

Article

Fault Diagnosis of Hydro-Power Units Using BP Neural Network and XGBoost Algorithm for Enhanced Operational Safety

Lei Kuang ¹, Yangyang Zeng ¹, Changxiong Hu ¹, Wenjun Yao ^{1,*}, Jiang Guo ²  and Junjie Hu ³

¹ Hubei Qingjiang Hydropower Development Co., Ltd., Yichang 443000, China

² School of Power and Mechanical Engineering, Wuhan University, Wuhan 430000, China

³ Faculty of Engineering, China University of Geosciences, Wuhan 430000, China

* Correspondence: wenjunyao@cypc.com.cn

Abstract

To enhance operational safety and reduce maintenance costs, this study investigates the fault diagnosis of hydro-power units, where the BP neural network and XGBoost algorithm are employed. To filter environmental noise, a combination of the least squares method and dispersion analysis is utilized to filter out irrelevant and erratic operational data. Following this, the random forest algorithm is applied to rank the significance of characteristic parameters, ensuring that only the most relevant features are selected for fault diagnosis. The BP neural network, integrated with expert knowledge, is then used to extract fault characteristics, improving model accuracy. To further refine fault detection and reflect the hydro-power unit's real-time operation, the XGBoost algorithm is employed for fault identification. A case study demonstrates the model's ability to predict fault characteristics 16 h in advance, confirming the effectiveness and reliability of the proposed diagnostic approach.

Keywords: hydro-power unit; fault diagnosis; BP neural network; XGBoost algorithm

1. Introduction

With the increasing global emphasis on environmental protection, the reliance on fossil fuels has been steadily decreasing. In this context, water energy has emerged as a promising alternative and is being widely developed to meet growing energy demands. Accordingly, water energy development and its environmental impacts have been studied [1–4]. Hydro-power, as a clean and renewable energy source, plays a significant role in reducing carbon emissions and mitigating the environmental impact of traditional energy production methods. However, hydro-power units are often installed in remote locations, which presents a unique set of operational challenges. In the event of a fault, the cost of maintenance is typically high, exacerbated by the difficulty of accessing these units for repairs and diagnostics.

As such, ensuring the safe operation of hydro-power units and minimizing maintenance costs are paramount concerns for the industry. This has led to the growing importance of fault diagnosis research in hydro-power units. By developing reliable fault diagnosis models, it is possible to identify potential issues early, allowing for proactive maintenance strategies that can reduce both downtime and operational costs. Under the background of intelligent operation and maintenance, developing effective hydro-power unit fault diagnosis methods is becoming a major area of research, and various approaches have been proposed to improve fault diagnosis models for hydro-power units [5–14]. For instance, Gabriel et al. [5] developed a fault diagnosis model based on the hidden



Academic Editor: Jun-Qiang Wang

Received: 28 December 2025

Revised: 28 January 2026

Accepted: 31 January 2026

Published: 2 February 2026

Copyright: © 2026 by the authors.

Licensee MDPI, Basel, Switzerland.

This article is an open access article distributed under the terms and conditions of the [Creative Commons Attribution \(CC BY\)](https://creativecommons.org/licenses/by/4.0/) license.

Markov model, which has shown effectiveness in dealing with uncertain system behaviors. Mugarra et al. [7] utilized frequency response analysis to detect faults, focusing on the hydro-power unit's dynamic response to operating conditions. Additionally, Saleem et al. [9] introduced a predictive maintenance framework that aims to optimize maintenance schedules and minimize unplanned downtime. Dao et al. [11] employed chaotic quadratic interpolation optimization in a deep learning model to enhance the accuracy of fault diagnosis predictions. While these models have demonstrated promising results, they often face challenges due to the complex and variable operating environments of hydro-power units. For example, the stability of prediction residuals in some models is poor, which limits their practical application and effectiveness in real-world settings.

Given these challenges, this study focuses on improving the fault diagnosis accuracy for hydro-power units by integrating two advanced machine learning techniques: the BP neural network and the XGBoost algorithm. Compared to the above fault diagnosis algorithms, the advantages of the BP neural network and XGBoost algorithm are obvious, including the ability to handle nonlinear data, integration of expert knowledge, high classification accuracy, and real-time applicability. We acknowledge the effectiveness of the hidden Markov model (HMM) in handling uncertainty. However, the BP neural network combined with XGBoost in our study is chosen for its ability to extract deep nonlinear features and perform high-accuracy classification in complex operational environments. Unlike HMM, which is more suited for sequential and state-based uncertainty, our approach integrates expert knowledge and real-time adaptability, making it more suitable for multi-parameter fault diagnosis. The BP neural network is employed to extract fault characteristics, leveraging both the model's learning capability and expert knowledge to enhance fault detection accuracy. Meanwhile, the XGBoost algorithm is used to reflect the real-time operational status of the hydro-power unit, identifying fault characteristics with higher precision. By combining these two methods, the proposed fault diagnosis model aims to overcome the limitations of previous models and improve both the accuracy and reliability of fault detection. Of course, there are some drawbacks to the BP neural network, for example, the need for large labeled datasets, the sensitivity to noise, and parameter tuning. To overcome these drawbacks, the environment noise of the dataset is removed before training the model, and expert experience is integrated to guide the parameter tuning.

A case study is conducted to validate the proposed fault diagnosis model. The results demonstrate that the model can accurately identify fault characteristics up to 16 h in advance, showcasing its potential for real-world application and its contribution to enhancing the operational safety and cost-effectiveness of hydro-power units.

2. Framework of Fault Diagnosis Model

Figure 1 illustrates the framework of the proposed fault diagnosis model, which integrates the BP neural network and XGBoost algorithm. The fault diagnosis process begins with the collection of real-time operational data from the hydro-power unit through a data acquisition system. This system captures three types of data: normal, fault, and warning data. Normal data correspond to the hydro-power unit's standard operating conditions, while fault data represent abnormal operating states, and warning data indicate transient conditions between normal and abnormal operations.

