Article

A Decision-Making Method for Machinery Abnormalities Based on Neural Network Prediction and Bayesian Hypothesis Testing

Gaojun Liu ¹, Shan Yang ^{2,*}, Gaixia Wang ¹, Fenglei Li ² and Dongdong You ^{2,*}

¹ State Key Laboratory of Nuclear Power Safety Monitoring Technology and Equipment, China Nuclear Power Engineering Company Ltd., Shenzhen 518172, China; liugaojun@cgnpc.com.cn (G.L.); wanggaixia@cgnpc.com.cn (G.W.)

² School of Mechanical and Automotive Engineering, South China University of Technology, Guangzhou 510640, China; flli@scut.edu.cn

* Correspondence: 202021002904@mail.scut.edu.cn (S.Y.); youdd@scut.edu.cn (D.Y.)

Abstract: For anomaly identification of predicted data in machinery condition monitoring, traditional threshold methods have problems during residual testing. It is difficult to make decisions when the residuals are close to the threshold and fluctuate. This paper proposes a Bayesian dynamic thresholding method that combines Bayesian inference with neural network signal prediction. The method makes full use of historical prior data to build an anomaly identification and warning model applicable under single variable or multidimensional variables. A long short-term memory signal prediction model is established, and then a Bayesian hypothesis testing-based anomaly identification strategy is presented to quantify the probability of anomaly occurrence and issue early warnings for anomalies beyond a certain probability. The model was applied to open data sets of a pumping station and actual operating data of a nuclear power turbine. The results indicate that the model successfully predicts the failure probability and failure time. The effectiveness of the proposed method is verified.

Keywords: Bayesian hypothesis testing; abnormality identification; long short-term memory; nuclear power turbine

Citation: Liu, G.; Yang, S.; Wang, G.; Li, F.; You, D. A Decision-Making Method for Machinery Abnormalities Based on Neural Network Prediction and Bayesian Hypothesis Testing. *Electronics* **2021**, *10*, 1610. <https://doi.org/10.3390/electronics10141610>

Academic Editor: Davide Astolfi

Received: 3 June 2021

Accepted: 29 June 2021

Published: 6 July 2021

Publisher's Note: MDPI stays neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Copyright: © 2021 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) license (<https://creativecommons.org/licenses/by/4.0/>).

1. Introduction

With the development of mechanical equipment fault diagnosis technology, using signals to analyze faults is a common technology. During the production process, various data on the status of equipment can be collected by sensors and monitored in real time [1–4]. In recent years, with the rapid development of machine learning, especially deep learning, various models based on data-driven algorithms capable of predicting future data using historical data have been developed and are widely used in the field of signal prediction. Saufi et al. [5] introduced a deep learning architecture for automatic feature learning to solve the problems of traditional fault detection and diagnosis (FDD) systems and discussed the challenges of deep learning and future work related to mechanical fault diagnosis systems. Among the various deep learning time-series prediction methods, recurrent neural networks (RNNs) and long short-term memory (LSTM) networks are the most widely used. The input of the RNN at each time includes both the input data of the current time and the hidden layer unit data of the previous time. Therefore, these hidden layer units play the role of a memory module, preserving the information of historical data and updating with the input of new data. Zhao et al. [6] reviewed the emerging research work on deep learning in machine health monitoring systems in terms of autoencoders and their variants, constrained Boltzmann machines and their variants, including deep belief networks (DBNs) and deep Boltzmann machines (DBMs), convolutional neural networks (CNNs), and RNNs. Liu et al. [7] presented the low-delay lightweight RNN (LLRNN) model for mechanical fault diagnosis, which occupies less memory space than previous

methods and has low computational latency. The LSTM network is an improvement on the RNN, which mainly solves the problem of gradient disappearance or explosion that exists in RNNs [8] and makes effective use of long-range sequence information. Tian et al. [9] proposed a model that combines self-correlated local feature scale decomposition and an improved LSTM neural network as a hybrid prediction modeling strategy to build a reciprocating compressor vibration signal prediction model. Ma et al. [10] compared LSTM with other prediction algorithms via the mean absolute percentage error (MAPE) and mean squared error (MSE) as evaluation metrics and showed that the LSTM model has better accuracy and stability than the other algorithms. Kim and Won [11] combined LSTM models with various generalized autoregressive conditional heteroskedasticity (GARCH)-type models to significantly improve the forecasting performance for time-series data. Di Persio and Honchar [12] propose a novel approach to energy load time series forecasting, combining deep learning networks with the prior properties of Bayes to predict the data of interest. He et al. [13] used deep confidence network to carry out unsupervised fault diagnosis for gear transmission chains. In this study, LSTM was employed to establish a prediction model.

In anomaly identification, the construction of the prediction model is only one part of the process, and this study focuses on two other aspects: (1) the fault identification decision strategy, which plays a decisive role, i.e., the identification of the residual between the hypothetical health value of the prediction model and the actual monitored value; (2) the data processing method in the multivariate case. The general anomaly decision method is mainly the threshold method and its improved algorithm, i.e., the residuals are recognized as anomalous when they exceed a threshold value. Zeng et al. [14] calculated the performance index of a generic delayer to adjust the threshold appropriately to reduce unexpected alarms. Zadakbar et al. [15] developed a quantitative operational risk assessment model to control the activities of an early warning system by setting a threshold value for acceptable operational risk. Yu et al. [16] designed optimal alarm thresholds for univariate simulated process variables based on alarm probability maps. Gao et al. [17] proposed a new multivariate alarm threshold optimization method based on the consistency of the correlation between process and alarm data and applied a particle swarm optimization (PSO) algorithm to obtain optimal thresholds. Zhao et al. [18] proposed an adaptive threshold determined by extreme value theory to prevent false alarms caused by the extreme value distribution of the calculated detection index in the health monitoring of wind turbines. Aslansefat et al. [19] evaluated the performance metrics of an alarm system with variable alarm thresholds and provided a genetic algorithm-based optimization design process for optimizing the parameter settings to improve the performance metrics. Zhang et al. [20] optimized the objective function by the PSO algorithm to obtain the optimal alarm threshold, which significantly reduced the false alarm rate. Each of these methods for determining the threshold value has its own merits, but in general, they all still have a certain rate of missed diagnoses and false diagnoses, especially when the residual is close to the threshold value. There is no reference standard for setting the threshold value in practice, and finding the optimal threshold incurs a high computational cost and yields an insufficient accuracy of the results.

