Research Article

Dong Liu*, Lijun Kong, Jinghui Song, and Yiming Zhou

Predictive models for overall health of hydroelectric equipment based on multi-measurement point output

https://doi.org/10.1515/jisys-2024-0364
received July 30, 2024; accepted December 05, 2024

Abstract: With the increased operating time and usage frequency of hydroelectric equipment, monitoring and predicting its health state has become increasingly important. Traditional health assessment methods often rely on single measurement point data, which has problems such as insufficient model precision and poor real-time performance. These methods are difficult to reflect the overall health state of the equipment fully and cannot accurately capture the complex dynamic characteristics of the equipment, lacking real-time monitoring capabilities. This article proposed a comprehensive health prediction model for hydroelectric equipment based on the multimeasurement point output, which realized real-time monitoring and prediction of the health state of hydroelectric equipment, optimized maintenance strategies, and reduced maintenance difficulty. By collecting temperature signal data, vibration signal data, pressure signal data, and lubrication degree signal data from three systems including upper guide bearing, thrust guide bearing system, and water guide bearing in hydroelectric equipment, they were used as original sensor signal data. Data preprocessing is performed on original sensor signal data, including handling missing and outliers, stabilizing nonstationary time series data, and filtering temporal noise to address the impact of diverse types and large differences in data scales, achieving more accurate predictions. This article used sequential Bayesian method (SBM), hypersphere algorithm, and long short-term memory (LSTM) network to construct a prediction model. These algorithms had different advantages and applicable scenarios and could complement each other to improve the precision and robustness of prediction models. To further improve prediction precision, the model parameters were optimized through cross validation to avoid overfitting and improve model performance. By comparing and analyzing the predictive performance, error results, and real-time prediction performance before and after model optimization, it was concluded that the prediction model constructed by SBM, hypersphere algorithm, and LSTM network had an overall average improvement of 23.7% in the prediction precision of 12 parameters, including temperature, vibration frequency, pressure, and lubrication degree, for the optimized upper guide bearing, thrust guide bearing, and water guide bearing systems. The overall average error has decreased by 55.6%, and the real-time prediction performance has improved by 10.4%. The optimized model not only improves accuracy but also significantly enhances real-time performance, which can better adapt to real-time prediction needs in dynamic environments.

Keywords: hydroelectric equipment, health prediction, sequential Bayesian method, hypersphere algorithm, long short-term memory network

*** Corresponding author: Dong Liu**, China Yangtze Power Co., Ltd., Wuhan, 430000, Hubei, China, e-mail: LewistWHU@163.com
Lijun Kong: China Yangtze Power Co., Ltd., Wuhan, 430000, Hubei, China, e-mail: kong_lijun@ctg.com.cn
Jinghui Song: China Yangtze Power Co., Ltd., Wuhan, 430000, Hubei, China, e-mail: song_jinghui@ctg.com.cn
Yiming Zhou: China Yangtze Power Co., Ltd., Wuhan, 430000, Hubei, China, e-mail: zhou_yiming@ctg.com.cn

1 Introduction

With the continuous growth of modern industry and energy demand, the importance of hydroelectric equipment in power production is becoming increasingly prominent. The stable operation of hydroelectric equipment plays a crucial role in the reliability and economy of power supply. However, hydroelectric equipment [1,2] often faces various types of faults and wear issues due to its complex structure and working environment. The health state of the hydroelectric equipment is influenced by various factors [3–5]. Traditional methods for predicting the overall health of hydroelectric equipment often suffer from inability to respond promptly to changes in the equipment state [6]. This significantly increases maintenance costs. Therefore, achieving high-precision overall health prediction of hydroelectric equipment [7,8] has important research significance and application value.

In previous studies, health assessment models mainly relied on data from a single measurement point [9–11]. The data from a single measurement point can easily lead to one-sided and inaccurate evaluation results [12,13]. Traditional methods often fail to fully consider the complex dynamic characteristics of devices and the interactive effects of multidimensional data during the model construction process [14–16]. Traditional health assessment methods have limited capabilities in data preprocessing [17,18], making it difficult to effectively address these data quality issues. This article proposes an overall health prediction model for hydroelectric equipment based on multiple measurement point outputs [19,20].

This article focuses on the original sensor signal data of hydroelectric equipment [21–23] and conducts comprehensive data preprocessing, including handling missing values and outliers, stabilizing nonstationary time series data, and filtering temporal noise [24–26], to address the impact of diverse types and large differences in data scales, achieving more accurate predictions. The sequential Bayesian method (SBM), hypersphere algorithm, and long short-term memory (LSTM) network are selected to construct a health state prediction model [27–29]. The SBM captures the temporal variation patterns of device states by modeling the sequence of historical data; the hypersphere algorithm utilizes high-dimensional data clustering technology to identify the normal and abnormal states of devices; the LSTM predicts the future health state of devices by processing long-time series data. During the model training process, the cross validation [30] method is used to optimize the model parameters, avoiding overfitting. By combining real-time monitoring data, the entropy weight method [31] is used to rank the fault risk of each measurement point.

The major contributions of this article are as follows:

(1) A prediction model is constructed using the SBM, hypersphere algorithm, and LSTM network. The SBM captures the temporal variation patterns of device states by modeling the sequence of historical data. The hypersphere algorithm utilizes high-dimensional data clustering technology to identify the normal and abnormal states of devices. The LSTM predicts the future health state of devices by processing long time series data.
(2) During the model training process, cross-validation is used to optimize the model parameters, avoiding overfitting and enhancing the model's robustness.
(3) A series of data preprocessing techniques are implemented to handle missing values, detect and handle abnormal values, stabilize nonstationary time series data, and filter temporal noise. These steps address the impact of diverse types and large differences in data scales, further improving the accuracy of predictions.

2 Related works

Traditional health assessment methods mainly rely on data from a single measurement point, and these methods have significant shortcomings in model precision and real-time performance, leading to unsatisfactory model prediction precision. To improve the limitations of single measurement point data, researchers have proposed a multimeasurement point data fusion method, which improves the accuracy of health prediction by collecting and analyzing data from multiple sensors.

There are currently many research achievements in the construction of prediction models. Liu et al. [32] performed comprehensive diagnosis and decision-making for high-speed train bearings by taking both spatial and temporal dimensions. Zhang et al. [33] established a health benchmark model for pumped storage units. The study explored the complex interaction characteristics between multiple influencing factors. The aforementioned research methods provided a fundamental research approach for the study of this article.

There are also many research achievements in the field of health prediction of hydroelectric equipment. Jiang et al. [34] proposed a method for anomaly prediction of hydraulic turbines by combining neural networks with LSTM models. The study used the historical values of pressure parameters obtained from actual engineering applications to determine the operating conditions of the steam turbine. The state trend of water turbine units under different operating conditions was predicted, and the abnormal states at different collection points under different operating conditions were warned based on the correlation between measurement points. Cheng et al. [35] adopted the Bayesian network (BN) as the technical framework, applied expected utility theory, and innovated maintenance decision models. The aforementioned research provided specific research methods for the study of this article.

This article aims to construct a high-precision model for predicting the health state of hydroelectric equipment, optimizing maintenance strategies, and reducing maintenance difficulty. Finally, by comparing and analyzing the predictive performance, error results, and real-time prediction performance before and after model optimization, the superiority of the proposed model was verified. This study not only enriched the research system of health assessment methods for hydroelectric equipment in theory but also had important guiding significance in practical applications. By using a health prediction model based on multiple measurement point outputs, it can improve the reliability and economic benefits of equipment operation.