However, environmental factors inevitably interfere with the data collection system, leading to the introduction of noise in the operational data. To accurately reflect the true operating status of the hydro-power unit, it is essential to remove the environmental noise from the data [12,15–17]. In this study, a combination of the least squares method and

dispersion analysis is employed to effectively eliminate extraneous and erratic data points, thereby improving the quality of the operational data.

Once the noise is removed, the complexity and accuracy of the fault diagnosis model are carefully balanced using the random forest algorithm. This algorithm ranks the significance of various characteristic parameters, allowing for the selection of those most relevant to fault detection. To further enhance the diagnostic accuracy, the BP neural network is utilized, in conjunction with expert knowledge, to extract fault characteristics from the data.

After extracting these fault characteristics, they are combined to represent the true operational state of the hydro-power unit. Finally, the XGBoost algorithm is applied to identify these fault characteristics, offering an advanced, accurate approach to fault diagnosis.

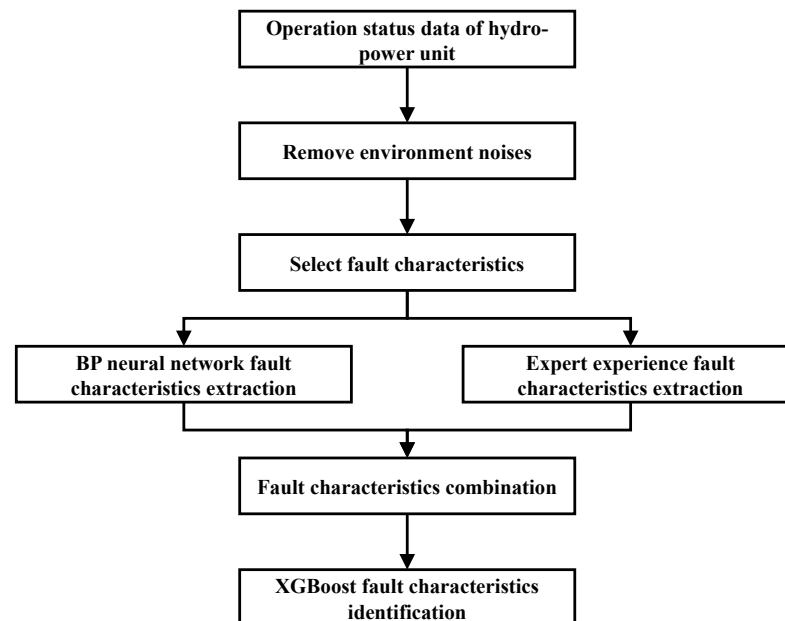


Figure 1. Framework of the fault diagnosis model.

3. Key Processes of Fault Diagnosis Model

3.1. Environment Noise Removal

To effectively eliminate environmental noise from the operational status data, a combination of the least squares method and dispersion analysis is employed [18–21]. The process begins by applying the least squares method to calculate the Euclidean distance between operational status data points, defined as

$$d_{ij} = \sqrt{(x_i - x_j)^2 + (y_i - y_j)^2}, \quad (1)$$

where d_{ij} is the Euclidean distance among the operation status data. The Euclidean distance reflects the relative position of the operation status data, namely, the normal data is close to each other, while the noise data is closer to each other. Next, the standard deviation among the operation status data is calculated, namely,

$$\sigma = \sqrt{\frac{\sum_{i=1}^N [(x_i - \mu_i^x)^2 + (y_i - \mu_i^y)^2]}{N}}, \quad (2)$$

where σ is the standard deviation among the operation status data,

$$\mu_i^x = \frac{\sum_{i=1}^N x_i}{N} \quad (3)$$

is the mean of the operation status data component x ,

$$\mu_i^y = \frac{\sum_{i=1}^N y_i}{N} \quad (4)$$

is the mean of the operation status data component y , and N is the number of the operation status data. Then, to remove the environment noise, the threshold of the operation status data is set, namely,

$$\sigma^* = 3\sigma, \quad (5)$$

where σ^* is the threshold of the operation status data. When the Euclidean distance is not more than the threshold, namely, $d_{ij} \leq \sigma^*$, the operation status data are retained, while the operation status data are removed when the Euclidean distance is larger than the threshold, namely, $d_{ij} > \sigma^*$. Through the least square method, the separate operation status data are removed. We recognize that the method could be more accurately described as a distance-based outlier removal filter with least squares trend fitting. The 3σ threshold was chosen based on empirical evidence from similar engineering applications and preliminary data analysis, which showed that it effectively isolated anomalous points without oversmoothing valid operational variations. To compensate for the shortcoming of the least square method, the dispersion analysis is adopted to remove the discrete operation status data. Furthermore, the operation status data is divided into small intervals, and the mean square error of each interval is calculated, namely,

$$\epsilon_i = \sqrt{\frac{\sum_{j=1}^{N_i} [(x_j - \mu_j^x)^2 + (y_j - \mu_j^y)^2]}{N_i}}, \quad (6)$$

where ϵ_i is the mean square error of the i -th interval. To remove the discrete operation status data, determining the range of each interval is critical. If the range is too small, the integrity of the operation status data is broken, while the discrete operation status data cannot be removed effectively if the range is too large. Lastly, the range of each interval is set to be $[-1.5\epsilon_i, 1.5\epsilon_i]$, within the range, the operation status data are retained, otherwise, the operation status data are removed.

Figure 2 shows the operation status data before and after preprocessing. Before preprocessing, the original operation status data are polluted by the environment noise, which makes it difficult to reflect the real operation status of the hydro-power unit. After preprocessing with the least square method and dispersion analysis, the separate and discrete operation status data are removed, namely, the environment noise is removed, and the retained operation status data are composed of normal data, fault data, and warning data, which are effective to train the fault diagnosis model. In our study, the 3σ filtering was applied only to remove extreme transient environmental artifacts (e.g., sensor spikes, transmission errors) that were uncorrelated with known fault patterns.

3.2. Fault Characteristics Selection

To train the fault diagnosis model, selecting reasonable characteristic parameters to reflect the operation status of the hydro-power unit is critical. If too many characteristic parameters are selected, the complexity of the fault diagnosis model is increased, while the accuracy of the fault diagnosis model is affected if too few characteristic parameters are selected. To balance the complexity and accuracy of the fault diagnosis model, the random forest algorithm is adopted to select the reasonable characteristic parameters. Compared to other algorithms, the random forest algorithm is advantageous for treating high-dimensional data and ranking the importance of the characteristic parameters [22–25].

