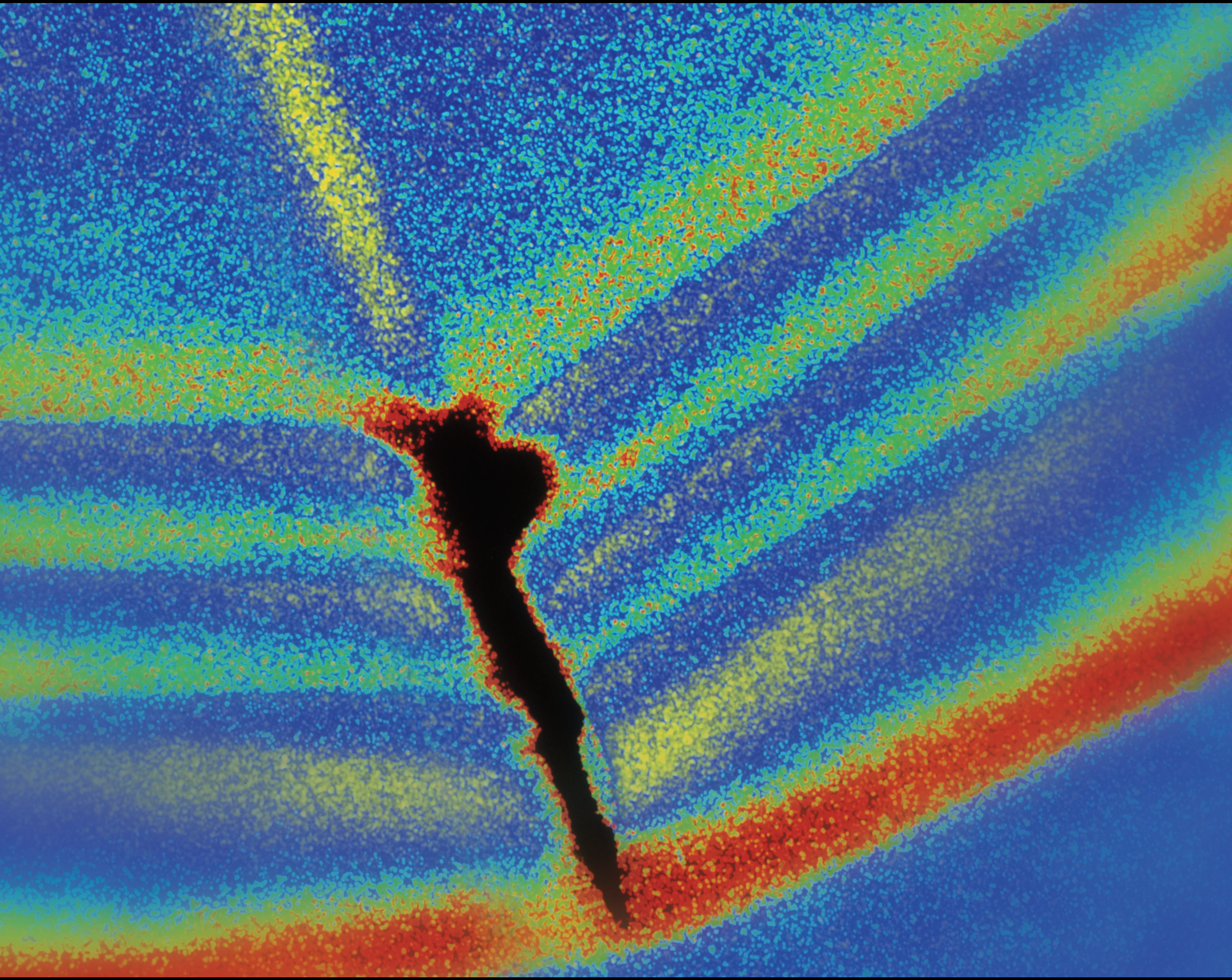
Shock and Vibration

Advances in Vibration Signal Time-Frequency Analysis for Defect Detection

Lead Guest Editor: Shengwei Fei
Guest Editors: Feng Lu, Chaoqun Duan, and Abolfazl Gharaei



Advances in Vibration Signal Time-Frequency Analysis for Defect Detection

Shock and Vibration

Advances in Vibration Signal Time-Frequency Analysis for Defect Detection

Lead Guest Editor: Shengwei Fei
Guest Editors: Feng Lu, Chaoqun Duan, and Abolfazl Gharaei

Chief Editor

Associate Editors

Academic Editors

Contents

Hindawi
Shock and Vibration
Volume 2020, Article ID 8810047, 9 pages
https://doi.org/10.1155/2020/8810047

Research Article

Estimation and Control for Industrial Products Reliability under Failure Trials

Kui Wang[1] **and Lili Ding**[2]

[1] *School of Mechanical Science and Engineering, Huazhong University of Science and Technology, Wuhan 430074, China*
[2] *Yichang Meteorological Bureau, Yichang 443000, China*

Correspondence should be addressed to Lili Ding; dinglili1986@126.com

Received 29 September 2020; Revised 28 October 2020; Accepted 17 November 2020; Published 8 December 2020

Academic Editor: Chaoqun Duan

In practices, most industrial products are subject to sudden failure and only failure information can be collected, which presents a great challenge for reliability prediction of modern devices. To address this issue, our paper proposes a dynamic reliability estimation and control for industrial products under regular failure trials. The failure trial is performed at different operational time points of the products, which provides sole data source for evaluating the status of industrial products. We use Bayesian approach to dynamically estimate the industrial products when the failure trial is available. The estimated reliability is updated using a point estimate with new available data. To maintain the reliability of products at a desirable status, a reliability control method is presented to monitor the confidence interval of reliability distribution. The lower limit of confidence interval is maintained above a control limit, which indicates that a corresponding quality-assurance action is preferable. The proposed reliability estimation and control approach is demonstrated using a case of light-emitting diodes under failure trials at production process. The obtained results indicate the effectiveness of our estimation and control model.

1. Introduction

With development of industrial engineering, small factories are becoming large-scale plants, which prompts the applications of prognostics [1]. For prognostics of the actual status of industrial products, reliability estimation is a crucial step, which provides risks information for the maintenance engineers [2, 3]. The reliability of industrial products is the ability to complete their specified functions under specified time [4]. This index is also the prerequisite of implementation of the condition-based maintenance. To eliminate the failure occurrences of the industrial products, the reliability estimation is required to ensure the production quality in most industries [5].

Most reliability estimation works were mainly focused on the prediction stage; however, there was no practical or useful guidance for the real applications [2, 6]. When we obtained estimated reliability level of industrial products, we rarely have useful policy to improve the reliability during the production process, where the reliability control approach is very much required in quality control process [7]. Therefore, this paper aims to propose a reliability estimation and control approach to address this problem existing in modern industries.

The consciousness of reliability applications has been noticeably improved in the recent years [8–10]. The reliability evaluation system can ensure many critical and heavy-duty machines with a relatively safe operating conditions, and it has become an indispensable part of modern industry [11]. This system is also widely used in manufacturing engineering, aviation, and nuclear industries [11, 12]. To properly develop an effective reliability evaluation system for industrial products, the reliability prediction and control techniques should be explored and studied. Currently, there are some approaches related to reliability evaluation, which are the statistics approaches for reliability modeling including reliability block diagrams, degradation-based reliability estimation, and hazard-based analysis [13–15]. These existing works often modeled degradation path or hazard pattern as fixed model and used some ideal assumptions to compute the reliability of industrial products. Besides, the distribution of reliability is not provided in these mentioned research works. In a classical reliability framework, it is often difficult or impossible to apply single-unit estimation model for a complex system consisting of multiple units [16]. The reliability distribution and mean reliability are completely different for these two systems. This hinders the application of current estimation methods on complex systems. Furthermore, the parameters evaluated are unknown constants and they are evaluated only based on large amount of recorded data sample [17]. In some engineering practices, however, there could be no perfect sources of reliability data. For example, testing of rare and valuable products such as missiles and satellite systems, the sample size will be small or even missing [18]. To achieve an accurate reliability estimation from these data sources through classical statistics approaches can be a great challenge.

To overcome the above challenge, researchers have employed Bayesian statistical techniques for reliability estimation. The Bayesian statistical techniques use an updating process to renew the operating reliability of industrial products, and when the new observation data are collected, more accurate estimation is provided. The Bayesian reliability estimation technique is especially useful for system with small size of data sample [11]. Cole [19] proposed a reliability assessment method for industrial systems. In Cole's work, Bayes' theorem is used to determine posterior distribution of system reliability and update system condition level. Mastran et al. [20] proposed a reliability assessment method for coherent structure of industrial products. Failure data from structure, components, and system are used for updating the instant reliability. Sharma and Bhutani [21] studied the Bayesian reliability analysis of the parallel industrial system, and the prior information of the failure rate of the products is used for reliability assessment. Bao et al. [22] used the processing reliability data to analyze the reliability of high flux engineering test reactor. The data obtained from similar research reactors are selected as the Bayesian prior distribution to compute the Bayesian posterior distribution of the reliability. Gardner et al. [23] presented a Bayesian history matching technique in a structural dynamics context to infer the model discrepancy. The Gaussian process regression is utilized to map the simulator output and for training observation data. For summarizing the existing Bayesian statistical techniques, Insua et al. [11] have reviewed the existing Bayesian approach in reliability decision-making. They illustrated how the Bayesian approach is applied in life testing, experimental design, reliability certification, and preventive maintenance areas.

To achieve more accurate prediction, Peng et al. [24] proposed a generalized multivariate hybrid degradation model to incorporate and analyze the dynamic degradation. A two-step Bayesian framework is presented for parameter estimation. Moreover, Mahadevan et al. [25] presented a method of artificial intelligence called Bayesian network, which has been an effective tool for modeling uncertainty in reliability analysis and is commonly used in reliability analysis due to its powerful model structure [26]. Boudali et al. [27] have studied an improved reliability analysis method based on Bayesian network to model the dynamic degradation process of industrial systems. In this research, a discrete-time Bayesian network framework has been defined and proved to be a powerful tool for modeling and analyzing the behavior and interaction of various system components. Lee and Pan [28] employed the nonparametric Bayesian network, which is suitable to model any type of continuous and discrete random variables and overcomes the limitation provided by the discrete Bayesian network. Also, a Bayesian method combining the prior knowledge with expert opinions was proposed to evaluate the system reliability at the early design stage and is proved to be effective in a real case study. As an extension, Bayesian network has also been applied in the mechanical product quality improvement [29]. A Bayesian principle-empirical model fusing the principal knowledge and empirical data is presented to discover the relations of quality characteristics. Later, Cai et al. [30] have concluded the Bayesian network applications in reliability evaluation. This work mainly focused on the reliability modeling procedures with Bayesian network and provided guide for practitioners in practical implementations. More recently, Rebello et al. [31] combined a hidden Markov process and dynamic Bayesian network to present a hybrid reliability prediction method to overcome the difficulty in discretization and multiple variables representation. In this method, process data can be used in continuous time domain without discretization. These Bayesian approaches used in reliability engineering are verified to be an effective tool to use paucity of data in reliability prediction under ideal assumptions.

To fully use limited testing data in the production process, in this paper, we use a Bayesian procedure to dynamically estimate the reliability of the industrial products and present a reliability control method to maintain the lower limit of prediction interval above a certain level. This is the first paper to provide reliability estimation and control approach using Bayesian procedures. The main contributions of this work are as follows: (1) Interpretation of the past and present failure trials using inconstant general distribution, (2) development of a dynamic estimation procedure for updating products reliability, (3) development of a novel reliability control scheme based on confidence interval, and (4) validation of the improved performance using a real case study.

The remaining parts of this paper are organized as follows. In Section 2, industrial products failure trial is described. In Section 3, a reliability estimation model using Bayesian approach is presented for industrial products. In Section 4, a reliability control method is developed to maintain the reliability level of the products. In Section 5, a case of light-emitting diodes under failure trials at production process is illustrated to show the effectiveness of the proposed model. Finally, conclusions and future work directions are provided in the last section.

2. Industrial Products Failure Trials

The industrial products are subject to regular failure inspection throughout the production process. The failure inspection is performed at discrete time points to reveal whether the examined product can be functional. The observed failure rate at different time point is a crucial index for quality assurance in production department of industry. This is a common practice for most modern industrial manufacturers, where the failure trials at regular time epochs provide important information for the reliability estimation and control of the industrial products. Figure 1 shows the flowchart of the failure trial in industry. The new manufactured products are inspected at different time with trial 1 to trial n to examine the qualification rate of the product. The aim of the trial is to know the quality of the new products at each service time.

The failure trial samples a constant number of products and examines the quality of the products at each trial. Through the failure trial, the number of qualified products is obtained, which provides the information for maintenance engineers to estimate the reliability of the batch of products. Figure 2 shows the results of failure trials at each trial epoch. It can be observed that the qualified products at each trial present a shock tend, which signifies that the quality of new products is not stable and strict control should be applied.

We assume that, for a given trial, the overall number of tested products is S, and the qualified number of the tested products is q. For that specified configuration, the probability that such a test result is obtained is conditioned on previous configuration x:

$$P(S,q|X=x)=x^{q}\cdot(1-x)^{S-q}, \tag{1}$$

where X denotes the probability that a product is successful on a given trial.

To make full use of the failure trials data, we consider the Bayesian approach to use the discrete trial information to dynamically update the reliability of the tested products, and the procedure is presented in the next section.

3. Reliability Estimation

For a given trial, the Bayesian approach can infer an approximate value of the product reliability. As the trial continues, the value is renewed, and more accurate reliability estimation can be obtained. For Bayesian approach, the typical Bayes' theorem can be expressed as

$$P(A_i|B)=\frac{P(B|A_i)P(A_i)}{\sum_{\text{all } j}P(B|A_j)P(A_j)}, \tag{2}$$

where P indicates the probability of , and B is any trial with positive probability. For the case where X and Y are both continuous random variables, equation (2) can be equivalent to

$$f_{X|Y}(x|y)=\frac{f_{Y|X}(y|x)f_X(x)}{\int_{-\infty}^{\infty}f_{Y|X}(y|x)f_X(x)\mathrm{d}x}, \tag{3}$$

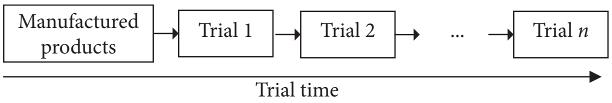


Figure 1: Flowchart of failure trial process in industry.

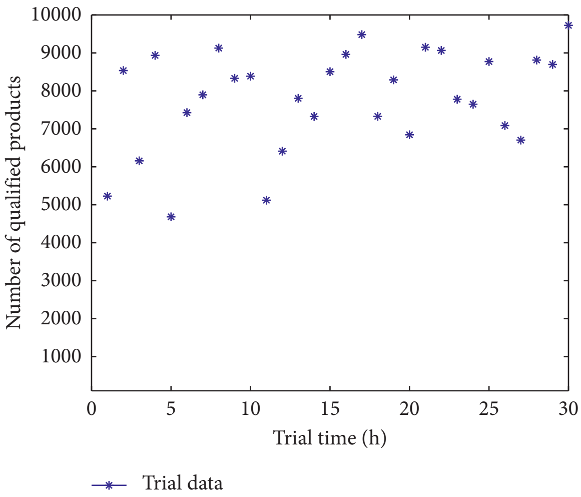


Figure 2: Failure trial of the industrial products.

where f_X represents the probability density function of the random variable X and $f_{X|Y}$ is a conditional probability density function, which is conditioned by a value of the random variable Y.

If X is continuous and Y is discrete, equation (2) can be written as

$$f_{X|Y}(x|y)=\frac{P(Y=y|X=x)f_X(x)}{\int_{-\infty}^{\infty}P(Y=y|X=x)f_X(x)\mathrm{d}x}. \tag{4}$$

For the failure trial of industrial products, $f_X(x)$ in equation (4) represents the density of the random variable of X before the next trial data are collected, and it is called the a priori density of X. $P(Y=y|X=x)$ is the probability of specified trial data given that the random variable has the value x. $f_{X|Y}$ is the a posteriori probability density of X given Y.

In the application of Bayes' theorem to reliability, the probability that, for a given trial, a product is qualified for manufacturing is denoted by R. R is treated as a random variable, since the product reliability is not constant and will be decreased with service time.

We suppose that the product reliability is a continuous random variable between 0 and 1. For overall S tested products, the reliability of the batch of product can be inferred as

$$f_{R|r_S}(r|q;S)=\frac{r^{q}(1-r)^{S-q}f_R(r)}{\int_0^1\beta^{q}(1-\beta)^{S-q}f_R(\beta)\mathrm{d}\beta}, \tag{5}$$

where r for $0\le r\le 1$ represents the value of estimated reliability. r_S represents the qualified number of products in all S tested products. f_R is the a priori density of the random variable R. Recalling equation (1), to compute with the

Bayesian approach, $f_R(r)$ can be considered as a random distribution with general form, such as Beta distribution and uniform distribution. Since the support interval of Beta distribution lies in [0, 1], it should be a natural candidate to model $f_R(r)$. Then $f_R(r)$ can be of the following form:

$$f_R(r) = \frac{r^k(1-r)^l}{\int_0^1 \beta^k(1-\beta)^l \mathrm{d}\beta}, \tag{6}$$

where k and l are the positive integers. The denominator of the above equation can be written in the following form:

$$\begin{aligned}\int_0^1 \beta^k(1-\beta)^l \mathrm{d}\beta &= \frac{\Gamma(k+l+2)}{\Gamma(k+1)\Gamma(l+1)},\\ &= B(k+1, l+1),\\ &= \frac{k!l!}{(k+l+1)!}.\end{aligned} \tag{7}$$

Therefore, the reliability conditioned on the available trial data can be derived using Bayes' theorem, which is given by

$$f_{R|r_S}(r|q;S) = \left(\frac{(S+k+l+1)!}{(q+k)(S+l-q)}\right) r^{q+k}(1-r)^{S-q+l}. \tag{8}$$

The reliability of the industrial products is obtained using equation (8), which indicates the estimated reliability distribution. To make the Bayesian updates less complicated, we use point estimates to represent the renewed mean reliability level of each trial. The point estimates use the expected value of the distribution of Bayesian a posteriori probability. In the following, we present a point estimation approach to simplify the Bayesian updating process.

The representative of the Bayesian reliability is considered here, and using differential calculus, equation (8) has the point estimate given by

$$\frac{k+q}{k+l+S}. \tag{9}$$

The above expression represents the updated mean reliability level at each sampling time epoch. When the a priori density is chosen as uniform, that is, $k = l = 0$, the estimate is simply (q/S). The uniform a priori distribution is an ideal situation that may not happen in reality. Thus, a second point estimate with more practical characteristics is presented, which is called the Neyman estimate [32]. The mean of the a posteriori density derived from equation (8) is given by

$$\frac{k+q+1}{k+l+S+2}. \tag{10}$$

In this instance, equation (10) is commonly referred to as the Bayesian estimate of the product reliability. We use equation (10) to calculate the reliability of industrial products at each failure trial.

4. Reliability Control

When we have obtained the expression of Bayesian estimate for the product reliability, we aim to maintain the satisfied level of product reliability. To maintain the reliability of product above a certain level, we propose a method to find optimal range of reliability distribution. The prediction interval of reliability distribution indicates the cumulative survival probability of the products. The method controls the cumulative survival probability at each failure trial to ensure that the risk of hazard is below a certain level α. Since the upper limit of range of reliability distribution is 1, we need to monitor the lower limit of the reliability distribution range, which ensures the cumulative survival probability above $1-\alpha$. When the lower limit of reliability distribution falls out of the range, the method alarms to indicate that a corresponding quality-assurance action is required. Using a Bayes technique, the two-sided $100(1-\alpha)$-precent confidence limits on R are defined as

$$P\left(R_{\text{Lower}} \le R \le R_{\text{Upper}}\right) = \int_{R_{\text{Lower}}}^{R_{\text{Upper}}} f_{R|r_S}(r|q;S)\mathrm{d}r = 1-\alpha. \tag{11}$$

subject to $R_{\text{Upper}} - R_{\text{Lower}}$ being a minimum, where R_{Upper} denotes the upper limit of estimate distribution of reliability and R_{Lower} denotes the lower limit of the distribution.

In reliability engineering, the high-quality product requires that the upper limit of reliability be equal to one, which means that the probability of survival product at certain time might be one; that is, $R_{\text{Upper}} = 1$. The lower one-sided limit is given by

$$P\left(R_{\text{Lower}} \le R\right) = \int_{R_{\text{Lower}}}^{1} f_{R|r_S}(r|q;S)\mathrm{d}r = 1-\alpha. \tag{12}$$

Referring to equation (8), we consider the a priori density of R as a uniform [0, 1]; that is, $k = l = 0$. Equation (12) can be derived as

$$P\left(R_{\text{Lower}} \le R\right) = \int_{R_{\text{Lower}}}^{1} \frac{r^q(1-r)^{S-q}}{B(q+1, S-q+1)}\mathrm{d}r = 1-\alpha. \tag{13}$$

The integral in equation (13) is the incomplete Beta function $I_{R_{\text{Lower}}}(q+1, S-q+1)$, and the values of incomplete Beta function are presented in [33]. For any $I_R(q+1, S-q+1)$, the expression can be further extended as

$$I_R(q+1, S-q+1) = \sum_{x=q+1}^{S+1} \binom{S+1}{x} R^X(1-R)^{S-X+1}. \tag{14}$$

Therefore, for the 100α-percent lower limit for R, R_{Lower} is given by the root of the following expression:

$$\sum_{x=q+1}^{S+1} \binom{S+1}{x} R^X(1-R)^{S-X+1} = \alpha. \tag{15}$$

At this point, we can use the existing table for incomplete Beta function to give limits on $1 - R$ rather than R. Using a similar method, the expression for the 100α-percent upper limit on $U = 1 - R$ can be obtained by the root of

$$\sum_{y=S-q+1}^{S+1} \binom{S+1}{y} U^{y} (1-U)^{S-y+1} = 1-\alpha, \quad (16)$$

where $y = S - q + 1$. To further derive the confidence limit of U, we use Neyman interval [29], and the 100α-percent upper limit of confidence can be computed using the root of the following equation:

$$\sum_{y=S-q+1}^{S} \binom{S}{y} U^{y} (1-U)^{S-y} = 1-\alpha. \quad (17)$$

Contrary to equation (16), it is shown that the Neyman interval with $S - q$ failures (unqualified products) in $S + 1$ trials is the same as Bayes estimate interval with assumptions of a uniform a priori density on R under $S - q$ failures in S trials. In the case that the a priori density of R is chosen as a Beta density, we have

$$f_R(r) = \frac{r^{k}(1-r)^{l}}{B(k+1, l+1)}. \quad (18)$$

For the lower 100α-percent confidence limit on R, R_{Lower} is defined as

$$\begin{gathered}\int_{R_{\text{Lower}}}^{1} \frac{r^{q+k}(1-r)^{S-q+l}}{B(q+k+1, S-q+l+1)} \mathrm{d}r = 1-\alpha, \\ 1 - I_{R_{\text{Lower}}}(q+k+1, S-q+l+1) = 1-\alpha.\end{gathered} \quad (19)$$

It is very clear that, with simple computation, the 100α-percent upper limit on $U = 1 - R$ can be determined by the solution of the following equation:

$$\sum_{y=S-q+l+1}^{S+k+l+1} \binom{S+k+l+1}{y} U^{y} (1-U)^{S-y+k+l+1} = 1-\alpha. \quad (20)$$

Now we have obtained the range of confidence interval. To maintain the reliability level of the industrial products, we set $\overline{R}$ as control limit of the reliability. When the lower limit of reliability distribution decreases to be below R_{Lower}, that is, $\overline{R} < R_{\text{Lower}}$, a corresponding quality assurance is required. The determination of $\overline{R}$ can be achieved using an algorithm presented in [34].

The confidence interval of reliability on a given operational time is updated at every trial, and the lower limit of the confidence interval should be maintained above a certain level $\overline{R}$, which guarantees the quality of industrial production process. In the next section, we use a case to illustrate how to estimate the product reliability and control the reliability level in the production process.

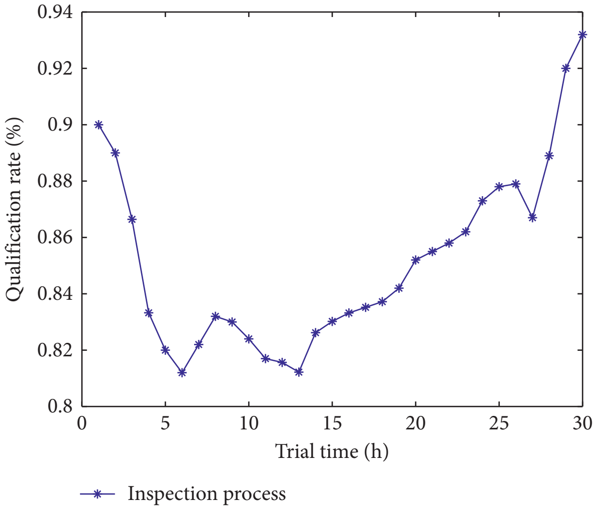


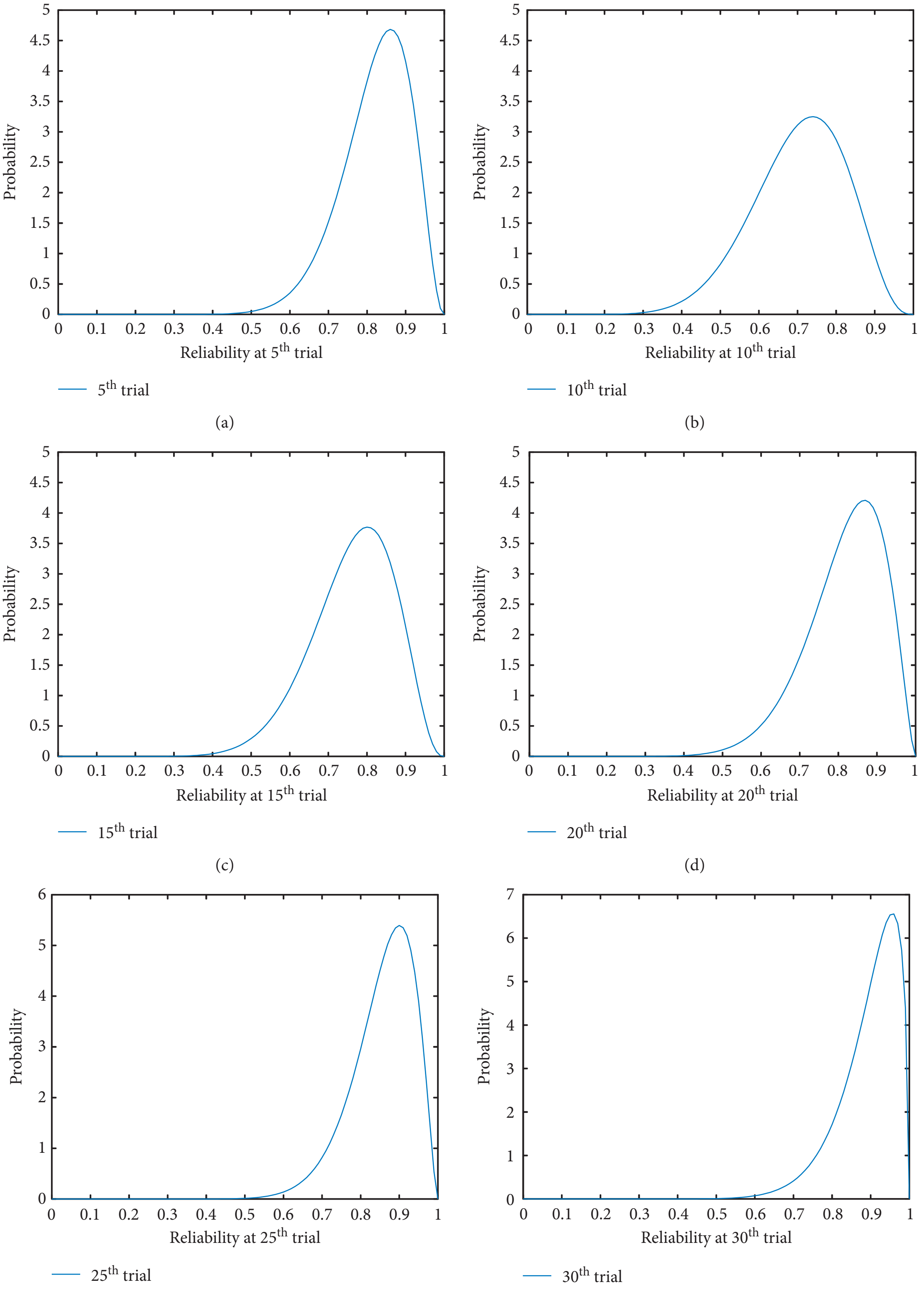
Figure 3: Qualification rate of each trial.

5. Case Study

5.1. Case Preprocessing. To illustrate the proposed approach, a case of light-emitting diodes was chosen to test the performance of our model. The failure trial (i.e., qualification test) was performed every 20 hours after the light-emitting diodes were produced. The failure trial is to ensure that the quality of the light-emitting diodes is maintained at a desired level. When the reliability of the light-emitting diodes derived from failure trial drops to a danger level, quality assurance action is initiated to check and guarantee the qualification rate of the production process. In each trial, the quality engineer took a sample size of products, and the sample size was not fixed for each trial. The light-emitting diodes were tested with result of “qualify” or “unqualify/failure.” The qualification rate (i.e., (q/S)) of each trial is computed and indicated in Figure 3. In Figure 3, the failure trial is performed 30 times, up to 600 testing hours. We use the data presented in Figure 3 to estimate the product reliability in each trial.

We consider the initial value of Bayesian estimate as the qualification rate of 1st failure trial, which is set to 0.902. The upper limit of reliability prediction range is considered as 1. The reliability of light-emitting diodes at each trial is updated using (9) in Section 3.

5.2. Reliability Prediction. Using the approach presented in Section 3, the reliability of light-emitting diodes can be obtained by equations (8)-(9). The estimated reliabilities at each trial are illustrated in Figure 4. In the first row of Figure 4, the mean reliability of the light-emitting diodes at the 5th trial is 0.816, which has been decreasing since the 1st trial. At the 10th trial, the distribution of reliability range is wider and scatter, which means that the quality of light-emitting diodes is not stable at this stage. In the second row of Figure 4, the reliability of light-emitting diodes has been slightly improved and increased to 0.9 at



(a)

(b)

(c)

(d)

(e)

(f)

FIGURE 4: Reliability prediction at each trial. (a) Reliability estimated at the 5th trial. (b) Reliability estimated at the 10th trial. (c) Reliability estimated at the 15th trial. (d) Reliability estimated at the 20th trial. (e) Reliability estimated at the 25th trial. (f) Reliability estimated at the 30th trial.

the 20th trial. This signifies that the failure rate of the products has been decreasing with the operational time. In the last row of Figure 4, the reliability of light-emitting diodes is increased noticeably, which is close to a value of 0.96. This means that the light-emitting diodes achieve a very satisfactory level of quality through the testing process.

Figure 5 indicates the estimated reliability throughout the trial. It is observed that the estimated reliability decreases at the first five trials and reaches the lowest level at the 5th trial. During the early stage of the testing process, that is, from the 4th trial to the 10th trial, the reliability of light-emitting diodes is at relatively low level, and this situation has been changing since the 19th trial. The predicted reliability of light-emitting diodes eventually remains at a desirable level at the end of testing process.

This situation indicates that the new produced products are not stable regrading the operating quality at the beginning of service time and gradually become stable with a desirable level after some operational time. The result suggests that the industrial factories should spend this unstable time in production before the products come into service.

5.3. Reliability Control.

5.3. Reliability Control. Through the approach presented in Section 4, the 90% confidence intervals for each trial are calculated (Figure 6). The upper limit of the prediction interval is 1 and the lower limit is plotted in Figure 6. We can see that the lower limit of the estimated reliability has the same evolution pattern as the mean reliability. In our reliability control approach, the control limit of lower limit of prediction interval is computed as 0.758.

Figure 7 shows the reliability control process of the light-emitting diodes. In the figure, the control limit of reliability prediction interval is between 0.758 and 1. The lower limit of the light-emitting diodes drops below the control limit within the 4th trial to the 6th trial, which indicates that a corresponding reliability control is required during the industrial production. After the 7th trial, the lower limit of 90% interval always remains above the control limit.

Through the above investigations, we have the two following conclusions: (1) The reliability of light-emitting diodes at early stage of production is not stable and will be stable with satisfactory level after some operational time, which signifies that the quality department should avoid this stage before the products come into service. (2) The reliability control method can maintain the distribution of light-emitting diodes reliability above a certain level, which helps to improve the production quality of the light-emitting diodes.

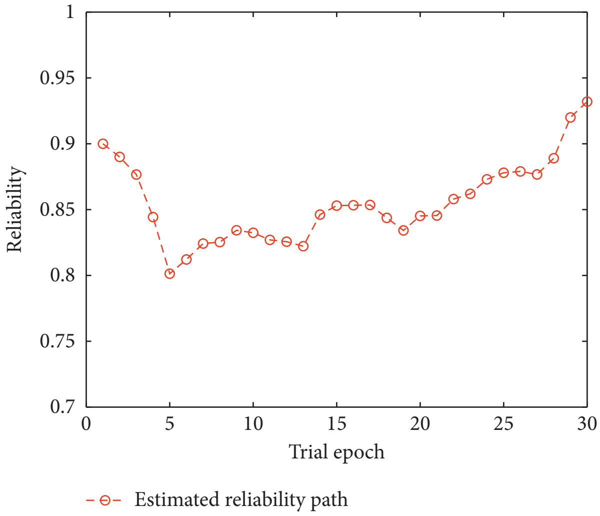


Figure 5: Estimated reliability path throughout the trial.

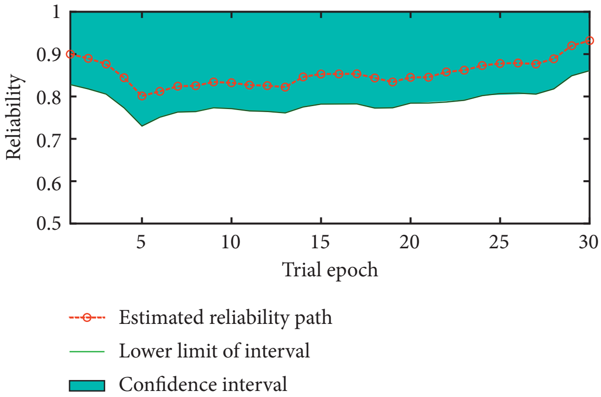


Figure 6: Estimated reliability interval for each trial.

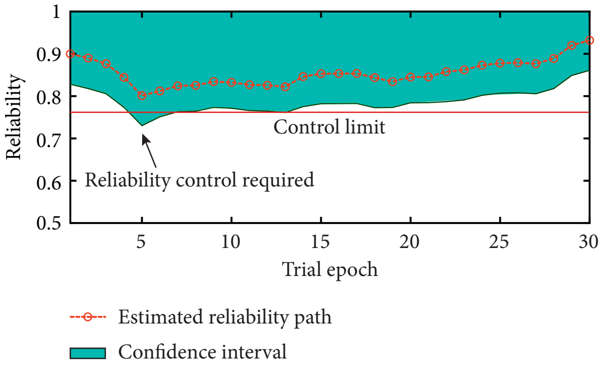


Figure 7: Illustration of reliability control process.

6. Conclusion

This paper proposes a reliability estimation and control approach for deteriorating products under production process. The new manufactured products experience a series of failure testings at different operational times before real usage. To make full use of failure testing data, a Bayesian procedure is presented to estimate the reliability distribution of the industrial products. A point estimate is derived to represent the mean reliability of the products. The estimated reliability distribution and corresponding mean value are updated with new available testing data. Afterwards, a reliability control method is proposed to monitor the 90% prediction interval. When the lower limit of the 90% prediction interval drops below a control limit, a corresponding quality-assurance action is required. The control limit policy maintains the reliability of industrial products at a satisfactory level. Finally, a case of light-emitting diodes under failure trial at production process is illustrated for validating the proposed approach.

There are two interesting directions for future research works. The first direction is to consider the heterogeneity of the industrial products in current reliability estimation model, which would be valuable for real applications. Another interesting direction of our work is to consider the missing data existing among the failure trials which commonly occur in some factories and would be a useful and appealing topic for future study.

Data Availability

The data that support the findings of the research are available upon request from the corresponding author.

Conflicts of Interest

The authors declare that they have no conflicts of interest.

Acknowledgments

This research was supported by the National Natural Science Foundation of China (NSFC) under Grant no. 62033001.

References

[1] W. Booyse, D. N. Wilke, and S. Heyns, “Deep digital twins for detection, diagnostics and prognostics,” *Mechanical Systems and Signal Processing*, vol. 140, Article ID 106612, 2020.

[2] E. Zio, “Reliability engineering: old problems and new challenges,” *Reliability Engineering & System Safety*, vol. 94, no. 2, pp. 125–141, 2009.

[3] S. Petchrompo and A. K. Parlikad, “A review of asset management literature on multi-asset systems,” *Reliability Engineering & System Safety*, vol. 181, pp. 181–201, 2019.

[4] Y. Chen, Z. Wang, Q. Ma, and K. Liang, “Reliability evaluation and failure behavior modeling of IMS considering functional and physical isolation effects,” *IEEE/ASME Transactions on Mechatronics*, vol. 24, no. 6, pp. 2441–2452, 2019.

[5] E. Chiodo and D. Lauria, “Some basic properties of the failure rate of redundant reliability systems in industrial electronics applications,” *IEEE Transactions on Industrial Electronics*, vol. 62, no. 8, pp. 5055–5062, 2015.

[6] J. Z. Sikorska, M. Hodkiewicz, and L. Ma, “Prognostic modelling options for remaining useful life estimation by industry,” *Mechanical Systems and Signal Processing*, vol. 25, no. 5, pp. 1803–1836, 2011.

[7] T. Aven, “Bayesian analysis: critical issues related to its scope and boundaries in a risk context,” *Reliability Engineering & System Safety*, vol. 204, Article ID 107209, 2020.

[8] B. Sun, X. Jiang, K. C. Yung, J. Fan, and M. G. Pecht, “A review of prognostic techniques for high-power white LEDs,” *IEEE Transactions on Power Electronics*, vol. 32, no. 8, pp. 6338–6362, 2016.

[9] S. Alaswad and Y. Xiang, “A review on condition-based maintenance optimization models for stochastically deteriorating system,” *Reliability Engineering & System Safety*, vol. 157, pp. 54–63, 2017.

[10] C. Duan and C. Deng, “Prognostics of health measures for machines with aging and dynamic cumulative damage,” *IEEE/ASME Transactions on Mechatronics*, vol. 25, no. 5, pp. 2264–2275, 2020.

[11] D. R. Insua, F. Ruggeri, R. Soyer, and S. Wilson, “Advances in Bayesian decision making in reliability,” *European Journal of Operational Research*, vol. 282, no. 1, pp. 1–18, 2020.

[12] Y. Lei, N. Li, L. Guo, N. Li, T. Yan, and J. Lin, “Machinery health prognostics: a systematic review from data acquisition to RUL prediction,” *Mechanical Systems and Signal Processing*, vol. 104, pp. 799–834, 2018.

[13] J. G. Torres-Toledano and L. E. Sucar, “Bayesian networks for reliability analysis of complex systems,” in *Proceedings of the Ibero-American Conference on Artificial Intelligence*, pp. 195–206, Springer, Lisbon, Portugal, October 1998.

[14] M. Asadi, N. Ebrahimi, and E. S. Soofi, “Optimal hazard models based on partial information,” *European Journal of Operational Research*, vol. 270, no. 2, pp. 723–733, 2018.

[15] A. Heng, S. Zhang, A. C. C. Tan, and J. Mathew, “Rotating machinery prognostics: state of the art, challenges and opportunities,” *Mechanical Systems and Signal Processing*, vol. 23, no. 3, pp. 724–739, 2009.

[16] M. C. Keizer, S. D. Flapper, and R. H. Teunter, “Condition-based maintenance policies for systems with multiple dependent components: a review,” *European Journal of Operational Research*, vol. 261, no. 2, pp. 405–420, 2017.

[17] S. Song, D. W. Coit, Q. Feng, and H. Peng, “Reliability analysis for multi-component systems subject to multiple dependent competing failure processes,” *IEEE Transactions on Reliability*, vol. 63, no. 1, pp. 331–345, 2014.

[18] W. Yao, X. Chen, Y. Huang, and M. Van Tooren, “An enhanced unified uncertainty analysis approach based on first order reliability method with single-level optimization,” *Reliability Engineering & System Safety*, vol. 116, pp. 28–37, 2013.

[19] P. V. Z. Cole, “A Bayesian reliability assessment of complex systems for binomial sampling,” *IEEE Transactions on Reliability*, vol. R-24, no. 2, pp. 114–117, 1975.

[20] D. V. Mastran and N. D. Singpurwalla, “A Bayesian estimation of the reliability of coherent structures,” *Operations Research*, vol. 26, no. 4, pp. 663–672, 1978.

[21] K. K. Sharma and R. K. Bhutani, “Bayesian reliability analysis of a parallel system,” *Reliability Engineering & System Safety*, vol. 37, no. 3, pp. 227–230, 1992.

[22] H. Bao, Y. Guo, H. Zhang, C. Peng, and J. Lu, “Bayesian analysis method on processing reliability data of high flux engineering test reactor,” *Reliability Engineering & System Safety*, vol. 199, Article ID 106912, 2020.

[23] P. Gardner, C. Lord, and R. J. Barthorpe, "Bayesian history matching for structural dynamics applications," *Mechanical Systems and Signal Processing*, vol. 143, Article ID 106828, 2020.

[24] W. Peng, Y.-F. Li, J. Mi, L. Yu, and H.-Z. Huang, "Reliability of complex systems under dynamic conditions: a Bayesian multivariate degradation perspective," *Reliability Engineering & System Safety*, vol. 153, pp. 75–87, 2016.

[25] S. Mahadevan, R. Zhang, and N. Smith, "Bayesian networks for system reliability reassessment," *Structural Safety*, vol. 23, no. 3, pp. 31–51, 2001.

[26] H. Langseth and L. Portinale, "Bayesian networks in reliability," *Reliability Engineering & System Safety*, vol. 92, no. 1, pp. 92–108, 2007.

[27] H. Boudali and J. B. Dugan, "A discrete-time Bayesian network reliability modeling and analysis framework," *Reliability Engineering & System Safety*, vol. 87, no. 3, pp. 337–349, 2005.

[28] D. Lee and R. Pan, "A nonparametric Bayesian network approach to assessing system reliability at early design stages," *Reliability Engineering & System Safety*, vol. 171, pp. 57–66, 2018.

[29] T.-T. Liu, R. Liu, and G.-J. Duan, "A principle-empirical model based on Bayesian network for quality improvement in mechanical products development," *Computers & Industrial Engineering*, vol. 149, Article ID 106807, 2020.

[30] B. Cai, X. Kong, Y. Liu et al., "Application of Bayesian networks in reliability evaluation," *IEEE Transactions on Industrial Informatics*, vol. 15, no. 4, pp. 2146–2157, 2019.

[31] S. Rebello, H. Yu, and L. Ma, "An integrated approach for system functional reliability assessment using dynamic Bayesian network and Hidden Markov model," *Reliability Engineering & System Safety*, vol. 180, pp. 124–135, 2018.

[32] J. Neyman, *Lectures and Conferences on Mathematical Statistics and Probability*, Graduate School, US Department of Agriculture, Washington, DC, USA, 1952.

[33] C. M. Thompson, E. S. Pearson, L. J. Comrie, and H. O. Hartley, *Tables of Percentage Points of the Incomplete Beta-Function*, Biometrika, Cattolica, Italy, 1941.

[34] H. C. Tijms, *A First Course in Stochastic Models*, John Wiley and Sons, Hoboken, NJ, USA, 2003.

Hindawi
Shock and Vibration
Volume 2020, Article ID 8851310, 18 pages
https://doi.org/10.1155/2020/8851310

Research Article

An Integrated Fault Identification Approach for Rolling Bearings Based on Dual-Tree Complex Wavelet Packet Transform and Generalized Composite Multiscale Amplitude-Aware Permutation Entropy

Wuqiang Liu, Xiaoqiang Yang, and Shen Jinxing

Field Engineering College, Army Engineering University of PLA, Nanjing 210007, Jiangsu, China

Correspondence should be addressed to Shen Jinxing; 565423803@qq.com

Received 13 July 2020; Revised 25 September 2020; Accepted 26 September 2020; Published 27 November 2020

Academic Editor: Chaoqun Duan

The health condition of rolling bearings, as a widely used part in rotating machineries, directly influences the working efficiency of the equipment. Consequently, timely detection and judgment of the current working status of the bearing is the key to improving productivity. This paper proposes an integrated fault identification technology for rolling bearings, which contains two parts: the fault predetection and the fault recognition. In the part of fault predetection, the threshold based on amplitude-aware permutation entropy (AAPE) is defined to judge whether the bearing currently has a fault. If there is a fault in the bearing, the fault feature is adequately extracted using the feature extraction method combined with dual-tree complex wavelet packet transform (DTCWPT) and generalized composite multiscale amplitude-aware permutation entropy (GCMAAPE). Firstly, the method decomposes the fault vibration signal into a set of subband components through the DTCWPT with good time-frequency decomposing capability. Secondly, the GCMAAPE values of each subband component are computed to generate the initial candidate feature. Next, a low-dimensional feature sample is established using the t-distributed stochastic neighbor embedding (t-SNE) with good nonlinear dimensionality reduction performance to choose sensitive features from the initial high-dimensional features. Afterwards, the featured specimen representing fault information is fed into the deep belief network (DBN) model to judge the fault type. In the end, the superiority of the proposed solution is verified by analyzing the collected experimental data. Detection and classification experiments indicate that the proposed solution can not only accurately detect whether there is a fault but also effectively determine the fault type of the bearing. Besides, this solution can judge the different faults more accurately compared with other ordinary methods.

1. Introduction

The working condition of bearings, as a vital part in rotating machinery, is closely related to the stable operation of equipment [1, 2]. Hence, real-time monitoring and prediction of the working status of the bearing is quite important to ensure safe production [3]. At present, there are a lot of mature and reliable methods to realize the fault diagnosis of bearing, such as vibration analysis, acoustic analysis [4, 5], oil analysis, and temperature analysis. Vibration signal is easy to collect and analyze, so it has been widely used and researched in the field of fault diagnosis. The fault diagnosis procedure of mechanical equipment based on vibration signal normally includes three steps: (1) collecting vibration data of equipment; (2) extracting the feature of a vibration signal to engender the initial feature; and (3) feeding the feature sample into the classifier for fault identification. Among them, the most important is the feature extraction, which is also the hotspot of current research. The quality of the extracted features directly affects the subsequent fault classification. The vibration signals of bearings are generally nonlinear [6]. Thus, it is the focus to explore the appropriate method to analyze the nonlinear vibration signal.

Due to the adverse factors such as friction, impact, and structural deformation in the working environment of bearing, the vibration signal is nonlinear and nonstationary. Therefore, how can reliable features be obtained from nonlinear data is the focus of research. With the further study of entropy-based theory [7], it becomes possible to process and analyze nonlinear data. For instance, Yan and Gao used approximate entropy (APE) for the first time in fault diagnosis to monitor the running status of bearings [8]. However, APE relies heavily on the length of data when processing signals with shorter data length, and the calculated entropy value may be less than the real. Richman and Moorman proposed sample entropy (SE) to settle the defect [9]. Unfortunately, SE may produce inaccurate estimations and undefined values. Afterwards, Bandt and Pompe presented permutation entropy (PE) [10], which measures the complexity by comparing the differences between adjacent data. Compared with other entropy-based methods, PE depends less on the model, and the calculation speed is fast and simple [11], whereas the amplitude information is neglected in the process of computing the PE. Consequently, the two time series would have with significantly different amplitudes but possibly with the same sort mode; meanwhile, the calculated PE value has an apparent error. To introduce amplitude information into the calculation process of permutation entropy, Azami and Escudero put forward amplitude-aware permutation entropy (AAPE) [12, 13] by introducing crucial information such as the amplitude and frequency of the signal into the calculation. Compared with PE, AAPE algorithm adds the deviation between amplitude and mean value of signal into the calculation process, contributing to further enhancing the stability and robustness of the algorithm.

However, both PE and AAPE are single-scale analysis methods. The actual vibration signal to be analyzed often contains information at multiple scales. The loss of a large amount of potentially useful information will be inevitably caused if only a single scale of analysis is conducted. Given the shortcomings of single-scale analysis, Costa et al. proposed multiscale entropy (MSE) [14], which can quantify the complexity of time series from multiple scales by dividing the original signal into multiple coarse-grained time series. Nevertheless, the multiscale approach adopted by MSE still has some defects: for example, the stability of the conventional multiscale computing method relies on the appropriate data length. Regarding short time series and larger-scale factors, a large entropy deviation will appear and cause the calculation result to be unreliable. Therefore, a composite multiscale method is employed in this paper to resolve the shortcomings of the traditional coarse-grained method. Meanwhile, the first-order moment (mean value) is expanded to the second-order moment (variance) during the process of coarse grained [15, 16]. Combined with AAPE, a generalized composite multiscale amplitude-aware permutation entropy (GCMAAPE) is proposed and utilized to subsequently extract the fault feature.

However, the direct usage of GCMAAPE to analyze the original signal cannot reveal the inherent characteristics such as the impact component contained in the vibration signal [17]. Thus, the entropy-based method is usually connected with the time-frequency processing method to reach a more comprehensive and detailed analysis for sake of highlighting the inherent characteristics of the vibration signal while extracting multiscale features [18]. Fourier transform (FT) is a commonly used signal analysis method while it cannot analyze the signal's time domain and frequency domain part simultaneously due to the uncertainty principle. The wavelet transform (WT) is a multiresolution analysis approach that can amplify the instantaneous changes in the signal through the window function. Nevertheless, WT is unable to analyze the high-frequency components of the signal. The wavelet packet transform (WPT) is an improved algorithm of the wavelet transform [19], which can process signals adequately and carefully, with a preferable time-frequency positioning capability. However, both WT and WPT have the same limitations, that is, the criteria for selecting the wavelet basis function cannot be well determined and it is difficult to set an appropriate decomposition level, resulting in restricting their further usage. Empirical mode decomposition (EMD) is an adaptive signal processing approach that can decompose complex signals into multiple intrinsic mode functions (IMFs). Each IMF includes features on different time scales of the raw signal [20] while EMD has serious disadvantage such as mode mixing and end effects. Besides, dual-tree complex wavelet transform (DTCWT) proposed by Kingsbury [21] possesses excellent performance such as nearly shift invariance and excellent directional selectivity while it cannot segment the high-frequency part of the signal. With the purpose of resolving the shortcoming of insufficient signal decomposition of DTCWT, dual-tree complex wavelet packet transform [22] (DTCWPT) is put forward. It can decompose the high-frequency part of the signal and solve the frequency aliasing effect of DTCWT.

Generally, the feature sample acquired through the time-frequency multiscale approach is high-dimensional and redundant, containing plenty of features that have no concern with the fault information. If directly used for classification, it not only reduces the classification efficiency but also seriously affects the identification accuracy. Consequently, it is necessary to select the features with a strong correlation to acquire sensitive low-dimensional fault features. T-distributed stochastic neighbor embedding (t-SNE) is a manifold dimensionality reduction algorithm with high nonlinear dimensionality reduction property [23]. Based on the probability distribution of random walk on the neighborhood graph, the structural relationship can be discovered in the raw data. Thus, this paper adopts t-SNE to reduce the dimensionality of the features to make up the final feature. After the final feature sample is obtained, it needs to be identified to determine the fault state. In terms of accurately estimating the current condition of the bearing, the accuracy of recognition is the first task needing to be considered. BP neural network is liable to get trapped in the local optimal value, and the convergence speed is slow. The approximation and generalization of the model are too dependent on the typicality of the selected sample. In the artificial neural network (ANN) [24], a large number of parameters need to be set, such as weight values and initial thresholds; besides, the learning time is too long and even caught in a loop without learning purpose. Support vector machine (SVM) is widely utilized owing to its excellent generalization ability and the advantage of processing small samples. SVM has parameter optimization problems that the classification performance would be severely affected by the selection of penalty coefficient and kernel function parameters [25]. With the deepening of artificial intelligence research, deep learning has been extensively studied. Deep network is a neural network simulating human brain processing information and has multiple hidden layers and multiple perceptrons. The research achievement has been successfully applied to the fields such as image recognition, speech processing, and text processing. However, these applications are all aimed at big data, and the application in small sample recognition is deficient. Deep belief network (DBN) [26–30] is a typical structure of the deep network, which is composed of multilayer restricted Boltzmann machines. The difficulty of parameter selection can be effectively avoided by adopting pretraining and fine-tuning training procedures [31]. Meanwhile, it can be trained with only a few samples, exhibiting obvious advantages in small sample recognition.

In summary, the focus of this paper is to propose an integrated fault diagnosis method based on DTCWPT, GCMAAPE, t-SNE, and DBN. The four tools (DTCWPT, GCMAAPE, t-SNE, and DBN) are employed to implement its four main targets (fault predetection, signal preprocessing, fault feature extraction, and fault pattern recognition), respectively. The main contributions and innovations of this paper can be summarized as follows:

(1) A fault diagnosis method integrating fault predetection and fault identification is presented. Different from most single step fault diagnosis methods, the proposed stepwise fault diagnosis strategy realizes the non-disassembly health detection of rolling bearing and avoids the secondary damage caused by the uncertainty of the subsequent pattern recognition, making it more consistent with the practical engineering applications.

(2) After a fault is detected in the rolling bearing, DTCWPT is used to process the vibration signal of the fault bearing, eliminating the noise and highlighting the vibration characteristics.

(3) GCMAAPE, a new nonlinear dynamic method, was used to extract fault characteristics at multiple scales from bearing vibration signals, not only overcoming shortcomings of MAAPE in the description of signal complexity but also better extracting fault characteristics by adopting generalized coarse-grained methods.

(4) DBN is introduced to automatically identify different fault types and severity and exhibits excellent generalization performance and classification efficiency compared to RF and SVM.

(5) The experimental data are used to verify the proposed method by comparing it with other methods. The effectiveness and operability of the proposed method are demonstrated by the experimental result.

2. Theory of Experimental Methods

2.1. Dual-Tree Complex Wavelet Packet Transform (DTCWPT). The DTCWPT is a modified algorithm based on the theory of DTCWT by scholars Bayram and Selesnick, overcoming the shortcomings of the DTCWT algorithm that cannot decompose the high-frequency component of signal. DTCWPT is an extension of the traditional wavelet packet transform and adopts two parallel and independent discrete wavelet packets of the low-pass filter and high-pass filter to implement signal decomposition and reconstruction. The two discrete wavelet packets are called DTCWPT's real tree and virtual tree. During the signal decomposition and reconstruction, the delay interval between the real tree and the virtual tree filter is exactly a sample value, and the sampling point of the virtual tree is kept exactly in the middle of the real tree to form the complementarity of the information, contributing to obtaining a nearly shift invariance and reduced loss of information. Besides, DTCWPT utilizes two parallel and independent discrete wavelet packets to decompose the low-frequency and high-frequency parts, exhibiting extremely high resolution while also effectively suppressing the frequency aliasing phenomenon. The decomposition and reconstruction of DTCWPT [32] are presented in Figure 1.

$S(t)$ is the input original signal; $\widehat{S}(t)$ is the reconstructed signal; $f_{1\text{-}1}$ is the high-frequency filter of the first layer decomposition of the real tree; $f_{1\text{-}0}$ is the low-frequency filter of the first layer decomposition of the real tree;$f_{2\text{-}1}$ is the high-frequency filter of the first layer of the virtual tree; $f_{2\text{-}0}$ is the low-frequency filter of the first layer of the virtual tree; $a_R(1,2)$, $a_R(1,1)$ are the decomposed components of the first layer of the real tree; $a_{\mathrm{Im}}(1,2)$, $a_{\mathrm{Im}}(1,1)$ are the decomposed components of the first layer of the virtual tree; h_1,h_0 are the filters for real tree decomposition after the second layer, g_1, g_0 are filters for virtual tree decomposition after the second layer; $a_R(2,4),\ldots,a_R(2,1)$ are the components of the second-level decomposition of the real tree; $a_{\mathrm{Im}}(2,4),\ldots,a_{\mathrm{Im}}(2,1)$ are the components of the second-level decomposition of the virtual tree; h_1', h_0' are filters for real tree reconstruction outside the second layer; g_1', g_0' are filters for virtual tree reconstruction other than the second layer; $f'_{1\text{-}1}$ is the reconstructed high-frequency filter of the first layer of the real tree; $f'_{1\text{-}0}$ is the low-frequency filter reconstructed from the first layer of the real tree; $f'_{2\text{-}1}$ is the high-frequency filter reconstructed from the first layer of the virtual tree; and $f'_{2\text{-}0}$ is the low-frequency filter reconstructed from the first layer of the virtual tree; 2↓ besides, 2↓ denotes interval sampling, and 2↑ indicates incremental sampling.

2.2. AAPE and GCMAAPE

2.2.1. AAPE. PE was proposed by Bandt in 2002 to analyze the complexity of time series, revealing that the more complex the signal being analyzed, the larger the

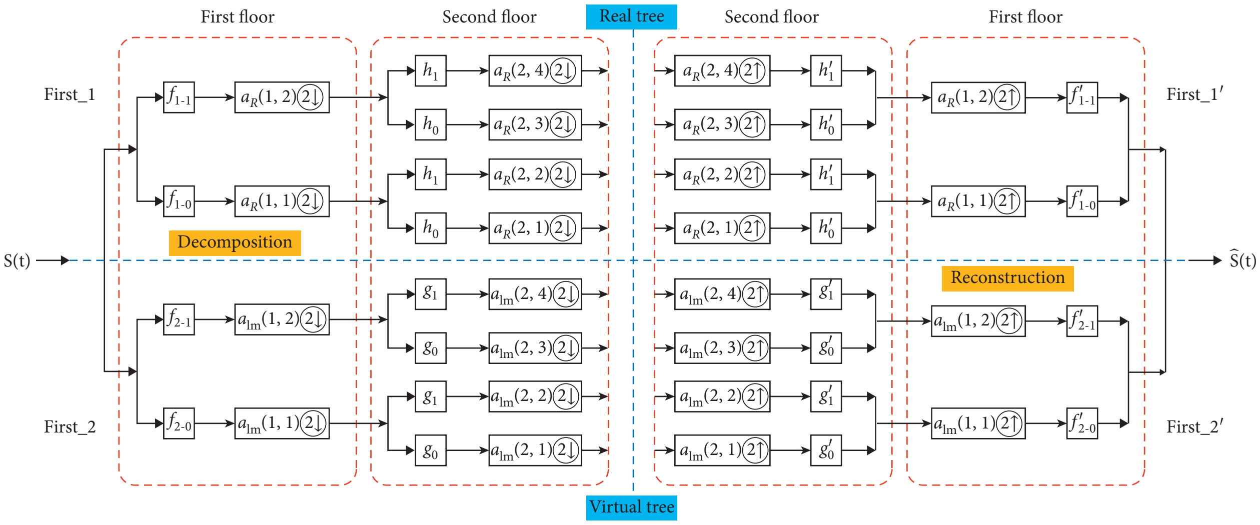


Figure 1: The decomposition and reconstruction process of DTCWPT.

permutation entropy value. For instance, the permutation entropy of white noise is greater than the permutation entropy of the cosine signal. The realization regulation of PE is described as follows [10].

Assuming a time series $x = \{x_1, x_2, \ldots, x_N\}$ of length N, a m-dimensional vector at any time t is constructed.

$$x_t^m = \{x(t), x(t+\tau), \ldots, x(t+(m-1)\tau)\},\quad t = 1, 2, \ldots, N-(m-1)\tau, \tag{1}$$

where m and τ denote embedding dimension and time delay, respectively. Define the permutation $\pi_j = (r_1, r_2, \ldots, r_{m-1})$ of $\{0, 1, \ldots, m-1\}$ when fulfils

$$x(t + r_o\tau) \le x(t + r_1\tau) \le \cdots \le x(t + r_{m-1}\tau). \tag{2}$$

When formula (2) holds, x_t^m has a permutation of π_j, where $0 \le r_n \le m-1$; $r_{n-1} < r_n$ holds when $x(t + r_{n-1}\tau) = x(t + r_n\tau)$.

For each permutation π_j, the relative frequency of $1 \le j \le m!$ can be calculated as

$$p(\pi_j) = \frac{\#\{t \mid t \le N-(m-1)\tau, x_t^m \text{ has type } \pi_j\}}{N-(m-1)\tau}, \tag{3}$$

where # represents the number of x_t^m belonging to type π_j. The PE of time series can be defined as

$$H_{\text{PE}}(x, m, \tau) = -\sum_{i}^{m!} p(\pi_j) \ln p(\pi_j). \tag{4}$$

According to the principles introduced above, PE ignores the diversity of amplitude in the uniform sort mode and may lose the signal's amplitude information. Regarding the vibration signal acquired from the rolling bearings, the amplitude includes a great deal of information related to the working state, which is the most important feature that represents the current operating condition; thus it cannot be ignored. For instance, the sequences {1, 2, 3, 4, 5, 6} and {1, 2, 3, 4, 5, 96} are the same when mapping while the mapping differences between 5-6 and 5–96 are very large. A case where different types of time series are mapped to the same sort mode by a mapping function under the embedding dimension $m = 4$ is illustrated in Figure 2. It can be observed from the figure that the distances between the four points of diverse types of time series are not equal, suggesting that the amplitudes of the vibration signals are not consistent. Nevertheless, the sort mode (1 and 2) should be the same according to the principle of PE [33, 34].

Given the problems of PE, AAPE is proposed to increase the impact of key information such as amplitude and frequency on PE calculations to enhance the stability and robustness of PE. The calculation flowchart of AAPE algorithm [17] is presented in Figure 3.

Assuming that the initial value of $p(\pi_j^{m,\tau})$ is 0, the probability of its occurrence $p(\pi_j^{m,\tau})$ for the time series $X_t^{m,\tau}$ should be recalculated whenever $\pi_j^{m\tau}$ appears when t gradually increases from 1 to $N-m+1$.

$$p(\pi_j^{m,\tau}) = p(\pi_j^{m,\tau}) + \left(\frac{A}{m}\sum_{k=1}^{m}\left|x_{t+(k-1)\tau}\right| + \frac{1-A}{m-1}\sum_{k=2}^{m}\left|x_{t+(k-1)\tau} - x_{t+(k-2)\tau}\right|\right), \tag{5}$$

where $A \in [0, 1]$ denotes the adjustment coefficient to adjust the weight of the signal amplitude mean and the deviation between the amplitudes. Thus, the probability of $p(\pi_j^{m,\tau})$ appearing in the entire time series is $\pi_j^{m,\tau}$.

