Shunting Locomotive Health Assessment and Management System

Jincheng Zhang, Jitao Li, Jianyu Zhang

***Abstract*— To address the low intelligent level of operation and maintenance management for shunting locomotives in coal enterprises, which often results in over-maintenance or delayed maintenance issues, this study analyzed historical operation and maintenance data of shunting locomotives in depth. Based on this analysis, a health index prediction model for shunting locomotives was established. By training the model using the random forest algorithm, the health status of shunting locomotives can be assessed without adding real-time monitoring devices. This provides support for managers to formulate maintenance strategies and reduce life-cycle operation and maintenance costs. The research results demonstrate that the model can accurately predict the health condition of shunting locomotives. The implementation of the health assessment system for shunting locomotives in coal enterprises has provided strong support for the intelligent development of railway transportation in coal enterprises, driving the transformation of shunting locomotive operation and maintenance management towards intelligence.**

***Index Terms*—Health Assessment, Shunting Locomotives, Random Forest, Management System.**

I. Introduction

Shunting locomotives are essential equipment in the transportation and production processes of coal mining enterprises, and their health status has a direct impact on both corporate profitability and operational safety. Meanwhile, there is a growing need to update and enhance current maintenance practices for these locomotives. Traditional approaches such as “ planned maintenance ” or “ mileage-based maintenance ” have long been the mainstay for ensuring locomotive operation, yet they often lead to excessive maintenance or delayed repairs, thereby escalating operating costs and introducing potential safety risks. Consequently, one pressing issue for coal mining enterprises is how to transition to a “ condition-based maintenance” model[1], accurately determine the real-time health status of locomotives, formulate economically viable maintenance strategies, and reduce total life-cycle maintenance expenses.

With the rise of Prognostics and Health Management (PHM) technologies [2], employing specialized sensors and data collection devices to monitor and assess locomotive conditions in real-time has become both a research focus and an industry trend. PHM primarily integrates real-time sensor data with machine learning and data analytics to evaluate equipment health and predict potential failures, thereby providing fault warning in advance. For example, Li Chenglong et al. [3] developed a fault prediction and health management system for Hexie locomotives, successfully diagnosing faults in key components such as traction converters, while Yan Ying et al. [4] achieved a health assessment of locomotive running gear. However, implementing PHM requires large volumes of precise, real-time data, which coal mining enterprises currently struggle to obtain. On one hand, they lack specialized, high-precision monitoring networks and integrated analysis platforms (e.g., the 5T system for railway vehicles or the CMD system for Chinese locomotives) [5]; on the other hand, the elevated costs and harsh operating conditions of specialized sensors make comprehensive and continuous data collection for shunting locomotives impractical. Therefore, finding a solution that leverages the readily available data within coal mining enterprises to assess locomotive health remains an urgent challenge.

To address this, the present study aims to design and develop a shunting locomotive health assessment and maintenance management system tailored to the practical conditions of coal mining enterprises. Unlike mainstream methods that rely heavily on real-time sensor deployment, our system utilizes the historical operational and maintenance data of shunting locomotives and applies machine learning alongside big data analytics to build a health assessment model. This approach offers a viable means of evaluating locomotive health while mitigating the problem of excessive maintenance. In doing so, it provides managers with scientific decision support for maintenance scheduling, thereby reducing overall operational costs and facilitating the transition to condition-based maintenance within the constraints of existing enterprise resources.

II. System Architecture

A. Figures and Tables

The overall architecture of the Shunting Locomotive Health Assessment and Maintenance Management System is illustrated in Figure 1.

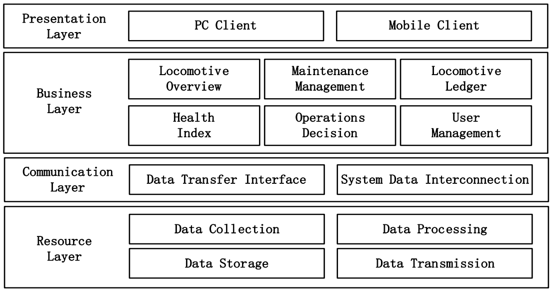


Fig 1. System Overall Architecture

Manuscript received March 13, 2025

Jincheng Zhang, School of Transportation Engineering, Dalian Jiaotong University, Dalian, China

Jitao Li, School of Transportation Engineering, Dalian Jiaotong University, Dalian, China.