In recent years, threshold analysis methods combined with Bayesian theory have become popular [21–25]. Cho et al. [26] used Bayesian networks to evaluate residuals and applied the method to fault detection of three-phase asynchronous motors. Nezhad and Niaki [27] developed a heuristic threshold strategy to detect and classify the state of a multivariate quality control system. The posterior confidence values of runaway features were updated using Bayesian rules and compared with decision thresholds to determine the minimum acceptable confidence value. Trachi [28] developed a fault testing strategy for electric motors that was based on likelihood ratio hypothesis testing, but the Bayesian approach was used only for parameter estimation. Jiang et al. [29] used a Bayesian hypothesis testing approach to develop anomaly identification criteria that are determined by testing whether the mean of the residuals is zero. However, if the prediction model is

not accurate enough, those minor anomalies will be covered by the model's own residuals and will be difficult to identify.

As known from above, although the existing thresholding method combined with Bayesian analysis is an improvement on the traditional judgment method, it is not adequate in the study and use of a priori information and has difficulty making judgments in some cases, such as with multivariate data or data with slight anomalies. In this paper, further research will be carried out in the following three aspects: (1) processing anomaly identification for time-series data with multiple variables, (2) making full use of the a priori information for decision-making, which is more in line with the actual application situation, and (3) calculating the confidence level for the identified anomalies and providing early warning based on the quantified probability. The fault anomaly decision process is shown in Figure 1.

Figure 1. Fault anomaly decision-making flow chart.

2. Data Preprocessing in the Multivariate Case

Before constructing the LSTM prediction model, the time-series signal needs to be processed, generally including noise reduction [30], normalization, and phase space reconstruction. The details of wavelet packet denoising and normalization can be found in our previous study [31]. For the multivariate time-series data, the new composite variables are output by dimensionality reduction in this study.

2.1. Dimensionality Reduction

During the operation of mechanical equipment, the number of variables can reach several hundred, and it is necessary to extract the main information by dimensionality reduction to simplify the subsequent calculation process and retain most of the essential information. Probabilistic principal component analysis (PPCA) [32] is employed to reduce the dimensionality and uncertainty of the a priori data, which is based on the Gaussian latent variable model by introducing a probabilistic framework into principal component

analysis to attenuate the influence of noisy variables on the structural characteristics of the data. This approach ensures the full excavation and utilization of limited information and avoids the "dimensional disaster" problem. The expression of the PPCA model is as follows:

$$\mathbf{X} = \mathbf{W}\mathbf{z} + \boldsymbol{\mu} + \boldsymbol{\psi} \quad (1)$$

where $\mathbf{X}$ is a p -dimensional vector, and the data points in each dimension are set to n , i.e., $x_i = \{x_{i1}, x_{i2}, \dots, x_{in}\}$. $\mathbf{W}$ is a weight vector of dimension $p \times r$ ($r \leq p$), $\mathbf{z}$ represents a latent variable of dimension r ($\mathbf{z} \sim N(0, \mathbf{I}_r)$), $\boldsymbol{\mu}$ is the sample mean vector, and $\boldsymbol{\psi}$ is a noise vector ($\boldsymbol{\psi} \sim N(0, \sigma^2 \mathbf{I}_p)$), where $\mathbf{I}$ is the identity matrix. The posterior distribution of $P(\mathbf{z}|\mathbf{X}) \sim N(\mathbf{M}^{-1}\mathbf{W}^T(\mathbf{X} - \boldsymbol{\mu}), \sigma^2\mathbf{M}^{-1})$ for $\mathbf{z}$ is obtained from the Bayesian equation, where $\mathbf{M} = (\sigma^2\mathbf{I}_r + \mathbf{W}^T\mathbf{W})^{-1}$. $\mathbf{W}$ and σ^2 in the model are estimated from the maximum likelihood method as follows:

$$\begin{aligned} \sigma^2 &= \frac{1}{p-r} \sum_{j=r+1}^p \lambda_j \\ \hat{\mathbf{W}} &= \mathbf{U}_r(\Lambda_r - \sigma^2\mathbf{I}_r)^{1/2}\mathbf{R} \end{aligned} \quad (2)$$

where λ_j is obtained by decomposing the covariance matrix of sample $\mathbf{X}$ according to the eigenvalue, i.e., $\mathbf{C}_{vj} = \lambda_j v_j$ and $\mathbf{U}_r = (v_1, v_2, \dots, v_r)$, where v_j is the eigenvector of $\Lambda_r = \text{diag}(\lambda_1, \lambda_2, \dots, \lambda_r)$. The conditional expectation of potential variables is obtained as follows:

$$\mathbf{z} = \mathbf{M}^{-1}\hat{\mathbf{W}}(\mathbf{X} - \boldsymbol{\mu}) \quad (3)$$

The number of principal components of the data can be determined by the cumulative variance contribution rate, which is generally specified to be not less than 80%.