Therefore, the random forest algorithm is adopted to select the characteristic parameters closely related to the fault diagnosis model.

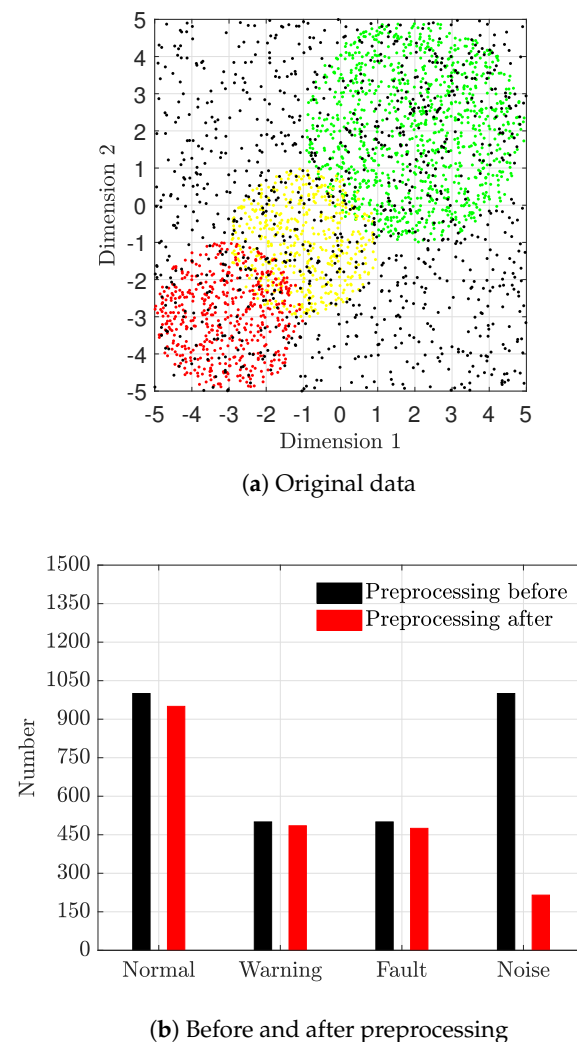


Figure 2. Operation status data before and after preprocessing.

To rank the importance of the characteristic parameters, Figure 3 shows the steps of the random forest algorithm. Firstly, we determine all input and output variables. The operation status of the hydro-power unit is affected widely, including the flow, generator speed, spindle speed, actual torque, set torque, cabin angle, cable angle, pitch angle, yaw angle, and temperature. Among the characteristic parameters, the importance of each characteristic parameter is different; some of them are related to the operation status of the hydro-power unit closely, while others are insignificant. Next, we select some input variables and predict the output variables. Then, according to the selected input variables, we calculate the average error between the predicted output variables and actual output variables. Lastly, we add the noise to each characteristic parameter separately and calculate the average error change before and after adding the noise. The larger the average error change is, the more important the characteristic parameter is. To rank the importance of the characteristic parameters, the average error change (AEC) is calculated, namely,

$$AEC = \frac{1}{N} \sum_{i=1}^N (\Delta\epsilon - \Delta\epsilon^*), \quad (7)$$

where N is the number of the decision trees, $\Delta\epsilon$ is the average error change before adding the noise, and $\Delta\epsilon^*$ is the average error change after adding the noise.

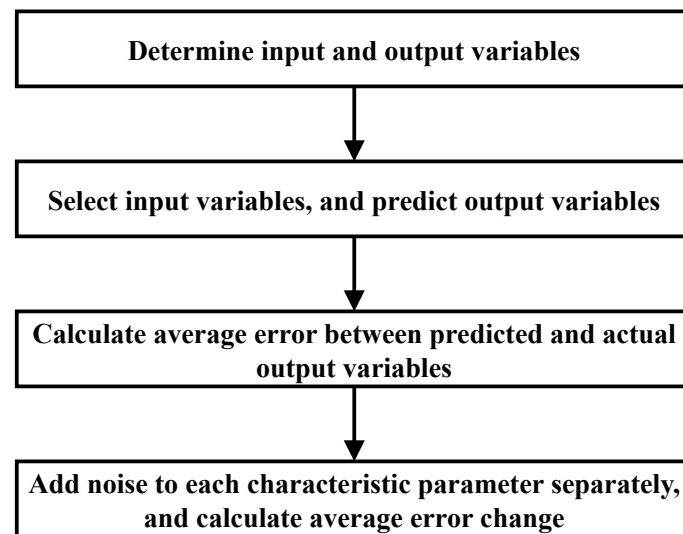


Figure 3. Steps of the random forest algorithm.

Figure 4 shows the characteristic importance ranking of the input variables. In the fault diagnosis of the hydro-power unit, the input variables are multi-dimensional, and the unit of each input variable is different. To rank the characteristic importance of the input variables, the input variables are normalized with min–max scaling, namely,

$$\phi_i^* = \frac{\phi_i - \phi_i^{\min}}{\phi_i^{\max} - \phi_i^{\min}}, \quad (8)$$

where ϕ_i is the actual value with unit, $\phi_i^{\max}$ is the maximum value of the variable, and $\phi_i^{\min}$ is the minimum value of the variable. By conducting the normalization, the input variables take on a dimensionless score between 0 and 1. According to the characteristic importance ranking, the characteristic importance of each input variable is different. Among the input variables, the characteristic importance of the flow is the highest, followed by the set torque, generator speed, spindle speed, and actual torque. Comparatively, the characteristic importance of the cable angle, cabin angle, yaw angle, temperature, and pitch angle is relatively low. In fact, the fluid–turbine interaction is significant, which reflects the operation status of the hydro-power unit directly. When the hydro-power unit is under an abnormal status, the flow and torque are affected first, which are relevant to the fluid–turbine interaction. Therefore, the characteristic importance of the flow and torque are the highest. The higher the characteristic importance is, the closer the relationship between the input variable and operation status of the hydro-power unit is. To balance the complexity and accuracy of the fault diagnosis model, the top five characteristic parameters are selected to reflect the operation status of the hydro-power unit, including the flow, set torque, generator speed, spindle speed, and actual torque. To present the fault detection performance with and without the filtering step, we conduct a comparative analysis. Compared to the operation status data without filtering, the characteristic importance ranking of the input variables with filtering are the same, namely, the integrity of the operation status data is unaffected by the environment noise removal.