Jianyu Zhang, Transportation Department, Fushun Mining Group Co. Ltd., Fushun , China.

(1) Resource Layer

The resource layer forms the foundational tier of the system architecture, responsible for data acquisition, processing, storage, and transmission. Relevant shunting locomotive data are obtained from associated systems or locomotive logs. Subsequently, raw data undergo cleaning, conversion, and processing to ensure standardized formats. The processed data are then stored in a SQL Server database and accurately transmitted to the upper layers to support the functionalities of the business layer.

(2) Communication Layer

This layer comprises data transfer interfaces and data interconnection. Standardized data interfaces are built based on RESTful APIs, and lightweight JSON protocols are employed for data encapsulation and routing, thereby enabling bidirectional communication with related systems such as the locomotive big-data statistical platform. A unified data model is also defined to achieve consistent format conversion for incoming and outgoing data, ensuring efficient and reliable inter-system data exchanges.

(3) Business Layer

Aimed at managerial personnel, the business layer provides key functionalities including Locomotive Overview, Health Assessment, Maintenance Decision-Making, Maintenance Management, and User Management. It offers real-time displays of critical operational parameters through graphical interfaces, leverages operational data and work records to assess the health status of shunting locomotives, assists managers in formulating maintenance strategies, and provides unified account management for the system. These capabilities collectively offer robust technical support for efficient operations and maintenance management.

(4) Presentation Layer

The presentation layer serves as the user-facing component of the system, converting internal data and functionalities into visualized interfaces. Users can access the system via either a PC client or a mobile client, catering to diverse work settings. The PC client offers comprehensive features to administrative staff, whereas the mobile client targets on-site personnel, providing basic data retrieval and status overviews.

B. Data Architecture

The data architecture of this system is shown in Figure 2.

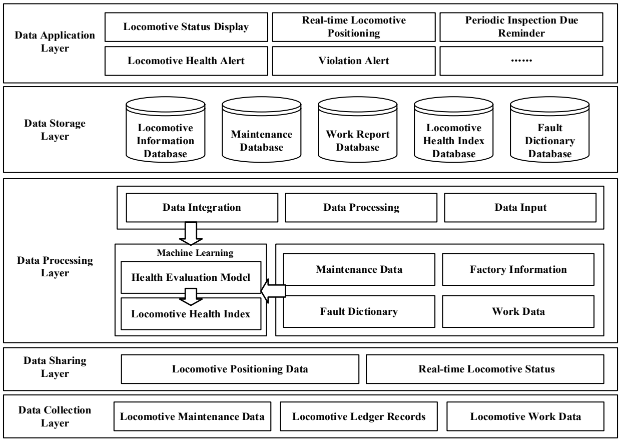


Fig 2. System Data Architecture

The data architecture of the system proposed in this study is divided into five layers: the Data Acquisition Layer, the Data Sharing Layer, the Data Processing Layer, the Data Storage Layer, and the Data Application Layer.

(1) Data Acquisition Layer

This layer acquires shunting locomotive-related data through the enterprise data interface, encompassing factors such as factory information, operational data, work data, and environmental parameters. Specifically, it includes details on locomotive models, technical specifications, and factory configurations, as well as maintenance and fault records, load and working hours, and other relevant data.

(2) Data Sharing Layer

Data are retrieved in real time from the big data statistical system via system interconnectivity, providing locomotive positioning and real-time status information. These data are subsequently transferred to the processing layer.