2.2. Phase Space Reconstruction

The predicted data of the signal are likely to be chaotic time-series data, and phase space reconstruction of the data is needed to make the output applicable to the neural network model. The C-C algorithm [33] is used to estimate the delay time τ and the embedding dimension m for phase space reconstruction. For a single-variable time-series $X = \{x_1, x_2, x_3, \dots, x_n\}$, the reconstructed matrix is as follows:

$$\mathbf{X} = \begin{Bmatrix} x_1, x_{1+\tau}, \dots, x_{1+(m-1)\tau} \\ x_2, x_{2+\tau}, \dots, x_{2+(m-1)\tau} \\ \dots \\ x_{N_m}, x_{N_m+\tau}, \dots, x_{N_m+(m-1)\tau} \end{Bmatrix} \quad (4)$$

where $N_m = n - (m-1)\tau$ is the number of vectors after reconstructing the phase space. Each vector represents a reconstructed input data $[x_i, x_{i+\tau}, \dots, x_{i+(m-1)\tau}]$. That is to say, there are m data input nodes.

3. LSTM Prediction Model

By using the dataset after data preprocessing as the input of the LSTM model, the predicted data can be output after determining the number of layers and related parameters. The larger τ is, the longer the time span for predicting future data. The predicted data Y are shown in Equation (5):

$$\mathbf{Y} = \begin{Bmatrix} x_{1+m\tau} \\ x_{2+m\tau} \\ \dots \\ x_{N_m+m\tau} \end{Bmatrix} \quad (5)$$

The detailed structure of the LSTM network is described in our previous work [31], and in this study, single-step prediction is used, i.e., each set of input data contains m values, a total of N_m groups of values are input, and the data input mode of group i is

shown in Figure 2, where h_0 and c_0 are the random values of the initial input, h_{m-1} is the transferred output of the last state, c_{m-1} is the historical information output of the last state, and the circle represents merging the input. After training, a new time series is input, and its output is the corresponding predicted value. It is worth noting that to achieve a contribution rate of 80%, there is a high probability of more than one virtual signal after dimensionality reduction, so a separate decision is made for each reduced virtual signal in the prediction process, and then a comprehensive analysis is performed with Bayesian hypothesis testing.

Figure 2. LSTM training process.

4. Real-Time Bayesian Hypothesis Testing

The Bayesian hypothesis testing method is developed to identify anomalies in real time. Different from the traditional threshold method, Bayesian hypothesis testing can extract useful information from historical prior data, which can be used for subsequent abnormality judgment. In addition, the application of univariate and multivariate data types in the data processing method mentioned above means that the model can deal with most of the data from different industries and can solve various situations with insufficient information conditions, uncertain data, and chaotic multivariate data in actual production processes. The following is the specific construction method.

4.1. Setting the Hypotheses

The threshold approach is to monitor whether the data residuals exceed a set threshold to determine whether to raise the alarm, as shown in Figure 3. In this study, instead of setting the threshold through complex calculation, Bayesian inference is used to establish two hypotheses of the mean μ_0 and variance σ_0^2 of the residuals. Through the form of a sliding window, the dataset $X = \{x_1, x_2, \dots, x_k\}$ is composed of the current point and the previous $k-1$ points, where k represents the length of the sliding window. It can be known that n data points have $n-k+1$ data sets. Assuming that it obeys a normal distribution $N(\mu, \sigma^2)$, the corresponding mean value and variance value can be obtained. Figure 4 shows the residuals and the corresponding means and variances, which fluctuate when the residuals fluctuate dramatically. Therefore, using the mean and variance can be a good substitute for the residuals.

Let the floating values of their mean and variance be ε and δ . The floating values are obtained from the prior information, and the original and alternative hypotheses of the mean and variance are set up in terms of whether the values of both exceed the floating values. Taking the mean μ as an example, where the null hypothesis is $H_0 : \mu \in [\hat{\mu} - \varepsilon, \hat{\mu} + \varepsilon]$ and the alternative hypothesis is $H_1 : \mu \notin [\hat{\mu} - \varepsilon, \hat{\mu} + \varepsilon]$, where $\hat{\mu}$ is the estimate of μ , the hypothesized probabilities are as follows [34]:

$$\begin{aligned} H_0 : \alpha_0 &= \int_{\hat{\mu}-\varepsilon}^{\hat{\mu}+\varepsilon} g(\mu) d\mu \\ H_1 : \alpha_1 &= \int_{-\infty}^{\hat{\mu}-\varepsilon} g(\mu) d\mu + \int_{\hat{\mu}+\varepsilon}^{\infty} g(\mu) d\mu \end{aligned} \quad (6)$$

where $g(\mu)$ is the posterior probability distribution. In the following, $g(\mu)$ and estimate $\hat{\mu}$ will be solved.

Figure 3. Setting the principle of signal anomaly recognition of threshold value.

Figure 4. Residual and the corresponding mean and variance relationships.

4.2. Solving the Parameters

4.2.1. Posterior Distribution Probability Determination

Assuming that dataset X in the above section obeys a normal distribution, its joint density function is:

$$p(X|\mu, \sigma^2) = (2\pi\sigma^2)^{-n/2} \exp\left\{-\frac{1}{2\sigma^2} \sum_{k=1}^n (x_k - \mu)^2\right\} \propto \sigma^{-n} \exp\left\{-\frac{1}{2\sigma^2} [(n-1)s^2 + n(\bar{x} - \mu)^2]\right\} \quad (7)$$

$$\text{where } \bar{x} = \frac{1}{n} \sum_{k=1}^n x_k, s^2 = \frac{1}{n-1} \sum_{k=1}^n (x_k - \bar{x})^2.$$

Considering the positions of μ and σ^2 in the joint density, it is reasonable to choose the normal distribution as the prior distribution of μ and the inverse gamma distribution as the prior distribution of σ^2 , so the independent prior distribution of σ^2 and the conditional prior distribution of μ are:

$$\sigma^2 \sim \text{IGa}(v/2, v\rho^2/2), \mu|\sigma^2 \sim N(\tau, \sigma^2/\varphi) \quad (8)$$

This step easily leads to the joint prior density function of (μ, σ^2) :

$$p(\mu, \sigma^2) = \frac{1}{\sqrt{2\pi\sigma^2/\varphi}} \frac{(v\rho^2/2)^{v/2}}{\Gamma(v/2)} \left(\frac{1}{\sigma^2}\right)^{v/2+1} \exp\left\{-\frac{1}{2\sigma^2}[v\rho^2 + \varphi(\mu - \tau)^2]\right\} \quad (9)$$

where v, τ, ρ, φ are the hyperparameters, which are assumed to be given here first.

