Journal of Computers Vol. 36 No. 5, October 2025, pp. 261-279
https://doi.org/10.63367/199115992025103605018

Design of Intelligent Decision-Making System for the Remaining Life of Various Hard Disks

Quanhai Wang[1], Sanyuan Wang[2], Xiaohao Wang[2], Fahai Wang[2], Yiqun Wang[2], and Wenbai Chen[2*]

[1] Northeastern University at Qinhuangdao, Qinhuangdao 066004, China
1000777@neuq.edu.cn

[2] School of automation, Beijing Information Science and Technology University, Beijing 100192, China
chenwb@bistu.edu.cn

Received 1 July 2025; Revised 24 July 2025; Accepted 2 August 2025

Abstract. As the demand for data storage grows, preventing hard disk failures becomes critical to ensure data security and business continuity. Traditional Remaining Useful Life (RUL) prediction methods often struggle with handling heterogeneous data and sample imbalances, particularly when applied to both hard disk drives (HDD) and solid-state drives (SSD). This paper proposes an effective RUL prediction and intelligent decision-making methodology tailored to multi-class HDD in data center applications. Using Backblaze's publicly available SMART dataset, we apply this methodology to both HDD and SSD. To tackle data heterogeneity and imbalance, we adopt XGBoost for selecting the critical and common SMART attributes of both HDD and SSD, while utilizing a Transformer model for multidimensional time series analysis. For the training strategy, we combine pre-training and transfer learning. The model is first pre-trained on HDD data and then fine-tuned for SSD tasks, resulting in a unified multi-class RUL prediction model. Our experimental results demonstrate that this approach outperforms traditional models in several evaluation metrics, especially in scenarios where SSD samples are limited. At the system level, we develop an intelligent decision-making system that integrates data collection, model training, failure prediction, and visual display, enabling a closed-loop process from data collection to operational decision-making. This study presents an efficient solution for managing the health of multiple types of hard disks, offering high engineering application value.

Keywords: life RUL prediction, hard disk drive, solid state drive, transformer, intelligent decision-making system

1 Introduction

With the advent of the information age, data has become a core asset, and both individual users and enterprise data centers rely extensively on hard drives to store massive amounts of data [1-3]. A hard disk failure can lead to irreversible data loss and business disruptions. Therefore, identifying potential failures and implementing preventive maintenance measures before a hard disk reaches the end of its lifespan is critical. Hard disk Remaining Useful Life (RUL) prediction technology analyzes operational data to estimate the time remaining before a disk failure occurs, allowing operation and maintenance personnel to complete data migration, backup, and replacement in advance, thereby ensuring data security and business continuity [4].

Hard drives are mainly classified into two categories: traditional mechanical hard disk drives (HDD) and solid-state drives (SSD). HDD and SSD differ significantly in terms of hardware structure and aging mechanisms, leading to heterogeneity in their monitoring data. Specifically, different vendors provide distinct sets of SMART (Self-Monitoring, Analysis, and Reporting) attributes, with HDD typically offering a broader range of monitoring parameters, while SSD have relatively fewer [5]. This inconsistency in monitoring attributes presents challenges in constructing a unified predictive model. Moreover, HDD operational status data is highly imbalanced. On the one hand, failed HDD samples are relatively scarce, while a large volume of normally functioning HDD lacks clear fault markers. On the other hand, publicly available HDD failure datasets contain significantly more failure samples for HDD than for SSD, making models prone to biases favoring HDD failure patterns. Traditional research often focuses on modeling using failure samples while neglecting healthy hard disk data, which results in models that are less effective at predicting the lifespan of healthy drives. Addressing the heterogeneity and im-

* Corresponding Author

balance of hard disk data requires novel approaches.

In recent years, data-driven methods for hard disk failure prediction and lifetime estimation have made significant progress [6-8]. Early approaches primarily employed statistical analysis and machine learning techniques to perform threshold-based assessments or simple regression on SMART parameters. However, these methods struggle to capture complex nonlinear degradation relationships. Deep learning techniques have been widely adopted for hard disk failure detection and RUL prediction due to their ability to automatically extract features and learn complex patterns [9]. For instance, one study applied Long Short-Term Memory (LSTM) networks to model hard disk health indicator sequences, achieving high prediction accuracy within tens of days before failure [10]. However, recurrent neural networks (RNN) suffer from gradient vanishing issues and inefficiencies when processing long sequences, while convolutional neural networks (CNN) are computationally efficient but struggle to capture long-range dependencies in time series data. Originally developed for machine translation, the Transformer model utilizes a self-attention mechanism to efficiently capture global dependencies in long sequences [11]. Integrating the Transformer model into multidimensional time series modeling enhances the characterization of hard disk performance degradation over long operational periods while maintaining computational efficiency. Given these advantages, this study adopts the Transformer model as the core framework for RUL prediction across multiple types of hard drives.

This paper makes three key contributions. First, to address the issue of limited SSD failure samples, a pretraining-based transfer learning strategy using the Transformer model is proposed. Large-scale HDD failure data is used for pretraining, followed by fine-tuning with a small amount of SSD data, significantly improving the accuracy of SSD lifespan predictions. Second, to mitigate the imbalance in HDD data, a unified RUL prediction model applicable to both HDD and SSD is developed. This model identifies common key monitoring indicators across both drive types through feature selection and incorporates an improved loss function to reduce prediction bias caused by the uneven distribution of samples across different lifespan intervals, thereby enhancing model robustness. Third, an intelligent decision-making system for HDD lifespan prediction is developed and deployed in practical applications. The proposed model is embedded into a data center operation and maintenance platform, enabling real-time hard disk health monitoring, early fault warnings, and maintenance recommendations, thereby significantly improving intelligent maintenance decision-making.

2 Hard Disk Condition Monitoring Data Analysis

Hard disk condition monitoring data is primarily derived from SMART logs and fault records collected during actual operation [12]. To develop a reliable hard disk RUL prediction model, it is essential first to acquire and analyze this monitoring data. In this study, we utilize the large-scale SMART dataset publicly released by Backblaze [13], a well-known cloud storage provider. This dataset contains daily operational records of tens of thousands of HDD in a data center from 2013 to 2022, covering approximately 200,000 HDD and thousands of SSD. Each record includes fundamental information about the drive (e.g., model number, storage capacity), the date of operation, a failure flag, and multiple SMART monitoring attributes—each recorded in both raw and normalized formats. The SMART attributes capture various aspects of the drive's operational status, such as read error rate, remapped sector count, power-on hours, and temperature, with hundreds of potential attribute identifiers. However, due to differences in implementation across manufacturers and drive types, the actual attributes recorded for each drive constitute only a subset of the complete set. Statistical analysis of the dataset reveals a total of 87 distinct SMART attributes recorded over the eight-year period, with 77 attributes associated with HDD and 43 with SSD. Only 45 attributes are common to both drive types, as illustrated in Fig. 1. This highlights the significant heterogeneity in the monitoring data of HDD and SSD. If all available attributes are directly used to train a unified prediction model, challenges such as high-dimensional sparsity and missing attribute values must be addressed to ensure model reliability and generalizability.

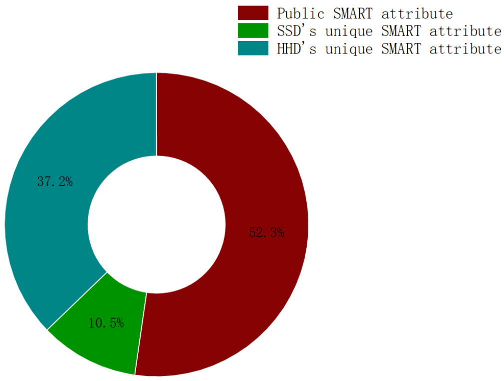


Fig. 1. Comparison of two hard disk monitoring properties

Feature engineering is essential to filter out attributes that are both useful for prediction and common to both types of hard disks [14]. On the one hand, the attributes that are too high in missing rate on most hard disks are eliminated to reduce invalid information. On the other hand, a data-driven approach is used to evaluate the correlation of attributes to hard drive failures. The model construction process utilizes the XGBoost gradient boosting tree algorithm to select features separately for HDD and SSD failure data. XGBoost aggregates the results of multiple decision trees to evaluate feature importance [15], thereby identifying SMART attributes most closely associated with HDD failures. After ranking attribute importance for HDD and SSD failure samples independently, the top-ranked attributes and their common subset are selected as input features for the unified model. Following this process, nine key SMART attributes—such as read error rate, remapped sector count, power-on hours, temperature, and cumulative read/write operations—are identified. These attributes are recorded for both HDD and SSD and exhibit a strong correlation with failures. Table 1 presents the final SMART attributes and their corresponding meanings used to construct the unified HDD failure prediction model, incorporating both raw and normalized values. The use of these shared features not only reduces data dimensionality and redundancy but also forms the foundation for the unified model. The structure of the unified HDD remaining useful life (RUL) prediction model is illustrated in Fig. 2.