(3) Data Processing Layer

Primarily responsible for processing, analyzing, and evaluating locomotive data, the data processing layer leverages machine learning to build a health assessment algorithm. Based on historical maintenance data and related locomotive information, it generates real-time health scores and continuously refines its model parameters to enhance the accuracy of the status evaluations.

(4) Data Storage Layer

System data are stored in SQL Server, capturing core information such as basic locomotive profiles, maintenance data, historical work reports, historical health indices, and a locomotive fault dictionary. These datasets directly support the operations of the data application layer.

(5) Data Application Layer

Providing intuitive and intelligent data services for coal mining enterprises, this layer offers real-time access to locomotive operating data. It displays key operational metrics such as locomotive location and running status, and pushes alerts for issues such as rule violations or upcoming maintenance deadlines, thereby bolstering efficient locomotive management.

III. Shunting Locomotive Health Assessment

A. Algorithm Principle

Random forest is a machine learning algorithm based on ensemble learning, specifically a type of Bootstrap aggregating method[6]. As a powerful predictive tool, its core concept involves building multiple decision trees and aggregating their prediction outcomes, thereby enhancing the model’s generalization capability, robustness, and resistance to overfitting.

The principle of random forest regression can be summarized as follows:

(1) Partition the original dataset into a training set and a test set.

(2) Perform multiple rounds of Bootstrap sampling on the training set, each time drawing a new subsample of the training data (denoted as Training Set 1 to Training Set N).

(3) For each subsample, randomly select m features out of M total features (where M > m) and build a decision tree based on these features. By calculating node impurity measures, the algorithm determines the optimal split points and branching directions, resulting in n decision tree

regression models and generating n regression predictions.

(4) The final prediction outcome is obtained by averaging or taking the median of all these n tree predictions[7].

Figure 3 illustrates the workflow of the random forest regression algorithm for shunting locomotive health assessment.

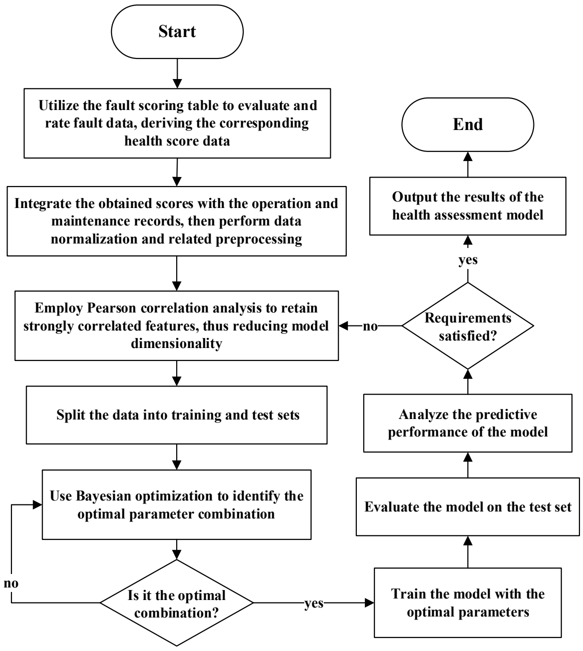

Start
Utilize the fault scoring table to evaluate and rate fault data, deriving the corresponding health score data
Integrate the obtained scores with the operation and maintenance records, then perform data normalization and related preprocessing
Employ Pearson correlation analysis to retain strongly correlated features, thus reducing model dimensionality
Split the data into training and test sets
Use Bayesian optimization to identify the optimal parameter combination
no
Is it the optimal combination?
yes
Train the model with the optimal parameters
Evaluate the model on the test set
Analyze the predictive performance of the model
Requirements satisfied?
no
yes
Output the results of the health assessment model
End

Fig 3. Training process of the health assessment model

B. Data Acquisition and Preprocessing

In this study, the health score is treated as the target variable, derived by quantifying the severity of faults within a given period. Specifically, the maximum health score is 100 points. Based on a fault scoring table for each subsystem, faults identified during the period receive corresponding deductions from the maximum score, yielding a final health score for that period.

$$G = 100 - \sum f_i \times d_i^2 \qquad (1)$$

where G is the health score, f_i denotes the score for the i-th fault, and d_i represents the severity level of the i-th fault.