Multiplying a priori density by the joint a priori density immediately yields the kernel for a posteriori density.

$$p(\mu, \sigma^2 | X) \propto p(X | \mu, \sigma^2) p(\mu, \sigma^2) \propto \sigma^{-1} (\sigma^2)^{-(v+n)/2-1} \exp\left\{-\frac{1}{2\sigma^2}[v\rho^2 + \varphi(\mu - \tau)^2 + (n-1)s^2 + n(\bar{x} - \mu)^2]\right\} \quad (10)$$

To simplify Equation (10), first write:

$$\begin{aligned} \tau_n &= \frac{\delta\tau + n\bar{x}}{\varphi + n} \\ \varphi_n &= \varphi + n \\ v_n &= v + n \\ v_n\rho_n^2 &= v\rho^2 + (n-1)s^2 + \frac{\varphi n(\tau - \bar{x})^2}{\varphi + n} \end{aligned} \quad (11)$$

Substituting these equations back into the original Equation (10) yields:

$$p(\mu, \sigma^2 | X) \propto (\sigma^2)^{-\frac{v_n+1}{2}-1} \exp\left\{-\frac{1}{2\sigma^2}[v_n\rho_n^2 + \varphi_n(\mu - \tau_n)^2]\right\} \quad (12)$$

This posterior density is formally the same as the prior density (Equation (9)), which shows that the normal-inverse-gamma distribution is the joint conjugate prior distribution of the normal mean of μ and the normal variance of σ^2 , so the joint posterior distribution can also be decomposed into the product of the independent posterior distribution of σ^2 and the conditional posterior distribution of μ . Thus, the probability distributions of the mean and variance are obtained as follows:

$$\sigma^2 \sim IGa(v_n/2, v_n\rho_n^2/2), \mu | \sigma^2 \sim N(\tau_n, \sigma^2/\varphi_n) \quad (13)$$

where the independent posterior distribution of σ^2 is the inverse-gamma distribution, and the independent posterior distribution of mean μ is the t -distribution with degrees of freedom v_n , location parameter τ_n , and scale parameter $\rho_n/\sqrt{\varphi_n}$.

The hyperparameters given above were solved by using a priori moments. To eliminate the effect of randomness in the training process when training the prediction model using historical data, it is usually trained several times to evaluate the accuracy comprehensively. Assuming that the training is performed k times, the mean and variance of the i -th training residual set are (μ_i, σ_i^2) . Then, the k overall sets can be expressed as $\{(\mu_1, \sigma_1^2), (\mu_2, \sigma_2^2), \dots, (\mu_k, \sigma_k^2)\}$, and the estimated values of the hyperparameters are obtained using Equation (8) as follows:

$$\begin{cases} \bar{\sigma}^2 = \frac{\rho^2 v}{v-2} = \bar{\sigma}_z^2 \\ D(\sigma^2) = \frac{2}{v-4} (\bar{\sigma}^2)^2 = S_{\sigma_z^2}^{*2} \\ \tau = \bar{\mu}_z \\ D(\mu) = \bar{\sigma}^2/\varphi = S_{\mu z}^{*2} \end{cases} \begin{cases} \hat{v} = 4 + \frac{2(\bar{\sigma}_z^2)^2}{S_{\sigma_z^2}^{*2}} \\ \hat{\rho}^2 = \bar{\sigma}_z^2 - \frac{S_{\sigma_z^2}^{*2} \bar{\sigma}_z^2}{(\bar{\sigma}_z^2)^2 + 2S_{\sigma_z^2}^{*2}} \\ \hat{\tau} = \bar{\mu}_z \\ \hat{\varphi} = \frac{\bar{\sigma}_z^2}{S_{\mu z}^{*2}} \end{cases} \quad (14)$$

where $\bar{\sigma}_z^2 = \frac{1}{n} \sum_{k=1}^n \sigma_k^2$ is the mean of the variance sample set $\{\sigma_i^2\}$ and $S_{\sigma_z^2}^{*2} = \frac{1}{n-1} \sum_{k=1}^n (\sigma_k^2 - \bar{\sigma}_z^2)^2$ is the corrected variance of the variance sample set $\{\sigma_i^2\}$. $\bar{\mu}_z = \frac{1}{n} \sum_{k=1}^n \mu_k$ is the mean of the

mean sample set $\{\mu_i\}$, and $S_{\mu_z}^{*2} = \frac{1}{n-1} \sum_{k=1}^n (\mu_k - \bar{\mu}_z)^2$ is the variance of the mean sample set $\{\mu_i\}$. Substituting the estimated values into Equation (11), the estimated values of $v_n, \rho_n, \tau_n, \varphi_n$ can be obtained.

4.2.2. Mean and Variance Estimation

After obtaining the posterior probability distribution of the mean μ and variance σ^2 , the median, mode or expectation of the posterior distribution is taken as the parameter estimate, and the equation of the mode M_o and expectation E corresponding to the mean value of the t -distribution is:

$$M_o = E = \mu \quad (15)$$

The equations of the mode M_o and expectation E corresponding to the variance of the inverse-gamma distribution are:

$$M_o = \beta/(\alpha + 1), \quad E = \beta/(\alpha - 1), \alpha > 1 \quad (16)$$

The estimates of the mean μ and variance σ^2 were obtained from the mode and expectation equations of the inverse gamma and t -distributions:

$$\begin{cases} \hat{\sigma}_M^2 = \frac{v_n \rho_n^2}{v_n + 2} \\ \hat{\sigma}_E^2 = \frac{v_n \rho_n^2}{v_n - 2}, v_n > 2 \\ \hat{\mu}_M = \mu_E = \tau_n \end{cases} \quad (17)$$

where $\hat{\sigma}_M^2$ and $\hat{\mu}_M$ are the mode estimates, and $\hat{\sigma}_E^2$ and μ_E are the expectation estimates. From the equation, the t -distribution corresponding to the mean makes no difference in choosing the mode or the expectation as the mean estimate because its mode and expectation are equal. However, for the variance estimate, since the mode value is smaller than the expectation, if the mode is chosen as the expectation, it will make the evaluation strategy stricter and reduce the underdiagnosis rate, but at the same time, it may also identify signals that are in a healthy state as abnormal and increase the misdiagnosis rate. Meanwhile, the opposite is true for the choice of the expectation, so the actual situation of the corresponding data should be considered in the selection process for analysis.