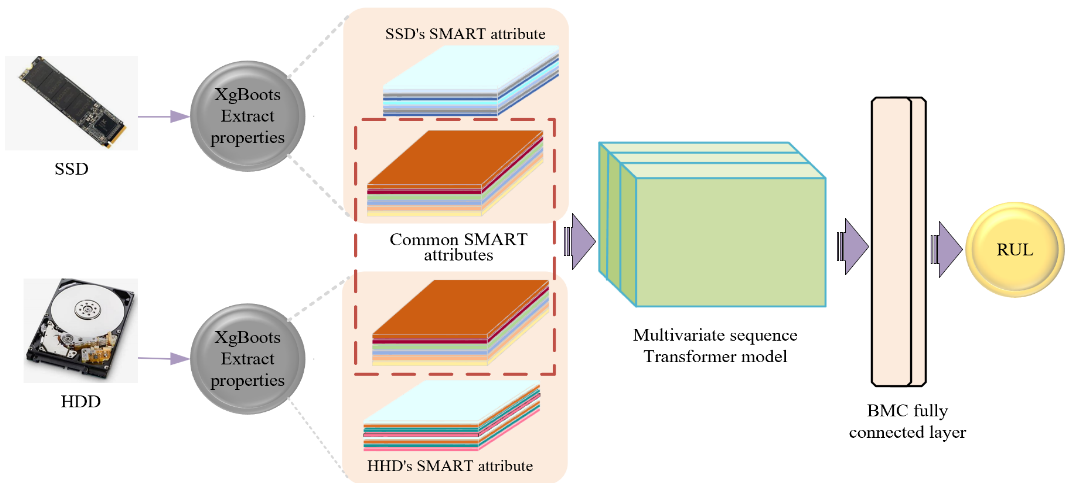


Fig. 2. Unified HDD remaining useful life prediction model structure

Table 1. Unifying the SMART attributes of hard disk remaining useful life prediction models

Causality	Name	Meaning
SMART_1	Read Error Rate	Hard disk raw read error rate
SMART_5	Reallocated Sector Count	Hard Disk Remap Sector Count
SMART_9	Power-On Hours	Hard disk power-up time since shipment
SMART_12	Power Cycle Count	Power on/off cycle counting
SMRT_190	Airflow Temperature	Air temperature
SMART_192	Power-off Retract Count	Power failure return count
SMART_194	Temperature	Celsius temperature of hard disk
SMART_241	Total LBAs Written	Total number of logical block writes
SMART_242	Total LBAs Read	Total number of logical block reads

In addition to feature heterogeneity, sample imbalance is another critical challenge in hard disk data analysis. First, failed hard disks constitute a minority class relative to normal hard disks [16]. If the training set includes only the life cycle data of failed hard disks for regression training, the model may produce significant errors when predicting the remaining useful life (RUL) of healthy hard disks (i.e., those far from failure). To enhance the model's understanding of the health state, this study incorporates a certain percentage of healthy hard disk data into the training process. A predefined RUL threshold is set, and failure samples exceeding this threshold are categorized as "sub-healthy." These sub-healthy samples are then combined with data from long-running, failure-free hard disks to form a healthy sample set. By integrating both healthy and failure data into the training process, the model can learn the differences in HDD performance indicators between normal operating conditions and near-failure states, thereby mitigating excessive reliance on the limited failure samples. Second, there is a substantial imbalance between the number of HDD and SSD failure samples. Directly combining these datasets during training can result in the model being biased toward the dominant HDD failure patterns. To mitigate this issue, HDD and SSD data are sampled in appropriate proportions within the training set, or SSD samples are assigned higher weights to ensure that the model considers the prediction accuracy for both types of hard disks. Additionally, the model optimization objective is refined by adopting an improved loss function to address the uneven distribution of regression labels. Specifically, Balanced Mean Squared Error (Balanced MSE) loss is introduced as an enhancement to traditional Mean Squared Error (MSE) [17]. This approach assigns different penalty weights to residuals in different RUL intervals, enabling the model to achieve robust training across both short-life (near failure) and long-life (healthy) regions. This adjustment statistically balances prediction biases. Through the aforementioned feature engineering and data balancing strategies, high-quality input data and a well-structured training sample distribution are ensured, facilitating the development of a unified RUL prediction model applicable to multiple types of hard disks.

3 Transformer-based RUL Prediction Method for Multi-species Hard Disks

3.1 Transformer-based Time Series Prediction

The Transformer is an advanced neural network architecture proposed by Vaswani et al. in 2017. It has been widely applied in sequence processing tasks, demonstrating significantly superior performance compared to traditional recurrent neural networks (RNN) and temporal convolutional networks (TCN), particularly in fields such as machine translation, language understanding, sequence prediction, music generation, and multimodal sentiment analysis. Unlike RNN and TCN, the Transformer does not rely on recursion or convolution to process sequences. Instead, it is based on a feedforward model that processes the entire sequence in parallel. The Transformer follows an encoder-decoder architecture, consisting of multiple stacked layers of encoders and decoders. Its key components include Multi-Head Self-Attention, Position-Wise Feed-Forward Networks, Layer Normalization, and Residual Connections. Among these components, the self-attention mechanism simulates the visual attention behavior of biological organisms, enabling the model to focus on the most relevant parts of the input sequence. The multi-head self-attention mechanism allows the model to capture multiple relationships between elements in the sequence simultaneously in different representation spaces, thereby enhancing feature extraction capabilities.

In the Transformer architecture, the multi-head self-attention mechanism functions similarly to the multi-kernel convolution operation in Convolutional Neural Networks (CNN) [18]. However, it is better suited for capturing global contextual information, addressing CNN' limitation of a restricted receptive field, and enhancing the model's ability to handle long-range dependencies within sequences, as illustrated in Fig. 3. Additionally, the Transformer employs a masking technique to regulate the flow of input and output information during model training and inference. In the decoder, this masking technique ensures that the model only utilizes previously generated information when predicting the next token in a sequence. This prevents the use of future, yet-to-be-generated information, thereby mitigating information leakage. This masking mechanism is particularly crucial in autoregressive models, such as language models, as it ensures the causality of predictions and maintains the validity of model inferences.

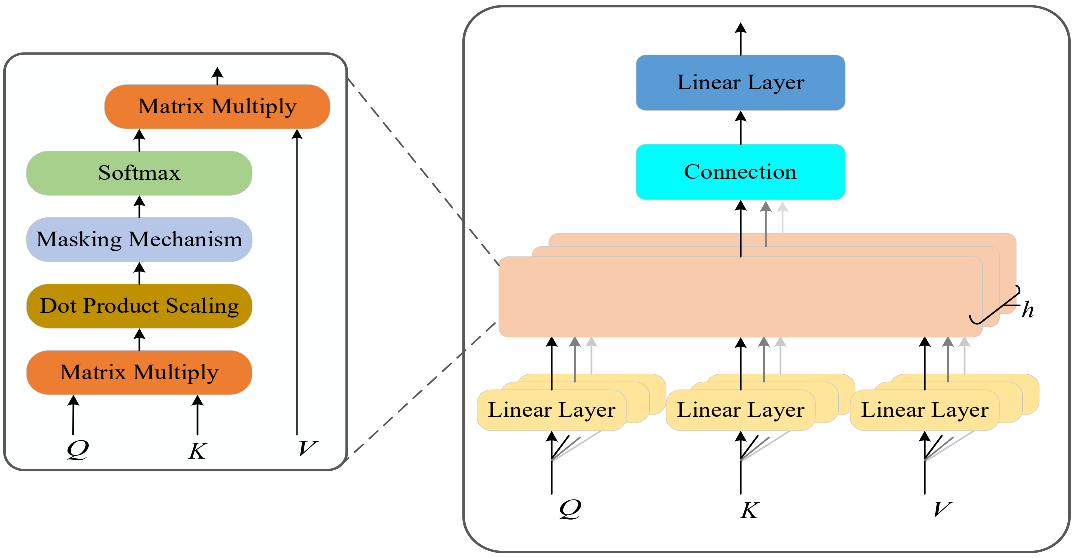


Fig. 3. Transformer multi-head attention mechanism

In the field of Remaining Useful Life (RUL) prediction for hard disks, the Transformer model has gradually become one of the mainstream approaches due to its powerful parallel processing capabilities and superior performance in modeling multi-dimensional and long-sequence data. Similar to convolutional neural networks, the Transformer benefits from efficient parallel computation. However, it significantly outperforms recurrent neural networks in capturing long-range dependencies and processing multi-dimensional time series data [19]. By leveraging these advantages, the Transformer can extract feature information from SMART monitoring data more accurately and efficiently, thereby providing a more precise assessment of hard disk health status.

3.2 RUL Prediction Model for Various Hard Disk Types

After completing data preprocessing and feature selection, this paper constructs a hard disk RUL prediction model based on the Transformer architecture. As a sequence modeling framework centered on the self-attention mechanism, Transformer is capable of capturing long-range dependencies between time steps in a sequence while processing time series. It offers high parallel efficiency and is well-suited for multidimensional long-sequence regression analysis. In this study, hard disk SMART monitoring data is treated as a multivariate time series. For each hard disk, a set of daily monitoring features—including the raw and normalized values of the nine selected attributes (totaling 18 dimensions)—is arranged in chronological order, forming a sequence of feature vectors evolving over time. The model processes these sequences using a sliding time window of fixed length, collecting recent monitoring data as input and outputting the predicted remaining useful life (RUL) at the corresponding moment. Since hard disk failures occur progressively, the model must capture potential performance degradation patterns from the input sequences and map them to RUL values. Through its multi-head self-attention mechanism, Transformer effectively accounts for the feature correlations across the entire time window, automatically emphasizing key signals indicative of remaining lifespan (e.g., the continuous deterioration of specific attributes). Compared to traditional recurrent neural networks, the Transformer-based approach achieves more accurate RUL regression by better identifying critical degradation trends in the data.