3 Operation and maintenance data processing of hydroelectric units

3.1 Data collection

This article aims to implement a comprehensive health prediction model for hydroelectric equipment based on the multimeasurement point output. The research equipment includes the upper guide bearing system, thrust guide bearing system, and water guide bearing system of hydroelectric generator units. The upper guide bearing is located at the top of the rotor of the hydroelectric generator set, mainly used to support and guide the upper end of the rotor. Thrust guide bearings are located in the middle or near the tail of the rotating shaft, used to guide and support the rotational motion of the rotating shaft, as well as reduce axial and radial vibrations. The water guide bearing is located at the upper or lower part of the hydroelectric generator unit, and sometimes may also be located on the side of the unit. Its main function is to support and guide the rotational movement of the turbine shaft to reduce friction and ensure the smooth operation of the unit. The schematic diagram of the structure of the hydroelectric generator set is shown in Figure 1.

To achieve health prediction of the systems of a hydroelectric generator set, this article starts from four aspects: temperature, vibration frequency, pressure, and lubrication degree, and uses temperature sensors, vibration sensors, pressure sensors, and oil quality sensors to measure and collect data. By collecting data from various aspects of the systems during normal operation of the hydroelectric generator set, a sample of normal and healthy original sensor signal data is obtained. The timing diagram of the original data is shown in Figure 2. Figure 2 shows the temperature, vibration frequency, pressure, and lubrication degree of the systems collected during 10 h of normal operation of the hydroelectric generator set.

It can be seen that the parameter levels of the upper guide bearing system and the thrust guide bearing system are similar because the working environment of them are similar. The parameter level of the water guide bearing system is significantly different from that of the upper and thrust guide bearing systems because

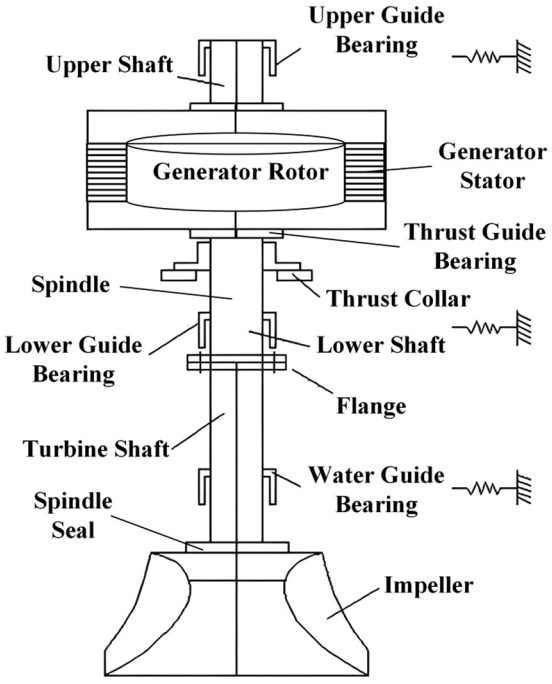

Figure 1: Structural schematic diagram of hydroelectric generator set. Source: Created by the authors.

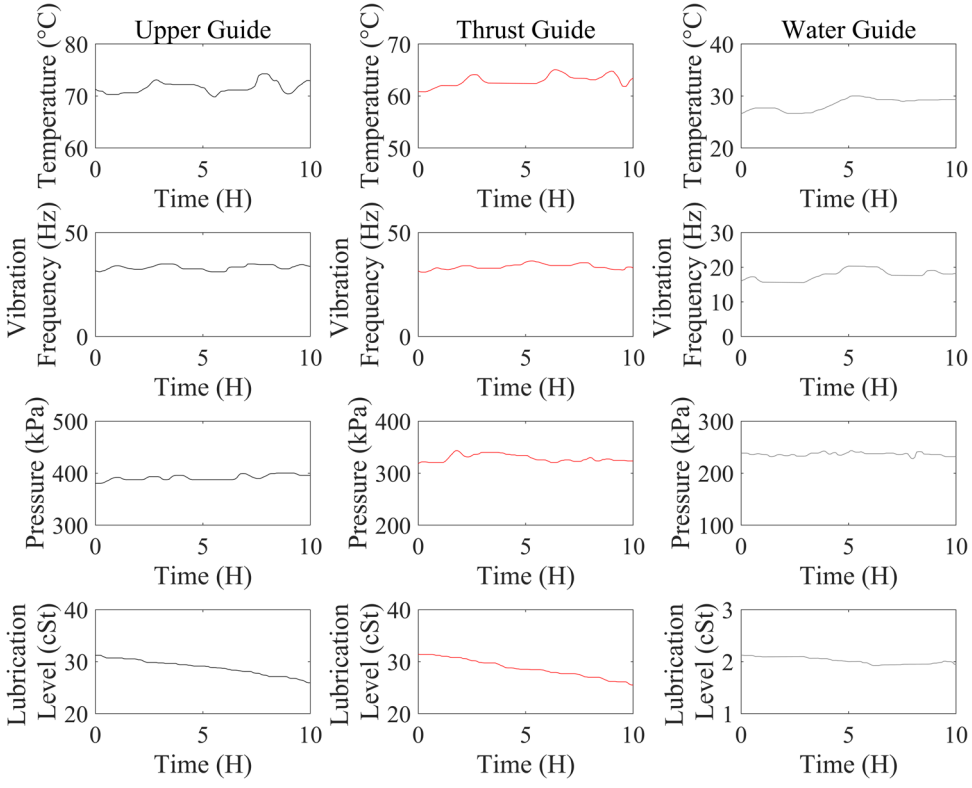

Figure 2: Timing diagram of original sensor signal data. Source: Created by the authors.

the water guide bearing system works in water and the working environment is very different from that of the upper and thrust guide bearing systems.

3.2 Data preprocessing

The original sensor signal data of systems collected in the aforementioned steps may have many missing data in the time dimension at some measurement points, as well as a large number of dead points and data drift due to sensor anomalies and other reasons. At the same time, its diverse types and large differences in data scales also have a certain degree of impact on subsequent research. To solve the aforementioned problems, it is necessary to preprocess the collected original sensor signal data, including missing value processing, outlier detection and processing, nonstationary sequence stabilization, and temporal noise filtering.

3.2.1 Missing value handling

The K-nearest neighbor (KNN) imputation algorithm is used to address missing values. The KNN algorithm calculates the Euclidean distance between each missing value sample in the dataset and other samples, finds the nearest neighbors to the sample, and fills in the missing values by weighted averaging the feature values of these neighbors. By utilizing the local similarity of the data, the rationality and accuracy of the filling results are ensured, and the impact of missing values on model training is solved.

First, for each missing value sample, its Euclidean distance from other samples is calculated as shown in formula (1):

$$d(x_i, x_j) = \sqrt{\sum_{k=1}^{n} (x_{ik} - x_{jk})^2}. \tag{1}$$

Among them, x_i and x_j are the feature vectors of samples i and j, respectively, and n is the number of features.

The nearest k samples are selected, and their feature values are recorded. The weighted average of these sample feature values is calculated as the filling value for missing values, as shown in formula (2):

$$\hat{x}_i = \frac{1}{K} \sum_{j \in N_i} x_j. \tag{2}$$

Among them, N_i is the set of k neighbors of sample i.

3.2.2 Abnormal value detection and handling

Autoencoder is used for outlier detection. Autoencoder compresses input data into a low-dimensional representation by training a neural network and then reconstructs the data through a decoder. Autoencoder can effectively identify outliers and solve the problem of outliers in sensor data by learning the internal structure of the data.