4.2.3. Normality Test

In the above solution process, the residuals are assumed to obey a normal distribution. To ensure the rigor of the decision, this paper uses the Anderson–Darling (A-D) goodness-of-fit test to determine whether the residuals obey a normal distribution. For residuals $X = \{x_1, x_2, \dots, x_n\}$, a hypothesis test can be performed such that the null hypothesis H_0 : X obeys a normal distribution and the alternative hypothesis H_1 : X does not obey a normal distribution. First, the decision statistic A^2 is calculated [35]:

$$A^2 = -\frac{1}{n} \sum_{k=1}^n (2k-1) \left[\ln N\left(\frac{x_k - \bar{x}}{S}\right) + \ln \left(1 - N\left(\frac{x_{n+1-k} - \bar{x}}{S}\right)\right) \right] - n \quad (18)$$

where $N(x)$ is the standard normal distribution, $\bar{x}$ is the mean value of the residual, and S is the standard deviation of the residual. Using Equation (18) to find A^2 , the critical value A_α^2 is obtained by checking the table according to the given significance level α . If $A^2 < A_\alpha^2$, it is concluded that X obeys a normal distribution and vice versa. If it does not obey a normal distribution, the Box–Cox power transformation model [36] is used to transform the nonnormal data.

4.3. Quantification and Anomaly Identification

After finding the posterior probability of the hypothesis, the failure probability can be quantified by setting the failure probability as P and the failure probabilities of the mean and variance as P_μ and P_{σ^2} , respectively.

$$P = 1 - \int_{-|\hat{\mu}|}^{|\hat{\mu}|} g(\mu) d\mu \quad (19)$$

Similarly, the value of P_{σ^2} can be obtained, so the overall failure probability is:

$$P = \frac{P_\mu + P_{\sigma^2}}{2} \quad (20)$$

Next, a suitable strategy is developed to determine when anomalies occur. Traditional judgments are usually made as posterior probability ratios, e.g., for the mean value:

$$\lambda_\mu = \frac{\beta_0}{\beta_1} = \frac{\int_{-|\hat{\mu}|}^{|\hat{\mu}|} g(\mu) d\mu}{\int_{-\infty}^{-|\hat{\mu}|} g(\mu) d\mu + \int_{|\hat{\mu}|}^{\infty} g(\mu) d\mu} \quad (21)$$

when $\lambda_\mu > 1$, H_0 is accepted, i.e., the signal is considered currently healthy. On this basis, its logarithm $\log_{10} \lambda_\mu$ is taken as the Bayes factor π_μ to compress the data. H_0 is accepted when $\pi_\mu > 0$. Similarly, the Bayes factor π_{σ^2} for the variance is obtained.

Finally, a health value is set up to synthesize the judgment conditions:

$$\omega = \begin{cases} 0, & \pi_\mu > 0 \text{ and } \pi_{\sigma^2} > 0 \text{ and } P < 0.05 \\ 1, & \text{else} \end{cases} \quad (22)$$

when the health value ω is 0, this signal is considered abnormal, and when it is 1, it is healthy.

For multivariate data, after i virtual signals are obtained by processing, such as dimensionality reduction, the above processing can be performed for each virtual signal similarly to obtain the respective health value ω_i . Then, the judgment for the overall health value is modified as follows:

$$\omega = \begin{cases} 0, & \omega_i = 0 \\ 1, & \text{else} \end{cases} \quad (23)$$

The whole anomaly identification and early warning model is divided into three parts. The preprocessed data are predicted by LSTM and combined with real-time Bayesian anomaly recognition to achieve early warning, as shown in Figure 1.

5. Experimental Results and Discussion

Machinery operation data of two cases are selected to verify the accuracy and generalization of the model. The sampling frequency of data is different and both of them are multidimensional.

5.1. Water Pumping Station

The public pump_sensor_data dataset [37] on the Kaggle website is taken as an example to illustrate the realization process of the decision strategy. The dataset is the operation data of a large water pumping station with a large-town water supply. The pump is equipped with nearly sixty sensors to detect different parts and variables. The sampling frequency of the data is 1 h, and the selected time is from 26 May to 31 July, which includes both healthy and fault data. The running state of the entire dataset is shown in Figure 5. The status value is 1 when the pump station is faulty and 0 when the operation is normal. Each signal is governed by a simple code, and this example also demonstrates the universality of the model.

Figure 5. Water pumping station operation status.

The entire processing procedure of the dataset is shown in Figure 1. These data need to be analyzed in a multivariate situation. After the missing values are filled and outliers are processed on the exported data, the dimensionality is reduced to simplify the data information, and the contribution of each principal component (PC1, PC2, PC3) is shown in Figure 6. The first two principal components are selected as two independent virtual signals according to the principle that the contribution is greater than 80%, and then these virtual signals are individually subjected to wavelet packet noise reduction and normalization to obtain the preprocessed dataset, as shown in Figure 7. In Figure 7, the data are divided into training, validation, and prediction data. The training data are used to train the LSTM model, while the validation data are used to validate the prediction model and as a priori information for parameter estimation. From the figure, the two independent virtual signals show significant fluctuations in both time periods in the prediction dataset and are consistent with the operation of the pumping station in the real state, indicating that the Bayesian principal component analysis in this model retains most of the information of the original data. Moreover, the failure times in the prediction set are marked in detail, which will be used to verify the prediction accuracy of this model.

