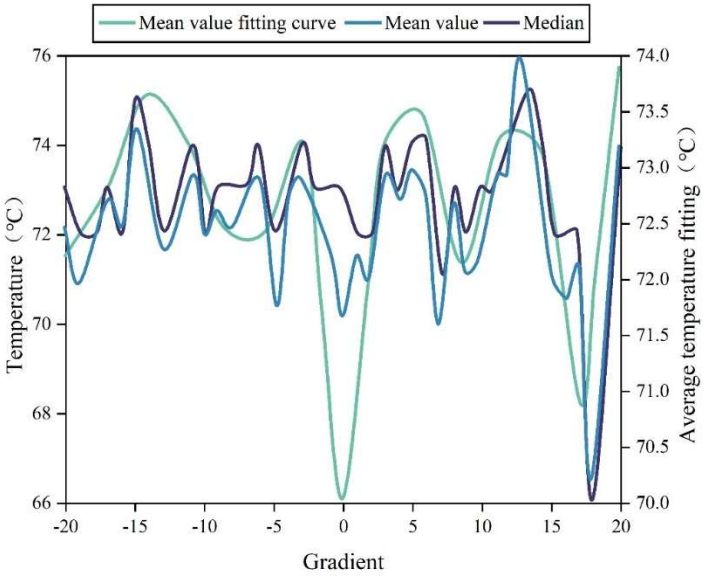



Figure 6: Curves of mean and median with slope change

2) Comparison of bearing temperatures between slopes under low-speed driving conditions

Under low-speed driving conditions, the average and median values of gearbox bearing temperatures as they change with slope are shown in Figure 7. As can be seen from the figure, bearing temperatures increase relatively slowly when driving downhill. Bearing temperatures increase more when driving uphill than when driving downhill.

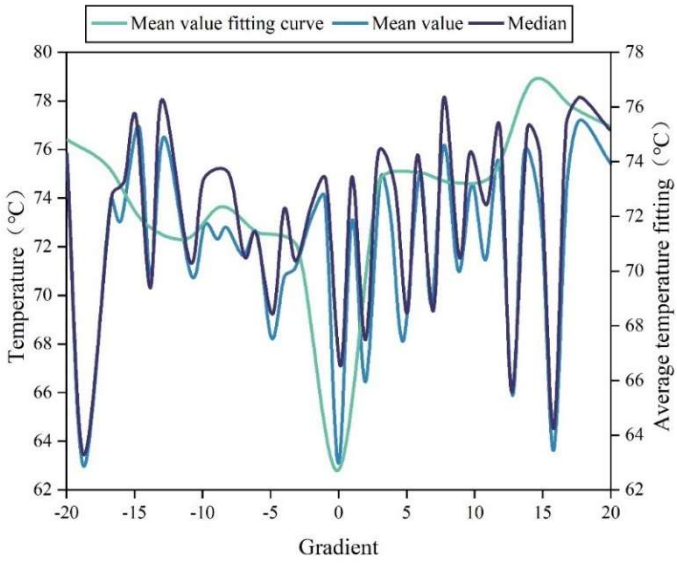



Figure 7: Curve of average and median with slope under low speed vehicle condition

## IV. Prediction of EMU wheel dimensions based on VMD-PSO-MKELM

A well-maintained high-speed train wheel profile not only ensures the train remains in the correct position on the track for safe operation but also reduces wear between the wheels and the track, thereby extending the service life of the wheels and lowering manufacturing and maintenance costs. Accurately predicting high-speed train wheel set data holds significant practical importance for train maintenance and railway safety. This chapter will

establish a high-speed train wheel dimension detection model based on VMD-PSO-MKELM to provide data support for intelligent detection and early warning of high-speed train wheels.

### *IV. A. Multi-core Extreme Learning Machine*

The KELM model builds upon the ELM model by replacing random mapping with kernel mapping, transforming the problem in low-dimensional space into a complete inner product space for solution. This significantly reduces network complexity, enhancing the model's predictive capability for regression problems and improving its generalization performance [27]. However, the predictive performance of different kernel functions varies significantly on the same dataset, indicating that the single kernel function in the standard KELM algorithm struggles to adapt to the diverse range of high-speed train wheel sizes. To address this, this paper modifies the KELM model to propose a multi-kernel function kernel extreme learning machine, resulting in the MKELM model, which overcomes the limitations of the kernel extreme learning machine and addresses the issue of insufficient regression capability.