First, the autoencoder model is constructed. The encoder maps the input data x to a low-dimensional representation, and the decoder reconstructs it as $\hat{x}$, as shown in formula (3):

$$z = f(x), \quad \hat{x} = g(z). \tag{3}$$

Among them, f and g represent the encoder and decoder, respectively.

The autoencoder is trained to minimize the reconstruction error between the input data x and the reconstructed data $\hat{x}$, as shown in formula (4):

$$L = \frac{1}{n} \sum_{i=1}^{n} \|x_i - \hat{x}_i\|^2. \tag{4}$$

All sample data are reconstructed, and the reconstruction error of each sample is calculated as shown in formula (5):

$$e_i = \|x_i - \hat{x}_i\|. \tag{5}$$

The reconstruction error threshold ϵ is set, and samples exceeding the threshold are marked as outliers, as shown in formula (6):

$$e_i > \epsilon \Rightarrow x_i \text{ is an anomaly.} \tag{6}$$

Assuming that the marked abnormal data point is x_{anom}, its corresponding normal data point is x_{norm}. It is necessary to find an alternative data point $\hat{x}$ that is as close as possible. To find the optimal alternative data point $\hat{x}$, a modified loss function $L_{\text{correction}}$ is defined using mean square error (MSE), as shown in formula (7):

$$L_{\text{correction}} = \|x_{\text{anom}} - \hat{x}\|^2. \tag{7}$$

By using the gradient descent method, the aforementioned loss function is optimized to find a substitute data point $\hat{x}$ that minimizes the loss. The optimization process is shown in formula (8):

$$\hat{x}^{(t+1)} = \hat{x}^{(t)} - \eta \nabla L_{\text{correction}}. \tag{8}$$

Among them, $\hat{x}^{(t)}$ is the substitute data point for the tth iteration, η is the learning rate, and $\nabla L_{\text{correction}}$ is the gradient of the loss function.

The optimized alternative data point $\hat{x}$ is replaced with the original outlier data point x_{anom} to obtain the corrected dataset $X_{\text{corrected}}$. The replacement process is shown in formula (9):

$$X_{\text{corrected}} = (X \backslash \{x_{\text{anom}}\}) \cup \{\hat{x}\}. \tag{9}$$

3.2.3 Stationarization of nonstationary sequences

The first-order difference method is used to stabilize nonstationary time series data. The first-order difference eliminates trend and periodic components in the time series by calculating the difference between adjacent time point data, making the time series data more stationary. First, for each original time series data, the data difference between adjacent time points is calculated, as shown in formula (10):

$$y_t' = y_t - y_{t-1}. \tag{10}$$

Among them, y_t' is the differential data and y_t is the original data.

After performing differential processing on temporal data, stationarity testing is required to ensure the stationarity of the data. This article uses augmented Dickey–Fuller (ADF) test to test the stationarity of the data. The principle of ADF test is to test that the data are nonstationary of the existence in time series data. The ADF test is conducted through a regression model, as shown in formula (11):

$$\Delta y_t' = \alpha + \beta t + \gamma y_{t-1}' + \sum_{i=1}^{p} \delta_i \Delta y_{t-i}' + \epsilon_t. \tag{11}$$

Among them, $\Delta y_t'$ is the difference of y_t'; α is a constant term; βt is a time trend term; γ is the coefficient to be estimated, used to test the hypothesis of $\gamma = 0$; δ_i is the coefficient of the autoregressive term; and ϵ_t is the error term.

The appropriate number of lagged terms p is selected to ensure the white noise characteristics of the error term ϵ_t. The regression analysis is performed on the differential data based on the selected number of lagged terms p, and the coefficients of each item in the regression formula are estimated. Afterward, the ADF statistic is calculated as shown in formula (12):

$$\text{ADF} = \frac{\hat{\gamma}}{\text{SE}(\hat{\gamma})}. \tag{12}$$

Among them, $\hat{\gamma}$ is the estimated coefficient of y_{t-1}' in the regression model and $\text{SE}(\hat{\gamma})$ is the standard deviation of $\hat{\gamma}$.

The calculated ADF statistic is compared with the critical value, or its corresponding p value is directly calculated.

3.2.4 Time series noise filtering

The moving average algorithm is used to smooth time series data, filter out high-frequency noise, and achieve smoothing of time series data. The moving average algorithm reduces data volatility by calculating the average value of data within a certain window. First, the size of the moving window is set to n, and for each time point, the average value of the data from the first n time points is calculated, as shown in formula (13):

$$\bar{x}_t = \frac{1}{n} \sum_{i=t-n+1}^{t} x_i. \tag{13}$$

The calculated average value is used to replace the original dataset to obtain the time series noise-filtered time series graph. After filtering, the time series data become smoother, reducing the interference of high-frequency noise on the prediction model and achieving more accurate health prediction. The processed data sequence diagram is shown in Figure 3.

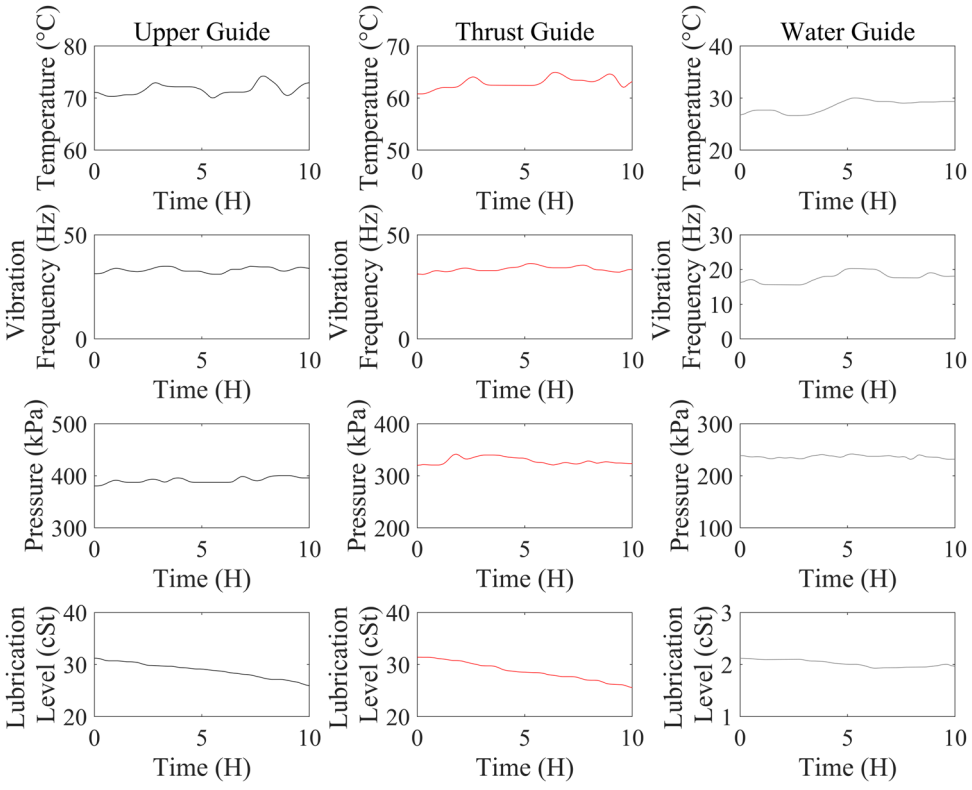

Figure 3: Time sequence diagram after filtering temporal noise. Source: Created by the authors.