To address the challenge of poor direct training performance due to the limited number of SSD failure samples, this paper adopts a pre-training and transfer fine-tuning strategy for model training. First, a large amount of HDD failure data from the Backblaze dataset is used to pre-train the Transformer model. The HDD dataset provides a sufficient volume of data, with failure samples covering a more comprehensive range of life intervals. Through pre-training, the model learns the general degradation patterns leading to HDD failure, resulting in an initial life prediction model. During pre-training, the model takes the monitoring sequences of HDD as input, with the target output being the number of days remaining until HDD failure (i.e., the true RUL) at each corresponding time step. The Transformer parameters are optimized by minimizing the error between the predicted RUL and the ground truth. After several training iterations, once the HDD training error converges, the model acquires the foundational ability to map HDD SMART sequences to RUL values. Next, the pre-trained model parameters are transferred to the SSD prediction task, where the model undergoes fine-tuning using a limited set of SSD failure data. During fine-tuning, the model structure remains unchanged, but certain network weights are adjusted to adapt to the specific characteristics of SSD. Since SSD and HDD share many health monitoring indicators and exhibit similar failure mechanisms to some extent, this transfer learning strategy effectively leverages the data-rich HDD samples to compensate for the scarcity of SSD failure samples, thereby significantly enhancing the accuracy of SSD RUL prediction. Experimental results demonstrate that the model fine-tuned with HDD pre-training achieves lower RUL prediction errors on the SSD test set. Additionally, several evaluation metrics show significant improvements compared to the model trained directly on SSD data alone.

Building upon the improved prediction capability for SSD, this paper further proposes a unified RUL prediction model for multiple types of hard disks, employing a single Transformer network to simultaneously predict the remaining useful life (RUL) of both HDD and SSD. To achieve this, the challenges of data heterogeneity and imbalance must be addressed. First, a common set of attributes, derived through feature engineering, is used as the model input, enabling the Transformer to process monitoring data from both HDD and SSD. Second, during model training, HDD and SSD data are integrated, and strategies such as healthy sample mixing and loss balancing are applied to ensure fair training. The unified model is trained on a comprehensive dataset comprising HDD and SSD failure samples as well as healthy samples, as compiled in the previous section. Within each training batch, HDD and SSD samples are mixed and fed into the model according to a predefined ratio. The loss function employs Balanced MSE, which separately weights and accumulates errors from different types of samples to simultaneously optimize the prediction performance for both HDD and SSD. It is important to note that due to the different actual lifespan distributions of HDD and SSD (e.g., SSD generally operate for a shorter duration before experiencing failure compared to the longest-lived HDD in the dataset), a maximum RUL threshold (e.g., 30 days) is imposed. This truncates excessively long prediction ranges during unified training, allowing the model to focus on short-term degradation patterns rather than minor variations in long-term lifespan segments. By implementing these measures, the resulting Transformer model can be trained to predict the RUL of both HDD and SSD using a unified set of network weights. The model demonstrates strong uniformity and generalization capability, as it can directly output the predicted RUL of any given hard disk without requiring prior knowledge of its type. While ensuring high prediction accuracy, this approach enhances model generality across multiple device types, facilitating practical deployment.

To provide a more rigorous understanding of the underlying mechanism of the proposed model, we present a mathematical formulation of the Transformer-based time series representation. This is particularly helpful for understanding how the model captures temporal dependencies and contextual interactions in hard disk performance data.

Transformer-based Time Series Modeling:

Let $X = [x_1, x_2, ..., x_t]$, where $x_t \in R^d$, be the input time series. The self-attention mechanism computes attention as:

$$Attention(Q, K, V) = softmax\left(QK^T \big/ \sqrt{d_K} \cdot V \right) \quad \textbf{(1)}$$

Let: $Q = X \cdot W^Q$, $K = X \cdot W^K$, $V = X \cdot W^V$, where W^Q, W^K, $W^V \in R^{(d \times d_k)}$. These projections transform the input into query, key, and value vectors for context-aware learning.

The scaled dot-product attention mechanism enables the model to weigh each time step according to its relevance to others, allowing both short- and long-term dependencies to be captured. The resulting contextualized representation is then passed through feedforward and normalization layers, forming the basis of the model's prediction capability.

4 Experimentation and Evaluation

4.1 Model Training Setup

To validate the effectiveness of the proposed multi-type HDD RUL prediction method, extensive experimental analyses were conducted. The dataset used in the experiments is derived from Backblaze HDD monitoring records [20-21]. Specifically, HDD data from 2015–2017 was selected for model pre-training, while SSD data from 2018–2022 was used for fine-tuning and testing. During the pre-training phase, 4,387 failed HDD samples were extracted, with each HDD containing a sequential record of its daily monitoring data from initial operation to failure. In the fine-tuning phase, 145 SSD failure samples were selected, and 19 SMART attributes (corresponding to the raw and normalized values of 9 attributes) common to both HDD and SSD were retained as input features. Among these, monitoring sequences from 129 SSD were used for training, while the remaining samples were reserved for test evaluation. For unified model training, data from both HDD and SSD were combined. Given that the number of SSD samples is significantly smaller than that of HDD and that SSD exhibit shorter lifespan distributions, SSD samples were appropriately oversampled. Additionally, the maximum RUL prediction was capped at 30 days to emphasize short-term lifespan prediction. The parameter settings for the Transformer-based unified HDD RUL prediction model are detailed in Table 2.

Table 2. Unified hard disk remaining useful life prediction model parameters

Parameters	Value	Parameters	Value
Number of features/ *m*	18	Dimensions of feed forward networks	512
Sample length/ *w*	30	Activation function	gelu
Number of validation sets	200	Learning rate	0.001
Input vector length	128	Batch size	64
Number of long attention spans	16	Maximum parameters for gradient trimming	0.7
Number of subencoders	9	Normalization	batch norm

In order to address the imbalance between failure and non-failure samples, we adopt a class-weighted loss function to penalize misclassification of rare events more heavily. This helps the model focus on learning patterns from underrepresented failure cases.

Loss Function Weighting Under Class Imbalance:

$$L\left(y, \hat{y}\right) = -\sum_{i} \alpha_i \cdot y_i \cdot \log\left(\hat{y}_i\right) \quad (2)$$

where the class weight α_i is defined as:

$$\alpha_i = \frac{1}{\log\left(c + p_i\right)} \tag{3}$$

Here, p_i is the prior probability of class i, and c is a smoothing constant used to prevent division by zero (typically c = 1.02).This weighting scheme enhances gradient contributions from minority classes while suppressing the dominance of frequent ones, thereby improving prediction robustness.

4.2 Assessment of Indicators

To comprehensively evaluate the performance of the model, several commonly used regression metrics are employed, including mean squared error (MSE), root mean squared error (RMSE), mean absolute error (MAE), mean absolute percentage error (MAPE), symmetric mean absolute percentage error (SMAPE), and the coefficient of determination (R^2). The formulas for each evaluation metric are presented below:

$$\mathcal{L}_{\text{MSE}}\left(S, S^*\right) = \frac{1}{N}\sum_{i=1}^{N}\left(x_i - y_i\right)^2 \tag{4}$$

$$\mathcal{L}_{\text{RMSE}}\left(S, S^*\right) = \sqrt{\frac{1}{N}\sum_{i=1}^{N}\left(x_i - y_i\right)^2} \tag{5}$$

$$\mathcal{L}_{\text{MAE}}\left(S, S^*\right) = \frac{1}{N}\sum_{i=1}^{N}\left|x_i - y_i\right| \tag{6}$$

$$\mathcal{L}_{\text{MAPE}}\left(S, S^*\right) = \frac{1}{N}\sum_{i=1}^{N}\left|\frac{y_i - x_i}{y_i}\right| \tag{7}$$

$$\mathcal{L}_{\text{R}^2}\left(S, S^*\right) = 1 - \frac{\sum_{i=1}^{N}\left(x_i - y_i\right)^2}{\sum_{i=1}^{N}\left(y_i - \overline{y}\right)^2}, \overline{y} = \frac{1}{N}\sum_{i=1}^{N} y_i \tag{8}$$

Here, $S = [x_1, ..., x_i, ..., x_N]$ and $S^* = [\mathrm{y}_1, ..., y_i, ..., y_N]$ represent the prediction vector and the actual observation, respectively. x_i denotes the ith predicted value, y_i represents the ith actual value, and N is the total number of prediction vectors. The range of MAE, MSE, RMSE, and MAPE in these formulas is $[0, +\infty]$.

When used as a loss function in regression tasks, a smaller output value generally indicates a smaller discrepancy between the predicted and actual values, suggesting higher model accuracy. Additionally, the range of the R2 evaluation metric is [0,1]. Typically, the closer the R2 value is to 1, the better the model's fit.