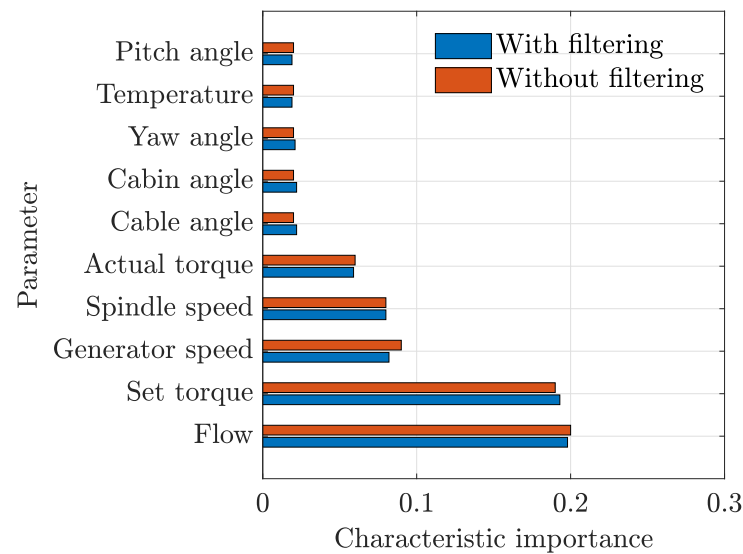


Figure 4. Characteristic importance ranking of the input variables.

3.3. Bp Neural Network

To develop the fault diagnosis model of the hydro-power unit, the BP neural network and expert experience are combined to extract the fault characteristics. Figure 5 shows the structure of the back-propagation neural network. The back-propagation neural network is composed of an input layer, hidden layer, and output layer, which is advantageous to extract fault characteristics [26–29]. Meanwhile, expert experience is valuable, helping to improve the effectiveness of the fault characteristics extraction. In our study, the expert experience refers to the domain knowledge from hydro-power engineers, including fault symptom prioritization, namely, which parameters (e.g., flow, torque) are most indicative of specific faults, BP neural network tuning, namely, guiding the selection of hidden layers and activation functions based on historical fault patterns, and labeling validation, namely, verifying the annotated fault/warning/normal data categories.

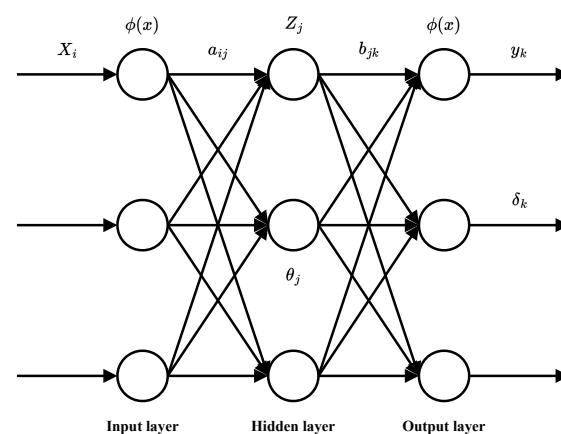


Figure 5. Structure of the back-propagation neural network.

To adopt the back-propagation neural network to extract the fault characteristics, determining the number of the hidden layer neurons is critical. If there are too many hidden layer neurons, the model complexity is increased, while the model accuracy is affected if there are too few hidden layer neurons. To balance the complexity and accuracy, determining the optimal number of the hidden layer neurons is critical, and expert experience is important. To improve the effectiveness of the fault characteristics extraction, expert

experience is valuable. The fault diagnosis of hydro-power units is complex, and there are no mature methods and theories to determine the optimal hidden layer neuron number. Therefore, referring to expert experience to determine the optimal hidden layer neuron number is more effective [30,31]. Empirically, the number of the hidden layer neurons is calculated by

$$h = \sqrt{m + n} + a, \quad (9)$$

where h is the number of the hidden layer neurons, m is the node number of the input layer, n is the node number of the output layer, and a is the constant between 1 and 10. After the characteristic parameter selection, the node number of the input layer is $m = 5$, and the node number of the output layer is $n = 1$; therefore, the number of hidden layer neurons is between 3 and 13.

Figure 6 shows the average errors and training times under different numbers of hidden layer neurons. With the increase in the number of hidden layer neurons, the average error decreases monotonically, namely, the model accuracy is improved. Meanwhile, the model becomes more complex with the increase in the hidden layer neuron number; therefore, the training time is increasing. To balance the complexity and accuracy, the number of hidden layer neurons is set to 10. Therefore, the back-propagation neural network is composed of 5 inputs, 1 output, and 10 hidden layer neurons.

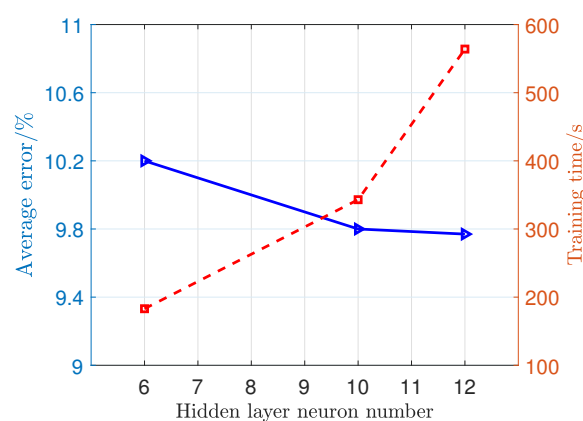


Figure 6. Average errors and training times under different numbers of hidden layer neurons.