The window size of the moving average method has a significant effect on the data smoothing effect. A larger window smoothes the curve but may lose some detail information, while a smaller window retains more detail but has less smoothing effect. The moving average method with a window length of 5 is chosen in this study, as shown in Figure 3. Compared to Figure 2, the time series curve appears smoother while retaining specific details. This choice ensures that the data smoothing process does not overly blur the details and avoids any impact on the subsequent research analysis.

This section significantly enhances the robustness of the model to outliers and noise by implementing a series of data preprocessing techniques. Through regularization and network structure optimization, the improved autoencoder improves the ability of the model to recognize and deal with outliers. The window

size of the moving average algorithm is optimized to effectively filter out the high-frequency noise in the time series data, while retaining the key features of the data.

After completing the basic data preprocessing step, other types of sensor data are expanded further. The sound signal can reveal the abnormal vibration pattern of the device, and the electrical signal data can reflect the health of the electrical system. By integrating multimodal data, the model can capture the complex dynamic behavior of the device from different angles, thus enhancing its generalization ability under different operating conditions.

4 Construction of device overall health prediction model based on multimeasurement point output

4.1 Construction of predictive model for device health state

This article uses SBM, hypersphere algorithm, and LSTM network to construct a prediction model. These algorithms have different advantages and applicable scenarios.

The temperature, vibration frequency, pressure, and lubrication degree data are segmented according to the fixed windows. The three states of a device are defined: normal, minor fault, and major fault. Markov chains are used to describe state transitions, assuming that the transition probability matrix is P, as shown in formula (14):

$$P(X_{t+1}|X_t) = \begin{bmatrix} p_{11} & p_{12} & \cdots & p_{1j} \\ p_{21} & p_{22} & \cdots & p_{2j} \\ \vdots & \vdots & \ddots & \vdots \\ p_{i1} & p_{i2} & \cdots & p_{ij} \end{bmatrix}. \tag{14}$$

Among them, $p_{ij} = P(X_{t+1} = j|X_t = i)$ represents the probability of state i transitioning to state j.

The Bayesian formula is used to update the health state, as shown in formula (15):

$$P(X_t|Y_t) = \frac{P(Y_t|X_t)\cdot P(X_t)}{P(Y_t)}. \tag{15}$$

Among them, Y_t represents observation data, $P(X_t|Y_t)$ is a posterior probability, $P(Y_t|X_t)$ is the likelihood function, $P(X_t)$ is a prior probability, and $P(Y_t)$ is the standardization factor.

The posterior probability of each health state is calculated, and the state corresponding to the maximum posterior probability is selected as the prediction result $\hat{X}_t$, as shown in formula (16):

$$\hat{X}_t = \arg\max_{X_t} P(X_t|Y_t). \tag{16}$$

The normal operation dataset of the device is input into the support vector machine to generate hypersphere boundaries, as shown in formula (17):

$$S = \{x \in R^n | \|x - c\|^2 \le R^2\}. \tag{17}$$

Among them, c is the center of the hypersphere and R is the radius of the hypersphere.

Whether the new observation data x_{new} is located within the hypersphere is determined as shown in formula (18):

$$\text{if} \quad \|x_{\text{new}} - c\|^2 \le R^2, \quad \text{then} \quad x_{\text{new}} \in \text{Normal}. \tag{18}$$

The LSTM network structure is designed, including an input layer, multiple LSTM layers, and an output layer. The input sequence length is set to T, and the feature dimension is set to n. The state update formula of the LSTM unit is defined, as shown in formula (19):

$$
\begin{aligned}
f_t &= \sigma(W_f \cdot [h_{t-1}, x_t] + b_f),\\
i_t &= \sigma(W_i \cdot [h_{t-1}, x_t] + b_i),\\
\tilde{C}_t &= \tanh(W_C \cdot [h_{t-1}, x_t] + b_C),\\
C_t &= f_t \cdot C_{t-1} + i_t \cdot \tilde{C}_t,\\
o_t &= \sigma(W_o \cdot [h_{t-1}, x_t] + b_o),\\
h_t &= o_t \cdot \tanh(C_t).
\end{aligned} \tag{19}
$$

Among them, f_t, i_t, and o_t are the activation functions of the forget gate, input gate, and output gate, respectively; C_t is the cellular state; and h_t is in a hidden state.

The backpropagation algorithm is adopted to train the LSTM network, and the minimum loss function is shown in formula (20):

$$L = \frac{1}{N}\sum_{i=1}^{N}(\hat{y}_i - y_i)^2. \tag{20}$$

Among them, $\hat{y}_i$ is the predicted value and y_i is the true value.

Through the aforementioned methods, this article constructs a device overall health prediction model based on the multimeasurement point output. The SBM captures the dynamic changes of the device state through time series modeling; the hypersphere algorithm detects abnormal states through high-dimensional clustering; LSTM networks predict future health state through temporal modeling. The combination of these methods can fully utilize the multimeasurement point data, improve the prediction precision and robustness of the model, and provide strong support for health monitoring and fault prediction of hydroelectric equipment.

The decision logic of the model is based on features such as temperature, vibration, pressure, and lubricity signals. It uses the SBM, the hypersphere algorithm, and the LSTM network to identify equipment status and predict faults. The SBM captures time-varying patterns, the hypersphere algorithm identifies abnormal states, and the LSTM predicts future health states. The decision process selects the most likely state based on the posterior probability. Cross-validation is used to optimize model parameters, enhance prediction accuracy and robustness, and support health management of hydropower equipment.

Under the multitask learning framework, the model is represented by sharing underlying features, while predicting temperature anomalies, vibration frequency deviations, and pressure instability. Using the deep structure of the LSTM network, the model can learn the interrelationships and dependencies between different fault modes, thus improving the accuracy and efficiency of prediction.

4.2 Model optimization

During the model training process, cross validation is used to optimize the model parameters. Cross validation improves model performance by dividing the dataset into multiple subsets and iteratively training and validating to avoid overfitting.

First, the dataset is divided into k subsets, where k typically takes a value of 5 or 10. Each time, one subset is selected as the validation set, and the remaining $k - 1$ subsets is selected as the training set. The aforementioned steps are repeated k times, and a different validation set is selected each time. The average of k validation results is calculated as the final performance indicator of the model. The model parameters are adjusted based on the cross validation results and the parameter configuration with the best performance is selected, as shown in formula (21):

$$\mathrm{MSE} = \frac{1}{k}\sum_{i=1}^{k}\frac{1}{n_i}\sum_{j=1}^{n_i}(\hat{y}_{ij} - y_{ij})^2. \tag{21}$$

Among them, $\hat{y}_{ij}$ is the predicted value in the ith validation, y_{ij} is the true value, and n_i is the number of samples in the ith validation.

Model parameters should be adjusted for different sizes of equipment to ensure prediction accuracy. By changing the number of layers and neurons in the LSTM network, as well as optimizing the clustering accuracy of the hypersphere algorithm, the model can adapt to the diverse needs of small-scale to large-scale devices. This flexibility ensures that the model can provide accurate health state predictions for hydro power plants of different sizes, enhancing the utility and universality of the model.

In the process of model optimization, the influence of data characteristics of different measurement points on the sensitivity of model parameters is particularly considered. By analyzing the mean, variance, and distribution of each measuring point data, the adaptive parameter adjustment strategy is adopted to optimize the model for different data characteristics.

By integrating online learning mechanisms, the model can continuously learn from the data accumulated during the operation of the equipment, so as to adapt to long-term changing operating states. The model allows the model parameters to be updated in real time without retraining the entire model.

4.3 Assessment and ranking of fault risks

To effectively evaluate the fault risk of hydroelectric equipment and rank the risks, this article adopts the Bayesian inference method, random forest algorithm, and entropy weight method, combined with real-time monitoring data, to construct a risk assessment model.