For the sample set $(x_i, t_i)$, $i = 1, 2, \ldots, n$, the standard KELM output is:

$$f_{KELM}(x) = \begin{bmatrix} K(x, x_1) \\ K(x, x_2) \\ \vdots \\ K(x, x_n) \end{bmatrix} \left( K_{ELM} + \frac{I}{C} \right)^{-1} T \tag{4}$$

In the equation: $K(x, x_i)$ is the kernel function; $K_{ELM}$ is the kernel matrix; $I$ and $T$ are the diagonal matrix and target vector matrix, respectively; $C$ is the regularization coefficient. The larger $C$ is, the higher the model accuracy; the smaller $C$ is, the stronger the generalization ability. An appropriate value of $C$ is crucial for the model.

From equation (4), it can be seen that when the kernel function $K(x, x_i)$ is determined, the prediction result can be obtained. Therefore, the selection of $K(x, x_i)$ is critical to the prediction accuracy of the model. Kernel functions can be classified into global kernels and local kernels. Common kernel functions include the Poly kernel, RBF kernel, and Lin kernel, with the following kernel function forms:

$$K_{Poly}(x, x_i) = (x, x_i + c_1)^d \tag{5}$$

$$K_{RBF}(x, x_i) = \exp\left( -\frac{\|x - x_i\|^2}{\sigma^2} \right) \tag{6}$$

$$K_{Lin}(x, x_i) = x^T x_i \tag{7}$$

In the formula: $\sigma$ and $c_1$ are the kernel parameters of the RBF kernel; $d$ is the kernel parameter of the Poly kernel.

The above kernel functions are added together to form a hybrid kernel function, that is:

$$K_{MKELM}(x, x_i) = c_2 K_{Poly}(x, x_i) + c_3 K_{RBF}(x, x_i) + c_4 K_{Lin}(x, x_i) \tag{8}$$

In the formula, $c_2$, $c_3$, and $c_4$ are the weighting coefficients of each kernel function, whose values range from $[0, 1]$ and satisfy the following constraints:

$$c_4 = 1 - c_2 - c_3 \tag{9}$$

Substituting equations (5), (7), and (9) into equation (8) yields:

$$K_{MKELM}(x, x_i) = c_2 (x, x_i + c_{12})^d + c_3 \exp\left( -\frac{\|x - x_i\|^2}{\sigma^2} \right) + (1 - c_2 - c_3) x^T x_i \tag{10}$$

Substituting equation (10) into equation (4) yields the MKELM model. The hybrid kernel function in this model combines the advantages of global and local kernels, enabling it to exhibit not only excellent local search

capabilities but also enhanced global search capabilities under different parameter settings. Since multi-kernel functions have many parameters, manually determining parameters is inefficient. Therefore, the PSO algorithm is used to optimize the six parameters in MKELM: $c_1$, $c_2$, $c_3$, $\sigma$, $d$, $C$.

### *IV. B. Prediction model for EMU wheel dimensions*

To accurately analyze the characteristics of high-speed train wheels, VMD is used to decompose the wheel dimensions into IMF modes with higher regularity. An MKELM model is established for each mode, and the PSO algorithm is used to optimize each model to obtain the optimal solution. Finally, the prediction results of each IMF mode are weighted and summed to obtain the final prediction result [28].

#### IV. B. 1) Data Preprocessing and Initialization

1) Substitute the original EMU wheel dimensions into the above VMD process, and use the weather, day type, and decomposed EMU wheel dimension sequence to form several input samples, including:

$$X:\left(x_1,x_2,\ldots,x_k,\ldots,x_K\right) \tag{11}$$

$$x_k=\left[P_k\ W\ D\right] \tag{12}$$

In the formula: $1\le k\le K$; $P_k$ is the $k$ th dimension sequence after decomposition; $W$ represents weather; $D$ is the day type.

2) Data normalization. To avoid the influence of different units and sizes of various input data on the prediction results, it is necessary to eliminate the dimensions and accelerate the algorithm optimization process. Therefore, the data is normalized. The normalization formula is:

$$x_k^{ni}=\frac{x_k^i-x_k^{\min}}{x_k^{\max}-x_k^{\min}} \tag{13}$$

In the formula: $k$ is the $k$ th input sample; $x_k^{ni}$ is the normalized value, $x_k^i$ is the original variable; $x_k^{\min}$ is the minimum value of the original variable; $x_k^{\max}$ is the maximum value of the original variable.