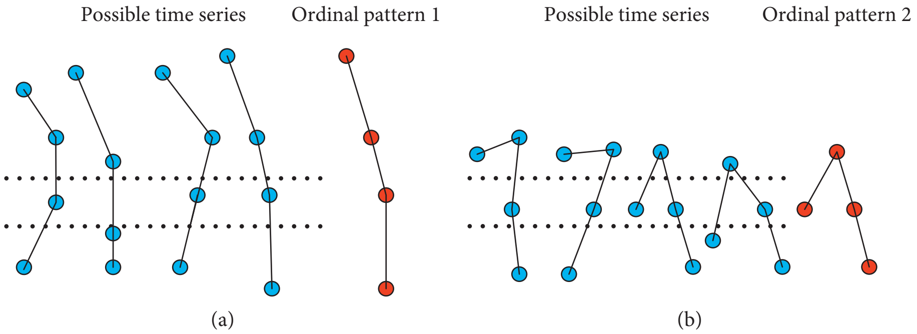


Figure 2: Two examples of diverse time series with the same sort mode.

$$p\left(\pi_j^{m,\tau}\right) = \frac{p\left(\pi_j^{m,\tau}\right)}{\sum_{t=1}^{N-m+1}\left((A/m)\sum_{k=1}^{m}\left|x_{t+(k-1)\tau}\right| + ((1-A)/(m-1))\sum_{k=2}^{m}\left|x_{t+(k-1)\tau} - x_{t+(k-2)\tau}\right|\right)}. \qquad (6)$$

The AAPE value is computed as

$$\mathrm{AAPE}\,(m,\tau,n) = -\sum_{\pi_k=1}^{\pi_k=m!} p\left(\pi_k\right)\ln p\left(\pi_k\right). \qquad (7)$$

2.2.2. GCMAAPE. The fault information contained in the vibration signals of rolling bearings usually appears on multiple scales, and a large amount of fault information will be lost if only single-scale analysis is performed. Thus, it is necessary to perform multiscale analysis to adequately extract the fault feature, and a multiscale AAPE (MAAPE) is accordingly put forward. Nevertheless, the coarse-grained approach adopted by MAAPE has the following defects. The coarse-graining process divides a time series into equal-length nonoverlapping segments and calculates the average of all data points in each segment. Therefore, the sequence of different scales of the original signal obtained by using only a single feature of the data mean will inevitably cause the loss of many potentially useful information. Consequently, a generalized composite multiscale method to address the defects of MAAPE is adopted to resolve the deficiencies of the traditional coarse-grained method. Specific steps are described as follows [16]:

(1) For the time series $x = \{x_1, x_2, \ldots, x_N\}$, the defined generalized coarse-grained time series $y_k^{(s)} = \left\{y_{k,j_1}^{(s)}, y_{k,j_2}^{(s)}, \ldots, y_{k,j_s}^{(s)}\right\}$ for scale factor s can be calculated as follows.

$$y_{k,j}^{(s)} = \frac{1}{s}\sum_{i=(j-1)s+k}^{js+k-1}\left(x_i - \overline{x_i}\right)^2,$$
$$1 \le j \le \frac{N}{S},\ 2 \le k \le s,\ \overline{x_i} = \frac{1}{s}\sum_{k=0}^{s-1} x_{i+k}. \qquad (8)$$

(2) For the scale factor s, the AAPE values of s generalized coarse-grained time series $y_k^{(s)}$ $(k = 1, 2, \ldots, s)$ are computed.

(3) The average of the s PE values is taken as the GCMAAPE value of the raw time series at the scale factor s, calculated as follows:

$$\mathrm{GCMAAPE}\,(x,m,A,\tau,s) = \frac{1}{s}\sum_{k=1}^{s}\mathrm{AAPE}\left(y_k^{(s)}, m, A, \tau\right). \qquad (9)$$

The PE value obtained by formula (9) is drawn as a function of the scale factor, called generalized composite multiscale amplitude-aware permutation entropy analysis. GCMAAPE not only synthesizes the information of multiple coarse-grained time series at the same scale but also generalizes the first-order moment (mean) to the second-order moment (variance). Theoretically, the performance of GCMAAPE is better than that of MAAPE method. Different from AAPE with single-scale analysis, GCMAAPE and MAAPE analyze time series from multiple scales. If GCMAAPE value of one time series is larger than another at most scales, indicating that the former is more random than the latter and has a higher probability of dynamic mutation.

2.2.3. The Parameter Choice Analysis for GCMAAPE. In the GCMAAPE, four vital parameters are required to be selected beforehand: embedding dimension m, adjustment coefficient A, time delay t, and scale factor s. Specifically, if the value of embedding dimension m is too small, the reconstructed vector includes too few states, and the algorithm loses its significance and effectiveness. However, the phase space reconstruction will homogenize the time series when m is too large; this not only spends much time to calculate but also fails to reflect the subtle transformation in the time series. Therefore, the embedding dimension is generally 3–7.

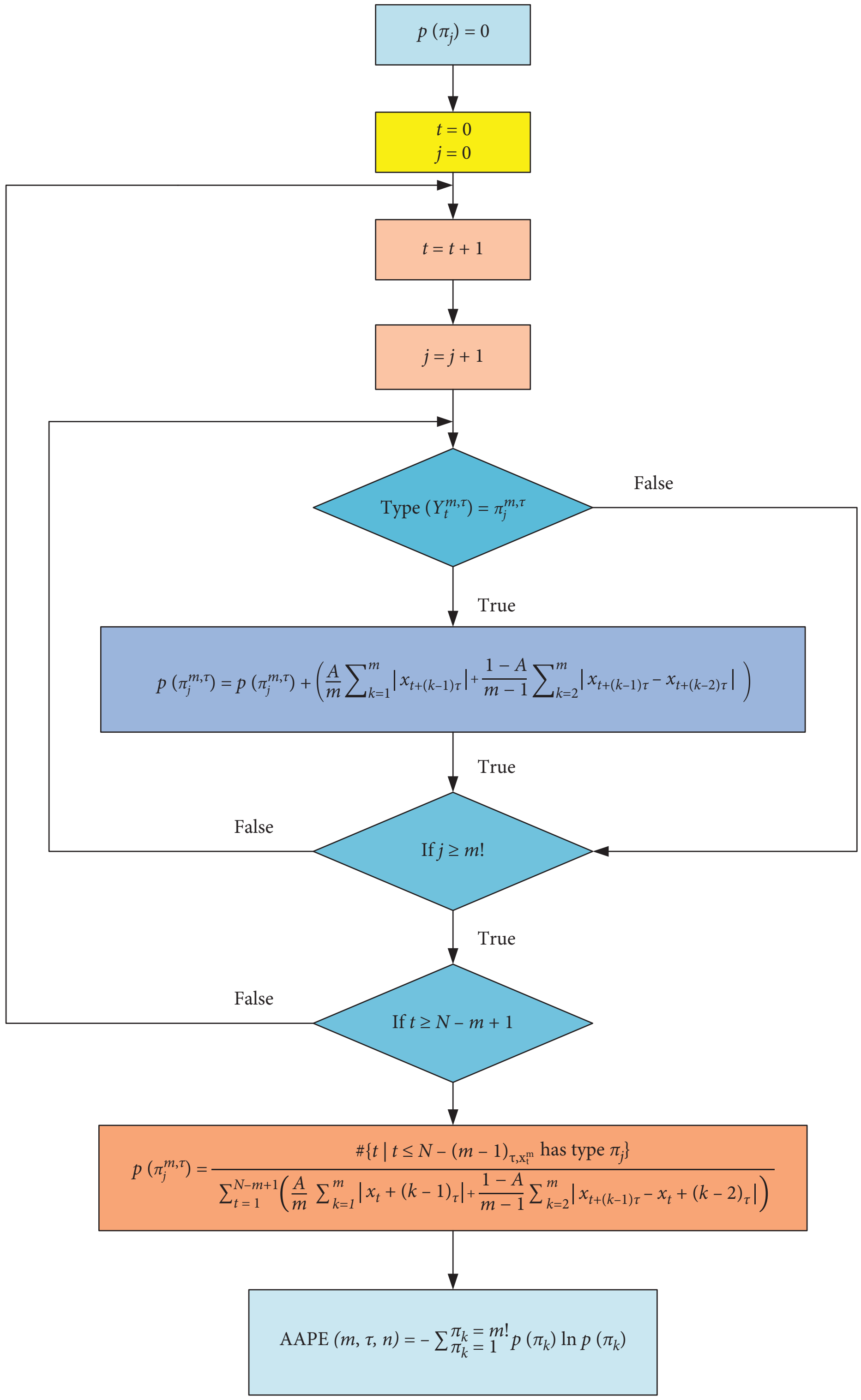

$p(\pi_j) = 0$
$t = 0$
$j = 0$
$t = t + 1$
$j = j + 1$
Type $(Y_t^{m,\tau}) = \pi_j^{m,\tau}$
False
True
$p(\pi_j^{m,\tau}) = p(\pi_j^{m,\tau}) + \left(\frac{A}{m}\sum_{k=1}^{m}\left|x_{t+(k-1)\tau}\right| + \frac{1-A}{m-1}\sum_{k=2}^{m}\left|x_{t+(k-1)\tau} - x_{t+(k-2)\tau}\right|\right)$
True
False
If $j \geq m!$
True
False
If $t \geq N - m + 1$
$p(\pi_j^{m,\tau}) = \frac{\#\{t \mid t \leq N-(m-1)_{\tau,x_t^m} \text{ has type } \pi_j\}}{\sum_{t=1}^{N-m+1}\left(\frac{A}{m}\sum_{k=1}^{m}\left|x_t+(k-1)_\tau\right| + \frac{1-A}{m-1}\sum_{k=2}^{m}\left|x_{t+(k-1)\tau} - x_t + (k-2)_\tau\right|\right)}$
AAPE $(m, \tau, n) = -\sum_{\pi_k=1}^{\pi_k=m!} p(\pi_k)\ln p(\pi_k)$

Figure 3: The flowchart of the AAPE algorithm.

Besides, the adjustment factor A is set to 0.5 according to the literature [12]. The time delay t has a small effect on the performance of GCMAAPE and is normally set to 1. Moreover, scale factor s is generally set to be larger than 10; there are no specific selection criteria while too small scale factor will lead to insufficient feature extraction and make it hard to effectively quantify the fault feature. However, too large scale factor will cause a large increase in the amount of calculation, as well as redundancy of features.

The vibration signals of rolling bearing under normal condition and the slight inner race fault condition were analyzed to understand the effect of the embedding

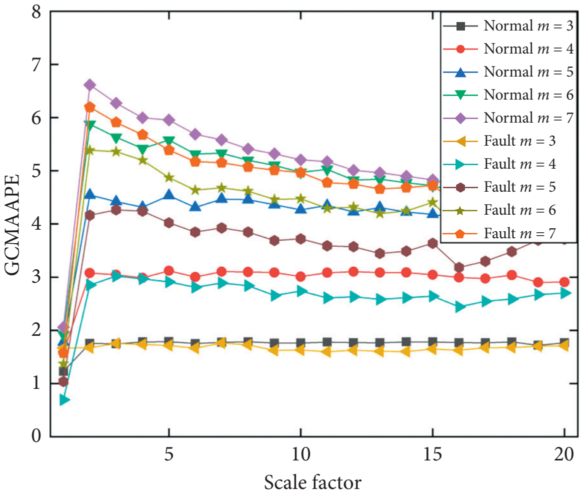


Figure 4: Effect of embedding dimension on the performance of GCMAAPE.

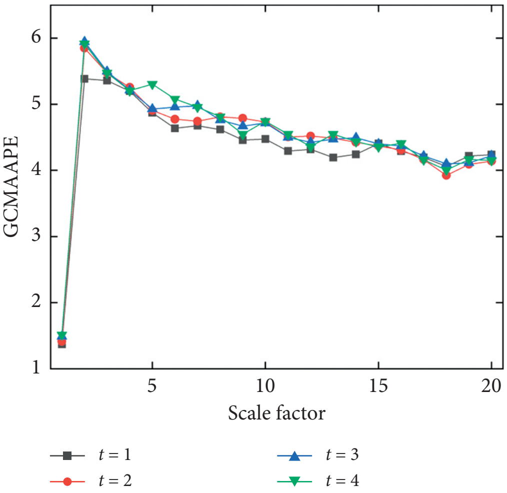


Figure 5: Effect of time delay on the performance of GCMAAPE.

dimension on GCMAAPE; the sampling points were 2400. The effect of embedding dimension ($m = 3$, 4, 5, 6, 7) on GCMAAPE performance when the time delay $t = 1$ and adjustment coefficient $A = 0.5$ is presented in Figure 4. It can be observed that the performance of GCMAAPE is the best when the embedding dimension $m = 6$. At this time, the entropy value of each scale factor has the largest difference; besides, the fault and normal states can be clearly distinguished. Consequently, the embedding dimension m is set to 6.

Secondly, the time delay $t = 1$, 2, 3, 4 when $m = 6$ is selected to test the effect of time delay t on the performance stability of GCMAAPE. The test results of the normal state vibration signal of the bearing under different time delay are illustrated in Figure 5. The different curves almost overlap together, and the difference in entropy value is very small. This indicates that the time delay t has little effect on GCMAAPE. Thus, t is set to 1 in this research according to the recommendations [13].

Finally, depending on the above analysis and the suggestions of the literature [13], the parameter settings are set as $m = 6$, $t = 1$, $A = 0.5$, and $s = 10$ in the subsequent experiments of this article.

2.3. Deep Belief Network (DBN). The deep belief network (DBN) is a probabilistic generation model consisting of multiple layers of restricted Boltzmann machines (RBMs), each of which can be simply considered independent. The layers are connected to each other, and each layer is an abstract representation of the visual layer data. A DBN network model that can be used is obtained through pretraining and fine-tuning. The detailed training procedure is divided into two steps [35].

> Step 1: each layer of RBM is trained separately to allow each layer to contain as many features of the input data as possible. Specifically, the input vector is first mapped to the output through the weight. Then, the output is obtained, and the input vector is reconstructed in turn. The reconstructed deviation is used as the basis for updating the weight. This process is repeated until the deviation between the input vector and the output vector is tiny. The procedure of forward and backward is the learning process of RBM.
>
> Step 2: in the former process, each RBM network only optimizes the mapping relationship between the input and output of its own layer, while it does not make the entire network structure reach the optimal. Therefore, it is necessary to establish a softmax classifier in the ultimate layer of DBN; then, the output feature vector of RBM is taken as the input feature vector of the softmax classifier to train the softmax classifier with supervision; next, the error between output and input is propagated to RBM of each layer from top to bottom; finally, all parameters of the network can be finely turned. The structure of the DBN classifier is presented in Figure 6.

In the learning process of DBN, the training of RBM is the core. The network parameters are initialized using the layer-by-layer learning of RBM. Although the resulting initialized network parameters are not optimal parameters, they are generally in the vicinity of the optimal parameters, avoiding the BP algorithm to easily fall into the local optimal and training time and other defects caused by the random initialization of network parameters during training DBN.

3. The Proposed Approach

This paper proposes an integrated fault identification technology for rolling bearings based on the advantages of DTCWPT, GCMAAPE, and t-SNE in rolling bearing fault feature extraction and combining DBN models that can handle high-dimensional data classification problems, this paper proposes an integrated fault identification technology for rolling bearings. The integrated technology includes the following two parts.

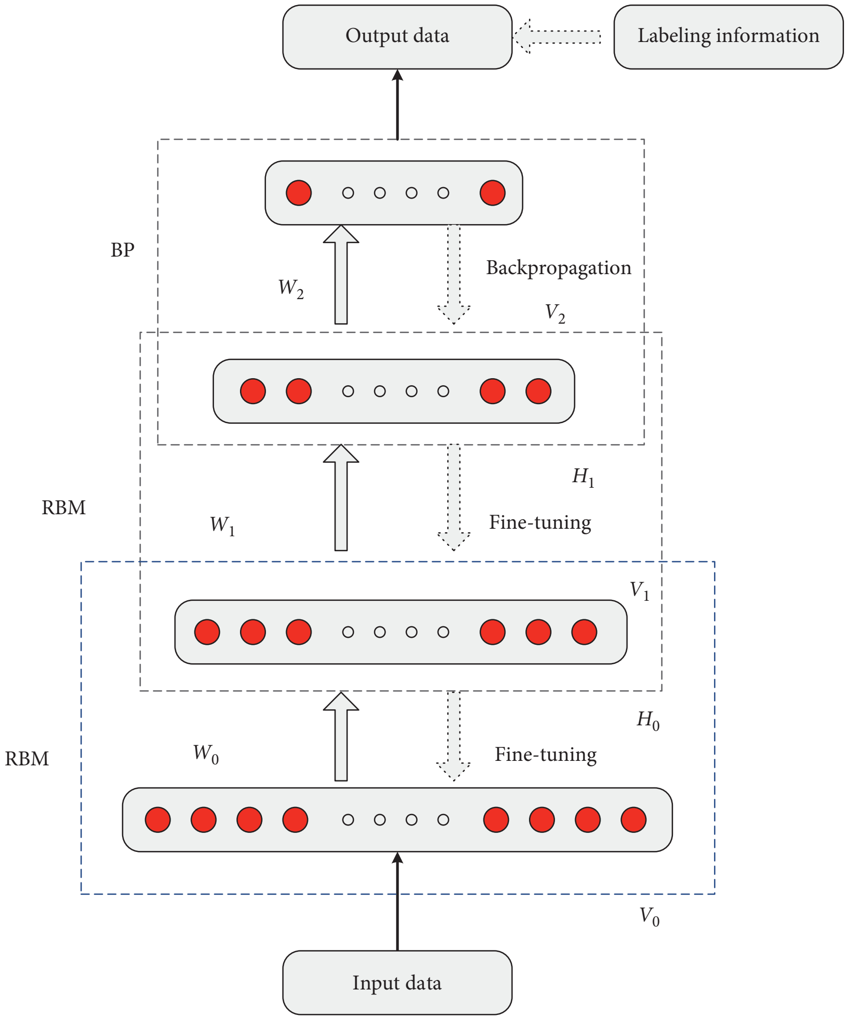


Figure 6: Structure diagram of double hidden layer DBN network.

3.1. Fault Predetection. As an improvement on the permutation entropy, AAPE has similar functions to PE. AAPE can detect the failure of equipment, indicating the ability to detect the failure. The sensitivity of AAPE to normal and the fault was used to screen the normal and fault of bearing.

When the scale factor is 1, the GCMAAPE value of the normal vibration signal is smaller than that of the fault vibration signal, with an obvious difference value. Therefore, the GCMAAPE value when the scale factor is 1 is used to differentiate the normal and fault states. Therefore, a threshold value is designed to detect the current health condition of the bearing to screen more intuitively.

3.2. Fault Classification. After the predetection, it is necessary to make further analysis to judge the bearing fault type and severity if the bearing fault is detected. A novel time-frequency multiscale feature extraction approach based on DTCWPT, GCMAAPE, and t-SNE was proposed. There are two traditional multiscale feature extraction methods: (1) the nonlinear time-frequency algorithm is adopted to decompose the signal into multiple components, and then the single entropy of multiple components is acquired; (2) the MSE of a single component is computed. Compared with the traditional method, the new time-frequency multiscale feature extraction method avoids the problem of insufficient feature extraction by extracting the multiscale entropy of multiple components to highlight the impact and resonant components in the fault vibration signal. The basic principle is to decompose the fault vibration signal into several components of different frequency bands and then use GCMAAPE to extract the fault characteristics of each component. Next, t-SNE is utilized to select sensitive features to acquire low-dimensional final feature vector. Finally, the DBN classification model is trained and tested with the final feature vector to classify different fault states. The technical route of the presented means is illustrated in Figure 7. The process of implementing the integrated fault diagnosis method includes the following six steps.

Step 1. Vibration data acquisition: collect vibration signals of running rolling bearings under different work conditions with sensors. Divide the collected experimental data into multiple samples of length N, which have no overlap between the sequences.

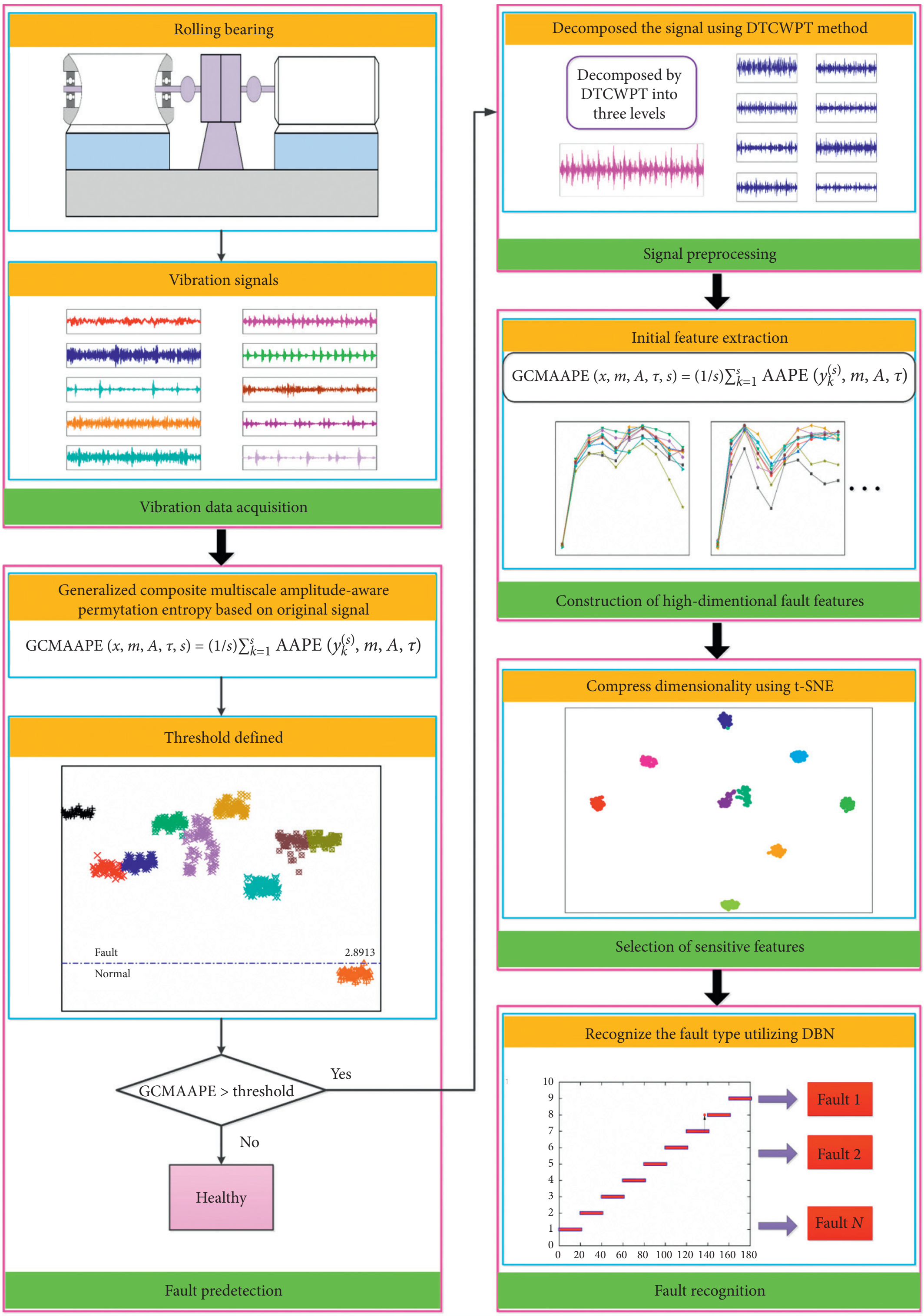

Rolling bearing
Vibration signals
Vibration data acquisition
Generalized composite multiscale amplitude-aware permytation entropy based on original signal
GCMAAPE $(x, m, A, \tau, s) = (1/s)\sum_{k=1}^{s}$ AAPE $(y_k^{(s)}, m, A, \tau)$
Threshold defined
Fault
2.8913
Normal
GCMAAPE > threshold
Yes
No
Healthy
Fault predetection
Decomposed the signal using DTCWPT method
Decomposed by DTCWPT into three levels
Signal preprocessing
Initial feature extraction
GCMAAPE $(x, m, A, \tau, s) = (1/s)\sum_{k=1}^{s}$ AAPE $(y_k^{(s)}, m, A, \tau)$
. . .
Construction of high-dimentional fault features
Compress dimensionality using t-SNE
Selection of sensitive features
Recognize the fault type utilizing DBN
10
9
8
7
6
5
4
3
2
1
0
0 20 40 60 80 100 120 140 160 180
Fault 1
Fault 2
Fault N
Fault recognition

Figure 7: The technical route of the presented approach.

Step 2. Fault predetection: calculate the GCMAAPE value when the scale factor is 1 and set a threshold value based on the GCMAAPE value to judge whether the bearing is healthy. If the GCMAAPE value of the bearing vibration signal to be detected is less than the threshold value, the bearing is healthy, the output is normal, and the diagnosis ends. Otherwise, proceed to the next step to judge the type and severity of the bearing failure.

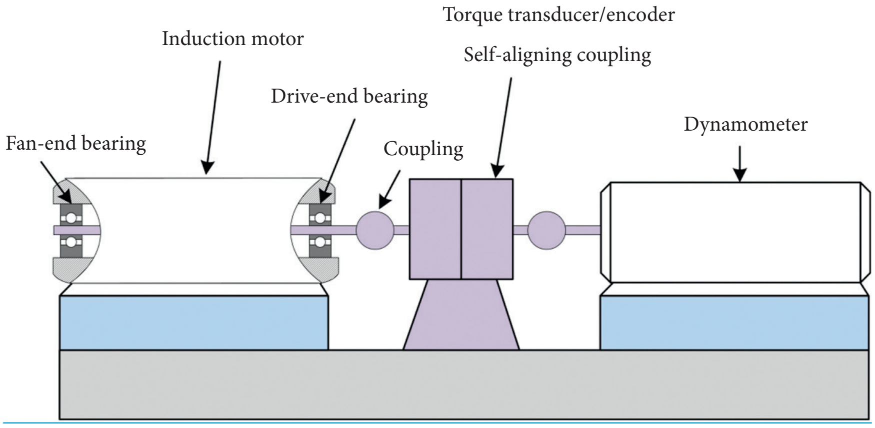


Figure 8: Bearing failure simulation test bench.

Step 3. Signal preprocessing: decompose the fault vibration signal with DTCWPT to obtain several subbands of different frequency bands.

Step 4. Construction of high-dimensional fault features: calculate the GCMAAPE value of each subband component to form the initial candidate feature vector set and perform normalization processing as input for the next step.

Step 5. Selection of sensitive features: choose sensitive features from the normalized initial features using the t-SNE to construct a final feature vector.

Step 6. Fault recognition: divide the normalized feature vector into a training sample and a test sample and establish an optimal DBN classification model through the training and test.

4. Experimental Process and Results

4.1. Experimental Data. The bearing vibration signal is adopted to verify the performance of the proposed DTCWPT-GCMAAPE-t-SNE-DBN model. The rolling bearing vibration signal data used in the experiment are collected from the rolling bearing failure simulation test bench of the Electrical Engineering Laboratory of Case Western Reserve University in the United States [36]. The rolling bearing model used for the test is 6205-2RS-SKF deep groove ball bearings. A bearing failure simulation test bench is exhibited in Figure 8.

The bearing data used include vibration data of the drive end bearing under 10 operating conditions, which are normal working conditions (labeled NM), inner race fault conditions (labeled IRF1, IRF2, and IRF3), outer race fault conditions (labeled ORF1, ORF2, and ORF3), and ball fault conditions (labeled BF1, BF2, and BF3). The fault diameters of the three fault types are 0.1778 mm, 0.3556 mm, and 0.5334 mm. The different fault diameters indicate the severity of bearing damage. In this test, the sampling frequency is 12 kHz, the rotating speed of the motor is 1797 rpm, and the load is 0 HP. The details of the data used in the experiment are listed in Table 1. Each group of signals is divided into multiple groups of nonoverlapping samples. Since each sample consists of 2400 sampling points, each state contains 50 samples. Therefore, the experimental data used are composed of 10 working conditions, each of which contains 50 sets of samples. Among them, the samples of each state use 30 groups as the training set for the DBN classification model and the remaining 20 groups are used as the testing set.

4.2. Results and Analysis. The time-domain waveforms of vibration signals of bearing with different fault types and severity are illustrated in Figure 9. The waveform of the vibration signal lacks regularity, making it difficult to determine the working condition of the bearing directly from the time-domain waveform. Therefore, further measures need to be taken to determine the working state of the bearing. Similar to PE, AAPE can detect the state of the bearing to avoid secondary damage to the bearing, according to the previous theoretical analysis. The AAPE value of all samples is presented in Figure 10. It can be observed that the AAPE value of the bearing in the fault state is generally large, and the AAPE value of the bearing in the normal state is small; this is significantly different from the AAPE value of the fault state. Therefore, this method can be used to detect the normal state of the bearing. The value at the red dotted line is defined as the AAPE threshold (2.8913). Besides, the normal and fault states can be clearly distinguished by comparing the AAPE value of the vibration signal with the threshold. Moreover, the two indicators of detection accuracy (DTA) and missed detection rate (MDR) are used to make it more intuitive to evaluate the efficiency of the approach in the predetection. As demonstrated in Figure 10, the entropy values of all fault samples are distributed above the threshold, and all normal samples are distributed below the threshold. According to mathematical statistical analysis, the indicator DTA has reached 100%, and the MDR is 0%. Depending on the definition, the larger the value of DTA, the higher the accuracy of detecting normal samples, and the smaller the MDR, the lower the probability of misdiagnosis. In summary, the larger the DTA and the smaller the MDR, the more effective the method. Furthermore, it can be verified

Table 1: The brief description of experimental data.

Health state	Fault diameter (mm)	Fault severity	Abbreviation	Training sample number	Testing sample number	Classification label
Normal	None		NM	30	20	0
Inner race fault	0.1778	Minor	IRF1	30	20	1
	0.3556	Medium	IRF2	30	20	2
	0.5334	Severe	IRF3	30	20	3
Outer race fault	0.1778	Minor	ORF1	30	20	4
	0.3556	Medium	ORF2	30	20	5
	0.5334	Severe	ORF3	30	20	6
Ball fault	0.1778	Minor	BF1	30	20	7
	0.3556	Medium	BF2	30	20	8
	0.5334	Severe	BF3	30	20	9

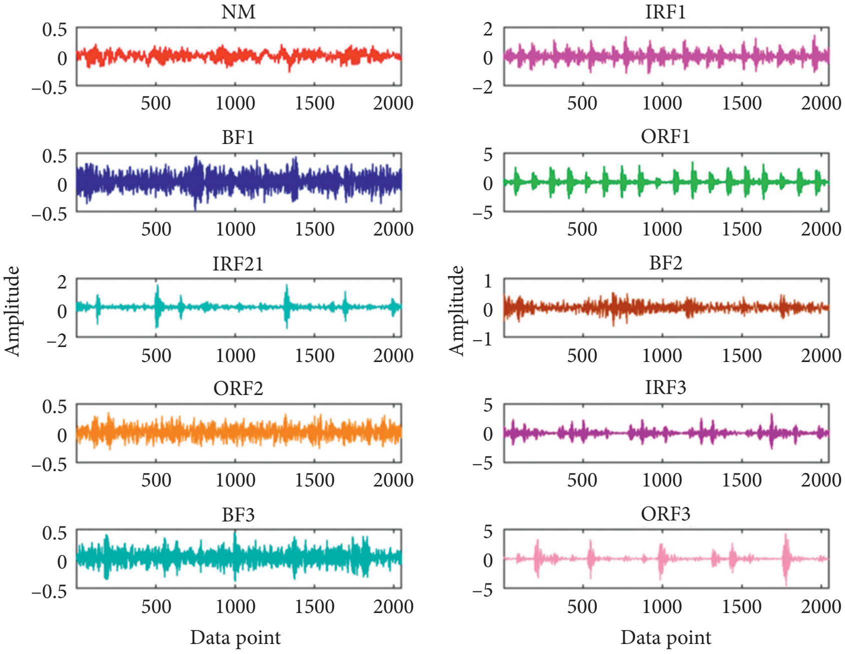


Figure 9: The time-domain waveforms of bearings with different types and fault severity.

by the experimental results that the proposed approach has an excellent performance in the bearing predetection stage.

After predetecting the working condition of the bearing, the fault recognition approach is used to diagnose the type and severity of the fault if the bearing does have a fault. First, each sample is decomposed to three levels adopting DTCWPT to highlight the impact components in the fault vibration signal and reduce the interference between the components. Next, eight subband components including diverse frequency band information can be acquired. Besides, only the DTCWPT decomposition results of minor outer race faults (ORF1) are used as representatives to reduce the space footprint. The results of the decomposition are exhibited in Figure 11.

After decomposing the signal, the GCMAAPE algorithm is used to fetch features of diverse scales from each subband signal. Considering space limitations, the GCMAAPE values of only subband signal 1, subband signal 2, subband signal 3, and subband signal 4 over scale factor of 10 under 9 working conditions are illustrated in Figure 12. It can be observed from Figure 12 that the GCMAAPE curves of the four subbands almost overlap when the scale factor is 1–4, demonstrating a poor separability of each state at this time. Moreover, the curve of each state has a significant difference when the scale factor is 4–10, indicating a strong separability. However, it is not possible to directly determine the fault condition based solely on the curve distribution in the figure, and further analysis is required to make the features obvious enough so as to recognize the diverse type of the bearing.

After completing the initial feature extraction, the R^{450*80} dimensional feature can be obtained. Apparently, the original feature is high-dimensional and redundant. It will not only reduce the efficiency but also guarantee the recognition accuracy if used directly for classification. Thus, it is indispensable to reduce the dimension of the fault feature. The t-SNE is adopted to select sensitive features for the

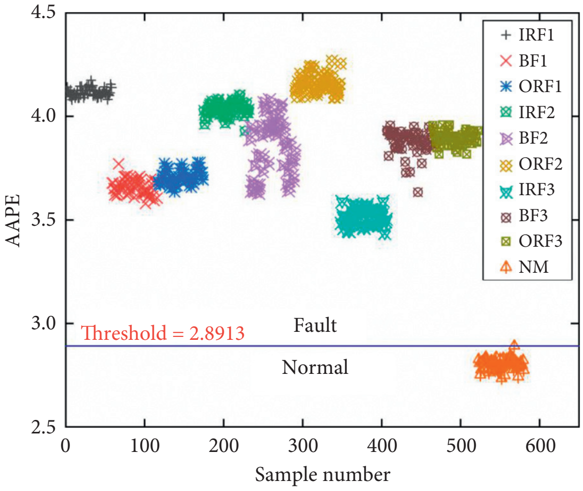


Figure 10: The amplitude-aware permutation entropy distribution of all samples.

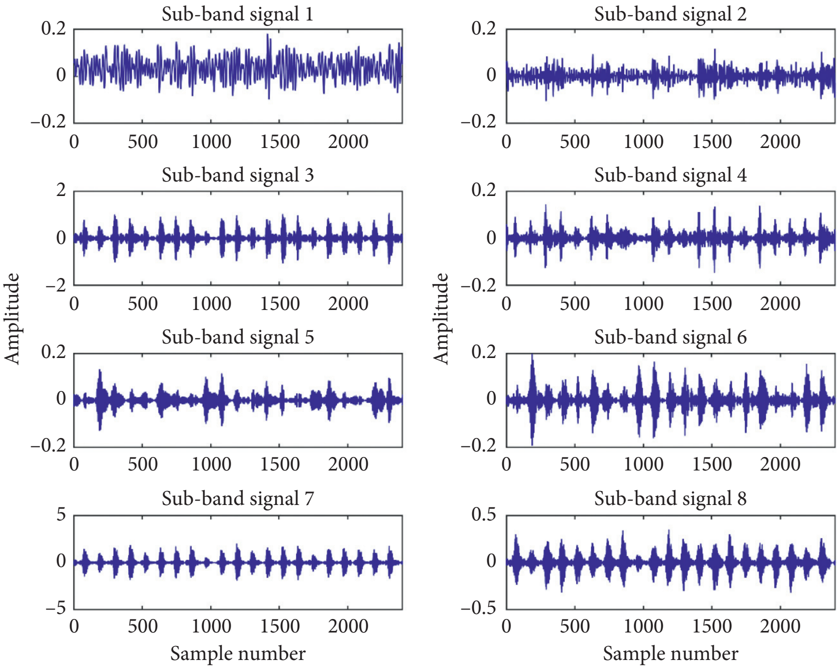


Figure 11: DTCWPT results of the sample with the minor outer race fault.

original high-dimensional features. The feature vector after dimensionality reduction is R^{450*3}. Then, the DBN classification model is trained using 270 groups of samples, and the property of the trained model is verified employing the remaining as test samples. The classification results of the proposed diagnostic model for nine fault conditions under one trial are illustrated in Figure 13. In the 180 sets of test samples, one BF1 sample was misclassified as BF2 sample. It can be revealed that the classification accuracy of minor outer race fault (BF2) conditions is 95% while the classification accuracy of the other eight states reaches 100%. Besides, the recognition rate of the proposed diagnosis model reaches as high as 99.44% for all working conditions, fully demonstrating that the solution possesses the best diagnostic performance and good robustness for diverse working conditions and fault severity.

The proposed approach is compared with the other three approaches used on the bearing dataset to verify its excellent performance. The other three methods are DTCWPT and GCMPE (DTCWPT-GCMPE), DTCWPT and MAAPE (DTCWPT-MAAPE), and DTCWPT and MSE (DTCWPT-MSE). For DTCWPT-GCMPE [37], $r = 3$, $m = 6$, $t = 1$, and $s_{\max} = 10$. For DTCWPT-MAAPE, $r = 3$, $m = 6$, $t = 1$, $A = 0.5$, and $s_{\max} = 10$. For DTCWPT-MSE, $r = 3$, $m = 2$, $t = 1$, $\mathrm{rd} = 0.15$, and $s_{\max} = 10$. Among them, r denotes the number of decomposition layers of DTCWPT, m refers to the embedding dimension, t is the time delay, rd represents the tolerance of the signal, A is the adjustment coefficient, and

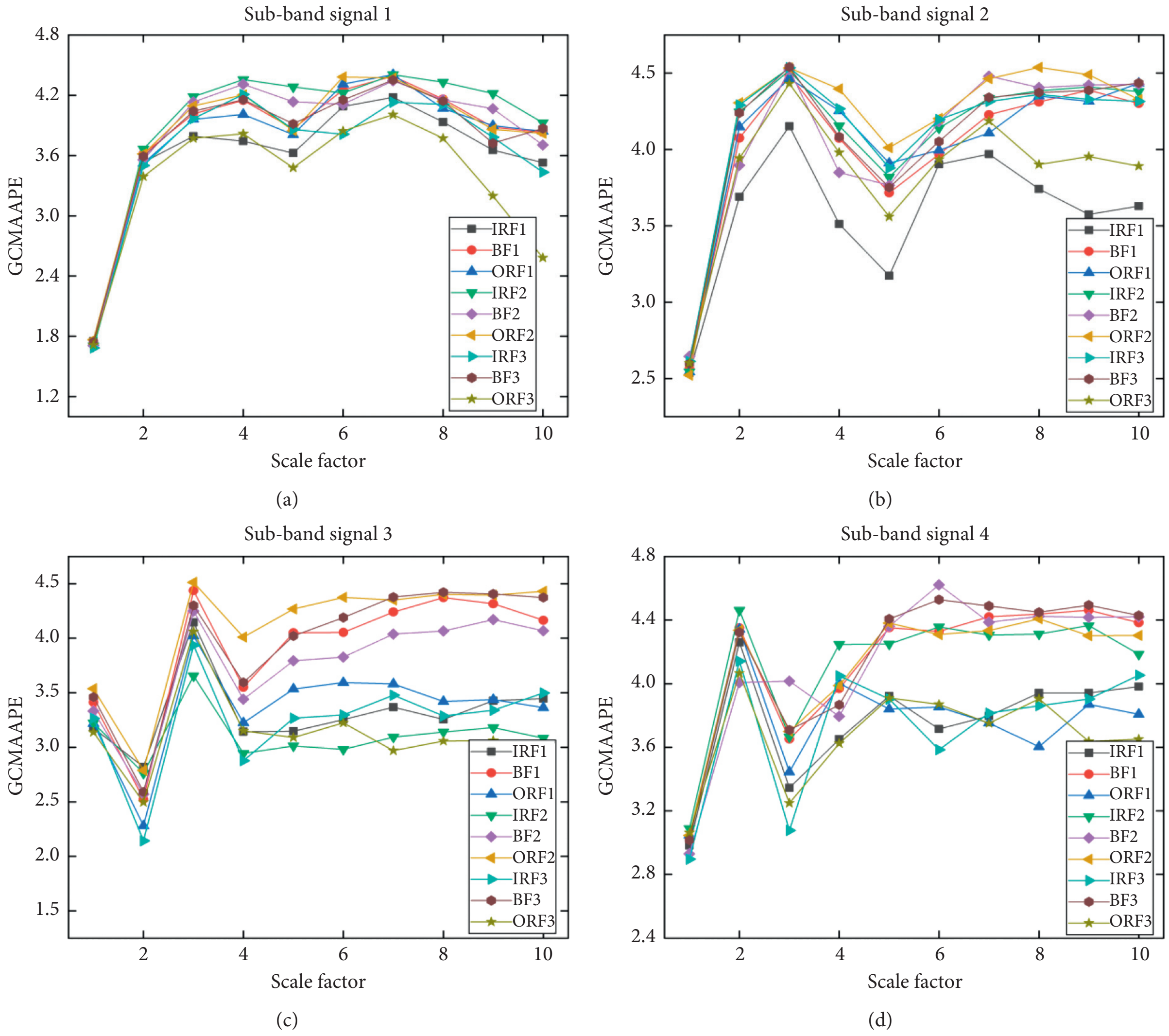


Figure 12: GCMAAPE value of four subband signals over 10 scales under nine different working conditions. (a) Subband signal 1. (b) Subband signal 2. (c) Subband signal 3. (d) Subband signal 4.

$s_{\max}$ indicates the scale factor. The classifier used in each method is DBN. Each method is repeated 25 times to avoid errors caused by random factors such as accidental factors. The experimental results of diverse approaches are presented in Figure 14 and Table 2. It can be clearly observed from Figure 14 and Table 2 that the classification effect of the proposed method is obviously superior to several other methods. The proposed method exhibits a maximum recognition rate of 100% and an average classification accuracy of 98.82%. After replacing GCMAAPE with GCMPE, the classification accuracy reaches 98.33%, and the average classification accuracy is 96.58%, lower than that of GCMAAPE. The main reason is that GCMAAPE introduces the amplitude information of the vibration signal into the calculation process, contributing to improving the utilization rate of fault information and obtaining higher-quality features. Besides, the average correct rate of MAAPE is lower than that of GCMAAPE, verifying that the generalized coarse-grained method adopted is superior to the traditional coarse-grained method used by MAAPE. Generally, the GCMAAPE method used is superior to several common entropy-based methods in performance, which is directly reflected in the higher recognition rate. Thus, the robustness of the approach to fault classification problems is fully demonstrated.

The advantages of this method in feature extraction are verified by comparing the performance of the DTCWPT-GCMAAPE feature extraction method with the following feature extraction methods. GCMAAPE acts on the raw vibration signal, the approach based on EEMD and GCMAAPE (EEMD-GCMAAPE) which can be found in the literature [38, 39] and the approach based on WT and GCMAAPE (WT-GCMAAPE). The parameters of these three methods are set as follows. For EEMD-GCMAAPE, $M = 100$, $sd = 0.2$, $m = 6$, $A = 0.5$, $t = 1$, and $s_{\max} = 10$. For WT-GCMAAPE, $r = 3$, wavelet basis function is db4, $m = 6$, $A = 0.5$, $t = 1$, and $s_{\max} = 10$. Among them, M is the ensemble number of EEMD, and sd is the standard deviation of added white noise in EEMD. The fault features extracted by the above three methods are finally inputted to the DBN

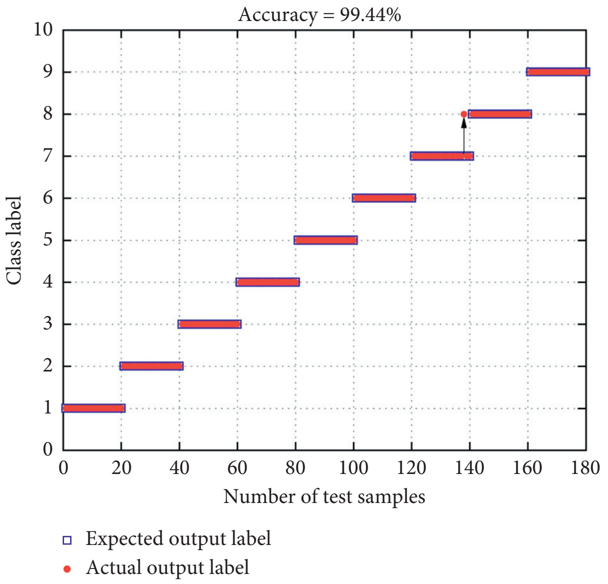


Figure 13: Classification result of the proposed model.

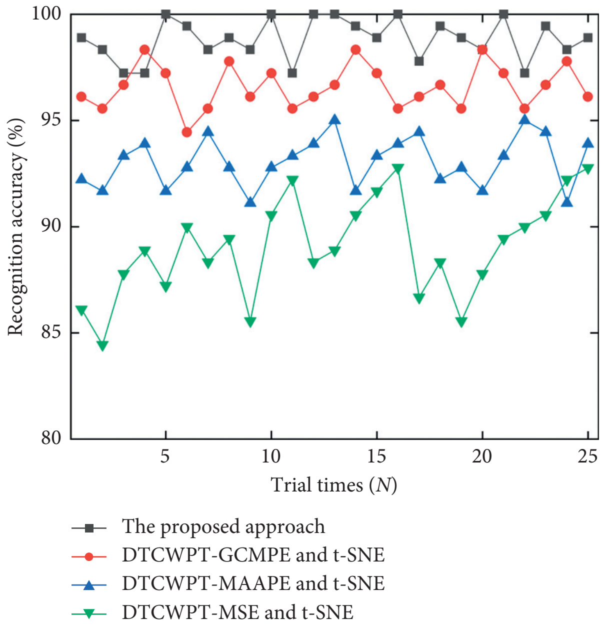


Figure 14: The classification results of five approaches during 25 trials.

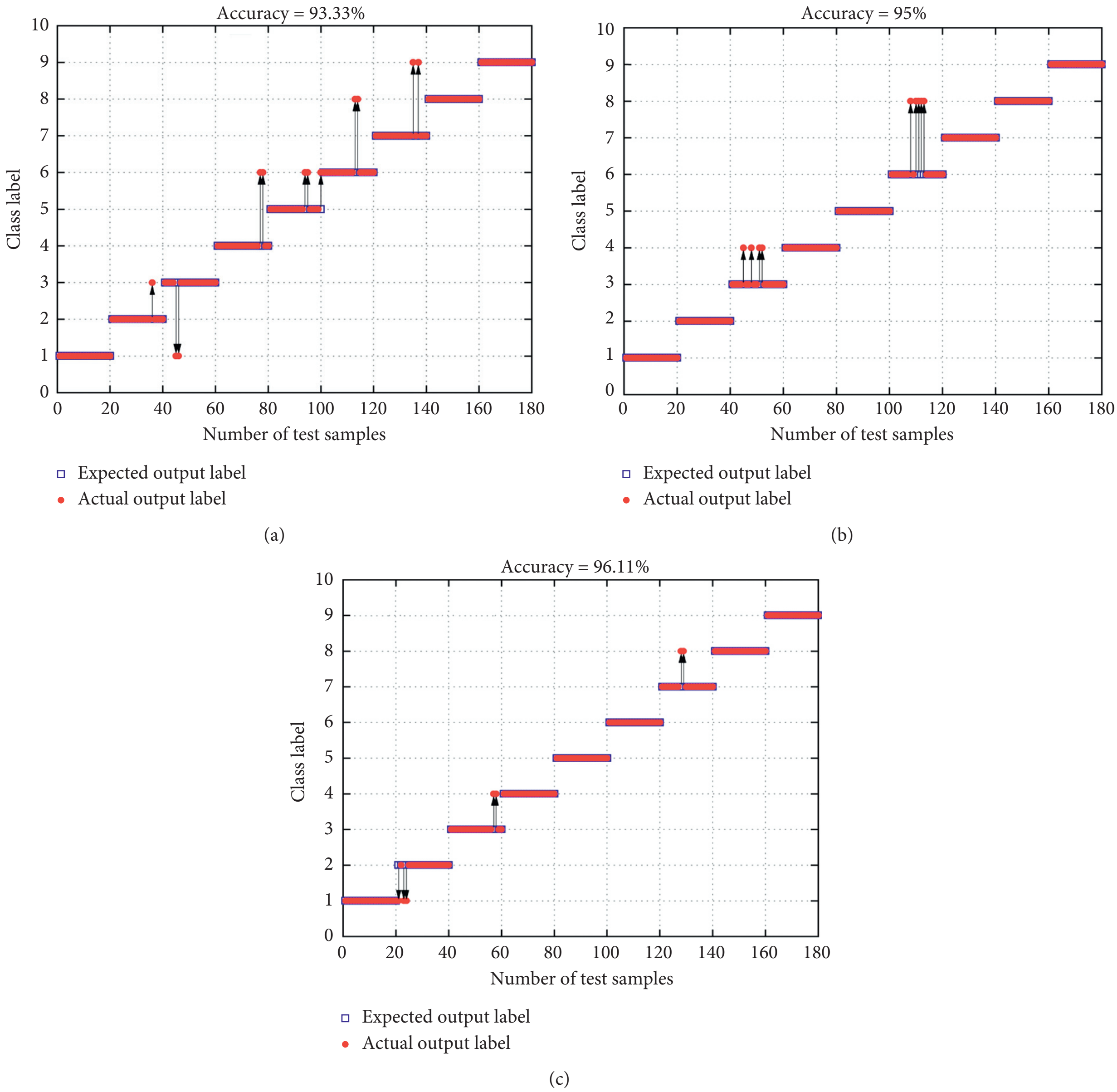
Table 2: The classification results of different approaches.

Approaches	Accuracy (%)		
	Max	Min	Mean
The proposed approach	100	97.22	98.82
DTCWPT-GCMPE	98.33	94.44	96.58
DTCWPT-MAAPE	95	91.11	93.07
DTCWPT-MSE	92.78	84.44	89.04

classifier to classify the fault type. The selection of parameters and the proportion of samples are consistent with the proposed method. The fault classification results of the three methods are presented in Figure 15, and the classification accuracy rates reach 93.33%, 95%, and 96.11%, respectively, lower than the proposed method in Figure 13. This is because when the GCMAAPE method is directly applied to the original vibration signal, the information such as the fault frequency in the fault vibration signal cannot be extracted, and the fault information is not sufficiently analyzed, resulting in a reduction in the quality of the feature. Both WT-GCMAAPE and EEMD-GCMAAPE are a multiscale analysis method based on time-frequency analysis. These two methods also have some problems limiting the quality of feature extraction. Besides, the WT method cannot effectively decompose the high-frequency part of the signal when analyzing the signal. EEMD has a serious mode aliasing effect, making the decomposed IMF have a large interference component. Compared with these three methods, the proposed DTCWPT-GCMAAPE feature extraction method is a multiscale analysis method based on time-frequency preprocessing and can reflect more fault information by highlighting different frequency components of vibration signals and by multiscale analysis, contributing to improving the quality of extracted features for better classification.

The classification experiments of the four methods in two conditions are compared to explore the necessity of dimensionality reduction and validate the performance of t-SNE in feature dimensionality reduction. The two conditions are without dimension reduction and with LDA dimension reduction. The experimental results of the four methods under different dimensionality reduction conditions are provided in Table 3. It can be observed from Table 3 that DTCWPT-GCMAAPE has achieved the best classification effect in both conditions while the accuracy rate of using LDA dimensionality reduction and nondimensionality reduction is lower than that presented in Table 2. It indicates that dimensionality reduction is necessary, and the dimensionality reduction performance of t-SNE is better than that of LDA. Without loss of the generality, two features are selected from the original features, as illustrated in Figure 16(a), and the two-dimensional visualization after LDA dimensionality reduction is presented in Figure 16(b). Moreover, the two-dimensional visualization after t-SNE dimensionality reduction is exhibited in Figure 16(c). Apparently, the different fault states in Figure 16(a) are well separated compared to Figures 16(a) and 16(b). Simultaneously, a BF2 sample is erroneously divided into IRF3 samples, confirming the classification result of Figure 13. Through dimension reduction, the subsequent classification becomes faster and efficiency and accuracy are improved.

The feature samples obtained by the proposed method are sent to different classifications for comparison (namely, support vector machine (SVM) and random forest (RF)) to verify the performance of the selected DBN classifier. The classification results and average running time are presented in Table 4. Besides, each method runs for 30 times to ensure that it is not affected by random factors. It can be revealed from Table 3 that the DBN classifier achieves the best classification effect, and the RF classifier requires the least running time. Thus, DBN has the highest classification accuracy, even though the DBN classifier has more running



(a)

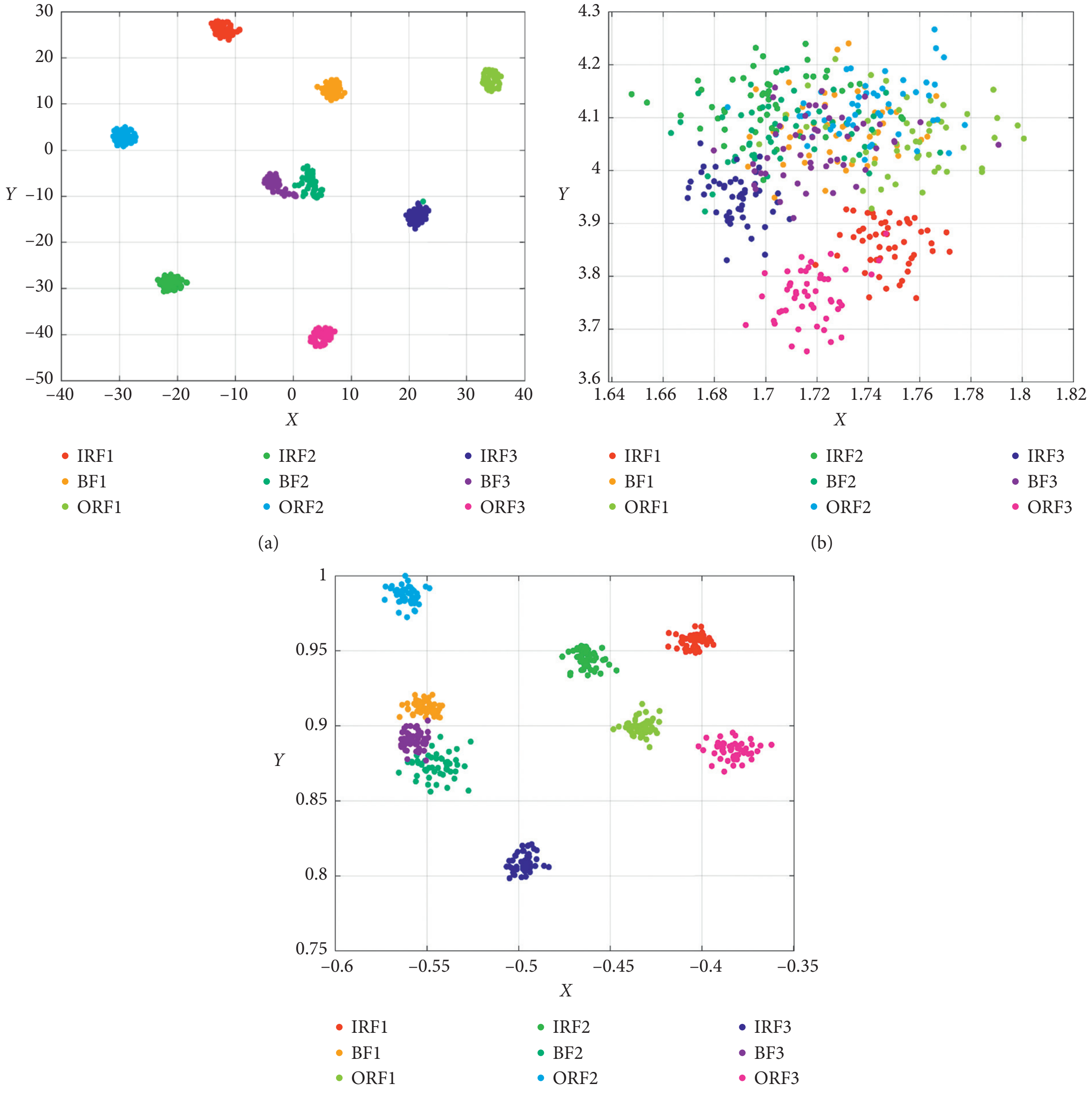
(b)

(c)

Figure 15: Outputs of GCMAAPE, WT-GCMAAPE, and EEMD-GCMAAPE based fault diagnosis approaches. (a) GCMAAPE. (b) WT-GCMAAPE. (c) EEMD-GCMAAPE.

Table 3: The recognition rate of four methods under different conditions.

Different conditions	Different methods	Accuracy (%)		
		Max	Min	Mean
LDA	DTCWPT-GCMAAPE	96.67	92.78	94.51
	DTCWPT-GCMPE	95.56	91.11	92.74
	DTCWPT-MAAPE	91.67	87.22	89.86
	DTCWPT-MSE	86.11	81.67	83.72
Without t-SNE	DTCWPT-GCMAAPE	93.33	90	91.58
	DTCWPT-GCMPE	92.22	88.89	90.14
	DTCWPT-MAAPE	89.44	85	86.95
	DTCWPT-MSE	82.78	78.89	81.29



(a)

(b)

(c)

Figure 16: The two-dimensional view of the first two features by DTCWPT-GCMAAPE: (a) adopting t-SNE; (b) without adopting t-SNE; (c) adopting LDA.

Table 4: The classification results of different classification models for 30 trials.

Classifier	Accuracy (%)			Time (s)
	Max	Min	Mean	
DBN	100	96	98.47	7.98
RF	100	93	96.23	5.61
SVM	95	88	91.25	11.36

time than RF. DBN achieves better performance by sacrificing part of the running time, which is acceptable to a certain extent. The running time of SVM is the longest among the three methods while its classification effect is not ideal. This is mainly because SVM is suitable for processing small sample data and cannot achieve the best performance when processing large batches of high-dimensional data. Generally, the DBN classification model not only has a shorter running time but also exhibits better performance.

5. Conclusion

In this research, a synthesized fault identification technology including bearing predetection and fault classification is proposed to detect and identify the status of the bearing. At the stage of predetection, an AAPE threshold value that can differentiate between normal and fault conditions is defined by calculating the AAPE value of the bearing vibration signal under different working conditions so as to screen out the bearings with normal working conditions. Specifically, the time-frequency multiscale method DTCWPT-GCMAAPE-t-SNE is utilized to extract the fault feature and generate a fault feature sample if the bearing is detected to be faulty. Finally, a DBN classifier with powerful classification performance is used to classify the acquired high-dimensional features. The classification effects of WT-GCMAAPE, EEMD-GCMAAPE, and GCMAAPE are compared. The results indicate that the proposed approach can accurately highlight the fault information in the vibration signal and improve the quality of the features extracted subsequently. Simultaneously, it is compared with MAAPE, GCMPE, and MSE, demonstrating that GCMAAPE can effectively extract fault features from the DTCWPT processed signal and has better robustness. Generally, compared with other common fault diagnosis methods, this paper introduces bearing predetection, which avoids the subsequent model classification with uncertainty and improves the diagnosis efficiency. It has practical engineering significance and is more suitable for practical engineering application.

Data Availability

The data used to support the findings of this study are available from the corresponding author upon request.

Conflicts of Interest

The authors declare that there are no conflicts of interest regarding the publication of this paper.

References

[1] G. Tang, X. Wang, and Y. He, "A novel method of fault diagnosis for rolling bearing based on dual tree complex wavelet packet transform and improved multiscale permutation entropy," *Mathematical Problems in Engineering*, vol. 2016, Article ID 5432648, 13 pages, 2016.

[2] R. Liu, B. Yang, E. zio, and X. Chen, "Artificial intelligence for fault diagnosis of rotating machinery: a review," *Mechanical Systems and Signal Processing*, vol. 108, pp. 33–47, 2018.

[3] W. Caesarendra, M. Pratama, B. Kosasih et al., "Parsimonious network based on a fuzzy inference system (PANFIS) for time series feature prediction of low speed slew bearing prognosis," *Applied Sciences*, vol. 8, no. 12, 2018.

[4] A. Glowacz, "Acoustic fault analysis of three commutator motors," *Mechanical Systems and Signal Processing*, vol. 133, 2019.

[5] A. Glowacz, "Fault detection of electric impact drills and coffee grinders using acoustic signals," *Sensors*, vol. 19, no. 2, 2019.

[6] J. Zhang, Y. Zhao, X. Li, and M. Liu, "Bearing Fault diagnosis with kernel sparse representation classification based on adaptive local iterative filtering-enhanced multiscale entropy features," *Mathematical Problems in Engineering*, vol. 2019, Article ID 7905674, 17 pages, 2019.

[7] M.-a. Li, H.-n. Liu, W. Zhu et al., "Applying improved multiscale fuzzy entropy for feature extraction of MI-EEG," *Applied Sciences*, vol. 7, no. 1, 2017.

[8] R. Yan and R. X. Gao, "Approximate entropy as a diagnostic tool for machine health monitoring," *Mechanical Systems and Signal Processing*, vol. 21, no. 2, pp. 824–839, 2007.

[9] J. S. Richman and J. R. Moorman, "Physiological time-series analysis using approximate entropy and sample entropy," *American Journal of Physiology-Heart and Circulatory Physiology*, vol. 278, no. 6, Article ID H2039, 2000.

[10] C. Bandt and B. Pompe, "Permutation entropy: a natural complexity measure for time series," *Physical Review Letters*, vol. 88, no. 17, pp. 1741021–1741024, 2002.

[11] J. Xingmeng, W. Li, P. Liwu et al., "Rolling Bearing Fault diagnosis based on ELCD permutation entropy and RVM," *Journal of Engineering*, vol. 2016, Article ID 1308108, 7 pages, 2016.

[12] H. Azami and J. Escudero, "Amplitude-aware permutation entropy: illustration in spike detection and signal segmentation," *Computer Methods and Programs in Biomedicine*, vol. 128, pp. 40–51, 2016.

[13] Y. S. Chen, T. H. Zhang, W. J. zhao et al., "Fault Diagnosis of rolling bearing using multiscale Amplitude-aware permutation entropy and random forest," *Algorithms*, vol. 12, no. 9, p. 184, 2019.

[14] M. Costa, A. L. Goldberger, and C.-K. Peng, "Multiscale entropy analysis of complex physiologic time series," *Physical Review Letters*, vol. 89, no. 6, 2002.