Using the subsystem fault data, the system evaluates and scores each locomotive's faults according to the fault scoring table, and calculates the final health score for each period based on the above rule. The resulting health scores are then integrated with the corresponding operational and maintenance records, followed by data cleaning (handling outliers and missing values), splitting specific fields such as date, performing normalization, and removing weakly correlated features via correlation analysis.

Ultimately, the dataset retains eight features: total mileage, post-maintenance mileage, post-maintenance total hours, load, active ratio, month, locomotive model, and factory year—yielding 1,634 complete data entries. Of these, 80% are chosen as the training set, and 20% as the test set.

C. Hyperparameter Settings

The model employs the sklearn library in PyCharm for random forest regression[8], with parameters configured as follows: the data split is 0.8 (80% training data and 20% testing data); the splitting criterion for internal nodes is Mean Squared Error (MSE); the minimum number of samples required for node splitting is 2; the minimum number of samples for each leaf node is 1; the maximum tree depth is 10; the maximum number of leaf nodes is 50; the threshold for node impurity splitting is 0; the number of decision trees is 100; and sampling is performed with replacement (True).

D. Feature Importance and Model Optimization

The model is first trained on the training set of the original dataset, after which feature importance is computed, as shown in Table 1. The most significant features for shunting locomotive health assessment include total mileage (0.3384), factory year (0.2140), post-maintenance mileage (0.1975), post-maintenance total hours (0.1322), and month (0.0577). The cumulative importance of these features exceeds 90%, indicating their strong contribution to health status prediction. By contrast, features such as the active ratio or locomotive model show relatively minor impacts. Accordingly, only these five selected key features are retained for a retraining process to reduce model input dimensionality[9].

Tab 1. Feature importance table

Feature Name	Feature Importance	Feature Name	Feature Importance
Total Mileage	0.338376	Month	0.057728
Factory Year	0.213985	Active Ratio	0.044856
Post-Maintenance Mileage	0.197456	Load	0.012408
Post-Maintenance Total Hours	0.132172	Locomotive Model	0.012408

E. Analysis of Model Prediction Performance

To evaluate model performance, this study adopts Mean Squared Error (MSE), Mean Absolute Error (MAE), Mean Absolute Percentage Error (MAPE), and the Coefficient of Determination (R²) as evaluation metrics[10], corresponding to Equations (2-5):

$$MSE = \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2 \qquad (2)$$

$$MAE = \frac{1}{n}\sum_{i=1}^{n}|y_i - \hat{y}_i| \qquad (3)$$

$$MAPE = \frac{100}{n}\sum_{i=1}^{n}\left|\frac{y_i-\hat{y}_i}{y_i}\right| \qquad (4)$$

$$R^2 = 1 - \frac{\sum_{i=1}^{n}(y_i-\hat{y}_i)^2}{\sum_{i=1}^{n}(y_i-\bar{y})^2} \qquad (5)$$

where $\bar{y}$ is the mean of the sample data, n is the number of samples, y_i is the true value, and $\hat{y}_i$ is the predicted value.

After evaluating the test set with the trained model, the results are shown in Table 2. The forecast error is relatively low, indicating high accuracy and strong generalization capability.

Tab 2. Model Evaluation Indicator Results

	MSE	MAE	MAPE	R²
Training set	2.7992	0.6830	1.3469%	0.8961
Test set	3.5779	1.4505	2.8742%	0.8586

Figures 4 and 5 compare the predicted and actual values for the test set and selected features, respectively, demonstrating that the predicted results are generally close to the true values. These outcomes confirm that the random forest regression model attains desirable performance and fits well under the experimental setting.