In addition to conventional loss metrics, we introduce a Balanced Mean Squared Error (BMSE) to improve the model's sensitivity to early failures and outliers. BMSE accounts for the variability in each sample by normalizing prediction errors using a learned or estimated variance term. Mathematical Justification for Balanced MSE:

$$BMSE = \frac{1}{n}\sum_{i}\left(\frac{y_i - \hat{y}_i}{\sigma_i}\right)^2 \tag{9}$$

where σ_i is the standard deviation of the i-th sample, estimated from historical SMART features or fitted residuals.This formulation prevents the loss function from being overly biased by high-variance samples and facilitates more balanced training convergence across varying data distributions.

4.3 Experimental Results and Analysis

To mitigate dataset imbalance, the final training dataset for hard disk drives consisted of 5,000 samples, including 4,000 faulty drives and 1,000 normal, yet-to-fail drives, with a more even distribution across different manufacturers. The SSD training dataset contained 200 samples, comprising 100 faulty and 100 healthy drives. For healthy drives, at least 15 consecutive days of monitoring data were randomly selected. A masking operation was then applied to extend the sequence to 30 days, setting the remaining useful life (RUL) to 30. For failed drives, input data were primarily taken from the 45-day period preceding failure. A minimum of 15 days of data were extracted, and the masking mechanism was similarly applied to extend sequences to 30 days. The final day of valid data was used as the ground truth for RUL. Additionally, 200 samples were selected for validation, consisting of 50 healthy SSD, 50 healthy HDD, 50 failed SSD, and 50 failed HDD. After 4,000 training iterations, the training and validation errors are presented in Fig. 4.

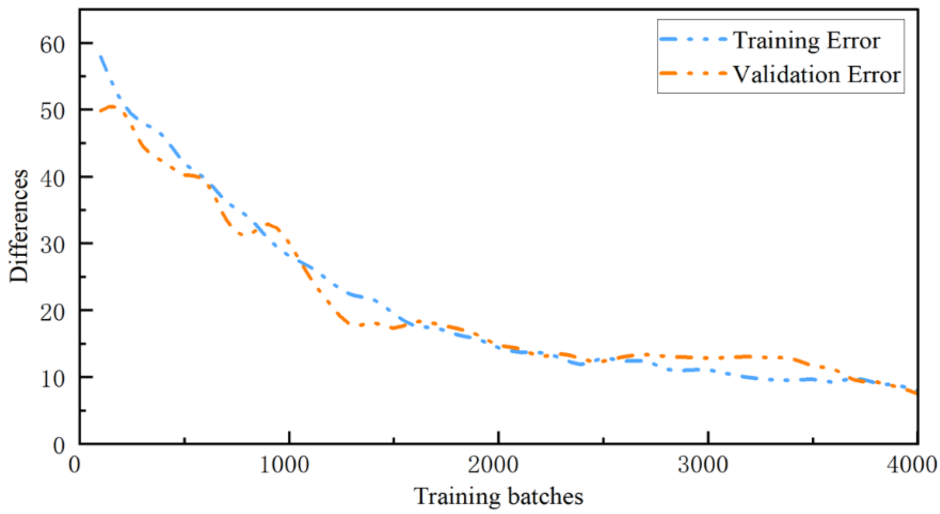
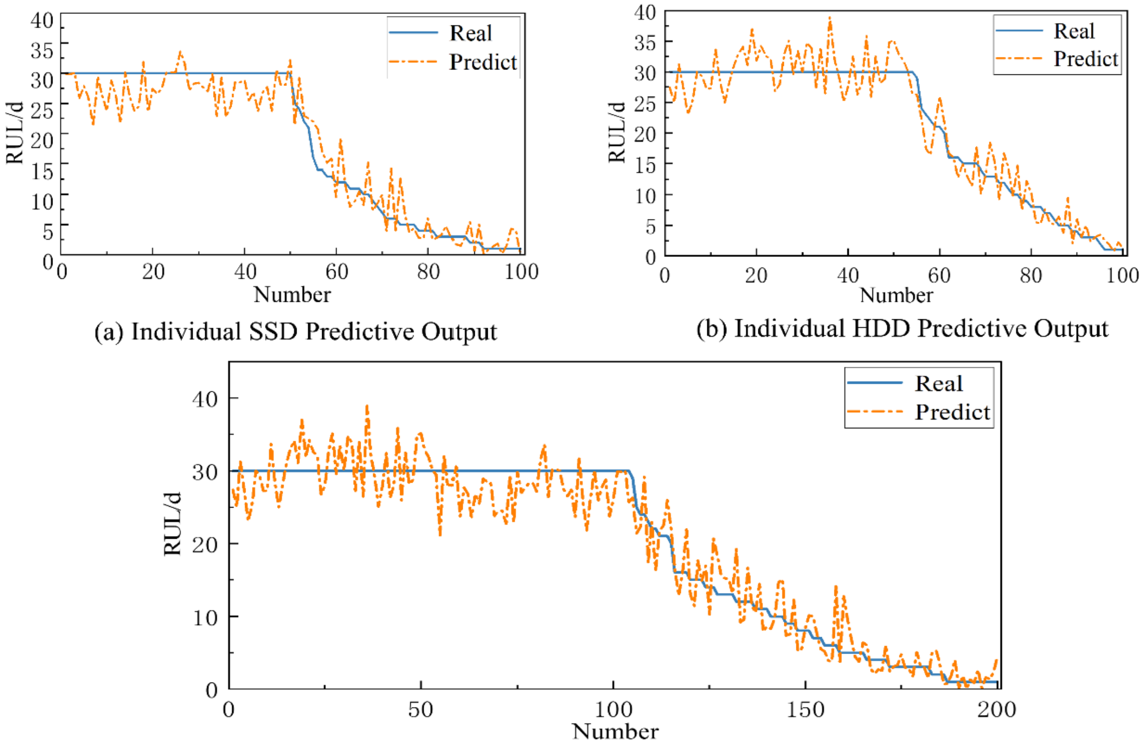


Fig. 4. Predicting model training error and validation error

The training error and validation error gradually decrease and eventually plateau, indicating that the unified model training has been completed. Subsequently, a test dataset comprising 200 samples is constructed, including 50 SSD failure samples, 50 normal SSD samples, 50 hard disk drive (HDD) failure samples, and 50 healthy HDD samples. The prediction results for individual SSD, individual HDD, and the overall hard disk performance are illustrated in Fig. 5.

To demonstrate the feasibility of the proposed Transformer-based remaining useful life (RUL) prediction method, which has been improved to address data imbalance, this study compares the proposed approach with recent RUL prediction methods for hard disk drives (HDD). The model's performance is evaluated using multiple evaluation metrics, and the comparative results with commonly used RUL prediction methods are presented in Table 3.



(a) Individual SSD Predictive Output

(b) Individual HDD Predictive Output

(c) Two Kinds of Hard Disk Overall Prediction Output

Fig. 5. Model predicted output results

Table 3. Comparison results of unified HDD remaining useful life prediction models

Model	Maximum RUL	Evaluation index				
		MSE	RMSE	MAE	MAPE	R^2
LSTM (HDD)	100	-	-	9.310	-	-
CNN (HDD)	350	-	64.394	-	-	-
LSTM (HDD)	40	-	-	7.940	-	-
RNN (HDD)	40	-	-	10.860	-	-
Improved Random Forest	-	-	104.970	-	-	0.887
Direct Training (SSD)	40	4.678	2.163	1.435	0.247	0.896
Pre-Training (SSD)	40	3.272	1.809	1.113	0.171	0.950
Unified Model (SSD)	30	11.955	3.458	2.643	0.322	0.920
UnifiedModel (HDD)	30	10.162	3.188	2.632	0.199	0.914
Unified Model (All)	30	11.059	3.325	2.638	0.261	0.918

As shown in Table 3, the proposed unified remaining useful life (RUL) prediction model effectively addresses data imbalance in hard disk lifespan prediction. The model not only achieves high accuracy in predicting the RUL of hard disk drives (HDD) but also performs well in estimating the RUL of solid-state drives (SSD), despite the limited availability of SSD failure data. By tackling both the imbalance between SMART failure and health data as well as the disparity between HDD and SSD monitoring data, the model demonstrates outstanding predictive accuracy, particularly in short-term RUL estimation. The proposed imbalance-handling strategy proves to be highly promising for RUL prediction tasks in hard disk systems.

5 Design and Implementation of an Intelligent Decision-Making System

To facilitate the practical application of the HDD RUL prediction model in real-world data center operations and maintenance, this paper designs and implements an intelligent decision-making system. This system is centered on hard disk remaining useful life (RUL) prediction and integrates hardware monitoring, data processing, and a human-computer interaction interface. It provides failure warnings and maintenance decision support for operations and maintenance personnel. By offering timely alerts and enabling proactive intervention, the intelligent decision-making system helps prevent potential failures, minimizes data loss, and enhances the reliability of hard disks and other storage systems within data center operations.