[15] Y. Li, G. Li, Y. Wei, B Liu, and X Liang, "Health condition identification of planetary gearboxes based on variational mode decomposition and generalized composite multi-scale symbolic dynamic entropy," *ISA Transactions*, vol. 81, pp. 329–341, 2018.

[16] Y. Wei, Y. Li, M. Xu, and W. Huang, "Intelligent Fault Diagnosis of rotating machinery using ICD and generalized composite multi-scale fuzzy entropy," *IEEE Access*, vol. 7, pp. 38983–38995, 2019.

[17] Y. Chen, T. Zhang, W. Zhao, Z. Luo, and H. Lin, "Rotating machinery Fault Diagnosis based on improved multiscale Amplitude-aware permutation entropy and multiclass relevance vector machine," *Sensors*, vol. 19, no. 20, p. 4542, 2019.

[18] W. B. Zhang and J. Z. Zhou, "A comprehensive Fault Diagnosis method for rolling bearings based on refined composite multiscale dispersion entropy and fast ensemble empirical mode decomposition," *Entropy*, vol. 21, no. 7, p. 680, 2019.

[19] S. Wan and X. Zhang, "Teager energy entropy ratio of wavelet packet transform and its application in Bearing Fault diagnosis," *Entropy*, vol. 20, no. 5, p. 388, 2018.

[20] J. S. Chen, D. J. Yu, J. S. Tang, and Y. Yang, "Application of SVM and SVD technique based on EMD to The fault diagnosis of the rotating machinery," *Shock and Vibration*, vol. 16, no. 1, pp. 89–98, 2009.

[21] N. G. Kingsbury, "The dual-tree complex wavelet transform: a new technique for shift invariance and directional filters," in *Proceedings of the 8th IEEE Digital Signal Processing Workshop*, 1998.

[22] I. Bayram and I. W. Selesnick, "On the dual-tree complex wavelet packet and *M*-Band transforms," *IEEE Transactions on Signal Processing*, vol. 56, no. 6, pp. 2298–2310, 2008.

[23] M.-F. Ge, Z. Ge, H. Pan et al., "A deep condition feature learning approach for rotating machinery based on MMSDE and optimized SAEs," *Measurement Science and Technology*, 2020, In press.

[24] Y. Yu and C. Junsheng, "A roller bearing fault diagnosis method based on EMD energy entropy and ANN," *Journal of Sound and Vibration*, vol. 294, no. 1-2, pp. 269–277, 2006.

[25] H. Ao, J. Cheng, K. Li, and T. K. Truong, "A roller Bearing Fault diagnosis method based on LCD energy entropy and ACROA-SVM," *Shock and Vibration*, vol. 2014, Article ID 825825, 12 pages, 2014.

[26] G. E. Hinton, S. Osindero, and Y. W. The, "A fast learning algorithm for deep belief nets," *Neural Comput," Neural Computation*, vol. 18, no. 7, pp. 1257–2544, 2006.

[27] W. Deng, H. Liu, J. Xu et al., "An improved quantum-inspired differential evolution algorithm for deep belief network," *IEEE Transactions on Instrumentation and Measurement*, vol. 69, no. 10, pp. 7319–7327, 2020.

[28] W. Deng, J. Xu, and H. Zhao, "An improved ant colony optimization algorithm based on hybrid strategies for scheduling problem," *IEEE Access*, vol. 7, pp. 20281–20292, 2019.

[29] H. Zhao, H. Liu, J. Xu, and W. Deng, "Performance prediction using high-order differential mathematical morphology gradient spectrum entropy and extreme learning machine," *Ieee Transactions on Instrumentation and Measurement*, vol. 69, no. 7, pp. 4165–4172, 2020.

[30] H. Zhao, J. Zheng, W. Deng, and Y. Song, "Semi-supervised broad learning system based on manifold regularization and broad network," *IEEE Transactions on Circuits and Systems I: Regular Papers*, vol. 67, no. 3, pp. 983–994, 2020.

[31] J. Gai, J. Shen, H. Wang, and Y. Hu, "A parameter-optimized DBN using GOA and its application in fault diagnosis of gearbox," *Shock and Vibration*, vol. 2020, Article ID 4294095, 11 pages, 2020.

[32] Q. Tong, J. Cao, B. Han et al., "A fault diagnosis approach for rolling element bearings based on dual-tree complex wavelet packet transform-improved intrinsic time-scale decomposition, singular value decomposition, and online sequential extreme learning machine," *Advances in Mechanical Engineering*, vol. 9, no. 12, 2017.

[33] S. Zhou, S. Qian, W. Chang et al., "A novel bearing multi-Fault Diagnosis approach based on weighted permutation entropy and an improved SVM ensemble classifier," *Sensors*, vol. 18, no. 6, 2018.

[34] B. Fadlallah, B. Chen, A. Keil et al., "Weighted-permutation entropy: a complexity measure for time series incorporating amplitude information," *Physical Review E: Statistical, Nonlinear, and Soft Matter Physics*, vol. 87, no. 2, Article ID 022911, 2013.

[35] G. Hinton, "A practical guide to training restricted Boltzmann machines," *Momentum*, vol. 9, pp. 1–21, 2010.

[36] Case Western Reserve University, *Bearing Data Center*, Case Western Reserve University, Cleveland, OH, USA, 2019, http://csegroups.case.edu/bearingdatacenter/pages/download-data-fifile.

[37] J. D. Zheng, T. Liu, R. Meng et al., "Generalized composite multiscale permutation entropy and PCA based fault diagnosis of rolling bearings," *Journal of Vibration and Shock*, vol. 37, no. 20, pp. 61–66, 2018, in Chinese.

[38] Y. Ye, Y. Zhang, Q. Wang, Z. Wang, Z. Teng, and H. Zhang, "Fault diagnosis of high-speed train suspension systems using multiscale permutation entropy and linear local tangent space alignment," *Mechanical Systems and Signal Processing*, vol. 138, Article ID 106565, 2020.

[39] X. Zhang, Y. Liang, J. Zhou, and Y. zang, "A novel bearing fault diagnosis model integrated permutation entropy, ensemble empirical mode decomposition and optimized SVM," *Measurement*, vol. 69, pp. 164–179, 2015.

[40] Q. Tong, J. Cao, B. Han et al., "a fault diagnosis approach for rolling element bearings based on RSGWPT-LCD bilayer screening and extreme learning machine," *IEEE Access*, vol. 5, pp. 5515–5530, 2017.

Hindawi
Shock and Vibration
Volume 2020, Article ID 8893381, 7 pages
https://doi.org/10.1155/2020/8893381

Research Article

Estimating Network Flowing over Edges by Recursive Network Embedding

Liqun Yu,[1] **Hongqi Wang**,[1] **and Haoran Mo**[2]

[1]*School of Economics and Management, Harbin University of Science and Technology, Harbin 150080, China*
[2]*Innopolis University, Innopolis, Russia*

Correspondence should be addressed to Liqun Yu; yuliqun@hrbust.edu.cn

Received 29 September 2020; Revised 21 October 2020; Accepted 2 November 2020; Published 26 November 2020

Academic Editor: Chaoqun Duan

In this paper, we propose a novel semisupervised learning framework to learn the flows of edges over a graph. Given the flow values of the labeled edges, the task of this paper is to learn the unknown flow values of the remaining unlabeled edges. To this end, we introduce a value amount hold by each node and impose that the amount of values flowing from the conjunctive edges of each node to be consistent with the node's own value. We propose to embed the nodes to a continuous vector space so that the embedding vector of each node can be reconstructed from its neighbors by a recursive neural network model, linear normalized long short-term memory. Moreover, we argue that the value of each node is also embedded in the embedding vectors of its neighbors, thus propose to approximate the node value from the output of the neighborhood recursive network. We build a unified learning framework by formulating a minimization problem. To construct the learning problem, we build three subproblems of minimization: (1) the embedding error of each node from the recursive network, (2) the loss of the construction for the amount of value of each node, and (3) the difference between the value amount of each node and the estimated value from the edge flows. We develop an iterative algorithm to learn the node embeddings, edge flows, and node values jointly. We perform experiments based on the datasets of some network data, including the transportation network and innovation. The experimental results indicate that our algorithm is more effective than the state-of-the-arts.

1. Introduction

1.1. Background. Learning the flow direction and amount of the edges for a network has been a critical problem in the network analysis. The edges connect two nodes, and the flow is defined from one node to another connected by the corresponding edge [1–3]. This problem is called the edge flow estimation. In this problem, we already know the network structure, including the sets of the nodes, and the edges between the nodes. We also know the directions and amounts of the flows of some edges, but for the rest edges, the flows are still unknown. The target is to predict the flow direction and amounts of these edges. Some examples of the applications of the edge estimation are given as follows:

(i) For example, in the transportation network, each intersection is a node, and each road is an edge, while the flows along the roads need to be estimated for the purpose of traffic control. According to the historical data of the traffic flows, we know the flows of some roads; however, other read flows need to be estimated. This use case raises the problem of learning edge flows from both the traffic network and existing flows of roads of the network [4, 5].

(ii) Another example is the innovation network analysis, where each research article is a node, and each citation is an edge between two articles. For the task of innovation, it is important to know the flowing of knowledge from an article to other articles. Please note that one article cites another one that does not necessarily mean there is an amount of "knowledge" flowing from the cited article to the citing article. By reading the content of the article which cites the others, we can annotate the "knowledge flowing" if one work is inspired or a following work of the other research works. However, such annotation by reading is time-consuming and subject to individual annotators, thus it is very necessary to develop an automatic system to estimate the knowledge flow from the article citation network [6–14].

Although the edge flow prediction/estimation is such a critical problem, the works on this direction are very few. Most recently, Ransom et al. [1] designed a new algorithm to predict the edge flow by balancing the amount of flows moving into and out of the nodes. Another condition of this algorithm is that the predicted flow of edges should be consistent to the known flows, for these edges whose flows are already known. In this paper, we proposed a novel edge flow estimation algorithm based on both semisupervised learning and network embedding.

1.2. Our Contributions. In this paper, we build a novel method for the problem of edge flow prediction. This method is both for network embedding and edge flow prediction. The two problems are solved at the same time. We designed a new framework for the learning problem. In this framework, for the first time, we bridge the node embedding and edge flow estimation, by introducing a node value for each node. On the one side, the node value is used to balance the flowing-in and flowing-out amount of a node. It plays the role of measuring the balance of the amount of the node at any moment of the flowing process, i.e., with the incoming flow and outgoing flow changing, the amount of the value in this node remains as the node value. One the other side, the node is employed to regularize the learning of the embedding vectors, thus we impose that the node value can be estimated from the embedding vector by a linear function.

We propose a novel algorithm to learn the embedding vectors and edge flows simultaneously. We model the learning problem as a minimization problem where the embedding vectors reconstruction error of the LSTM embedding model, the node value estimation error from the embedding vectors, and the node value flow amount error are minimized. In the iterative algorithm, the node embeddings, node values, and the edge flow amounts are optimized alternately.

We evaluated the proposed method over benchmark datasets of networks, conducted experiments to reveal the properties of the proposed algorithms, and show its advantage over state-of-the-art methods.

Remark: the superiority of the proposed semisupervised edge flow learning method compared with the traditional semisupervised learning method is listed as follows:

(1) The traditional semisupervised learning methods can only predict the labels of the nodes but are not able to predict the flows of the edges. For the applications discussed in this paper, the traditional semisupervised learning methods are not suitable. Our method is especially designed for these applications.

(2) Traditional semisupervised learning methods can only use the edge information of a graph but ignore the flow information, which is critical for both the node and flow label prediction. However, our method can effectively use them to learn better node embeddings and flow amounts.

1.3. Paper Origination. This paper is organized as follows. In Section 2, we introduce the joint learning framework, with an objective and optimization solution. In Section 3, we experimentally evaluate the performance of the proposed method and conduct studies over its properties. In Section 4, we conclude this paper with some future works.

2. Proposed Method

In this section, we introduce the proposed joint network embedding and edge flow learning framework. In this framework, we propose a deep learning-based network embedding method [15–23] and further use the embedding vectors to estimate the flow amount regarding each node [15]. The flow amount is used to regularize the edge flow learning process.

2.1. Problem Definition. Suppose we have input network, denoted as $G = \{V, E\}$, where $V = \{1, \ldots, n\}$ is the set of n nodes, and $\varepsilon_k = (i, j) \in E$ is an edge linking the i-th and j-th nodes. Here, we assume that $i < j$. For a group of edges, we already know their flows. This set of edges is denoted as $\varepsilon_k \in E_L$. For such an edge ε_k, we define its flow as $f_k \in R$. The direction of the flow is the sign of f_k, and the amount of the flow is the absolute value of f_k. For the other edges, the flows are unknown, and we want to predict them. For these flows, we define a set as E_U. We define a vector as the flows of all the nodes, as $f = [f_1, \ldots, f_{|E|}]^T \in R^{|E|}$. Thus, the prediction of the flows of the nodes is transformed to the solving of f.

2.2. Problem Modeling. We firstly embed each node $i \in V$ to a vector. The dimension of the vector is denoted as d. Moreover, for each node, we define a node value, ϕ_i. We also propose to calculate the node value from the node's embedding vector. After ϕ_i is given, we use it to regularize the estimation of the flow of edges connected to the node. The flow chart of the proposed semisupervised edge flow learning method is given in Figure 1.

2.2.1. Recursive Node Embedding. The network embedding converts the nodes to a set of vectors, denoted as $x_1, \ldots, x_n$, where $x_i \in R^d$. To this end, we want to reconstruct the embedding vector of one node from its linked nodes. The linked nodes are presented as a sequence of nodes. Accordingly, the sequence of neighbouring nodes of the i-th node are given as

$$N_i = \{j | (i, j) \in E \text{ or } (j, i) \in E\}. \quad (1)$$

We firstly sort the nodes in the neighbouring set, N_i, and the sorted set's embedding vector set is

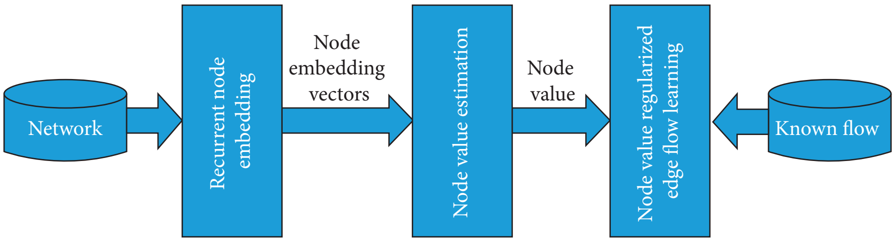


Figure 1: Flow chart of proposed semisupervised flow learning method.

$$S_i = \left\{ x_{i,l}, \ldots, x_{i|Ni|} \right\}. \tag{2}$$

We apply aln-LSTM to this sequence of embedding vectors. In this model, for an input node's embedding vector, $x_t \in S_i$, its inputs are x_t itself and the output of its previous h_{t-1},

$$h_t = g\left(x_t, h_{t-1}; \theta\right), \quad t = 1, \ldots, |\mathrm{Ni}|, \tag{3}$$

where g is the cell function of ln-LSTM and θ is its parameter.

Then, we try to use the output of the cell function from the last node to approximate the embedding vector of the i-th node itself,

$$\mathrm{LSTM}\left(x_{il}, \ldots, x_{i|Ni|}; \theta\right) = h_{|Ni|}. \tag{4}$$

Thus, the minimization of the error of the approximation is modelled as

$$\min_{x|_{i=1}^{n}, \theta} \sum_{i=1}^{n} \left\| x_i - \mathrm{LSTM}\left(x_{i1}, \ldots, x_{i|Ni|}; \theta\right) \right\|_F^2. \tag{5}$$

By solving this problem, the learning of the embedding vector and the parameter of the cell function are solved together. With the good quality embedding and cell function parameter, we should be able to approximate the embedding vectors for the neighbouring nodes's embedding vectors.

2.2.2. Node Value Estimation from Recursive Embedding.

In out learning framework, the embedding of each node approximated by the recursive model plays two roles, representing the node's neighbor structure and estimating the amount of value hold by the nodes. We define a node value for the i-th node, $\phi_i \in R$. The function of this value is to balance the incoming and outgoing flow of the node. Moreover, it can somehow measure the nature of the node. For example, a node in traffic network at the working time may have more incoming flow, because it is an office area. This value can indicate this node's nature of being in office area. Moreover, we want to use the embedding vector of the node to calculate the value as follows:

$$\phi_i \longleftarrow \varphi\left(\mathrm{LSTM}\left(x_{il}, \ldots, x_{i|Ni|}; \theta\right)\right), \tag{6}$$

where $\varphi(x) = \sigma(w^T h)$ is the function of node value approximation. This function is actually a single-layer neural network. We proposed to minimize the square error of the approximation over all the nodes,

$$min_{x|_{i=1}^{n}, \theta, w} \sum_{i=1}^{n} \left\| \phi_i - \varphi\left(LSTM\left(x_{i1}, \cdots, x_{i|Ni|}; \theta\right)\right) \right\|_F^2. \tag{7}$$

The minimization is performed regarding the node value amount, embedding vectors, recursive model parameters, and the single-layer neural network parameter. In this way, we bridge the learning of the flow amount and embedding of nodes by using an LSTM recursive model.

2.2.3. Flow Prediction by Using Node Value.

Given an edge $\varepsilon_k \in E$, we want to predict its flow f_k. To this end, we use the node value as reference. Our assumption is that for any node, the value of the node is controlling the incoming and outgoing flow. To be more specific, for a node, the difference between its incoming flow and outgoing flow should be equal to the node value itself. For this purpose, we define two sets of edges for a node $R_i^+ = \{\varepsilon_k \mid \varepsilon_k \in E,\ \varepsilon_k = (u, i)\}$ and $R_i^- = \{\varepsilon_k \mid \varepsilon_k \in E,\ \varepsilon_k = (i, v)\}$. The amount of the incoming flow is $\sum_{k \in R_i^+} f_k - \sum_{k \in R_i^-} f_k$. Since we hope this amount is equal to the node value,

$$\begin{gathered} \phi_i = \sum_{k \in R_i^+} f_k - \sum_{k \in R_i^-} f_k = \sum_{k=1}^{|E|} \tau_{ik} f_k, \\ \text{where} \\ \tau_{ik} = \begin{cases} +1, & \text{if } k \in R_i^+, \\ -1, & \text{if } k \in R_i^-, \\ 0, & \text{otherwise.} \end{cases} \end{gathered} \tag{8}$$

We also denote the node value of all nodes in a vector $\phi = [\phi_1, \ldots, \phi_n]^T \in R^n$. A matrix is also proposed as $\Phi = [\tau_{ik}] \in \{+1, -1, 0\}^{n \times |E|}$. It is the node-flow mapping matrix. So, equation (8) is transformed to

$$\phi = \Phi f. \tag{9}$$

We minimize the squared approximation error regarding both the edge flow vector and the node value vector,

$$\min_{f, \phi} \|\phi - \Phi f\|_F^2. \tag{10}$$

The final objective and minimization problem is the combination of (5), (7), and (10) as follows:

$$\begin{aligned}
&\min_{f,\phi,x_i|_{i=1}^{n},\theta,w}\left\{\sum_{i=1}^{n}\left\|x_i-\mathrm{LSTM}\left(x_{i1},\cdots,x_{i|Ni|};\theta\right)\right\|_F^2\right.\\
&+\lambda_1\sum_{i=1}^{n}\left\|\phi_i-\varphi\left(\mathrm{LSTM}\left(x_{i1},\cdots,x_{i|Ni|};\theta\right)\right)\right\|_F^2+\lambda_2\|\phi-\Phi f\|_F^2\\
&\left.+\lambda_3\left(\|f\|_F^2+\|\phi\|_F^2+\sum_{i=1}^{n}\left\|x_i\right\|_F^2+\|\theta\|_F^2+\|w\|_F^2\right)\right\}\\
&s.t.\quad f_k=\overline{f}_k,\forall k:\ \varepsilon_k\in E_L.
\end{aligned}\tag{11}$$

Here, we also add the l_2 norm regularization terms to the parameters to prevent the over-fitting problem.

2.3. Problem Solution. The solving of the minimization problem is conducted in an iterative algorithm. In each iteration, every parameter is updated sequentially. The others are not updated when one parameter is being updated. In the following subsections, we introduce the optimization of each parameter one by one, while the others are fixed.

2.3.1. Optimization of LSTM Parameters, θ. When the other parameters are fixed, and only the LSTM model parameters are considered, we have the following minimization problem:

$$\min_\theta\left\{o_1(\theta)=\sum_{i=1}^{n}\left\|x_i-\mathrm{LSTM}\left(x_{i1},\ldots,x_{i|Ni|};\theta\right)\right\|_F^2+\lambda_1\sum_{i=1}^{n}\left\|\phi_i-\varphi\left(\mathrm{LSTM}\left(x_{i1},\ldots,x_{i|Ni|};\theta\right)\right)\right\|_F^2\right\},\tag{12}$$

which can be solved by the ADAM algorithm. To use the ADAM algorithm, we calculate the gradient function of the objective $o_1(\theta)$ regarding θ as follows:

$$\begin{aligned}
\nabla_\theta o_1(\theta)=&-2\sum_{i=1}^{n}\left(x_i-\mathrm{LSTM}\left(x_{i1},\ldots,x_{i|Ni|};\theta\right)\right)\times\frac{\partial\,\mathrm{LSTM}\left(x_{i1},\ldots,x_{i|Ni|};\theta\right)}{\partial\,\theta}\\
&-2\lambda_1\sum_{i=1}^{n}\left(\phi_i-\varphi\left(\mathrm{LSTM}\left(x_{i1},\ldots,x_{i|Ni|};\theta\right)\right)\right)\times\frac{\partial\varphi\left(\mathrm{LSTM}\left(x_{i1},\ldots,x_{i|Ni|};\theta\right)\right)}{\partial\mathrm{LSTM}\left(x_{i1},\ldots,x_{i|Ni|};\theta\right)}\times\frac{\partial\mathrm{LSTM}\left(x_{i1},\ldots,x_{i|Ni|};\theta\right)}{\partial\theta}+2\lambda_3\theta,
\end{aligned}\tag{13}$$

where $\partial f(x)/\partial x$ is the derivative function of f regarding variable x.

2.3.2. Optimization of Embedding Vectors, $\mathbf{x_i}|_{\mathbf{i=1}}^{\mathbf{n}}$. When the other variables are fixed and only the embedding vectors are considered, we have the following minimization problem:

$$\min_{x_i|_{i=1}^{n}}\left\{\sum_{i=1}^{n}\left\|x_i-\mathrm{LSTM}\left(x_{i1},\ldots,x_{i|Ni|};\theta\right)\right\|_F^2+\lambda_1\sum_{i=1}^{n}\left\|x_i-\mathrm{LSTM}\left(x_{i1},\ldots,x_{i|Ni|};\theta\right)\right\|_F^2+\lambda_3\sum_{i=1}^{n}\left\|x_i\right\|_F^2\right\}.\tag{14}$$

To solve the vectors, we apply the sequential optimation method and optimize the embedding vectors of the nodes one by one. When one node embedding vector is optimized, others are fixed. When the i-th vector, x_i, is considered, we have the following problem for minimization:

$$\begin{aligned}
&\min_{x_i}\left\{o_2(x_i)=\left\|x_i-\mathrm{LSTM}\left(x_{i1},\ldots,x_{i|Ni|};\theta\right)\right\|_F^2\right.\\
&\left.+\sum_{j:\ i\in N_i}\left\|x_j-\mathrm{LSTM}\left(x_{j1},\ldots,x_{j|Ni|};\theta\right)\right\|_F^2+\lambda_1\sum_{j:\ i\in N_i}\left\|\phi_i-\varphi\left(\mathrm{LSTM}\left(x_{i1},\ldots,x_{i|Ni|};\theta\right)\right)\right\|_F^2+\lambda_3\left\|x_i\right\|_F^2\right\}.
\end{aligned}\tag{15}$$

Table 1: Dataset summary.

Dataset	# Node	# Edge	Node	Edge	Reference
Minnesota road network	2642	3303	Intersection	Road	[19]
US power grid network of KONECT	4941	6593	Consumer	Transmission line	[24]
Water irrigation network of Balerma, Spain	447	454	Water supply/hydrant	Water pipe	[25]
Innovation flow network	10782	39741	Author	Directed citation link	[26, 27]

Again, we use the ADAM algorithm to solve this problem similar to the optimization of θ.

2.3.3. Optimization of Edge Flow Vector, f. When the edge flow vector is considered, we have the following minimization problem:

$$\begin{aligned} &\min_{f}\left\{o_3(f) = \lambda_2\|\phi - \Phi f\|_F^2 + \lambda_3\|f\|_F^2\right\} \\ &s.t.\ f_k = \overline{f}_k, \forall k:\ \varepsilon_k \in E_L. \end{aligned} \quad (16)$$

This is a linear constrained quadratic programming problem; we employ the active-set algorithm to solve it.

2.3.4. Optimization of Node Value Vector, ϕ. To solve the node value vector, we fix the other variables and have the following minimization problem:

$$\min_{\phi}\left\{o_4(\phi) = \lambda_2\|\phi - \Phi f\|_F^2 + \lambda_3\|\phi\|_F^2\right\}. \quad (17)$$

We set the derivative of o_4 regarding ϕ to zero and have the solution of ϕ as

$$\begin{aligned} &\frac{\partial o_4(\phi)}{\partial \phi} = 2\lambda_2(\phi - \Phi f) + 2\lambda_3\phi = 0 \\ &\Longrightarrow \phi = \frac{\lambda_2}{\lambda_2 + \lambda_3}\Phi f. \end{aligned} \quad (18)$$

2.3.5. Optimization of w. To solve the parameter of function ϕ, we have the following minimization problem:

$$\min_{w}\left\{\lambda_1 \sum_{i=1}^{n}\left\|\phi_i - \varphi\left(LSTM\left(x_{i1}, \cdots, x_{i|Ni|}; \theta\right)\right)\right\|_F^2 + \lambda_3\|w\|_F^2\right\}. \quad (19)$$

We still employ the ADAM algorithm to solve this minimization problem.

3. Experiments

In this section, we give the experimental setting and results. The algorithm tested and developed is called edge flow prediction by network embedding (EFPNF).

3.1. Benchmark Dataset and Experimental Setting. Four datasets of network are used in our setting. We summarize the datasets in Table 1.

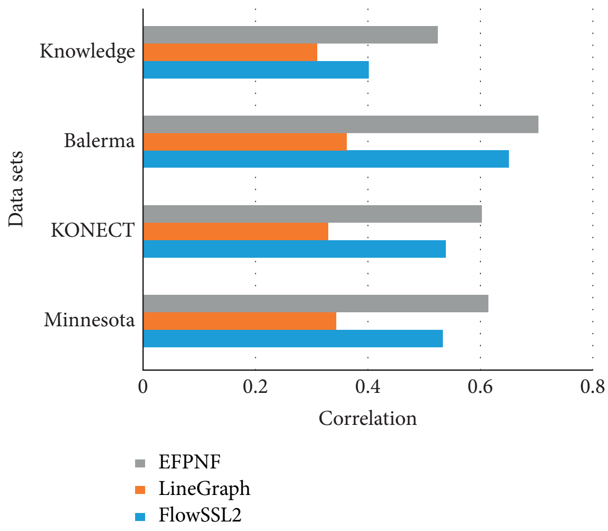


Figure 2: Comparison results over state-of-the-arts.

The 10-fold cross-validation is used to generate the training and test set. The Pearson correlation coefficient is used to measure the quality of the flow predicted by the algorithm [28].

3.2. Experimental Results

3.2.1. Comparison to State-of-the-Arts. As shown in Figure 2, the algorithm is compared to the others which are also used to predict the flows of edges. The compared algorithms are flow SSL algorithm proposed by Jia et al. [1] and the LineGraph algorithm. According to the figure, the proposed algorithm EFPNF has better performance for all four datasets than the compared algorithms. Especially for the Balerma dataset, the EFPNF is the only algorithm which has the correlation score larger than 0.7. We also can see that the knowledge network dataset is the hardest task, and the EFPNF still gives the best performance.

3.2.2. Convergence Analysis. Since our algorithm is an iterative algorithm, we are also interested in the convergence of the algorithm. Thus, we plot the curve of correlation while the number of iterations is increasing (Figure 3). We can see from these curves that our algorithm's performance is boosted when the iteration number is increased from 5 to 20. Since our algorithm aims to minimize the objective function, more iteration numbers result in a smaller value of the objective. This verifies the effectiveness of the objective to

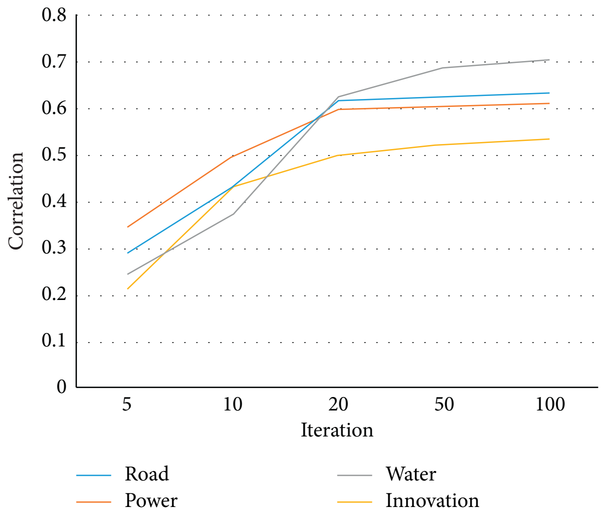


Figure 3: Convergence curve.

achieve a better edge flow estimation performance. However, when the iteration number further increases from 20 to 100, the performance improvement is not significant, meaning our algorithm does not need a large number of iterations to give a good performance.

4. Conclusion

We developed an iterative algorithm to learn the node embeddings and missing flows of the edges for a network. This algorithm is based on the deep recurrent network for the embedding purpose. Moreover, it uses the embeddings to calculate the node values and further uses the node values to approximate the flows around the node. A unified learning framework is built for the learning of embedding network, node value, and flows of the edges. The learning process is guided by the minimization of the reconstruction errors of embedding vectors, node values, and incoming/outgoing flows. Experimental results show the advantage of the proposed algorithm.

Data Availability

All the datasets used in this paper to produce the experimental results are publicly accessed online.

Conflicts of Interest

The authors of this paper claim no conflicts of interest regarding the work reported in this paper.

Acknowledgments

This paper was funded by the National Natural Science Foundation of China (Project nos. 71704036 and 71473062); Social Science Foundation of Ministry of Education of China (Project no. 19YJA790087); and the Talents Plan of Harbin University of Science and Technology: Outstanding Youth Project (Project no. 2019-KYYWF-0216).

References

[1] P. Russom, “Big data analytics,” *TDWI Best Practices Report*, vol. 19, no. 4, pp. 1–34, 2011.

[2] M. Chen, S. Mao, and Y. Liu, “Big data: a survey,” *Mobile Networks and Applications*, vol. 19, no. 2, pp. 171–209, 2014.

[3] G. Liang, H. Mo, Z. Wang, C. Q. Dong, and J. Y. Wang, “Joint deep recurrent network embedding and edge flow estimation,” in *Proceedings of the International Conference on Intelligent Computing*, pp. 467–475, Bari, Italy, October 2020.

[4] E. Alpaydin, *Introduction to Machine Learning*, MIT press, Cambridge, MA, USA, 2020.

[5] X. Zhu and A. B. Goldberg, “Introduction to semi-supervised learning,” *Synthesis Lectures on Artificial Intelligence and Machine Learning*, vol. 3, no. 1, pp. 1–130, 2009.

[6] X. J. Zhu, *Semi-supervised Learning Literature Survey. Technical Report*, University of Wisconsin-Madison Department of Computer Sciences, Madison, WI, USA, 2005.

[7] S. J. Pan and Q. Yang, “A survey on transfer learning,” *IEEE Transactions on Knowledge and Data Engineering*, vol. 22, no. 10, pp. 1345–1359, 2009.

[8] M. E. Taylor and P. Stone, “Transfer learning for reinforcement learning domains: a survey,” *Journal of Machine Learning Research*, vol. 10, no. Jul, pp. 1633–1685, 2009.

[9] V. Mahadevan, W. Li, V. Bhalodia, and N. Vasconcelos, “Anomaly detection in crowded scenes,” in *Proceedings of the 2010 IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, pp. 1975–1981, IEEE, San Francisco, CA, USA, June 2010.

[10] S. Bansod and A. Nandedkar, “Transfer learning for video anomaly detection,” *Journal of Intelligent and Fuzzy Systems*, vol. 36, no. 3, pp. 1967–1975, 2019.

[11] T. U. Haque, N. N. Saber, and F. M. Shah, “Sentiment analysis on large scale amazon product reviews,” in *Proceedings of the 2018 IEEE International Conference on Innovative Research and Develop- Ment (ICIRD)*, pp. 1–6, IEEE, Bangkok, Thailand, May 2018.

[12] A. Bhatt, A. Patel, H. Chheda, and K. Gawande, “Amazon review classification and senti- ment analysis,” *International Journal of Computer Science and Information Technologies*, vol. 6, no. 6, pp. 5107–5110, 2015.

[13] T. N. Sainath, A. R.. Mohamed, B. Kingsbury, and B. Ramabhadran, “Deep convolutional neural networks for lvcsr,” in *Proceedings of the 2013 IEEE International Conference on Acoustics, Speech and Signal Processing*, pp. 8614–8618, IEEE, Vancouver, Canada, May 2013.

[14] M. Yamada, L. Sigal, M. Raptis, M. Toyoda, Y. Chang, and M. Sugiyama, “Cross-domain matching with squared-loss mutual information,” *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 37, no. 9, pp. 1764–1776, 2015.

[15] G. Zhang, G. Liang, F. Su, F. Qu, and J. Y. Wang, “Cross-domain attribute representa- tion based on convolutional neural network,” in *Proceedings of the International Conference on Intelligent Computing*, pp. 134–142, Springer, Wuhan, China, August 2018.

[16] Y. Geng, R. Z. Liang, W. Li et al., “Learning convolutional neural network to maximize pos@ top performance measure,” in *Proceedings of the ESANN 2017*, pp. 589–594, Bruges, Belgium, April 2016.

[17] G. Zhang, G. Liang, W. Li et al., “Learning con- volutional ranking-score function by query preference regularization,” in *Proceedings of the International Conference on Intelligent Data Engineering and Automated Learning*, pp. 1–8, Springer, Madrid, Spain, November 2017.

[18] Y. Geng, G. Zhang, W. Li et al., "A novel image tag completion method based on convolutional neural transformation," in *Proceedings of the International Conference on Artificial Neural Networks*, pp. 539–546, Springer, Sardinia, Italy, September 2017.

[19] K. Wang, J. Liu, and J. Y. Wang, "Learning domain-independent deep representations by mutual information minimization," *Computational Intelligence and Neuroscience*, vol. 2019, Article ID 9414539, 14 pages, 2019.

[20] M. Xiao and Y. Guo, "Feature space independent semi-supervised domain adaptation via kernel matching," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 37, no. 1, pp. 54–66, 2014.

[21] M. Long, Y. Cao, Z. Cao, J. Wang, and M. I. Jordan, "Transferable representation learning with deep adaptation networks," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 41, no. 12, pp. 3071–3085, 2018.

[22] L. Duan, I. W. Tsang, and D. Xu, "Domain transfer multiple kernel learning," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 34, no. 3, pp. 465–479, 2012.

[23] M. Long, H. Zhu, J. Wang, and M. I. Jordan, "Deep transfer learning with joint adaptation networks," in *Proceedings of the 34th International Conference on Machine Learning*, vol. 70, pp. 2208–2217, Sydney, Australia, August 2017.

[24] W. Ge and Y. Yu, "Borrowing treasures from the wealthy: deep transfer learning through selective joint fine-tuning," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 1086–1095, Honolulu, HI, USA, July 2017.

[25] T. Suzuki and M. Sugiyama, "Sufficient dimension reduction via squared-loss mutual in- formation estimation," in *Proceedings of the Thirteenth International Conference on Artificial Intelligence and Statistics*, pp. 804–811, Sardinia, Italy, May 2010.

[26] G. Niu, W. Jitkrittum, B. Dai, H. Hachiya, and M. Sugiyama, "Squared-loss mutual in- formation regularization: a novel information-theoretic approach to semi-supervised learning," in *Proceedings of the International Conference on Machine Learning*, pp. 10–18, Atlanta, USA, June 2013.

[27] S. Boyd, N. Parikh, E. Chu, B. Peleato, and J. Eckstein, "Distributed optimization and statistical learning via the alternating direction method of multipliers," *Foundations and Trends ® in Machine Learning*, vol. 3, no. 1, pp. 1–122, 2011.

[28] B. Jacob, J. Chen, Y. Huang, and I. Cohen, "Pearson correlation coefficient," in *Noise Reduction in Speech Processing*, pp. 1–4, Springer, Berlin, Germany, 2009.

Hindawi
Shock and Vibration
Volume 2020, Article ID 8816453, 21 pages
https://doi.org/10.1155/2020/8816453

Research Article

Multiphysics Modeling of Gas Turbine Based on CADSS Technology

Tian Wang, Ping Xi, and Bifu Hu

School of Mechanical Engineering and Automation, Beihang University, Beijing 100192, China

Correspondence should be addressed to Tian Wang; wangt703@buaa.edu.cn

Received 12 September 2020; Revised 29 September 2020; Accepted 30 September 2020; Published 23 October 2020

Academic Editor: Chaoqun Duan

Product modeling has been applied in product engineering with success for geometric representation. With the application of multidisciplinary analysis, application-driven models need specific knowledge and time-consuming adjustment work based on the geometric model. This paper proposes a novel modeling technology named computer-aided design-supporting-simulation (CADSS) to generate multiphysics domain models to support multidisciplinary design optimization processes. Multiphysics model representation was analyzed to verify gaps among different domain models' parameters. Therefore, multiphysics domain model architecture was integrated by optimization model, design model, and simulation model in consideration of domain model's parameters. Besides, CADSS uses requirement space, domain knowledge, and software technology to describe the multidisciplinary model's parameters and its transition. Depending on the domain requirements, the CADSS system extracts the required knowledge by decomposing product functions and then embeds the domain knowledge into functional features using software technology. This research aims to effectively complete the design cycle and improve the design quality by providing a consistent and concurrent modeling environment to generate an adaptable model for multiphysics simulation. This system is demonstrated by modeling turbine blade design with multiphysics simulations including computational fluid dynamics (CFD), conjugate heat transfer (CHT), and finite element analysis (FEA). Moreover, the blade multiphysics simulation model is validated by the optimization design of the film hole. The results show that the high-fidelity multiphysics simulation model generated through CADSS can be adapted to subsequent simulations.

1. Introduction

Computer-aided design (CAD) and computer-aided engineering (CAE) have been ubiquitously used in product development. However, design and analysis have been performed as two separate modules. According to product performance requirements, mechanical products' structure was designed through modeling software. The simulation analysis verifies whether the designed structure meets the requirements in dedicated simulation analysis software. Effectively integrating design and analysis processes can shorten product development cycle time and reduce the times of expensive physical experiments [1].

In general, different tools were utilized to build different models for product development in different domains during product development. The differences in both tools and expertise forced the analyst to preprocessing the CAD model, and any modifications made to the CAD model would cause an additional tedious geometry preparation [2]. Evidently, whether it is for process development or feature recognition in manufacturing automation, CAD modeling pays more attention to lower-level geometry and geometric parameters [3] and does not retain performance parameters and analysis parameters. The digital design does not pay much attention to the subsequent simulation, which is especially crucial for the preprocessing of the simulation analysis model and the modification of the simulation results [4]. The model is verified by the qualitative or quantitative performance parameter calculation of the simulation analysis. The lower-level geometric information operation is not intuitive for the analyst, and the qualitative or quantitative analysis results need the analyst to perform the corresponding conversion [5]. Subsequently, the designer can modify the CAD model, especially when it comes to the multidisciplinary fields such as dynamics, aerodynamics, structure, and thermodynamics. The modification requirements of the model are more frequent and complicated [6].

Therefore, multidisciplinary design optimization (MDO) approach is used to solve problems concerning multiple disciplines or subsystems [7]. Multiple disciplines or subsystems typically use analytical tools in their respective areas of expertise and establish a multiphysics simulation model that simulates the performance and behavior of the overall integrated design [8]. However, sophisticated engineering modeling across multiple domains is a daunting task. First, it takes a lot of time and effort to build a single discipline or domain simulation model. When it comes to multiphysics simulation model, multiple integrated information from many disciplines needs to be dealt with [9]. Second, high-fidelity geometry information and other detailed configuration parameters are required when using computational fluid dynamics (CFD) or finite element method (FEM), even conjugated heat transfer (CHT) [10]. Then, the delay in creating a single subject analysis model by manual adjustment may slow down the entire MDO process [11]. Moreover, multidisciplinary optimization typically sets one or more goals [12]. When the design goals change, all simulation models should be updated synchronously rather than manually.

Product design essentially needs to meet the structural and performance requirements at the same time [12], but the existing feature modeling technology lacks support for product structure performance analysis, and performance information cannot be reused in the feature-based modeling method, especially for multiphysics domain modeling [13]. The problem leads to repeated and cumbersome CAE analysis, which significantly affects the design efficiency. In the existing product design and analysis process, the preprocessing is time-consuming, such as the conversion of the CAD model to CAE model through simplification and dimensionality reduction method [14], defining material properties and loading boundary conditions (displacement constraints and loads) [15]. Moreover, most of the above processes directly operate geometric models based on B-rep expression.

In order to improve the efficiency of preprocessing, some scholars introduced feature technology in the field of CAE analysis and proposed the concept of analysis features. Nizar et al. [16] present a representation method for mechanical analysis features, but this method is only suitable for engineering analysis in the early conceptual design stage of the product, and it is challenging to express structural performance analysis features. Shephard et al. [17] used different parts of the CAD model as the analysis features and used them as the basic unit of the model simplification process. This method brings convenience to the simplification of the model, but it is challenging to support other processing related to the analysis. Lee et al. [18] simplified the design features [19] as an analysis feature and proposed an integrated model for the unified expression of design and analysis features. This model can derive different levels of geometry and analysis models according to different needs. This method requires the analysis features to be mapped one-to-one with the design features. However, such mappings do not necessarily exist, making it difficult to handle complex features. Aifao et al. [20] used a set of analysis processes as analysis features and proposed a physical model including information such as load distribution, a simulation model that can evaluate physical model methods, and an observation model that can quantitatively evaluate product behavior. This method can achieve the reuse of analytical features to a certain extent. However, the lack of fixed template in engineering analysis leads to low flexibility [21].

For the CAD-CAE integration solution, the basic idea is to integrate the modeling and analysis module using unified software [22]. Through the object-linking technology (OTL) of commercial software, the CAD system can directly call the built-in CAE calculation module, such as CATIA module and NX and MSC solvers. On the one side, the modeling and analysis integration interface is unified and no intermediate model conversion is required. On the other side, the analysis module integrated under the CAD system platform has limited solution ability and cannot handle complex analysis problems.

Standard exchange files can be utilized to carry out the model between platforms [23], such as DXF, STEP, and IGES. The common data model (CDM) modeling and analysis integration uses the method of customizing intermediate files and deals with the data exchange problem of modeling and simulation platform [15]. A unified integration platform can guarantee the data compatibility between the CAD model and CAE model to ensure the flexibility of selecting the modeling and simulation platform [24]. However, the standard file interface sacrificed high-level design information such as features to ensure simplicity and versatility. Resulted in loss of model data, standard files cannot meet the relationship between the CAD model and CAE model. Consequently, it is challenging to implement the analysis and verification process of frequent modification iteration automatically or semiautomatically. Constructing the intermediate platform method analyzes the CAD model and the CAE model and realizes the data exchange through the intermediate model. To a certain extent, the CAD model can be transmitted to change the associated process of the CAE model, and the proprietary algorithm and program design are realized. However, model management and data exchange rely heavily on expert experience and the development of special algorithms. These limitations result in the lack of universality in the process of design-analysis integration.

In terms of integrated framework solutions, Peak et al. [25] proposed a multirepresentation architecture (MRA) to address data sharing between CAD and CAE applications. Sudarsan et al. [26] proposed a design-analytical integration framework based on the concept of PLM, using the expression of the main model and functional model to realize the construction of the framework model. Smit et al. [4] realized the association between design and analysis models by extending the multiview feature-based modeling concept to the paradigm of the analysis domain. Kim et al. [14] proposed a simplified method based on the feature-based boundary representation (B-rep) model, which uses a sequential iterative volume decomposition. The method generates a feature-based model from the B-rep model by applying volume decomposition methods sequentially and iteratively. Gujarathi et al. [15] proposed the integration of design-analysis using a unified data model. Li et al. [27] present a practical CAD/CFD integration framework that seamlessly integrates CAD and CFD tools to facilitate the cycle product development process and simplify the CFD solver setup.

In addition, Iso-geometric Analysis (IGA) in view of tighter CAD-CAE integration considered the whole geometric model directly as the analysis model using the same basis function to express the CAD geometry [28], and the geometric model can be directly used for analysis and design based on common data [29]. Simulation calculations and model modifications can be completed in one step, instead of having to undergo complex model processing and grid discretization like the current popular finite element method [30], thereby effectively reducing the design grid discretization [31]. However, these design attempts are only experimental and have not been applied to industrial practice at present.

The above literature concerns the construction of design-analysis integration framework and the analysis feature modeling method. Few researchers pay attention to the combination of the preprocessing stage and CAD modeling for the CAE process. The application of manufacturing-oriented product models to numerical simulation relies mainly on manual processing, while the integration of CAD modeling and CAE analysis focuses on constructing an integrated framework for modeling and analysis. Dealing with modeling product in multiphysics simulation modeling is not discussed and researched.

Due to the complex dependencies and lack of management tools [32], it is challenging to guarantee model integrity and adaptability. Ideally, analysts only spend time on analysis work rather than these repetitive preprocessing tasks. A rapid parameter configuration method for two dimensions simplified model was generated for pipe-net calculation [33], and the results obtained are in good agreement with the results of the three-dimensional model simulation [34]. Therefore, in order to effectively complete the design cycle and improve the design quality, it is necessary to establish a reasonable modeling technology to ensure that the CAD model of the design can adapt to the subsequent simulation, and the analysis results of the numerical simulation can be quickly fed back to the modification of the CAD model.

This paper proposes a novel modeling technology called CADSS (computer-aided design-supporting-simulation). The CADSS model focuses on the functional realization of the product and constructs the relationship between features and parameters in design, modeling, and analysis in order to solve the problem of modeling efficiency in the design and analysis process. The CADSS features can reflect engineering knowledge, design principles, and simulation analysis requirements. Finally, as a typical application of multiphysics simulation, the design of the turbine blade was studied as a meaningful case to illustrate the application of CADSS technology.

2. Multiphysics Modeling for Simulation

2.1. Feature-Based Modeling System's Model Representation Method Analysis. The existing feature-based CAD system provides a series of modeling operations through features shown in Figure 1 and defines the relationship with Boolean features for parts or assembling between parts by assembling features. Finally, a product model is generated for simulation or manufacturing.

Based on the process, feature-based modeling technology can be generally defined as three main steps: the setting of reference, justification of dimension and constraint, and preservation of nongeometric information. Therefore, it can be expressed as follows [35]:

$$F = (g, d, m) = \left(\bigcup_{1}^{u} g_m\right) \cup \left(\bigcup_{1}^{v} d_v\right)\left(\bigcup_{1}^{k} m_k\right), \tag{1}$$

where g represents geometric information, d represents design intent, and m represents nongeometric information.

For example, to complete a hole feature design, CAD system internal function is CreateCylinder $(\mathrm{DP}, R, H, V, O_{ip})$. The input parameters required for the modeling process can be expressed as follows:

$$P_m(f) = \{p(x, y, z), r, h, v(u, v, k)\}, \tag{2}$$

where $P_m(f)$ is modeling results using input parameters, $p(x, y, z)$ is datum point, r is radius, h is height, and $v(u, v, k)$ is extruding direction.

The modification and updating of features are relevant to input parameters, i.e., modeling parameters. However, output parameters in the CAD system can be defined as follows:

$$P_c(f) = \{g_c, d_c, m_c\}, \tag{3}$$

where g_c represents cylinder entity, d_c represents the design intent of cylinder, and m_c represents the cylinder's nongeometry information such as material or color.

According to the above expression, the input parameters and output parameters of the model are inclusive but not identical. Secondly, the definition of design intent is not clear. These parameters may be dimensional such as r in equation (2) or direction constraint such as $v(u, v, k)$ in equation (2). To express a model for simulation requires additional work by engineers to define its semantic information such as d_c and m_c in equation (3). Nongeometric information about materials and colors is stored in annotations in the model and cannot be updated with the model. These drawbacks undoubtedly affect the quality and efficiency of the CAD model to the subsequent simulation process. Therefore, how to transfer the CAD model and its parameters to the analysis process quickly and accurately needs further exploration.

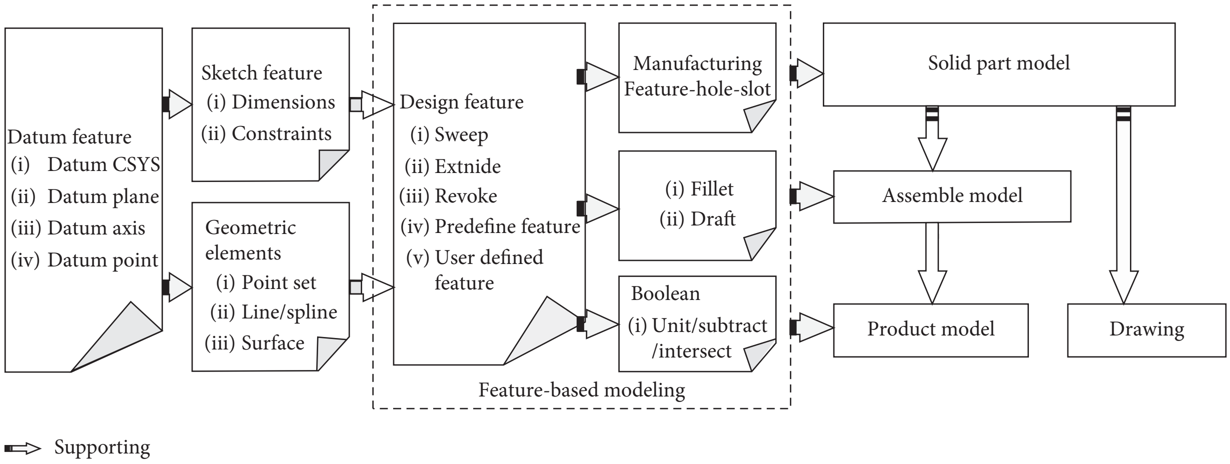


Figure 1: Feature-based CAD system modeling process.

2.2. Multiphysics Domain Model Representation Method Analysis. The existing multiphysics domain modeling also encounters the parameter conversion problem mentioned, especially even when it comes to optimization. Though we have adopted a more simplified way for modeling and simulation [36], the use of complex models has more verification functions and industrial application value [37]. The preprocessing, calculation, and postprocessing of simulation still require repeated work for modifying model and calculation conditions even after determining the analysis goals and methods. Furthermore, different methods for describing the same object or using different definitions for the domain can cause that domain model is not adapting to the design-analysis iterative process.

A typical example of computational fluid dynamics and heat transfer analysis for film cooling [38] is shown in Figure 2. Its analysis goal can be defined as follows:

$$\eta_f = k\left(\frac{X}{\mathrm{MS}}\right)^{\xi} + c, \tag{4}$$

$$M = \frac{\rho_g v_g}{\rho_c v_c}, \tag{5}$$

where η_f represents film hole cooling efficiency, k is the coefficient of experience related to the location of the hole, M is the blowing ratio, S is the equivalent radius, $\rho_g v_g$ is gas mass flow, and $\rho_c v_c$ is cooling air mass flow.

According to the expected cooling efficiency and a certain blowing ratio, the equivalent radius of the desired design can be obtained through equations (4) and (5). When the circular hole is specified in Figure 2, the calculation formula can be given by

$$S = \frac{n}{H}\frac{\pi d^2}{4}, \tag{6}$$

$$H = (n-1)t, \tag{7}$$

where H is the height of one-row hole, n is the number of holes, d is hole radius, and t is spacing of adjacent film hole.

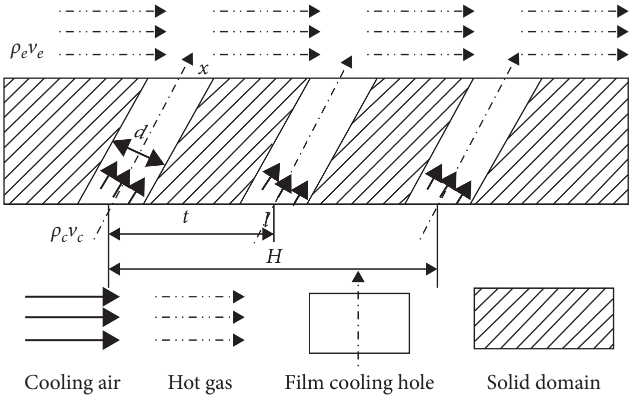


Figure 2: Flat film cooling design diagram.

By extruding, hole operation, and instant hole copy, the geometric entities to be analyzed are initially formed. From the CAD model expression, modeling parameters can be expressed as follows:

$$P_m(f) = \{p(x,y,z), r, h, v(u,v,k), n, t\}, \tag{8}$$

where $p(x,y,z)$ is datum point, r is radius, h is the depth, $v(u,v,k)$ is extruding direction, n is the number of copies, and t is spacing of adjacent film hole.

On the other hand, CFD software needs to construct a complete fluid calculation domain for analysis. The usual practice is to construct a complete inclusion solid to simulate the range of fluid flow and then is to form a fluid domain containing cold air and gas by subtracting the solid portion as shown in Figure 3.

Therefore, in order to achieve multiphysics simulation, simulation parameters are needed as follows:

$$P_S(f) = P_{\text{domain}} \cup P_{\text{boundary}} \cup P_{\text{material}} \cup P_{\text{mesh}} \cup P_{\text{compute}} \cup P_{\text{post}}, \tag{9}$$

where P_{domain} represents computational domain parameters including fluid domain and solid domain, P_{boundary} represents boundary parameters including inlet, outlet, wall,

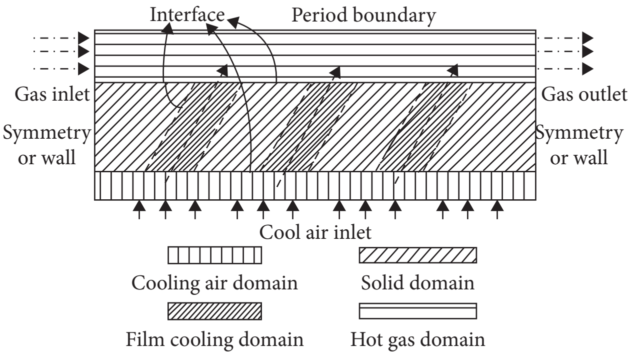


Figure 3: Multiphysics model for film cooling flat simulation.

symmetry, opening, interface or interior, and their dependent surface and physics variable. P_{material} represents material parameters for multiphysics domain such as rigid elastic modulus, fluid thermal conductivity, and specific heat capacity. P_{mesh} represents parameters of mesh method, mesh size, and mesh control such as tetrahedron, hexahedron, prism, and their appending geometry. P_{compute} represents parameters of different domain physics model, calculation accuracy, and iteration times. P_{post} depends on analysis goal and performance evaluation. Specifically, reference geometry and parameters can be made to the illustration of Figure 3.

From the geometrical point of view, the conventional method is to aggregate a number of geometric entities with geometric shapes and forms, and there is no integration of the properties of the relevant features. The computational domain geometry's modification requires geometric adjustment from the CAD or even the initial design parameters. Boolean operation usually causes errors or unexpected results due to the complexity of the model. The designer needs to perform geometric inspection and repair even sewing.

Table 1 shows the parameter difference for CAD and simulation. The setting of the boundary conditions and the division of the mesh in the preprocessing of the numerical simulation are heavily dependent on the geometry of the computational domain as equation (9) defines. However, cooling air and gas domain is a negative entity when modeling in the CAD system shown in Figure 2. They do not exist or are not preserved during the modeling process. Besides, these fluid regions in Figure 3 should have been molded for meshing, setting boundaries, and physics model for numerical simulation.

According to the above analysis, the simulation parameters are dependent on the geometry of the multiphysics model. Therefore, it is necessary to construct a multiphysics model and adopt a unified parameter expression method for the same object in the modeling stage, which is necessary to solve the problems of model adaptation and efficiency for multiphysics simulation or MDO.

2.3. Multiphysics Domain Model Architecture. Through the analysis of the model representation method, the relationship between multiphysics domain model parameters (MMDP) and modeling parameters (MP), design parameters (DP), simulation parameters (SP), and optimization parameters (OP) was summarized, as shown in Figure 4.

Updating domain model depends on its domain parameters. Through product requirements, the rational design model (DM) was designed using physical parameters determined by multiple disciplines' knowledge and was evaluated through performance parameters. Then, a CAD model (MM) is generated, and optimization objective function and constraint function can be generated as an optimization model (OP). Moreover, the multiphysics domain model (MDM) is carried out and should be relevant to the designed CAD model. The feature parameters are needed to be inherited. When it comes to multiphysics domain numerical simulation, simulation model (SM) generation is generally divided into three stages: preprocessing, calculation, and postprocessing. Three stage parameters are associated with multiphysics domain geometric and are summarized in three aspects as follows:

(1) Preprocessing: multianalysis domain model, meshing (boundary layer mesh, local refinement, and mesh match), material properties (elastic modulus), boundary conditions (inlet and outlet, wall, symmetry, and periodic boundary), displacement constraints, and boundary conditions
(2) Calculation: determination of the analytical analysis model (fluid domain calculation model, residual, and number of iterations)
(3) Posttreatment: temperature cloud map, flow rate map, equivalent stress, etc.

Therefore, when building a CAD model, the designer should understand the changes in the foreseeable geometric model and make the geometric model robust enough to cope with the changes. Besides, according to the numerical simulation evaluation results of the physical model, the analyst should also be able to modify the geometric model quickly. Through multiphysics domain model architecture, there are still works to achieve.

(1) Implement requirement decomposition and functional decomposition [37] for appropriate domain parameter extraction
(2) Modeling method to achieve different domain knowledge storing and updating in one model
(3) Achieve multiphysics domain parameters transferring to each other

3. Computer-Aided Design-Supporting-Simulation Technology

3.1. Proposed Method. The product's function refers to the ability to meet users' requirements while users need to form a demand space. Each requirement corresponds to the function. The definition and description of each function should not only meet the specific user requirements, but also reflect the technical principle of the function realization. Although the functional definitions of products in different fields are different, the functional decomposition performed

Table 1: Parameters difference for CAD and simulation.

Application type	Input parameters	Output parameters
Modeling in CAD	Datum (i) Datum CSYS, datum plane, datum axis, datum point Dimension (ii) Length, angle, number	Geometric entity
Modeling for simulation	Domain body (i) Solid domain, fluid domain Mesh (ii) Boundary layer mesh, local refinement, and mesh match Boundary condition (iii) Inlet, outlet, wall, symmetrical plane, periodic boundary Material Solver setting	Simulation results (i) Temperature and pressure fields files (ii) Maximum stress (iii) Objective functions' values

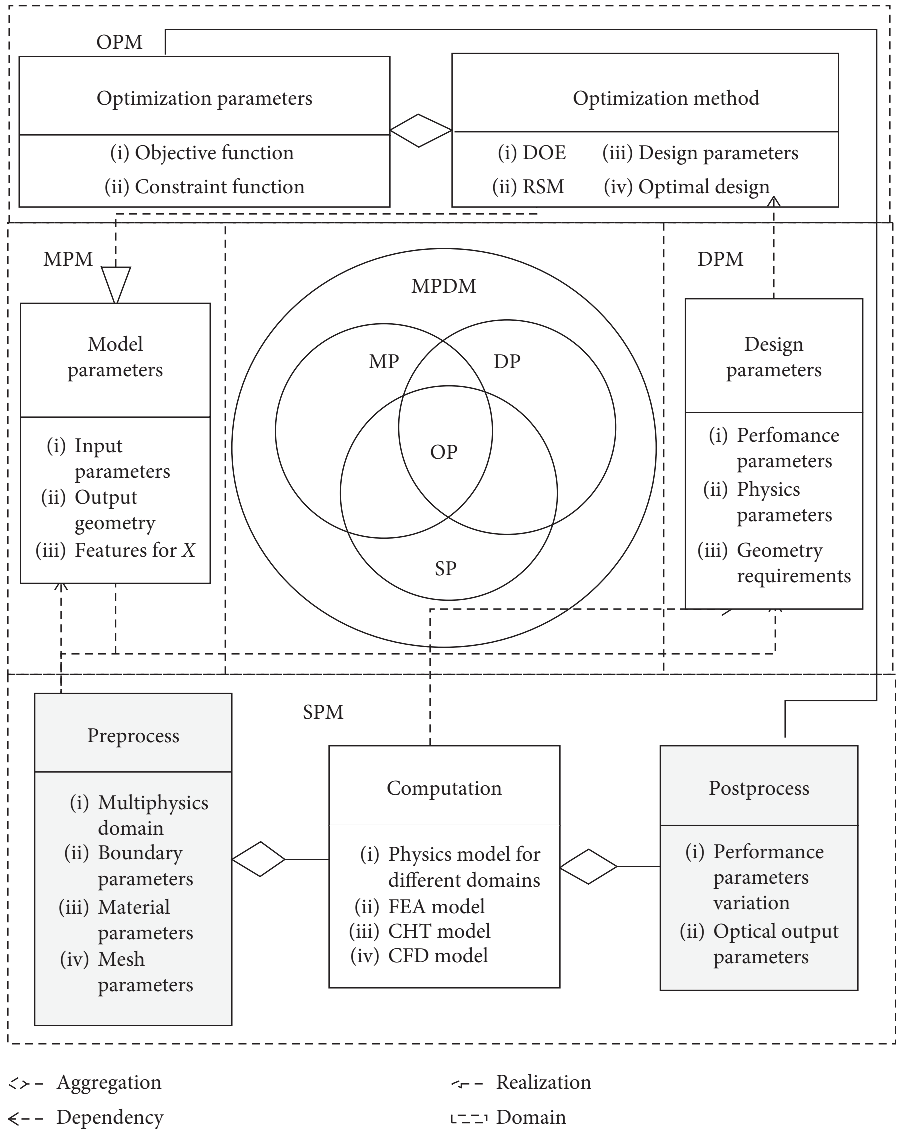


Figure 4: Structure of multiphysics domain model architecture.

under the CADSS framework is based on the knowledge space as a carrier to consider the subsequent verification process from top to bottom in the process of modeling. As shown in Figure 5, CADSS system contains three aspects including requirement space, domain knowledge, and software technology as follows:

$$S_{\mathrm{CADSS}} = (S_R, S_K, S_T). \tag{10}$$

A collection of all requirements can be expressed as follows:

$$S_r = \{r_i | i = 1, 2, \ldots, n\}. \tag{11}$$

Generally, each requirement corresponds to a function or a set of features, and each function may also correspond to a combination of multiple features. Domain knowledge includes different analysis domains and mathematical models. Firstly, analyzing and decomposing product functions are required to embed the domain knowledge into functional features. And then using relevant software technologies (such as specific geometric algorithms, parametric modeling methods, and numerical simulation methods) can achieve simulation requirements. The next section elaborates on the use of functional features for CADSS technology.