The time series data of multiple measurement points are standardized to ensure that the data of each measurement point has the same scale. From temporal data, features are extracted, such as mean, standard deviation, maximum, minimum, and frequency domain features. The processed data are input into a random forest model for training, and the training samples are labeled with historical fault data. The contribution of each measurement point feature is evaluated to fault prediction by calculating its importance index. The importance of features is calculated by the contribution of their split nodes in the decision tree, as shown in formula (22):

$$\text{Importance}(X_i) = \sum_{t=1}^{T} \sum_{j \in \text{nodes}(t)} I(s_j = X_i) \cdot \Delta R_j. \tag{22}$$

Among them, $I(s_j = X_i)$ represents node j splitting using feature X_i; ΔR_j is the information gain of node j; and T is the number of decision numbers.

The entropy weight method can allocate weights based on the entropy value of measurement point information, reflecting the relative importance of each measurement point. First, the entropy value of the fault probability distribution for each measurement point is calculated, as shown in formula (23):

$$E_i = -\frac{1}{\ln(n)} \sum_{j=1}^{n} p_{ij} \ln(p_{ij}). \tag{23}$$

Among them, p_{ij} represents the probability of measurement point i in state j and n represents the total number of states.

The weight w_i is calculated based on the entropy value of each measurement point using formula (24), and then the fault probability of each measurement point is multiplied with its weight using formula (25) to calculate the comprehensive fault risk. Formulas (24) and (25) are as follows.

$$w_i = \frac{1/E_i}{\sum_{i=1}^{n}(1/E_i)}, \tag{24}$$

$$R_{\text{total}} = \sum_{i=1}^{m} w_i \cdot P(F_i|D_i). \tag{25}$$

According to the comprehensive fault risk R_{total}, each measurement point is sorted, and the highest risk measurement point is prioritized for processing. This achieves the risk assessment and ranking of faults in hydroelectric equipment.

5 Results

5.1 Model performance evaluation

After completing the construction and optimization of the model, the models before and after optimization are subjected to performance testing and evaluation. This article evaluates the performance of the model by calculating its prediction precision and recall, and drawing the Precision-Recall (PR) curves of the model before and after optimization.

First, the calculation of precision and recall is carried out, as shown in formulas (26) and (27):

$$\text{Precision} = \frac{\text{TP}}{\text{TP} + \text{FP}}, \tag{26}$$

$$\text{Recall} = \frac{\text{TP}}{\text{TP} + \text{FN}}. \tag{27}$$

Among them, TP is the number of correctly predicted positive samples; FP is the number of negative samples incorrectly predicted as positive; and FN is the number of positive samples incorrectly predicted as negative.

In addition, this article also evaluates the performance of the model by calculating the area under curve (AUC) index of different system parameters before and after model optimization. By comparing the AUC values, it can be determined whether the model performs well in predicting positive and negative class samples. An AUC value greater than 0.9 usually indicates that the model has strong discriminative ability; 0.7–0.9 indicates good model performance; 0.5–0.7 indicates average model performance; and less than 0.5 indicates poor model performance.

After obtaining the data of precision and recall, the PR curve of the model is drawn, including the PR curves of temperature, vibration, pressure, and lubrication degree of the upper guide bearing, thrust guide bearing, and water guide bearing systems, and the AUC value of each curve is calculated, as shown in Figures 4 and 5.

From Figure 4, it can be seen that before optimization, the AUC values of 12 parameters, including temperature, vibration frequency, pressure, and lubrication degree, for the three systems are in the range of 0.7–0.9 for 9 and 0.5–0.7 for 3. The overall performance of the model before optimization is generally good.

From Figure 5, it can be seen that the AUC values of 12 parameters, including temperature, vibration frequency, pressure, and lubrication degree, for the three systems are greater than 0.9 in 6 of them, and within the range of 0.7–0.9 in 6 of them. The optimized model performance shows an overall good to excellent level.

By comparing the AUC indicators before and after model optimization, it can be seen that the performance of the optimized model has significantly improved in all indicators. Overall, the optimized model shows an average improvement of 23.7% in the prediction precision of 12 parameters, including temperature, vibration frequency, pressure, and lubrication degree, for the upper guide bearing, thrust guide bearing, and water guide bearing systems. It can be seen that the optimized model is clearly more suitable for practical applications and can better support the implementation of maintenance and preventive measures.

The energy consumption of the model is analyzed. It is found that the computational complexity of the model can be reduced effectively by optimizing the structure of the algorithm and adjusting the parameters, thus reducing the energy consumption. Especially in the model optimization stage, the use of lightweight deep learning models and efficient data processing algorithms significantly reduces the energy consumption in the process of model training and prediction.

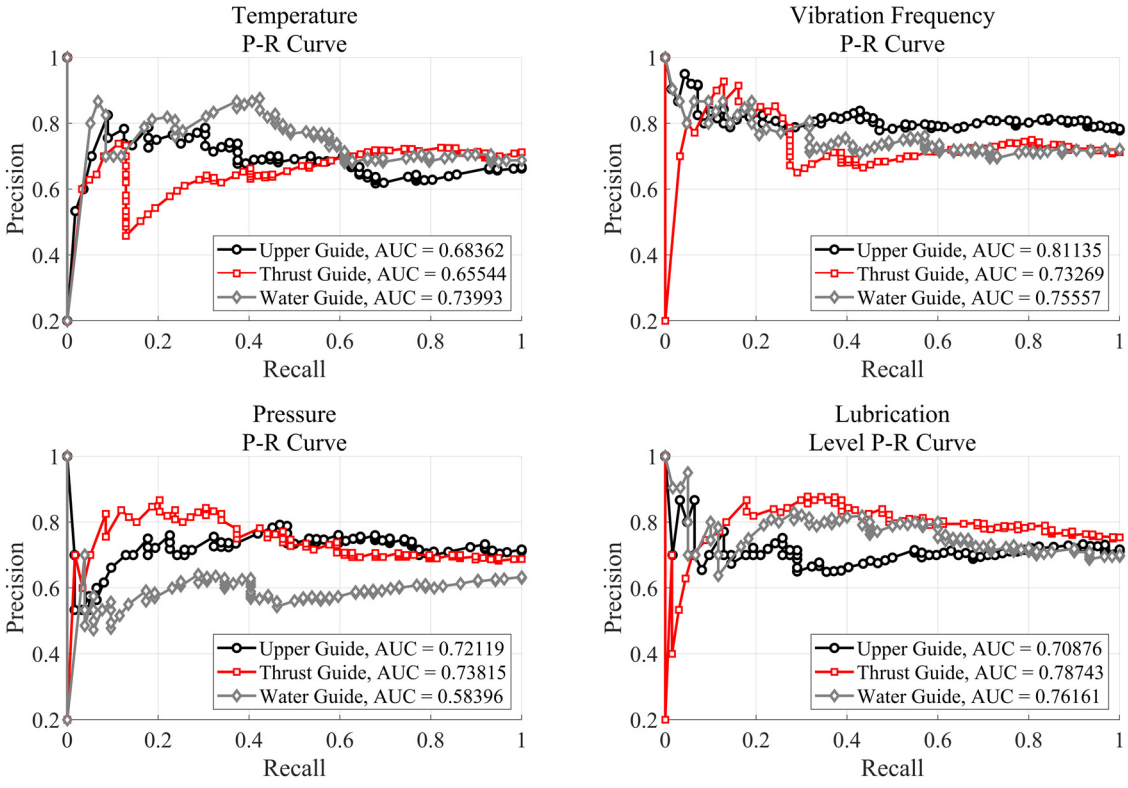

Figure 4: PR curve of the model before optimization. Source: Created by the authors.