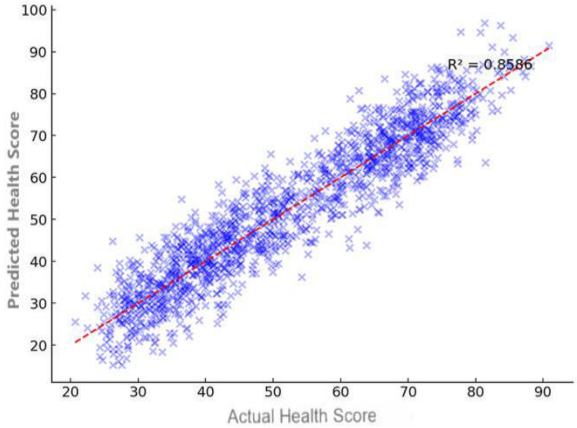


Fig 4. Test Set Regression Comparison Chart

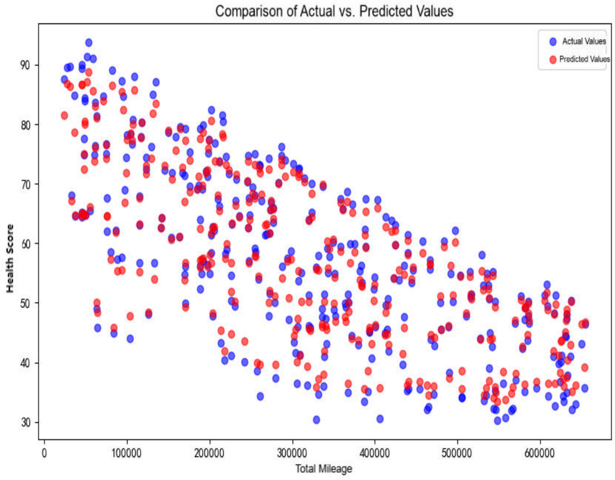


Fig 5. Comparison of Actual vs. Predicted Values

IV. Development of a Shunting Locomotive Health Assessment System

The goal of this system’s development is to employ a random forest algorithmic model to conduct intelligent assessments of shunting locomotive health within coal mining enterprises. By constructing a comprehensive health index model, the system provides real-time oversight of shunting locomotive conditions, offering scientific guidance for maintenance and management decisions. In doing so, it aims to address shortcomings in traditional locomotive health management practices — namely, the insufficient precision of manual assessments, excessive maintenance, and delayed repairs. Through intelligent decision support, the system strives to reduce enterprise maintenance and operational costs while advancing shunting locomotive maintenance management toward greater intelligence.

A. System Functions

(1) Locomotive Overview

a. Status Monitoring: Primarily displays the current operating status of shunting locomotives, including daily mileage, daily running hours, daily standby hours, and other key parameters essential for managerial observation. The system continuously tracks each locomotive’s operational and maintenance status, as well as verifying daily violations and fault reports. Critical information is prominently shown on the main interface.

b. Health Alerts: Utilizing the health assessment algorithm, the system generates a real-time health score for each locomotive. If a locomotive’s health score falls below a specified threshold, the main interface alerts the maintenance staff to pay closer attention to that unit.

c. Inspection Alerts: The system can define monthly or periodic maintenance intervals for each locomotive based on default settings or user-defined inspection plans. When a locomotive approaches its scheduled maintenance date or has exceeded it without proper servicing, the system issues a reminder encouraging timely maintenance checks.

d. Position Display: By collecting real-time locomotive location data, the system provides real-time positioning and a map-based display. Meanwhile, it records the locomotive’s running trajectory, allowing users to retrieve operational history for specific time periods.

(2) Maintenance Management

a. Maintenance Records: Users can conveniently browse the historical maintenance and fault records of various locomotives. The system enforces different levels of data access authority based on user roles. Only administrators and authorized general users are permitted to modify records, thereby ensuring both data security and convenient access.

b. Operations Decision-Making: Taking into account a locomotive’s health score, fault history, and current condition, the system supports flexible maintenance planning under different priorities or operational demands. This approach facilitates resource optimization by helping managers appropriately schedule maintenance efforts and allocate repair resources efficiently.

c. Report Generation: Users can generate daily work reports or operation charts on demand, enabling a quick, visual representation of the locomotive’s workflow, content, and location data. This functionality streamlines the maintenance staff’s understanding of the locomotive’s daily operations. Additionally, the system supports date-based retrieval of historical reports.