3.2. Multiphysics Model Representation in CADSS. CADSS aims to establish a functional space for product design. Therefore, parameter relationship between design, modeling, and simulation can be bridged. According to Figure 5, all functional elements can be expressed as follows:

$$F = \{f_i | i = 1, 2, \ldots, n\}, \tag{12}$$

$$f_i = S_f(r_i, k, t), \tag{13}$$

where S_f represents the design process, requirements $r_i \in S(R)$, domain knowledge $k \in S_k$, and software technology $t \in S_T$.

The realization of functional features after decomposition by means of parameter decomposition is as follows:

$$P_{\mathrm{CADSS}}(f) = P_d(f) \cup P_m(f) \cup P_s(f), \tag{14}$$

where $P_d(f)$ is defined in equation (2), $P_m(f)$ is defined in equation (3), and $P_s(f)$ is defined in equation (9).

The way defined by the above parameters can be used to build the connection between design, modeling, and simulation, thus expanding the scope of application of the feature. Through the domain knowledge, the mapping relationship of the three types of parameters can be realized as follows:

$$\begin{cases} P_m(f) = K_1(P_d(f)), P_d(f) = K_1^{-1}(P_m(f)), \\ P_s(f) = K_2(P_m(f)), P_m(f) = K_2^{-1}(P_s(f)), \\ P_s(f) = K_2(K_1(P_d(f))), P_d(f) = K_1^{-1}\left(K_2^{-1}(P_s(f))\right). \end{cases} \tag{15}$$

In general, the meaning of K can be expressed by the following triplet:

$$(X, K, Y), \tag{16}$$

where $X = (x_1, x_2, \ldots, x_n)$, $Y = (y_1, y_1, \ldots, y_n)$, and X is a set of parameters of the $P_{\mathrm{CADSS}}(f)$ space. K represents inference conditions, reasoning processes, or geometric algorithms.

According to the definition of K, design reasoning can be divided into parameter calculation, rules reasoning, chart query, and empirical solution.

3.2.1. Parameter Calculation. A set of parameters is known and is a scalar parameter, and the unknown parameters can be calculated by solving the following equations:

$$Y = E(X), \tag{17}$$

$$E = (e_1(X_1), e_2(X_2), K \ldots, e_n(X_n)). \tag{18}$$

For example, via equations (4)–(7), the number, spacing, and radius of holes can be calculated by the estimated cooling efficiency of the heat transfer design.

3.2.2. Rules Reasoning. Representation of composite features by logical expressions is shown in the following equation:

$$(X, L_c, Y). \tag{19}$$

For instance, feature-based shape inference rules can be expressed as follows:

$$\left(f_{\mathrm{gas}}, f_{\mathrm{cool}} = f_{\mathrm{fluid}} Y\right), \tag{20}$$

where f_{gas} represents gas domain features, f_{cool} represents cooling air domain features, f_{fluid} represents fluid domain features, and Y represents interface or Boolean operation depending on domain knowledge.

3.2.3. Chart Query and Empirical Solution. Under this condition, K represents experiment data or experience. Therefore, given X, with expertise and experience K, Y can be determined by C_Y in an empirical chart:

$$Y = C_Y. \tag{21}$$

The conversion of the above design parameters, modeling parameters, and simulation parameters is carried out by different engineering personnel. In order to ensure the accurate and efficient conversion of the above mapping relationship, the knowledge of the above domain needs to be solidified by the corresponding software technology, which can be expressed as follows:

$$T_{\mathrm{CADSS}} = T\left(K_{f_1} \cup K_{f_1} \cup, \ldots, \cup K_{f_1}\right). \tag{22}$$

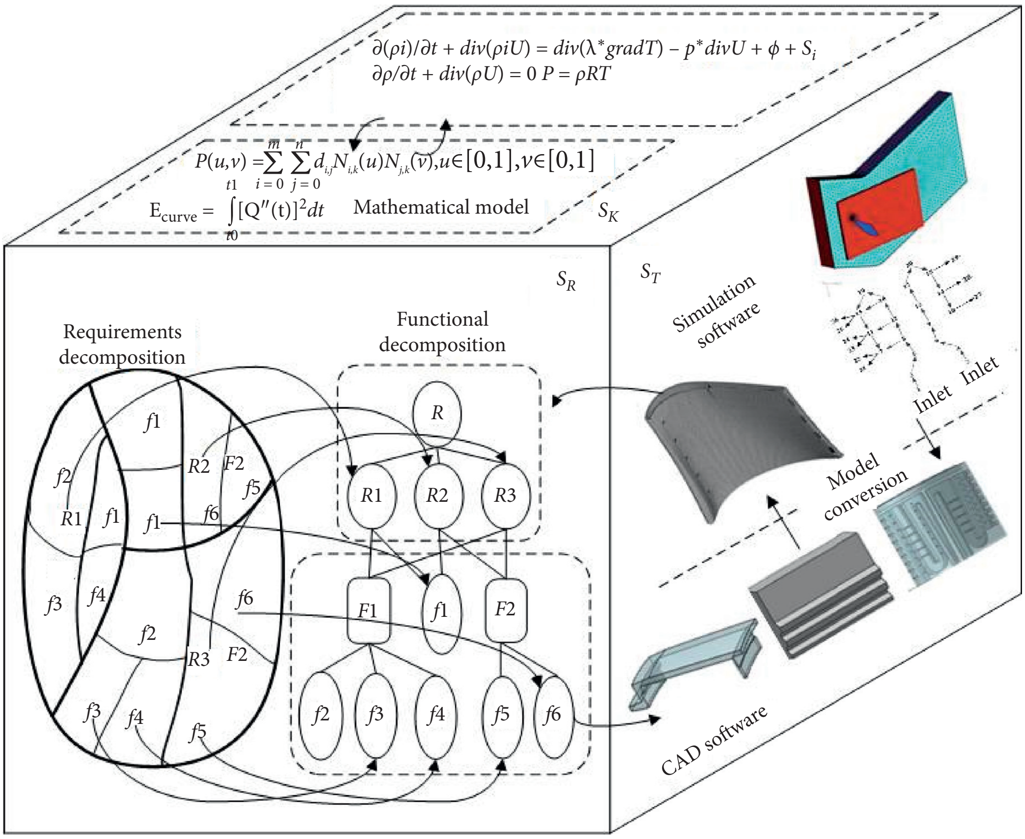

$\partial(\rho i)/\partial t + div(\rho iU) = div(\lambda^* gradT) - p^* divU + \phi + S_i$
$\partial\rho/\partial t + div(\rho U) = 0$ $P = \rho RT$
$P(u,v) = \sum_{i=0}^{m}\sum_{j=0}^{n} d_{i,j}N_{i,k}(u)N_{j,k}(v), u \in [0,1], v \in [0,1]$
$E_{curve} = \int_{t0}^{t1}[Q''(t)]^2 dt$
Mathematical model
S_K
S_R
S_T
Requirements decomposition
Functional decomposition
f1
f2
R2
F2
f5
R1
f1
f1
f6
f3
f4
f6
f2
R3
F2
f3
f4
f5
R
R1
R2
R3
F1
f1
F2
f2
f3
f4
f5
f6
Simulation software
Inlet
Inlet
Model conversion
CAD software

Figure 5: CADSS framework.

It can be realized through programming and related algorithms through CAD software, CAE analysis software, and MDO software. The next section will show how CADSS technology is conducted in a typical design case with multiple physical modeling, simulation, and optimization.

The design of the turbine blade is chosen to be the case study in this paper because it shows how the product requirements and function are decomposed to modeling features. It demonstrates how the concept of MMDP is transferred among DM, MM, SM, and OM and how they are embedded into a CAD model.

3.3. CADSS Implementation for Turbine Blade

3.3.1. Requirement Space and Functional Decomposition of Turbine Blades via CADSS Technology. The working environment of turbine blades is becoming increasingly harsh with the continuous increment of turbine inlet temperature. And the coupling effects between aerodynamics, heat transfer, and strength are becoming more and more significant for turbine blade design. It is urgent to start research on multidisciplinary design optimization methods for turbine blades compared with multiobjective optimization of discrete film hole [39]. The design optimization flowchart is shown in Figure 6.

Firstly, according to the overall design requirements of the original designed turbine blade, the initial design variables and the objective functions are determined. Then, a preliminary DOE is performed to generate several design points for simulations shown in the right of Figure 6, which is the general process for multifield coupling analysis [40]. Secondly, according to the various design points by the DOE design, the geometric model was reconstructed in the CAD software according to design variables. On the one hand, the fluid domain and solid domain geometric models are constructed for CHT simulation, and then the fluid mesh is generated based on the solid and fluid models in the meshing software. The fluid grid and solid grid are numerically solved in the computational fluid dynamics solver, and the temperature and pressure fields of the turbine blade are predicted to determine whether the objective functions are met. On the other hand, in the finite element calculation, based on the solid model of the turbine blade, the temperature field and pressure field data need to be input to the finite element analysis as boundary conditions for strength check to further determine whether the maximum stress of the blade meets the objective functions. Thirdly, the construction and improvement of the approximation model is generated with DOE results. Finally, a multidisciplinary and multiobjective optimization algorithm is used to find the optimal turbine blade model for the current design variables and objective functions.

Turbine blade overall design requirements are based on both strength and longevity conditions, using cooling air flow as less as possible to achieve higher cooling efficiency.

The design requirements can be expressed mathematically as equations (23)–(25):

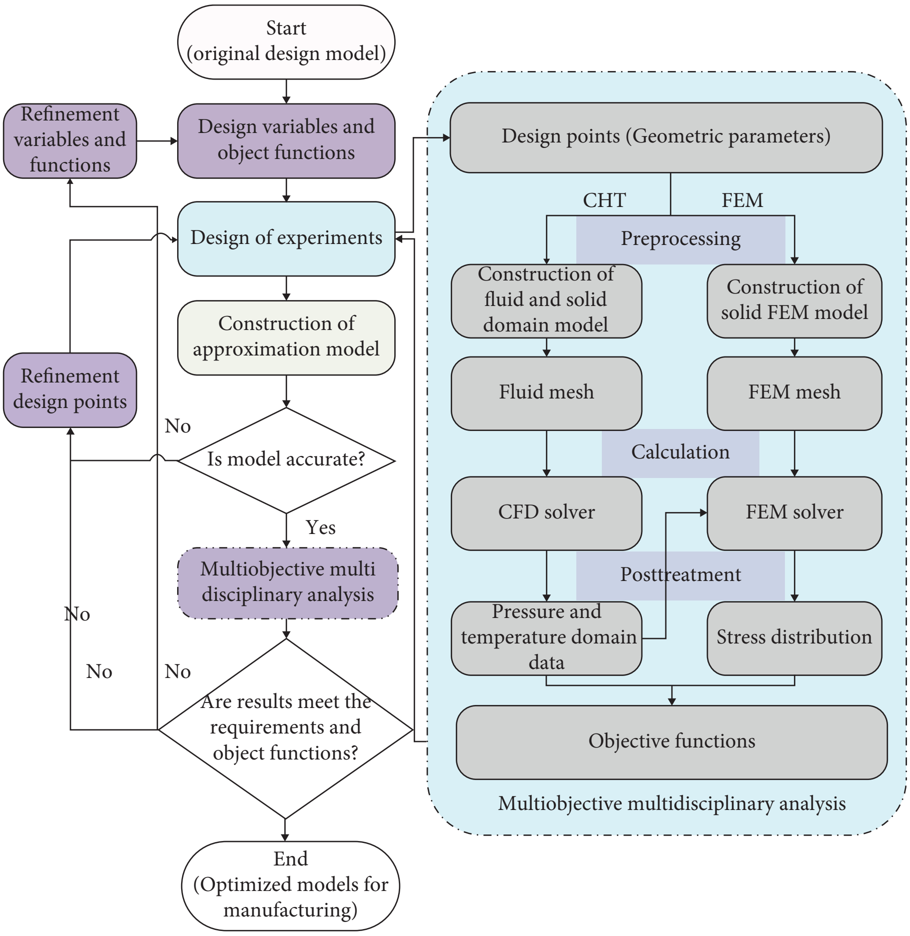

Start
(original design model)
Refinement variables and functions
Design variables and object functions
Design of experiments
Construction of approximation model
Refinement design points
No
Is model accurate?
Yes
Multiobjective multi disciplinary analysis
No
No
No
Are results meet the requirements and object functions?
End
(Optimized models for manufacturing)
Design points (Geometric parameters)
CHT
FEM
Preprocessing
Construction of fluid and solid domain model
Construction of solid FEM model
Fluid mesh
FEM mesh
Calculation
CFD solver
FEM solver
Posttreatment
Pressure and temperature domain data
Stress distribution
Objective functions
Multiobjective multidisciplinary analysis

Figure 6: Design optimization flowchart.

$$\eta_1 = \frac{T_{g,r} - T_{aw}}{T_{g,r} - T_c}, \quad (23)$$

where η_1 is cooling efficiency, $T_{g,r}$ is gas recovery temperature, T_{aw} is adiabatic wall temperature, and T_c is cooling air temperature.

$$\eta_2 = \frac{T_{\max}}{T_{\text{ref}}} \frac{A_{tb}}{A_b} \frac{1}{1-\chi}, \quad (24)$$

$$\chi = \frac{T_b}{T_*}, \quad (25)$$

where T_{ref} is the reference temperature, $T_{\max}$ is the maximum temperature on the blade surface, A_b is the blade surface area. A_{tb} is the area where the temperature is higher than T_b. χ is temperature coefficient, and T_* is gas total temperature.

$$\sigma_{\max} \leq 0.8 * \sigma_{0.2}^{T}, \quad (26)$$

where $\sigma_{0.2}^{T}$ represents 0.2% yield strength of the material at normal blade operating temperatures. η_1 reflects the cooling efficiency of the blade cooling structure. η_2 reflects the worst-case temperature field during blade operation, which is related to blade lifetime. $\sigma_{\max}$ reflects the maximum strength experienced by the blade.

According to requirements, the composite cooling structure is adopted for reducing weight and enhancing cooling air efficiency. As shown in Figure 7, according to a different subject domain, the functional analysis and decomposition of the turbine blades are carried out to obtain a functional decomposition tree. The functional structure decomposition tree is organized in the form of functional features, and the top-down design is used to perform functional features management. Finally, the turbine blade 3D CAD model and multiphysics simulation preprocessing model are generated.

After the requirement decomposition, the next step is to customize the blade model according to the requirement decomposition design scheme to quickly complete the generation of the calculation domain. The traditional design process does not include the modeling history and only knows the modeling results, and the results are all nonparametric solid model; fluid domain generation contains a lot of engineering design experience and frequent manual interaction, and the modeling process is irregular and does not support the rapid modification and serial design of the blade model. The numerical analysis process is an iterative process of modifying and reanalysing the blade model according to the calculation results and constructing and adjusting a nonparametric solid model that lacks a modeling history, repeated work, and complicated adjustment process, and the quality of the model cannot be guaranteed.

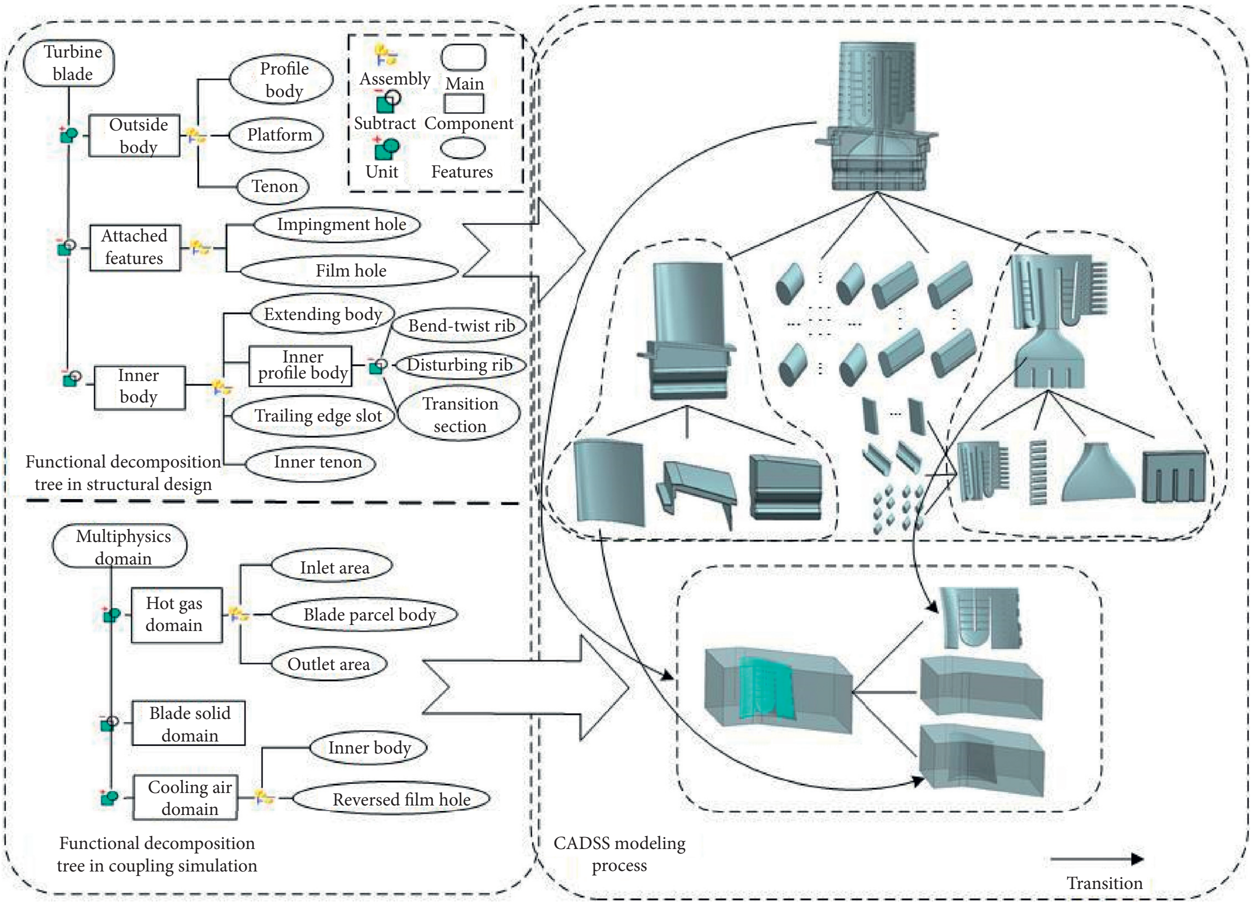


Figure 7: Functional decomposition of turbine blades via CADSS technology.

In addition, the above modeling method only considers the construction of the geometric model of the calculation domain and does not analyze the entire process of gas-heat coupling numerical simulation and considers the requirements of subsequent analysis for modeling. In fact, in the postprocessing stage, in order to obtain the temperature distribution, pressure distribution of the critical position of the blade, and the cooling air flow, speed, and pressure of the cooling structure, it is necessary to extract the geometric elements in the calculation domain model and then observe and analyze the calculation results, and the location and generation of complex geometric features in the analysis software are difficult, require a large amount of work, and cannot guarantee accuracy.

According to the above analysis, the requirements of the gas-heat coupling numerical simulation process for the computational domain modeling are as follows:

(1) A general method for the rapid construction of the turbine blade inflow cold air region

(2) Extraction of analysis information such as the positioning and generation of key geometric features

(3) The generation of outflow gas domain should be compatible with the blade model and can be flexibly adjusted

Among them, (1) need to design methods to avoid the need to rely on experience in the process of generating cold air regions; (2) need to extract methods for key geometric features according to the requirements of the postprocessing stage; (3) mean that according to the boundary of gas-heat coupling analysis. The conditions construct a method of generating outflow gas domains that can adapt to different blade shapes and cooling structures.

3.3.2. Model Design with CADSS

(1) Different Analysis Domain Model Generation Scheme. According to the above decomposition and design scheme, turbine cooling blades were divided into typical features as shown in the following equation:

$$F_m = \{f_i(P_{\text{CADSS}}) | i = 1, 2, \ldots, 13\}, \tag{27}$$

where $f_1, f_2, \ldots, f_{13}$ is shown in Figure 8.

The basic feature set of any turbine cooling blade was F'_m, $F'_m \subseteq F_m$. Blade model is represented as follows:

$$M_b = (M_o, M_i, f_{12}, f_{13}, e_b), \tag{28}$$

where $e_b = \left\{e_b^j | j = 1, 2, 3\right\}$, when subtraction is operated, e_b equals zero, and when uniting, e_b equals one.

Blade inner body is represented as follows:

$$M_i = (f_6, f_7, f_8, f_9, f_{10}, f_{11}, f_5, f_4, e_i). \qquad (29)$$

Blade outside the body can be represented as follows:

$$M_o = (f_1, f_2, f_3, e_o). \qquad (30)$$

Take f_{13} for example, its design parameters depend on η_f, relevance variables in equation (4), and estimated gas temperature and cooling air temperature. Tables 1 and 2 show the parameters of the blade designed in this paper. Through equation (6), the modeling parameters are calculated. In the CAD system, f_{13}^{-1}, the film hole fluid domain corresponding to f_{13}, can be generated and its simulation parameters such as mesh size control and inflating method can be assigned to f_{13}^{-1}.

Therefore, it is possible to design a turbine blade computational domain modeling solution for gas-heat coupling (Figure 9), generate the inner shape and cooling structure of the blade body according to the inner section line and modeling parameters, and then automatically locate the outer cooling features to achieve the cold air domain. The rapid modeling meets the requirements (1); according to the profile section line and blade modeling parameters, the blade shape is created, and then the adaptive blade periodic boundary and the blade inclusion are created, and then according to the blade periodic boundary and the wall boundary as well as the direction and length of the entrance and exit, adaptive cutting inclusions for rapidly modeling gas field is generated, which meets the requirements (3); in the process of modeling the cold gas field and the gas field, according to the relevant modeling parameters, the extraction of geometric features and the preservation of nongeometric information meet the requirements (3).

(2) Computational Domain Automatic Generation Algorithm. The calculation model of the turbine blade in multiphysics numerical simulation mainly includes the fluid calculation domain and the solid calculation domain. According to the flow field characteristics and the mesh division strategy, the fluid calculation domain can be divided into the gas domain and the cold gas domain.

Specifically as shown in Figure 10, according to the flow path entity M_b created by the existing blade solid model M_b and the blade flow path information, the calculation domain model is constructed by artificial interaction. The fluid calculation domain M_f is formed by subtracting between M_p and M_b. With the method of segmentation and combination, a gas domain M_g and a cooling air domain M_c were divided by the fluid domain.

During the modeling of fluid domain, the simulation parameters can be generated. The boundary attached on M_f includes a gas inlet, gas outlet, cooling air inlet, period surface, and wall. Mesh control parameters can also be embedded into the fluid entity. The interface generated by M_f and M_b needs to assign inflation mesh control for the requirement of y-plus depending on the physics model. Material parameters are appended to the target entity for simulation. Blade surface and fluid surface for postprocess evaluating performance parameters such as temperature and pressure are marked by name to be identified in the postprocess module.

The creation of the gas domain model needs to adapt to different blade profiles, and it is necessary to combine the common geometrical characteristics of the blades with the characteristics of the outflow gas domain. The midblade arc is not only an important parameter for aerodynamic shape design but also an important design reference for the inner shape of the blade. It can reflect the bending and torsional changes of the blade in the flow channel, which meets the conditions of periodic boundary selection. The main difficulties are the automatic matching of periodic boundaries and the adaptive cutting of blade inclusions.

Adaptive pipeline intersection algorithm is shown in Figure 11. First of all, in the past, the method of creating intersection arcs in the pipeline intersection was mainly based on empirical data to obtain the pipeline radius, and the generation of periodic boundaries needs to adapt to different blade shapes. The pipeline radius data cannot automatically match the blade shape, so the improved adaptive pipeline is used. The intersection algorithm can solve this problem; secondly, due to the intersection of the pipeline and the partial encryption and streamlining of the blade cross section line, a variety of curves shown in Figure 11 will be generated, and it is necessary to use an automatic recognition algorithm to identify the middle arc from the curve of Figure 11.

Step 1: traverse the blade profile section line, according to the height and curve type, automatically match each layer of the leaf back curve L_k and the leaf pot curve L_k (the number k of section line layers), and discretize the L_k and L_k' according to the parameter method of equal arc length to obtain the layer respectively point set.

Step 2: calculate the distance between the points in the two points separately $|\overrightarrow{P_i S_j}|$. The maximum wall thickness of the airfoil is obtained by trial calculation $C_{\max}' \longleftarrow \max\{\min(D_i) | i = 0, 1, 2, \ldots, n-1\}$. The diameter of the pipeline in the pipeline intersection is $0.75C_{\max}'$, which is the radius of the pipeline of each layer of cross section line.

Step 3: with the L_k and L_k' starting point as the center of the circle, create a circle with a diameter d_k in the normal plane of the starting point of the curve, and as the guide line, and after sweeping, the two pipes intersect to obtain C_P and C_P', and project onto the plane of the section line of this layer to obtain the midarc C_k.

Adaptive generation method of gas domain is shown in Figure 12. The blade model obtained by improving the pipeline intersection algorithm is processed by the periodic boundary automatic matching method and the outflow gas

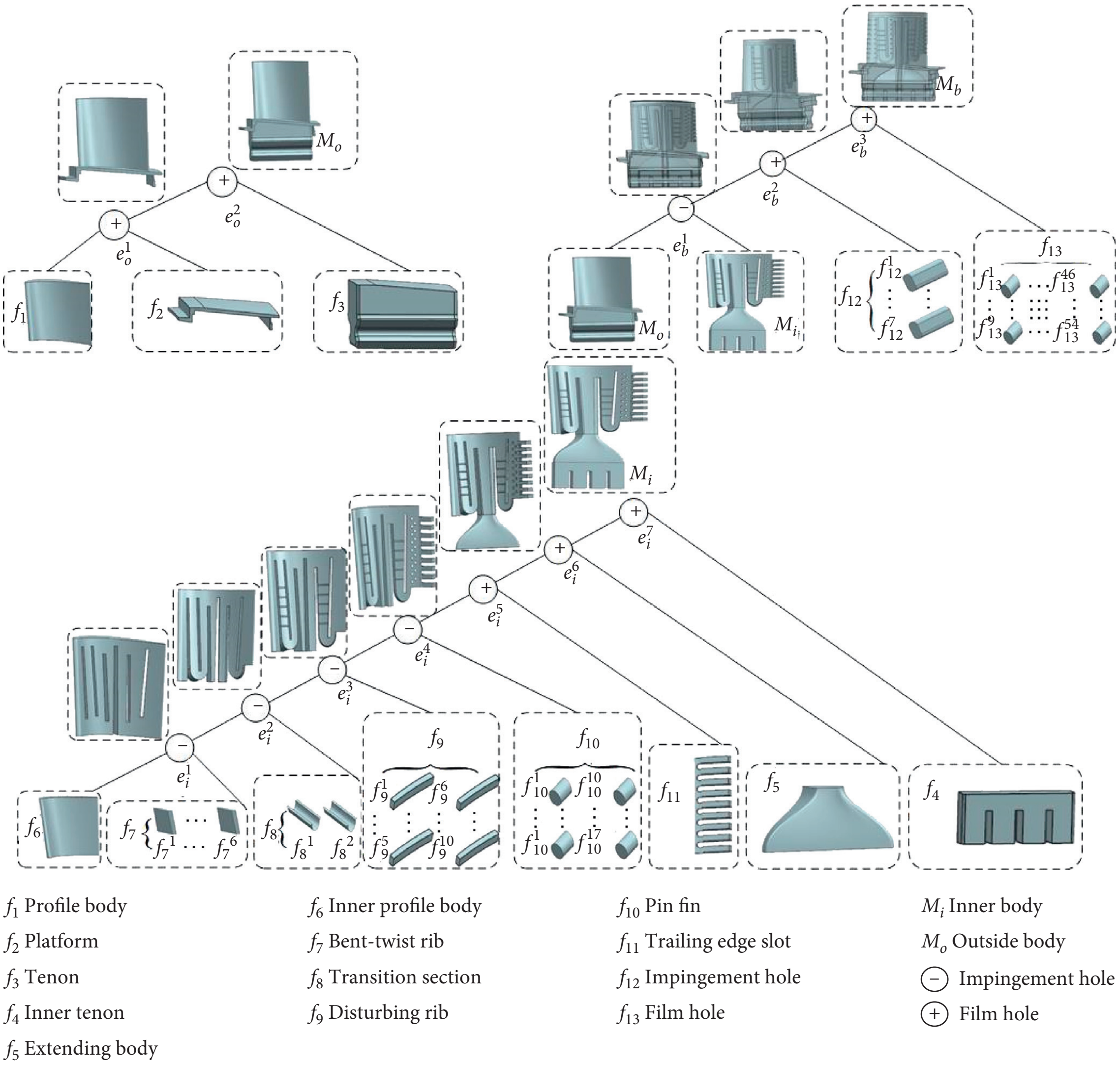


Figure 8: CADSS feature representation in CADSS.

Table 2: Tags in CAD software for simulation.

Tag and type	Attribute	Boundary condition
NS_I_1, NS_I_2,. . ., NS_I_n	Inlet	Velocity or pressure inlet
NS_O_1, NS_O_2,. . ., NS_O_n	Outlet	Velocity or pressure inlet
NS_W_1, NS_W_2,. . ., NS_W_n	Wall	No-slip wall
NS_S_1, NS_S_2,. . ., NS_S_n	Symmetrical plane	Symmetry
NS_P_1, NS_P_2,. . ., NS_P_n	Periodic boundary	Periodic surface
NS_D_1, NS_D_2,. . ., NS_D_n	Domain or subdomain	Fluid or solid domain

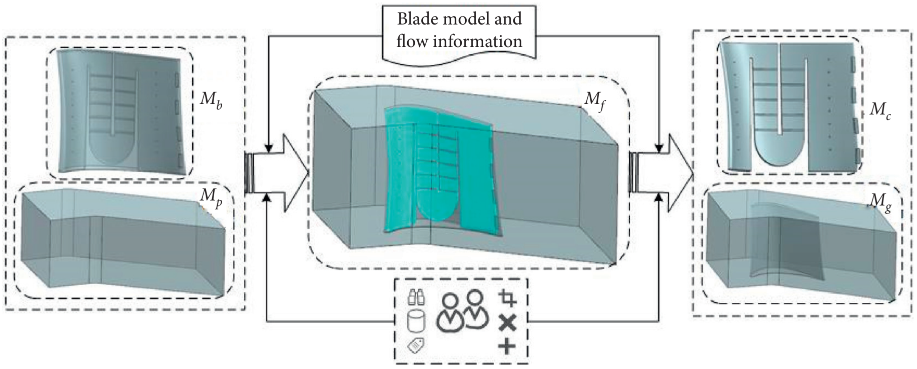


Figure 9: Multiphysics numerical simulation model.

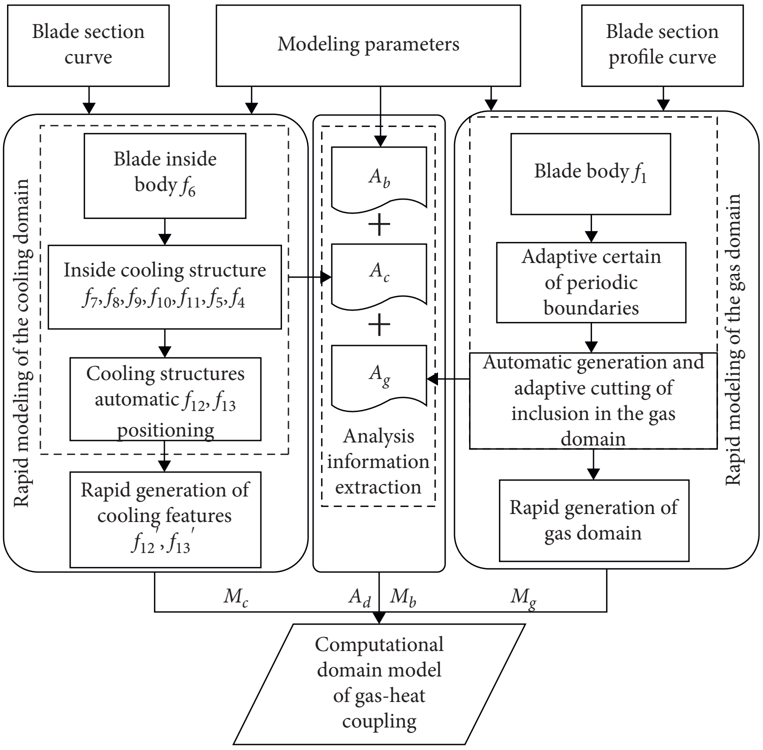

Blade section curve
Modeling parameters
Blade section profile curve
Rapid modeling of the cooling domain
Blade inside body f_6
Inside cooling structure $f_7, f_8, f_9, f_{10}, f_{11}, f_5, f_4$
Cooling structures automatic f_{12}, f_{13} positioning
Rapid generation of cooling features f_{12}', f_{13}'
A_b
+
A_c
+
A_g
Analysis information extraction
Blade body f_1
Adaptive certain of periodic boundaries
Automatic generation and adaptive cutting of inclusion in the gas domain
Rapid generation of gas domain
Rapid modeling of the gas domain
M_c
A_d
M_b
M_g
Computational domain model of gas-heat coupling

Figure 10: Conjugated heat transfer simulation-oriented turbine blade computation domain modeling scheme.

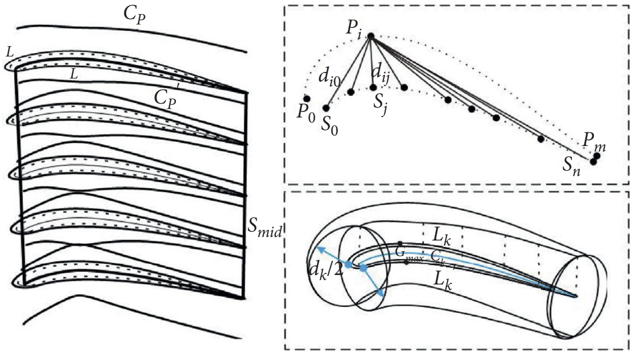


Figure 11: Adaptive pipeline intersection for turbine blade.

domain adaptive cropping algorithm, which can adaptively generate the gas domain.

Step 1: according to the height and curve type, filter the pipes in Figure 7 to find the intersection curve C_P and C_p', and group the remaining outer section line, inner section line, and middle arc line according to height, query the minimum bounding box of each group of section line, and compare the smallest surrounding. The bounding box can be divided into outer section line, inner section line, and middle arc line from inside to inside in sequence.

Step 2: according to the grouped midarc lines, loft them in order of height to get the midarc surface and extend the midarc surface S_m' and S_m''. According to the number of blades on the turbine, the midarc surface rotation angle is obtained.

Step 3: simultaneously calculate the leaf pot, the backside periodic boundary, and the minimum bounding box of the leaf, and create a leaf inclusion E_{cubic}, the minimum enclosing box B_c ($B_{\text{blade}} \subset B_c$), and within the area formed by the backside of the pot, to ensure rotation curved surface and ability to cut inclusions,

Step 4: continue to cut according to the wall boundary; input the inlet and outlet gas angles α_1 and α_2, lengths l_1 and l_2, stretch and wrap the inlet and outlet surface to generate the blade flow channel entity, and minus the blade entity and the cooling air domain to obtain the outflow gas domain part M_g.

Through the adaptive pipeline intersection Algorithm 1, the arc in the blade can adapt to the changes of different blades. The periodic boundary automatic matching method and the outflow gas domain adaptive cropping Algorithm 2 both solve the problem of outflow gas domain and blade model adaptation. The flexibility of its modeling requirements has been met through Algorithms 1 and 2.

(3) Generation of Analysis Domain Model for Simulation and Optimization. The modeling part and algorithm described above can be realized by building a dedicated GUI for analysis-oriented CAD models. For example, in Siemens NX®, engineers can use Product Template Studio or Block UI Styler, depending on the complexity of the current product and the user's skill set. If users want to use Block UI Styler, they need to use the application programming interface for programming. Such design may be function-

```
Inputs: blade section line $L_k$ and $L_k'$
Outputs: midcurve $C_k$
$k \longleftarrow 0$, initialize and traverse the blade profile section line $L_k$ and $L_k'$
while $H_{L_k} \longleftarrow h$ do
    Discretize the $L_k$ and $L_k'$ according to the parameter method of equal arc length to obtain the layer respectively point sets $P = \{P_i(x, y, z) | i = 0, 1, 2, \ldots, n-2, n-1\}$, $S = \{S_j(x, y, z) | j = 0, 1, 2, \ldots, n-2, n-1\}$.
    for all point sets $P$ and $S$
        Calculate the distance between the points in the two points separately $|\overrightarrow{P_i S_j}|$
        $C_{\max}' \longleftarrow \max\{\min(D_i) | i = 0, 1, 2, \ldots, n-1\}$
        $d_k \longleftarrow 0.75 * C_{\max}'$
        Create a circle with a diameter $d_k$ in the normal plane of the starting point of the curve, and as the guide line, and after sweeping, the two pipes intersect to obtain $C_P$ and $C_P'$ and project onto the plane of the section line of this layer to obtain the midarc $C_k$
    end for
    $k \longleftarrow k + 1$
end while
return $C_k$
```

Algorithm 1: Adaptive pipeline intersection algorithm.

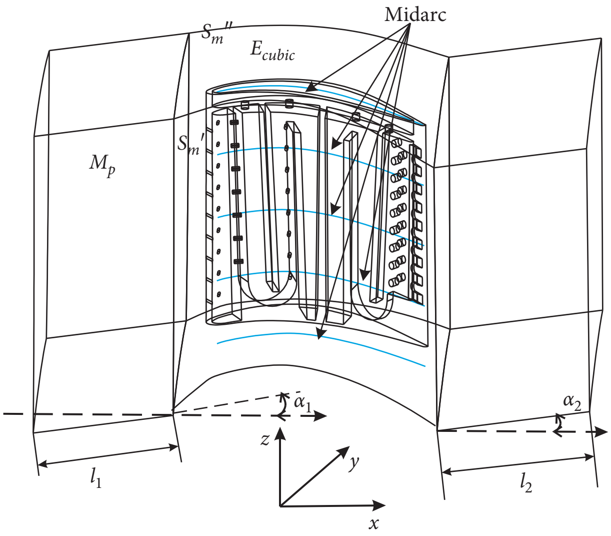


Figure 12: Adaptive generation method of gas domain.

centric in the future because users do not need to rebuild CAD models, but provide some functional parameters to automatically generate as shown in Figures 8 and 9. Beneficially, the flow space can be easily updated subject to design changes. Tags with attributes, similar to named selections, are assigned to the fluid geometrical faces to transmit boundary information in CAD/CFD conversion. By using the CAD Configuration Manager provided by ANSYS Workbench, the simulation platform can visit and modify the geometry file constructed by SolidWorks. The attributes attached by tags are used to guide the mesh generation as shown in Figure 13. Consequently, CAE boundary features are established, resulting in the generation of the fluid flow space, which is the input of the intelligent CFD solver.

According to the geometry and parameters automatically generated by the CAD system shown in Tables 2 and 3, they are stored in files or models in the form of text or expressions. The simulation part is implemented based on ANSYS Workbench. The platform provides two script levels. For application-level task automation, especially in CFX, the CFX Command Language (CCL) ("Workbench Scripting Guide" 2012) is used as a session to operate CFX-Pre and CFD-Post. Therefore, CCL can be used to perform physical model selection. After the simulation is completed, CCL can also automatically perform postprocessing. For project-level task automation, Workbench scripts ("Workbench Scripting Guide“ 2013) is used to create the entire project and call various applications to complete the created project. Actions performed through the GUI are recorded as a log of Python-based scripts. Such scripts can be customized for specific purposes. Therefore, the function of the entire system has been greatly expanded without excessive scripting. Based on the provided tools, fluid physical characteristics and dynamic physical characteristics can be edited through the program CCL. In addition, the intelligent solver function is

```
Inputs: midcurve $C_k$
Outputs: outflow gas domain $M_c$ and $M_g$
i ← 0, initialize and filter to group $L_k$ and $L_k'$
Loft them in order of height to get the midarc surface and extend the midarc surface $S_m$, blade number $N$←n, $\alpha_{\text{rotate}}$←360°/2•$N$,
    symmetrical rotating midarc $S_m$ to get $S_m'$ and $S_m''$,
Calculate the leaf pot, the dorsal periodic boundary, and the minimum bounding box of the leaf $B_p$, $B_s$, and $B_{\text{blade}}$, and then create
    the blade inclusion $E_{\text{cubic}}$,
while $B_{\text{blade}} \subset B_c$ do
if $B_c \in B_p$ && $B_c \in B_s$
Using rotating midsurface $S_m'$ and $S_m''$ to cut inclusions $E_{\text{cubic}}$,
else adjust the size midsurface $S_m'$ and $S_m''$
Input the inlet and outlet gas angles $\alpha_1$ and $\alpha_2$, lengths $l_1$ and $l_2$, and stretch the entrance and exit surfaces of the package to
    generate the blade flow channel entity $M_p$,
Subtract the blade solids $M_b$ and the cooling domain $M_c$ to get outflow gas domain $M_g$
end while
return $\{M_c, M_g\}$
```

Algorithm 2: Adaptive pipeline intersection algorithm.

installed into the CAD/CFD integrated system through Workbench scripts.

4. Model Validation and Optimization Cases

4.1. Numerical Method and Experimental Data Verification. Conjugated heat transfer (CHT) numerical simulation method was adopted to optimize the blade cooling structure. Further experimental data are required to verify method validity and find an appropriate numerical model.

The C3Xvane analytical and experimental evaluation has extensive experimental data under high pressure and temperature conditions.

The C3X experiment has detailed data. In this paper, the CHT method is used to simulate the C3X blade, and the simulation results are also compared. The data of blade geometry and experimental conditions 4521 are from the literature [41]. The calculation model contains a total of 12 computational domains: 1 gas domain, 1 solid domain, and 10 cooling air domains. The boundary layer grid is added to the fluid-solid interface to simulate the close avoidance. The y-plus is less than 1, and the inlet Mach number is equal to 0.17, the total inlet temperature is 818 K, and the gas-thermal coupling numerical simulation is solved in ANSYS CFX using the RANS method, and the SST turbulence model is applied.

The numerical simulation results were compared with the experimental data. According to Figure 14, the gas-thermal coupling numerical simulation can well predict the pressure distribution on the blade surface. The SST turbulence model can give an accurate prediction of the temperature distribution near the leading edge, as shown in Figure 15. However, the CHT model tends to overestimate the convective heat transfer intensity after the suction side shock wave, so the calculated blade temperature is significantly higher than the experimental data. Besides, the calculation accuracy of the pressure surface is higher than the calculation accuracy of the suction surface because the suction phenomenon such as surface transition and flow separation occurs on the suction surface, and the full turbulence model is unable to predict the transition. Thus, using the SST turbulence model can provide an acceptable prediction of the cooling performance of the blade with an average error of less than 22.8 K, and the SST model can give an acceptable prediction of convective heat transfer on the blade surface.

4.2. Multiphysics Model for CHT and FEA. This study focuses on improving the internal cooling efficiency, not paying much attention to the Tenon and platform design. Therefore, the Tenon and platform of the blade are simplified as a block, which could reduce the computational cost. At the same time, simulation parameters are preserved during the modeling process as equation (15) represented and the parameters through the custom script file transfer to ANSYS Workbench. CAD configuration manager provided by ANSYS Workbench can be used to update the blade geometry model parameters. Table 4 gives the optimization parameters' initial value, and Table 5 gives the film holes' design parameters and their design values.

The boundary conditions of the conjugate heat transfer are shown in Figure 16. The inlet total pressure is equal to 1.2 MPa, the inlet total temperature is 1200 K, and the flow direction is in the axial of the turbine engine. Coolant flow inlet mass flow rate is equal to 0.02 kg/s, and its total temperature is 650 K. At the outlet, the static pressure is equal to 0.9 MPa.

4.3. Optimization Design of Turbine Blade Film Hole and Cooling Mass Flow

4.3.1. Direct Optimization. In this case, the optimal parameters of the film hole are set for the two optimization objectives η_1 and η_2 as introduced in Section 4.2, Table 4.

Optimization variables can be expressed as follows:

$$X* = [\alpha_1, \alpha_2, r, M], \tag{31}$$

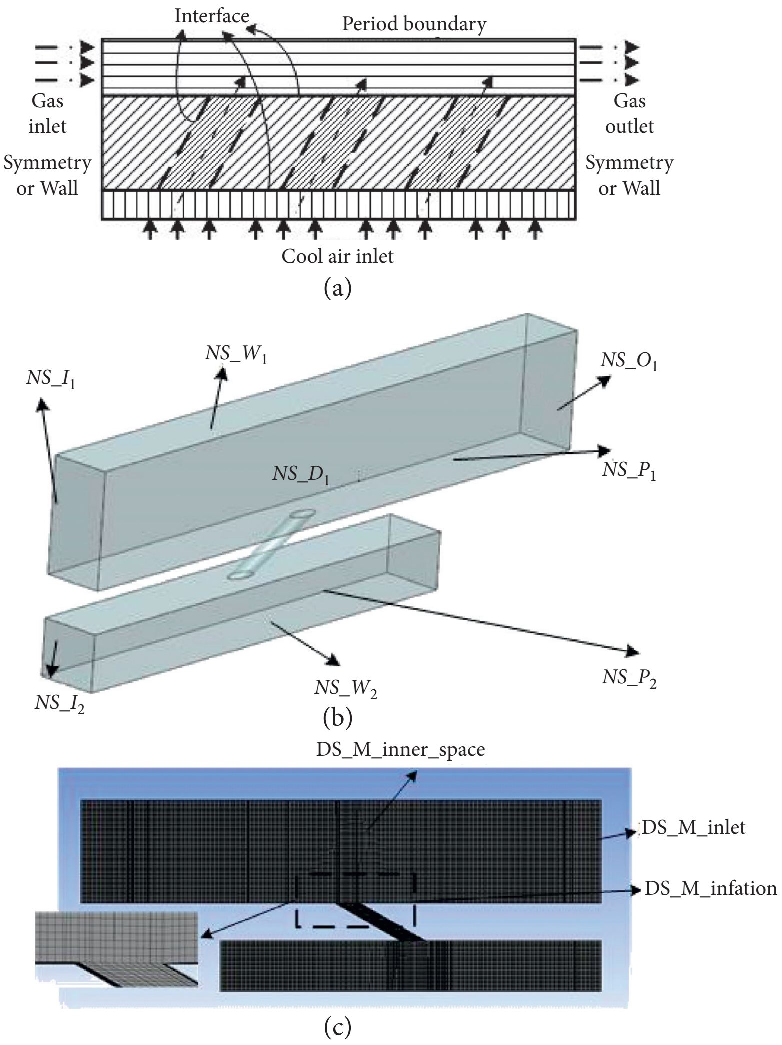


Figure 13: Model conversion in CAD and CFD.

Table 3: Multiphysics domain model parameters generated in CAD software for simulation.

Tag and type	Attribute	Application
DS_O_1, DS_O_2,. . ., DS_O_n	Optimization parameters	DOE and optimization
DS_Mp_1, DS_Mp_2,. . ., DS_Mp_n	Material properties	Simulation or optimization
DS_P_1, DS_P_2,. . ., DS_P_n	Physical parameter	Simulation boundary condition
DS_M_1, DS_M_2,. . ., DS_M_n	Mesh type	Simulation

where α_1 is film hole row 1and row 2 composition angle on blade suction surface, α_2 is film hole row 3 and row 4 composition angle on blade pressure surface, r is film hole radius, and q is cooling air mass flow.

Figure 17 shows the schematic design of film hole-related parameters and conversion relationship between the film hole design angles. In order to save computing space and improve optimization efficiency, response surface methods (RSMs) with Latin hypercube sampling design is dedicated to approximate the empirical relationship between the objective function and the design variable. Table 6 shows twenty-five design points and design variables.

For the two objective functions in equations (23) and (24) and one constraint function equation (26), the optimization problem is expressed as the following mathematical model:

$$\min F(X*) = \{\eta_1, -\eta_2\} \text{S.t.} g_m(X*) \leq 0 \sigma_{\max} \leq 0.8 * \sigma_{0.2}^{T}. \tag{32}$$

Optimization results are shown in Table 7.

Compared with the original design point, maximum temperature and average temperature of the blade are reduced with little promotion of cooling air mass flow. The overall cooling efficiency is improved with the constraint that the yield strength is under $0.8 * \sigma_{0.2}^{T}$ (750 MPa). Each of the above examples takes only about 4 hours from design,

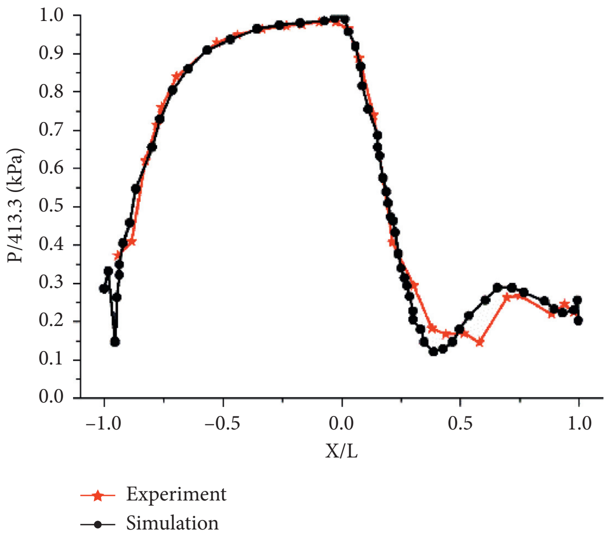


Figure 14: Pressure distribution on the midspan section of the blade surface.

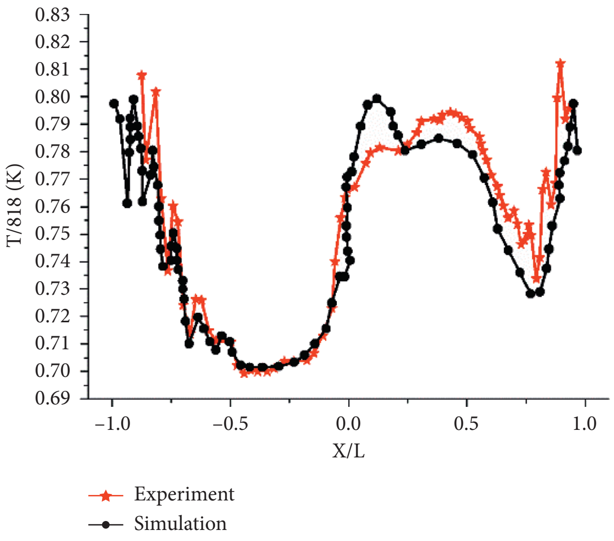


Figure 15: Temperature distribution on the midspan section of the blade surface.

Table 4: Optimization parameters' initial value.

Optimization parameters	η_1	$T_{g,r}$ (K)	T_{aw} (K)	T_c (K)
Initial value	72.72%	1200	800	750

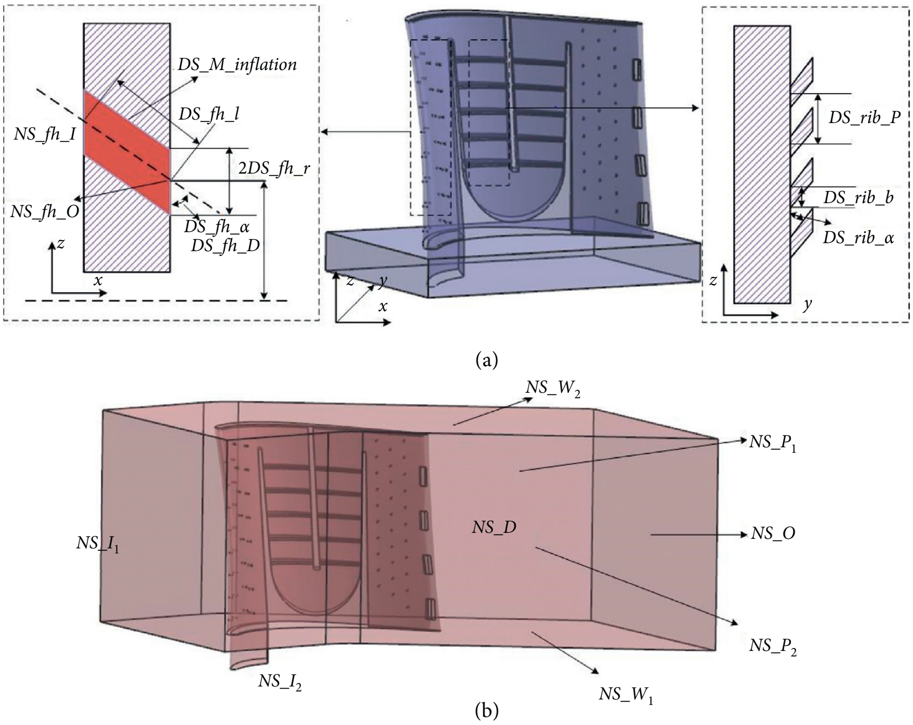
Table 5: Film holes' design parameters and their design values.

Design parameters	η_f	k	c	X (mm)	M	S (mm)
Design value	27.2%	k_f	c_f	1	2	0.4

modeling to simulation, which meets engineering design requirements.

4.3.2. Indirect Optimization. Through the generation of summary of cooling structure introduced in Section 4.2, a simplified cooling model can also be generated for batch optimization. Due to the relationship between curved and twisted blades, when the angle range of the gas film hole is too large, not only is it difficult to model, but also difficult to process, and the gas film cooling effect reaches a certain degree corresponding to a specific injection angle range, so the cooling effect can be performed first. The gas film hole can be simplified first, and the cooling channel is simplified to a straight channel, thereby reducing the range of design parameters.

The main parameters of the blade film hole are extracted according to the method in Section 3.3. The gas film hole cooling performance was studied, and the gas film cooling angle and the length of the gas film hole were determined.



(a)

(b)

Figure 16: Multiphysics model and simulation parameters.

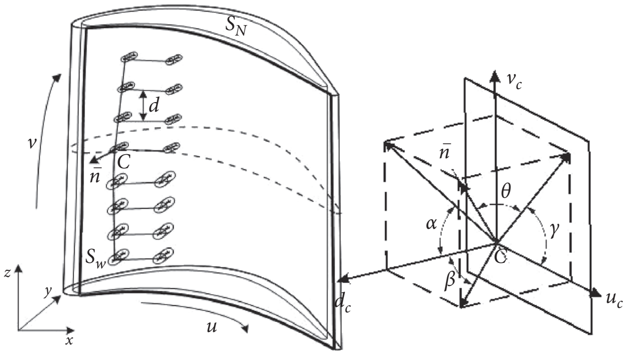


Figure 17: Schematic design of film hole-related parameters and conversion relationship between the film hole design angles.

Due to the fixed blade thickness, when the gas film hole angle was determined, the length of the gas film hole was also determined. Boundary conditions intercept the main boundary conditions of the blade.

Latin hypercube is a sampling method that is widely used in optimization, especially in approximate simulation calculations where the position of the sample point data in the entire design space is high. It can be implemented by the LHS design command in MATLAB. Parameterization mainly uses the script file of ANSYS ICEM to realize the call to ICEM software. The final parameterization program can generate a .cfx5 format file, which can be directly read by the CFX preprocessing macro file. In order to reduce the amount of calculation in the optimization process, this paper also introduces the RBNN neural network commonly used in approximate models in the optimization system, which can be implemented by MATLAB commands. The optimization algorithm selects the genetic algorithm that researchers mostly use. Through the control of the main program, the above modules are combined, and the optimal cooling structure can be obtained directly for the given aerodynamic conditions, optimized design variables, and optimized goals. The entire process does not require human intervention. The jet angle of the gas film hole is an important factor affecting its cooling performance. Reducing the jet angle will reduce the velocity component of the cooling air flow perpendicular to the cooling surface, and the effect of the air flow infiltrating into the mainstream will be weakened. More cooling air flow is distributed on the cooling surface, and the cooling air flow can extend

Table 6: Design points and its variables.

Design points	α_1	α_2	d	q
DP1	−87.3	120.7	0.11	0.046
DP2	−60.3	118.9	0.1388	0.06
DP3	−53.1	108.1	0.0236	0.05
DP4	−65.7	117.1	0.1244	0.072
DP5	−78.3	135.1	0.038	0.034
DP6	−80.1	138.7	0.1892	0.056
DP7	−58.5	136.9	0.1172	0.062
DP8	−63.9	122.5	0.182	0.038
DP9	−76.5	131.5	0.074	0.064
DP10	−45.9	144.1	0.0596	0.036
DP11	−89.1	102.7	0.0884	0.03
DP12	−69.3	124.3	0.0524	0.07
DP13	−74.7	113.5	0.1028	0.068
DP14	−51.3	104.5	0.146	0.044
DP15	−81.9	140.5	0.0812	0.074
DP16	−56.7	115.3	0.1676	0.042
DP17	−71.1	142.3	0.1604	0.04
DP18	−49.5	106.3	0.1316	0.052
DP19	−72.9	109.9	0.0452	0.028
DP20	−47.7	111.7	0.0308	0.058
DP21	−62.1	133.3	0.0956	0.054
DP22	−85.5	100.9	0.1748	0.066
DP23	−54.9	126.1	0.1964	0.048
DP24	−83.7	129.7	0.0668	0.032
DP25	−67.5	127.9	0.1532	0.026

Table 7: Optimization results.

Design point	η_1 (%)	η_2 (%)	$\sigma_{\max}$ (MPa)	$T_{\max}$ (K)	T_{ave} (K)	q (kg s^{-1})
ODP	79.47	98	762	910	762	0.04
CDP1	82.53	54	748	861	746	0.0525
CDP2	82.49	36	744	857	746	0.0536
CDP3	82.30	38	740	851	747	0.056

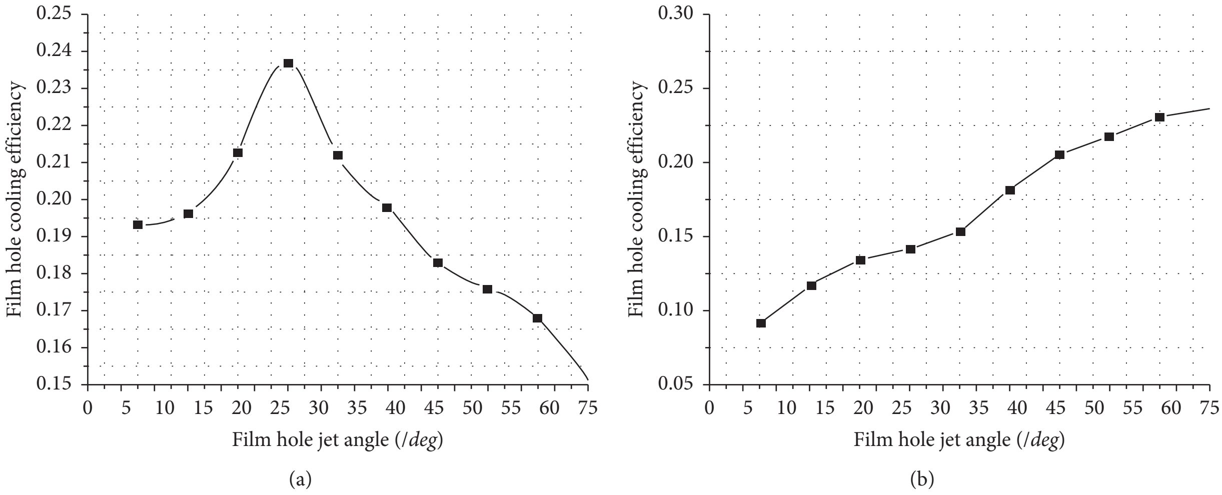


Figure 18: The effect of the compound angle on the average film hole cooling efficiency of the film hole area.

Table 8: Optimization results.

Parameters	Optimization results (design range)	Units
DS_rib_e	1.75 (0.5–2)	Mm
DS_rib_P	5.94 (5–10)	Mm
DS_rib_α	59.6 (50–130)	Deg

further away in the flow direction. However, in actual blades, due to the constraints of structure and manufacturing, it is impossible to produce gas film holes with a very small jet angle.

As shown in Figure 18, it can be seen that the film cooling efficiency increases with the increase of the compound angle. The increase in amplitude indicates that the compound angle has a greater effect at high blowing ratios. After the recombination angle is greater than 45°, the cooling efficiency of the gas film hole at a high blowing ratio becomes higher than that at a low blowing ratio, indicating that a large composite angle can fully exert the cooling capacity of the cooling air flow at a high blowing ratio.

According to the analysis results, it can be concluded that the gas film cooling efficiency should reach 0.2, and the injection angle and composite angle range are (18, 45) and (45, 75).

Simplify the cooling channel of the blade, through the parametric modeling module, enter the cooling channel parameters, and the simplified geometric model and analysis model are shown in the figure. The design range and optimization results of design parameters are shown in Table 8.

By optimizing the range of film cooling parameters and the optimization of cooling channels, the blades are optimized. The film cooling range has been effectively improved. The average cooling effect of the blade body cooling channel is better than the initial design.

5. Conclusions

(1) According to the model adaptability problem of complex product model in multiphysics simulation, the differences of model expression methods in design, modeling, and simulation were analyzed to improve the adaptability of complex product models to multiphysics simulation.

(2) The product design space and analytical space mapping method adapted to multiphysics simulation was studied. According to the existing principles and empirical formulas, the relationship between design parameters modeling parameters and simulation parameters was established by unified CADSS parameters, so that the analysis results can be quickly fed back to the modification of geometric models to break the isolation between design and analysis.

(3) According to the analysis and optimization requirements, blade model for CHT and FEA simulation was built from the top down under the CADSS framework. The multiphysics blade model containing simulation parameters adapts well to multidisciplinary simulation and can generate batch design point models for optimization.

(4) Through the blade's MDM by CADSS and optimization design of the film hole, 25 design points were generated for about 2 hours for each case which takes more time through conventional design and simulation process. Three candidate design points were chosen by MOGA and effectively promoted cooling efficiency with less promotion of cooling mass flow. The case studied in the paper shows the method proposed can improve design efficiency and shorten the design cycle.

Data Availability

The data that support the findings of the research are available from the corresponding author.

Conflicts of Interest

The authors declare that there are no conflicts of interest regarding the publication of this paper.

References

[1] M. S. Shephard, M. W. Beall, R. M. O'bara, and B. E. Webster, "Toward simulation-based design," *Finite Elements in Analysis and Design*, vol. 40, no. 12, pp. 1575–1598, 2004.