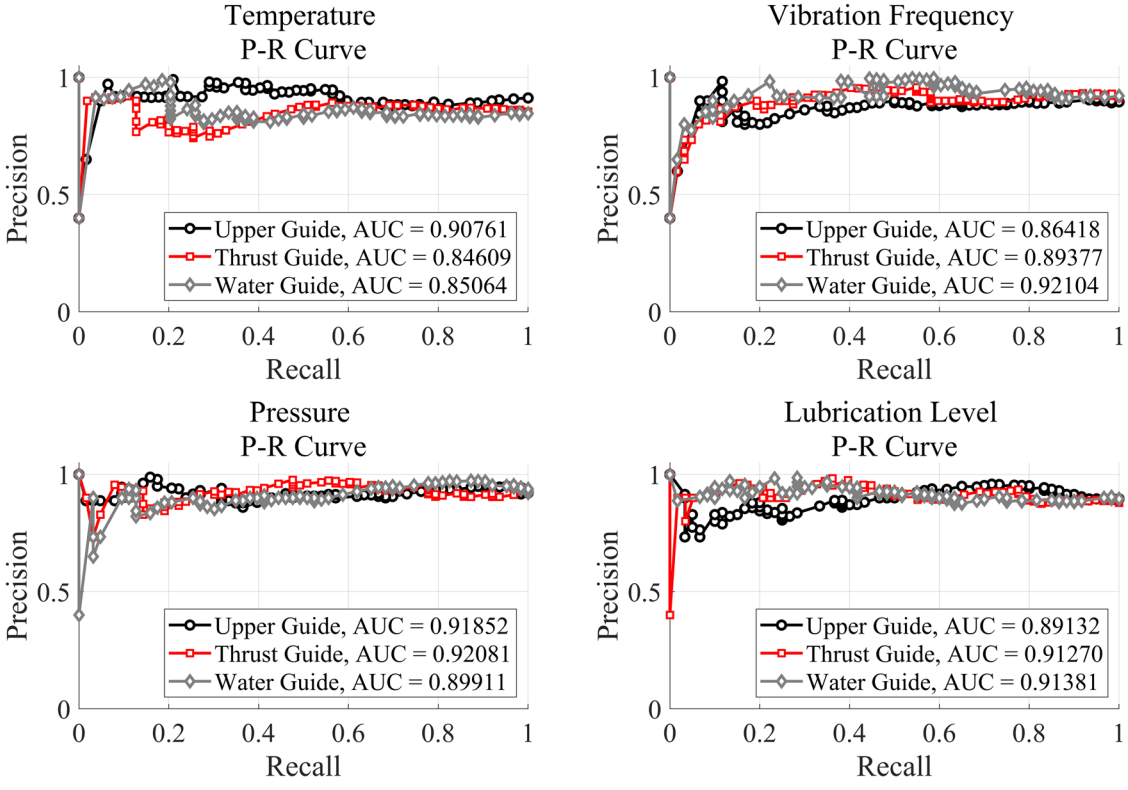

Figure 5: PR curve of the model after optimization. Source: Created by the authors.

5.2 Prediction error

It is crucial to precisely quantify errors to comprehensively evaluate model performance and analyze the performance improvement before and after model optimization. This article uses MSE and mean absolute error as the main indicators. The calculation is shown in formulas (28) and (29):

$$\mathrm{MSE} = \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2, \tag{28}$$

$$\mathrm{MAE} = \frac{1}{n}\sum_{i=1}^{n}|y_i - \hat{y}_i|. \tag{29}$$

Among them, y_i is the actual value; $\hat{y}_i$ is the predicted value; and n is the sample size.

This article visualizes and compares the error results before and after model optimization by drawing error box plots. From the error box plots, the distribution and changes of errors before and after optimization can be observed intuitively, as shown in Figure 6.

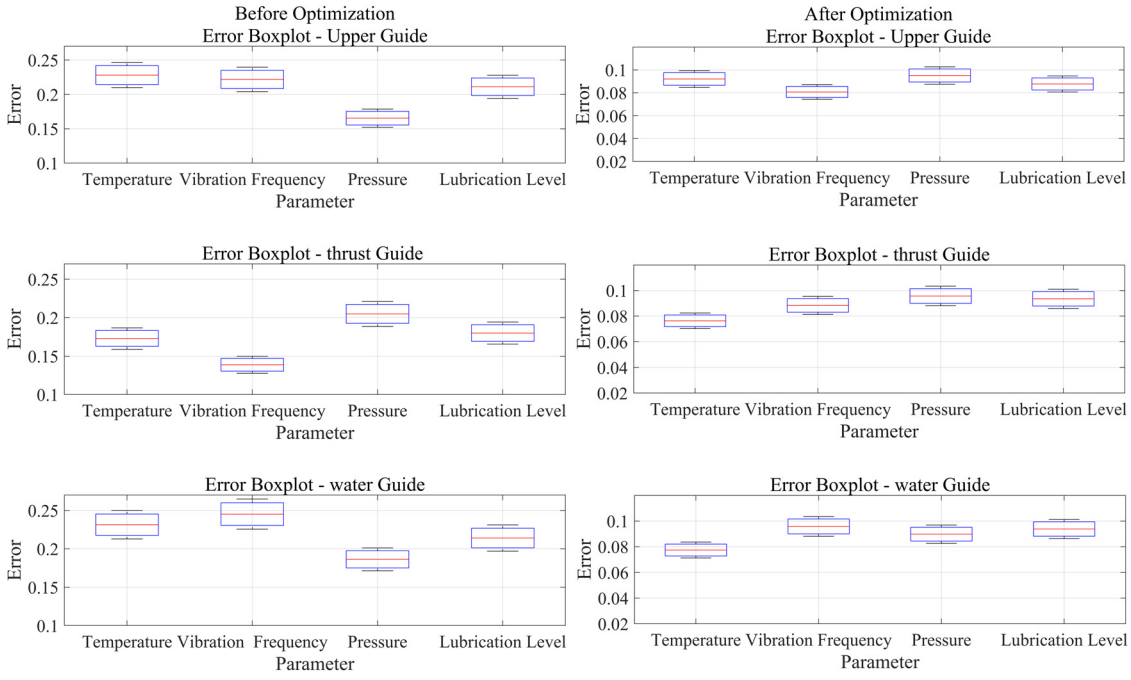

Figure 6: Comparison of error box plots before and after model optimization. Source: Created by the authors.

By analyzing the error data before and after optimization, it can be seen that the model has significant optimization effects on various measurement points and dimensions. The optimized model not only shows a significant decrease in mean error but also significantly improves the concentration and robustness of error distribution. Overall, the optimized model shows an average reduction of 55.6% in error performance levels for 12 parameters including temperature, vibration frequency, pressure, and lubrication degree in the upper guide bearing, thrust guide bearing, and water guide bearing systems. These results indicate that the accuracy and stability of the model in predicting the overall health of hydroelectric equipment have been significantly improved, providing a more reliable basis for real-time monitoring and maintenance of equipment.

To assess the environmental adaptability of the model, the model was tested extensively under a variety of geographical and climatic conditions. By collecting data from hydropower stations in different regions, the model's predictive performance was validated in environments with significant differences in temperature,

humidity, and elevation. The results show that the model can adapt to different environmental conditions and maintain high prediction accuracy.