After the back-propagation neural network model is constructed, Figure 7 shows the actual operation status data and predicted operation status data. The predicted operation status data reflect the actual operation status data, namely, the BP neural network can show the real operation status of the hydro-power unit effectively.

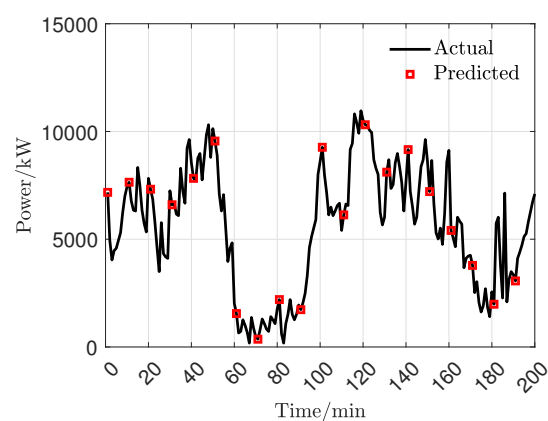


Figure 7. Actual operation status data and predicted operation status data.

3.4. Xgboost Algorithm

The XGBoost algorithm is an ensemble learning method, which is composed of the decision trees with different weights. Compared to other classification algorithms, due to the accuracy and robustness, the XGBoost algorithm is advantageous in identifying the fault characteristics. To avoid over-fitting, a regularization term is introduced, and the objective function of the XGBoost algorithm is

$$f_{\text{obj}}^t = \sum_i^n [y_i - (y_i^* + f_t)]^2 + \Omega(f_t) + C, f_t \in F, t = 1, 2, \dots, i, \quad (10)$$

where n is the number of the operation status data, y_i^* is the predicted operation status data, F is the set of the decision trees, $\Omega(f_t)$ is the regularization term of the decision tree complexity, and C is the constant term.

Identifying the fault characteristics accurately is critical to the fault diagnosis model; the fault diagnosis model should reflect the real operation status of the hydro-power unit, including normal operation status, fault operation status, and warning operation status. Figure 8 shows the normalized confusion matrix. The XGBoost algorithm can distinguish the real operation status of the hydro-power unit effectively: the classification accuracy of the normal operation status data is 0.85, the classification accuracy of the fault operation status data is 1, and the classification accuracy of the warning operation status data is 0.8.

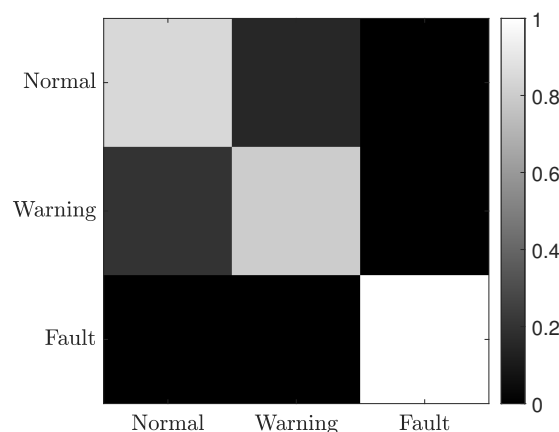


Figure 8. Normalized confusion matrix.

Furthermore, the performance of two models is compared: one is the present combination model, namely, XGBoost+BP neural network, and the other is XGBoost without a BP neural network for feature extraction. Figure 9 shows the curves of the accuracy–confidence and accuracy–recall rate. Obviously, with the increase in the accuracy, the confidence of the two models is increasing, meaning that using the XGBoost model to identify the fault characteristics is reliable. Meanwhile, the recall rate of the two models is decreasing with the increase in the accuracy; the XGBoost algorithm can distinguish the real operation status of the hydro-power unit effectively. Regarding the XGBoost model as the baseline model, the XGBoost+BP combination model is more advantageous; the confidence is higher, while the recall rate is lower. By introducing the BP neural network, the fault characteristics are extracted first, which is helpful in identifying the fault characteristics for the XGBoost algorithm.

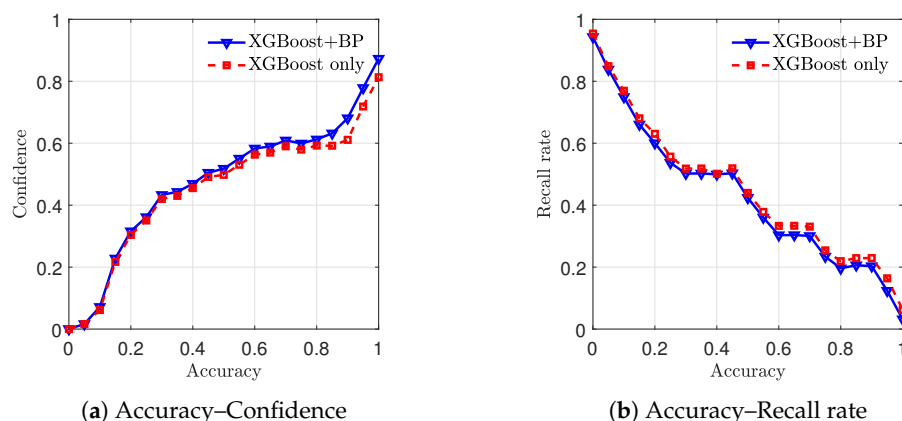


Figure 9. Curves of accuracy–confidence and accuracy–recall rate.

4. Case Study

To demonstrate the effectiveness of the proposed fault diagnosis model, a comprehensive case study is conducted. The historical operational status data from a 1.5 MW hydro-power unit located in Hubei Province, China, are utilized as the basis for this study. This dataset includes various operational parameters collected over an extended period, providing valuable insights into the performance and fault behaviors of the hydro-power unit. The real-time operation status data of the hydro-power unit is recorded with the data collection system, including the time, power, flow, generator speed, cabin angle, yaw angle, temperature, and so on. The sampling time is from 0:00 on 1 January 2024 to 0:00 on 1 January 2025, and the sampling period is 1 min. Table 1 shows the real-time operation status data of the hydro-power unit.

Table 1. Real-time operation status data of the hydro-power unit.