(3) Locomotive Management

a. Status Display: Presents detailed information for each shunting locomotive, covering mileage, hours of operation, factory date, model specifications, current speed, location, and recent key maintenance history.

b. Ledger Management: Displays and manages fundamental data, factory information, and maintenance records for individual locomotives, ensuring timely updates and archival maintenance by managerial personnel.

c. Health Scoring: Employs a health assessment algorithm to generate a quantified health index for each locomotive, integrating its real-time condition with historical maintenance data. The results are shown in an intuitive format, enabling managers to swiftly gauge locomotive status.

d. History Display: By using a timeline layout, the system visually highlights significant events and dates throughout each locomotive’s lifecycle. This feature aids in analyzing long-term performance and predicting future health levels.

(4) User Management

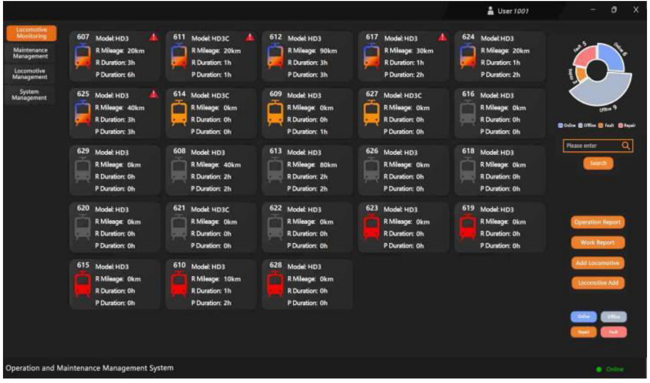
a. Account Management: Allows administrator users to manage other system accounts—creating new accounts, updating existing information, recovering lost credentials, and resetting passwords. Administrators may also freeze or remove accounts belonging to employees on extended leave or those who have departed the organization.

b. Permission Management: Administrators can assign appropriate permissions to different users, configuring which system functionalities, data, and operations can be accessed or modified. The system supports operation auditing and rollback to prevent errors arising from misuse or mismanagement.

c. System Settings: Configures inspection reminders and alerts for different shunting locomotives. Users may tailor maintenance intervals according to specific operational needs and opt for the most suitable alert formats as required.

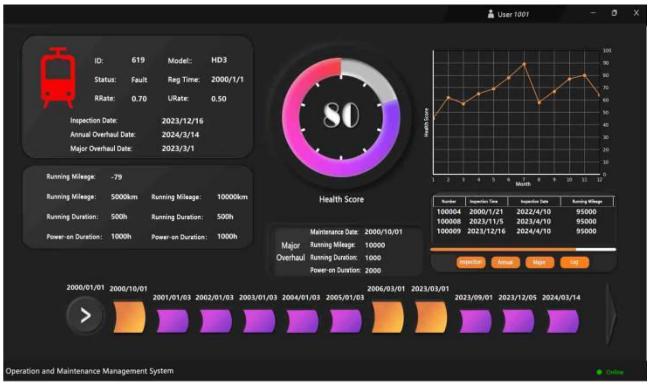
B. System Application

The shunting locomotive health assessment and maintenance management system for coal mining enterprises has been fully designed and entered a trial phase. Partial system interfaces are shown in Figures 6-8.