3) Divide the normalized data into training and testing sets.

#### IV. B. 2) PSO stage

1) Initialize the particle swarm: Set the population size to 20, the number of iterations to 100, and the weights $\omega_{start}$ and $\omega_{end}$ at the start and end of optimization to 0.9 and 0.4, respectively. The velocity and position ranges of the particles are updated within the intervals $[-1,1]$ and $[-5,5]$, respectively.

2) Use the MKELM parameters associated with the PSO algorithm, train the MKELM using the training samples, and obtain the fitness of the particles. The calculation formula is:

$$fit=\left(\sum_{i=1}^{N}\left(x_k^{i,predicted}-x_k^{i,real}\right)^2\right)\Big/N \tag{14}$$

In the formula, $x_k^{i,predicted}$ and $x_k^{i,real}$ are the predicted value and actual value of the $k$ th sample mode, respectively.

3) Update the MKELM parameters until the fitness criterion or maximum number of iterations is satisfied, and obtain the optimal MKELM parameters.

#### IV. B. 3) MKELM prediction phase

1) Substitute the optimal parameters and test samples obtained from data preprocessing into the MKELM model, then calculate the training set kernel matrix, output weights, and test set kernel matrix. Next, multiply the test set kernel matrix and output weights to obtain the prediction results for each modality. Finally, weight and sum the results of the $K$ modalities to obtain the prediction results for the wheel size of the EMU:

$$x_{predicted}=\sum_{k=1}^{K}x_k^{predicted} \tag{15}$$

In the equation: $x_k^{predicted}$ is the predicted value of the $k$ th sample mode; $x_{predicted}$ is the predicted result of the wheel size of the EMU.

2) Result evaluation. To evaluate the performance of the VMD-PSO-MKELM combined algorithm model, three metrics are used: percentage error (PE), mean absolute percentage error (MAPE), and root mean square error (RMSE), represented as follows:

$$PE = \frac{x_{predicted} - x_{real}}{x_{real}} \times 100\% \tag{16}$$

$$MAPE = \frac{1}{n}\sum_{i=1}^{n}\frac{\left|x_{predicted} - x_{real}\right|}{x_{real}} \times 100\% \tag{17}$$

$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}\left(x_{predicted} - x_{real}\right)^2} \tag{18}$$

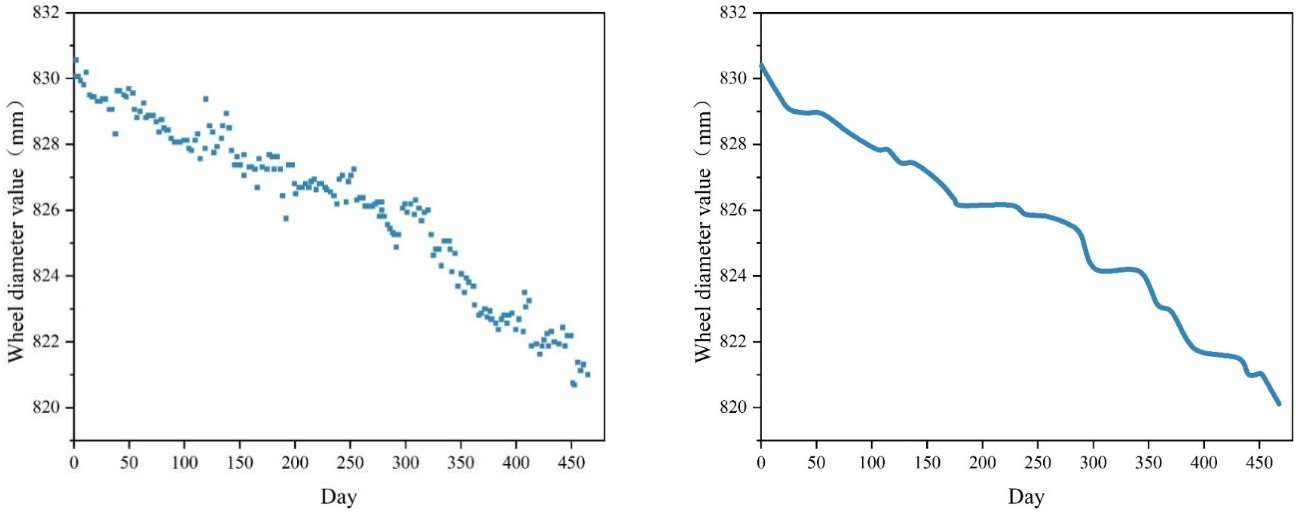
In the equation, $x_{real}$ is the actual value of the wheel size of the EMU.