Time	Power/kW	Flow/(m ³ /s)	Generator Speed/(r/min)	...	Temperature/°C
2024/1/1 0:00	3424.8	65.1	1491.2	...	16.7
2024/1/1 0:01	3625.5	58.4	1505.7	...	16.7
2024/1/1 0:02	2526.6	43.4	1332.2	...	16.7
⋮	⋮	⋮	⋮	⋮	⋮

The first step in the case study involves preprocessing the data to eliminate environmental noise that could interfere with accurate fault detection. To achieve this, a combination of the least squares method and dispersion analysis is employed. These techniques effectively filter out irrelevant and noisy data, ensuring that the operational status data reflect the true performance of the hydro-power unit. This noise removal process is critical, as it enhances the quality of the data, enabling more precise fault diagnosis.

Once the noise is removed, the next step is to optimize the complexity and accuracy of the fault diagnosis model. To achieve this, the random forest algorithm is utilized to rank the importance of various characteristic parameters, which are critical to the fault diagnosis process. These parameters include the flow rate, set torque, generator speed, spindle speed, and actual torque, variables that directly influence the performance and reliability of the hydro-power unit. By ranking these parameters, the random forest algorithm helps to identify the most relevant features, improving the model's efficiency by selecting only the most impactful variables.

To further enhance the diagnostic accuracy, the BP neural network is combined with expert knowledge to extract fault characteristics from the operational data. This hybrid approach leverages the computational power of the BP neural network, which excels in identifying complex patterns and relationships within the data, while expert experience provides valuable insights that can guide the model in detecting subtle anomalies that may indicate faults.

Finally, to ensure that the fault diagnosis model accurately reflects the real-time operational status of the hydro-power unit, the XGBoost algorithm is employed. XGBoost, renowned for its high performance and robustness in classification tasks, is used to classify and identify the fault characteristics in the operational data. This algorithm enhances the model's ability to detect faults with high precision, ensuring that it can provide reliable fault identification under various operating conditions.

Through this case study, the proposed fault diagnosis model demonstrates its ability to effectively detect and classify faults in the hydro-power unit. By combining data preprocessing, feature ranking, neural network-based fault extraction, and advanced classification techniques, the model achieves high diagnostic accuracy, offering significant potential for improving the operational safety and reliability of hydro-power units.

To validate the fault diagnosis model, the fault record table of the hydro-power unit is checked, where there was a pitch system fault at 15:30 on 9 May 2024. Figure 10 shows the operation status index before and after the fault occurrence. When the hydro-power unit is under a normal operation status, the operation status index is below the threshold; while the operation status index is over the threshold, the hydro-power unit is under an abnormal operation status. The operation status index is a composite score derived from the normalized outputs of the BP neural network and XGBoost classification probabilities. In the revised manuscript, we explicitly define the index mathematically, namely,

$$\text{OSI} = \omega_1 \cdot \text{BP}_{\text{output}} + \omega_2 \cdot \text{XGBoost}_{\text{condifer}}, \quad (11)$$

where weights ω_1 , ω_2 are determined via grid search, $\text{BP}_{\text{output}}$ is the output of the BP neural network, and $\text{XGBoost}_{\text{condifer}}$ is the XGBoost classification probability. To reflect the operation status of the hydro-power unit, the fault characteristics are extracted and identified, and the fault warning signal is generated when the operation status index exceeds the threshold for the first time. At 23:30 on May 8, the operation status index was over the threshold, and the fault warning signal was generated. At the time, the hydro-power unit was under an abnormal operation status, until 15:30 on May 9, when the fault occurred. Before the fault occurrence, the fault characteristics are extracted and identified 16 h in advance, proving the effectiveness of the fault diagnosis model.

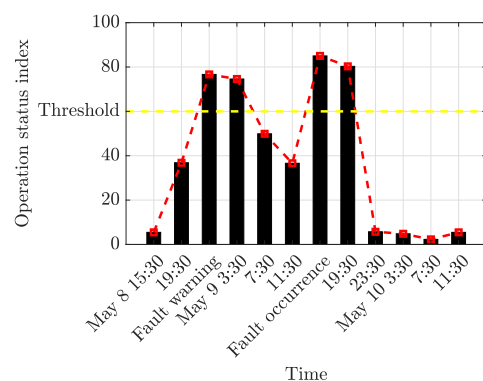


Figure 10. Operation status index before and after the fault occurrence.

Furthermore, Figure 11 illustrates the actual operation status and the predicted operation status of the hydro-power unit. The operation status data are categorized into three types: normal data, fault data, and warning data. Each of these data types reflects a different operational state of the hydro-power unit, with normal data representing standard operating conditions, fault data indicating abnormal operational behavior, and warning data signaling potential transitions between normal and abnormal states.

Accurately reflecting the true operation status of the hydro-power unit requires the precise extraction and identification of fault characteristics. This step is crucial, because identifying fault features allows for the detection of early signs of potential failures, enabling proactive maintenance and minimizing downtime. The fault diagnosis model presented in this study is specifically designed to extract and identify these fault characteristics with high accuracy.

By effectively distinguishing between normal, fault, and warning states, the model plays a significant role in enhancing the operational safety of the hydro-power unit. The ability to detect faults early not only helps prevent catastrophic failures but also supports optimized maintenance schedules, ultimately contributing to the long-term efficiency and reliability of the system.

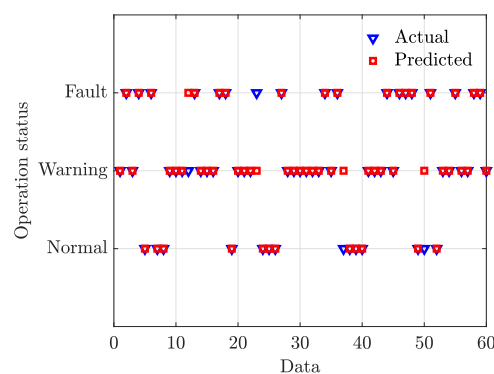


Figure 11. Actual operation status and predicted operation status.