Figure 6. Contribution of principal components.

Figure 7. The signal value.

Using the data of the first principal component as an example, the data of the training and validation sets are used as the training data of the LSTM neural network. A phase space reconstruction is performed by trying different combinations of m and τ , and the prediction model is found to have the highest accuracy when $m=4$, $\tau=2$ with an MSE of 0.0054. Figure 8a. compares the actual value of the first principal component with the predicted value, and Figure 8b shows the predicted residuals. The residuals in the training and validation sets were intercepted and tested for normality, and the results are shown in Figure 9. The confidence level α was taken as 0.05, the mean value was -0.0248 , the variance was 0.0548, and the p-value was 0.092 (greater than 0.05), which means that the residuals were found to obey a normal distribution and the training model was valid.

Figure 8. LSTM prediction results.

Figure 9. Normality test.

The mean and variance estimates of health signal residuals and the hyperparameters were obtained from the validation set. The mean $\bar{x} = 0.00716$ and variance $s^2 = 0.002$ of the validation set residuals of the trained LSTM model. The prediction model was retrained six times, and the mean and variance of each time are shown in Table 1. The final estimated mean $\hat{\mu} = 0.006265$ and variance $\hat{\sigma}^2 = 0.002181$ were used as the criteria for the subsequent evaluation. The variation in the mean was within ± 0.004 of the estimate, and the variation in the residuals was within ± 0.0005 . The hyperparameters can then be solved according to Equations (14) and (20), as shown in Table 2.

Table 1. Mean and variance of the residuals for the 6 sets of training data.

Group	1	2	3	4	5	6	Average
μ	0.002156	0.004781	0.003737	0.009898	0.009301	0.007718	0.006265
σ^2	0.001640	0.002045	0.002020	0.002548	0.002603	0.002232	0.002181

Table 2. Hyperparameter values.

Parameter	v	v_n	ρ	ρ_n	τ	τ_n	δ	δ_n
value	4.74080	446.74080	−0.01032	0.04141	0.00351	0.00582	84.88947	526.88947

The signal is tested in real time based on the obtained residual estimates and the mean estimates. The current signal value and the previous 11 h of data are taken as a sliding window. The obtained testing result is shown in Figure 10, where (a) and (b) show the Bayes factor and the failure probability for the mean test, respectively, and (c) and (d) show the Bayes factor and the failure probability for the variance test, respectively. As seen from the Bayes factor diagram, the mean graph appeared negative at 4:00 on 28 June, and the variance graph also appeared negative at 3:00 on the same day. The corresponding failure probability of the two parameters also began to rise rapidly to 100% at this moment, indicating that the equipment failed. In addition, the Bayes factor value of the mean diagram dropped rapidly at 7:00 on 25 July, while the factor value of the variance diagram also dropped rapidly at that moment, and the corresponding failure probability climbed to 100% just as rapidly. Both times, the change in the signal before the failure was successfully identified, and the earliest alarm was issued 18 h earlier.

Figure 10. Bayes factor and probability of failure.

The overall failure probability and health values of the first principal component are obtained from Equations (20) and (22), as shown in Figure 11a,b. Similarly, the overall

failure probability and health values of the second principal component are obtained in Figure 11c,d. As seen from the figures, the trends in the first and second principal components are basically the same, so a comprehensive judgment by using Equation (23) is unnecessary in this example. However, in the comparison of the health value plots, the first principal component predicts the fault signal at an earlier time, and the failure probability value is always 100% during an actual fault of the pump station, while the second principal component has larger fluctuations, indicating that the first principal component contains most of the information of the signal and can predict the signal fault more accurately. Additionally, comparing the actual health values in Figure 5 shows that the proposed model accurately predicts the subsequent faults and is able to issue alerts in advance.

Figure 11. Overall failure probability and health values of the two principal component signals.

5.2. Nuclear Power Turbine

To verify the developed decision method in practical applications, a motor with faults of a turbine unit in a nuclear power plant is used as the research object, and the operation data from 00:00 on 25 November to 00:00 on 30 November are extracted from the monitoring system. The data consist of vibration signals on both sides of the driving end and the nondriving end of the motor, including both health data and fault data. The data are collected at a frequency of minutes. The four vibration signals are shown in Figure 12. According to the accident report from the nuclear power plant, the monitoring system found that the vibration of the motor's driving end was out of limit. The vibration value quickly rose to 2.48 mm/s, exceeding the alarm value by 2.3 mm/s, and then triggered the alarm at 11:26 on November 28.

Figure 12. Operation data of a nuclear power turbine.

The decision model is applied to the abnormal prediction of operation data. The data are preprocessed and reduced in dimensionality, and the contribution rate of the first principal component after dimensionality reduction is as high as 98%, indicating that there is a strong correlation between the four vibration signals. Therefore, the first principal component PCA1 data are only taken for further analysis. The data curve is consistent with the trend of the curve of the original data, as shown in Figure 13. Subsequently, according to the decision flow, PCA1 is imported into the LSTM model for training and prediction to obtain the residuals. After the Bayesian hypothesis testing, the failure probability is obtained in Figure 14. The failure probability is basically 0 in the first half of time, i.e., normal operation, and rises slightly to 0.0543 at 16:30 on the 17th. This trend may be caused by the sensitive algorithm. Then, at 19:51 on the 17th, the failure probability quickly climbed to 100 percent. Compared with PCA1 in Figure 13, although each signal in Figure 13 did not exceed the alarm value before the monitoring system alarm, there was a significant change at 19:59 on 17 November, and the fluctuation range of the subsequent data deviated from the normal value area. This anomaly was consistent with the failure probability graph in this time period, and this decision model predicted this anomaly in advance and accurately. According to this decision model, the alarm will be given before the occurrence of abnormality, and the actual results also prove that machinery failure will occur within 12 h after the abnormality is identified. However, the actual threshold alarm system of the nuclear power plant did not identify this anomaly, and the fault was not found until the vibration value reached the alarm value, resulting in alarm delay. At the same time, the failure probability reached 100% again at 9:59 on 28 November, corresponding to the alarm of the monitoring system at 11:26, which indicated that the decision model still predicted the fault 1.5 h in advance, even without considering the previous abnormality.