5.1 Overall System Architecture

The intelligent decision-making system employs a layered modular design, with its logical hierarchy illustrated in Fig. 6. It consists of four layers, organized from bottom to top: the managed object layer, the data collection layer, the business processing layer, and the operation display layer. The managed object layer comprises the monitored and managed devices within the data center, serving as the system's primary data source. This layer primarily includes various storage devices, such as hard disk drives (HDD), solid-state drives (SSD), and storage arrays, as well as related network switches. These devices continuously generate and provide operational status data. The data collection layer is deployed either on-site at the equipment location or on a remote server. It gathers state information from the managed objects and transmits it at regular intervals via a data collection agent. For instance, in the case of hard disks, this layer collects Self-Monitoring, Analysis, and Reporting Technology (SMART) attributes, including power cycle count, temperature, and read/write error rate, as well as log information. Ensuring both data accuracy and real-time acquisition, this layer also supports customizable collection plug-ins, enabling seamless integration of data from various device types.

The business processing layer serves as the core of the system, handling the higher-level application logic. This layer comprises several functional modules, including data preprocessing, model management, fault prediction, report generation, and alarm configuration [22]. The data preprocessing module is responsible for cleaning, transforming, and storing the collected raw data to ensure that its quality meets the model's requirements. The fault prediction model management module maintains the parameters and versions of the hard disk RUL (Remaining Useful Life) prediction model, supporting both offline training and model deployment. The online fault prediction module processes real-time collected data using the latest model, generating current RUL predictions and fault risk assessments for hard disks. These results are stored in a database and further used for report generation and alarm triggering. The report and alarm module produces comprehensive and interpretable analysis reports based on the prediction results and triggers notifications according to predefined rules. The operation display layer, located at the top of the hierarchy, serves as the user interaction interface. Through a visualization dashboard, operations and maintenance (O&M) personnel can monitor an overview of the operational status of each hard disk, including remaining life expectancy, failure probability, and system-generated maintenance recommendations. This layer presents data in the form of intuitive charts and reports, allowing users to quickly identify at-risk hard drives and take appropriate actions, thereby establishing a closed-loop intelligent decision-making system.

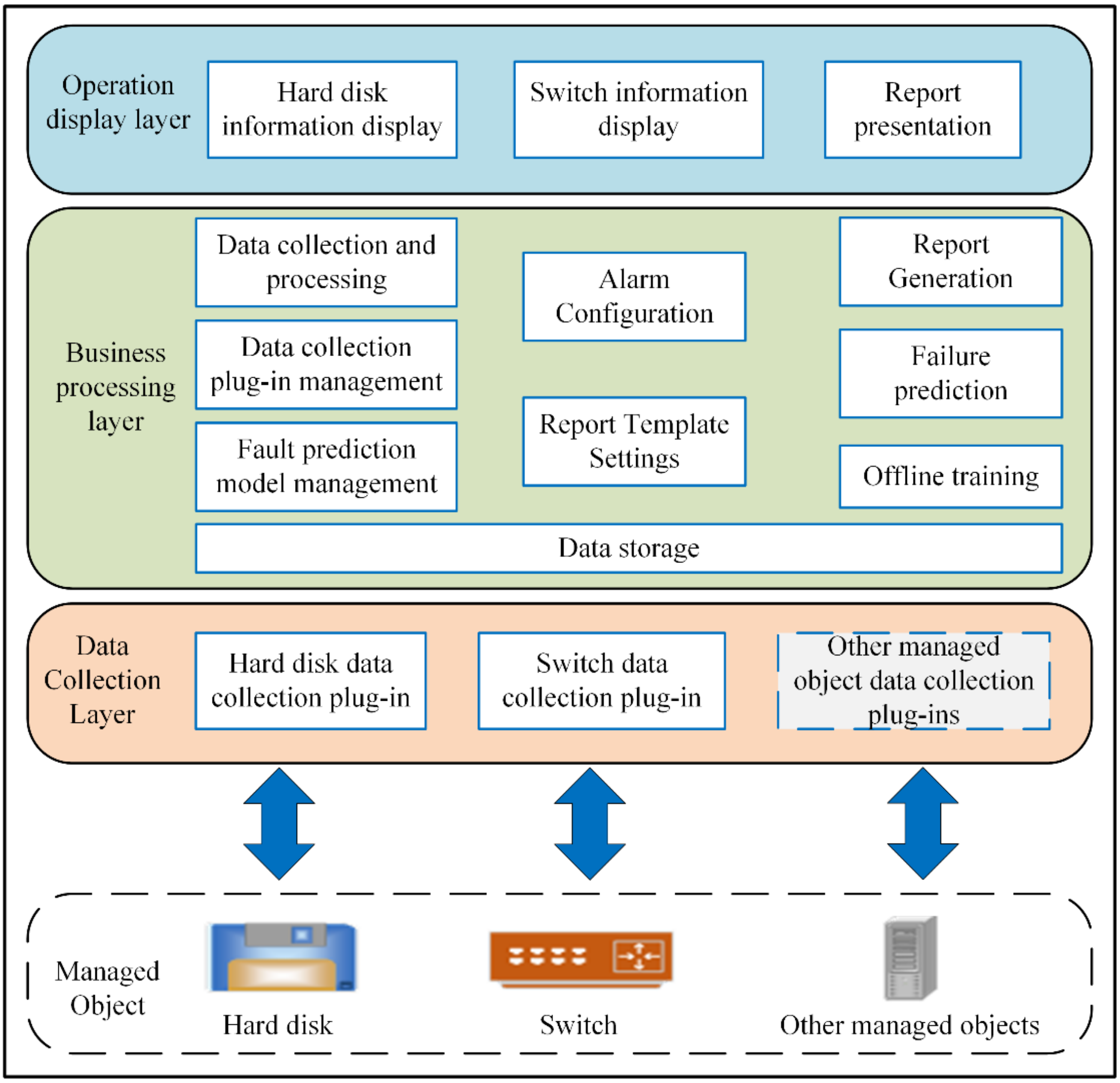


Fig. 6. Overall architecture diagram of the intelligent decision-making system

5.2 Functions and Realization

In the overall framework of the intelligent auxiliary maintenance decision-making system, the seamless integration and fusion of the managed object layer, data acquisition layer, business processing layer, and operational display layer enable the system to be structured into four logical modules: the data acquisition module, offline training module, fault prediction module, and report generation module. The specific design of each system sub-module within these modules is illustrated in Fig. 7.

In the design of the intelligent assisted maintenance decision-making system, the data acquisition module, offline training module, fault prediction module, and report generation module each have relatively independent functions and tasks. However, they must also collaborate seamlessly with other modules to form a comprehensive and cohesive intelligent maintenance decision-making system.

Acquisition Module. The acquisition module serves as the foundation of the entire system. First, the system defines the managed objects requiring data collection and configures the relevant collection parameters, including data acquisition intervals and log collection periods. Through task scheduling, instruction execution, and status monitoring, the module gathers hard disk SMART data, switch log data, and other relevant information. A collection task plan is then established, allowing for flexible scheduling—such as daily or hourly data collection—to systematically acquire the status information of managed objects. Finally, the collected raw data undergoes standardized processing before being stored in the database and pushed to the fault prediction modules.