### *IV. C. Analysis of wheel set data prediction results*

Collect historical wheel set dimension data from a CRH2A model EMU from July 2022 to February 2024 as a sample. Wheel diameter and flange thickness are the most important wheel dimension parameters in wheel turning strategies. Therefore, these two parameters were selected as data samples for predictive analysis.

#### IV. C. 1) Wheel diameter prediction and result analysis

First, the historical wheel diameter values of a certain wheel are used as data samples. The original data records and the pre-processed wheel diameter values are shown in Figure 8. Figures (a) and (b) correspond to the wheel diameter value sequences before and after denoising, respectively. The denoised wheel diameter value sequence reflects the monotonic decreasing trend of the wheel tread diameter as the running time increases.




(a) Raw data (b) Preprocessing data

Figure 8: Raw data record value and wheel diameter pretreatment value

Reconstruct the wheel diameter value sequence into 250 sets of input-output datasets. Simultaneously, perform learning and prediction based on a training set to test set data ratio of 4:1, where the first 200 sets serve as the training sample set and the remaining 50 sets as the test sample set.

Train the network using the VMD-PSO-MKELM algorithm proposed in this paper, employing PSO iteration to identify the optimal model parameters. The optimal fitness function changes of the wheel diameter value iteration for 100 times in the training of the high-speed train wheel size detection model proposed in this paper are shown in Figure 9. As shown in the figure, the RMSE of the training set reaches its minimum value when the network

model is iterated to the 30th time. The network parameters output at this time are the optimal parameters of the model.

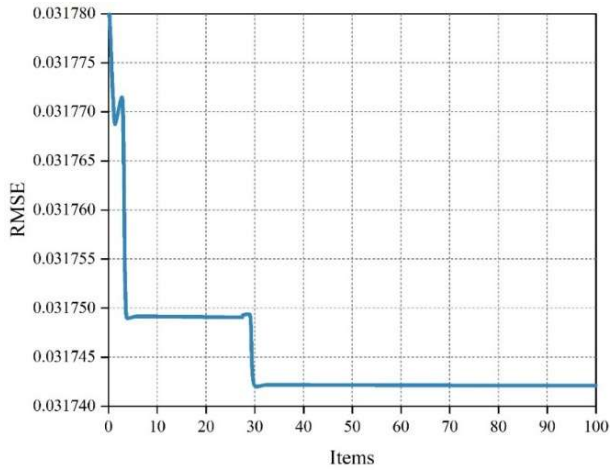



Figure 9: 100 iteration best fitness of VMD-PSO-MKELM model

The model performance was evaluated using the test set data, and the prediction results are shown in Figure 10. As can be seen from the figure, the results of predicting the wheel diameter values using the VMD-PSO-MKELM model show good consistency with the actual situation. The calculated $R^2$ value for the predicted values is 0.9968, the standard error (SE) is 0.0012, the mean absolute error (MAE) is 0.0294, and the mean absolute percentage error (MAPE) is 0.0004%, indicating high accuracy and generalization capability.