5. Discussion

In response to the ongoing advancement of intelligent hydro-power stations, this study presents a fault diagnosis model for hydro-power units that integrates the BP neural network and XGBoost algorithm. The need for such models is driven by the increasing complexity of hydro-power unit operations and the critical importance of maintaining efficient, reliable, and safe performance. Under normal operational conditions, the operation status index remains below a predefined threshold, whereas, during abnormal operation, the index exceeds this threshold, indicating potential faults or system malfunctions.

The first step in the fault diagnosis process involves preprocessing the operation status data to eliminate noise and irrelevant information. The least squares method and dispersion analysis are combined to effectively remove separate and discrete operation status data points, ensuring the dataset reflects only the relevant operational patterns. This preprocessing step is crucial for ensuring the accuracy and reliability of the fault detection process.

To balance the complexity and accuracy of the fault diagnosis model, the random forest algorithm is utilized to rank the importance of various characteristic parameters. These parameters, including the flow rate, set torque, generator speed, spindle speed, and actual torque, are closely associated with the performance and fault detection of the hydro-power unit. By ranking these parameters, the model identifies the most critical variables that contribute to accurate fault diagnosis.

Next, to further enhance the precision of the fault diagnosis model, a hybrid approach is employed, combining the BP neural network with expert knowledge to extract fault characteristics from the operation status data. This combination leverages the computational power of the neural network and the practical insights of expert experience, enabling the model to identify subtle patterns and anomalies indicative of impending failures.

Finally, to ensure that the fault diagnosis model reflects the real-time operational status of the hydro-power unit, the XGBoost algorithm is employed to identify the fault characteristics accurately. XGBoost, known for its robustness in classification tasks, further improves the model's ability to detect faults with high accuracy and reliability.

However, we should point out that there are some limitations of the present study. Firstly, only a single object (1.5 MW, one location) is concerned in the present study, which is not the case in the real hydro-power station. Usually, there are several hydro-power units in one hydro-power station, and the fault diagnosis system should monitor and predict the operation status of all these units, and the interaction among the units should be taken into account, which is more complicated than the single unit. Additionally, in a real hydro-power production scenario, the types of hydro-power units can be different, and their fault characteristics can be different from each other. The present study only considers a single unit, which does not cover other types of units. The present study provides a framework to develop the hydro-power unit fault diagnosis system; by transfer learning or recalibration, the present model can be applied to other types of units.

6. Conclusions

In response to the ongoing advancement of intelligent hydro-power stations, this study presents a fault diagnosis model for hydro-power units that integrates the BP neural network and XGBoost algorithm. The first step in the fault diagnosis process involves preprocessing the operation status data to eliminate noise and irrelevant information. To balance the complexity and accuracy of the fault diagnosis model, the random forest algorithm is utilized to rank the importance of various characteristic parameters. Next, to further enhance the precision of the fault diagnosis model, a hybrid approach is employed, combining the BP neural network with expert knowledge to extract fault characteristics from the operation status data. Finally, to ensure that the fault diagnosis model reflects the real-time operational status of the hydro-power unit, the XGBoost algorithm is employed to identify the fault characteristics accurately. To validate the effectiveness of the proposed fault diagnosis model, a comprehensive case study is conducted. The model successfully extracts and identifies fault characteristics up to 16 h in advance, demonstrating its potential for proactive maintenance and early fault detection. Furthermore, the model achieves a classification accuracy higher than 95%, highlighting its capacity to significantly enhance the operational safety of hydro-power units.

In the future, to describe the operation status of the hydro-power unit accurately, considering the fluid–turbine interaction is significant. Introducing computational fluid dynamics (CFD) technologies and establishing a digital twin model of the hydro-power unit are new research areas to develop the intelligent operation and maintenance management of a hydro-power station.

Author Contributions: L.K., Writing—Original Draft Preparation; Y.Z., Software; C.H., Software; W.Y., Resources; J.G., Conceptualisation; J.H., Conceptualisation. All authors have read and agreed to the published version of the manuscript.

Funding: The present study was supported by the Research and Demonstration Application of the Unit Condition Monitoring System and Intelligent Early Warning Method for Geheyan Power Plant, Qingjiang Company (No. ENQJ-GY1-HW-2024036).

Data Availability Statement: The original contributions presented in this study are included in the article. Further inquiries can be directed to the corresponding author.

Conflicts of Interest: Authors Lei Kuang, Yangyang Zeng, Changxiong Hu, Wenjun Yao were employed by the company Hubei Qingjiang Hydropower Development Co., Ltd. The remaining authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest. The company had no role in the design of the study; in the collection, analyses, or interpretation of data; in the writing of the manuscript, or in the decision to publish the results.