Figure 13. First principal component data after dimensionality reduction.

Figure 14. Failure probability diagram.

6. Conclusions

A multivariate abnormality identification and early warning model based on Bayesian hypothesis testing is proposed to address the shortcomings of traditional thresholds with many missed diagnoses and late alarms. The model predicts time-series data by an LSTM neural network and makes full use of historical data to dynamically determine future abnormal status in real time while proposing a processing method for multivariate cases. The proposed method has the following features: (1) based on data-driven methods, it makes full use of a priori information to improve the identification accuracy, (2) it can be effectively applied to abnormality identification in multivariate situations, and (3) it uses Bayesian hypothesis testing as the fault judgment method to quantify the failure probability, making the judgment results more rigorous and reliable and enabling earlier detection of potential faults. The proposed model is validated using monitoring data from pumping stations and actual failure cases in nuclear power plants. The experimental study shows that this model successfully predicts the occurrence of faults in the multivariate case over a long period of time, which validates the accuracy and reliability of the model.

Author Contributions: Conceptualization, G.L. and D.Y.; Methodology, D.Y. and S.Y.; Resources, G.L. and G.W.; Software, S.Y. and F.L.; Validation, G.W. and F.L.; Visualization, S.Y.; Writing—original draft, S.Y.; Writing—review & editing, D.Y. and G.L. All authors have read and agreed to the published version of the manuscript.

Funding: This research was funded by the National Natural Science Foundation of China, grant number 51875209, the Open Funds of State Key Laboratory of Nuclear Power Safety Monitoring Technology and Equipment, grant number K-A2020.408, the Science and Technology Planning Project of Guangdong Province, grant number 2021A0505030005, and the Guangdong Basic and Applied Basic Research Foundation, grant number 2019B1515120060.

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Dao, P.B.; Staszewski, W.J.; Barszcz, T.; Uhl, T. Condition monitoring and fault detection in wind turbines based on cointegration analysis of SCADA data. *Renew. Energy* **2018**, *116*, 107–122. [\[CrossRef\]](#)
2. Abdeljaber, O.; Sassi, S.; Avci, O.; Kiranyaz, S.; Ibrahim, A.A.; Gabbouj, M. Fault detection and severity identification of ball bearings by online condition monitoring. *IEEE Trans. Ind. Electron.* **2018**, *66*, 8136–8147. [\[CrossRef\]](#)
3. Ling, J.; Liu, G.J.; Li, J.L.; Shen, X.C.; You, D.D. Fault prediction method for nuclear power machinery based on Bayesian PPCA recurrent neural network model. *Nucl. Sci. Tech.* **2020**, *31*, 75. [\[CrossRef\]](#)
4. Tautz-Weinert, J.; Watson, S.J. Using SCADA data for wind turbine condition monitoring—A review. *IET Renew. Power Gener.* **2016**, *11*, 382–394. [\[CrossRef\]](#)
5. Saufi, S.R.; Ahmad, Z.A.B.; Leong, M.S.; Lim, M.H. Challenges and opportunities of deep learning models for machinery fault detection and diagnosis: A review. *IEEE Access* **2019**, *7*, 122644–122662. [\[CrossRef\]](#)
6. Zhao, R.; Yan, R.; Chen, Z.; Mao, K.; Wang, P.; Gao, R.X. Deep learning and its applications to machine health monitoring. *Mech. Syst. Signal Process.* **2019**, *115*, 213–237. [\[CrossRef\]](#)
7. Liu, W.; Guo, P.; Ye, L. A low-delay lightweight recurrent neural network (LLRNN) for rotating machinery fault diagnosis. *Sensors* **2019**, *19*, 3109. [\[CrossRef\]](#)
8. Hochreiter, S.; Schmidhuber, J. Long short-term memory. *Neural Comput.* **1997**, *9*, 1735–1780. [\[CrossRef\]](#)
9. Tian, H.X.; Ren, D.X.; Li, K. A hybrid vibration signal prediction model using autocorrelation local characteristic-scale decomposition and improved long short term memory. *IEEE Access* **2019**, *7*, 60995–61007. [\[CrossRef\]](#)
10. Ma, X.; Tao, Z.; Wang, Y.; Yu, H.; Wang, Y. Long short-term memory neural network for traffic speed prediction using remote microwave sensor data. *Transp. Res. Part C Emerg. Technol.* **2015**, *54*, 187–197. [\[CrossRef\]](#)
11. Kim, H.Y.; Won, C.H. Forecasting the volatility of stock price index: A hybrid model integrating LSTM with multiple GARCH-type models. *Expert Syst. Appl.* **2018**, *103*, 25–37. [\[CrossRef\]](#)
12. Di Persio, L.; Honchar, O. Bayesian approach to energy load forecast with neural networks. In *Handbook of Energy Finance: Theories, Practices and Simulations*; World Scientific Publishing Company: Singapore, 2020; pp. 73–92.
13. He, J.; Yang, S.; Gan, C. Unsupervised fault diagnosis of a gear transmission chain using a deep belief network. *Sensors* **2017**, *17*, 1564. [\[CrossRef\]](#)
14. Zeng, Z.; Tan, W.; Zhou, R. Performance assessment for generalized delay-timers in alarm configuration. *J. Process Control* **2017**, *57*, 80–101. [\[CrossRef\]](#)
15. Zadakbar, O.; Imtiaz, S.; Khan, F. Dynamic risk assessment and fault detection using principal component analysis. *Ind. Eng. Chem. Res.* **2013**, *52*, 809–816. [\[CrossRef\]](#)
16. Yu, Y.; Wang, J.; Yang, Z. Design of alarm trippoints for univariate analog process variables based on alarm probability plots. *IEEE Trans. Ind. Electron.* **2017**, *64*, 6496–6505. [\[CrossRef\]](#)
17. Gao, H.; Liu, F.; Zhu, Q. A correlation consistency based multivariate alarm thresholds optimization approach. *ISA Trans.* **2016**, *65*, 37–43. [\[CrossRef\]](#)
18. Zhao, H.; Liu, H.; Hu, W.; Yan, X. Anomaly detection and fault analysis of wind turbine components based on deep learning network. *Renew. Energy* **2018**, *127*, 825–834. [\[CrossRef\]](#)
19. Aslansefat, K.; Gogani, M.B.; Kabir, S.; Shoorehdeli, M.A.; Yari, M. Performance evaluation and design for variable threshold alarm systems through semi-Markov process. *ISA Trans.* **2020**, *97*, 282–295. [\[CrossRef\]](#)
20. Zhang, G.; Wang, Z.; Mei, H. Sensitivity clustering and ROC curve based alarm threshold optimization. *Process Saf. Environ. Prot.* **2020**, *141*, 83–94. [\[CrossRef\]](#)
21. Asr, M.Y.; Ettetfagh, M.M.; Hassannejad, R.; Razavi, S.N. Diagnosis of combined faults in Rotary Machinery by Non-Naive Bayesian approach. *Mech. Syst. Signal Process.* **2017**, *85*, 56–70. [\[CrossRef\]](#)
22. Ayadi, A.; Ghorbel, O.; Bensaleh, M.S.; Obeid, A.; Abid, M. Data classification in water pipeline based on wireless sensors networks. In Proceedings of the IEEE/ACS International Conference on Computer Systems and Applications AICCSA, Hammamet, Tunisia, 30 October–3 November 2018; pp. 1212–1217.
23. Khalid, A.J.; Wang, J.; Nurudeen, M. A new fault classification model for prognosis and diagnosis in CNC machine. In Proceedings of the 2013 25th Chinese Control and Decision Conference CCDC, Guiyang, China, 25–27 May 2013; pp. 3538–3543.
24. Lu, Y.; Wang, Z.; Xie, R.; Zhang, J.; Pan, Z.; Liang, S.Y. Bayesian optimized deep convolutional network for bearing diagnosis. *Int. J. Adv. Manuf. Technol.* **2020**, *108*, 313–322. [\[CrossRef\]](#)
25. Li, C.; Ledo, L.; Delgado, M.; Cerrada, M.; Pacheco, F.; Cabrera, D.; Sánchez, R.V.; de Oliveira, J.V. A Bayesian approach to consequent parameter estimation in probabilistic fuzzy systems and its application to bearing fault classification. *Knowl. Based Syst.* **2017**, *129*, 39–60. [\[CrossRef\]](#)