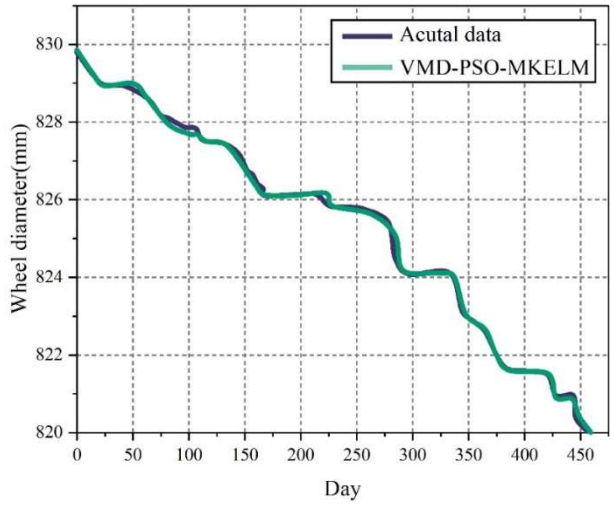



Figure 10: Prediction results of wheel diameter

To comprehensively evaluate the performance of multi-kernel extreme learning machines in predicting wheel size data, we selected traditional BP neural networks, extreme learning machines (ELM), and linear kernel extreme learning machines (L-ELM), polynomial kernel extreme learning machines (P-ELM), RBF kernel extreme learning machine (R-ELM) for comparison with the prediction results of the VMD-PSO-MKELM model proposed in this paper. The evaluation metrics used for analysis include $R^2$, MSE, MAE, and MAPE, with the prediction results shown in Table 3. As shown in the table, the MSE, MAE, and MAPE values of the VMD-PSO-MKELM model's prediction results are lower than those of other models, at 0.0012, 0.0294, and 0.0004%, respectively. This indicates that the optimized model achieves the highest precision and accuracy in predicting wheel diameter values. Additionally, the model's $R^2$ value is high, reaching 0.9968, which also reflects its strong fitting and generalization capabilities.

Table 3: Comparison of prediction results of different algorithms for wheel diameter

| Dataset | Algorithm | R2 | MSE | MAE | MAPE (%) |
|---|---|---|---|---|---|
| Wheel diameter | BP | 0.9409 | 0.0383 | 0.1622 | 0.0038 |
| | ELM | 0.9581 | 0.0228 | 0.1242 | 0.003 |
| | L-ELM | 0.9595 | 0.0095 | 0.0886 | 0.0008 |
| | P-ELM | 0.9956 | 0.0164 | 0.1217 | 0.0037 |
| | R-ELM | 0.9452 | 0.0204 | 0.1132 | 0.0006 |
| | VMD-PSO-MKELM | 0.9968 | 0.0012 | 0.0294 | 0.0004 |

**IV. C. 2) Flange thickness prediction and result analysis**

The same method was used to predict the flange thickness, and the results are shown in Figure 11. As can be seen from the figure, the calculation results of the VMD-PSO-MKELM model are in good agreement with the actual data.

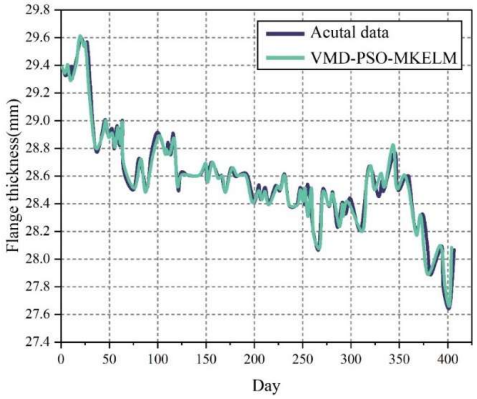



Figure 11: Prediction results of flange thickness

Similarly, the prediction results of the VMD-PSO-MKELM model were compared with those of the BP, ELM, L-ELM, P-ELM, and R-ELM algorithms. The specific comparison of prediction results among different algorithms for different rim thicknesses is shown in Table 4. From the flange thickness prediction results, it can also be seen that the MSE, MAE, and MAPE of the VMD-PSO-MKELM model's prediction results are 0.0081, 0.0741, and 0.0005%, respectively, all of which are lower than those of other models, indicating that the optimized model has higher precision and accuracy in predicting flange thickness. The model's $R^2$ value is also the highest among all models, reaching 0.9251, which reflects the model's strong generalization capability.

Table 4: Comparison of prediction results of different algorithms for flange thickness

| Dataset | Algorithm | R2 | MSE | MAE | MAPE (%) |
|---|---|---|---|---|---|
| Wheel flange thickness | BP | 0.8418 | 0.0587 | 0.195 | 0.1548 |
| | ELM | 0.9005 | 0.0234 | 0.1194 | 0.0752 |
| | L-ELM | 0.8971 | 0.0134 | 0.1001 | 0.0268 |
| | P-ELM | 0.9154 | 0.0137 | 0.0931 | 0.0536 |
| | R-ELM | 0.9071 | 0.0103 | 0.0852 | 0.0175 |
| | VMD-PSO-MKELM | 0.9251 | 0.0081 | 0.0741 | 0.0005 |

From the prediction results of the two sets of data, wheel diameter and flange thickness, it can be seen that the VMD-PSO-MKELM model proposed in this paper has higher prediction accuracy than other models such as ELM, L-ELM, P-ELM, and R-ELM. Compared with the BP model, it not only saves the complex network training process but also avoids the instability of traditional neural networks, making it more practical.