[2] Y. Zeng and I. Horváth, "Fundamentals of next generation CAD/E systems," *Computer-Aided Design*, vol. 44, no. 10, pp. 875–878, 2012.

[3] Y. Bodein, B. Rose, and E. Caillaud, "Explicit reference modeling methodology in parametric CAD system," *Computers in Industry*, vol. 65, no. 1, pp. 136–147, 2014.

[4] M. S. Smit and W. F. Bronsvoort, "Integration of design and analysis models," *Computer-Aided Design and Applications*, vol. 6, no. 6, pp. 795–808, 2009.

[5] S. Ando, H. Aoyama, and N. Sano, *CAD System for Utilizing Design Information and Know-How Based on Function Features*, American Society of Mechanical Engineers (ASME), London, UK, 2012.

[6] F. Tian and M. Voskuijl, "Automated generation of multiphysics simulation models to support multidisciplinary design optimization," *Advanced Engineering Informatics*, vol. 29, no. 4, pp. 1110–1125, 2015.

[7] J. R. R. A. Martins and A. B. Lambe, "Multidisciplinary design optimization: a survey of architectures," *AIAA Journal*, vol. 51, no. 9, pp. 2049–2075, 2013.

[8] X.-J. Ma, G.-F. Ding, S.-F. Qin et al., "Transforming multidisciplinary customer requirements to product design specifications," *Chinese Journal of Mechanical Engineering*, vol. 30, no. 5, pp. 1069–1080, 2017.

[9] O. Hamri, J. C. Léon, F. Giannini et al., "Software environment for CAD/CAE integration," *Advances in Engineering Software*, vol. 41, no. 10-11, pp. 1211–1222, 2010.

[10] Z. Chi, J. Ren, and H. Jiang, "Coupled aerothermodynamics optimization for the cooling system of a turbine vane," *Journal of Turbomachinery*, vol. 136, no. 5, 2014.

[11] Q. Li, P. Liu, and G. He, "Fluid-solid coupled simulation of the ignition transient of solid rocket motor," *Acta Astronautica*, vol. 110, pp. 180–190, 2015.

[12] B. Chen and Y. Xie, “A computational approach for the optimal conceptual design synthesis based on the distributed resource environment,” *International Journal of Production Research*, vol. 55, no. 20, pp. 5881–5901, 2017.

[13] X. Jiao, G. Zheng, P. A. Alexander et al., “A system integration framework for coupled multiphysics simulations,” *Engineering with Computers*, vol. 22, no. 3-4, pp. 293–309, 2006.

[14] B. C. Kim and D. Mun, “Feature-based simplification of boundary representation models using sequential iterative volume decomposition,” *Computers & Graphics*, vol. 38, pp. 97–107, 2014.

[15] G. P. Gujarathi and Y.-S. Ma, “Parametric CAD/CAE integration using a common data model,” *Journal of Manufacturing Systems*, vol. 30, no. 3, pp. 118–132, 2011.

[16] N. Aifaoui, D. Deneux, and R. Soenen, “Feature-based interoperability between design and analysis processes,” *Journal of Intelligent Manufacturing*, vol. 17, no. 1, pp. 13–27, 2006.

[17] M. S. Shephard, P. L. Baehmann, M. K. Georges et al., “Framework for the reliable generation and control of analysis idealizations,” *Computer Methods in Applied Mechanics and Engineering*, vol. 82, no. 1–3, pp. 257–280, 1990.

[18] S. H. Lee, “A CAD-CAE integration approach using feature-based multi-resolution and multi-abstraction modelling techniques,” *Computer-Aided Design*, vol. 37, no. 9, pp. 941–955, 2005.

[19] B. A. Szabó, “Geometric idealizations in finite element computations,” *Communications in Applied Numerical Methods*, vol. 4, no. 3, pp. 393–400, 1988.

[20] N. Aifaoui, D. Deneux, A. Benamara et al., “Mechanical analysis process modeling based on analysis features,” *IEEE*, vol. 3, p. 6, 2002.

[21] Z. Wang, L. Tian, and W. Duan, “Annotation and retrieval system of CAD models based on functional semantics,” *Chinese Journal of Mechanical Engineering*, vol. 27, no. 6, pp. 1112–1124, 2014.

[22] Z. Pan, X. Wang, R. Teng, and X. Cao, “Computer-aided design-while-engineering technology in top-down modeling of mechanical product,” *Computers in Industry*, vol. 75, pp. 151–161, 2016.

[23] V. Shapiro, I. Tsukanov, and A. Grishin, “Geometric issues in computer aided design/computer aided engineering integration,” *Journal of Computing and Information Science in Engineering*, vol. 11, no. 2, 2011.

[24] I. Matin, M. Hadzistevic, J. Hodolic et al., “A CAD/CAE-integrated injection mold design system for plastic products,” *The International Journal of Advanced Manufacturing Technology*, vol. 63, no. 5–8, pp. 595–607, 2012.

[25] R. S. Peak, R. E. Fulton, I. Nishigaki, and N. Okamoto, “Integrating engineering design and analysis using a multi-representation approach,” *Engineering with Computers*, vol. 14, no. 2, pp. 93–114, 1998.

[26] R. Sudarsan, S. J. Fenves, R. D. Sriram, and F. Wang, “A product information modeling framework for product lifecycle management,” *Computer-Aided Design*, vol. 37, no. 13, pp. 1399–1411, 2005.

[27] L. Li, C. F. Lange, and Y. Ma, “Association of design and computational fluid dynamics simulation intent in flow control product optimization,” *Proceedings of the Institution of Mechanical Engineers, Part B: Journal of Engineering Manufacture*, vol. 232, no. 13, pp. 2309–2322, 2018.

[28] M.-C. Hsu, C. Wang, A. J. Herrema, D. Schillinger, A. Ghoshal, and Y. Bazilevs, “An interactive geometry modeling and parametric design platform for isogeometric analysis,” *Computers & Mathematics with Applications*, vol. 70, no. 7, pp. 1481–1500, 2015.

[29] A. J. Herrema, N. M. Wiese, C. N. Darling, B. Ganapathysubramanian, A. Krishnamurthy, and M.-C. Hsu, “A framework for parametric design optimization using isogeometric analysis,” *Computer Methods in Applied Mechanics and Engineering*, vol. 316, pp. 944–965, 2017.

[30] E. Ferede, M. M. Abdalla, and G. J. W. van Bussel, “Isogeometric based framework for aeroelastic wind turbine blade analysis,” *Wind Energy*, vol. 20, no. 2, pp. 193–210, 2017.

[31] P. Kang and S.-K. Youn, “Isogeometric topology optimization of shell structures using trimmed NURBS surfaces,” *Finite Elements in Analysis and Design*, vol. 120, pp. 18–40, 2016.

[32] X. Li, C. D. LeiWei, and W. WangLi, “Statistical process monitoring with biogeography-based optimization independent component analysis,” *Mathematical Problems in Engineering*, vol. 2018, no. 1, pp. 1–14, 2018.

[33] J. Li, T. Ning, T. Wang, B. Hu, P. Xi, and J. Xu, “A rapid parameter configuration method for film hole component in pipe-net calculation,” *Proceedings of the Institution of Mechanical Engineers, Part A: Journal of Power and Energy*, vol. 234, no. 7, pp. 915–933, 2020.

[34] C. Li, *A Two-Dimensional Analytical Method for Turbine Blade Cooling Design*, American Society of Mechanical Engineers (ASME), London, UK, 2017.

[35] Z. Cheng and Y. Ma, “Explicit function-based design modelling methodology with features,” *Journal of Engineering Design*, vol. 28, no. 3, pp. 205–231, 2017.

[36] J. Li, T. Ning, P. Xi, B. Hu, and T. Wang, “An analysis-oriented parameter extraction method for features on freeform surface,” *Proceedings of the Institution of Mechanical Engineers, Part C: Journal of Mechanical Engineering Science*, vol. 233, no. 17, pp. 6005–6025, 2019.

[37] F. Wagner, “Multi-objective optimization of the cooling configuration of a high pressure turbine blade,” American Society of Mechanical Engineers (ASME), London, UK, 2018.

[38] R. J. Goldstein, “Film cooling,” *Advances in Heat Transfer*, vol. 23, pp. 321–379, 1971.

[39] X. Wang, H. Xu, J. Wang, W. Song, and M. Wang, “Multi-objective optimization of discrete film hole arrangement on a high pressure turbine end-wall with conjugate heat transfer simulations,” *International Journal of Heat and Fluid Flow*, vol. 78, 2019.

[40] Z. Wang, W.-J. C. Du, and S.-J. Li, “Multi-field coupling analysis on the film-cooling with transverse and arched trenches,” *Journal of Thermal Science and Technology*, vol. 14, no. 1, 2019.

[41] L. D. Hylton, M. S. Mihelc, E. R. Turner et al., “Analytical and Experimental Evaluation of the Heat Transfer Distribution over the Surfaces of Turbine vanes,” 1983.

Hindawi
Shock and Vibration
Volume 2020, Article ID 8873137, 15 pages
https://doi.org/10.1155/2020/8873137

Research Article

Learning Transferable Convolutional Proxy by SMI-Based Matching Technique

Wei Jin [1,2] and Nan Jia [3]

[1] *Beijing Academy of Science and Technology, Beijing 100094, China*
[2] *Beijing Institute of New Technology Applications, Beijing 100094, China*
[3] *School of Informatics and Cyber Security, People's Public Security University of China, Beijing 100038, China*

Correspondence should be addressed to Wei Jin; jinwei@bjast.ac.cn

Received 20 August 2020; Revised 22 September 2020; Accepted 28 September 2020; Published 14 October 2020

Academic Editor: Chaoqun Duan

Domain-transfer learning is a machine learning task to explore a source domain data set to help the learning problem in a target domain. Usually, the source domain has sufficient labeled data, while the target domain does not. In this paper, we propose a novel domain-transfer convolutional model by mapping a target domain data sample to a proxy in the source domain and applying a source domain model to the proxy for the purpose of prediction. In our framework, we firstly represent both source and target domains to feature vectors by two convolutional neural networks and then construct a proxy for each target domain sample in the source domain space. The proxy is supposed to be matched to the corresponding target domain sample convolutional representation vector well. To measure the matching quality, we proposed to maximize their squared-loss mutual information (SMI) between the proxy and target domain samples. We further develop a novel neural SMI estimator based on a parametric density ratio estimation function. Moreover, we also propose to minimize the classification error of both source domain samples and target domain proxies. The classification responses are also smoothened by manifolds of both the source domain and proxy space. By minimizing an objective function of SMI, classification error, and manifold regularization, we learn the convolutional networks of both source and target domains. In this way, the proxy of a target domain sample can be matched to the source domain data and thus benefits from the rich supervision information of the source domain. We design an iterative algorithm to update the parameters alternately and test it over benchmark data sets of abnormal behavior detection in video, Amazon product reviews sentiment analysis, etc.

1. Introduction

1.1. Background. With the rapid development in Internet technologies, more and more people are using the Internet and generating large-scale sets of behavior data [1, 2]. Nowadays, machine learning technologies are widely used for the automatic annotation of the data to extract insights into the purpose of decision-making [3, 4]. To learn the machine learning models for this purpose, we need sufficient data and the corresponding labels so that the model can learn from the data-label pairs to build the mapping from the data to the labels. Moreover, the size of such data-label pairs should be large enough to cover most of the patterns of the data. However, the labeling of the data is sometimes expensive. When the data are generated, it is usually not labeled automatically, and thus, the label is missing. To fill this gap, human label work is required to provide the label. In many cases, the labeling of data is time-consuming and costly. As a result, we can only label a limited amount of data and train the model with a large amount of unlabeled data with a small amount of labeled data [5, 6]. The lack of labeled data is a bottleneck for the machine learning tasks over big data. To solve this problem, various methods are developed, among which the most popular ones are the semisupervised learning and transfer learning [7, 8].

Remark 1. The research method of this paper is a new method of transfer learning, and the application objects include many different applications. Their close relations are explained as follows.

Transfer learning aims to borrow the sufficiently labeled data from another domain to help the learning of the model in a domain where the data are not sufficient. This problem is defined over a source domain with sufficient labeled data and a target domain with insufficient labeled data, but a large set of unlabeled data. The source domain and target domain have the same data space and label space. However, the data distribution of the two domains is different; thus, it is not suitable to directly merge them to train a model for the target domain. Due to the mismatch of the data distributions of source and target domains, it is important to transfer the knowledge from the source domain to the target domain. Transfer learning algorithms are designed for this purpose. For example, in the problem of video anomaly detection, the problem is to detect the anomalies from the frames of the video, and the anomalies include cycle, skater, truck, car, wheelchair, and baby cart. In the real world, the data of video are usually collected from subway stations, communities, shopping centers, etc., which are treated as the target domain. Moreover, these data are usually not labeled at all, or only a small part is labeled. Meanwhile, we can use a fully labeled data set collected and annotated by the University of California and San Diego (UCSD) [9, 10] as the source domain data to help the training of the target domain model. However, the video of UCSD is usually captured on the campus of the university, which is different from the target domain's locations, such as shopping centers and subway stations. Thus, the data of source and target domains do not match perfectly, and the mismatching of the two domain's video requires transferring the model trained over the source domain to the target domain to adapt the target domain data. Otherwise, the model trained over the source domain cannot fit the target domain.

Another example is the Amazon product review sentiment analysis [11, 12]. This task is to detect the sentiment (positive/negative) from the text content of a review of a product. For different product types, the buyers may have different review styles. Some product types have sufficient labeled data, while other types are in lack of labeled data. Thus, we want to use the well-labeled product reviews as a source domain to help the learning of sentiment detectors for the limited-labeled reviews of other products. Again, the domain transfer is needed to adopt the model from the source domain product to the target domain product.

In the transfer learning models, the deep convolutional neural network (CNN) [13–17] is the most popular base models. It is composed of multiple convolutional layers and max-pooling layers. The convolutional layers use a filter bank to extract the location features by sliding over the input data, such as image or sentence. This model can extract both local and hierarchical features from data and thus can perform automatic engineering effectively. In this paper, we study the problem of transfer learning over a source domain and a target domain. We proposed a novel transfer learning method based on a convolutional neural network (CNN) [13] and use the squared-loss mutual information (SMI) [18] to measure the matching between the source and the target domain.

1.2. Existing Works. In traditional methods of transfer learning, to match the source and target domains, various methods are applied, such as mutual information minimization (MIM) [19], Hilbert Schmidt independence criterion (HSIC) maximization [20], and maximum mean discrepancy (MMD) minimization [21, 22]. Moreover, to train the transfer learning models, the most popular base model is the deep CNN model, which is also adapted in our paper. In this paper, we briefly introduce some state-of-the-arts relevant to our work of CNN-based transfer learning.

Long et al. [21, 23] proposed a deep CNN model for transfer learning. This model is a two-branch network. The first four convolutional layers are shared by both source and target domains, and the last three full-connection layers are domain-specific. However, in the last three layers, the output features are forced to be in the same spaces by the MMD minimization. The MMD minimization measures mismatch in the features of each layer of two domain domains by the squared ℓ_2 distance of the means of two domains.

Zhang et al. [14] proposed a CNN model learning method for transfer learning to present both data features and attributes. The structure of this network is composed of a shared attribute-embedding CNN across domains, a domain-independent CNN, and a domain-dependent CNN for the embedding of the original data of each data sample. The outputs of these three CNN models are concatenated as the overall features of one sample and used to predict the class label. The attribute-embedding CNN is imposed to approximate the transformed attributes accurately, and the domain-independent CNN outputs are supposed to minimize the MMD of two domains.

Wang et al. [19] developed a novel CNN model learning method for both source and target domains. The CNN outputs are imposed to be independent of the domain; i.e., from the CNN model output of a data sample, we should not be able to tell which domain it is from. To measure the independence of the CNN outputs and the domain indicator, the mutual information is applied. The CNN models are learned by minimizing the mutual information between the CNN outputs and the domain indicators to obtain the domain-independent CNN representations.

Geand and Yu [24] designed a learning method of deep CNN model for transfer learning. This method is based on the source-target selective joint fine-tuning. This method is not to use all the data samples of the source domain but only use a subset of the training samples of the source domain. The selected subset is similar to the target domain samples at the low-level characteristics. To this end, the algorithm constructs the descriptors from the responses of the filter bank over the training samples. These descriptors are used to search for the subset of training samples used for the learning problem.

1.3. Our Contributions. In this paper, we propose a novel CNN-based deep learning framework for transfer learning. Our work is motivated by a phenomenon: the mismatch between the source domain and target domain is usually at the entire data set level but at the data sample level. Moreover, it is also difficult to find the one-to-one matching between the samples of the source and target domains. Thus, we propose to construct a proxy of each target domain in the source domain. This proxy is constructed by a linear combination of the source domain samples. We also argue that the CNN model is a powerful tool to extract the features of both source and target domains and an optimal choice for the proxy construction. Thus, we first extract the convolutional features from both source and target samples and then apply the proxy construction. We hope the constructed proxy of each target domain can match the original target sample convolutional representation as well as possible. In this paper, we propose to use an information theory measure to measure the quality of matching between the proxy and convolutional features of target domain samples. This measure is the squared-loss mutual information (SMI) [25, 26], which measures the Pearson divergence of the joint probability of proxies and convolutional features and the product of the independent profanities of proxies and convolutional features. We further release the calculation of the SMI by firstly designing a parametric density ratio estimator of proxy and convolutional feature of target domain samples and then approximate it as the squared matching error of the true density ratio against the density ratio estimator.

To learn the proxy construction coefficients, we model the learning problem as a minimization problem. In this minimization problem, we minimize the SMI, the classification loss of labeled samples, and the classification response entropy of the unlabeled samples of both domains. Moreover, the manifold regularization and the model complexity are also considered in the minimization problem. The contributions of this paper are listed as follows:

(1) We build a novel transfer learning schema, which firstly represents the source and target domains by CNN models and then constructs proxies for the target samples from the source domain. The construction is guild by the SMI minimization. In this way, the matching of the two domains is directly applied at the proxy level, while the CNN representations can still keep the domain characteristics. A novel SMI approximation is proposed as the squared matching error between the density ratio function and its parametric estimator.

(2) We model the problem as a minimization problem with SMI, classification loss, classification entropy, manifold regularization, and model complexity regularization. This minimization is a joint learning framework to optimize the CNN model parameters and proxy construction coefficients simultaneously.

(3) We develop an iterative algorithm to solve the minimization problem based on the alternating direction method of multipliers (ADMM) [27]. Moreover, the parameter of the density ratio estimator of the SMI is also updated in this algorithm. With this algorithm, the SMI parameter is also automatically approximated together with the other variables of the objective.

1.4. Paper Organization. This paper is organized as follows: in Section 2, we introduce the proposed transfer learning method based on CNN and proxy learning; in Section 3, we evaluate the proposed method experimentally over some benchmark data sets; in Section 4, we give the conclusion of this paper and some future works.

2. Proposed Method

In this section, we will propose the novel transfer convolutional neural network learning method. Firstly, we will build the objective function for the learning of the source and target domain convolutional network, and then, we will discuss how to minimize this objective function and finally develop an iterative algorithm based on the optimization results.

2.1. Objective Function. Suppose we have a training set of two domains. In the target domain, we have n training samples $\{x_1^t, \dots, x_n^t\}$, where x_i^t is the i-th training sample of the target domain. To represent a target domain sample x^t, we use a deep CNN model, f, to convert it to a d_t-dimensional vector:

$$f = f\left(x^t; \Theta\right) \in R^{d_t}. \quad (1)$$

Meanwhile, we have a set of source domain training set $\{x_1^s, \dots, x_m^s\}$, which has m samples and x_j^s is the j-th source domain training sample. Similar to the target domain, we use another deep CNN model, g, to convert a source domain sample, x^s to a d_s-dimensional vector:

$$g = g\left(x^s; \Phi\right) \in R^{d_s}. \quad (2)$$

To learn the network parameters Θ and Φ, we discuss the following problems.

2.1.1. Squared-Loss Mutual Information. Due to the mismatch of the source and target domains, we propose to leverage the two domains by constructing a proxy for each target domain sample in the source domain space. To this send, for a target domain sample, *xti*, we denote its proxy $z_i \in R^{d_s}$, as a d_s-dimensional vector. Moreover, we impose that it can be reconstructed by a learning combination of the convolutional representations of source domain samples:

$$z_i = \sum_{j=1}^{m} \alpha_{ij} g_j, \quad \text{s.t.} \sum_{j=1}^{m} \alpha_{ij} = 1, \quad 1 \geq \alpha_{ij} \geq 0, \quad (3)$$

where $g_j = g(x_j^s; \Phi)$ is the convolutional vector of the j-th source domain sample and α_{ij} is the nonnegative weight of the j-th source domain sample for the construction of the i-th proxy.

To encourage the matching of the source and target domains, we argue to measure their quality of matching by the density ration of the proxies and convolutional vectors of the target domain samples z and f. For this purpose, we firstly measure the density ratio of z and f, denoted as $r(f,z)$. According to the definition of density ratio [18],

$$r(f,z)=\frac{p(f,z)}{p(f)p(z)}, \tag{4}$$

where $p(f,z)$ is the joint probability of f and z, $p(f)$ is the probability of f, and $p(z)$ is the probability of z. Since it is difficult to directly estimate the density ratio, we propose to learn a parametric density ratio estimator function:

$$\hat{r}(f,z;\phi)=h\left(\phi^{\top}\begin{bmatrix} f\\ z\\ f-z\end{bmatrix}\right), \tag{5}$$

where $p(f,z)$ is the joint probability of f and z, $p(f)$ is the probability of f, and $p(z)$ is the probability of z. Since it is difficult to directly estimate the density ratio, we propose to learn a parametric density ratio estimator function. $\begin{bmatrix} f\\ z\\ f-z\end{bmatrix}$ is the concatenation of the f, z, and their element-wise difference vector, $\phi\in\mathbb{R}^{3\mathbf{d}_s}$ is the parameter vector, and $h(x)=1/(1+\exp(-x))$ is the Sigmoid activation function. To learn ϕ, we minimize the SMI between $r(f,z)$ and $\hat{r}(f,z;\phi)$, defined as follows:

$$\begin{aligned}
SMI(\phi) &= \int_f\int_z [r(f,z)-\hat{r}(f,z;\phi)]^2 p(f)p(z)\mathrm{d}f\mathrm{d}z\\
&= \int_f\int_z \left[r(f,z)^2-2r(f,z)\hat{r}(f,z;\phi)+\hat{r}(f,z;\phi)^2\right]p(f)p(z)\mathrm{d}f\mathrm{d}z\\
&= \int_f\int_z r(f,z)^2 p(f)p(z)\mathrm{d}f\mathrm{d}z-2\int_f\int_z r(f,z)\hat{r}(f,z;\phi)p(f)p(z)\mathrm{d}f\mathrm{d}z\\
&\quad+\int_f\int_z \hat{r}(f,z;\phi)^2 p(f)p(z)\mathrm{d}f\mathrm{d}z\\
&= \int_f\int_z r(f,z)^2 p(f)p(z)\mathrm{d}f\mathrm{d}z-2\int_f\int_z p(f,z)\hat{r}(f,z;\phi)\mathrm{d}f\mathrm{d}z\\
&\quad+\int_f\int_z \hat{r}(f,z;\phi)^2 p(f)p(z)\mathrm{d}f\mathrm{d}z.
\end{aligned} \tag{6}$$

To estimate the SMI given the training samples of target domain and their proxies, we collect the set of convolutional vectors and corresponding proxies $\{(f_1,z_1),\ldots,(f_n,z_n)\}$. We assume the distributions of f, z, and (f,z) are uniform distributions, which lead the probability functions as

$$\begin{aligned}
p(f)&=\frac{1}{n},\\
p(z)&=\frac{1}{n},\\
p(f,z)&=\frac{1}{n}.
\end{aligned} \tag{7}$$

With probabilities in (7) and the empirical approximation, we rewrite the second and third terms of (6) as follows:

$$\begin{aligned}
\int_f\int_z p(f,z)r(f,z;\phi)\mathrm{d}f\mathrm{d}z &= \frac{1}{n}\int_f\int_z r(f,z;\phi)\mathrm{d}f\mathrm{d}z=\frac{1}{n}\sum_{i=1}^{n} r(f_i,z_i;\phi),\\
\int_f\int_z \hat{r}(f,z;\phi)^2 p(f)p(z)\mathrm{d}f\mathrm{d}z &= \frac{1}{n^2}\sum_{i,i_\prime=1}^{n}\hat{r}(f_i,z_i;\phi)\hat{r}\left(f_{i_\prime},z_{i_\prime};\phi\right).
\end{aligned} \tag{8}$$

By substituting (8) to (6), we have the approximated SMI as

$$\mathrm{SMI}(\phi) = C - \frac{2}{n}\sum_{i=1}^{n} r(f_i, z_i; \phi) + \frac{1}{n^2}\sum_{i,i'=1}^{n} r(f_i, z_i; \phi) r(f_{i'}, z_{i'}; \phi), \tag{9}$$

where $C = \int\int r(f,z)^2 p(f) p(z) \mathrm{d}f \mathrm{d}z$ is a constant. To obtain a good quality estimator, we propose to minimize the squared-loss mutual information learned from the parameter of $\hat{r}(f, z; \phi)$:

$$\phi^* = \mathrm{argmin}_{\phi}\mathrm{SMI}(\phi). \tag{10}$$

With this optimal ϕ^*, we have to use the SMI as a measure of the matching quality between f and z. Using this measure as a term of loss for the purpose of learning the convolutional representations f and proxy $\mathbf{z}$, we organize these parameters as matrices as $F = [f_1, \ldots, f_n]$ and $Z = [z_1, \ldots, z_n]$. With these variables and the SMI as the matching measure, we seek to learn them to maximize the SMI. In other words, we minimize a loss function derived from SMI, defined as $\mathscr{M}(F, Z; \phi^*)$:

$$\begin{aligned} \mathscr{M}(F, Z; \phi^*) &= \frac{2}{n}\sum_{i=1}^{n} \hat{r}(f_i, z_i; \phi^*) - \frac{1}{n^2}\sum_{i,i_{\prime}=1}^{n} \hat{r}(f_i, z_i; \phi^*)\hat{r}\left(f_{i_{\prime}}, z_{i_{\prime}}; \phi^*\right) \\ &= -\frac{2}{n}\sum_{i=1}^{n} h(\phi^{*\top}[f_i z_i f_i - z_i]) \\ &\quad + \frac{1}{n^2}\sum_{i,i'=1}^{n} h(\phi^{*\top}[f_i z_i f_i - z_i]) h(\phi^{*\top}[f_{i'} z_{i'} f_{i'} - z_{i'}]). \end{aligned} \tag{11}$$

We rewrite

$$\phi^{*\top}[f_i z_i f_i - z_i] = \phi_1^\top f_i + \phi_2^\top z_i + \phi_3^\top (f_i - z_i) = (\phi_1 + \phi_3)^\top f_i + (\phi_2 - \phi_3)^\top z_i, \tag{12}$$

where $\phi^* = \begin{bmatrix} \phi_1 \\ \phi_2 \\ \phi_3 \end{bmatrix}$ accordingly.

$$\begin{aligned} \mathscr{M}(F, Z; \phi^*) = &-\frac{2}{n}\sum_{i=1}^{n} h\left((\phi_1 + \phi_3)^\top f_i + (\phi_2 - \phi_3)^\top z_i\right) \\ &+ \frac{1}{n^2}\sum_{i,i_{\prime}=1}^{n} h\left((\phi_1 + \phi_3)^\top f_i + (\phi_2 - \phi_3)^\top z_i\right) h\left((\phi_1 + \phi_3)^\top f_{i_{\prime}} + (\phi_2 - \phi_3)^\top z_{i_{\prime}}\right). \end{aligned} \tag{13}$$

The squared-loss mutual information regularization problem is modelled as

$$\min_{F,Z} \mathscr{M}(F, Z; \phi^*). \tag{14}$$

By minimizing this objective, we seek to learn the convolutional representation vectors of both source and target domain samples, target domain sample proxies, and other parameters to match the two domains as much as possible. Please note that, in this problem, we are not matching the source and target domain samples directly but seek to match target domain samples and the proxies of target domain samples in the source domain space.

2.1.2. Semisupervised Classification. In both the source and target domains, there are both labeled and unlabeled data samples. To learn from them, we use the semisupervised method. We discuss the semisupervised learning problems in both source and target domains as follows.

Source Domain. To approximate the class label vector, $y^s \in \{1, 0\}^{\mathbf{c}}$, of a source domain sample, x^s, from its convolutional feature vector, g, we design a linear classification function:

$$y^s = W^\top g, \tag{15}$$

where $W \in \mathbb{R}^{d_s \times c}$ is the parameter vector of the classifier and $\hat{y}^s \in \mathbb{R}^c$ is the classification response. To learn the classifier parameter W, we utilize both the labeled and unlabeled samples of the source domain. We denote the set of the labeled samples of source domain as $\mathscr{L}^s$ and the set of unlabeled as $\boldsymbol{\mathscr{U}}^s$. For the samples in $\mathscr{L}^s$, we minimize their classification errors measured by the squared loss:

$$\min_{W,G}\left\{O_1(W,G)=\sum_{j:\ x_j^s\in L^s}\left|y_j^s-W^\top g_j\right|_2^2\right\},\tag{16}$$

where y_j^s is the class label vector of the x_j and $G=[g_1,\ldots,g_m]$. For the samples in $\mathscr{L}^s$, even we do not have the available labels to constrain the classification responses in $\widehat{y}^s$, and we still can regularize the learning of the parameter by its entropy. For a sample x_j^s, we hope the uncertainty of its classification responses in $\widehat{y}_j^s$ is as small as possible. The uncertainty is measured by its entropy:

$$H\left(y_j^s\right)=-\sum_{k=1}^{c}y_{jk}^s\log\left(y_{jk}^s\right)=-\sum_{k=1}^{c}\left(w_k^\top g_j\right)\log\left(w_k^\top g_j\right),\tag{17}$$

where $\widehat{y}_{jk}^s=w_k^\top g_j$ is the k-th dimension $\widehat{y}_j^s$ and w_k is the k-th column of W. We minimize the entropy of the source samples in $\boldsymbol{\mathscr{U}}^s$ as follows:

$$\min_{W,G}\left\{O_2(W,G)=-\sum_{j:\ x_j^s\in U^s}\left(\sum_{k=1}^{c}\left(w_k^\top g_j\right)\log\left(w_k^\top g_j\right)\right)\right\}.\tag{18}$$

Target Domain. In the target domain, we also have labeled set denoted as $\mathscr{L}^t$ and unlabeled set denoted as $\boldsymbol{\mathscr{U}}^t$. To use the labeled samples of the target domain, instead of learning a classifier in the target domain space, we design a classifier in the target domain sample proxy space. For a proxy z, we apply a linear classifier function to approximate its label vector:

$$\widehat{y}^t=V^\top z,\tag{19}$$

where $V\in R^{d_t\times c}$ is the classifier parameter and $\widehat{y}^t\in\mathbb{R}^c$ is the estimated class label response vector. For a labeled sample, $x_i^t\in\mathscr{L}^t$, its class label vector $y_i^t\in\{1,0\}^c$ is known, and we minimize the classification loss measured by the squared loss between y_i^t and $\widehat{y}_i^t$:

$$\min_{V,Z}\left\{O_3(V,Z)=\sum_{qi:\ x_i^t\in L^t}\left|y_i^t-V^\top z_i\right|_2^2\right\}.\tag{20}$$

Meanwhile, for the samples in the target domain $\boldsymbol{\mathscr{U}}^t$, we also hope their uncertainty can be minimized, and accordingly, we minimize the entropy of the classification responses:

$$\min_{V,Z}\left\{O_4(V,Z)=-\sum_{i:\ x_i^t\in U^t}\left(\sum_{k=1}^{c}\left(v_k^\top z_i\right)\log\left(v_k^\top z_i\right)\right)\right\}.\tag{21}$$

2.1.3. Manifold Regularization. Moreover, we also hope the classification responses of samples in both source and target domains can be smooth over their manifold. To be specific, the classification responses of two neighboring samples should also be similar. To this end, we first construct the nearest neighbor graphs for both the source and target domains and then use the graphs to regularize their classification responses.

Source Domain. We firstly build a convolutional graph from the convolutional representations of the source domain samples $\{g_1,\ldots,g_m\}$. For each sample x_j^s, its nearest neighbor set is defined as $\boldsymbol{\mathscr{N}}_j^s$, and we calculate its affinity between itself and each neighbor $x_{j'}^s\in\boldsymbol{\mathscr{N}}_j^s$ based on a Gaussian kernel:

$$A_{j'j}^s=\begin{cases}\dfrac{\psi\left(g_j-g_{j'}\right)}{\sum_{j'':\ x_{j''}^s\in N_j^s}\psi\left(g_j-g_{j''}\right)}, & if\ x_{'j}^s\in N_j^s,\\[2ex] 0, & \text{otherwise},\end{cases}\tag{22}$$

where $\psi(x)=\exp-\left(x^2/\sigma^2\right)$ is the Gaussian kernel function. Based on the neighborhood affinity, we hope the squared ℓ_2 norm distance between g_j and its neighbors can be minimized so that the neighborhood structure can be kept in the convolutional representation space:

$$\min_{G}\left\{O_5(G)=\sum_{j,j'=1}^{m}A_{jj'}^s\left|g_j-g_{j'}\right|_2^2\right\}.\tag{23}$$

Target Domain. In the target domain, we firstly construct the graph from the original convolutional representations of the target samples $\{f_1,\ldots,f_n\}$, not the proxies. Given a target sample x_i^t, we find its nearest neighbors and denote the set of nearest neighbors $\boldsymbol{\mathscr{N}}_i^t$. We calculate the similarity between x_i^t and a sample $x_{i'}^t\in\boldsymbol{\mathscr{N}}_i^t$ also by kernel function:

$$A_{ii'}^t=\begin{cases}\dfrac{\psi\left(f_i-f_{i'}\right)}{\sum_{i'':\ x_{i''}^t\in N_i^t}\psi\left(f_j-f_{i''}\right)}, & \text{if } x_{i'}^t\in N_i^t,\\[2ex] 0, & \text{otherwise}.\end{cases}\tag{24}$$

With this affinity measure, different from the source domain, we use it to regularize the learning of the proxies of the target domain, $\{z_1,\ldots,z_n\}$, i.e.,

$$\min_{Z}\left\{O_6(Z)=\sum_{i,i'=1}^{m}A_{ii'}^t\left|z_i-z_{i'}\right|_2^2\right\}.\tag{25}$$

2.1.4. Overall Objective Function. Our overall object is the combination of the above objectives:

$$\begin{aligned}
O(F,G,Z,A,W,V,\theta,\Phi) &= M\left(F,G,Z,A;\phi^{*}\right) + C_1O_1(W,G) + C_2O_2(W,G)\\
&\quad + C_3O_3(V,Z) + C_4O_4(V,Z) + C_5O_5(G) + C_6O_6(Z) + C_7\left(|W|_2^2 + |V|_2^2 + |\Theta|_2^2 + |\Phi|_2^2\right)\\
&= \frac{2}{n}\sum_{i=1}^{n}\left(h\left(\phi_1+\phi_3\right)^T f_i + \left(\phi_2-\phi_3\right)^T Z_i\right)\\
&\quad + \frac{1}{n^2}\sum_{i,i'=1}^{n}\left(h\left(\phi_1+\phi_3\right)^T f_i + \left(\phi_2-\phi_3\right)^T Z_i\right)h\left(\phi_1+\phi_3\right)^T f_{i'}\\
&\quad + \left(\phi_2-\phi_3\right)^T + C_1\sum_{j:\ x_J^s\ \in\ L^s}\left|y_j^s - W^T - g_j\right|_2^2 - C_2\sum_{j:\ x_J^s\in U^s}\left(\sum_{k=1}^{c}\left(W_k^T g_j\right)\log\left(W_k^T g_j\right)\right)\\
&\quad + C_3\sum_{i:\ x_i^t\in L^t}\left|y_i^t - V^T Z_i\right|_2^2 - C_4\sum_{i:\ x_i^t\in U^t}\left(\sum_{k=1}^{c}\left(V_k^T Z_i\right)\log\left(V_k^T Z_i\right)\right) + C_5\sum_{j,j'-1}^{m}A_{jj'}^s\left|g_j - g_{j'}\right|_2^2\\
&\quad + C_6\sum_{i,i'=1}^{m}A_{ii'}^t\left|Z - Zi'\right|_2^2 + C_7\left(|W|_2^2 + |V|_2^2 + |\Theta|_2^2 + |\Phi|_2^2\right),\\
&s.t.\ \forall i = 1,\ldots,n\colon\ f_i = f\left(x_i^t;\Theta\right), z_i = \sum_{j=1}^{m}\alpha_{ij}g_j, \sum_{j=1}^{m}\alpha_{ij} = 1, \text{ and } 1 \ge \alpha_{ij} \ge 0,\\
&\forall j = 1,\ldots,m\colon\ g_j = g\left(x_j^s;\Phi\right),
\end{aligned} \tag{26}$$

where $A = \begin{bmatrix}\alpha_{11}, & \cdots, & \alpha_{n1}\\ \vdots & \ddots & \vdots\\ \alpha_{1m}, & \cdots, & \alpha_{nm},\end{bmatrix} \in \mathbb{R}_{+}^{m\times n}$ and $C_i, i = 1,\ldots,7$ are the tradeoff parameters, which weight the loss terms of the objective. The last term is the squared ℓ_2 norms of the mode parameters to prevent the overfitting problem. In this objective, we impose the following condition:

(1) In the target domain, we hope the SMI between the samples and their proxies constructed from the source domain can be maximized so that the source and target domains can be matched well.

(2) In both target and source domains, we hope the classification function can approximate the ground truth of labeled well, while the uncertainty of classification results of unlabeled can be minimized.

(3) Again, in both source and target domains, we hope the classification results can be similar among the neighborhoods.

(4) Finally, we hope the overall complexity of the model can be minimized as much as possible; thus, we minimize the squared ℓ_2 norm of the parameters of the model in the last term.

With this objective, we model the learning problem as a minimization problem:

$$\begin{aligned}
&\min_{F,G,Z,A,W,V,\Theta,\Phi}\ \mathcal{O}(F,G,Z,A,W,V,\Theta,\Phi),\\
&s.t.\ \forall i = 1,\ldots,n\colon\ f_i = f\left(x_i^t;\Theta\right), z_i = \sum_{j=1}^{m}\alpha_{ij}g_j, \sum_{j=1}^{m}\alpha_{ij} = 1,\ an\,d\ 1 \ge \alpha_{ij} \ge 0,\\
&\forall j = 1,\ldots,m\colon\ g_j = g\left(x_j^s;\Phi\right).
\end{aligned} \tag{27}$$

By solving this problem, we can learn the source and target CNN models W and V and the classification layer parameters Θ and Φ. Moreover, the proxy parameters of the target samples, A, are also optimized together with the model parameters to match the source and target domains.

2.2. Problem Optimization. To solve the problem at (27), we adopt the ADMM algorithm, and according, we have the following argued Lagrangian function:

$$\begin{aligned}
&\mathscr{L}(F, G, Z, A, W, V, \Theta, \Phi, \pi, \Psi, \Xi, \varsigma) \\
&\quad = \frac{2}{n} \sum_{i=1}^{n} h\left((\phi_1 + \phi_3)^T f_i + (\phi_2 + \phi_3)^T z_i \right) \\
&\quad + \frac{1}{n^2} \sum_{i,i'=1}^{n} h\left((\phi_1 + \phi_3)^T f_i + (\phi_2 + \phi_3)^T z_i \right) h\left((\phi_1 + \phi_3)^T f_{i'} + (\phi_2 + \phi_3)^T z_{i'} \right) \\
&\quad + C_1 \sum_{j:\ x_s^j \in L^s} \left| y_j^s - W^T g_j \right|_2^2 + C_2 \sum_{j:\ x_s^j \in U^s} \left(\sum_{k=1}^{c} \left(W_k^T g_j \right) \log\left(W_k^T g_j \right) \right) \\
&\quad + C_3 \sum_{i:\ x_i^t \in L^T} \left| y_i^t - V^T z_i \right|_2^2 \\
&\quad - C_4 \sum_{i:\ x_i^t \in U^T} \sum_{k=1}^{c} \left(V_k^T z_i \right) \log\left(V_k^T z_i \right) + C_5 \sum_{j,j'=1}^{m} A_{jj'}^s \left| g_j - g_{j'} \right|_2^2 \\
&\quad + C_6 \sum_{i,i'=1}^{m} A_{ii'}^t \left| g_z - g_{z'} \right|_2^2 + C_7 \left(|W|_2^2 + |V|_2^2 + |\Theta|_2^2 + |\Phi|_2^2 \right) \\
&\quad + \sum_{i=1}^{n} \pi_i^T \left(f_i - f\left(x_i^t; \Theta \right) \right) + \frac{\rho_1}{2} \sum_{i=1}^{n} \left| f_i - f\left(x_i^t; \Theta \right) \right|_2^2 + \sum_{j=1}^{m} \varpi_j^T \left(g_j - g\left(x_j^s; \Phi \right) \right) \\
&\quad + \frac{\rho_2}{2} \sum_{j=1}^{n} \left| g_j - g\left(x_j^s; \Phi \right) \right|_2^2 + \sum_{i=1}^{n} \tau_i^T \left(z_i - \sum_{j=1}^{m} \alpha_{ij} g_j \right) \\
&\quad + \frac{\rho_3}{2} \sum_{i=1}^{n} \left| z_i - \sum_{j=1}^{m} \alpha_{ij} g_j \right|_2^2 + \sum_{i=1}^{n} \varsigma_i^T \left(\sum_{j=1}^{m} \alpha_{ij} - 1 \right) \\
&\quad + \frac{\rho_4}{2} \sum_{i=1}^{n} \left| \sum_{j=1}^{m} \alpha_{ij} - 1 \right| \\
&\text{s.t. } \forall i = 1, \ldots, n,\ j = 1, \ldots, m\colon\ 1 \geq \alpha_{ij} \geq 0,
\end{aligned} \tag{28}$$

where the following holds:

(i) $\pi_{\mathbf{i}}$ is the dual variable of constraint $f_i = f(x_i^t; \Theta)$, ϖ_j is the dual variable of constraint $g_j = g(x_j^s; \Phi)$, $\tau_{\mathbf{i}}$ is the dual variable of constraint $z_i = \sum_{j=1}^{m} \alpha_{ij} g_j$, and ς_i is the dual variable of constraint $\sum_{j=1}^{m} \alpha_{ij} = 1$.

(ii)
$\Pi = [\pi_1, \ldots, \pi_n], \Psi = [\varpi_1, \ldots, \varpi_m], \Xi = [\tau_1, \ldots, \tau_n]$, and $= [\varsigma_1, \cdots, \varsigma_n]^\top$.

(iii) $\rho_k, k = 1, \ldots, 4$ are the corresponding penalty parameters of the constraints.

According to the ADMM algorithm, we optimize the parameters and the dual variables iteratively as follows.

Optimizing F. To optimize the target convolutional vectors of F, we update f_i one by one. The subgradient of $\mathscr{L}$ with regard to f_i is

$$\nabla_{f_i}\mathscr{L}(f_i) = \left(\frac{1}{n^2}\sum_{i'=1}^{n} h\Big(\left(\phi_1+\phi_3\right)^{\top} f_{i'} + \left(\phi_2-\phi_3\right)^{\top} z_{i'}\Big) - \frac{2}{n}\right)$$
$$\times \nabla h\Big(\left(\phi_1+\phi_3\right)^{\top} f_i + \left(\phi_2-\phi_3\right)^{\top} z_i\Big)\left(\phi_1+\phi_3\right) + \pi_i + \rho_1\Big(f_i - f\big(x_i^t;\Theta\big)\Big), \tag{29}$$
$$\textit{where}\ \nabla h(x) = h(x)\times(1-h(x)).$$

The subgradient descent step to update f_i is

$$f_i \leftarrow f_i - \eta\nabla_{f_i}\mathscr{L}(f_i). \tag{30}$$

Optimizing G. Similar to F, we also optimize g_j of G one by one, and the following subgradient algorithm is used to update g_j:

$$g_j \leftarrow g_j - \eta\nabla_{g_j}\mathscr{L}\big(g_j\big),$$
$$\text{where}\ \nabla_{g_j}\mathscr{L}\big(g_j\big) = 2C_1 I\big(x_j^s \in \mathscr{L}^s\big)W\big(y_j^s - W^{\top} g_j\big) - C_2 I\big(x_j^s \in \boldsymbol{\mathscr{U}}^s\big)\left(\sum_{k=1}^{c}\Big(\log\big(w_k^{\top} g_j\big) + \big(w_k^{\top} g_j\big)\log\big(w_k^{\top} g_j\big)\Big)w_k\right) \tag{31}$$
$$+\, 2C_5\sum_{j_{\prime}=1}^{m} +\varpi_j + \rho_2\sum_{j=1}^{m} A_{j'j}^{s}\Big(g_j - g_{j_{\prime}}\Big)\Big(g_j - g\big(x_j^s;\Phi\big)\Big) - \sum_{i=1}^{n}\alpha_{ij}\tau_i + \frac{\rho_3}{2}\sum_{i=1}^{n}\left(\sum_{j_{\prime}=1}^{m}\alpha_{ij}\alpha_{ij_{\prime}}g_{j_{\prime}} - 2\alpha_{ij}z_i\right),$$

where $I(x) = 1$ if x is true and 0 otherwise.

Optimizing Z. We use the subgradient descent algorithm to update $\mathbf{z_i}$ one by one, and the updating rule is

$$z_i \leftarrow z_i - \eta\nabla_{z_i}\mathscr{L}(z_i)$$
$$\text{where}\ \nabla_{z_i}\mathscr{L}(z_i) = \left(\frac{1}{n^2}\sum_{i_{\prime}=1}^{n} h\Big(\left(\phi_1+\phi_3\right)^{\top} f_{i_{\prime}} + \left(\phi_2-\phi_3\right)^{\top} z_{i_{\prime}}\Big) - \frac{2}{n}\right)\times\nabla h\Big(\left(\phi_1+\phi_3\right)^{\top} f_i + \left(\phi_2-\phi_3\right)^{\top} z_i\Big)\left(\phi_2-\phi_3\right)$$
$$-\, 2C_3 I\big(x_i^t \in L^t\big)V\big(y_i^t - V^{\top} z_i\big) - C_4 I\big(x_i^t \in U^t\big)\left(\sum_{k=1}^{c}\left(\log\left(v_k^{\top} z_i\right) + \left(v_k^{\top} z_i\right)\log\left(v_k^{\top} z_i\right)\right)v_k\right) \tag{32}$$
$$+\, 2C_6\sum_{i_{\prime}=1}^{m} A_{i_{\prime}i}^{t}\Big(z_i - z_{i_{\prime}}\Big) + \tau_i + \rho_3\left(z_i - \sum_{j=1}^{m}\alpha_{ij}g_j\right).$$

Optimizing Θ. The gradient descent algorithm is applied to update Θ:

$$\Theta \leftarrow \Theta - \eta\nabla_{\Theta}\mathscr{L}(\Theta),$$
$$\text{where}\ \nabla_{\Theta}\mathscr{L}(\Theta) = \sum_{i=1}^{n}\nabla_{f(x_i^t;\Theta)}\mathscr{H}\Big(f\big(x_i^t;\Theta\big)\Big)\times\nabla_{\Theta} f\big(x_i^t;\Theta\big) + 2C_7\Theta,\ \mathscr{H}\Big(f\big(x_i^t;\Theta\big)\Big) = \pi_i^{\top}\Big(f_i - f\big(x_i^t;\Theta\big)\Big) \tag{33}$$
$$+\, \frac{\rho_1}{2}\Big|f_i - f\big(x_i^t;\Theta\big)\Big|_2^2,\ \nabla_{f(x_i^t;\Theta)}\mathscr{H}\Big(f\big(x_i^t;\Theta\big)\Big) = -\pi_i - \rho_1\Big(f_i - f\big(x_i^t;\Theta\big)\Big),$$

and $\nabla_{\Theta} f(x_i^t;\Theta)$ is the gradient function of the convolutional network function with regard to the network parameters, usually the convolutional filters.

Optimizing Φ. We use the gradient descent algorithm to optimize Φ to minimize $\mathscr{L}(\Phi)$:

$$\begin{aligned}&\Phi \leftarrow \Phi - \eta_{\Phi}\nabla_{\Phi}\mathscr{L}(\Phi),\\ &\quad\text{where } \nabla_{\Phi}\mathscr{L}_3(\Phi) = \nabla_{g\left(x_j^s;\Phi\right)}\boldsymbol{\mathscr{G}}\left(g\left(x_j^s;\Phi\right)\right) \times \nabla_{\Phi} g\left(x_j^s;\Phi\right) + 2C_7\Phi\boldsymbol{\mathscr{G}}\left(g\left(x_j^s;\Phi\right)\right) = \varpi_j^{\top}\left(g_j - g\left(x_j^s;\Phi\right)\right) + \frac{\rho_2}{2}\left|g_j - g\left(x_j^s;\Phi\right)\right|_2^2,\\ &\nabla_{g\left(x_j^s;\Phi\right)}\boldsymbol{\mathscr{G}}\left(g\left(x_j^s;\Phi\right)\right) = -\sum_{j=1}^{m}\varpi_j - \rho_2\sum_{j=1}^{m}\left(g_j - g\left(x_j^s;\Phi\right)\right).\end{aligned} \tag{34}$$

Optimizing W. To optimize the classifier parameter in W, we also update the columns one by one. To optimize the k-th column, we use the following subgradient descent step:

$$\begin{aligned}&w_k \leftarrow w_k - \eta\nabla_{w_k}\mathscr{L}(w_k),\\ &\text{where } \nabla_{w_k}\mathscr{L}(w_k) = -2C_1\sum_{j:\, x_j^s\in L^s}\left(y_j^s(k) - w_k^{\top}g_j\right)g_j - C_2\sum_{j:\, x_j^s\in U^s}\left(\left(w_k^{\top}g_j + 1\right)\log\left(w_k^{\top}g_j\right)\right)g_j + 2C_7 w_k.\end{aligned} \tag{35}$$

Optimizing V. To optimize $\mathbf{V}$, we also update the columns one by one. To update the k-th column v_k, we have the following subgradient descent step:

$$\begin{aligned}&v_k \leftarrow v_k - \eta\nabla_{v_k}\mathscr{O}(v_k),\\ &\text{where } \nabla_{v_k}\mathscr{O}(v_k) = -2C_1\sum_{j:\, x_i^t\in L^t}\left(y_i^t(k) - v_k^{\top}z_i\right)z_i - C_2\sum_{j:\, x_i^t\in U^s}\left(\left(v_k^{\top}z_i + 1\right)\log\left(v_k^{\top}z_i\right)\right)z_i + 2C_7 v_k.\end{aligned} \tag{36}$$

Optimizing A. The optimization of the proxy construction coefficient matrix is a constrained minimization problem as follows:

$$\begin{aligned}&\min_{A}\left\{L(A) = \sum_{i=1}^{n}\tau_i^{\top}\left(z_i - \sum_{j=1}^{m}\alpha_{ij}g_j\right) + \frac{\rho_3}{2}\sum_{i=1}^{n}\left|z_i - \sum_{j=1}^{m}\alpha_{ij}g_j\right|_2^2 + \sum_{i=1}^{n}\varsigma_i^{\top}\left(\sum_{j=1}^{m}\alpha_{ij} - 1\right) + \frac{\rho_4}{2}\sum_{i=1}^{n}\left|\sum_{j=1}^{m}\alpha_{ij} - 1\right|_2^2\right\},\\ &s.t.\ \forall i = 1,\ldots,n,\ j = 1,\ldots,m\colon\ 1 \geq \alpha_{ij} \geq 0.\end{aligned} \tag{37}$$

Again, we optimize the columns of A one by one. To optimize the i-th column α_i of A, we have the following subproblem:

$$\begin{aligned}&\min_{\alpha_i}\left\{\tau_i^{\top}\left(z_i - G\alpha_i\right) + \frac{\rho_3}{2}\left|z_i - G\alpha_i\right|_2^2 + \varsigma_i^{\top}\left(1^{\top}\alpha_i - 1\right) + \frac{\rho_4}{2}\left|1^{\top}\alpha_i - 1\right|_2^2\right\},\\ &\text{s.t. } 1 \geq \alpha_i \geq 0,\end{aligned} \tag{38}$$

where 1 is an all-one vector and 1 is an all-zero vector. This is a typical quadratic programming (QP) problem, and we use the active set algorithm to solve it.

Optimizing $\Pi, \Psi, \Xi,$ andς. To optimize the dual variables, we have the following maximization problem:

$$\begin{aligned}
&\max_{\Pi,\Psi,\Xi,\varsigma}\Bigg\{\mathscr{L}(\Pi,\Psi,\Xi,\varsigma)=\sum_{i=1}^{n}\pi_i^{\top}\left(f_i-f\left(x_i^t;\Theta\right)\right)\\
&\quad+\sum_{j=1}^{m}\varpi_j^{\top}\left(g_j-g\left(x_j^s;\Phi\right)\right)+\sum_{i=1}^{n}\tau_i^{\top}\left(z_i-\sum_{j=1}^{m}\alpha_{ij}g_j\right)\\
&\quad+\sum_{i=1}^{n}\varsigma_i^{\top}\left(\sum_{j=1}^{m}\alpha_{ij}-1\right)=Tr\left(\Pi^{\top}N\right)+Tr\left(\Psi^{\top}J\right)+Tr\left(\Xi^{\top}D\right)+\varsigma^{\top}e\Bigg\},\\
&\text{where } N=\left[\left(f_1-f\left(x_1^t;\Theta\right)\right),\ldots,\left(f_n-f\left(x_n^t;\Theta\right)\right)\right],\\
&J=\left[\left(g_1-g\left(x_1^s;\Phi\right)\right),\ldots,\left(g_m-g\left(x_m^s;\Phi\right)\right)\right],\\
&D=\left[\left(z_1-\sum_{j=1}^{m}\alpha_{1j}g_j\right),\ldots,\left(z_n-\sum_{j=1}^{m}\alpha_{nj}g_j\right)\right],\text{ and}\\
&e=\left[\left(\sum_{j=1}^{m}\alpha_{1j}-1\right),\ldots,\left(\sum_{j=1}^{m}\alpha_{nj}-1\right)\right],
\end{aligned}\tag{39}$$

are the residual matrices and vector and $Tr(X)$ is the trace of matrix X. To solve this problem, we use the gradient-ascent algorithm, and the updating rules are

$$\begin{aligned}
\Pi\longleftarrow\Pi+\varepsilon\nabla_{\Pi}\mathscr{L}(\Pi,\Psi,\Xi,\varsigma),\ \nabla_{\Pi}\mathscr{L}(\Pi,\Psi,\Xi,\varsigma)&=N,\\
\Psi\longleftarrow\Psi+\varepsilon\nabla_{\Psi}\mathscr{L}(\Pi,\Psi,\Xi,\varsigma),\ \nabla_{\Psi}\mathscr{L}(\Pi,\Psi,\Xi,\varsigma)&=J,\\
\Xi\longleftarrow\Xi+\varepsilon\nabla_{\Xi}\mathscr{L}(\Pi,\Psi,\Xi,\varsigma),\ \nabla_{\Xi}\mathscr{L}(\Pi,\Psi,\Xi,\varsigma)&=D,\\
\varsigma\longleftarrow\varsigma+\varepsilon\nabla_{\varsigma}\mathscr{L}(\Pi,\Psi,\Xi,\varsigma),\text{ and }\nabla_{\varsigma}\mathscr{L}(\Pi,\Psi,\Xi,\varsigma)&=e,
\end{aligned}\tag{40}$$

where ϵ is the ascent step size.

Besides solving the minimization problem, we also need to solve the parameter of the SMI term of (13) ϕ^*. To learn the parameter of $\widehat{\mathbf{SMI}}, \phi^*$, according to (10), we have the following minimization problem:

$$\begin{aligned}
&\phi^*=\min_{\phi}\left\{S(\phi)=-\frac{2}{n}\sum_{i=1}^{n}r\left(f_i,z_i;\phi\right)+\frac{1}{n^2}\sum_{i,i'=1}^{n}r\left(f_i,z_i;\phi\right)r\left(f_{i'},z_{i'};\phi\right)=-\frac{2}{n}\sum_{i=1}^{n}h\left(\phi^{\top}r_i\right)+\frac{1}{n^2}\sum_{i,i'=1}^{n}h\left(\phi^{\top}r_i\right)h\left(\phi^{\top}r_{i'}\right)\right\},\\
&\text{where } r_i=\left[f_i;z_i;f_i-z_i\right].
\end{aligned}\tag{41}$$

The gradient function of $\mathcal{S}(\phi)$ regarding ϕ is

$$\nabla\mathcal{S}_{\phi}(\phi)=-\frac{2}{n}\sum_{i=1}^{n}\nabla_{\phi}h\left(\phi^{\top}r_i\right)r_i+\frac{1}{n^2}\sum_{i,i_l=1}^{n}\left(\nabla_{\phi}h\left(\phi^{\top}r_i\right)h\left(\phi^{\top}r_{i_l}\right)r_i+h\left(\phi^{\top}r_i\right)\nabla_{\phi}h\left(\phi^{\top}r_{i_l}\right)r_{i_l}\right).\tag{42}$$

We use the gradient descent algorithm to update ϕ to minimize $\mathcal{S}(\phi)$:

$$\phi\longleftarrow\phi-\eta\nabla\mathcal{S}_{\phi}(\phi).\tag{43}$$

3. Experiments

In this section, we experimentally evaluate the proposed algorithm over benchmark data sets.

3.1. Benchmark Data Sets. In our experiment, we use the following benchmark data sets:

(i) Amazon review data set. This data set has four domains of four products, including books, DVD, and music. For each domain, there are 2,000 positive reviews and 2,000 negative reviews. To fit the content of the review to the CNN network, we firstly tokenize the review to a sequence of words and secondly represent each word to by word embedding and then use the 1D-CNN as the CNN model.

(ii) UCSD anomaly detection data set. This data set has 70 sequences of video, which contains 13,900 frames. This data set is used as a source domain. Meanwhile, we also collected 122 sequences of videos from subway stations, malls, and communities in China as a target domain. The target domain has 24,400 frames. Both target and source domain data sets have 6 classes of anomalies.

(iii) TRECVID video concept detection data set. This data set has two domains of data. One domain is the data of TRECVID 2005, which has 61,901 frames from video, while the other one domain is the data of TRECVID 2007, which has 21,532 frames. The classification problem is to categorize one frame to one of the 36 semantic concepts.

(iv) Protein subcellular localization data set. The last data set is a protein data set. It is composed of three domains. The first domain is the MultiLoc data set which has 5,859 proteins, the second one is the BaCelLoc date set with 4,286 proteins, and the last data set Euk-mPLoc has 5,618 proteins. The problem of this data is to predict the subcellular locations of each protein from the amino acid sequence. To this end, we first map each amino acid by embedding technology and then use the 1D-CNN model to extract the features.

The statistics of the data sets are summarized in Table 1.

Table 1: Summary of data sets.

Data set	# domain	# sample	Data type
Amazon	3	12,000	Text
UCSD	2	38,300	Video
TRECVID	2	83,433	Video
Protein	3	15,763	Protein

3.2. Experimental Setting. To perform the experiment, we design a leave-one-out protocol to use each domain as a target domain, while the other domains are combined as one single source domain. With each target domain-source domain configuration, we use the 10-fold cross-validation protocol to perform the training test process. The target domain data set is split into ten folds, one fold is used as the test set, and the remaining nine folds are combined with the source domain as the training set. The model is learned over the training set and tested over the test set. To measure the performance of the method, we calculate the average classification accuracy over different target domains.

3.3. Experimental Results. In our experiments, we firstly study the properties of the proposed method experimentally and then compare it to the other state-of-the-art methods.

3.4. Convergence Analysis. Since the algorithm is an iterative algorithm, we are interested in the convergence of the algorithm. We plot the curves of the objective values against the iteration number in Figure 1. From this figure, we can see that, for all four benchmark data sets, the algorithm converges well at 50 iterations. For the protein data set, the object value keeps decreasing but cannot reach a lower value after 50 iterations. The possible reason is that the algorithm finds a local minimum object instead of a global minimum. The overall conclusion for the convergence of the algorithm is for all the tested data sets, and the algorithm can converge at certain number of iterations. This is due to the effective alternate optimization method.

3.5. Tradeoff Parameter Analysis. We have seven tradeoff parameters in the objective, which control the weights of different objective terms. We also study how the performance of the algorithm changes with the change in these tradeoff parameters. The classification accuracies with different values of the tradeoff parameter values are shown in Figure 2.

From this figure, we have the following observations:

(i) The increasing C_1 and C_3 improve the performance significantly since it is the supervision term of the labeled data. This means the supervision of the labeled plays an important role in the learning of predictive models in the transfer learning problem.

(ii) The increasing C_2 and C_4 also boost the performance; however, it is not as significant as the class error terms' tradeoff parameters. The reduction of entropy of the unlabeled samples' classification responses from both domains also improves the performance.

(iii) C_5 and C_6 are the tradeoff parameters of two domains' manifold regularization terms. Their increasing values also improve the accuracy of the target domain prediction. This means the neighborhood smoothness is also important for the transfer learning problem.

(iv) The performance is stable to the change in C_7, the tradeoff value of the squared ℓ_2 norm term, which

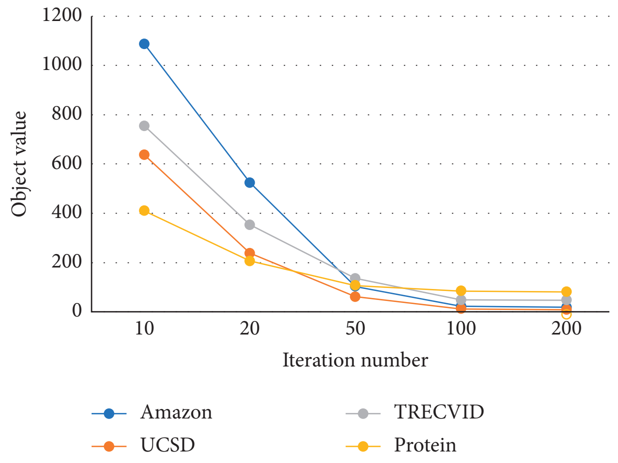


Figure 1: Convergence curves.