26. Cho, H.C.; Knowles, J.; Fadali, M.S.; Lee, K.S. Fault detection and isolation of induction motors using recurrent neural networks and dynamic Bayesian modeling. *IEEE Trans. Control Syst. Technol.* **2009**, *18*, 430–437. [[CrossRef](#)]
27. Nezhad, M.S.F.; Niaki, S.T.A. A heuristic threshold policy for fault detection and diagnosis in multivariate statistical quality control environments. *Int. J. Adv. Manuf. Technol.* **2013**, *67*, 1231–1243. [[CrossRef](#)]
28. Chen, X.; Ge, Z. Switching LDS-based approach for process fault detection and classification. *Chemom. Intell. Lab. Syst.* **2015**, *146*, 169–178. [[CrossRef](#)]
29. Jiang, X.; Mahadevan, S.; Yuan, Y. Fuzzy stochastic neural network model for structural system identification. *Mech. Syst. Signal Process.* **2017**, *82*, 394–411. [[CrossRef](#)]
30. El-Sheimy, N.; Nassar, S.; Noureldin, A. Wavelet de-noising for IMU alignment. *IEEE Aerosp. Electron. Syst. Mag.* **2004**, *19*, 32–39. [[CrossRef](#)]
31. Liu, G.; Gu, H.; Shen, X.; You, D. Bayesian long short-term memory model for fault early warning of nuclear power turbine. *IEEE Access* **2020**, *8*, 50801–50813. [[CrossRef](#)]
32. Tipping, M.E.; Bishop, C.M. Probabilistic principal component analysis. *J. R. Stat. Soc. Ser. B Stat. Methodol.* **1999**, *61*, 611–622. [[CrossRef](#)]
33. Kim, H.S.; Eykholt, R.; Salas, J.D. Nonlinear dynamics, delay times, and embedding windows. *Phys. D Nonlinear Phenom.* **1999**, *127*, 48–60. [[CrossRef](#)]
34. Berger, J.O.; Mortera, J. Default bayes factors for nonnested hypothesis testing. *J. Am. Stat. Assoc.* **1999**, *94*, 542–554. [[CrossRef](#)]
35. Anderson, T.W.; Darling, D.A. A test of goodness of fit. *J. Am. Stat. Assoc.* **1954**, *49*, 765–769. [[CrossRef](#)]
36. Box, G.E.P.; Cox, D.R. An analysis of transformations. *J. R. Stat. Soc. Ser. B Methodol.* **1964**, *26*, 211–243. [[CrossRef](#)]
37. Kaggle. Pump-Sensor-Data. 2019. Available online: <https://www.kaggle.com/nphantawee/pump-sensor-data> (accessed on 30 March 2021).