## V. Application of PHM Technology for Intelligent Detection and Early Warning of EMU Wheels

The intelligent detection and early warning system for high-speed train wheels requires real-time dynamic monitoring of the train while it is in operation, with the train's position tracked in real time. Train dispatchers utilize

a fault expert knowledge base to make emergency responses to faults or other urgent situations that pose a threat to train safety. This chapter will combine the high-speed train wheel size detection method proposed in this paper, based on VMD-PSO-MKELM, with PHM technology to achieve high-speed train wheel detection and early warning, thereby constructing a high-speed train wheel intelligent detection and early warning system. The functional content of the system is as follows:

1) Real-time operational status monitoring. By comprehensively aggregating real-time fault information, basic status information, and notification information from onboard data, the system comprehensively monitors high-speed train alarm information and status information across all departments and processes.

2) By integrating information on high-speed train operational status, maintenance status, spare vehicle information, maintenance warnings, and related technologies, the system provides robust auxiliary information support for vehicle scheduling.

3) During fault handling, remotely view driver screen fault and status parameter information, and directly contact the on-board mechanic and driver. Through system integration between the railway bureau and the railway corporation, provide emergency response recommendations for critical faults based on the fault expert knowledge base, and automatically record and back up the entire fault resolution process.

4) Obtain the cause of the fault and related environmental parameters, integrate with the railway bureau's high-speed train base management system fault module, and provide information on the fault resolution results.

5) Replay the operational status of the high-speed train and driver's operational actions to provide robust information support for accident cause analysis and liability determination.

6) Real-time tracking of the train set's location, providing a convenient browsing method to obtain information on the assigned train set's location and fault status.

### *V. A. Application of Safety Technology*

A public information security platform has been established to ensure the security of real-time data transmission. Non-real-time download data is transmitted via a wireless local area network (WLAN) deployed within the maintenance depot to the railway intranet, so its security must be given high priority. For wired networks, data is transmitted via cables to the destination. In environments where the physical link is disrupted during transmission, data leakage may occur; However, under the currently most widely covered Internet network, as long as the terminal is within the coverage range of a wireless access point (AP), it can receive the signal. Additionally, the wireless access point (AP) directs the signal to a specific receiving device, fully demonstrating its security.

To effectively ensure the security of communication between vehicles and ground stations, high-precision security technologies must be implemented in wireless networks, including system authentication, external access control, communication encryption, and data upload review. By comprehensively implementing security measures, robust protection is provided across all layers of network communication. The specific security protection scheme for vehicle-to-ground communication is illustrated in Figure 12. Based on this, thorough wireless local area network (WLAN) security testing was conducted on the wireless communication endpoints (STA), corresponding access points (AP), and access controllers (AC) to ensure the security and reliability of vehicle-to-ground communication.

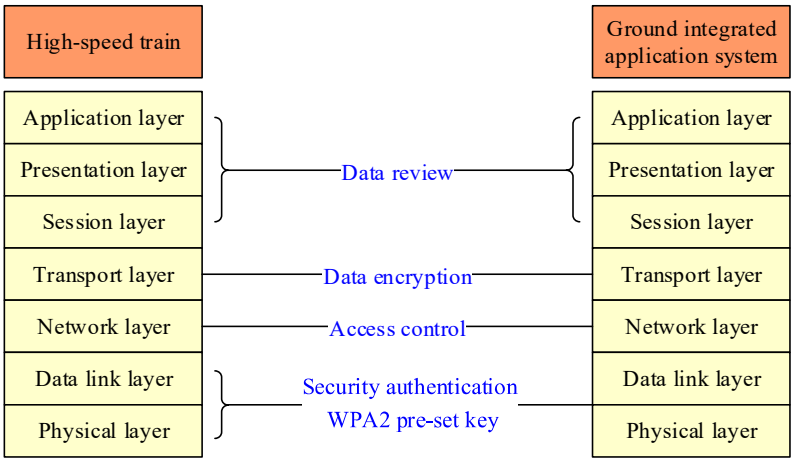



Figure 12: Safety protection scheme of vehicle-ground communication

### *V. B. Application of Monitoring Technology*

As more high-speed trains equipped with onboard information systems are put into service and the scope of ground-based integrated business applications continues to expand, IT system monitoring and management have become increasingly critical and prominent. It is essential to gradually optimize the system, with the most critical safeguard being a rigorous system monitoring mechanism.