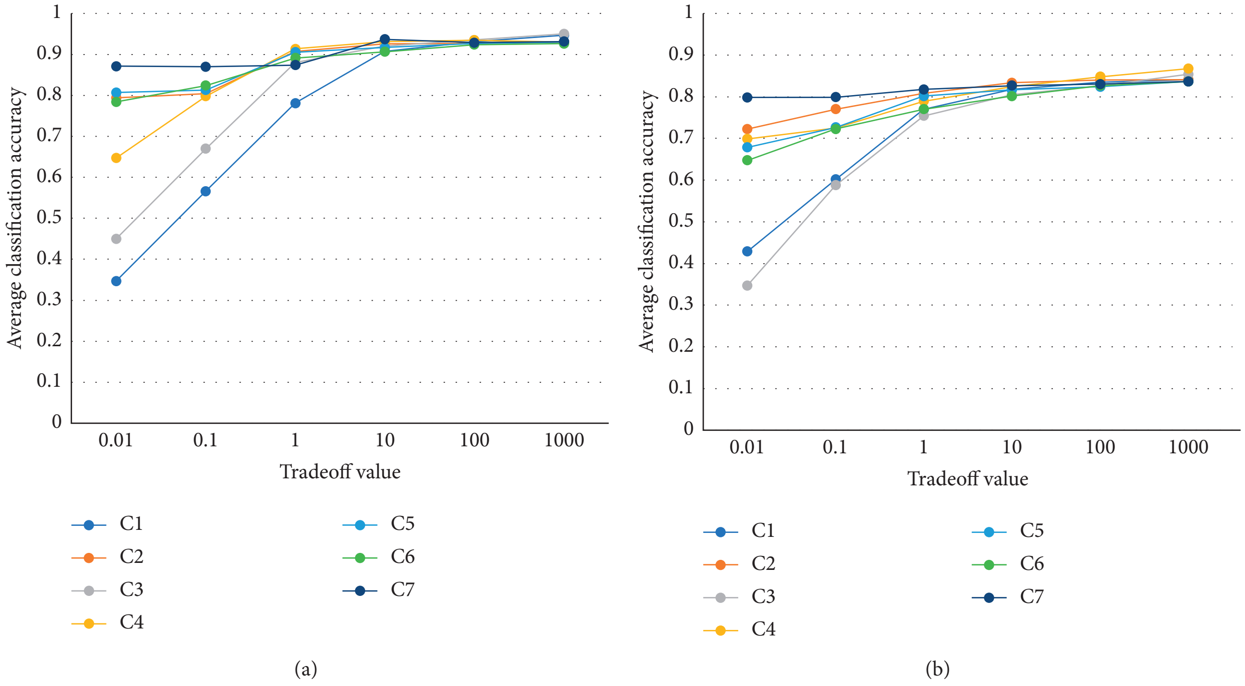


Figure 2: Tradeoff parameter analysis: (a) UCSD and (b) protein.

means this term did not play a critical role in the learning problem.

3.6. Comparison to Other Methods. We compare our proposed algorithm against several state-of-the-art CNN-based transfer learning methods, including the methods developed by Haque et al. [11, 12], Zhang et al. [14], Long et al. [21], and Pan and Yang [7]. The comparison results are shown in Figure 3.

From this figure, we can see the proposed method always outperforms the other methods in all cases. The second best method is Long et al. method. The results show the advantage of our method, especially the SMI-guided domain-transfer CNN learning and the proxy mapping mechanism. Among the four data sets, the most difficult one is the TRECVID, where all the classification accuracies of different methods are below 0.75. However, in this data set, our methods give the most significant improvements over the other methods. It is the only method which achieves an accuracy higher than 0.72.

4. Conclusions

In this paper, we proposed a novel CNN-based transfer learning method. We construct a proxy for each target domain sample in the source domain space and use the SMI

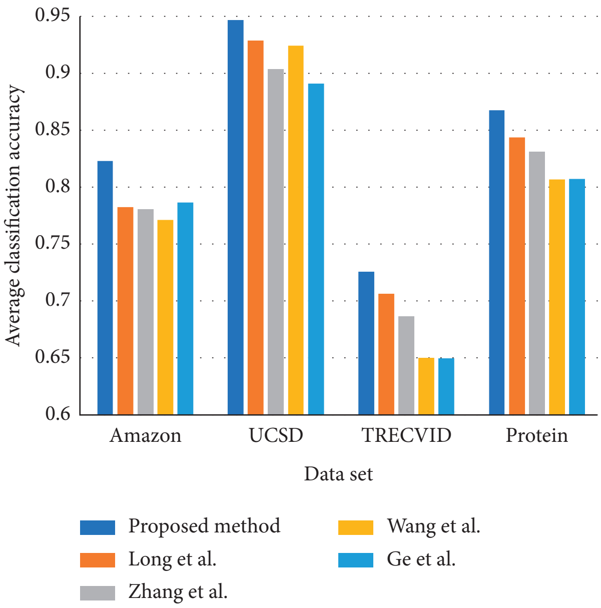


Figure 3: Comparison to state-of-the-arts.

as a measure to match the constructed proxy and the convolutional representation of a target domain sample. The learning of the proxy construction parameters and the CNN parameters are learned simultaneously by a unified learning framework. The estimation and parameter learning of the SMI is also driven by the proxy and CNN outputs. Thus, this framework can optimize the CNN, proxy, and SMI parameters jointly and automatically. Experimental results show the advantages of the proposed framework and algorithm.

Data Availability

All the data sets used in this paper to produce the experimental results are publicly accessed online.

Conflicts of Interest

The authors declare that there are no conflicts of interest regarding the work reported in this paper.

Acknowledgments

This work was supported by the Beijing Natural Science Foundation (no. 9194027), National Key R&D Program of China (Grant no. 2018YFC0704800), 2018-2019 Excellent Talent Program, Xicheng District, Beijing, and National Natural Science Foundation of China (no. 71904095).

References

[1] P. Russom, “Big data analytics,” *TDWI Best Practices Report*, vol. 19, no. 4, pp. 1–34, 2011.

[2] M. Chen, S. Mao, and Y. Liu, “Big data: A survey,” *Mobile Networks and Applications*, vol. 19, no. 2, pp. 171–209, 2014.

[3] T. M. Mitchell, *Machine Learning*, Springer, Berlin, Germany, 1997.

[4] E. Alpaydin, *Introduction to Machine Learning*, MIT press, Cambridge, MA, USA, 2020.

[5] X. Zhu and A. B. Goldberg, “Introduction to semi-supervised learning,” *Synthesis Lectures on Artificial Intelligence and Machine Learning*, vol. 3, no. 1, pp. 1–130, 2009.

[6] X. J. Zhu, *Semi-supervised Learning Literature Survey. Tech. Rep*, University of Wisconsin-Madison Department of Computer Sciences, Madison, Wisconsin, 2005.

[7] S. J. Pan and Q. Yang, “A survey on transfer learning,” *IEEE Transactions on Knowledge and Data Engineering*, vol. 22, no. 10, pp. 1345–1359, 2009.

[8] M. E. Taylor and P. Stone, “Transfer learning for reinforcement learning domains: A survey,” *Journal of Machine Learning Research*, vol. 10, pp. 1633–1685, 2009.

[9] V. Mahadevan, W. Li, V. Bhalodia, and N. Vasconcelos, “Anomaly detection in crowded scenes,” in *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, pp. 1975–1981, San Francisco, CA, USA, June 2010.

[10] S. Bansod and A. Nandedkar, “Transfer learning for video anomaly detection,” *Journal of Intelligent & Fuzzy Systems*, vol. 36, no. 3, pp. 1967–1975, 2019.

[11] T. U. Haque, N. N. Saber, and F. M. Shah, “Sentiment analysis on large scale amazon product reviews,” in *Proceedings of the IEEE International Conference on Innovative Research and Develop- Ment (ICIRD)*, pp. 1–6, Bangkok, Thailand., 2018.

[12] A. Bhatt, A. Patel, H. Chheda, and K. Gawande, “Amazon review classification and senti- ment analysis,” *International Journal of Computer Science and Information Technologies*, vol. 6, no. 6, pp. 5107–5110, 2015.

[13] T. N. Sainath, A. Mohamed, B. Kingsbury, and B. Ramabhadran, “Deep convolutional neural networks for lvcsr,” in *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing*, pp. 8614–8618, Vancouver, Canada, May 2013.

[14] G. Zhang, G. Liang, F. Su, F. Qu, and J.-Y. Wang, “Cross-domain attribute representa- tion based on convolutional neural network,” in *Proceedings of the International Conference on Intelligent Computing*, pp. 134–142, Wuhan, China, October 2018.

[15] Y. Geng, R. Z. Liang, W. Li et al., “Learning convolutional neural network to maximize pos@ top performance measure,” in *Proceedings of the ESANN 2017 - Proceedings*, pp. 589–594, Bruges, Belgium, April 2017.

[16] G. Zhang, G. Liang, W. Li et al., “Learning con- volutional ranking-score function by query preference regularization,” in *Proceedings of the International Conference on Intelligent Data Engineering and Automated Learning*, pp. 1–8, Guilin, China, October 2017.

[17] Y. Geng, G. Zhang, W. Li et al., “A novel image tag completion method based on convolutional neural transformation,” in *Proceedings of the International Conference on Artificial Neural Networks*, pp. 539–546, Sardinia, Italy, September 2017.

[18] M. Yamada, L. Sigal, M. Raptis, M. Toyoda, Y. Chang, and M. Sugiyama, “Cross-domain matching with squared-loss mutual information,” *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 37, no. 9, pp. 1764–1776, 2015.

[19] K. Wang, J. Liu, and J.-Y. Wang, “Learning domain-independent deep representations by mutual information minimization,” *Computational Intelligence and Neuroscience*, vol. 2019, pp. 1–14, Article ID 9414539, 2019.

[20] M. Xiao and Y. Guo, “Feature space independent semi-supervised domain adaptation via kernel matching,” *IEEE*

Transactions on Pattern Analysis and Machine Intelligence, vol. 37, no. 1, pp. 54–66, 2014.

[21] M. Long, Y. Cao, Z. Cao, J. Wang, and M. I. Jordan, "Transferable representation learning with deep adaptation networks," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 41, no. 12, pp. 3071–3085, 2018.

[22] L. Duan, I. W. Tsang, and D. Xu, "Domain transfer multiple kernel learning," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 34, no. 3, pp. 465–479, 2012.

[23] M. Long, H. Zhu, J. Wang, and M. I. Jordan, "Deep transfer learning with joint adaptation networks," in *Proceedings of the 34th International Conference on Machine Learning*, pp. 2208–2217, Sydney, Australia, August 2017.

[24] W. Ge and Y. Yu, "Borrowing treasures from the wealthy: Deep transfer learning through selective joint fine-tuning," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 1086–1095, Honolulu, HI, USA., July 2017.

[25] T. Suzuki and M. Sugiyama, "Sufficient dimension reduction via squared-loss mutual in- formation estimation," in *Proceedings of the Thirteenth International Conference on Artificial Intelligence and Statistics*, pp. 804–811, Sardinia, Italy, May 2010.

[26] G. Niu, W. Jitkrittum, B. Dai, H. Hachiya, and M. Sugiyama, "Squared-loss mutual in- formation regularization: A novel information-theoretic approach to semi-supervised learning," in *Proceedings of the International Conference on Machine Learning*, pp. 10–18, Atlanta, GA, USA, June 2013.

[27] S. Boyd, N. Parikh, E. Chu, B. Peleato, and J. Eckstein, "Distributed optimization and statistical learning via the alternating direction method of multipliers," *Foundations and Trends® in Machine Learning*, vol. 3, no. 1, pp. 1–122, 2011.

Hindawi
Shock and Vibration
Volume 2020, Article ID 8846429, 9 pages
https://doi.org/10.1155/2020/8846429

Research Article

Consensus Mechanism of IoT Based on Blockchain Technology

Yue Wu,[1,2] **Liangtu Song**,[1,2] **Lei Liu**,[1,2] **Jincheng Li**,[1,2] **Xuefei Li**,[1,2] **and Linli Zhou**[1,2]

[1]*Institute of Intelligent Machines, Hefei Institutes of Physical Science, Chinese Academy of Sciences, Hefei 230031, Anhui, China*
[2]*University of Science and Technology of China, Hefei 230026, Anhui, China*

Correspondence should be addressed to Yue Wu; wyw533k@mail.ustc.edu.cn

Received 10 August 2020; Revised 27 August 2020; Accepted 21 September 2020; Published 7 October 2020

Academic Editor: Chaoqun Duan

Applying blockchain technology to the Internet of Things (IoT) remains a huge challenge. To meet the actual needs of IoT, a lightweight and high-throughput consensus mechanism, combined with blockchain technology, is proposed in this study. Blockchain nodes use the Diffie–Hellman algorithm for key negotiation. Sensors and blockchain nodes can use the shared key to generate HMAC (Hash-based Message Authentication Code) signatures for sensor-aware transactions and use the Verifiable Random Function to implement block nodes. Offline fast election, which is the node that wins the election, becomes the block node. Machine learning methods are also introduced to identify or remove outliers in the sensor data before such data are uploaded to the chain. Experimental results show that the system throughput synchronously increases as the test load increases. Moreover, when the test load is 800 tps, the system throughput reaches the maximum, close to 600 tps. When the test load exceeds 800 tps, the actual system throughput starts to drop, and approximately 90% of transactions have a delay time within 5000 ms. This method can be used in a lightweight IoT system.

1. Introduction

With the continuous development and progress of technologies such as sensor technology, computer control technology, embedded technology, and wireless network data communication, the Internet of Things (IoT) has shown great development worldwide [1]. IoT is considered the third world information industry wave after the computer and the Internet and is developing rapidly. By 2020, more than 50 billion smart devices will be connected to the network [2]. Distributed processors and communication hardware send/receive data from the environment, thus generating huge data. However, at present, most IoT architectures require a central hub or server, which allows data storage and transmission between several devices in the network space. These smart devices use sensors, embedded processors, and communication hardware to send/receive data from the environment, thereby generating huge amounts of data. Therefore, security and privacy have become the greatest challenges of IoT [3].

According to the characteristics of the IoT platform, blockchain technology can solve the security and management problems that a large amount of smart device data can appear in a centralized system framework [4]. Blockchain is a new application model similar to an integrated distributed database, consensus mechanism, peer-to-peer (P2P) transmission, and asymmetric encryption algorithm [5]. Blockchain has the characteristics of decentralization, openness, autonomy, anonymity, and information tampering. The use of blockchain technology for IoT has attracted increasing attention in this direction. Literature [6] constructed a distributed data management system on the basis of blockchain and trusted execution environment. The system stores the encrypted hash value in the blockchain and stores the original encrypted data in a trusted execution environment. The system also provides corresponding data access control strategies through smart contracts to ensure the security, privacy, and integrity of the IoT dataset. Literature [7, 8] proposed a lightweight architecture on the basis of blockchain for IoT. Such an architecture maintains security and privacy while reducing blockchain overhead. IoT devices combine private ledger with centralized management similar to blockchain to optimize energy consumption. High-resource devices create an overlay network to implement a publicly accessible distributed blockchain, ensure end-to-end privacy and security, and use distributed authentication to reduce block verification processing time, which is ultimately implemented in smart home applications. Literature [9] proposed a new role and authority arbitration structure in IoT that is a fully distributed access control system for IoT based on blockchain technology and is evaluated in a real IoT scenario. The results provided in this article indicate that, in specific scalable IoT scenarios, blockchain technology can be used as an access control technology to enhance security. Literature [10, 11] evaluated the role of blockchain in strengthening network security and protecting privacy. Using practical applications and practical examples, the distributed sensitivity of the blockchain likely causes attack sensitivity problems. Ways to solve problems related to blockchain-based identity and access control system are also discussed. In addition, certain key challenges related to IoT security are put forward. However, the current mainstream blockchain implementation is mainly aimed at digital currency applications, such as Bitcoin and Ethereum [12]. These blockchains are directly used in IoT and have the following problems:

(i) The computing power and storage resource of IoT devices are generally low, making the performance of complex cryptographic calculations difficult [13]. Mainstream blockchain platforms usually use computationally intensive asymmetric key technology as user identification and transaction verification mechanism; one example is the secp256k1 elliptic curve asymmetric key pair used in Bitcoin and Ethereum [14], which requires more computing power than what most IoT devices can provide. Taking the common IoT hardware platform Arduino Uno (CPU frequency: 16 MHZ) as an example, performing an elliptic curve private key signature and an elliptic curve public key verification bell takes more than five and eight seconds, respectively [15]. Therefore, an appropriate cryptographic security mechanism must be selected to allow IoT devices to access the blockchain and ensure a certain degree of security [16].

(ii) IoT requires high data throughput. At present, mainstream blockchain technology still cannot easily support mainstream public chains that use digital currency applications as the main target scenarios, such as Bitcoin or Ethereum. The primary purpose of consensus is to avoid double-spending [17]; that is, a coin is consumed multiple times so that the blockchain can meet the needs of certain users at low throughput. For example, the throughput of Bitcoin is 7 tps [18], whereas that of Ethereum is 15 tps. Considering that the sampling period of sensors in IoT is usually in the order of seconds or even milliseconds and a large number [19], which means that the data on IoT devices have high-throughput requirements, a consensus mechanism suitable for IoT with high throughput should be designed [20].

(iii) IoT can improve the credibility of data on the chain, but it lacks effective safeguards. Although the blockchain can guarantee that such data cannot be tampered with, tampering with the data source is powerless. It affects the trustworthiness of data on the blockchain. IoT can be directly uploaded to the chain through sensor data, thereby maximizing the credibility of the data on the chain, which puts high requirements on the confirmation of transactions in the blockchain consensus algorithm to identify sensor data.

The research on this topic is applicable to the blockchain consensus mechanism of IoT. The main contents include the research on cryptographic algorithms and mechanisms suitable for lightweight IoT devices to access the blockchain. We attempt to avoid excessive communication overhead, and we introduce machine learning methods to identify or remove outliers in sensor data before data are uploaded to the chain.

2. System Model

As shown in Figure 1, the system model is divided into four layers, from bottom to top.

(i) Perception layer is mainly composed of lightweight IoT devices, such as sensors or drivers, and adjacent IoT devices are connected to IoT gateways.

(ii) Gateway layer mainly comprises IoT gateways, and each gateway is responsible for the access of the part of sensor/driver devices. Gateways are responsible for identifying the abnormal sensor data and submitting the processed transactions to the blockchain nodes in the network layer.

(iii) Network layer is mainly composed of blockchain nodes which form P2P networks with each other. Blockchain nodes are responsible for caching transactions (from the gateway or upper application), selecting master nodes, identifying blocks, and resolving the chain fork conflict.

(iv) Application layer mainly comprises various decentralized applications (DApps) based on blockchain. DApp accesses the data on the chain through blockchain nodes or submits transactions to the chain.

Compared with the traditional Internet of Things system model, the blockchain technology is not introduced in the network layer, but the centralized server is adopted as the core of the system. The upward application layer and the downward gateway layer directly interact with the centralized server for data. When data is collected continuously by multiple different devices on the network, the amount of data is very large. The traditional Internet of Things system is managed by a central service manager, so the system will be responsible for thousands of devices. In this way, for such a large amount of data, the system must have a storage device

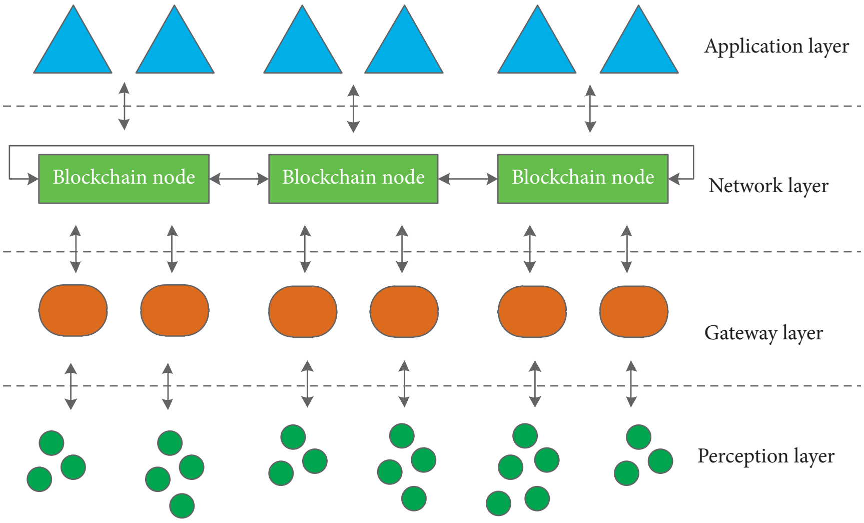


Figure 1: System model of IoT based on blockchain.

that can handle all the data, a powerful data processing system, and an efficient mobile phone and transmission system. After the introduction of the blockchain in the network layer, there is no central server and each node has its own copy of data, which can be processed independently. After the system obtains the transmission parameters, it compares the other nodes in the network with the transmission parameters. If they are in agreement, then, they will reach a consensus and can deal with trading blockchain technology to solve this problem to some extent. In addition, once the central server is hacked, all the data stored in the central server will be destroyed. Blockchain networks maximize their strengths and avoid their weaknesses. All data is encrypted and stored in the network, and only participating nodes can access it. And the nodes of network verification must be consistent to ensure that the data will not be maliciously modified.

The overall flow of the algorithm is shown in Figure 2.

After the system starts, the sensor first requests access to the IoT gateway, and the sensor periodically uploads data to the gateway. At the same time, after the sensor is started, it uses the Diffie–Hellman algorithm for key negotiation with the blockchain node and uses the shared key to sense transactions for the sensor to generate HMAC signature. After the HMAC signature verification of the sensor data and the abnormal data is removed, the IoT gateway starts to submit ordinary transactions. To avoid the computational overhead in PoW consensus and the communication overhead in BFT consensus, this study adopts Verifiable Random Function (VRF) to realize the offline fast election of block nodes. The node that wins the election becomes the block producer. Then, the dominant node packs the transaction and produces the block. When there are multiple legal block-producing nodes, multiple legal blocks will appear at the same height; that is, bifurcation occurs. The consensus algorithm in this article uses the longest chain rule to resolve fork conflicts; that is, the chain with the largest length is regarded as the optimal chain.

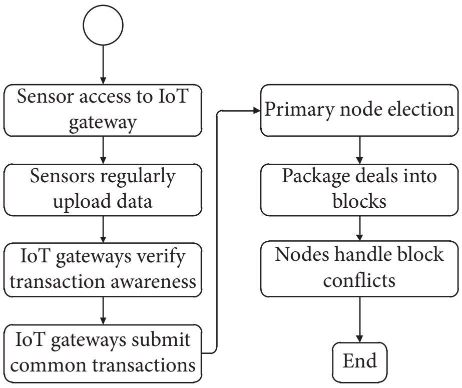

Sensor access to IoT gateway
Sensors regularly upload data
IoT gateways verify transaction awareness
IoT gateways submit common transactions
Primary node election
Package deals into blocks
Nodes handle block conflicts
End

Figure 2: Algorithm process.

As shown in Figure 2, the system has two types of transactions: sensor-aware transactions that are submitted to gateways, and normal transactions that occur on blockchain nodes or upper-level applications. Considering the performance difference between IoT terminal devices and node or application hosts, we design different verification mechanisms for both transactions.

(i) Transaction aware: from sensors to gateways, the symmetric encryption algorithm is used for transaction verification.

(ii) Ordinary transactions: the asymmetric encryption algorithm is used for transaction verification.

Gateways use machine learning models to detect the abnormality in the sensor perception transaction, attach the detection result and reencapsulate it, and broadcast it to the blockchain network. After receiving and verifying the broadcast transaction, blockchain nodes temporarily store it in the pending transaction pool and wait for block confirmation. We divide the continuous time into consensus rounds at equal intervals and blockchain nodes package pending transactions, produce blocks, and broadcast to the blockchain network in rounds. To improve throughput, in a round, certain nodes are randomly selected as dominant nodes, and only the selected dominant nodes can produce blocks. Given that multiple leading nodes may be selected in a round, nodes may receive blocks of the same height broadcast from multiple leading nodes. We use the longest chain rule to solve the blockchain fork conflict.

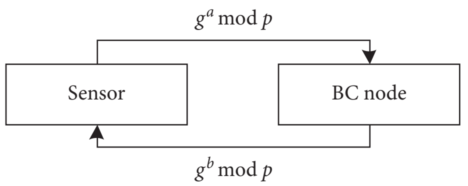


Figure 3: Sensor and node key agreement protocol.

Table 1: Sensor transaction structure.

#	Field	Type	Explanation
1	Sender	String	Sensor ID
2	Time	Int	Send time, Linux timestamp
3	Value	Float	Sensor sampling data
4	MAC	String	MAC code

3. Consensus Mechanism

3.1. Key Agreement Protocol. As shown in Figure 3, after sensors are started, the Diffie–Hellman algorithm is used for key negotiation with the blockchain nodes.

(1) Sensors and blockchain nodes share prime number p and base g.

(2) Sensors first select private key a and, then, send public key A's a to blockchain nodes:

$$A = g^a \text{mod } p. \tag{1}$$

(3) Nodes first select a private key b and, then, send b's public key B to sensor nodes:

$$B = g^b \text{mod } p. \tag{2}$$

(4) Sensors calculate the shared key:

$$K_{\text{sensor}} = B^a \text{mod } p. \tag{3}$$

(5) Nodes calculate the shared key:

$$K_{bc-\text{node}} = A^b \text{mod } p. \tag{4}$$

(6) Considering that $K_{\text{sensor}} = K_{bc-\text{node}}$, the shared key can be used by sensors and blockchain nodes to generate HMAC signatures for sensor-aware transactions.

3.2. Sensor Transaction Verification. After nodes receive sensor transactions, the verification process is divided into two steps:

(i) Verify the HMAC signatures of sensor data

(ii) Identify and label abnormal sensor data

3.2.1. Sensor Data Signature Verification. As shown in Table 1, the purpose of MAC, the message verification code, is to verify the data source and its integrity. In this article, we use an MAC based on a cryptographic hash function, or HMAC. HMAC requires a shared key between a sender and a receiver. For the specified data and key K, the HMAC calculation formula is as follows:

$$\text{HMAC}(K, \text{data}) = H\big((K_0 \oplus \text{opad}) \| H\big((K_0 \oplus \text{ipad}) \| \text{data}\big)\big). \tag{5}$$

(i) H: selected hash function,

(ii) K_0: key K preprocessed result,

(iii) opad: complete data outside, bytes of hash block length $0 \times 5c$,

(iv) ipad: fill in the data, the length of the hash block byte 0×36,

(v) : splicing operation symbol,

(vi) ⊕: XOR operation symbol.

Sensor perception transaction adopts the following structural definitions:

Before sensors calculate the HMAC codes of transaction structures, other parts, except the HMAC field, must first be serialized into data, and, then, the HMAC calculation can be performed. The algorithm for the serialized data of sensor transaction tx is as follows:

$$\text{data} = \text{bytes}(tx.\text{sender}) \| \text{bytes}(tx.\text{time}) \| \text{bytes}(tx.\text{value}). \tag{6}$$

(i) bytes: convert parameters to bytecode stream,

(ii) : splicing two streams.

3.2.2. Sensor Abnormal Data Detection

(1) Sensor Data Fusion. First is the sensor data fusion. We assume that the data from sensor i in time slot t are $d_{ti} \in R$, and the number of sensors connected to the same blockchain node is n. The sensor data $D_t = [d_1, \ldots d_n]$ accessed by this blockchain node in time slot t thus constitute the state of the external environment at time t, taking the data of k continuous time slots to form $D = [D_1, D_2, \ldots D_k]^T$, which is a $k \times n$ in the matrix. Each row represents the state of the external environment at a time, and each column represents the time series of a sensor. For example, the following figure shows matrix D which is formed by the data of six sensors in three time slots, which is shown in Figure 4.

According to the different characteristics of sensor data, we use two unsupervised machine learning algorithms to detect abnormal data in the sensor fusion data represented by the $k \times n$ matrix.

$$\begin{array}{c|c} & \text{Sensor} \rightarrow \\ \hline \text{Time} \downarrow & \begin{bmatrix} d_{11} & d_{12} & d_{13} & d_{14} & d_{15} & d_{16} \\ d_{21} & d_{22} & d_{23} & d_{24} & d_{25} & d_{26} \\ d_{31} & d_{32} & d_{33} & d_{34} & d_{35} & d_{36} \end{bmatrix} \end{array}$$

FIGURE 4: Time slot matrix.

(2) HBOS Abnormal Data Detection Algorithm. If the data of n sensors connected to a gateway is irrelevant, then, we can use the HBOS algorithm to calculate the anomaly score of each time slot sample using the statistical characteristics of the sensor data. The HBOS algorithm process is as follows:

(i) Use the time series data of a single sensor to calculate its histogram, count the number of occurrences of each value (domain), and calculate its relative frequency.

(ii) Normalize the histogram of each sensor, so that the maximum frequency is 1, to ensure that different sensors have the same weight on the final abnormality score.

(iii) Use the following formula to calculate the anomaly score of the sensor fusion data for each time slot:

$$\text{HBOS}_k = \sum_{i=1}^{n} \log\left(\frac{1}{\text{hist}_i(k)}\right), \tag{7}$$

where $\text{hist}_i(k)$ represents the density estimation of the k-th time slot data of the i-th sensor in the statistical histogram of the sensor and HBOS_k represents the anomaly score of the sensor fusion data of the k-th time slot. The HBOS algorithm assumes that the sensor data are uncorrelated. Therefore, the above HBOS score calculation formula is actually the application of the Naive Bayes probability model (Naive Bayes Model) in the log domain.

(3) Autoencoder Abnormal Data Detection Algorithm. If the sensor data are related, then the unsupervised deep learning model Autoencoder can be used to detect abnormal sensor fusion data.

As shown in Figure 5, autoencoder is a specially constructed neural network structure. By setting the neural network output as input, training data without annotation are obtained. The input data can be simplified/lessened after training Rank representation (Code) through certain regularization conditions. In the sensor data anomaly detection algorithm based on autoencoder, we mainly use the output and input residuals to determine the anomaly score of each sample. The algorithm is briefly described as follows:

(i) Construct the autoencoder model, the numbers of output layers, and output layer neurons to be n.

(ii) Train with sensor data D.

(iii) Calculate the anomaly score of each sample, where d represents the reconstructed data of the i-th sensor:

$$S_k = \sum_{i=1}^{n} \left(d_i - \widehat{d}_l\right)^2. \tag{8}$$

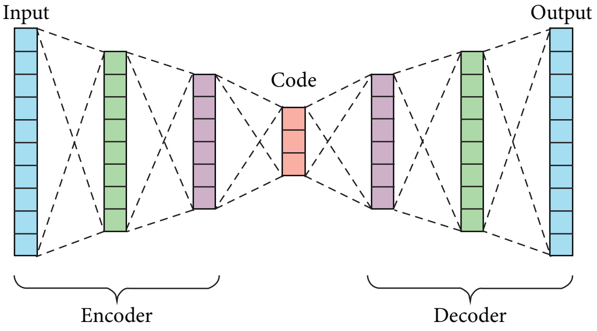


FIGURE 5: Autoencoder neural network.

3.3. Block Node Selection. To avoid the computational overhead in PoW consensus and the communication overhead in BFT consensus, this study adopts Verifiable Random Function (VRF) to realize the offline fast election of block nodes. VRF is a cryptographic hash of the public key version. Only the private key holder can calculate the hash, but any participant who knows the public key can verify the correctness of the hash.

The algorithm flow is summarized as follows:

(1) The private key holder uses the private key SK and the public input data alpha to calculate the hash beta and evidence pi:

$$\text{beta} = \text{vrf\_proof2hash}(\text{pi}). \tag{9}$$

(2) The verifier uses the hash provider's public key PK, evidence pi, and input data alpha to recalculate hash beta_2. If it matches the provider, the hash is correct:

$$\text{beta}_2 = \text{vrf\_verify}(\text{PK}, \text{alpha}, \text{pi}). \tag{10}$$

In this study, the consensus algorithm uses a round consensus mechanism, which divides the time into rounds of fixed length. In each round, VRF is used to determine whether the current node of the current round is selected as the block-generation node, and if so, the block is broadcasted. VRF requires all parties involved to hold a key pair, and the block selection algorithm flow of each round is as follows:

(1) Calculate the shared information alpha of the current round, where t represents the current time and T represents the duration of the round:

$$\text{alpha} = \frac{t}{T}. \tag{11}$$

(2) The node uses its private key and shared information alpha to calculate the hash beta and evidence pi:

$$\begin{gathered} \text{pi} = \text{vrf\_prove}(\text{SK}, \text{alpha}), \\ \text{beta} = \text{vrf\_proof2hash}(\text{pi}). \end{gathered} \tag{12}$$

Assuming that the probability of winning a node in an election is p, and the selection experiment for each node n times is repeated, the probability X_k of finally

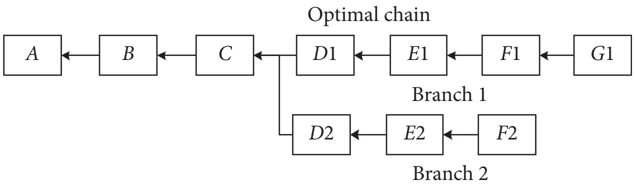


Figure 6: Longest chain rule.

selecting k nodes conforms to the binomial distribution, namely,

$$X_k \sim b(n, p). \tag{13}$$

Calculate the corresponding k when X_k is the following value, where hashlen represents the bit length of hash beta:

$$p_{\text{node}} = \frac{\text{beta}}{2^{\text{hashlen}}}. \tag{14}$$

If $k > 0$, then the node is selected.

(3) The selected block node broadcasts the block, with its public key and evidence.

(4) After receiving the block, the other nodes calculate the alpha value of the current round, use the public key PK of the block-producing node in the block and the evidence pi to calculate the hash, and perform Step 3. After verifying that the node is indeed a block-producing node, change the area. The block is also added to the local chain.

3.4. Chain Conflict Resolution. The aforementioned algorithm for selecting a block-generation node cannot guarantee that only one block-generation node is selected in the same round. When multiple legal block-producing nodes exist, multiple legal blocks can appear at the same height, that is, a fork. The consensus algorithm in this study uses the longest chain rule to solve the bifurcation conflict; that is, the chain with the maximum length is regarded as the optimal chain.

As shown in Figure 6, the development at height D has a fork, but the length of Branch 1 exceeds that of Branch 2. Thus, Branch 1 is regarded as the optimal chain.

4. Results

4.1. A Transaction Delay Time. Transaction delay time represents the time it takes for transactions to be submitted to the block confirmation. From the data sent by sensors to the final transactions entering blocks, the total delay time can be divided into two sections.

$$\text{Latency} = L_1 + L_2, \tag{15}$$

where L_1 represents the time from sensor submission to gateway detection completion and L_2 represents the time from gateway submission transaction to block packaging. The following figure shows the schematic structure of the transaction delay time simulation experiment, which is shown in Figure 7.

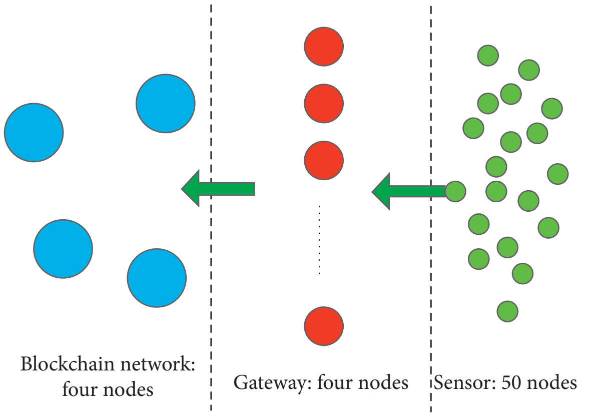


Figure 7: Transaction delay structure diagram.

The histogram of transaction delay statistics is shown in Figure 8.

Approximately 90% of transactions have a delay time within 5000 ms.

Figures 9 and 10 show the proportion of the total delay in the L_1 and L_2 stages (from sensors to gateways).

The average proportion of the overall delay from the sensor to the gateway confirmation stage is approximately 10%. Figure 11 displays a comparison chart of the two-stage delay.

The transaction delay in the first phase remains basically unchanged, whereas that in the second phase exhibits an obvious periodic characteristic. The reason is that the data submission of sensors is periodic; thus, the transaction submission of gateways also remains basically periodic. When the moment in which gateways submit transactions is close to the next block-generation round, only a short transaction delay can be confirmed; otherwise, a large transaction delay time is required.

4.2. Throughput. Throughput examines the total amount of blockchain transaction processing per unit time. The simulation experiment parameters are selected as follows:

(i) Blockchain nodes: four.

(ii) Workload node: 10 tps/node, 10–100, corresponding to 100–1000 input tps.

(iii) Leader node election: Vrf-PoW.

The schematic of the simulation experiment is shown in Figure 12.

The experimental results are shown in Figure 13.

Within a certain range (test load <800 tps), the system throughput synchronously increases as the test load increases, and when the test load is 800 tps, the system throughput reaches the maximum which is close to 600 tps. When the test load exceeds 800 tps, the actual system throughput begins to decrease, indicating that the simulation system has been overloaded.

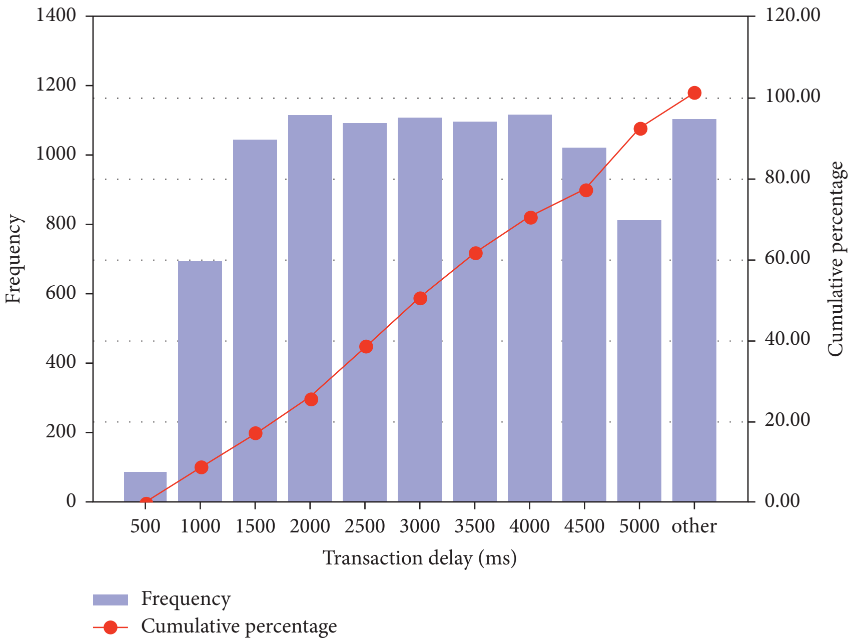


Figure 8: Transaction delay analysis.

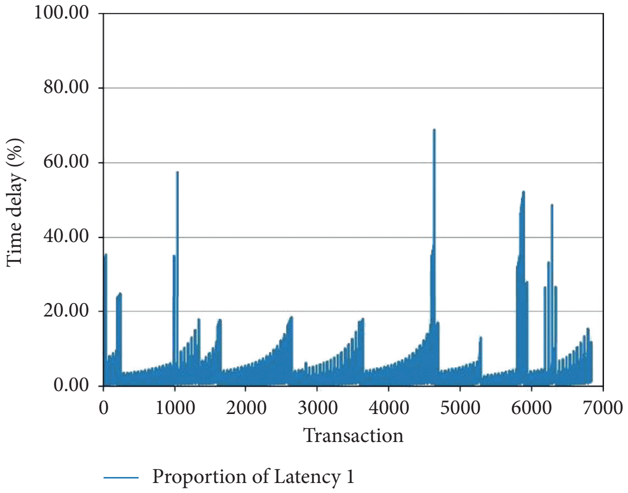


Figure 9: Proportion of Latency 1.

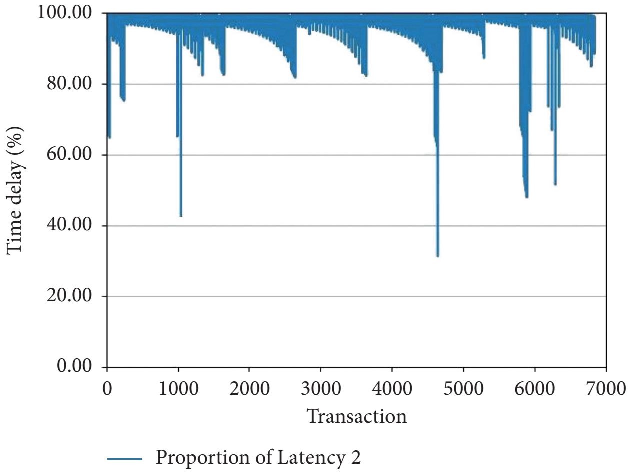


Figure 10: Proportion of Latency 2.

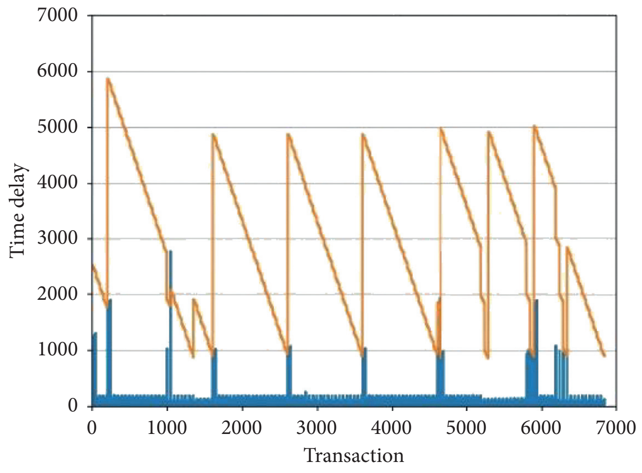


Figure 11: Comparison of transaction delay between two stages.

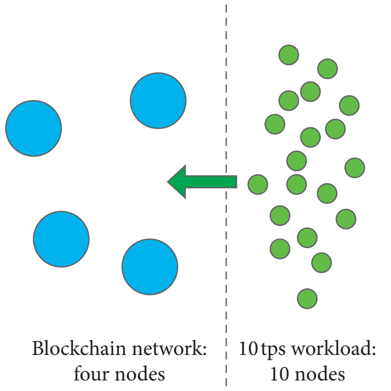


Figure 12: Throughput structure diagram.

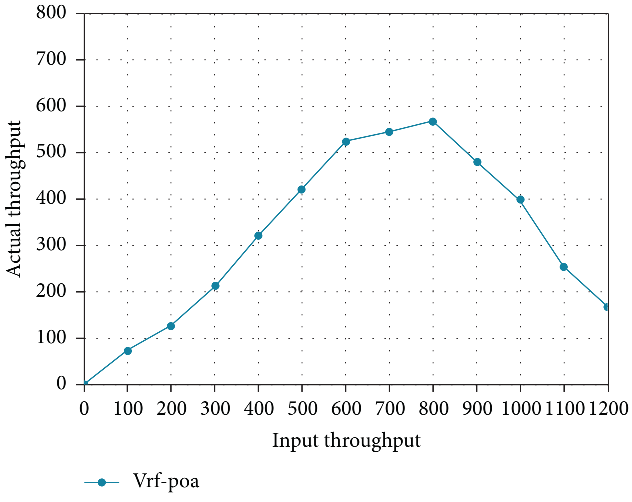


Figure 13: Throughput analysis.

5. Conclusion

Based on the blockchain theory and IoT system architecture, this study proposes a blockchain consensus mechanism suitable for lightweight IoT devices. The Diffie–Hellman algorithm is used for key negotiation with blockchain nodes, sensors, and zones. Blockchain nodes can use this shared key to generate HMAC signatures for sensor-aware transactions, including VRF, to achieve the offline fast election of block nodes. The winning node becomes the block node. Machine learning means are also introduced to identify or eliminate outliers in sensor data before the data are uploaded.

We believe that continuous research on blockchain and IoT can bring about major changes in the industry and promote continuous social development.

Data Availability

The data that support the findings of the research are available from the corresponding author.

Conflicts of Interest

The authors declare that there are no conflicts of interest.

Acknowledgments

This study was supported by Key Research and Development Plan of Anhui Province (201904a06020056).

References

[1] J. Gubbi, R. Buyya, S. Marusic, and M. Palaniswami, “Internet of things (IoT): a vision, architectural elements, and future directions,” *Future Generation Computer Systems*, vol. 29, no. 7, pp. 1645–1660, 2013.

[2] M. A. Khan and K. Salah, “IoT security: review, blockchain solutions, and open challenges,” *Future Generation Computer Systems*, vol. 82, pp. 395–411, 2018.

[3] K. Christidis and M. Devetsikiotis, “Blockchains and smart contracts for the internet of things,” *IEEE Access*, vol. 4, pp. 2292–2303, 2016.

[4] S. K. T. Mehedi, A. A. M. Shamim, and M. B. A. Miah, “Blockchain-based security management of IoT infrastructure with Ethereum transactions,” *Iran Journal of Computer Science*, vol. 2, no. 3, pp. 189–195, 2019.

[5] J. Wu and N. Tran, “Application of blockchain technology in sustainable energy systems: an overview,” *Sustainability*, vol. 10, no. 9, p. 3067, 2018.

[6] G. Ayoade, V. Karande, L. Khan, and K. Hamlen, “Decentralized IoT data management using blockchain and trusted execution environment,” in *Proceedings of IEEE International Conference on Information Reuse and Integration (IRI)*, pp. 15–22, IEEE, Salt Lake City, UT, USA, July 2018.

[7] A. Dorri, S. S. Kanhere, and R. Jurdak, “Towards an optimized blockchain for IoT,” in *Proceedings of 2017 IEEE/ACM Second International Conference on Internet-of-Things Design and Implementation (IoTDI)*, IEEE, Pittsburgh, PA, USA, pp. 173–178, April 2017.

[8] A. Dorri, S. S. Kanhere, R. Jurdak, and P. Gauravaram, "Lsb: a lightweight scalable blockchain for iot security and privacy," 2017, https://arxiv.org/abs/1712.02969.

[9] O. Novo, "Blockchain meets IoT: an architecture for scalable access management in IoT," *IEEE Internet of Things Journal*, vol. 5, no. 2, pp. 1184–1195, 2018.

[10] N. Kshetri, "Blockchain's roles in strengthening cybersecurity and protecting privacy," *Telecommunications Policy*, vol. 41, no. 10, pp. 1027–1038, 2017.

[11] A. Panarello, N. Tapas, G. Merlino, F. Longo, and A. Puliafito, "Blockchain and iot integration: a systematic survey," *Sensors*, vol. 18, no. 8, p. 2575, 2018.

[12] D. Vujičić, D. Jagodić, and S. Ranđić, "Blockchain technology, bitcoin, and ethereum: a brief overview," in *Proceedings of 17th International Symposium Infoteh-Jahorina (Infoteh)*, IEEE, East Sarajevo, Bosnia and Herzegovina, pp. 1–6, March 2018.

[13] Q. Jing, A. V. Vasilakos, J. Wan, J. Lu, and D. Qiu, "Security of the internet of things: perspectives and challenges," *Wireless Networks*, vol. 20, no. 8, pp. 2481–2501, 2014.

[14] H. Mayer, "ECDSA security in bitcoin and ethereum: a research survey," *CoinFaabrik*, vol. 28, p. 126, 2016.

[15] D. Mahto, D. A. Khan, and D. K. Yadav, "Security analysis of elliptic curve cryptography and RSA," in *Proceedings of the world congress on engineering*, vol. 1, pp. 419–422, London, UK, July 2016.

[16] L. Zhou, L. Wang, Y. Sun, and P. Lv, "Beekeeper: a blockchain-based iot system with secure storage and homomorphic computation," *IEEE Access*, vol. 6, pp. 43472–43488, 2018.

[17] J. L. Zhao, S. Fan, and J. Yan, *Overview of Business Innovations and Research Opportunities in Blockchain and Introduction to the Special Issue*, Springer Open, New York, NY, USA, 2016.

[18] A. Kiayias and G. Panagiotakos, "Speed-security tradeoffs in blockchain protocols," *IACR Cryptology ePrint Archive*, vol. 1019, 2015.

[19] D. Guinard, V. Trifa, and E. Wilde, "A resource oriented architecture for the web of things," in *Proceedings of 2010 Internet of Things (IOT)*, pp. 1–8, IEEE, Tokyo, Japan, December 2010.

[20] Q. He, N. Guan, M. Lv, and W. Yi, "On the consensus mechanisms of blockchain/dlt for internet of things," in *Proceedings of 2018 IEEE 13th International Symposium on Industrial Embedded Systems (SIES)*, IEEE, Graz, Austria, pp. 1–10, June 2018.

Hindawi
Shock and Vibration
Volume 2020, Article ID 8837071, 9 pages
https://doi.org/10.1155/2020/8837071

Research Article

A Novel Efficient Passive Spatial Orientation Detection Method of UMT Enabled by ISB

Zhengqiang Xiong,[1] Tao Sun ,[1] Zhengxing Wang,[2] Yuhao Wu,[1] and Jie Yin[1]

[1] *The Department of Electrical and Information, Wuhan University, Wuhan 430072, China*
[2] *State Key Laboratory of Bridge Structure's Health and Safety, Wuhan 430034, China*

Correspondence should be addressed to Tao Sun; suntao@whu.edu.cn

Received 15 June 2020; Revised 1 July 2020; Accepted 1 September 2020; Published 28 September 2020

Academic Editor: Chaoqun Duan

The passive detection and direction-of-arrival (DOA) estimation problem is of great importance in many underwater applications such as target reconnaissance and data collection. In this paper, an Efficient Correlation-based Orientation Detection (ECOD) method is proposed to achieve high efficiency. Without high computational complexity in any Transform Domain, the time consumption of ECOD is largely reduced, which is especially critical for underwater intrusion detection, territorial waters protection, and many other real-time underwater applications. To achieve good invisibility, we design an intelligent submerged buoy (ISB) structure, which consists of six embedded hydrophones and an in situ electronic control unit (IECU). As a supplement to solutions against complex underwater environments, a hybrid ECOD method is also developed by involving the cooperation from underwater distributed sensor networks. To be specific, when high SNR signals are not recorded by a single ISB node, other distributed sensors are scheduled to assist in cooperative sensing. Simulation experiments demonstrate the efficiency of the ECOD method in passive 3D spatial orientation of underwater acoustic target and show that the ECOD method has a better performance in time consumption compared with general DOA algorithms.

1. Introduction

Passive spatial detection and orientation estimation of a moving acoustic target is of great importance in many underwater applications [1, 2]. For example, tracking adversary targets, involving submarines and Autonomous Underwater Vehicles (AUV), are critical for security purposes [3, 4]. Also, with regard to biological research, it is beneficial to obtain the rough position and spatial behaviour features of marine creatures [5]. In addition, sensing underwater moving abnormal objects is potentially helpful to disaster forecast and relevant protection [6]. Hence, it is very important to develop effective underwater detection and spatial orientation technologies.

1.1. Motivations and Challenges. There are major considerations and challenges for underwater spatial orientation detection. Due to difficulties in deployment and maintenance, the devices carrying out underwater detection should feature extremely low complexity and energy consumption, which may undermine the accuracy of detection results [7]. Additionally, it is widely proved that the underwater acoustic channel suffers from limited bandwidth, serious multipath effect, and high latency, which leads to packet loss and error bits, interfering the interdevice communications [8]. Furthermore, for homeland security applications, the deployed devices and interdevice communication signals can be possibly be exposed to adversary targets [9]. Therefore, the underwater passive detection and spatial orientation strategy is supposed to be extremely efficient, i.e., low complexity, energy saving, and stable with favorable invisibility.

Several research studies have been carried out in the fields of underwater detection and orientation. In [10], performances of well-known techniques implemented by a single acoustic vector sensor (AVS) are studied. Though the performance of practical AVS-based systems provides a valuable insight, the cost of an acoustic vector sensor is still much higher than omnidirectional ones. Zou et al. [11] raises a two-step approach to reduce the complexity of AVS-based direction of arrival (DOA) estimation within a spatial sparse representation framework. In [12], the DOA estimation is achieved by a matrix-pencil pair derived from time-delayed signals collected from a single AVS. However, the very high computational payload significantly decreases the practicability of above methods. Disregarding the usage of AVS, the DOA and relevant techniques are common for precise localization. Shao et al. [13] develop efficient closed-form angle-of-arrival- (AOA-) based self-localization algorithms to improve the localization accuracy. In [14], the effectiveness of two novel positioning schemes based on n time-of-arrival (TOA) measurements is validated. A time difference of arrival (TDOA) algorithm for passive localization via estimating the delay of two correlated channels is proposed, which outperforms other TDOA algorithms [15]. However, most of AOA, TOA, and TDOA approaches are algorithms with high computational payload, which is a challenging difficulty to interdevice time synchronization. Moreover, both vector sensors and distributed sensors need to transmit recorded signals by wireless channel back to a control centre for decision, leading to more unreliability. In [16], the proposed trilateration algorithm achieves precise underwater target positioning by utilizing received signal strength indicator (RSSI) value, which is generally obtained by the signal transmission power. Nevertheless, due to lack of priori knowledge of noncooperative targets, we can hardly know the target's signal transmission power.

1.2. Contributions and Organization. Compared with existing research works, the specific contributions of this paper are given below:

An intelligent submerged buoy (ISB) structure is designed, which consists of an intelligent control board and three pairs of embedded hydrophones. Components of the ISB are integrated into a sphere structure, which contributes to receive signals from omni-directions and effectively reduce flow resistance for position stabilization.

An ECOD method is developed. Its efficient performance is achieved by a low-complexity crosscorrelation algorithm of input signals for each hydrophone pair. Without high computational complexity in any Transform Domain, the time consumption of the method is largely reduced, which is especially critical for some real-time applications including underwater intrusion detection and territorial waters protection.

A Hybrid ECOD method is proposed. Other ISBs distributed in surrounding underwater sensor networks will join in to improve the spatial orientation estimation accuracy if some ISBs do not provide good sensing performance or fail to work, aiming to improve the stability of the ECOD method further. The remainder of this paper is organized as follows. Section 2 describes the underwater orientation detection system. The ECOD method is proposed in Section 3, and Hybrid ECOD method, a cooperative detection, and orientation strategy are developed. Numerical experiment results are presented and discussed in Section 4. Section 5 draws a conclusion.

2. System Model

For traffic control in harbors or a homeland security sensitive sea area monitoring, a moving acoustic targets passive detection and spatial orientation system is designed for 3D underwater space, as shown in Figure 1. Several submerged buoys are anchored on the seabed, which are able to passively and continuously receive acoustic signals from underwater targets. Each submerged buoy is designed to be a compact sphere structure to provide stable connections with six embedded hydrophones.

Let equation (1) be the signal received by the m^{th} hydro phone in time domain:

$$y_n(t) = x_m(t) + n_m(t), \tag{1}$$

where $x_m(t)$ represented the received signal from the target and $n_m(t)$ denoted the additional noise. To simplify the calculation, continuous-time signals are truncated and discretely sampled. The acoustic signal monitored by the m^{th} sensor during $N\Delta t$ is then described as follows:

$$s_m = [y_m(1), \ldots, y_m(n), \ldots, y_m(N)], \tag{2}$$

where $y_m(n)$ is the n^{th} sample of the received signal and N is the sampling number.

In order to improve azimuth resolution and signal-noise ratio, the hydrophones are deployed as sensor arrays. The received signals are synchronously recorded by embedded hydrophones and processed simultaneously by the in situ control unit (IECU) on each submerged buoy. Because of varies noisy from marine animals, a prejudgment mechanism has also been developed to eliminate those incoherent signal, which is explained in Section 3.2. For energy conservation, orientation estimation is carried out on the basis of the above operation.

3. Intelligent Submerged Buoy-Enabled Target Orientation Detection

In this section, an efficient target orientation detection strategy enabled by intelligent submerged buoy (ISB) is proposed.

3.1. The Structure of the Intelligent Submerged Buoy. The basic idea for the submerged buoy enabled method is that the intelligent buoy is a relatively autonomous system to undertake the entire task during passive spatial detection and DOA detection independently. The control board sends commands to embedded hydrophones, as shown in Figure 1, makes decisions, and has wireless acoustic communications ability with the control centre on shore.

For high-effective signal sampling to analyze spatial orientation, an ISB consists of six embedded hydrophones

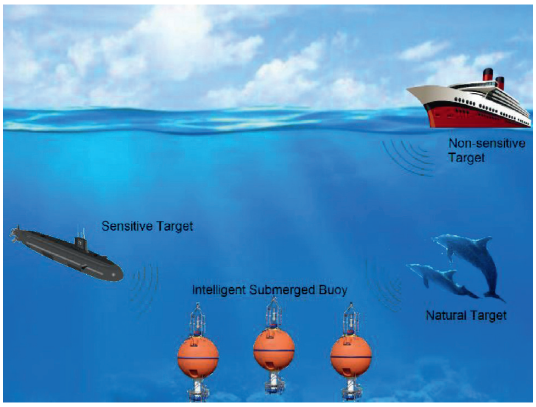


Figure 1: System model of the underwater acoustic detection system for moving target.

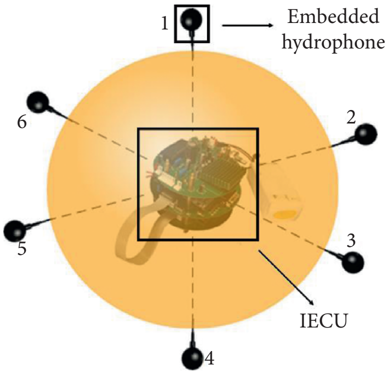


Figure 2: Components of an intelligent submerged buoy (ISB).

orthogonally distributed in 3D Cartesian coordinate, which is shown in Figure 2. The hydrophone array is divided into three pairs to be responsible for three dimensions, i.e., north-south (N-S), west-east (W-E), and up-down (U-D). Each pair of hydrophones collects the signal in one assigned dimension, which avoids redundancy in sensor deployment. Moreover, the invisibility of ISB is improved by aggregating IECU and hydrophones together into a sphere structure. The submerged buoy achieves independent low-complexity orientation computation enabled by the intelligent board and its received signals.

The structure design of the intelligent submerged buoy is inspired by the human auditory system. It is well known that the human auditory system consists of two ears and the auditory centre in the head, as shown in Figure 3. The sound source is localized by time delay estimation between each cochlea. Similarly, each pair of hydrophones receives acoustic signals transmitted from underwater targets and the in situ electronic control unit (IECU) functions as "auditory centre" by processing these received signals.

The shape of submerged buoy is approximately as a sphere, and the three pairs of hydrophones are exactly orthogonally distributed in three spatial dimensions, as illustrated in Figure 2. Each pair of hydrophones can judge that the target's location lies in the same side of the positive axis or the negative one in the dimension where they are distributed as far as a 3D spatial coordinate is established with the ISB as its origin. On this basis, the IECU estimates the orientation of the underwater target through a synthesis processing of these three groups of signals, which contains direction information in their dimension, respectively.

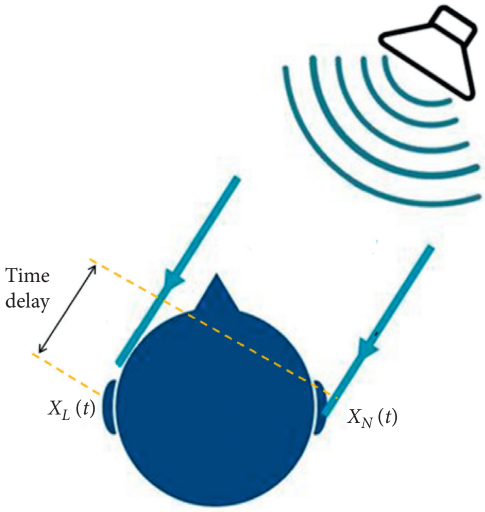


Figure 3: Orientation principle of the human auditory system.

Table 1: Frequency range of acoustical signals generated by some typical underwater targets.

Types of underwater target	Frequency range (Hz)
Whale	10–40
Toadfish	200–300
Ship	50–1000
Submarine	100–1000
Dolphin	7000–120000

3.2. Predetection of Underwater Acoustic Targets. Table 1 shows the frequency range of acoustical signals generated by some typical underwater targets. Since the power spectrum of pure background noises differs a lot from that of a signal of underwater acoustic targets, the power spectrum of received signals for the in-device hydrophones is applied to target detection (judge an underwater acoustic target is out of detection range or not, through its energy distribution in the power spectrum). For each received signal of in-device hydrophone, the power spectrum can be obtained through Short-Time Fourier Transform (STFT):

$$\mathrm{STFT}(n,\omega) = \sum_{m=-\infty}^{\infty} \left[y_i(m) \cdot w(n-m) \right] \cdot e^{-jm\omega}, \quad (3)$$

where$y_i(m)$ is the m^{th} sample of the signal received by the i^{th} hydrophone and $w(m)$ denotes the window function.

Considering the difficulty in battery replacement for the underwater system, several methods have been adopted for energy conservation. Since hydrophones receive only background noises without any available signal of target for most of the time, it is unnecessary for the IECU to keep

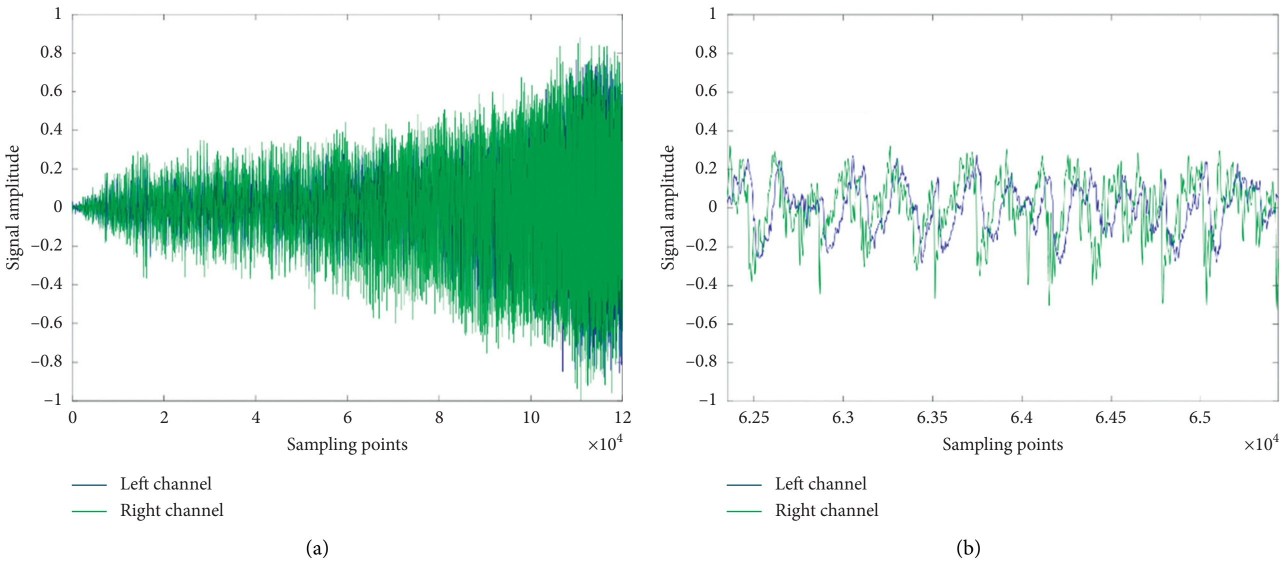


Figure 4: A dual-channel signal waveform of engine shipping noise recorded at East Sicily: (a) complete waveform and (b) a certain part of this signal.

working all the time. Accordingly, a threshold based on statistical characteristics of the overall spectrum is set to wake up the sense module. In this way, the central IECU is able to figure out potential threatens approaching the detection area in time, while ignoring some natural or cooperative targets to save energy.