607 Model: HD3 R Mileage: 20km R Duration: 3h P Duration: 6h
611 Model: HD3C R Mileage: 20km R Duration: 1h P Duration: 1h
612 Model: HD3 R Mileage: 90km R Duration: 3h P Duration: 3h
617 Model: HD3 R Mileage: 30km R Duration: 1h P Duration: 2h
624 Model: HD3 R Mileage: 20km R Duration: 1h P Duration: 2h
625 Model: HD3 R Mileage: 40km R Duration: 3h P Duration: 3h
614 Model: HD3C R Mileage: 0km R Duration: 0h P Duration: 0h
609 Model: HD3 R Mileage: 0km R Duration: 0h P Duration: 1h
627 Model: HD3C R Mileage: 0km R Duration: 0h P Duration: 0h
616 Model: HD3 R Mileage: 0km R Duration: 0h P Duration: 0h
629 Model: HD3 R Mileage: 0km R Duration: 0h P Duration: 0h
608 Model: HD3 R Mileage: 40km R Duration: 2h P Duration: 2h
613 Model: HD3 R Mileage: 80km R Duration: 2h P Duration: 2h
626 Model: HD3 R Mileage: 0km R Duration: 0h P Duration: 0h
618 Model: HD3 R Mileage: 0km R Duration: 0h P Duration: 0h
620 Model: HD3 R Mileage: 0km R Duration: 0h P Duration: 0h
621 Model: HD3C R Mileage: 0km R Duration: 0h P Duration: 0h
622 Model: HD3 R Mileage: 0km R Duration: 0h P Duration: 0h
623 Model: HD3 R Mileage: 0km R Duration: 0h P Duration: 0h
619 Model: HD3 R Mileage: 0km R Duration: 0h P Duration: 0h
615 Model: HD3 R Mileage: 0km R Duration: 0h P Duration: 0h
610 Model: HD3 R Mileage: 10km R Duration: 1h P Duration: 2h
628 Model: HD3 R Mileage: 0km R Duration: 0h P Duration: 0h

Fig 6. Locomotive Monitoring Interface



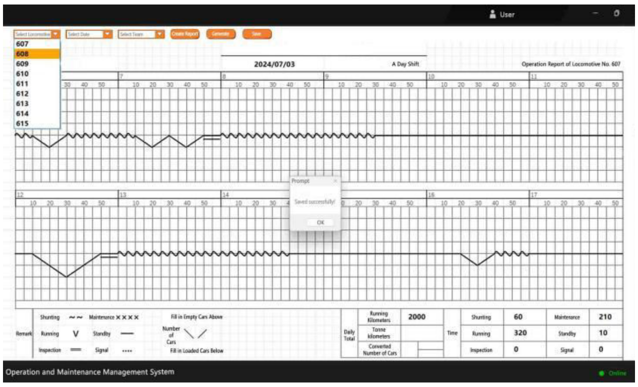
ID: 619 Model:: HD3
Status: Fault Reg Time: 2000/1/1
RRate: 0.70 URate: 0.50
Inspection Date: 2023/12/16
Annual Overhaul Date: 2024/3/14
Major Overhaul Date: 2023/3/1

Running Mileage: -79
Running Mileage: 5000km Running Mileage: 10000km
Running Duration: 500h Running Duration: 500h
Power-on Duration: 1000h Power-on Duration: 1000h

Major Overhaul
Maintenance Date: 2000/10/01
Running Mileage: 10000
Running Duration: 1000
Power-on Duration: 2000

Number	Inspection Time	Inspection Date	Running Mileage
100004	2000/1/21	2022/4/10	95000
100008	2023/11/5	2023/4/10	95000
100009	2023/12/16	2024/4/10	95000

Fig 7. Single-column monitoring interface



2024/07/03 A Day Shift Operation Report of Locomotive No. 607

Remark: Shunting ~~ Maintenance ×××× ; Running V Standby — ; Inspection ═ Signal ····
Fill in Empty Cars Above
Number of Cars
Fill in Loaded Cars Below

Daily Total	Running Kilometers	2000	Time	Shunting	60	Maintenance	210
	Tonne kilometers			Running	320	Standby	10
	Converted Number of Cars			Inspection	0	Signal	0

Fig 8. Work chart generation

(1) Digitized Locomotive Ledger Management: Effectively resolves issues of inconvenient information retrieval and unstandardized formats inherent to traditional locomotive ledger practices in coal mining enterprises. By consolidating relevant data within a centralized database, the system can integrate and synchronize with other locomotive monitoring platforms via data interfaces and shared databases, ensuring real-time connectivity of operation and maintenance data.