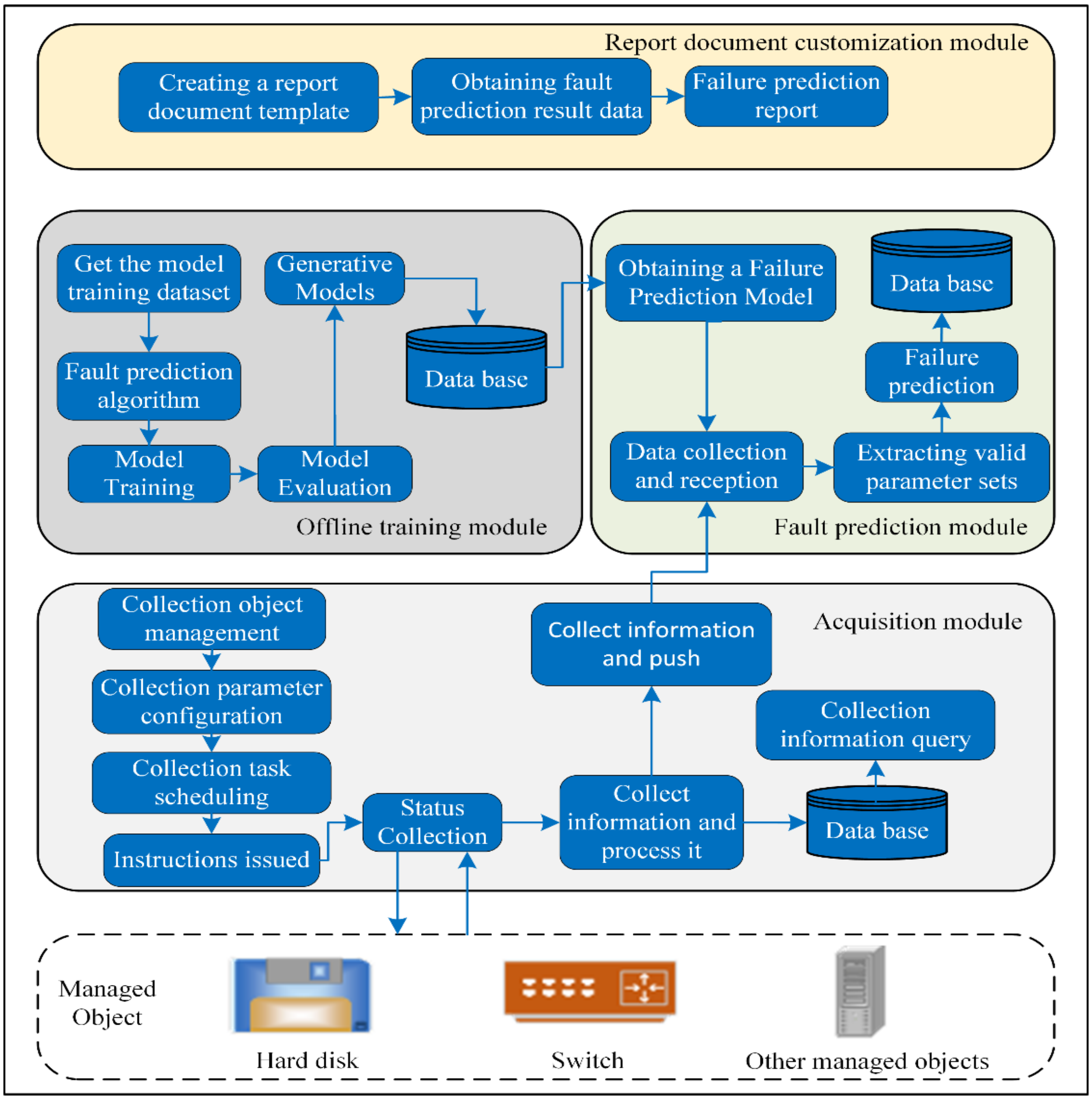


Fig. 7. System modules design structure diagram

Offline Training Module. The offline training module integrates the HDD remaining useful life (RUL) prediction methods discussed in the previous two chapters. It employs a Transformer-based neural network to train the HDD RUL prediction model using the SMART dataset. Once trained, the model is deployed to the prediction platform and stored as a backup in the database. This module primarily fine-tunes the pre-trained model using real-world data from the data center, iteratively updating and optimizing it based on predefined model settings. The optimized models serve as candidate HDD RUL prediction models for the failure prediction module.

Fault Prediction Module. The fault prediction module first selects a specific device fault prediction model from the offline training module's database. It then retrieves device status data received from the corresponding acquisition module and preprocesses it to extract meaningful attributes and features. Using the effective SMART dataset for hard disks and the remaining service life data from switch logs, the module performs failure predictions. The predicted results are then compared against predefined failure thresholds, and rule-based matching is conducted to determine potential faults.

5.3 Introduction to the Main Interface of the Intelligent Decision-Making System

The Intelligent Assisted Maintenance Decision System operates on a data-driven foundation, leveraging historical and real-time data to analyze and predict equipment status and performance. By enabling failure prediction and maintenance recommendations, the system significantly enhances equipment reliability and operational efficiency while reducing maintenance costs and failure rates, making it a valuable industrial intelligence technology. As a practical engineering application, this system can predict the remaining useful life of hard disks, switches, and other equipment, providing proactive maintenance recommendations. The software interface of the entire system is shown in Fig. 8.

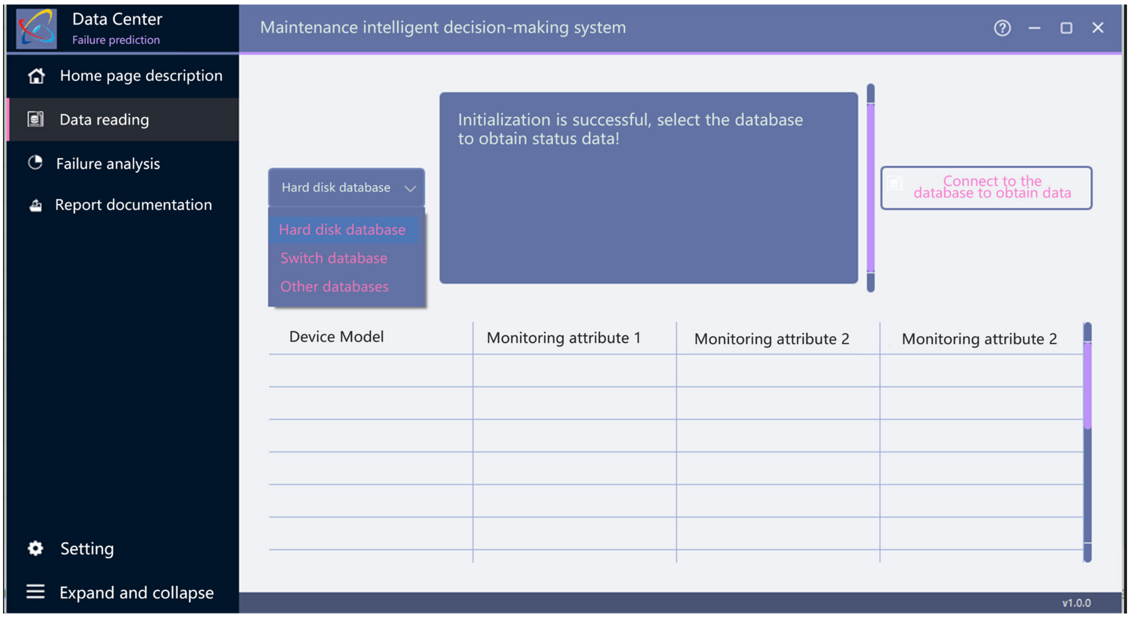


Fig. 8. Main interface of the intelligent decision-making system

The software interface of the Intelligent Assisted Maintenance Decision System is structured into five primary sections: Homepage Description, Data Reading, Fault Analysis, System Setup, and Report Documentation. The Homepage Description section serves as an introduction to the system, providing users and maintenance personnel with essential information on its functionalities and operation. This section typically includes: System Overview: A brief introduction to the system, highlighting its features and benefits. User Guide: Instructions on utilizing various system modules, including data reading, fault analysis, system configuration, and report generation. Technical Support: Access to FAQs, user manuals, and online support for troubleshooting and assistance. By navigating the Homepage Description section, users can gain a comprehensive understanding of the system's capabilities, allowing them to operate it more effectively. Additionally, the Technical Support subsection ensures users can access assistance as needed, improving both system efficiency and stability.

Secondly, the data reading section is a crucial component of the Intelligent Assisted Maintenance Decision System, enabling the system to retrieve data from multiple sources, including the latest SMART data, switch log data, and hard disk monitoring data. Users can import these datasets into the system for analysis and processing, allowing for a more comprehensive understanding of equipment status and potential faults.

Additionally, the fault analysis section represents one of the system's core functionalities, incorporating an advanced hard disk remaining useful life (RUL) prediction model and fault analysis capabilities. By setting thresholds, users can quickly identify faulty devices, analyze failure causes, and receive maintenance recommendations.

Finally, the report generation module is responsible for creating maintenance and data reports. This module typically includes various report templates and automated report generation tools, enabling users or maintenance personnel to efficiently produce different types of reports. These reports serve as records of equipment status and maintenance activities while also supporting data analysis and decision-making.

The system functions as a practical application software due to the distinct roles of its components and their seamless cooperation in realizing the intelligent assisted maintenance decision-making process. The data acquisition module enables users to retrieve information from multiple data sources, such as the latest SMART data and switch log data, ensuring a reliable foundation for fault analysis. The fault analysis module incorporates a remaining service life prediction model and analytical tools that assist users in rapidly identifying faulty equipment by setting thresholds, allowing for early fault detection and the formulation of appropriate maintenance strategies. This, in turn, enhances maintenance efficiency and accuracy. The system configuration module offers flexible settings, enabling users to tailor the system to diverse operational requirements. Finally, the report generation module facilitates the creation of various maintenance reports, including fault diagnosis reports, maintenance plans, and maintenance records, allowing users and maintenance personnel to efficiently track and manage maintenance tasks.

5.4 Status Data Display Interface

The status data display module in the Intelligent Assisted Maintenance Decision System serves as a crucial interactive component for predicting the remaining service life of hard disks. This section primarily introduces the data loading process and other relevant information within the system's hard disk failure prediction function. The hard disk status data display module consists of three main sections, as illustrated in Fig. 9. The first section is the database selection area, where users can choose the hard disk database to be accessed. The second section is the connection confirmation area, which allows users to establish a connection and retrieve real-time status data of the devices. Once successfully connected, the system continuously updates the displayed information. The third section is the status data display area, where real-time monitoring data of various hard disks are presented, enabling users to intuitively assess the operational status of different drives. The interface of the status data display module is designed to be both functional and user-friendly, ensuring that users and administrators can effortlessly access the necessary information for efficient system monitoring and maintenance.