In the design specifications for communication protocols and data content, all required functionalities and operational needs for data access have been considered. The system's control mechanisms have been studied to ensure normal operation, with real-time monitoring of all system operations. All changes in the operational cycles of the equipment are updated in real time, enabling control and command based on the latest status to swiftly address and resolve issues. Various monitoring tools and software are employed, along with a unified system management platform that integrates multiple monitoring tools and software, to manage the entire software and hardware platform environment. This enables performance monitoring of networks, systems, applications, data, and the environment, while also providing event-related interfaces for business management, ensuring scalability for future expansion.

## VI. Conclusion

This paper proposes a data processing framework for high-speed train PHM based on Spark Streaming and Kafka, and uses correlation algorithms to determine the factors affecting wheel set wear. Data from the Wuhan-Guangzhou Railway Line from January 2023 to July 2024 was selected to predict and analyze the health status of high-speed train wheels. The slope of the track affects bearing temperature, with steeper slopes resulting in higher bearing temperatures. Under high-speed conditions, bearing temperature increases with slope steepness, while it reaches its lowest value on non-sloped sections of track. During descents, bearing temperature increases at a relatively slower rate, indicating that temperature increases are more pronounced during ascents than descents. Under low-speed conditions, bearing temperature increases are similarly more pronounced during ascents than descents.

Using the identified factors influencing high-speed train wheel set wear as input parameters, a wheel size prediction model based on VMD-PSO-MKELM was constructed, and the effectiveness and practicality of the VMD-PSO-MKELM model were validated using wheel diameter and flange thickness data. In wheel diameter data, the VMD-PSO-MKELM model's MSE, MAE, and MAPE were 0.0012, 0.0294, and 0.0004%, respectively, all lower than those of other comparison models such as BP, ELM, L-ELM, P-ELM, and R-ELM, while $R^2$ reached the highest value of 0.9968. For flange thickness data, the VMD-PSO-MKELM model still had the lowest MSE, MAE, and MAPE among all models, at 0.0081, 0.0741, and 0.0005%, respectively. The $R^2$ value was also the highest among all models, reaching 0.9251. Overall, the VMD-PSO-MKELM model proposed in this paper avoids the instability issues of traditional neural networks while demonstrating high prediction accuracy, making it more practical.

Finally, by combining the VMD-PSO-MKELM-based high-speed train wheel size detection method proposed in this paper with PHM technology to achieve high-speed train wheel detection and early warning, an intelligent detection and early warning system for high-speed train wheels has been established. This system enables real-time monitoring and tracking of the operational status and location of high-speed trains, promptly identifies the causes of faults and related environmental parameters, and provides reliable auxiliary information support for the scheduling of high-speed train vehicles. In terms of safety technology, rigorous testing was conducted on the wireless communication end stations (STA), corresponding access points (AP), and access controllers (AC); in terms of monitoring technology, multiple monitoring tool software and a unified system management platform were adopted, integrating various monitoring tool software to efficiently manage the entire software and hardware platform environment.

## References

[1] Li, Z., & Chen, Z. (2023). Predicting the future development scale of high-speed rail through the urban scaling law. Transportation Research Part A: Policy and Practice, 174, 103755.

[2] Sansan, D., Dawei, C., & Jiali, L. (2021). Research, development and prospect of China high-speed train. Chinese journal of theoretical and applied mechanics, 53(1), 35-50.

[3] Zhao, H., Liang, J., & Liu, C. (2020). High-speed EMUs: characteristics of technological development and trends. Engineering, 6(3), 234-244.

[4] Zhao, P., Han, B., Li, D., & Li, Y. (2021). Model and algorithm for the first-level maintenance operation optimization of EMU trains. Journal of Intelligent & Fuzzy Systems, 41(1), 2145-2160.

[5] Zheng, W., Zhou, T., & Li, Y. F. (2020, December). An overview of the EMUs maintenance scheduling in China. In 2020 IEEE 20th International Conference on Software Quality, Reliability and Security Companion (QRS-C) (pp. 297-300). IEEE.