5.3 Real-time prediction

To achieve real-time prediction analysis of the model, this article conducts online data flow testing and simulates data input in actual operating environments. A real-time data collection system is built. The real-time data flow of device sensors is input into the model, and the entire process time from data input to prediction result output is recorded. To ensure the accuracy of the test, multiple measurement points and different types of sensor data are selected for comprehensive testing. In actual operating environments, real-time state data of devices, such as vibration signals, temperature signals, pressure signals, and lubrication degree signals, are collected and continuously input into the model. The predicted average response time is recorded and compared for the analysis. The analysis table of the real-time prediction results of the model before and after optimization is shown in Table 1.

Table 1: Real-time performance analysis of model prediction before and after optimization

System	Parameter	Prediction time of model before optimization (ms)	Prediction time of model after optimization (ms)	Optimization effect (%)
Upper guide bearing	Temperature	196	176	10.20
	Vibration frequency	237	213	10.13
	Pressure	211	189	10.43
	Lubrication level	201	181	9.95
Thrust guide bearing	Temperature	220	198	10.00
	Vibration frequency	228	203	10.96
	Pressure	285	256	10.18
	Lubrication level	261	233	10.73
Water guide bearing	Temperature	253	226	10.67
	Vibration frequency	232	208	10.34
	Pressure	244	219	10.25
	Lubrication level	273	243	10.99
Average		236.75	212.08	10.4

The prediction time of the model before and after optimization has significantly improved under various system parameters. Overall, by using multiple models such as SBM, hypersphere algorithm, and LSTM network to construct the model, and combining cross validation for model optimization, the optimized model achieves an overall average improvement of 10.4% in real-time prediction of 12 parameters including temperature, vibration frequency, pressure, and lubrication degree in the three systems. These results indicate that the optimized model not only improves accuracy but also significantly enhances real-time performance, which can better adapt to real-time prediction needs in dynamic environments.

By using the LSTM structure, the model can effectively adapt to the rapidly changing data flow. The gating mechanism enables the model to maintain stable prediction performance in the face of real-time fluctuations and sudden changes in data. The model's online learning capability allows it to constantly update parameters from new data, thus responding quickly to changes in the data.

5.4 Discussion

In this study, the stability and accuracy of the model over different time periods, including seasonal changes, were analyzed in depth. Through long-term data tracking and periodic performance evaluation, it is found that the model shows good time stability.

The model is feasible for deployment in real industrial environments, providing real-time health monitoring and predictive maintenance to optimize maintenance strategies and reduce downtime. The deployment process may encounter challenges such as data integration, hardware compatibility, and operating environment complexity. To overcome these challenges, custom development and rigorous field testing in close collaboration with industrial partners are required to ensure the stability of the model and match the actual needs of users.

By adopting a maintenance strategy based on model prediction, unnecessary maintenance activities can be significantly reduced, thereby reducing maintenance costs. The model can monitor the status of equipment in real time and predict potential failures, making maintenance work more accurate and timely, reducing the risk of unexpected downtime and related economic losses. Predictive maintenance strategies help extend the service life of equipment and reduce production interruptions caused by failures.

In this study, the sequential Bayes method is used to capture the time change pattern of the device state, the hypersphere algorithm is used to identify the normal and abnormal state of the device through high-dimensional data clustering technology, and the LSTM network is used to process long time series data to predict the future health state of the device. The advantages of SBM in processing historical data series, the high efficiency of hypersphere algorithm in anomaly detection, and the ability of LSTM in handling complex time series prediction problems together constitute a highly complementary and adaptable prediction model framework, which effectively improves the prediction accuracy and robustness of the model.

On the basis of the model performance evaluation, the performance difference of the model in predicting different fault types was further analyzed. Through a detailed analysis of the fault prediction performance of the guide bearing system, the thrust guide bearing system, and the water guide bearing system of the hydropower unit, it was found that the model's performance in identifying severe fault conditions is more prominent because the signal changes in severe fault conditions are more significant and easy to be captured and learned by the model. For the prediction of minor fault conditions, the performance of the model is worse, because the signal changes of minor faults are more subtle, and the accuracy of the model in distinguishing normal conditions from minor fault conditions is affected to a certain extent. The model parameters were optimized by the cross-validation method, and the prediction performance before and after optimization was compared and analyzed. The results show that the model's prediction ability for different fault types has been significantly improved after optimization. However, there is still room for improvement in the prediction accuracy of minor faults. Future work will focus on further optimizing model parameters and enhancing the model's ability to identify minor fault features.

Different time window sizes have a significant impact on model performance. Shorter time windows cannot capture the long-term trend of equipment state changes, and longer time windows contain too much historical noise. Selecting a moderate time window helps the model predict the health status of hydropower equipment more accurately. According to the characteristics of the equipment and the speed of fault development, a time window size that can best reflect the status of the equipment should be selected.

The model adopts a multilayer defense strategy, which has high security for potential networks. The data transmission process of the model is protected by a strong encryption protocol to prevent data from being intercepted during transmission. Second, the model is deployed in a protected network environment, and potential network attacks are identified and blocked through firewalls and intrusion detection systems. The model also conducts security audits regularly to patch security vulnerabilities in a timely manner.

6 Conclusions

This article aims to study the overall health prediction model of hydroelectric equipment based on the multimeasurement point output. The temperature signal data, vibration signal data, pressure signal data, and lubrication degree signal data of the three systems in hydroelectric equipment are collected as raw sensor signal data. Data preprocessing was performed on original sensor signal data, including handling missing and outliers, stabilizing nonstationary time series data, and filtering temporal noise to address the impact of diverse types and large differences in data scales, achieving more accurate predictions. This article used SBM, hypersphere algorithm, and LSTM network to construct a prediction model. These algorithms have different advantages and applicable scenarios. To further improve prediction precision, the model parameters were optimized through cross validation to avoid overfitting and improve model performance. By comparing and analyzing the predictive performance, error results, and real-time prediction performance before and after model optimization, it was concluded that the prediction model constructed by SBM, hypersphere algorithm, and LSTM network had an overall average improvement of 23.7% in the prediction precision of 12 parameters, including temperature, vibration frequency, pressure, and lubrication degree, for the three systems. The overall average error has decreased by 55.6%, and the real-time prediction performance has improved by 10.4%. The optimized model not only improves accuracy but also significantly enhances real-time performance, which can better adapt to real-time prediction needs in dynamic environments.

Funding information: Development and Demonstration Application of Intelligent Early-Warning Algorithms for Hydropower Units by Integrating Recursive Trees and Recurrent Neural Networks (1519020008).

Authors contributions: Dong Liu: data curation and writing – original draft. Lijun Kong: conceptualization. Jinghui Song: methodology. Yiming Zhou: validation.

Conflict of interest: These no potential competing interests in our paper. And all authors have seen the manuscript and approved to submit to your journal. We confirm that the content of the manuscript has not been published or submitted for publication elsewhere.

Data availability statement: The data that support the findings of this study are available from the corresponding author upon reasonable request.

References

[1] Bernardes Jr J, Santos M, Abreu T, Prado Jr L, Miranda D, Julio R, et al. Hydropower operation optimization using machine learning: A systematic review. AI. 2022;3(1):78–99. doi: 10.3390/ai3010006.

[2] Zhang G, Yu R, Dai L, Pan J. Condition monitoring and fault diagnosis of hydropower station units. Acad J Eng Technol Sci. 2019;2(2):89–95. doi: 10.25236/AJETS.020045.

[3] Zhang L, Qiao F, Wang J, Zhai X. Equipment health assessment based on improved incremental support vector data description. IEEE Trans Syst Man Cybern: Syst. 2019;51(5):3205–16. doi: 10.1109/TSMC.2019.2919468.