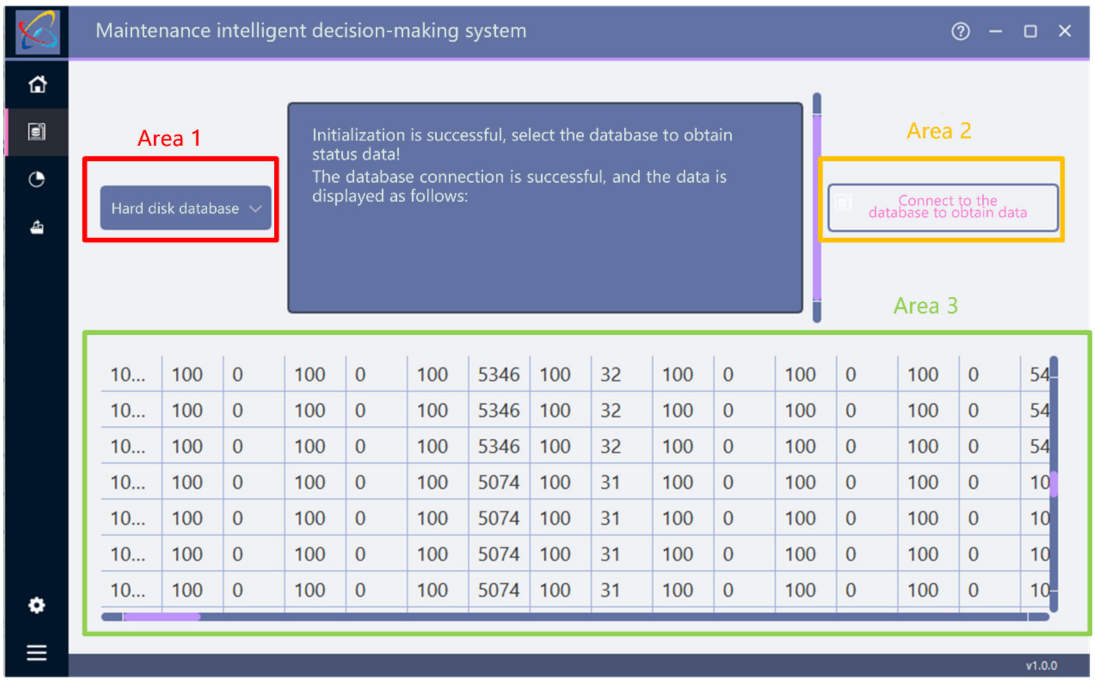


Fig. 9. The operation interface of the fault prediction module

The status data display module enables users or administrators in the data center to collect and visualize hard disk monitoring parameters at the plant layer over a specified period. These parameters include disk read and write speed, disk capacity, and disk error rate. By displaying this data, administrators can gain a clearer understanding of the operational status of hard disks. Additionally, the module provides real-time monitoring data to the fault prediction module of the Intelligent Assisted Maintenance Decision System. The fault prediction module analyzes potential failures based on the collected data and generates reports, allowing administrators to take proactive measures to address identified issues.

5.5 Fault Prediction Display Interface

As the core component of the Intelligent Assisted Maintenance Decision System, this module primarily collects data from hard disks selected by the data center administrator. Utilizing a Transformer-based remaining service life prediction model, it analyzes the processed data to estimate the remaining lifespan of the hard disk and provides corresponding maintenance recommendations. The interface layout of the hard disk fault prediction module within the system software, as shown in Fig. 10, is divided into three main sections. On the left, the Hard Disk Device List displays the identification numbers of all hard disks stored in the database, allowing administrators to search and select a specific device for analysis. The middle section, Prediction Results, presents the failure probability estimated by the predictive model. By analyzing the failure probability, administrators can assess the health status and remaining lifespan of a given hard disk. The right section, Maintenance Recommendations, provides suggested maintenance actions based on the prediction results. These recommendations assist administrators in formulating appropriate maintenance plans to ensure the continued optimal performance of hard disk devices.

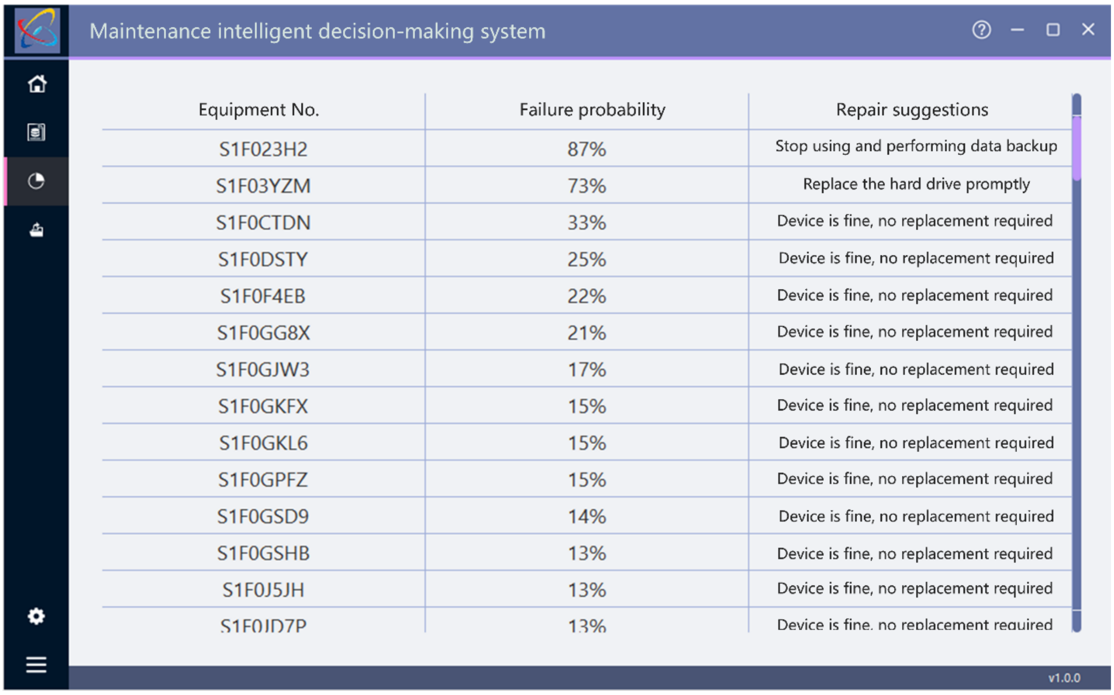


Fig. 10. The operation interface of the fault prediction module

The failure rate in the Failure Prediction module is calculated as follows:

$$F_{\mathrm{r}} = \frac{\left| y_{\mathrm{pred}} - y_{\mathrm{thres}} \right|}{y_{\mathrm{thres}}} \times 100\% \tag{10}$$

Here, y_{pred} represents the predicted remaining useful life (RUL) of the hard disk, while y_{thres} denotes the predefined threshold. The fault prediction module operates by selecting a disk dataset from the status data display module and applying a unified model for remaining service life prediction. In this system, a threshold of 30 days is predefined. Once the threshold is established, the system evaluates and analyzes the prediction results based on this threshold and subsequently generates corresponding maintenance recommendations.

The fault prediction module retrieves status data from the system and applies the offline-trained model to estimate the remaining useful life (RUL) of the hard disk device. If the predicted RUL falls below the predefined minimum replacement threshold, the system prompts maintenance personnel to take immediate action to prevent further deterioration, which could lead to equipment failure and extended downtime. Simultaneously, the system records the prediction results and corresponding recommendations in a database, providing a data foundation for generating comprehensive maintenance reports. This enables administrators to access real-time predictions regarding a hard drive's lifespan and suggested repair schedule, facilitating proactive maintenance planning to ensure optimal device performance and longevity. It is important to note that predictive models not only generate valuable insights and maintenance recommendations but also serve as a decision-support tool for human intervention. However, the suggested repair schedule may require adjustments based on specific circumstances, such as hard drive utilization, environmental conditions, and workload intensity.

5.6 Summary of System Applications

An intelligent auxiliary maintenance decision-making system, utilizing Transformer-based hard disk remaining useful life (RUL) prediction technology, has been designed and implemented in data center environments. Within the data center, the system's application software serves as a decision-support tool for administrators and maintenance personnel, providing recommendations for equipment repair and maintenance by analyzing equipment condition data and predictive models. To assess its effectiveness in real-world applications, the system software was deployed in a data center for a 180-day experimental period.

To evaluate the effectiveness of the intelligent maintenance decision-making system, the system utilizes hard disk status monitoring data to predict the remaining useful life (RUL) of each hard disk. Based on the prediction results, hard disks are categorized into different groups, with those having an RUL of less than five days classified as faulty. The corresponding hard disk device numbers are recorded for further analysis. Once the system completes data backup and replaces the identified faulty hard disks, the replaced disks undergo an in-depth evaluation. The hard disk analysis software CrystalDiskInfo is employed to assess their lifespan, and further verification is conducted through physical disassembly to inspect their internal condition. This process enables a comparative analysis between the system's maintenance decision results and the actual condition of the hard disks, providing insights into the system's predictive accuracy and reliability.

During the actual deployment of the intelligent auxiliary maintenance decision-making system, hard disk parameter index data from the data center is collected at 30, 60, 90, and 180-day intervals. The system then utilizes this data to predict the remaining useful life (RUL) of the hard disks. The predictive accuracy is evaluated by comparing the predicted results with the actual lifespan of the hard disks, as presented in Table 4.