References

1. Wei, M.; Zheng, Z.; Bai, X. Application of rough set and neural network in water energy utilization. *Front. Energy Res.* **2021**, *9*, 604660. [\[CrossRef\]](#)
2. Borhan, S.; Dariush, K.; Javad, E. Integrated optimization of multi-carrier energy systems: Water-energy nexus case. *Energy* **2022**, *257*, 124764. [\[CrossRef\]](#)
3. Alberto, B. The water-energy-environment-economy nexus progressing toward net zero. *Water-Energy Nexus* **2023**, *6*, 13. [\[CrossRef\]](#)
4. Zhilkashinova, A.; Ocheredko, I.; Azamatov, B. Modeling and Comparison of Design Features of Pendulum and Radial Micro-Hydropower Plants Considering the Influence of Variable Design Parameters. *Designs* **2024**, *8*, 101. [\[CrossRef\]](#)
5. Gabriel, P.; Roberto, R.; Fernandes, L. Bearing Diagnostics of Hydro Power Plants Using Wavelet Packet Transform and a Hidden Markov Model with Orbit Curves. *Shock. Vib.* **2018**, *2018*, 5981089. [\[CrossRef\]](#)
6. Gou, B.; Xu, Y.; Xia, Y. An Intelligent Time-Adaptive Data-Driven Method for Sensor Fault Diagnosis in Induction Motor Drive System. *IEEE Trans. Ind. Electron.* **2019**, *66*, 9817. [\[CrossRef\]](#)
7. Mugarra, A.; Platero, C. Validity of Frequency Response Analysis (FRA) for Diagnosing Large Salient Poles of Synchronous Machines. *IEEE Trans. Ind. Appl.* **2020**, *56*, 226. [\[CrossRef\]](#)
8. Liu, J.; Schmid, F.; Li, K. A knowledge graph-based approach for exploring railway operational accidents. *Reliab. Eng. Syst. Saf.* **2021**, *207*, 107352. [\[CrossRef\]](#)
9. Afridi, Y.S.; Ahmad, K.; Hassan, L. Artificial intelligence based prognostic maintenance of renewable energy systems: A review of techniques, challenges, and future research directions. *Int. J. Energy Res.* **2022**, *46*, 21619. [\[CrossRef\]](#)
10. Guo, B.; Wang, Y.; Pan, W. Fault diagnosis method for hydro-power plants with Bi-LSTM knowledge graph aided by attention scheme. *J. Vibroeng.* **2023**, *25*, 1629. [\[CrossRef\]](#)
11. Dao, F.; Zeng, Y.; Zou, Y. Fault diagnosis method for hydropower unit via the incorporation of chaotic quadratic interpolation optimized deep learning model. *Measurement* **2024**, *237*, 115199. [\[CrossRef\]](#)
12. Jad, S.; Desforges, X.; Medjaher, K.; Villard, P.Y.; Caussid ry, C. Condition monitoring and diagnostic of hydropower units based on machine learning techniques. *Renew. Energy* **2025**, *249*, 123210. [\[CrossRef\]](#)
13. Chen, Y.; Wang, J.; Chen, W. Fault diagnosis in hydropower units based on chaotic Kepler optimization algorithm-enhanced BiLSTM model. *Energy Rep.* **2024**, *12*, 5163. [\[CrossRef\]](#)
14. Tan, J.; Radhi, R.; Shirini, K. Innovative framework for fault detection and system resilience in hydropower operations using digital twins and deep learning. *Sci. Rep.* **2025**, *15*, 15669. [\[CrossRef\]](#)
15. Kumar, K.; Saini, R. A review on operation and maintenance of hydropower plants. *Sustain. Energy Technol. Assess.* **2022**, *49*, 101704. [\[CrossRef\]](#)
16. Liu, Y.; Xu, Y.; Liu, J. Real-time comprehensive health status assessment of hydropower units based on multi-source heterogeneous uncertainty information. *Measurement* **2023**, *216*, 112979. [\[CrossRef\]](#)
17. Liu, T.; Kong, F.; Yang, L. Operational risk assessment of hydropower units based on PSSCA-VMD-CNN-GBiLSTM and multi-feature fusion. *Comput. Electr. Eng.* **2024**, *118*, 109412. [\[CrossRef\]](#)
18. Ding, F. Least squares parameter estimation and multi-innovation least squares methods for linear fitting problems from noisy data. *J. Comput. Appl. Math.* **2023**, *426*, 115107. [\[CrossRef\]](#)
19. Ding, F.; Xu, L.; Zhang, X. Hierarchical gradient- and least-squares-based iterative estimation algorithms for input-nonlinear output-error systems from measurement information by using the over-parameterization. *Int. J. Robust Nonlinear Control.* **2024**, *34*, 1120–1147. [\[CrossRef\]](#)
20. Malburet, C.; Martin, M.; Leclercq, L. Optimization of limit of detection in Taylor dispersion analysis: Application to the size determination of vaccine antigens. *Talanta Open* **2023**, *7*, 100209. [\[CrossRef\]](#)
21. Huang, J.; Peng, W.; Yang, J. Overview of computation strategies on the dispersion analysis for explicit finite difference solution of acoustic wave equation. *Pet. Sci.* **2024**, *21*, 2311. [\[CrossRef\]](#)

22. Balyan, A.; Ahuja, S.; Lilhore, U. A Hybrid Intrusion Detection Model Using EGA-PSO and Improved Random Forest Method. *Sensors* **2022**, *22*, 5986. [[CrossRef](#)] [[PubMed](#)]
23. Ho, C.; Cheng, K.; Ang, C. Utilizing the Random Forest Method for Short-Term Wind Speed Forecasting in the Coastal Area of Central Taiwan. *Energies* **2023**, *16*, 1374. [[CrossRef](#)]
24. Tan, X.; Li, Y.; Wu, X. Identification of the key driving factors of flash flood based on different feature selection techniques coupled with random forest method. *J. Hydrol. Reg. Stud.* **2024**, *51*, 101624. [[CrossRef](#)]
25. Ahmed, I. Analyzing the impact of blockchain technology on banking transaction costs using the random forest method. *Front. Blockchain* **2025**, *8*, 1551970. [[CrossRef](#)]
26. Wen, J.; Chen, X.; Li, X. SOH prediction of lithium battery based on IC curve feature and BP neural network. *Energy* **2022**, *261*, 125234. [[CrossRef](#)]
27. Chen, J.; Liu, Z.; Yin, Z. Predict the effect of meteorological factors on haze using BP neural network. *Urban Clim.* **2023**, *51*, 101630. [[CrossRef](#)]
28. Li, X.; Ke, S.; Li, Y. Temperature compensation based on BP neural network with small sample data for chloride ions optical fiber probe. *Opt. Laser Technol.* **2024**, *176*, 110973. [[CrossRef](#)]
29. Peng, Y.; Li, M.; Li, Z. What is the impact of discrete memristor on the performance of neural network: A research on discrete memristor-based BP neural network. *Neural Netw.* **2025**, *185*, 107213. [[CrossRef](#)] [[PubMed](#)]
30. Li, Y.; Vong, C.; Chen, P. Analytical selection of hidden parameters through expanded enhancement matrix stability for functional-link neural networks and broad learning systems. *Knowl.-Based Syst.* **2025**, *310*, 112923. [[CrossRef](#)]
31. Thant, Y.; Manzhos, S.; Ihara, M. On the Sufficiency of a Single Hidden Layer in Feed-Forward Neural Networks Used for Machine Learning of Materials Properties. *Physchem* **2025**, *5*, 4. [[CrossRef](#)]

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.