[6] Wu, J., Lin, B., Wang, J., & Liu, S. (2017). A network-based method for the EMU train high-level maintenance planning problem. Applied Sciences, 8(1), 2.
[7] Qi, Y., & Zhou, L. (2020). The fuxing: the China standard EMU. Engineering, 6(3), 227-233.
[8] Lu, C., Zhang, B., & Zhao, H. (2023). CR-Fuxing high-speed EMU series. Frontiers of Engineering Management, 10(4), 742-748.
[9] Bai, Y. (2022, November). Research on Ethernet Communication of EMU Single Vehicle Commissioning System. In Proceedings of the 5th International Conference on Information Technologies and Electrical Engineering (pp. 526-533).
[10] Yang, W., Shi, J., & Huang, J. (2020, December). Static contact characteristic analysis of new axle box bearing based on EMU wheelset. In Proceedings of the 3rd International Conference on Information Technologies and Electrical Engineering (pp. 313-316).
[11] Zhang, G., & Ren, R. (2019). Study on typical failure forms and causes of high-speed railway wheels. Engineering Failure Analysis, 105, 1287-1295.
[12] Pecht, M. G., & Kang, M. (2018). Introduction to PHM. Prognostics and health management of electronics: fundamentals, machine learning, and the internet of things, 1-37.
[13] Lin, R., Yang, J., Huang, L., Liu, Z., Zhou, X., & Zhou, Z. (2023). Review of Launch Vehicle Engine PHM Technology and Analysis Methods Research. Aerospace, 10(6), 517.
[14] Gharib, H., & Kovács, G. (2023). A review of prognostic and health management (PHM) methods and limitations for marine diesel engines: New research directions. Machines, 11(7), 695.
[15] Liu, Z., Jia, Z., Vong, C. M., Han, J., Yan, C., & Pecht, M. (2018). A patent analysis of prognostics and health management (PHM) innovations for electrical systems. IEEE Access, 6, 18088-18107.
[16] Bo, J., Guangyue, X., Yifeng, H. U. A. N. G., Xiaoxuan, J., & Wei, L. (2017). Recent advances analysis and new problems research on PHM technology of military aircraft. Journal of Electronic Measurement and Instrumentation, 31(2), 161-169.
[17] Cheng, Q., Cao, Y., Liu, Z., Cui, L., Zhang, T., & Xu, L. (2024). A health management technology based on PHM for diagnosis, prediction of machine tool servo system failures. Applied Sciences, 14(6), 2656.
[18] Grosso, L. A., De Martin, A., Jacazio, G., & Sorli, M. (2020). Development of data-driven PHM solutions for robot hemming in automotive production lines. International Journal of Prognostics and Health Management, 11(1).
[19] Zhang, S. (2019, September). The Framework of EMU Wheelset Lifecycle RAMS Quality Management Platform. In Sixth International Conference on Transportation Engineering (pp. 1075-1081). Reston, VA: American Society of Civil Engineers.
[20] Tao, X., Guo, Z., Sun, P., & Li, Z. (2023, April). An integrated technical scheme for the wheel set flaw detection robot system of EMUs. In International Conference on Signal Processing, Computer Networks, and Communications (SPCNC 2022) (Vol. 12626, pp. 632-637). SPIE.
[21] Arena, F., & Pau, G. (2020). An overview of big data analysis. Bulletin of Electrical Engineering and Informatics, 9(4), 1646-1653.
[22] Wang, H., & Min, B. W. (2022). Research and Application of Fault Prediction Method for High-speed EMU Based on PHM Technology. Journal of Internet of Things and Convergence, 8(6), 55-63.
[23] Feng, D., Lin, S., He, Z., & Sun, X. (2017). A technical framework of PHM and active maintenance for modern high-speed railway traction power supply systems. International Journal of Rail Transportation, 5(3), 145-169.
[24] Min, B. W. (2023). Improved Fault Prediction Algorithm of High-Speed EMUs based on PHM Technology. International Journal of Contents, 19(2).
[25] Byung Won Min. (2023). Improved Fault Prediction Algorithm of High-Speed EMUs based on PHM Technology. International JOURNAL OF CONTENTS,19(2).
[26] Manoilov V. V.,Novikov L. V.,Belozertsev A. I.,Zarutskiy I. V.,Titov Yu. A.,Kuzmin A. G. & El Salim S. Z. (2020). Processing of Mass Spectra of Exhaled Gases Based on Correlation Algorithms. Journal of Analytical Chemistry,75(13),1678-1684.
[27] Nirjharinee Parida,Debahuti Mishra,Kaberi Das & Narendra Kumar Rout. (2019). Development and performance evaluation of hybrid KELM models for forecasting of agro-commodity price. Intelligence,14(prepublish),1-16.
[28] Ahmet Akkaya & Cemil Közkurt. (2025). An effective approach for adaptive operator selection and comparison for PSO algorithm. Cluster Computing,28(6),368-368.