3.3. The Efficient Correlation-Based Orientation Detection (ECOD) Method. Sound source localization algorithms mainly consist of two key components: azimuth and distance estimation. The range estimation is generally achieved by power attenuation calculation. However, it is arduous to figure out the signal strength of acoustic source due to the passive listening mechanism taken in our method. Hence, this work focuses on the orientation detection problem.

The spatial behaviour features (relative position, moving direction, etc.) of the underwater acoustic target are acquired by a dual-channel signal and its associated waveform time delay analysis. Considering that a target's relative position varied over time in realistic applications, the input signal is segmented into frame sequences. Let $y_L(n)$ and $y_R(n)$ be the time sequence of left and right channels, respectively; then, the crosscorrelation function of them is formulated as

$$R_{LR} = \frac{1}{N} \sum_{k=n}^{n+N-1} y_L(k) \cdot y_R(k-m), \tag{4}$$

where N is frame length and m represents the amount of displacement of $y_L(n)$ and $y_R(n)$.

In order to extract features of realistic underwater acoustic signals, some measured data are adopted in this paper. A part of engine shipping noise was recorded at East Sicily by the observatory NEMO set by the Laboratory of Applied Bioacoustics (LAB) of the Technical University of Catalonia (Barcelona Tech, UPC). Figure 4 shows the waveform of this dual-channel signal. We can see that the amplitudes of both the left and right channels grow gradually for the whole process. This indicates the target is getting closer to the ISB as time goes by. Moreover, the green curve corresponding to the right channel stays ahead of the blue one, which demonstrates that the target approaches the ISB from the right side.

Since structures of acoustic signal waveform of both channels are similar to each other except a slight time delay between them, the maximal crosscorrelation coefficient can be determined via time delay estimation (TDE). Due to their approximate waveforms, the displaced right channel signal is approximated as copy of left channel signal. That is to say $y_R(n-D) \approx y_L(n)$. Then, the corresponding crosscorrelation coefficient can be presented as follows:

$$R_{LR} = \frac{1}{N} \sum_{k=n}^{n+N-1} [y_L(k)]^2 = \frac{E_n}{N}, \tag{5}$$

where E_n denotes short-time energy.

The next step is to figure out the maximal crosscorrelation coefficient M and corresponding TDOA (Time Difference of Arrival) D of left and right channels. Since the received signals from each pair of hydrophone are similar in time domain, the crosscorrelation coefficient reaches the maximum as time delay from corresponding channels is zero. According to the analysis above, M can be employed to estimate the distance between the ISB and target. This operation is illustrated in Figure 5. Let D be the time delay between acoustic signals of discrete time received by the right and left channels. To acquire the maximal crosscorrelation coefficient, signals from the right channel are translated to eliminate the time delay. Afterwards, D can also be defined as the TDOA via this operation, which contributes to identify the target bearing. We can see from Figure 5 that the target is on the left side of the observation point.

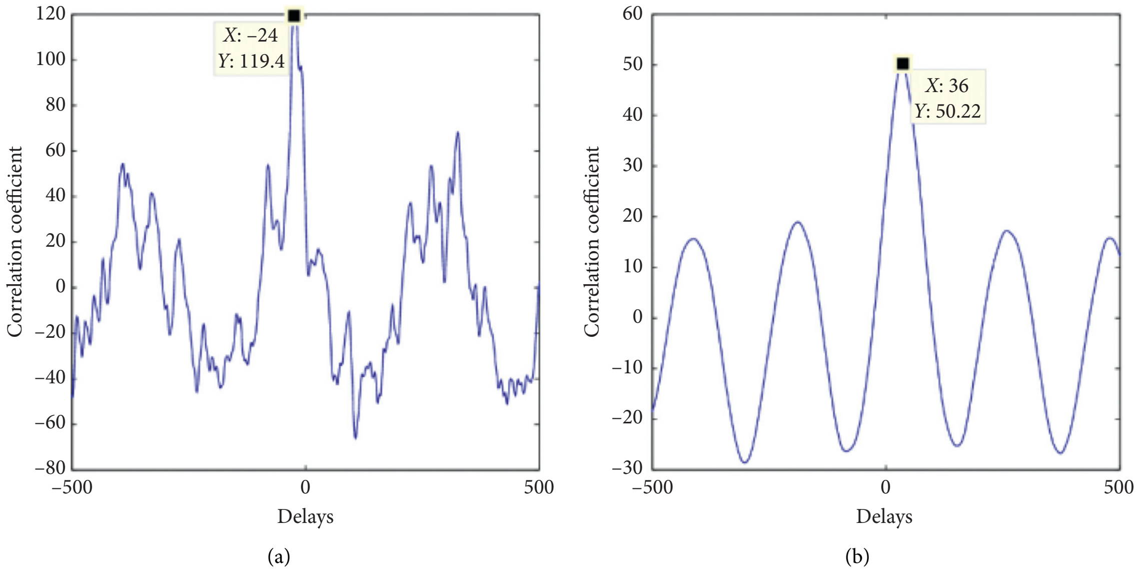


FIGURE 5: Peak search of the crosscorrelation coefficient: (a) the time delay of the maximal crosscorrelation $D > 0$ and (b) the time delay of the maximal crosscorrelation $D < 0$.

Since ISBs are fixed to the seabed, the possible location areas of underwater acoustic targets can be divided into 4 quadrants (quadrant I–quadrant IV in Figure 6). In far field applications, both the target and the ISB can be assumed to be a point source, since its dimensions are much smaller compared with the acoustic wavelength. Thus, the wavefront incident on the ISB can be considered as planar wave [10]. As shown in Figure 7, the 3D spatial orientation of an underwater target relative to the ISB is described by a direction vector, which can be expressed with azimuth angle θ and elevation angle α.

Elevation angle α is able to be acquired once the angle between target direction vector and Z-axis is calculated, as shown in Figure 8. Assume that ΔL_{14} denotes the transmission distance difference of the planar wave between 1sthydrophone and 4thhydrophone (set in Z-axis), as demonstrated in Figure 8. The elevation angle α is given by

$$\alpha = \arcsin\left(\frac{\Delta L_{14}}{\Phi}\right) = \arcsin\left(\frac{V_w \Delta T_{14}}{\Phi}\right), \tag{6}$$

where V_w denotes the velocity of sound traveling underwater, ΔT_{14} denotes time difference of arrival between the 1st hydrophone and the 4thhydrophone, and $\varnothing$ is the diameter of an ISB.

In order to estimate the azimuth angle θ, both angles with X-axis and Y-axis are needed. Let θ_i $(i = x, y)$ be the angle between target direction vector and axis, as shown in Figure 9, when the target locates in area I, the azimuth angle θ is

$$\theta = \arctan\left(\frac{\cos\theta_y}{\cos\theta_x}\right) = \arctan\left(\frac{\Delta T_{25}}{\Delta T_{36}}\right). \tag{7}$$

Finally, the azimuth angle θ is described as follows:

$$\theta = \begin{cases} \arctan\left(\dfrac{\Delta T_{25}}{\Delta T_{36}}\right), & \text{in area I,} \\[2ex] 180^\circ - \arctan\left(\dfrac{\Delta T_{25}}{\Delta T_{36}}\right), & \text{in area II,} \\[2ex] \arctan\left(\dfrac{\Delta T_{25}}{\Delta T_{36}}\right) - 180^\circ, & \text{in area III,} \\[2ex] -\arctan\left(\dfrac{\Delta T_{25}}{\Delta T_{36}}\right), & \text{in area IV,} \end{cases} \tag{8}$$

where ΔT_{ij} denotes time difference of arrival between the ith hydrophone and the jth hydrophone. arget can also be approximated, combining with short-time energy analysis of the signal in different time periods. Although the exact transition energy intensity is unknown, our method is still able to efficiently identify the direction and movement trajectory relative to ISB. As it has been widely proved that the amplitude of crosscorrelation coefficient has a positive correlation with the target signal, the received signal power at each time is normalized. By this, the relative position can be determined.

3.4. *A Hybrid ECOD Method.*

The proposed ECOD method is potentially vulnerable when the time delay between two channels is too short to be distinguished. To solve this, a robust orientation detection strategy enabled by cooperative sensing of underwater intersensor-network is developed. There are minor limitations for ECOD method enabled by a single ISB. Firstly, the time delay D would be too small to be distinguished when the target moves just in one of the three axis directions in the 3D Cartesian coordinate with an ISB

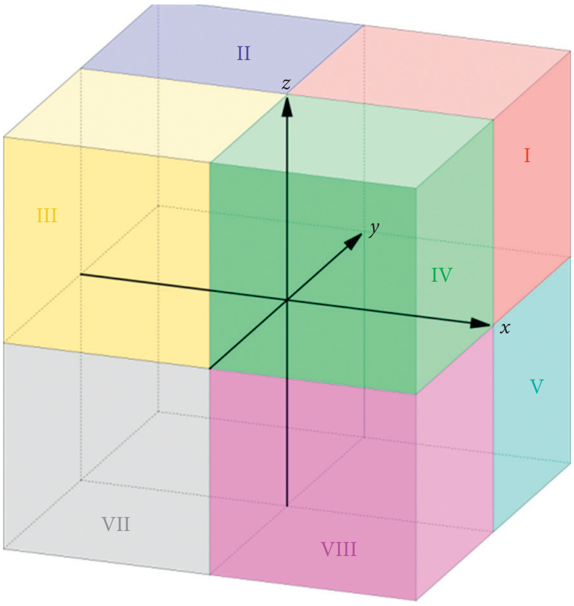


Figure 6: Preparatory division of underwater acoustic target's location areas.

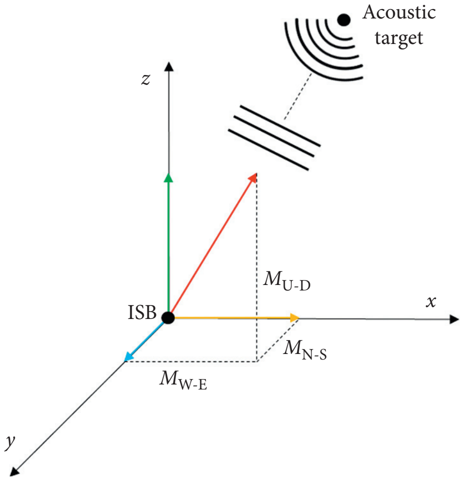


Figure 7: Explanation of a spatial direction vector.

located at the origin point. In the case of low SNR, the sign of D could even be the opposite from its real value. Furthermore, Method 1 for the comprehensive orientation detection would fail when one hydrophone in the submerged buoy is broken down. Hence, it is dispensable to take some assistant measures.

The distributed ISB nodes in underwater sensor networks provide critical assistance. Although it is highly costly for the distributed sensors to undertake all orientation detection tasks, it is reasonable to "ask" one or two sensor nodes for cooperation when necessary. When $|D| < \delta$, it indicates the target is nearly equally distant to those associated hydrophones in the submerged buoy. In this case, the ISB communicates with other ISB nodes distributed in the nearby underwater sensor network for cooperative sensing. In this way, the accuracy of orientation detection in corresponding dimension is significantly improved. This Strategy can also be applied in the situation that one ISB breaks down. To be specific, the principle of Hybrid ECOD method is depicted in Figure 10.

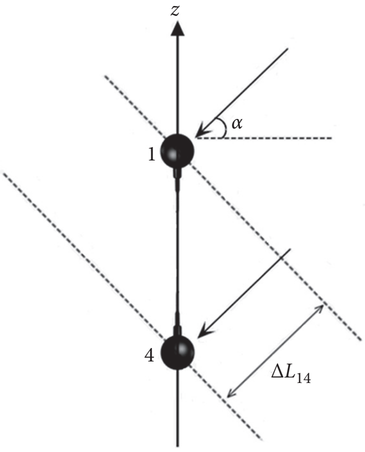


Figure 8: Estimation of elevation angle for spatial direction vector.

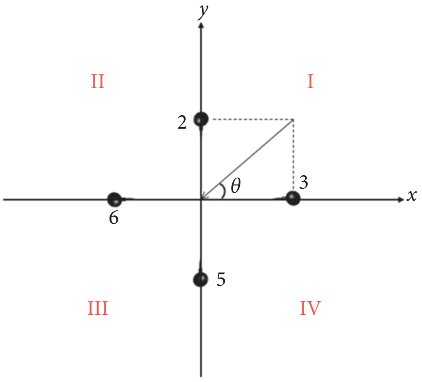


Figure 9: Estimation of azimuth angle for spatial direction vector.

4. Experiments and Analysis

4.1. The Efficiency Validation.

To evaluate the performance of the proposed algorithm, simulated experiments were carried out to locate a moving target. To demonstrate the superiority of the proposed algorithm, typical azimuth estimation method including MUSIC was compared with the proposed algorithm.

Since targets are expected to be detected online, the real-time performance should be firstly to be considered. Thus, the average time consumption is counted for both algorithms. With the length increment of the input signal, the time consumption of each method becomes longer accordingly. Figure 11 shows the average time consumption of each method. It can be seen that ECOD keeps 50% average time consumption less than that of MUSIC. The result shows that ECOD is still more suitable for underwater detection and monitoring.

4.2. Target Tracking Simulation.

Besides orientation, the ECOD method is also able to obtain the spatial behaviour of the underwater acoustic target. As demonstrated in Section

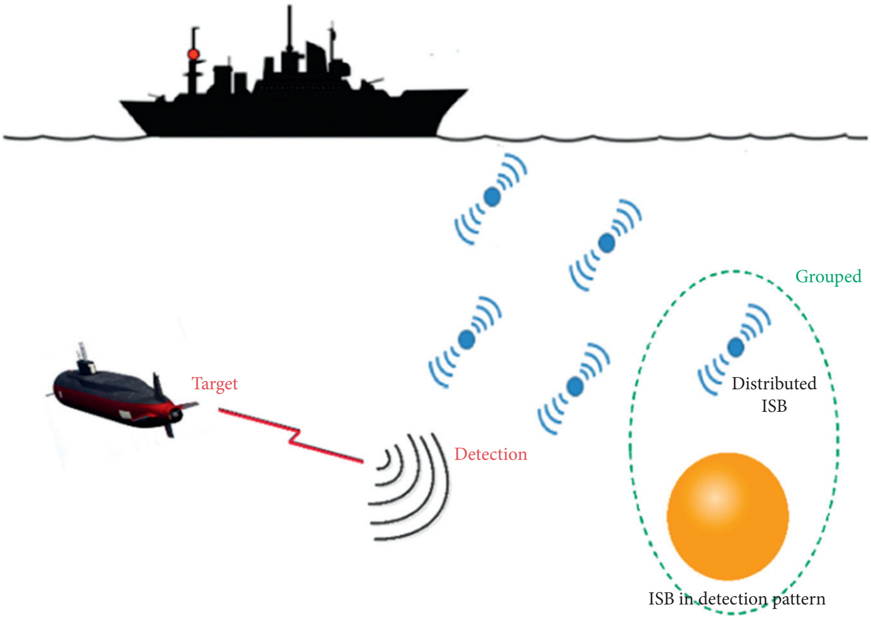


Figure 10: An application scenario of the Hybrid ECOD Method.

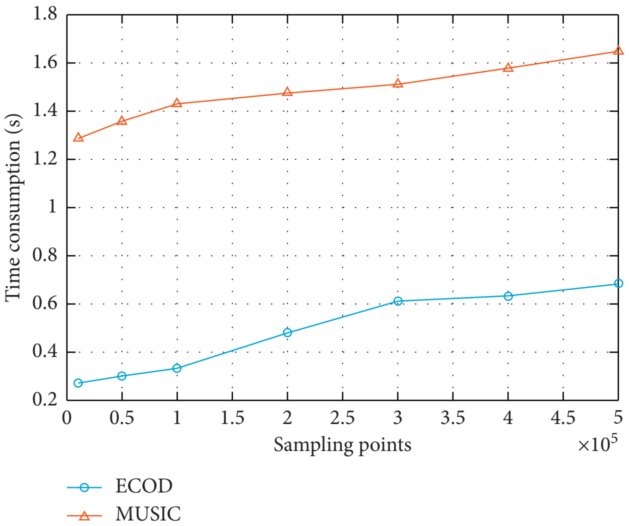


Figure 11: The time consumption comparison between the ECOD method and MUSIC method.

3, the maximal crosscorrelation coefficient M is approximately equal to the average power of each signal frame. Thus, the ECOD method can obtain the motion state of underwater acoustic target.

In order to test the performance of proposed ISB in underwater monitoring and tracking, the target is assumed to approach our underwater sensor networks from a distance, as described in Figure 12. Some measured dual-channel data are applied and processed to demonstrate the observed data received by three pairs of hydrophones in an ISB: dual-channel signal is developed into three groups of input signals corresponding to the three pairs of hydrophones, respectively, distributed in 3D space, by adjusting the time delay of the two channels. The three groups of input signals are segmented into 40 frames with a smooth window. The IECU obtains 40 groups of direction vectors using the ECOD method and assigns these vectors' relative values based on the maximal crosscorrelation coefficient M of each frame. 40 spatial location points in the coordinate system will be arranged according to time sequences, which is the target orientation information and its trajectory.

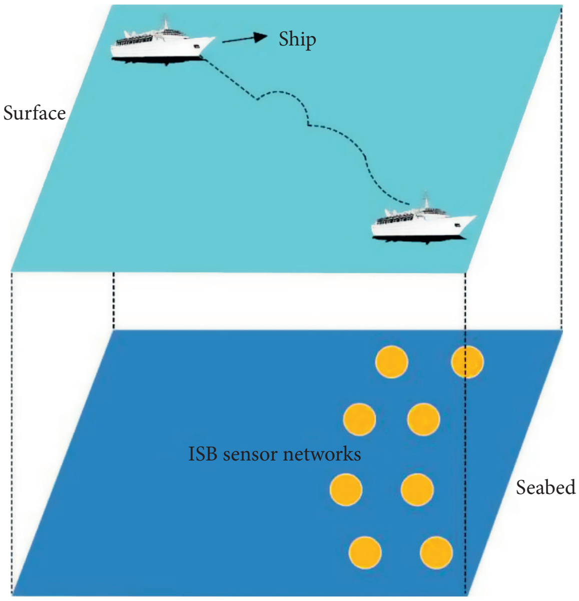


Figure 12: Simulation scenario: a ship is approaching to the ISB from a distance.

As described in Section 3, spatial feature parameters of the target are represented by 3D coordinates after simple vector synthesis computation. To achieve this, the 40 points in the coordinate system which represent the current orientation are connected together one by one

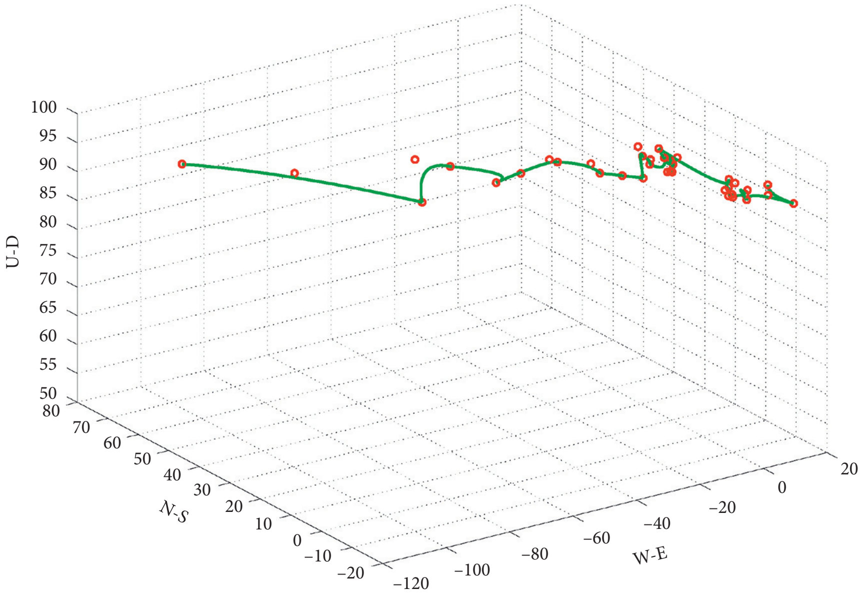


Figure 13: The 3D trajectory generated by the proposed ECOD method.

according to time sequences. In this way, the trajectory of the target is derived through Bezier curve fitting, as shown in Figure 13. The direction of the target in a particular time as well as its moving behaviour is then obtained in this relative trajectory.

5. Conclusion

An ECOD method of efficient 3D spatial orientation and motion estimation of an underwater target based on a novel intelligent submerged buoy structure is proposed. The high efficiency is achieved by low-complexity crosscorrelation algorithm and independent signal processing in the in situ electronic control unit of the intelligent submerged buoy. To achieve robustness, a distributed hybrid ECOD method featuring the cooperative sensing of underwater sensor networks is developed. Other distributed sensors are scheduled to provide cooperation for sensing when a separate pair of embedded hydrophones is unable to obtain highly distinguishable signals. Numerical simulations show that our ECOD method obtains the spatial behaviour features of the simulated underwater target and has better performance on efficiency than MUSIC algorithm. Because of passive detection pattern and real-time capability, the ECOD method is much more suitable for underwater detection and monitoring with low cost.

Data Availability

The data used to support the findings of this study are available from the corresponding author upon request.

Disclosure

The work presented here was carried out in collaboration among all authors.

Conflicts of Interest

The authors declare that there are conflicts of interest.

Authors' Contributions

All authors have contributed to, seen, and approved the manuscript.

Acknowledgments

This work was supported by Project of Technical Innovation of Hubei Province (2018AAA029).

References

[1] S. Zhang, D. Li, and J. Chen, "A link-state based adaptive feedback routing for underwater acoustic sensor networks," *IEEE Sensors Journal*, vol. 13, no. 11, pp. 4402–4412, 2013.

[2] H. Ramezani, H. Jamali-Rad, and G. Leus, "Target localization and tracking for an isogradient sound speed profile," *IEEE Transactions on Signal Processing*, vol. 61, no. 6, pp. 1434–1446, 2013.

[3] K. A. C. Baumgartner, S. Ferrari, and T. A. Wettergren, "Robust deployment of dynamic sensor networks for cooperative track detection," *IEEE Sensors Journal*, vol. 9, no. 9, pp. 1029–1048, 2009.

[4] Y. Zeng, J. Cao, J. Hong, S. Zhang, and L. Xie, "Secure localization and location verification in wireless sensor networks: a survey," *The Journal of Supercomputing*, vol. 64, no. 3, pp. 685–701, 2013.

[5] Y. Pailhas, C. Capus, and K. Brown, "Dolphin-inspired sonar system and its performance," *IET Radar, Sonar & Navigation*, vol. 6, no. 8, pp. 753–763, 2012.

[6] C. R. Krishna and P. S. Yadav, "A hybrid localization scheme for underwater wireless sensor networks," in *Proceedings of the IEEE 2016 International Conference on Inventive Computation Technologies (ICICT)*, pp. 1–4, Ghaziabad, India, February 2014.

[7] S. Sendra, J. Lloret, J. M. Jimenez, and L. Parra, “Underwater acoustic modems,” *IEEE Sensors Journal*, vol. 16, no. 11, pp. 4063–4071, 2016.

[8] E. Lattanzi, V. Freschi, M. Dromedari, and A. Bogliolo, “An acoustic complexity index sensor for underwater applications,” *IEEE Sensors Journal*, vol. 16, no. 11, pp. 4043–4050, 2016.

[9] V. Vadori, M. Scalabrin, A. V. Guglielmi, and L. Badia, “Jamming in underwater sensor networks as a Bayesian zero-sum game with position uncertainty,” in *Proceedings of the IEEE GLOBECOM*, pp. 1–6, San Diego, CA, USA, December 2015.

[10] A. Bereketli, M. B. Guldogan, T. Kolcak, T. Gudu, and A. L. Avsar, “Experimental results for direction of arrival estimation with a single acoustic vector sensor in shallow water,” *Journal of Sensors*, vol. 2015, Article ID 401353, 10 pages, 2015.

[11] Y.-X. Zou, B. Li, and C. H. Ritz, “Multi-source DOA estimation using an acoustic vector sensor array under a spatial sparse representation framework,” *Circuits, Systems, and Signal Processing*, vol. 35, no. 3, pp. 993–1020, 2015.

[12] X. Yuan and J. Huang, “Polynomial-phase signal direction-finding and source-tracking with a single acoustic vector sensor,” in *Proceedings of the International Conference on Acoustics, Speech, and Signal Processing*, pp. 2559–2563, Brisbane, Australia, August 2015.

[13] H. Shao, X. Zhang, and Z. Wang, “Efficient closed-form algorithms for AOA based self-localization of sensor nodes using auxiliary variables,” *IEEE Sensors Journal*, vol. 62, no. 10, pp. 2580–2594, 2014.

[14] H. Chen, B. Liu, P. Huang, J. Liang, and Y. Gu, “Mobility-assisted node localization based on TOA measurements without time synchronization in wireless sensor networks,” *Mobile Networks and Applications*, vol. 17, no. 1, pp. 90–99, 2012.

[15] Q. Liang, B. Zhang, C. Zhao, and Y. Pi, “TDoA for passive localization: underwater versus terrestrial Environment,” *IEEE Transactions on Parallel and Distributed Systems*, vol. 24, no. 10, pp. 2100–2108, 2013.

[16] C. Klungmontri, I. Nilkhamhang, W. Covanich, and T. Isshiki, “Underwater positioning systems for underwater robots using trilateration algorithm,” in *Proceedings of the IEEE Information and Communication Technology for Embedded Systems (IC-ICTES)*, pp. 1–5, Hua-Hin, Thailand, May 2015.

Hindawi
Shock and Vibration
Volume 2020, Article ID 8826802, 10 pages
https://doi.org/10.1155/2020/8826802

Research Article

Data Analysis for Predictive Maintenance of Servo Motors

Oguz Girit,[1] **Gurcan Atakok**,[1] **and Sezgin Ersoy**[2]

[1]*Marmara University, Faculty of Technology, Department of Mechanical Engineering, Istanbul 34722, Turkey*
[2]*Marmara University, Faculty of Technology, Department of Mechatronics Engineering, Istanbul 34722, Turkey*

Correspondence should be addressed to Gurcan Atakok; gatakok@marmara.edu.tr

Received 12 June 2020; Revised 20 July 2020; Accepted 20 August 2020; Published 11 September 2020

Academic Editor: Chaoqun Duan

Vibration and temperature data of a servo motor are analyzed with PLC which is widely used in the industry. With this system, power supply can be detected on the servo motors. In this way, undesirable situations such as disruptions in production and productivity loss can be prevented from occurring. It is an important problem for businesses to detect malfunctions that may occur in servo motor dysfunction. Previously, methods such as ultrasonic sound measurements, thermal cameras, endoscopy equipment, and energy analysis have been used and discussed in the literature. Our study offers a PLC-based vibration and temperature measurement system designed as a solution of this problem. In this system, vibration and temperature measurements were made while the servo motor was kept running. These measurements were measured with or without load, considering the operating ranges of the servo motor, and the compatibility of the data was evaluated.

1. Introduction

One of the biggest problems encountered in the automation sector is the loss suffered in production when the product dysfunctions. In order to prevent the unexpected malfunctions from occurring during the intensive production periods, the maintenance system must be well managed and organized. The implementation of maintenance activities is very important for the smooth operation of machines and work processes that operate continuously [1, 2]. In such cases, it is possible to work with an external company to carry out the relevant maintenance work [3]. Briefly, maintenance activity is a planned and programmed movement that all the partitions in the firm create in an organized manner to maintain the functions of the systems in the most efficient way [4]. Research carried out on the efficiency of maintenance work shows that 33% of the maintenance expenses are unnecessary or wasted due to disruption of their periodic maintenance [5, 6]. While maintenance strategy is created, for this reason, the selection of a business method that suits the maintenance requirements constitutes great importance [7]. Failure can occur while production is intensive. Maintenance and repair to be made during this process can cause high costs. Since the application of this method determines the malfunctions in advance, costs are minimized [6]. Even if the maintenance work is carried out in a comprehensive manner, it may become stereotyped over a long period, and it may not be possible to achieve the desired benefits proportional to the experience of the maintenance technician. Among the types of maintenance, predictive care occupies the last position, with an application rate of 2%. This shows that the predictive care type has been ignored although it provides many benefits for companies because it can foresee and prevent failure before the equipment malfunction occurs [8]. The origin and the development of malfunctions learned from the analyses made with the data obtained can be used in high capacity use of engines and in avoiding shutdowns caused by timeless failures. The predictive care applications are measurement, analysis, and, respectively, repair [9]. The predictive care uses the vibration measurement tool, ultrasonic sound measurement tool, thermal cameras, endoscopy equipment, and energy analyser tools. Benefits of predictive care are the increased life cycle of equipment, estimation of maintenance time, prevention of labour loss, quality, and more efficient use [10–12]. Although the benefits of predictive care are known, it has been determined that the proportion of firms that determine their malfunctions by applying fractional care in the world is 0.04% [13].

The systems used to determine the malfunction of products such as electric motors, generators, and transformers commonly used in the industry measure voltage and current signals to determine the malfunction status of related products [14, 15]. Depending on the operating conditions and the characteristics of the points to be measured using various parameters that characterize the behaviour of systems at runtime to monitor the operating conditions and performances of the machine, the working performance of the machine under various physical quantities can be observed at certain intervals [16]. These measurements are carried out by employing methods. These methods include monitoring actual data, the load set and frequencies, performance curves during offline use, and use of the existing net worth [17, 18]. However, the acquisition and processing of these data are disadvantageous due to the lack of an information boundary and the lack of continuity. System design and management of data traffic are important in the prevention and diagnosis of malfunctions. Writing-based system developments are among the attempted methods [19]. However, these system approaches are generally concentrated on vibration.

For predictive maintenance purposes, thermodynamic and dynamic analysis of a motor that operates between the attachment of the sharp ends of the cylinder, gamma type, free piston, and hot and cold welding temperature can be carried out by detecting the development of a running machine's failure [1] on asynchronous motors [2]. Also, real-time monitoring and evaluation of rotating machines can be done [20]. Today, some methods have been used to determine the friction problems of fixed and rotating parts. One of these methods is friction spectral analysis which is carried out by the measurement of the distribution of the characteristics of reaction. However, the disadvantage of this method is that the friction generating equipment produces noise in the current frequency band. A complete analysis of the temporary response of the rotor-stator interaction in which the friction process is represented by a linear product model (Coulomb friction), and the distribution of the cavity effects due to friction in the spiral vibrations which increases the stability of the system in the rubbing area, has been detected [21, 22].

Our study offers a PLC-based vibration and temperature measurement system designed for the solution of this problem. Vibration and temperature measurements were made while working on the servo motor made according to this system. These measurements were measured with and without load, while considering the operating ranges of the servo motor and while considering the compatibility of working with the graphical data analyses.

2. Materials and Methods

This system is designed to measure the vibration and temperature measurements in servo motors. As seen in Figure 1, the vibration values and temperature values of the surface area of the servo motor on the X and Y axes were controlled by PLC, which is widely used in the industry.

In this study, the automation systems (motor, moving elements, etc.) is used to measure the vibration and temperature values of the products to determine whether the products comply with the standards and to prevent malfunctions. The acceptance of vibration values obtained from the measurements of the machines has been determined with the international ISO 2372 standard. This standard has been used to evaluate the vibration intensity of machines operating between 10–200 Hz [23]. In this study, measurements and evaluations were applied to servo motors between 10–50 Hz by connecting them to any machine. All the harmonic movements that occur in simple harmonic vibrations are repeated periodically. The magnitude of the forces required for the vibration to occur is proportional to the intensity of the vibration [24, 25]. Displacement, velocity, and acceleration are units of amplitude. The unit to be used during the measurement is decided upon the system's work value Hz. The measurements done have been interpreted and evaluated according the ISO 2372 standards.

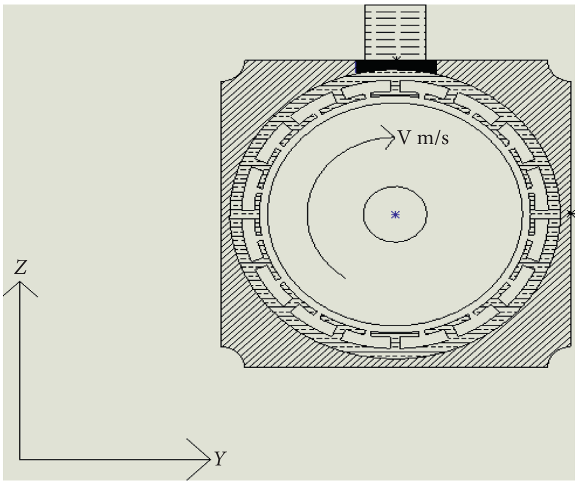


Figure 1: Servo motor working principle.

Temperature is a term connected to a system's molecule's average kinetic energy. It is a base magnitude and a scaler. As the temperature increases, the kinetic energy of the molecules also increases and they move faster; while temperature decreases, the molecules' kinetic energy decreases well and they move slower. If two or more objects are in contact, an energy transfer occurs from the hot objects to the cold objects until there is a thermal equilibrium [26]. Temperature measuring detectors, which are frequently used in industrial environments, are very important, because they determine the temperature range and process conditions made in industrial environments. These measuring systems are generally of low cost semiconductor (PTC-NTC and similar) materials. Today, analog temperature detectors such as NTC and PTC are also used alongside digital temperature detectors.

The operating algorithm of the system is shown in Figure 2. Press start to perform vibration and temperature analysis and select the servo motor model in the recipes. Measurement will start automatically after the recipe is selected. The measurement will be completed when the preset time has elapsed. If the servo motor is operating in accordance with normal operating conditions, the SCADA system will record the data and finish the operation. If the servo motor is not operating in accordance with normal operating conditions, the SCADA system will warn and the servo motor maintenance operation will be required.

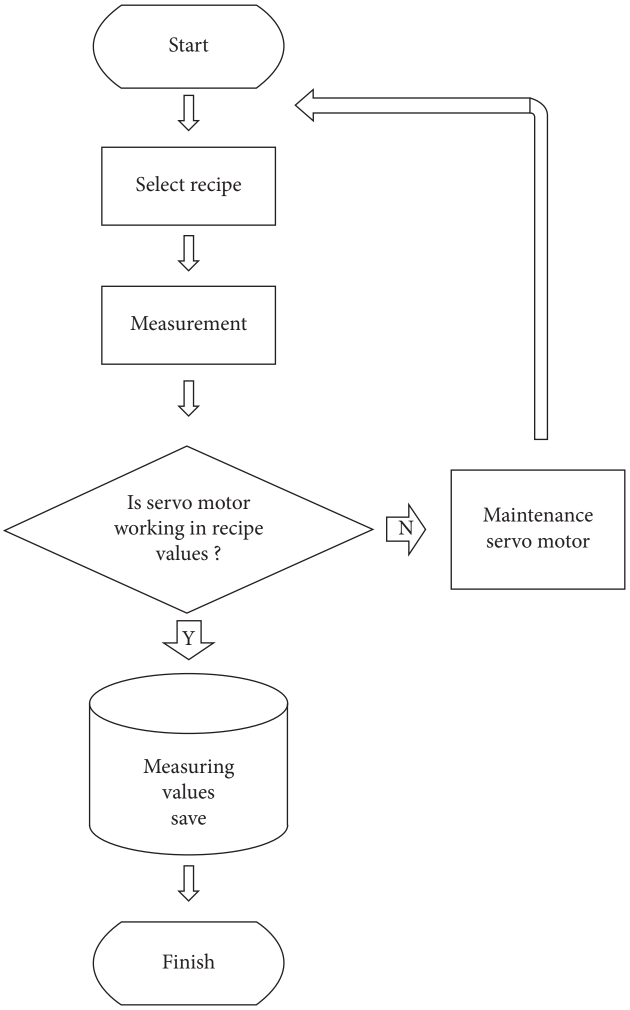


Figure 2: The operating algorithm of the system.

2.1. System Design. The system designed in this work is used to measure the vibration and temperature values of servo motors. When selecting a detector, the following aspects need be paid attention:

The reading sensitivity of the detector.

The minimum and maximum values to be measured.

The sensitivity limit against the highest temperature to be measured.

The reaction speed and the reading accuracy against the change of temperature in unit time.

The continuation time of determination and accuracy.

The restrictions of the environment.

The accuracy level of the application and the change in cost according to the way the detector has been mounted. As there are transducers that measure temperature without contact, temperature detectors usually work by being in contact with the surface that is going to be measured. The temperature detecting equipment consists of thermoelectric temperature elements and resistive temperature elements.

2.2. Vibration and Temperature Measurement. Vibration is expressed in two physical variables which are vibration frequency and intensity of vibration. The frequency of vibration is the number of vibration in a unit of time, and the unit is expressed in hertz (Hz). The severity of vibration is the current strength that occurs in unit of time in the environment where the vibration occurs perpendicular to the energy which comes from the vibration, and the unit of the severity of vibration is (W/cm^2) [27].

Analysis of the data obtained by measurement is important for maintenance and performance use. Accurate analysis of these data creates predictable data for machine failure. For example, the balancing of all forces on piston and rotating machines and the use of special montages decrease the stress. Likewise, the vibration characteristics of the system need to be understood, and the resonance condition analysis needs to be carried out to get an excellent working performance [28]. The goal here is to avoid resonance with the measurement of the vibrations created and the experiments. Moreover, thus decrease releases, many studies have been performed on this. As an example, the deduction below has been made about a ship's vibration in conclusion of the experiments made [29].

Alongside the measurements of vibration, the temperature measurements also carry importance to a large extent. Because it is a parameter that affects various properties and creates a deformation effect on materials, it is essential for the measurements to be done in specific periods to be controlled. Different temperature measurement devices can be designed by taking advantage of their various thermometric features in the measurement of materials' temperatures. Today, there are various temperature measurement devices that depend on the length, pressure, volume, electric resistance, electromotor force inside the electric circuit created by two different wires, and the changes of materials' external heat intensity. These along with devices usually measure by being in contact with the surface to be measured. Besides this method, there are devices used in measuring high temperatures that measure contact-free [16].

In an expertly designed algorithm, the number of transactions must be constant per data instance. Therefore, the total number of operations must be linear depending on N. In general, the processing time required for a collection is much shorter than the processing time required for a multiplication operation. Algorithms can be developed to make these complex operations quick and easy [30]. With this analysis, the signal can be seen in the frequency domain, and the frequency spectrum in the blocks can be calculated and displayed. Real-time data processing is used to calculate time, field signals, and the frequency domain signals obtained from these signals must also be higher than the data acquisition rate.

The control of vibration-temperature during operation and a suitable system for the servo motors to operate have been created. The system is run by the PLC. The sensors have been used to measure the temperature and the vibration values of the devices used in the systems are shown in Figure 3.

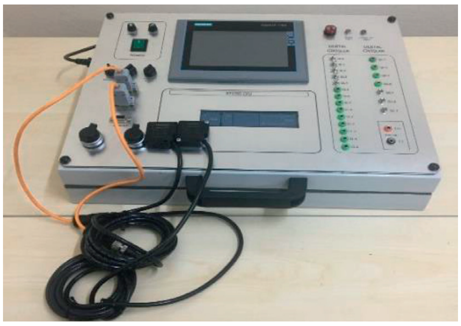


Figure 3: View of the designed system.

The vibration and temperature sensor code used in this process is QM42VT2. The vibration sensors have X- and Z-axis indicators on the surface. When X is parallel to the sensor, the Z axis moves through the sensor on a plane. The X axis is mounted on the same axis as the motor shaft or on it axially.

For best results, the sensor should be installed as close as possible to the motor mount. If this is not possible, the sensor must be mounted on a rigidly connected surface with the vibration characteristics of the motor. Using a surface with a cloth on or any another unstable mounting location to detect specific vibration characteristics can result in a reduced accuracy or capability.

2.3. Properties of the Designed Set. In order to evaluate the measured data in the designed measurement system, a Siemens S71200 PLC, a Siemens Comfort 9″ operator panel with a SCADA system, two vibration and temperature monitors, and a Siemens Brand MODBUS card have been used to measure the vibration and temperature. The measured values have been transferred via the Modbus RTU protocol to the PLC. Five units have been used on the measurement system to control the on/off switch. Besides, a fuse has been added to protect the led system and the high supply voltage from alerting the alarm. The design of the system is shown in Figure 4.

The PLC program in the measuring system has been created by programming the Siemens Tia Portal Professional V14 SP1 software and the SCADA program installed in the operator panel with the Siemens Win CC Comfort V14 SP1 software. Modbus RTU and Ethernet protocol was created to ensure communication between the products used in the system.

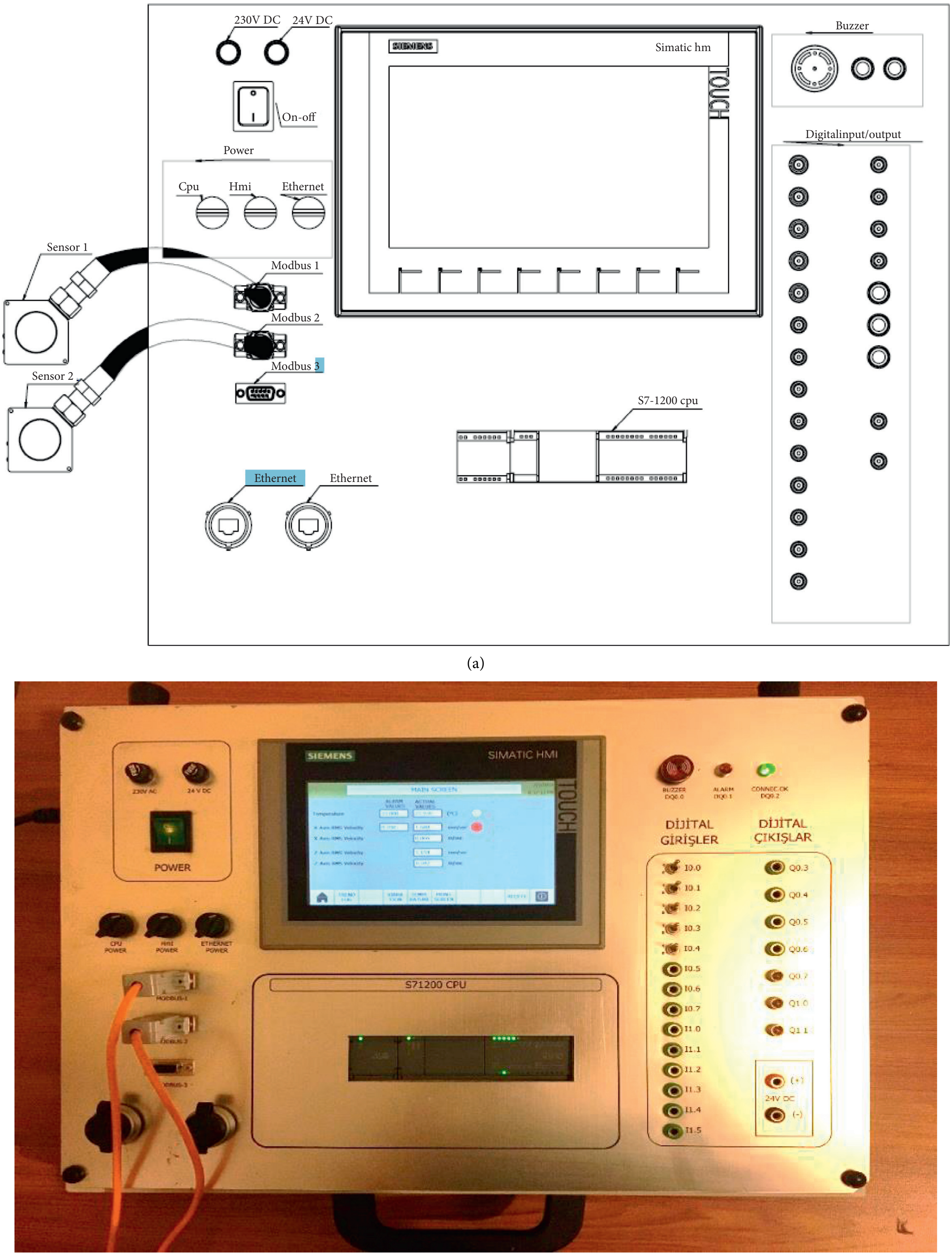
3. Results

Through the system designed in this paper, a servo motor used in the industry has been implemented. The measurements have been made between 0–3000 and 3000–0 RPM. Besides, vibration and temperature analyses have been performed by running the motor with and without load.

3.1. 0 to 3000 RPM Unloaded. The servo motor, with the label information seen, has values increased from 0 RPM to 3000 RPM with no load, and the vibration and temperature values have been recorded for 60 seconds. The vibration data have been taken on the body both in the vertical and in the horizontal axes; the temperature data have also been taken from the body (see the graphs from Figures 5(a)–5(c), respectively). According to the graphic values in Figures 5(a) and 5(b), an increase in vibration speed occurred as the speed value increased. The value of vibration was fixed and it has remained stable after reaching the speed of 3000 RPM. Even though there have been minimal fluctuations in temperature, no considerable change in the temperature value has been observed, as it can be seen in Figure 5(c). The maximum and the minimum vibration values that can be measured were between the values of 1.5 G and −1.5 G. The vibration values have never reached those values. The servo motor has worked with temperature under the maximum measurement of 35 degrees Celcius (Figure 5).

3.2. 3000 RPM to 0 RPM Unloaded. The servo motor was slowed down from 3000 RPM to 0 RPM with no load, and the vibration and temperature values have been recorded for 60 seconds. The vibration data taken from the vertical axis are seen in Figure 6(a), the vibration data taken from the horizontal axis are seen on Figure 6(b), and the temperature data obtained through the body are seen in the chart from Figure 6(c). According to the graphical data of Figures 6(a) and 6(b), it has been served that as the vibration value decreases, the speed value also decreases. The vibration value has remained constant after the speed value has dropped to 0 RPM and then it has been fixed. Even though there have been minimal value changes, the temperature value has decreased before the change in the value. After that, it has increased, as the motor speed has decreased (see Figure 6(c)). In this case, it is acknowledged that the generally known servo motors get warm when they operate at lower speeds. In the measurements made, depending on the measurement direction (X, Y, Z), the operating values of the servo motor are compared with the reference values. From the value of the normal operation of the corresponding servo motor, depending on the measurement direction according to the references, it is observed that it works between the maximum vibration value of 1.5 G and minimum vibration value of −1.5 G. It has been observed that the maximum temperature value of 45 servo motor working referenced to the standard operating value of the respective servo motor works under the maximum value.

3.3. 0 RPM to 3000 RPM Loaded. The servo motor, with the label information seen, is increased from 0 RPM to 3000 RPM with on-load, and the vibration and temperature values have been recorded for 60 seconds. The vibration data have been taken on the body in the vertical axis, as seen in Figure 7(a), data taken on the body in the horizontal axis are illustrated in Figure 7(b), and the temperature data taken on the body in Figure 7 are shown on the graph. According to the graphical data shown in Figures 7(a) and 7(b), when the speed value reaches 500 RPM, it is observed that the vibration values exceed over the limit that can be measured. A decrease in the vibration value has been observed when the speed value of 500 RPM increases to 1500 RPM. While the



(a)

(b)

Figure 4: (a) Design of the vibration and temperature measurement set and its appearance at the moment of energization ((b) alternative).

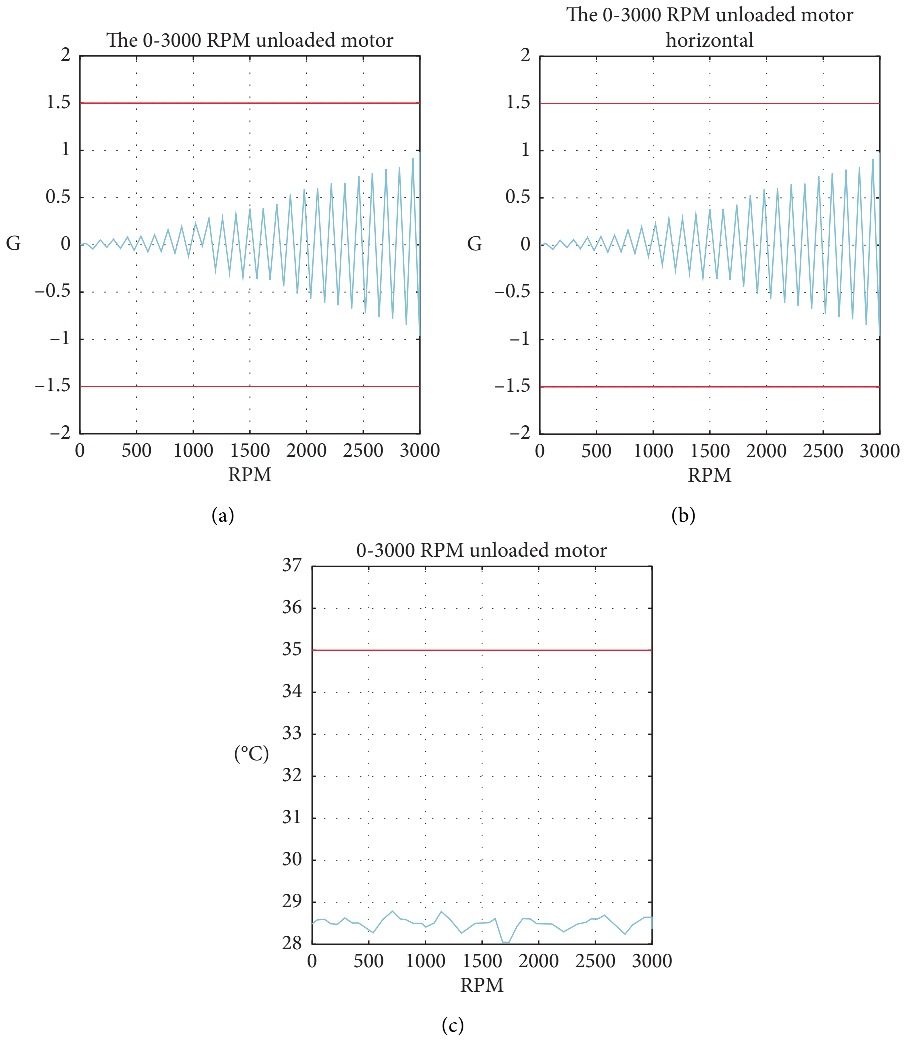


Figure 5: The 0–3000 RPM measurement-speed graph of the unloaded motor: (a) vertical axis, (b) horizontal axis, and (c) temperature-speed graph of the unloaded motor.

speed value has increased from 1500 RPM to 3000 RPM, the vibration value also has increased and exceeded the limit values. As a result of the vibration values measured loaded, it can be concluded that the servo motor does not work properly loaded and that maintenance should be applied. Even though there have been minimal fluctuations in the temperature of the motor, it can be seen that it has been working in an acceptable manner, in accordance with the temperature value, as illustrated in Figure 7(c).

3.4. 3000 RPM to 0 RPM Loaded. The servo motor, with the label information seen, is increased from 0 RPM to 3000 RPM with on-load, and the vibration and temperature values have been recorded for 60 seconds. The vibration data have been taken on the body in the vertical axis, as seen Figure 8(a), data taken on the body in the horizontal axis are illustrated in Figure 8(b), and the temperature data taken from the body are graphically shown in Figure 8(c). A decrease in the value of vibration has been observed in the graphic values in Figures 8(a) and 8(b) until the speed value has dropped from 3000 RPM to 1200 RPM. While the speed value has decreased from 1200 RPM to 700 RPM, an increase in the vibration value could have been observed, and when the speed value has reached 700 RPM, it has exceeded the limit values. While the speed value dropped from 700 RPM to 0 RPM, the vibration value also decreased and became 0 G. When the servo motor vibration values have exceeded the limit values at 700 RPM, it could have been observed that the servo motor needed maintenance. When observing the temperature graph from Figure 8(c), it can be seen that the servo motor works at typical temperature values.

4. Discussion

System performance may be dependent on the effects of vibration and temperature data such as the total operating time of the system, operating conditions of the system, and the external environment. So, the reference value can be obtained from an idle system. By applying the specified load,

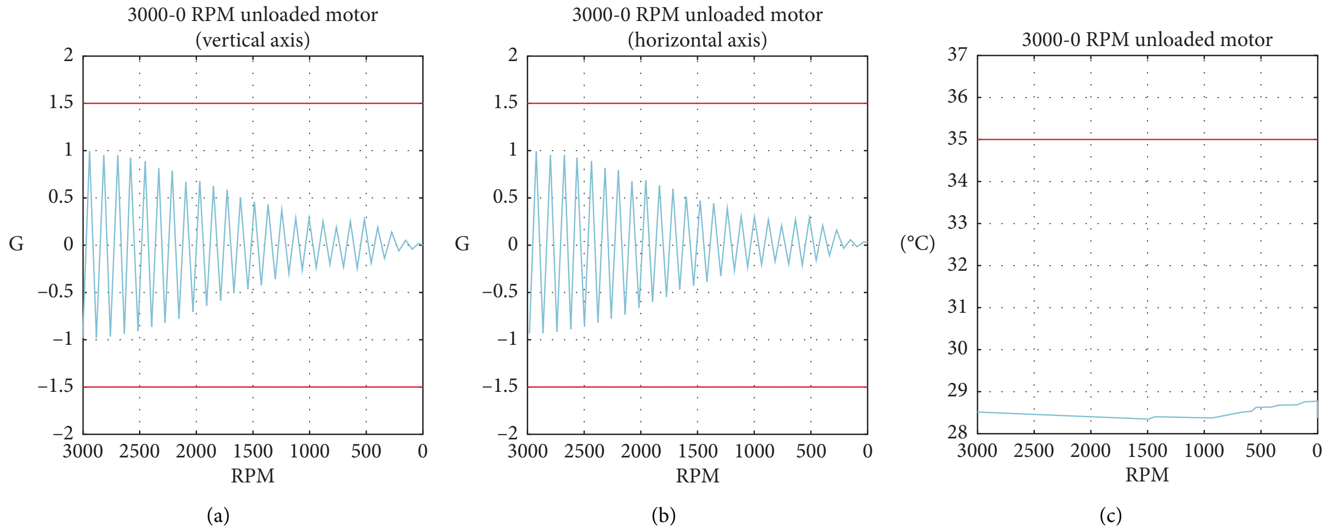


Figure 6: The 3000–0 RPM measurement-speed graph of the unloaded motor: (a) vertical axis, (b) horizontal axis, and (c) temperature-speed graph of the unloaded motor.

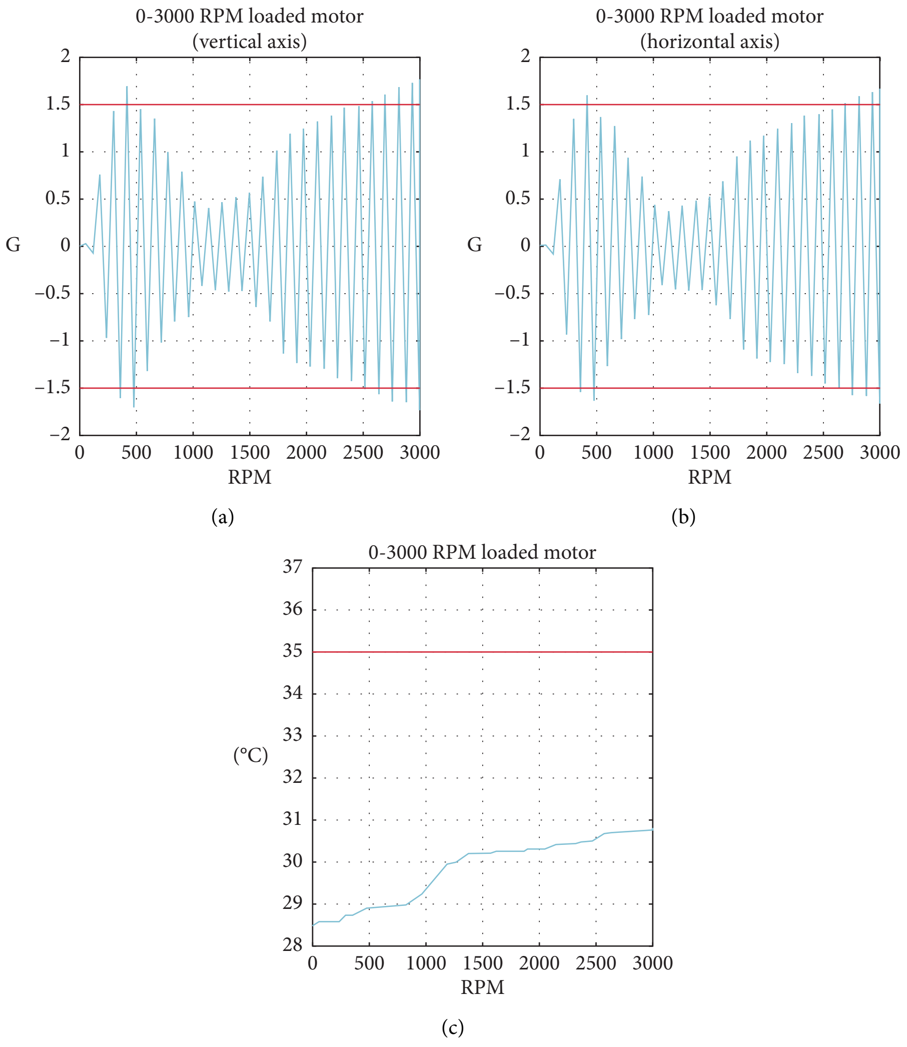


Figure 7: The 0–3000 RPM measurement-speed graph of the loaded motor: (a) vertical axis, (b) horizontal axis, and (c) temperature-speed graph of the loaded motor.

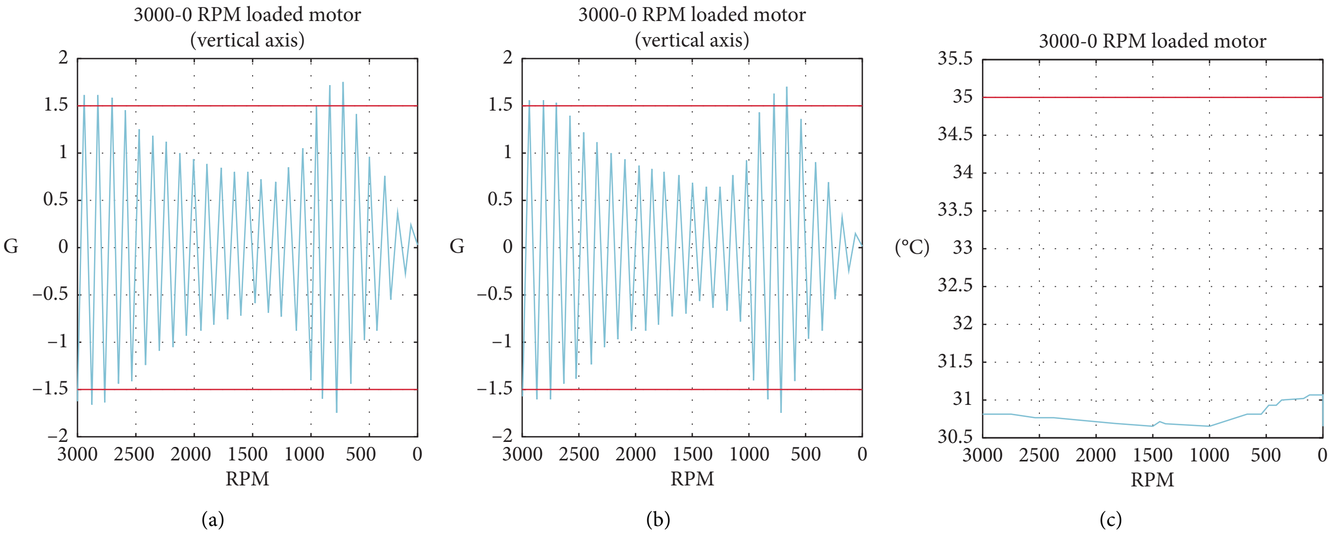


Figure 8: The 3000–0 RPM measurement-speed graph of the loaded motor: (a) vertical axis, (b) horizontal axis, and (c) temperature-speed graph of the loaded motor.

vibration and heat data can be monitored during operation and termination processes. As seen in the results, critical levels can be overcoming under load. This can make the framework require support before due time, and if this circumstance is not resolved in time, there may be a malfunction in the motor. Control used for detecting errors can allow it to give a warning when the limit values are exceeded and to keep values optimal.

Vibration values were measured for Siemens 1FK7063 coded servo motor at 1000, 2000, and 3000 RPM speeds, with horizontal and vertical axis displacement. The value of the temperature in degrees Celsius (°C) has been measured for a minute and it was graphically illustrated. The working time for all the measurements has been set at 60 s.

First, to create the sample system in terms of the reference value, data have been obtained from the system displayed in this article between 0–3000 RPM. The vibration values increase depending on the number of vibration cycles as expected. When the data has been examined, no abnormal value has been observed, and regular operations have been found according to the motor speed. However, when the heat has been measured, the reference value of 35°C has not been exceeded.

To test the validation of the first results of measurement made with 3000–0 RPM measurement cycles, observations were made that the vibration values have decreased according to the amount of vibration it received and it was kept within the working ranges. The temperature value at 500 RPM is within the boundary of the first variable and the temperature value reaches the critical point in the measurement range of 3000–0 RPM. It has been identified as a normal thermal behaviour, considering the running time of this system. The values of the system were recorded while kept idle under monitoring.

A loaded system has operated between 0–3000 RPM. 0–500 RPM and 2200–3000 RPM operating ranges have exceeded the limit of the values of the vibration data. It has been identified that the vibration data have remained at the desired level from this point on. It means that the machine needs to be taken into the maintenance. Otherwise, there is a possibility of causing friction-related malfunction in the system. On the other hand, temperature values were constant at 900–1200 RPM. This variable data remains within the boundary values. The measurements in the system, based on different parameters during the study, could be done in real-time and variable areas have been identified. The data obtained from the regular study are deemed to be inside the working range. When the 0–3000 RPM values of the motor was measured, it was observed that at 3000–2600 RPM that it exceeded over the vibration limits. The rise in these vibration values were seen at 1200–600 RPM. Between the measurements of 0–3000 RPM, the same consequences were not observed in said frequencies. This spike was observed in both measurements made with 0–3000 RPM and 3000–0 RPM and different frequencies in data are observed. Even with these vibration values, the system's temperature has never exceeded over the working temperature of 35 degrees Celsius. In conclusion, better data for predicting malfunction with 3000–0 RPM measurement method have been gained.

5. Conclusions

Disruptions during the mass production processes can cause a competitive disadvantage to companies in the industrial field.