Table 4. Evaluation of the results of the intelligent decision-making system

Smart auxiliary maintenance decision-making system usage days	Accuracy of hard disk failure prediction /%
30	81
60	89
90	93
180	98

As shown in Table 4, the prediction accuracy of the intelligent auxiliary maintenance decision-making system is relatively low during the initial 30-day period. This is primarily because, at the early stage, the prediction model is trained on a diverse range of SMART datasets. However, as the system operates over a longer period, the model undergoes offline training on real hard disk datasets from the data center, allowing it to be fine-tuned for improved alignment with actual working conditions.

In the practical application of the intelligent auxiliary maintenance decision-making system software, particularly in hard disk maintenance, the system assists administrators and maintenance personnel in accurately assessing the status and health of hard disk equipment, providing appropriate repair recommendations, and predicting the remaining lifespan of hard disks. Implementing an intelligent service decision-making system offers several advantages, including:

① Enhancing Equipment Reliability: The system provides timely maintenance recommendations, assisting administrators and maintenance personnel in detecting and resolving hard disk failures at an early stage, thereby improving the reliability and availability of hard disk equipment.

② Reducing Downtime and Costs: By providing maintenance recommendations and predicting the remaining lifespan of hard disks, the system helps administrators and maintenance personnel better plan maintenance and replacement schedules, thereby minimizing downtime and reducing maintenance costs.

③ Enhancing Decision-Making Efficiency: The system automates the process, enabling administrators and maintenance personnel to make decisions and plan maintenance schedules more quickly, thereby improving decision-making efficiency.

④ Improving Maintenance Quality: The system provides more accurate and reliable maintenance recommendations and predictive results, helping maintenance personnel better plan and execute maintenance tasks, thereby enhancing maintenance quality.

In summary, the intelligent auxiliary maintenance decision-making system can help administrators and maintenance personnel better manage hard disk devices, improve equipment reliability and availability, reduce downtime and costs, and enhance maintenance quality.

6 Conclusion

This paper proposes a complete set of hard disk remaining life (RUL) prediction and intelligent decision-making solutions. Firstly, a reliable feature set is constructed through feature engineering, and a Transformer-based prediction model is designed, and the self-attention mechanism is used to capture the degradation mode of the SMART sequence of the hard disk, so as to achieve unified and accurate prediction across models of hard disks. Experiments show that the proposed model significantly improves the prediction accuracy compared with the traditional method. Finally, the system integrates the prediction model to realize the whole process intelligence from data collection to operation and maintenance decision-making, and the simulation experiment verifies its fault early warning ability.

Future research can be improved in the following directions: at the model level, multi-source information such as workloads and environmental data can be fused, or hybrid architectures (such as Transformer+ graph neural networks) can be explored to improve generalization capabilities; In terms of algorithm optimization, transfer learning is studied to reduce the dependence on fault data. In terms of decision-making system, adaptive thresholds and cost models are introduced to optimize O&M strategies. In addition, it is necessary to verify the deployment effect in the actual production environment and promote the application of intelligent maintenance technology.

References

[1] C. Xu, G. Wang, X. Liu, D.D. Guo, T-Y. Liu, Health status assessment and failure prediction for hard drives with recurrent neural networks, IEEE Transactions on Computers 65(11)(2016) 3502-3508. https://doi.org/10.1109/TC.2016.2538237

[2] Y. Wang, E.W.M. Ma, T.W.S. Chow, K.-L. Tsui, A two-step parametric method for failure prediction in hard disk drives, IEEE Transactions on Industrial Informatics 10(1)(2014) 419-430. https://doi.org/10.1109/TII.2013.2264060

[3] K. Kaur, K. Kaur, Failure prediction, lead time estimation and health degree assessment for hard disk drives using voting based decision trees, Computers, Materials & Continua 60(3)(2019) 913-946. https://doi.org/10.32604/cmc.2019.07675

[4] M. Ma, Z. Mao, Deep-convolution-based LSTM network for remaining useful life prediction, IEEE Transactions on Industrial Informatics 17(3)(2021)1658-1667. https://doi.org/10.1109/TII.2020.2991796

[5] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A.N. Gomez, Ł. Kaiser, I. Polosukhin, Attention is all you need, in: Proc. 31st Conference on Neural Information Processing Systems, 2017.
https://proceedings.neurips.cc/paper_files/paper/2017/file/3f5ee243547dee91fbd053c1c4a845aa-Paper.pdf
[6] M. Pecht, E. Elburn, Commercial hard drive failures in a data center application and the role of SMART attribute information, Circuit World 47(4)(2020) 408-426.
https://doi.org/10.1108/CW-07-2020-0127
[7] R. Sakata, I. Ohama, T. Taniguchi, An extension of gradient boosted decision tree incorporating statistical tests, in: Proc. 2018 IEEE International Conference on Data Mining Workshops (ICDMW), 2018.
https://doi.org/10.1109/ICDMW.2018.00139
[8] S. Ganguly, A. Consul, A. Khan, B. Bussone, J. Richards, A. Miguel, A practical approach to hard disk failure prediction in cloud platforms: big data model for failure management in datacenters, in: Proc. 2016 IEEE Second International Conference on Big Data Computing Service and Applications (BigDataService), 2016.
https://doi.org/10.1109/BigDataService.2016.10
[9] J. Ren, M. Zhang, C. Yu, Z.W. Liu, Balanced MSE for imbalanced visual regression, in: Proc. of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2022.
https://doi.org/10.1109/CVPR52688.2022.00777
[10] C.J. Wang, Y.Y. Chen, S.Q. Zhang, Q.H. Zhang, Stock market index prediction using deep Transformer model, Expert Systems with Applications 208(2022) 118128.
https://doi.org/10.1016/j.eswa.2022.118128
[11] J.W. Bai, L. Yuan, S.-T. Xia, S.H. Yan, Z.F. Li, W. Liu, Improving vision transformers by revisiting high-frequency components, in: Proc. European Conference on Computer Vision (ECCV), 2022.
https://doi.org/10.1007/978-3-031-20053-3_1
[12] Z. Liu, Y.T. Lin, Y. Cao, H. Hu, Y.X. Wei, Z. Zhang, S. Lin, B.N. Guo, Swin Transformer: Hierarchical Vision Transformer using Shifted Windows, in: Proc. of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2021.
https://doi.org/10.1109/ICCV48922.2021.00986
[13] S.W. Zamir, A. Arora, S. Khan, M. Hayat, F.S. Khan, M. Yang, Restormer: Efficient Transformer for High-Resolution Image Restoration, in: Proc. 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2022.
https://doi.org/10.1109/CVPR52688.2022.00564
[14] Q.H. Wang, X.H. Wang, F.H. Wang, H.X. Liu, J.P. Gou, W.B. Chen, Remaining lifespan prediction on multiple types of hard disks under conditions of data imbalance, Computers and Electrical Engineering 116(2024) 109195.
https://doi.org/10.1016/j.compeleceng.2024.109195
[15] D.Y. Zhang, J. Shao, X.Y. Li, H.T. Shen, Remote sensing image super-resolution via mixed high-order attention network, IEEE Transactions on Geoscience and Remote Sensing 59(6)(2021) 5183-5196.
https://doi.org/10.1109/TGRS.2020.3009918
[16] S. Pang, Y.H. Jia, R. Stones, G. Wang, X.G. Liu, A combined Bayesian network method for predicting drive failure times from SMART attributes, in: Proc. 2016 International Joint Conference on Neural Networks (IJCNN), 2016.
https://doi.org/10.1109/IJCNN.2016.7727837
[17] N. Wang, X.L. Zhao, Time series forecasting based on convolution transformer, IEICE Transactions on Information and Systems E106.D(5)(2023) 976-985.
https://doi.org/10.1587/transinf.2022EDP7136
[18] F.H. Huang, M. Jiang, F. Liu, D. Xu, Z.M. Fan, Y.H. Wang, Classification of heads in multi-head attention mechanisms, in: Proc. Knowledge Science, Engineering and Management (KSEM), 2022.
https://doi.org/10.1007/978-3-031-10989-8_54
[19] Q. Li, H. Li, K. Zhang, Prediction of HDD failures by ensemble learning, in: Proc. 2019 IEEE 10th International Conference on Software Engineering and Service Science (ICSESS), 2019.
https://doi.org/10.1109/ICSESS47205.2019.9040739
[20] J. Chen, P.F. Zhang, J.H. Chen, T.S.C. Lau, Heterogeneity-aware clustering and intra-cluster uniform data sampling for federated learning, IEEE Transactions on Emerging Topics in Computational Intelligence 9(3)(2025) 2545-2556.
https://doi.org/10.1109/TETCI.2024.3515007
[21] Rajashekarappa, K.M.S. Soyjaudah, Self monitoring analysis and reporting technology (SMART) Copyback, in: Proc. Computer Networks and Intelligent Computing, 2011.
https://doi.org/10.1007/978-3-642-22786-8_58
[22] L.J. Wu, D.Y. Wu, J.Y. Chen, W.X. Li, An agent formal model for autonomous decision-making, in: Proc. 2012 Fourth International Conference on Multimedia Information Networking and Security, 2012.
https://doi.org/10.1109/MINES.2012.59