[4] Lee CY, Dong ZH. Hierarchical equipment health index framework. IEEE Trans Semicond Manuf. 2019;32(3):267–76. doi: 10.1109/TSM.2019.2925362.

[5] Bahreini R, Doshmangir L, Imani A. Influential factors on medical equipment maintenance management: In search of a framework. J Qual Maint Eng. 2019;25(1):128–43. doi: 10.1108/JQME-11-2017-0082.

[6] Collins GS, Moons KGM. Reporting of artificial intelligence prediction models. Lancet. 2019;393(10181):1577–9. doi: 10.1016/S0140-6736(19)30037-6.

[7] Zhang C, Li J, Wang H, Li S. Applications of big data in equipment health status prediction and spare parts replenishment. China Mech Eng. 2019;30(2):183–7. doi: 10.3969/j.issn.1004-132X.2019.02.008.

[8] Zhang W, Yang D, Wang H. Data-driven methods for predictive maintenance of industrial equipment: A survey. IEEE Syst J. 2019;13(3):2213–27. doi: 10.1109/JSYST.2019.2905565.

[9] Huang HY, Kueng R, Preskill J. Predicting many properties of a quantum system from very few measurements. Nat Phys. 2020;16(10):1050–7. doi: 10.1038/s41567-020-0932-7.
[10] Raissi M, Yazdani A, Karniadakis GE. Hidden fluid mechanics: Learning velocity and pressure fields from flow visualizations. Science. 2020;367(6481):1026–30. doi: 10.1126/science.aaw4741.
[11] Pratapa A, Jalihal AP, Law JN, Bharadwaj A, Murali TM. Benchmarking algorithms for gene regulatory network inference from single-cell transcriptomic data. Nat Methods. 2020;17(2):147–54. doi: 10.1038/S41592-019-0690-6.
[12] Turnbull A, Carroll J, McDonald A. Combining SCADA and vibration data into a single anomaly detection model to predict wind turbine component failure. Wind Energy. 2021;24(3):197–211. doi: 10.1002/we.2567.
[13] Fan ZY, Huang Q, Ren Y, Zhu ZY, Xu X. A cointegration approach for cable anomaly warning based on structural health monitoring data: An application to cable-stayed bridges. Adv Struct Eng. 2020;23(13):2789–802. doi: 10.1177/1369433220924793.
[14] Ma X, Si Y, Yuan Z, Qin Y, Wang Y. Multistep dynamic slow feature analysis for industrial process monitoring. IEEE Trans Instrum Meas. 2020;69(12):9535–48. doi: 10.1109/TIM.2020.3004681.
[15] Pingchao YU, Guo C, Lunxu LI. Modal analysis strategy and nonlinear dynamic characteristics of complicated aero-engine dual-rotor system with rub-impact. Chin J Aeronaut. 2022;35(1):184–203. doi: 10.1016/j.cja.2020.10.031.
[16] ur Rehman N, Aftab H. Multivariate variational mode decomposition. IEEE Trans Signal Process. 2019;67(23):6039–52. doi: 10.1109/TSP.2019.2951223.
[17] Maharana K, Mondal S, Nemade B. A review: Data pre-processing and data augmentation techniques. Glob Transit Proc. 2022;3(1):91–9. doi: 10.1016/j.gltp.2022.04.020.
[18] Rahman A. Statistics-based data preprocessing methods and machine learning algorithms for big data analysis. Int J Artif Intell. 2019;17(2):44–65.
[19] Odu GO. Weighting methods for multi-criteria decision making technique. J Appl Sci Environ Manag. 2019;23(8):1449–57. doi: 10.4314/jasem.v23i8.7.
[20] Feng D, Haase-Schutz C, Rosenbaum L, Hertlein H, Glaeser C, Timm F, et al. Deep multi-modal object detection and semantic segmentation for autonomous driving: Datasets, methods, and challenges. IEEE Trans Intell Transp Syst. 2020;22(3):1341–60. doi: 10.1109/TITS.2020.2972974.
[21] Singh G, Sundaram K. Methods to improve wind turbine generator bearing temperature imbalance for onshore wind turbines. Wind Eng. 2022;46(1):150–9. doi: 10.1177/0309524X211015292.
[22] Tu W, Liang J, Yu W, Shi Z, Liu C. Motion stability analysis of cage of rolling bearing under the variable-speed condition. Nonlinear Dyn. 2023;111(12):11045–63. doi: 10.1007/s11071-023-08432-8.
[23] He C, Zhang J, Geng K, Wang S, Luo M, Zhang X, et al. Advances in ultra-precision machining of bearing rolling elements. Int J Adv Manuf Technol. 2022;122(9):3493–524. doi: 10.1007/s00170-022-10086-6.
[24] de Cheveigne A, Nelken I. Filters: when, why, and how (not) to use them. Neuron. 2019;102(2):280–93. doi: 10.1016/j.neuron.2019.02.039.
[25] Chatterjee S, Thakur RS, Yadav RN, Gupta L. Review of noise removal techniques in ECG signals. IET Signal Process. 2020;14(9):569–90. doi: 10.1049/iet-spr.2020.0104.
[26] Pang G, Shen C, Cao L, Hengel AV. Deep learning for anomaly detection: A review. ACM Comput Surv (CSUR). 2021;54(2):1–38. doi: 10.1145/3439950.
[27] Naesseth CA, Lindsten F, Schon TB. Elements of sequential monte carlo. Found Trends Mach Learn. 2019;12(3):307–92. doi: 10.1561/2200000074.
[28] Stoyan Y, Yaskov G, Romanova T, Litvinchev I, Yakovlev S, Cantú JM. Optimized packing multidimensional hyperspheres: a unified approach. Math Biosci Eng. 2020;17(6):6601–30. doi: 10.3934/mbe.2020344.
[29] Van Houdt G, Mosquera C, Napoles G. A review on the long short-term memory model. Artif Intell Rev. 2020;53(8):5929–55. doi: 10.1007/s10462-020-09838-1.
[30] De Rooij M, Weeda W. Cross-validation: A method every psychologist should know. Adv Methods Pract Psychol Sci. 2020;3(2):248–63. doi: 10.1177/2515245919898466.
[31] Mukhametzyanov I. Specific character of objective methods for determining weights of criteria in MCDM problems: Entropy, CRITIC and SD. Decis Mak: Appl Manag Eng. 2021;4(2):76–105. doi: 10.31181/dmame210402076i.
[32] Liu YZ, Zou YS, Wu Y, Zhang HY, Ding GF. A novel abnormal detection method for bearing temperature based on spatiotemporal fusion. Proc Inst Mech Eng, Part F: J Rail Rapid Transit. 2022;236(3):317–33. doi: 10.1177/09544097211022105.
[33] Zhang X, Jiang Y, Wang XB, Li C, Zhang J. Health condition assessment for pumped storage units using multihead self-attentive mechanism and improved radar chart. IEEE Trans Ind Inform. 2022;18(11):8087–97. doi: 10.1109/TII.2022.3165642.
[34] Jiang X, Gao X, Wang Z, Wang L. Working condition analysis and state trend prediction of hydraulic turbine units. Int J Fluid Mach Syst. 2021;14(3):258–69. doi: 10.5293/IJFMS.2021.14.3.258.
[35] Cheng J, Zhu C, Fu W, Wang C, Sun J. An Imitation medical diagnosis method of hydro-turbine generating unit based on Bayesian network. Trans Inst Meas Control. 2019;41(12):3406–20. doi: 10.1177/0142331219826665.