(2) Automated Generation of Locomotive Operation Charts: Achieves automated production of daily reports and charts, enhancing accuracy while substantially reducing manual charting efforts. This approach boosts work efficiency and facilitates future lookups by date.

(3) Operations Data – Driven Health Assessment: Tailored to the practical environment of coal mining enterprises, the system leverages historical shunting locomotive operational data in conjunction with machine learning and big data analytics to construct a health assessment model. This feature aids maintenance personnel in precisely gauging a locomotive ’ s overall condition. Moreover, as more operational data accumulate over time, the model can employ incremental learning to iteratively update its parameters, further refining its alignment with real-world conditions.

V. CONCLUSION

This study designed and developed a health assessment system for shunting locomotives in coal mining enterprises, underpinned by a random forest algorithm. The system offers a novel approach to addressing the health evaluation challenges faced by coal mining enterprises. Future work will focus on further optimizing the random forest model and exploring new methods for multi-source data integration and intelligent maintenance decision-making. Such advances aim to deliver more precise assessments and warnings concerning shunting operations in coal mining enterprises, ultimately fostering smarter locomotive management practices.

REFERENCES

[1] H.-Q. Jin, P.-A. Chen, X. Xue, et al., "Condition-based maintenance analysis for intelligent operation and maintenance technology in rail transit vehicles (Periodical style—Chinese)," Technology & Market, vol. 31, no. 03, 2024, pp. 41–43+47..

[2] Z.-Y. Liao, J.-M. Deng, Y. Shu, et al., "Status and development trends of PHM technology for rail transit equipment (Periodical style —Chinese)," Electric Locomotives & Mass Transit Vehicles, vol. 47, no. 03, 2024, pp. 8 – 16, doi: 10.16212/j.cnki.1672-1187.2024.03.002.

[3] C.-L. Li, B.-C. Yu, X. Li, et al., "Overall design of the fault prediction and health management system for Hexie locomotives (Periodical style—Chinese)," Railway Computer Application, vol. 32, no. 02, 2023, pp. 23–27.

[4] Y. Yan, J. Wang, J.-G. Wang, et al., "Health assessment system for the rotating components of locomotive running gear (Periodical style —Chinese)," Railway Computer Application, vol. 33, no. 03, 2024, pp. 59–66.

[5] B.-C. Yu, Y.-B. Ning, W.-J. Zhang, et al., "Design and research on a health monitoring platform for Hexie locomotives (Periodical style—Chinese)," Railway Vehicle, vol. 59, no. 06, 2021, pp. 105−108.

[6] X.-R. Xiao, Analysis of factors affecting the severity of traffic accidents based on random forest (Doctoral dissertation, Book style—Chinese). Henan Agricultural University, 2024, doi: 10.27117/d.cnki.ghenu.2024.000164.

[7] W. Wang, ″Analysis and prediction of pumping unit well system efficiency based on random forest regression (Periodical style—Chinese),″ Petroleum & Petrochemical Energy & Management and Measurement, vol. 14, no. 08, 2024, pp. 1–5.

[8] W. C. Zhang, ″Compare linear regression, decision tree regressor, and random forest regressor based on Python, a restaurant company on Kaggle as a case (Periodical style),″ BCP Business & Management, no. 36, 2023, pp. 322–329.

[9] Z. Zheng, Q. Cheng, and Y. Zhang, ″A comparative analysis of linear regression, neural networks and random forest regression for predicting air ozone employing soft sensor models (Periodical style),″ Scientific Reports, vol. 13, no. 1, 2023, p. 22420.

[10] Z.-W. Wu, ″Study on the prediction model of wagon final stay time at railway freight stations (Periodical style—Chinese),″ Railway Computer Application, vol. 33, no. 09, 2024, pp. 12–16.