Furthermore, malfunctions can cause major setbacks during the manufacturing process since continuous production depends on how well the machines can perform. Particularly in remotely controlled processes, achieving both timely and accurate monitoring might prove difficult. The system designed and developed in this study provides a way to measure the heat and the vibration in real-time. This innovative system can be configured in accordance with future research regarding industry 4.0. The monitoring system can be used on demand to show measurements for heat and vibrations, remotely and on demand in real-time. The control systems are integrated to provide a safe and nonhazardous production. The system can display real-time data. With the data obtained, a suitable dataset is created for use in artificial intelligence applications. Cost advantages can be provided and the system has a structure that can be improved in all aspects. It can be easily moved anywhere and has a multifunctional structure that can be used for every device and machine of similar structure. The information requested in reporting can be selected by the user. Also, by adopting the understanding of intervention before failure occurs, the highest efficiency can be obtained and the production resulting from a halt or unplanned maintenance caused by the breakdown can be contributed to production by minimizing the cost losses. This system vibration and temperature values can be examined during the operation of the machines without damaging the systems and the data that will cause problems can be monitored. In addition, it can be examined in other motors without damaging machinery and equipment, and how real errors are reflected on the graphics as a result of vibration measurements can be examined.

Data Availability

The data are available on request through a data access committee or institutional review board or from the authors by sending e-mail.

Conflicts of Interest

The authors declare that there are no conflicts of interest regarding the publication of this paper.

References

[1] N. Arthur and J. Penman, "Induction machine condition monitoring with higher order spectra," *IEEE Transaction on Industrial Electronics*, vol. 47, no. 5, pp. 1051–1105, 2000.

[2] H. Arslan, M. Ranjbar, E. Seçgin, and V. Çelik, "Theoretical and experimental investigation of acoustic performance of multi-chamber reactive silencers," *Applied Acoustic*, vol. 157, pp. 543–552, 2020.

[3] N. Johansson, E. Roth, W. Reim, and A. Baraçlı, "Smart and sustainable maintenance: capabilities for digitalization of maintenance," *Sustainability*, vol. 11, no. 13, p. 3353, 2019.

[4] N. Saravanan, V. N. S. K. Siddabattuni, and K. I. Ramachandran, "Fault diagnosis of spur bevel gear box using artificial neural network (ANN), and proximal support vector machine (PSVM)," *Applied Soft Computing*, vol. 10, no. 1, pp. 344–360, 2010.

[5] M. Pophaley and R. K. Vyas, "Plant maintenance management practices in automobile industries: a retrospective and literature review," *Journal of Industrial Engineering and Management*, vol. 3, no. 3, pp. 512–541, 2010.

[6] C. Chengbin, J. M. Proth, and P. Wolff, "Predictive maintenance: the one-unit replacement model," *International Journal of Production Economics*, vol. 54, no. 3, pp. 285–295, 1998.

[7] U. Suadiye, M. Sönmez, and Öztürk, *Control of Three Dimensional Moving System Designed for Wide Field by PLC and Observing by SCADA*, AutomotionSempozium, Denizli, Turkey, 2005.

[8] R. K. Mobley, "An introduction to predictive maintenance," *Plant Engineering*, vol. 54, 2002.

[9] O. Sadettin and N. Aktürk, "Determination of physical faults in gearbox through vibration analysis," *Journal of the Faculty of Engineering and Architecture of Gazi University*, vol. 18, no. 3, 2003.

[10] H. C. Roy, *Principles of Planned Maintenance*, Edward Arnold, London, UK, 1974.

[11] B. P. Lientz and B. Swanson, *Software Maintenance Management*, Addison-Wesley Longman Publishing, Boston, MA, USA, 1980.

[12] A. S. Corder, *Maintenance Management Techniques*, McGraw-Hill, London, UK, 1976.

[13] H. Sun, Y. Zhang, D. Baleanu, W. Chen, and Y. Chen, "A new collection of real world applications of fractional calculus in science and engineering," *Communications in Nonlinear Science and Numerical Simulation*, vol. 64, pp. 213–321, 2018.

[14] A. Duyar and W. Merrill, "Fault diagnosis for the Space Shuttle main engine," *Journal of Guidance, Control, and Dynamics*, vol. 15, no. 2, pp. 384–389, 1992.

[15] J. Litt, M. Kurtkaya, and A. Duyar, "Sensor fault detection and diagnosis for a T700 turboshaft engine," *Journal of Guidance, Control, and Dynamics*, vol. 18, no. 3, pp. 640–642, 1995.

[16] A. Stetco, F. Dinmohammadi, X. Zhao et al., "Machine learning methods for wind turbine condition monitoring: a review," *Renewable Energy*, vol. 133, pp. 620–635, 2019.

[17] C. Li, R.-V. Sánchez, G. Zurita, M. Cerrada, and D. Cabrera, "Fault diagnosis for rotating machinery using vibration measurement deep statistical feature learning," *Sensors*, vol. 16, no. 6, p. 895, 2016.

[18] H. Zhengjia, S. Yudi, and Q. Liangsheng, "Rub failure signature analysis for large rotating machinery," *Mechanical Systems and Signal Processing*, vol. 4, no. 5, pp. 417–424, 1990.

[19] M. Behzad and M. Alvandi, "Unbalance-induced rub between rotor and compliant-segmented stator," *Journal of Sound and Vibration*, vol. 429, pp. 96–129, 2018.

[20] D. W. Childs and L. T. Jordan, *Clearance Effects on Spiral Vibrations Due to Rubbing*, ASME, New York, NY, USA, 1997.

[21] j. Yan and Y. Meng, "Industrial big data in an industry 4.0 environment: challenges, schmes, and application for predictive maintenance," *IEEE Access*, vol. 5, pp. 23484–23491, 2018.

[22] T. Govardhan, C. Achintya, and P. Deepak, "Numerical simulation and vibration analysis of dynamically loaded bearing with defect on rolling element," *International Journal of Acoustic and Vibration*, vol. 23, no. 3, pp. 332–342, 2018.

[23] S. Farhadi, "Acoustic radiation of rotating and non-rotating finite length cylinders," *Journal of Sound and Vibration*, vol. 428, pp. 59–71, 2018.

[24] I. Senjanovic and I. Ancic, "Validation of analytical methods for the estimation of the torsional vibration of ship power transmission system," *Ocean Engineering*, vol. 184, pp. 107–120, 2019.

[25] E. Berber, *Industrial Automatic Temperature Measurement and Control System with Microcontroller*, Yildiz Technical University, Istanbul, Turkey, 2008.

[26] Q. Huang, X. Yan, C. Zhang, and H. Zhu, "Coupled transverse and torsional vibrations of the marine propeller shaft with multiple impact factors," *Ocean Engineering*, vol. 178, pp. 48–58, 2019.

[27] H. Jona, *Implementation of a Real-Time Fast Fourier Transform on a Graphics Processing Unit with Data Streamed from a*

High-Performance Digitizer, Linköping University, Linköping, Sweden, 2015.

[28] M. Mutmaz and A. Revenga, *Design Aspects of Winterized and Arctic LNG Carriers: A Classification Perspective*, American Society of Mechanical Engineers (ASME), London, UK, 2006.

[29] J. I. Taylar, "Back to the basics of the rotating machinery vibration analysis," *Sound and Vibration*, vol. 29, no. 2, pp. 12–16, 1995.

[30] S. W. Winthrop and J. M. Smith, "Handbook of Real-Time Fast Fourier Transforms," IEEE Press, New York, NY, USA, 1995.

Hindawi
Shock and Vibration
Volume 2020, Article ID 8895000, 8 pages
https://doi.org/10.1155/2020/8895000

Research Article

Noise Elimination and Contour Detection Based on Innovative Target Image Contour Coding Algorithm

Siming Meng

Information Engineering Institute, Guangzhou Railway Polytechnic, Guangzhou 510430, China

Correspondence should be addressed to Siming Meng; mengsiming@gtxy.edu.cn

Received 1 July 2020; Revised 25 July 2020; Accepted 28 July 2020; Published 24 August 2020

Academic Editor: Chaoqun Duan

At present, the image mining is mainly based on its local and key features, which focuses on its texture and statistical grayscale features, but it focuses on its edge and shape features rarely. However, the contour is also an important feature for image shape recognition. In this paper, a good target image contour coding algorithm was adopted, and an LCV segmentation model with good image boundary acquisition capability that can reflect the target image contour features was selected for the original image contour segmentation. The detailed features analysis of the contour coding algorithm was carried out through the experiments; the experimental results showed that the algorithm was a significant technological breakthrough in image feature extraction and recognition.

1. Introduction

At present, the image mining is based on its texture and statistical grayscale features mostly. However, during the actual target image recognition process, the image features are not derived from its texture, grayscale, pixel density, and other features alone [1]. In fact, in a specific application field, the image contour is also a key target feature. Therefore, the specific processing of the target contour can bring a new concept in the image recognition field [2–8], especially in the physical image recognition field in a 3D space. There is still an insufficient mining depth in this field at home and abroad currently. Although the feature extraction of the target image contour was proposed in some references [9–13], the secondary conversion of the contour features was still not analyzed in depth; there was still a lack of mature and perfect model algorithms, there was a large error in the retrieval of the contour boundary when the target image contains some noises, and there was a poor discrimination among the similar images in the traditional secondary feature extraction, which resulted in failure to the wide applications of the target contour features in the image recognition field.

At present, most of the traditional image coding algorithms are based on its internal textures and pixels, and more image coding algorithms are applied mainly in the image compression and background prediction. Several traditional image coding algorithms are described as follows:

The entropy coding refers to a nonsemantic data stream compressed mainly with statistical information of data and is a lossless coding. The common entropy codes include *Shannon* code, *Huffman* code, and *arithmetic* code. The entropy code in video coding is a compressed code stream for storage or transmission transformed from the element symbols (representing the video sequence). The input symbols include additional information, header information, motion vectors, and transform coefficients. Several common entropy codes are described briefly as follows:

Shannon coding is a coding method called as *Shannon-Fano* algorithm obtained by *Robert Fano*, a mathematics professor from *Shannon* and MIT based on the information theory proposed by *Shannon* in 1948~1949 and it is a kind of symbol coding with variable lengths. *Shannon-Fano* algorithm is coded from top to bottom as follows: the first is to use the probability of the symbol occurrences as the sequence basis; the second is to divide the symbols into two parts with approximately equal frequencies from top to bottom recursively and mark their boundaries with 0 and 1, respectively; the experimental results 0 and 1 are the binary codes of the target.

Huffman coding is a brand-new coding method proposed by *David Albert Huffman* in the early 1950s. It has top-down feature and is optimal statistically. The shortest code is assigned to the most frequent symbol, and so on. The coding steps are as follows: the first step is to arrange the leaf nodes of the symbols from right to left in the probability sequence; the second step is to connect the two top-row nodes with the lowest probability to get the parent node and mark the two lines of the left and right child nodes with 0 and 1; the third step is to repeat the second step until the root node is obtained to get a binary tree; the fourth step is to obtain the binary coding of the symbols, i.e., leaf nodes 0/1 string of each symbol started from the root node. From the above coding steps, the codes are different but their average length is the same. The specific analysis can be carried out as follows: the detailed probability sequence can be from right to left or from left to right, because the symbols are only related to the probability. There is no effect when the left and right branches are marked with 0 or 1, so the coding result is not specified.

The prototype of *arithmetic* coding was proposed by *P. Elias* in 1960, and the algorithm was pioneered by *J. Rissanen* and *R. Pasco* in 1976, systemized and implemented by *G. G. Langdon* and *Rissanen* in 1979, and corrected by *Rissanen* to obtain a lossless compression algorithm in 1984. Based on the information theory, this coding method and *Huffman* coding are the optimal variable code length types. Its advantage is that it is no longer limited to *Huffman* coding integer bits. For example, if a symbol in *Huffman* coding needs to be represented by 0.1 bits, it can only be represented by 1 bit, which causes a waste of storage space. The symbol in the arithmetic coding is different from those of the above two coding methods. The symbol is indicated with a real interval with a width equal to its occurrence probability within [0,1), and then all the symbols in the symbol table can just fill the entire interval of [0,1), and the input symbol string (data stream) is mapped to a real value in the interval of [0,1).

The predictive coding is based on the feature that there is a certain correlation among the discrete signals to predict the unknown signals and then code the prediction errors. Obviously, the coding accuracy is closely related to the prediction error. If the predicted value is accurate, the error will be small. When the coding accuracy requirements are relatively similar, this method can also be used to compress the data. It has the features of strong error diffusion, simplified and fast algorithm, and easy hardware implementation.

Since the image data and sound data are sampled, the predictive coding method is more suitable for them, the differences between adjacent values are not obvious, and too many bits are not required. The standard predictive coding diagram is shown in Figure 1, and the corresponding coding steps are as follows:

Step 1: calculate the difference between $f(i, j)$ and the prediction value $f\prime(i, j)$ generated by the predictor at the transmitting end to obtain the prediction error $e(i, j)$;

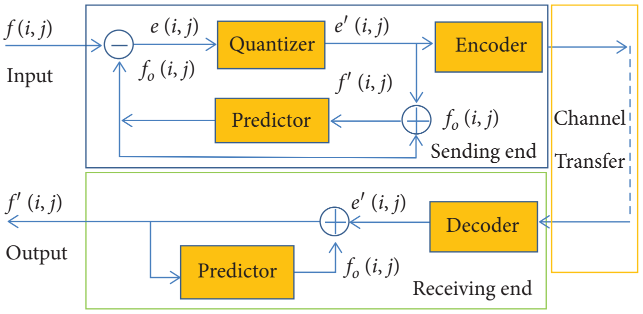


Figure 1: Flow chart of the predictive coding steps.

Step 2: $e(i, j)$ is quantized by the quantizer to $e\prime(i, j)$, and a quantization error is generated;

Step 3: $e'(i, j)$ is encoded by the encoder into a code word for sending, and $f'(i, j)$ is added into $e'(i, j)$ to restore the input signal $f'(i, j)$. Due to the existence of quantization error, $f(i, j)$ is not equal to $f'(i, j)$, but they are very close. The local decoder at the sending end is the predictor and its loop;

Step 4: set the memory in the predictor at the sending end in advance to store $f'(i, j)$ for prediction of the next pixel and then input the second pixel; repeat the abovementioned operation.

In 1968, the transform coding was proposed firstly by *Pratt* by applying Fourier transform, followed by the oblique transform, *Walsh* transform, the discrete cosine transform (DCT), and *k-l* transform. The essence of transform coding is to reduce the spatial correlation of the image signals at the frequency domain level, and its digital rate reduction effect is similar to that of the predictive coding. Since 1980, a hybrid coding scheme was emerged gradually, which combined with the transform coding and the motion compensation, to promote a great progress in the digital video coding technologies. In the early 1990s, the famous video coding proposal for videoconferencing applications with the hybrid coding scheme was proposed firstly by ITU. Afterwards, with the continuous improvement of the video coding standards and recommendations, the hybrid coding technology was slowly developed and stabilized and became a digital video coding technology with a relatively high application frequency. The transform coding is an indirect coding method and is generated by the orthogonal transformation of image information. Because the correlations among signals are omitted in the orthogonal transformation, the redundancy of the signals can be reduced, and the coding method is relatively easy. In this way, the coding of the image is transformed to the coding of transform coefficients. Because the amount of data is not large, and the parameters are not related to each other, a more compression ratio can be obtained. *Haar* transform, *Walsh Hadama* transform, discrete *Fourier* transform, and discrete cosine transform are all quasioptimal transforms, and the last one is used most frequently.

Thus, a good target image contour coding algorithm was adopted, and the feasibility analysis and related experimental analysis of the algorithm were carried out in this paper. The target image contour coding algorithm is a model algorithm based on the coding transformation of the contour feature information after extracting the contour features of the target image. Compared with other traditional coding algorithms, the proposed algorithm has a unique role in the secondary conversion of image contour features, which can extract the image boundary contour information from multiple perspectives. The detailed featuresanalysis of the contour coding algorithm was carried out through the experiments; the experimental results showed that the algorithm was a significant technological breakthrough in image feature extraction and recognition.

2. Target Image Contour Coding Algorithm

From the above traditional coding algorithms, it can be seen that the traditional image coding is mainly used in the compression and storage of images. In a certain sense, the image attribute features are not mined, that is, the image contour attribute features are not extracted, while the data quantity of the digital image is reduced effectively only under the premise that the saved image information is not missed as much as possible. Therefore, a transformation model of target contour quadratic feature codes was adopted and the characteristics and stability of the algorithm were analyzed through relevant verification experiments in this paper.

The target image contour coding algorithm is used mainly to obtain the target image boundary coordinate information, perform a certain image conversion, and then extract the target image contour information matrix. The model is different from the traditional coding algorithms in coding role, coding speed, and coding principle, and it has a unique role in the secondary conversion of image contour feature extraction and can be used to extract the image boundary contour information from multiple aspects. The coding algorithm steps are as follows:

Step 1: obtain the target image boundary information from the image analysis model, represented as (x, y) coordinates;

Step 2: normalize the coordinates of the contour boundary obtained in Step 1 with the interval [1,255], respectively, to eliminate the effect of the difference in absolute phase pixels of the image;

Step 3:convert the contour matrix data obtained in Step 2 into uint8 format, for subsequent calculation and processing;

Step 4: carry out the grayscale process for the contour matrix obtained in Step 3, and convert it to obtain the image boundary coding matrix;

Step 5: carry out the binarization of the image boundary coding matrix obtained in Step 4 by setting the corresponding threshold;

Step 6: in order to combine the image boundary coding matrix with other model algorithms, arrange and encode the binary coding matrix obtained in Step 5 in row, column, and oblique directions; finally, draw the corresponding coding chain, among which the coding in row direction is converted with each row of the binarized coding matrix into a row in order, the coding in column direction is converted with each column of the coding matrix into a row in order, and the coding in oblique direction is converted with each diagonal of the coding matrix from upper right to lower left into a row in order.

3. Feasibility Analysis of Target Image Contour Coding Algorithm

From the above coding model algorithm principle, it can be seen that the most critical feature of the target image contour coding algorithm adopted in this paper is the image contour boundary, that is, as long as the target image has a certain contour boundary, which can be obtained with a certain image processing method, the target features can be extracted to obtain the contour feature information of the image. It is worth noting that since the actual object is 3D, when it is converted into an image, only the projection information at one dimension can be obtained, and such an image contour cannot be correctly reflected on a 2D image. Therefore, the images with 3D features were not studied and only the typical images with 2D features were studied and analyzed in this paper.

At present, for most images, whether they are obtained from static or dynamic objects, the specific contours can be generated on the images. As long as the effective target object boundary contour is obtained from a reasonable model, the model algorithm adopted in this paper can be executed; at the same time, because it is transformed into a 1D 01 coding sequence during the subsequent process of the model algorithm, this algorithm can be directly integrated with other models and there is no integration problem in the subsequent calculation process. Therefore, the model is applicable actually.

4. Experimental Analysis of Target Image Contour Coding Algorithm

The target image contour coding algorithm adopted in this paper can be used to extract the secondary features of the target image contour in a targeted manner and can convert them into the key image information. The coding experiments of the contour features extracted with the row set model Local Chart-Vest (LCV) [1] segmentation were carried out to analyze the features of the target image contour coding algorithm as follows.

All experiments in this paper were completed under the experimental conditions of Intel I7-4712HQ 2.30 GHz CPU, NVIDIA 610M graphics card, 8 GB memory, and MATLAB R2011b.

Because the LCV model has better image boundary acquisition capability under uneven image grayscale and brightness conditions, and the extracted target image contour boundary is smooth and can reflect the detailed contour features of the target, the target image contour coding algorithm adopted in this paper was based on the image contour boundary extracted from the LCV model; at the same time, the algorithm was analyzed with actual images and artificial images in the experiments.

4.1. Analysis of Image Contour Coding Algorithms in Different Ways. The ultimate objective of the image coding adopted in this paper was to extract the features of the image. For different types of pictures, different coding chains can be obtained with different coding methods. The codes in different directions were analyzed as coding in row direction, coding in column direction, and coding in oblique direction. During the experimental process, the uniform contour boundary was set as [20,50], which could be modified with A1 = imresize(A1,[20 50]) program, and the binarized threshold coefficient was 0.9, which could be modified with A1 = im2bw(A1,0.9) program. The coding experiments with the global pixel points of the image and the contour boundary of the image from the row set model were carried out; the experimental results are shown in Figure 2.

From the coding results in Figure 2, the following conclusions were obtained.

In Figure 2(c), the global pixels of the image were coded simply; in the experimental resulted coding chain, there were more white value segments, which were small and thin and distributed unevenly. Such coding chain had a very low degree of interpretation for the features of the image, because the small and thin white value segments obtained by coding would disappear, resulting in a weak memory for the interpretation of the image when there was a slight change in a part of the image.

In Figure 2(d), the image boundary contour was extracted from the row set model and encoded in the row direction. The coding results showed that there were less white value segments in the coding chain, which were longer and more obvious. Therefore, the experimental result obtained by the coding had a high degree of interpretation for the features of the image. When the image is changed locally, the relative position and length of the coding chain would still not be changed significantly. Therefore, the coding in row direction has higher anti-interference ability in the image recognition.

In Figure 2(e), the coding chain result of the contour boundary in column direction showed that the white value segments were narrower and thinner and distributed more uniformly; in other words, the coding has better anti-interference ability, because only some white value segments disappeared in the coding chain when the image was changed locally, and there was only a small impact on the image features. Therefore, the coding in column direction still has an anti-interference ability in the image recognition.

In Figure 2(f), the coding chain result of the contour boundary in oblique direction showed that the white value segments were narrower and thinner and distributed more uniformly and the coding also had a certain anti-interference ability. Therefore, the coding in oblique direction can reflect the overall image features largely and has still an anti-interference ability in the image recognition.

From the experimental results with the coding methods in above directions, it can be summarized as follows. In the extraction of global pixels, the image was compressed only while the original image information was unchanged; when they were arranged in row direction, there were fewer, narrower, and unevenly distributed white value segments in the coding chain obtained with global pixel features. Therefore, these white value segments are greatly affected by the local changes of the image and are lost easily under the presence of other interference factors and the coding chain is not suitable for characterizing the image features. In the coding in row direction, there were wider and fewer white value segments. Therefore, it can reflect the single-phase image features obviously, has a higher anti-interference ability, and is suitable for single-phase image recognition. In the coding in column and oblique directions, there were thinner and more and evenly distributed white value segments. Therefore, they are suitable for the multiphase image recognition.

4.2. Analysis of Coding Algorithms for Contour Attributes with Different Parameters. From the coding algorithm flow, it can be seen that the core parameters that affect the image contour boundary coding include the binarized threshold coefficient and the image coding matrix size. Therefore, different image features can be extracted with different core parameters. In the actual applications, in order to extract the coding chain that reflects the image features, the relatively good coding chain parameters should be provided.

4.2.1. Effect of Threshold Coefficient. The coding experiments of the image contour boundary extracted from the row set model in different directions were shown as follows, in which the binarized threshold coefficient was set to 0.7. The coding result is shown in Figure 3.

From the coding results in different directions in Figure 3, it can be seen that the number and width of the white value segments of the image contour boundary encoded in all directions were increased when the binarized threshold coefficient is decreased during the coding process.

In the coding result in row direction, there were wider white value segments. However, compared with the coding result with the binarized threshold coefficient of 0.9, it can be seen that the white value segments of the original coding did not disappear, but only the width and number of the coding chain were increased on the original ones. Therefore, for special images, when the coding is not significant, the image contour coding chain can be enhanced by decreasing the binarized threshold coefficient.

In the coding in column and oblique directions, when the binarized threshold coefficient was decreased during the coding process, the number of white value segments of the coding chain was also increased, but the original white value segment positions were not changed. Therefore, the degree of discrimination of the image coding chain can also be

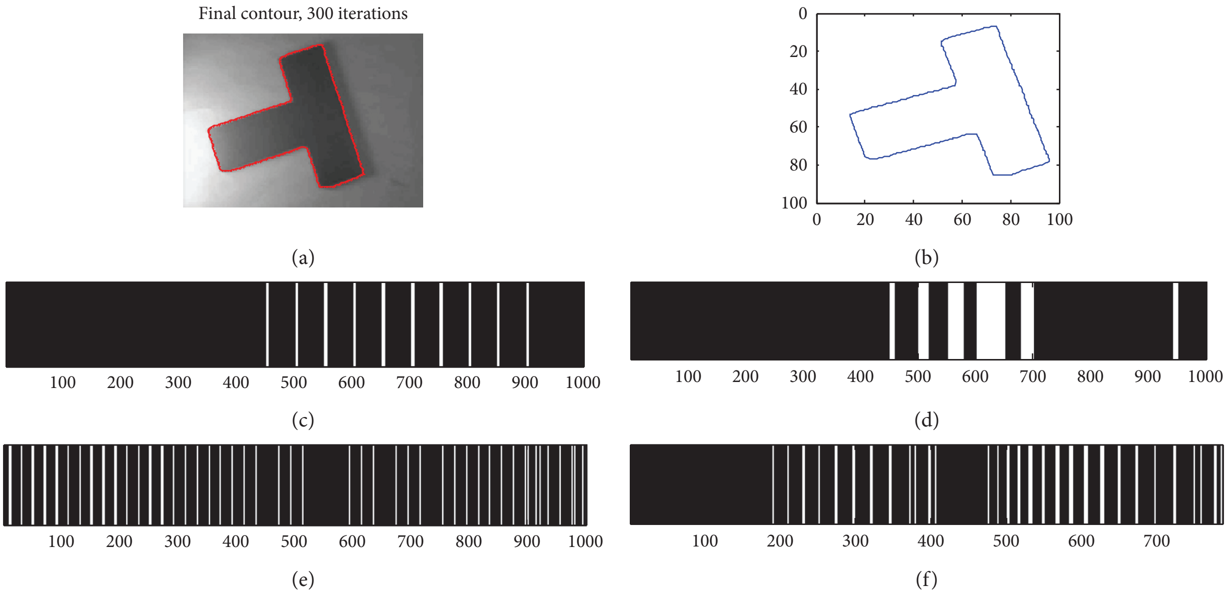


Figure 2: Result of global image pixels and image contour feature coding chain. (a) Row set LCV model image segmentation result. (b) Image contour. (c) Coding chain result of the global pixels in row direction. (d) Coding result of the contour features in row direction. (e) Coding result of the contour features in column direction. (f) Coding result of the contour features in oblique direction.

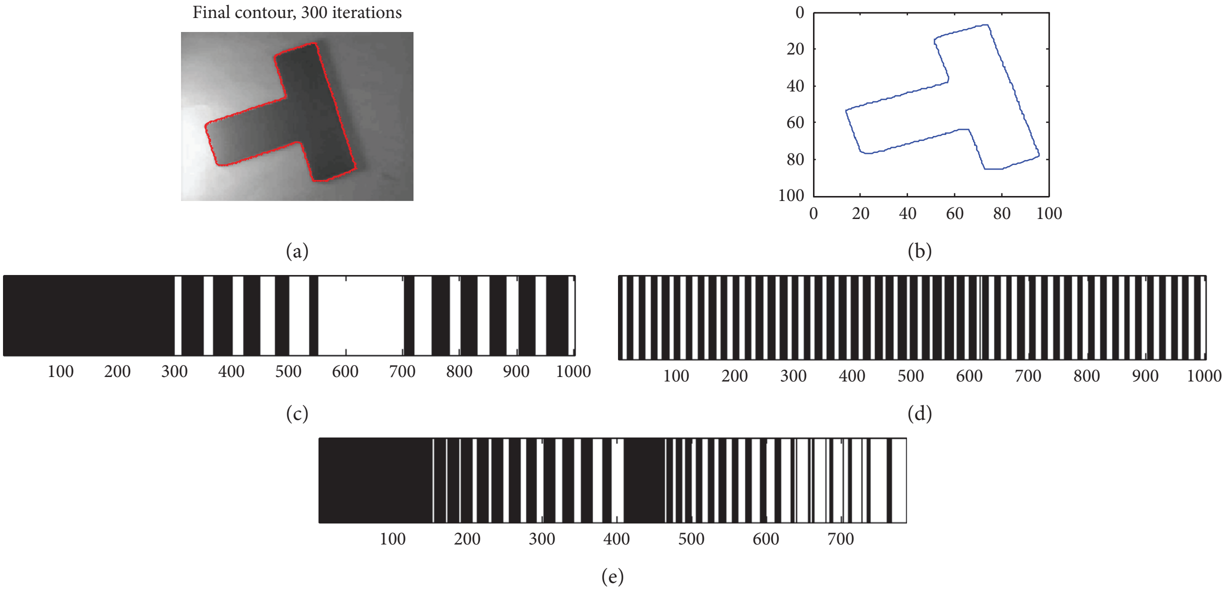


Figure 3: Image coding results with the binarized threshold coefficient of 0.7. (a) Row set LCV model image segmentation result. (b) Image contour. (c) Coding chain result of the contour boundary in row direction. (d) Coding chain result of the contour boundary in column direction. (e) Coding chain result of the contour boundary in oblique direction.

enhanced by decreasing the binarized threshold coefficient, that is, the similarity between the coding chains of two different images can be reduced and the difference between image features can be enhanced.

4.2.2. Effect of Coding Matrix. The coding experiments of the image contour boundary extracted from the row set model in different directions were shown as follows, in which the image coding matrix sizes were 10^*50 and 20^*50 uniformly, and the threshold coefficient was 0.9. The coding results in row and column directions are shown in Figure 4.

From the abovementioned experimental results, it can be seen that when the matrix of the coding process was reduced, the number and width of the white value segments of the image contour boundary coding in all directions would be

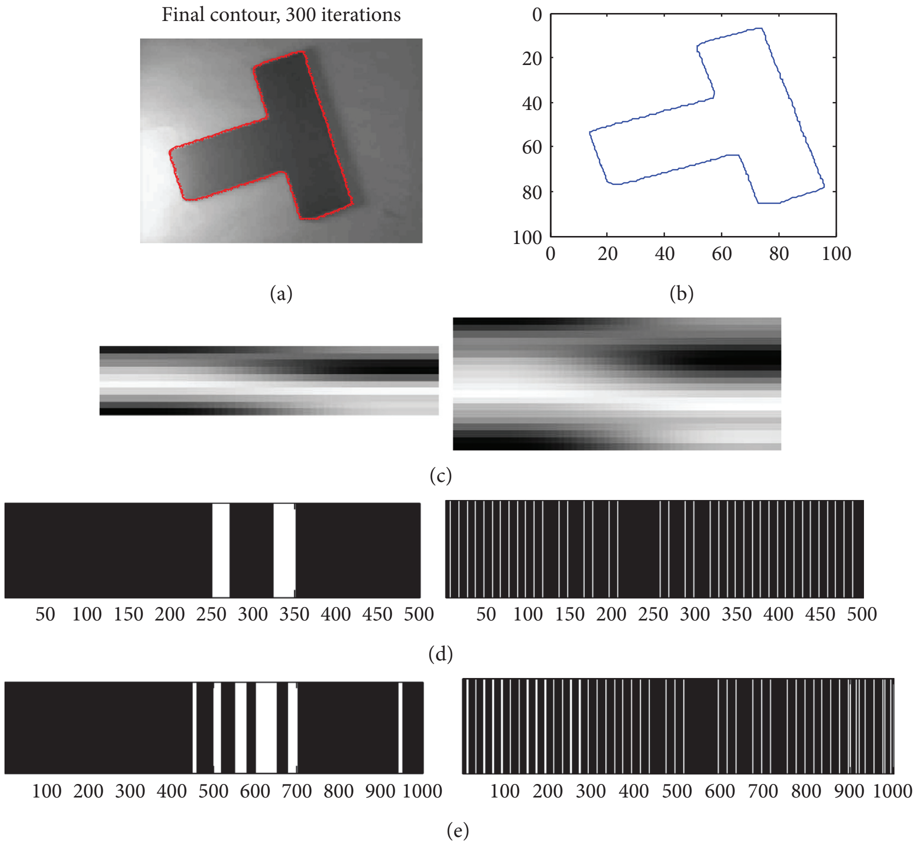


Figure 4: Image boundary feature coding results of different coding matrixes. (a) Row set LCV model image segmentation result. (b) Image contour. (c) Coding matrixes with sizes of 10∗50 and 20∗50. (d) Coding chain results of 10∗50 coding matrix in row and column directions. (e) Coding chain results of 20∗50 coding matrix in row and column directions.

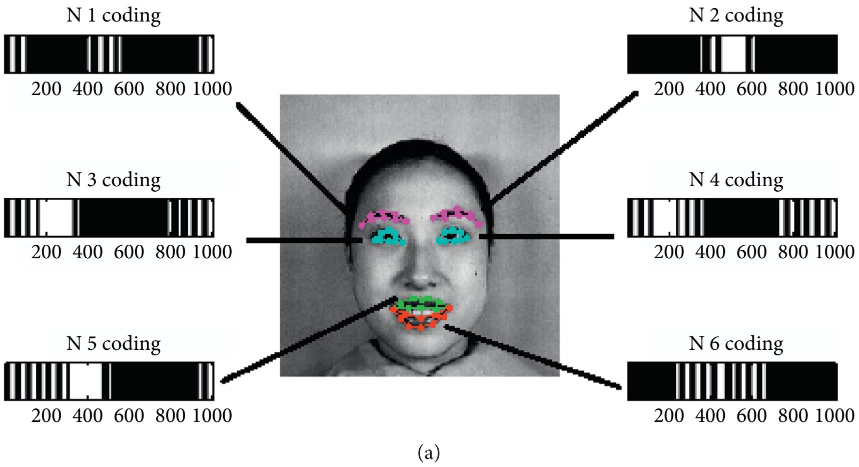
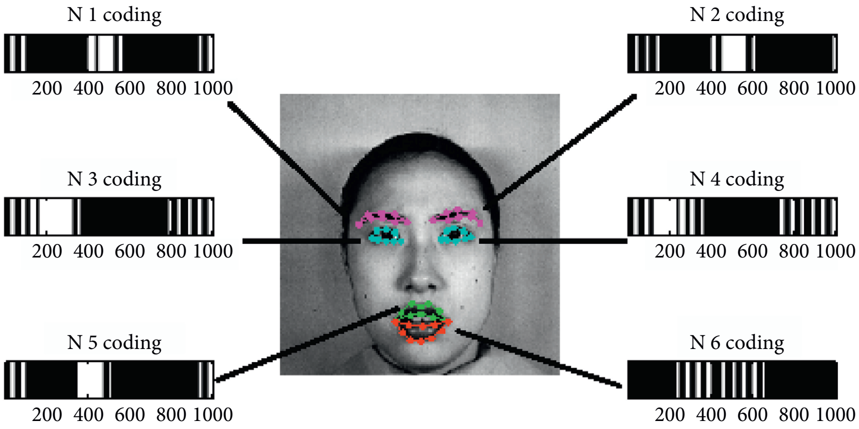


Figure 5: Continued.



(b)

Figure 5: Image facial expression feature extraction and the coding in row direction results in JAFFE. (a) Happy expressions. (b) Angry expressions.

decreased, and the relative positions of the white value segments would be changed. In the coding results in row and column directions, the relatively clear white value segments could be obtained, but the number of white value segments was only a half of that of 20*50 coding matrix. Therefore, when there are many types of images in the image recognition, the matrix size of the coding process can be reduced appropriately, so that the coding chain can accommodate more sample codes for distinguishing.

4.3. Experimental Analysis of Target Contour Coding Face Expression Recognition Based on Active Appearance Models (AAM) [14]. In this paper, the contours of eyebrows, eyes, and mouth, which had a great relationship with facial expressions, were extracted by AAM to obtain the contours of each part (N1–N6), and then the algorithm was used to convert and encode the contours to obtain the feature values of each part. The contours of each part were extracted from the AAM, and the algorithm was used to convert and encode the values of each part. The coding matrix size was 20*50, the threshold coefficient was 0.8, and the coding mode was coded in row directions. The obtained facial feature parts and coding results in row directions are shown in Figure 5. Finally, the comprehensive judgment result of facial expressions was given by the specific expression characteristics of each part in the face, and the expression recognition rate in the face database JAFFE reached 98.78%.

5. Conclusion

The contour is the important information for image recognition. The relevant researches showed that, for any object under specific conditions, the contour is a main factor to distinguish different objects; even for extremely similar objects, the local difference in the target contour is quite significant. For example, for the leaves under one tree, although their shapes are very similar, if all the leaves are unified, the features of their contours also have obvious differences. Therefore, compared with other commonly used image recognition algorithms, such as KL recognition algorithm, texture-based recognition algorithm, model-based recognition algorithm, and geometric feature-based recognition algorithm, only the image boundary shape information is required for the image contour-based recognition algorithm and different target images have different contours. Therefore, a transformation model with the secondary feature coding of the image boundary contour information from multiple aspects, the target image contour coding algorithm, was adopted in this paper.

From the above coding experiments with global pixels of the image and the coding experiments with the image boundary contours in all directions based on the row set energy, it can be seen that, the coding features obtained from the coding process with the global pixels of the image were weak, could not reflect the features of the image fully, and might be lost under external interference; therefore, the anti-interference ability was poor. Thought the experiments with actual images and artificialimages, it could be seen that, for the coding with image boundary features, the coding methods in row, column, and oblique directions could better reflect the image features. It was undoubtedly an effective experiment in image feature extraction and recognition.

Data Availability

The data that support the findings of this study are available from the corresponding author.

Conflicts of Interest

The author declares no conflicts of interest.

Acknowledgments

The research was supported by Guangdong General Colleges and Universities (Natural Science) Special Innovation Projects of China (no. 2018GKTSCX060).

References

[1] X. Wang, *Row Set Method and its Applications in Image Segmentation*, University of Science and Technology of China, Hefei, China, 2009.

[2] S. Aich, M. Yamazaki, Y. Taniguchi, and I. Stavness, "Multiscale weight sharing network for image recognition," *Pattern Recognition Letters*, vol. 131, pp. 348–354, 2020.

[3] L. A. d. Souza, A. N. Marana, and S. A. T. Weber, "Automatic frontal sinus recognition in computed tomography images for person identification," *Forensic Science International*, vol. 286, pp. 252–264, 2018.

[4] K. Mainzer, S. Killinger, R. McKenna, and W. Fichtner, "Assessment of rooftop photovoltaic potentials at the urban level using publicly available geodata and image recognition techniques," *Solar Energy*, vol. 155, pp. 561–573, 2017.

[5] A. Akula, A. K. Shah, and R. Ghosh, "Deep learning approach for human action recognition in infrared images," *Cognitive Systems Research*, vol. 50, pp. 146–154, 2018.

[6] M.-K. Tsai, "Automatically determining accidental falls in field surveying: a case study of integrating accelerometer determination and image recognition," *Safety Science*, vol. 66, pp. 19–26, 2014.

[7] G. Hu, C. Yin, M. Wan, Y. Zhang, and Y. Fang, "Recognition of diseased pinus trees in UAV images using deep learning and AdaBoost classifier," *Biosystems Engineering*, vol. 194, pp. 138–151, 2020.

[8] S.-B. Chen, Y.-L. Xu, C. H. Q. Ding, and B. Luo, "A non-negative locally linear KNN model for image recognition," *Pattern Recognition*, vol. 83, pp. 78–90, 2018.

[9] A. Relja and M. S. Tevfik, "Sketch recognition by fusion of temporal and image-based features," *Pattern Recognition*, vol. 44, pp. 1225–1234, 2011.

[10] T. Huynh-The, C.-H. Hua, T.-T. Ngo, and D.-S. Kim, "Image representation of pose-transition feature for 3D skeleton-based action recognition," *Information Sciences*, vol. 513, pp. 112–126, 2020.

[11] A. V. Savchenko, "Sequential three-way decisions in multi-category image recognition with deep features based on distance factor," *Information Sciences*, vol. 489, pp. 18–36, 2019.

[12] Š. Radoslav, B. Ivan, and Š. Júlia, "Object recognition in clutter color images using herarchical temporal memory combined with salient-region detection," *Neurocomputing*, vol. 307, pp. 172–183, 2018.

[13] K. Lee and K. Min, "An interactive image clipping system using hand motion recognition," *Information Systems*, vol. 48, pp. 296–300, 2015.

[14] T. Cootes, G. Edwards, and C. Taylor, "Comparing active shape models with active appearance models," in *Procedings of the British Machine Vision Conference 1999*, pp. 173–182, Nottingham, UK, September 1999.

Hindawi
Shock and Vibration
Volume 2020, Article ID 8873353, 7 pages
https://doi.org/10.1155/2020/8873353

Research Article

Precision Improvement for the Detection of TGC via RBF Network

Yue Yan [1,2]

[1] *Chongqing Engineering Laboratory for Detection, Control and Integrated System, Chongqing Technology and Business University, Chongqing 400067, China*
[2] *College of Artificial Intelligence, Chongqing Technology and Business University, Chongqing 400067, China*

Correspondence should be addressed to Yue Yan; 1998042@email.ctbu.edu.cn

Received 21 June 2020; Revised 11 July 2020; Accepted 13 July 2020; Published 31 July 2020

Academic Editor: Chaoqun Duan

The approach of multipoint measurement, with increasing hardware cost, should no longer be adopted against the problem of low detection precision on the quality and concentration measurement of large-caliber or irregular pipeline gas with the single platinum film probe. Alternatively, the data correction and improvement can be carried out through establishing an RBF model to detect sample gas after preprocessing. Furthermore, the computer simulation and error analysis can be implemented by taking actual SO_2 data emitted by one medium-sized coal-fired power plant in China as a training sample. Hence, it can be shown that this approach on improvement and analysis of continuous monitoring of the systematic integrated error against the instrument correction and flue gas emission has feasibility, and the comprehensive average error is less than 0.6%.

1. Introduction

The nondispersive infrared gas concentration analyser based on the principle of platinum film is mainly applied to gas mass flow and concentration detection. It has a larger measurement range, a lower loss of pressure, and a relatively smaller size, and it has no consumption parts compared to electrochemical analysis instruments and spectrographic analysis instruments. Furthermore, it is easy to match with an integrated circuit. Therefore, it is suitable for extremely low gas monitoring and control, and it can be widely applied in the area of industrial gas testing. However, in the actual working environment, there are many difficulties in how to improve the gas concentration detection accuracy of large diameter or irregular pipeline. Taking the test of SO_2 vehicle emissions discharged by the thermal power plant as an example, the problems to be studied and solved are mainly reflected in the following aspects.

For the construction difficulty, according to GB/T 16157-1996 [1], “the sampling aperture of circular flue shall be located at mutually perpendicular diametrical lines in various measuring points.” Meanwhile, “for chimney flue with the diameter which is greater than 4 m, homalographic cylinder number is 5, and the sampling number of test point is 10–20.” Apart from the method of equal annular area, the common test point setting also includes the Chebyshev method and the log-linear method. In fact, the diameter of the exhausting chimney in large- and medium-scaled thermal power plants in China is currently above 8 m, and the height is 180–210 m. The test usually trepans on the chimney directly, close to the actual discharge, and the height of the sample gas gathering always exceeds 90 m. Hence, detection precision is increased through setting multiple detection points on equal ring sections, and this has large constructional difficulties and high equipment costs.

For the detection principle, the realization of mutual interference of multiple groups of gases in the processing environment will cause the loss or degeneration of principal component features, but this is unavoidable in the actual working environment.

For the actual process flow, at present, the detection technology does not fully consider the impact of coal quality on the detection accuracy. On the one hand, since the production areas and quality of coal are different, the content of SO_2 in the actual off-gas after combustion (before treatment and purification) is approximately 20000 mg/m^3 to ~2000 mg/m^3, and it has a large fluctuation range and is very unstable. The actual working status of the exhaust gas purifying process equipment is greatly influenced by the condition of the hardware and the external environment. In consequence, the concentration fluctuation range of SO_2 gas entering into the chimney after purification treatment is still large; on the other hand, usually, the water content of exhaust gas discharged by the thermal power plant is 13%–15%. However, the continuous moisture absorption peak results in cross-interference of the absorption peak of multicomponent gas (zero moisture cannot be ensured in the actual working environment).

For the detection technology, the current detection technology does not consider the influence of the pretreatment system on detection accuracy. According to the emission standard of SO_2 in thermal power plants and other key regions is 35 mg/m^3 (standard state) [2]. This refers to the trace amount of SO_2 emitted into the atmosphere by thermal power plants after purification treatment. In fact, the off-gas discharged into the chimney by the thermal power plant is characterized by high temperatures (120°C to 50°C), high moisture content (13% to 15%), and high dust content (about 10 μm), and it contains corrosive gases and so on. In the actual working environment, the off-gas shall not be detected in the sensor directly after sampling. Meanwhile, the actual operating procedure of the whole process in the thermal power plant is not completely stable due to various factors such as coal product quality, burning temperature, and other procedure controls for achieving combustion quality of the product. As a consequence, the pressure, tar, benzene, naphthalene, moisture, fine particle, temperature, and gas flow rate of sample gases are different. Therefore, the preprocessing of sample gas shall be carried out to ensure that the typical sample gas can be obtained in the shortest retardation time. On the condition that the concentration of the tested gas is not lost, the state (temperature, pressure, flow, cleanliness, etc.) of the sample gas shall be suitable for the operating conditions required by the sensor. Hence, the representation and authenticity of the sample gas sent to the sensor after being processed by the sample processing system have a crucial influence on the ultimate detection accuracy. At the design stage, the atmospheric pressure, wind direction, geographic position, local climate, extreme climate, and other geographic information shall be taken into account in the CEMS (Continuous Emission Monitoring System) of the thermal power plant. These are specialized designs. Hence, it is difficult to accomplish quantitative error analysis assessment of the pretreatment system of the sample gas because the deviation caused by the pretreatment system of the sample gas cannot be estimated and revised with a mathematical model.

In addition, the current detection methods do not consider the interference of detection instruments and pretreatment system itself on the detection accuracy; for example, the platinum film probe has noise generated by the signal processing circuit and other random errors. It also exhibits temperature effect on zero and temperature drift of sensitivity, calibrated error, linearization error, error of signal processing circuit, and measuring error as a result of the temperature of the sample gas. The problems also including voltage fluctuation of the light source of trace amount and aged light of optical glasses still exist.

Therefore, SO_2 concentration detection interference in the thermal power plant mainly implies that the output value of concentration P is not only decided by one target parameter (the absorbed infrared energy, e); it is a multivariate function that is related to nontarget parameters, for instance, flow of sample gas (fr), noise of conditioning circuit (n), temperature (t), water content (w), calibrated error (c), various linear error (l), organic matter (o) (such as tar, benzene, and naphthalene), and others; namely,

$$P = f(fr, n, t, w, c, l, o, \ldots). \tag{1}$$

In recent years, neural networks have many applications in sensor signal processing, nonlinear correction, temperature compensation, and so on. BP (back propagation) neural network model has been brought into infrared temperature and humidity compensation [3, 4]. RBF (Radial Basis Function) neural network is applied to precision motion system [5] and neural network is used in the pressure analysis [6] and gas concentration measurement [7] in industrial environment; and a new method of Correction of Dynamic Errors of a Gas Sensor Based on Neural Network has been presented [8], etc. These studies adequately demonstrate that data fusion technology of sensor network on basis of neural network has effectiveness and super application prospect. However, there still exists great research space on multigroup gas analysis, comprehensive analysis, and processing of interference factors, normalization of gas sample, product engineering, and other directions under the constraints of hardware cost and construction condition.

The concentration measurement of large-caliber or irregular pipeline gas with the single platinum film probe has low-level detection precision, hardware cost shall be controlled, and constructional difficulties shall be reduced. Therefore, the method of combining the proceeding control of sample gas and data correction of neural network is adopted on the basis of existing analysers rather than simply enhancing the minimum range of the instrument. Further, test and effect analysis are also carried out through computer simulation in this study.

2. Materials and Methods

2.1. Description and Problem Formulation. According to the principle of consistency approximation of neural network, if it can ensure that the number of neurons in the hidden layer of surface web is plentiful enough or the deep network is deep enough, a surface web or deep network that approximates a nonlinear mapping at arbitrary precision will be found [9]. In consequence, the problems including analysis leakage with interference factor and mathematical modelling compensated by interference factor can be avoided through bringing neural network into error analysis and correction of thermal power plant based on the above analysis. The experimental model is shown in Figure 1. In this model, the

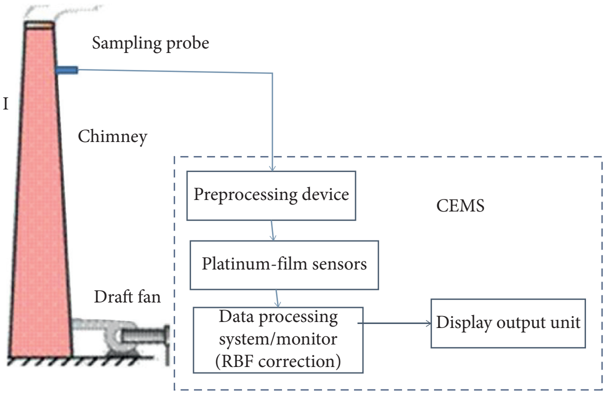


Figure 1: Schematic diagram of the experimental model.

preprocessed sample gas is sent to the platinum film sensor for concentration detection, and the data after detection are output after error correction of neural network.

The RBF neural network is selected to promote error performance by utilizing the characteristics of high convergence rate, strong self-learning ability, and easy achievement [10, 11]. Therefore, the generalization performance of the neural network will be enhanced, and misguidance of nonstandard training sample sets will be avoided. In the RBF network, a large number of neurons and other improved algorithms (genetic algorithm, PSO, etc.) can be utilized to further approximate system dynamics or precision [12]. However, this will definitely cause more complicated calculated load. The RBF network structure shall be simplified as far as possible under the premise of ensuring error precision, and the appropriate mean square error (MSE) shall be determined through experiment. Thus, the hardware-based design will be realized at a later period, and retardation time will also be shortened.

RBF network training data are actual data from the actual working environment. Numerous interference and custom attributes about gas concentration detection of the aforesaid thermal power plant shall be taken into full consideration rather than compensation of signal values detected by the sensor through adopting the function model calculation method and parameter calibration calculation method. Therefore, error improvement can be aimed at the performance of the whole CEMS.

2.2. Experiment and Experimental Data Specifications

2.2.1. Training Process Instructions. The network training process can be briefly described as follows. A single set of PA200 analysers provided by Chongqing Chuanyi Analytical Instrument Co., Ltd., shall be installed at the exhausting chimney of a large domestic thermal power plant (including the pretreatment device, and the mounting height is 90 m). The main relevant performance parameter indicators are linear error $\leq \pm 1\%$ FS (full range) and the minimum range of SO_2 of 0 to 0.01% (100 ppm). The network training is carried out by regarding 100 groups of actual test data from 2015 as the sample of RBF network training. Moreover, the method of computer simulation can be adopted to simulate 20 groups of experimental data under an equivalent process environment and for testing, evaluation, and analyses.

2.2.2. Training and Testing Sample Instructions. Under normal conditions, the temperature of the gas released from the thermal power plant after coal combustion is about 150°C, where the gas arrives at the exhausting chimney after processing through denitration, dust removal, desulfuration, and other environmental protection processes. The pretreatment system shall be set in the site environment to protect the sensitivity of the instrument. The systematic object is shown in Figure 2, and the structure is shown in Figure 3. The complete set is utilized with an instrument that can control the fluctuation range of temperature, dust particles, water content, and flow before the sample gas enters into the analysers. This practice also has the properties of high stability and low technique implementation cost.

According to the aforesaid interference factor analysis, water content, temperature, dust particles, flow, and output voltage are selected as sample input vectors of the neural network. The sensor that is applied to measure the corresponding interference factor is not installed in the field due to technique implementation cost. Through the pretreatment system, the sample gas has the following properties: temperature of 3°C–5°C, dust particle content of 0 to 0.5 μm, flow of 40 L/H–60 L/H, and water content <0.8. Hence, the value of the relevant zone can be determined by adopting a random generation mode of computer simulation with the fluctuation range.

In order to verify the trained RBF network characteristics, 20 groups of experimental data are generated in consideration of the multidimensional features of the sample data and with the method of two-dimensional interpolation, which is common in computer simulation as test samples for testing. Meanwhile, these data shall conform to the actual working environment of the thermal power plant and the equivalent interval (the range of water volume, temperature, dust particle, and flow is the input range of the pretreatment device for the sample gas, and the sensor output voltage is 300–800 mv). Furthermore, the mean square error, the absolute error, and the relative error can be determined on a performance analysis basis.

3. Simulations and Results

3.1. Training and Analysis of RBF. The basic starting point of this study is to simplify follow-up hardware design and control the cost. In consequence, the simplified RBF structure shall be selected.

An output layer and input vector are the above-mentioned 5×80 training sample data, a hidden layer, and output layer. Data normalization processing shall not be implemented. The adaptive adjustment of the threshold value shall not be carried out. The limiting number of neurons is 200. The training function is newrb (), of which the selection of the regression factor shall abide by the following algorithm.

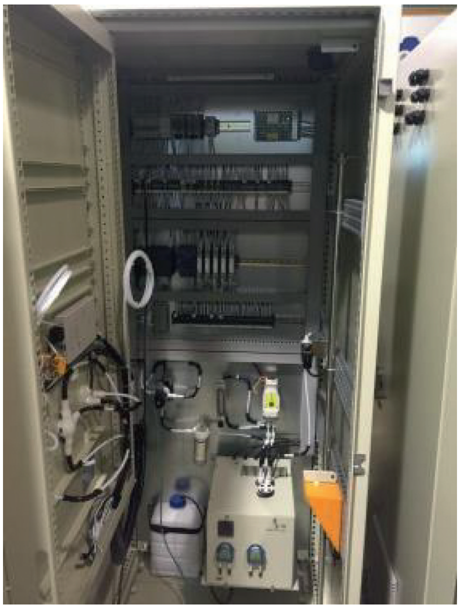


Figure 2: Photograph of the preprocessing device.

For the orthogonality between $i \neq j$, m, and w_j, the energy of $y\ (t)$ is

$$y^T y = \sum g_i^2 w_i^T w_i + E^T E. \tag{2}$$

After eliminating its mean value, y refers to the vector of desired output, and the variance of $y\ (t)$ is

$$N^{-1} y^T y = N^{-1} \sum g_i^2 w_i^T + N^{-1} E^T E. \tag{3}$$

It is found that $\sum g_i^2 w_i^T$ desired output variance can be explained with the regression factor, so the compression ratio of error generated by w can be defined:

$$[\mathrm{err}] = \frac{g_i^2 w^T w_i}{\left(y^T y\right)}, \quad 1 \leq i \leq M. \tag{4}$$

In terms of several optional regression factors, each regression factor has a corresponding compression ratio of error. The maximum compression ratio of error can be selected, and its corresponding regression factor is the final selected regression factor.

The unit of error margin is MSE (mean square error); relative error is defined as follows: (actual output of network–desired output)/desired output. When the MSE is set as $1e-3$, $1e-5$, $1e-10$, $1e-12$, and $1e-15$ and the simulation is carried out on the same computer, its time consumption, number of iterations, average relative error, and maximum relative error are as shown in Table 1.

3.2. Training and Analysis of BP

3.2.1. Using the Same Sample Training Data, the BP Neural Network Is Constructed for Comparison. The increasing number of hidden layers of BP neural network has no positive correlation with the improvement of system performance [13, 14]. To ensure hardware design as much as possible, there is only one hidden layer of the network. Trainlm function is the error return training function. Adopt batch training mode. The training of Levenberg-Marquardt direction propagation algorithm is selected.

To avoid the local minimum of BP neural network caused by training samples, the Adaboost.m1 algorithm based on boosting idea is used to concentrate on the weak learners with low prediction accuracy to improve accuracy.

When the number of neurons is 30, the MSE (mean squared error) value is as shown in Figure 4.

Besides, the best epochs value is 4, and the recognition accuracy (with the error range of 0.05 as the accurate recognition) is 92.5%.

Furthermore, 60-sample data are randomly selected from the same test set for testing, and the prediction error analysis is shown in Table 2.

It can be seen that the average relative error of BP neural network prediction is still less than 1%, but its overall performance is worse than the RBF network.

3.3. Further Analysis of RBF Network. According to the experimental data, the RBF network on data fitting of SO_2 gas shows superior performance and rate of convergence. Its average relative error precision is higher than the requirements of the existing 1%. When the precision requirements are relatively high (MSE decreases), the performance period of the RBF network will improve substantially, while the whole performance approaches a promising outcome.

When the error tolerance of MSE is $1e-15$, the network training process is as shown in Figure 5. When the iteration reaches up to 79 times, an evident MSE jump emerges. In consequence, the average relative error and maximum relative error increase. This is because the input sample vector components of the neural network are different types of data. Further combined with Table 1, under the premise of a certain number of training samples, the prediction accuracy of RBF network with the same structure will greatly increase the hardware cost, but the detection accuracy will not necessarily increase with the improvement of single error tolerance requirements and instability will even appear. From the perspective of overall performance, when MSE is $1e-5$, time consumption and error performance of the network are more appropriate for hardware design of the trace gas flow analyser.

Residual comparison of the test sample in the same group under the above five setting conditions is shown in Figure 6. It shows that when MSE is $1e-3$, error variation has a large interval range; when MSE is $1e-5$, the range of error variation reduces further. However, when MSE is $1e-10$, $1e-12$, and $1e-15$, the range of error variation is very close. Meanwhile, the residual values of error at a large number of test points almost coincide, and the residual error at a large number of test points is extremely low. The interval of error variation does not decrease and is even higher than that of the MSE of $1e-5$. Therefore, the MSE in the RBF network design should not be set too high. On the other hand, if the precision requirements are too high, it will be adverse to the hardware design, cost control, and stability of hardware from the perspective of hardware design.

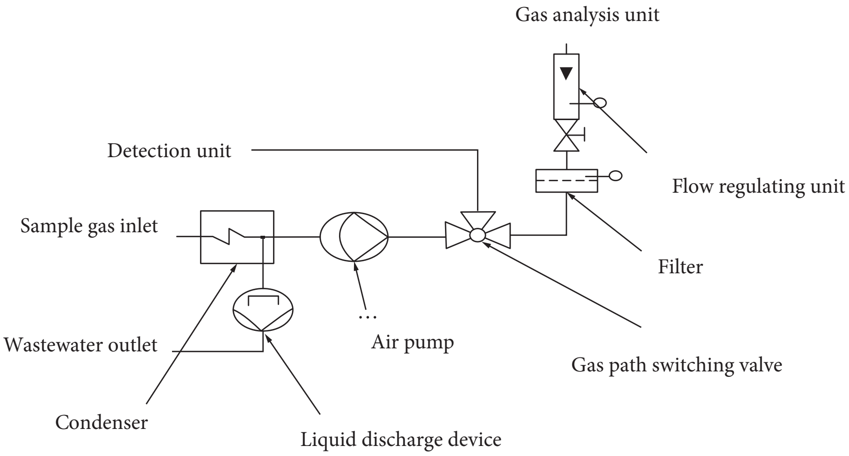


Figure 3: Structural diagram of the pretreatment device.

Table 1: Contrastive analysis under different MSE.

MSE	Number of iterations	Time consumption[1]	Average relative error[2]	Maximum relative error[3]
1e − 3	29	6.6157	0.647	1.975
1e − 5	43	9.389719	0.205	0.603
1e − 10	80	15.741	0.074	0.939
1e − 12	80	15.825	0.068	0.274
1e − 15	80	15.885	0.200	2.225

[1] T refers to unit time on the same computer. [2] Average relative error is the average value of the relative error of all test data. [3] Maximum relative error is the maximum value of the relative error of all test data.

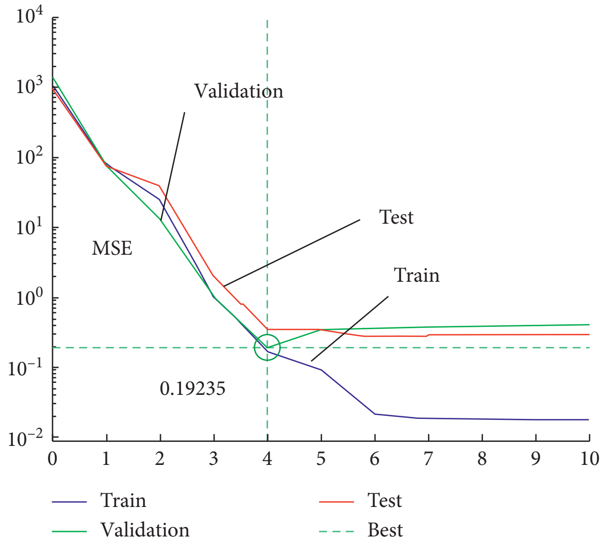
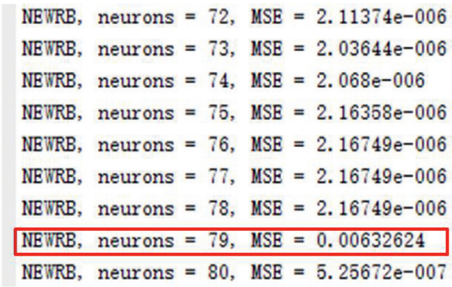


Figure 4: MSE of BP neural network.

Table 2: Prediction error analysis of BP.

Maximum absolute error (mg/m^3)	Minimum absolute error (mg/m^3)	Average absolute error (mg/m^3)	Maximum relative error (%)	Minimum absolute error (%)	Average relative error (%)
2.2493	0.2847	0.565	3.4	0.4358	0.8651



```
NEWRB, neurons = 72, MSE = 2.11374e-006
NEWRB, neurons = 73, MSE = 2.03644e-006
NEWRB, neurons = 74, MSE = 2.068e-006
NEWRB, neurons = 75, MSE = 2.16358e-006
NEWRB, neurons = 76, MSE = 2.16749e-006
NEWRB, neurons = 77, MSE = 2.16749e-006
NEWRB, neurons = 78, MSE = 2.16749e-006
NEWRB, neurons = 79, MSE = 0.00632624
NEWRB, neurons = 80, MSE = 5.25672e-007
```

Figure 5: Network training process.

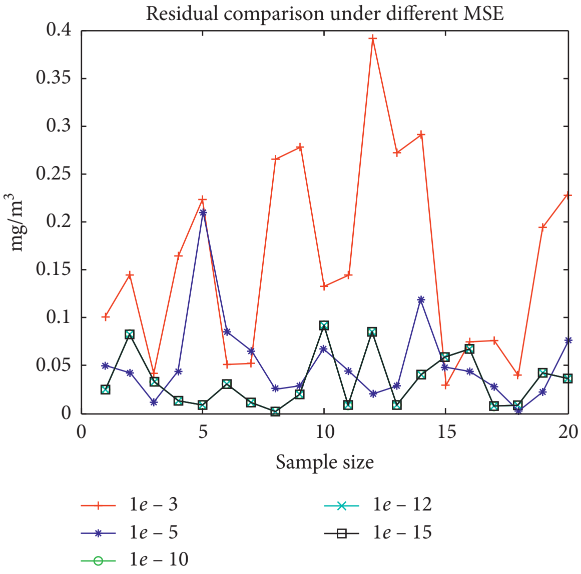


Figure 6: Residual comparison under different MSE.

4. Conclusions

From the perspective of multiple factors influencing concentration measurement of trace gas, an RBF neural network was designed in this study by regarding SO_2 concentration measurement of the thermal power plant as an example. Meanwhile, the training and simulation were carried out by adopting actual engineering environmental data to analyse and reveal its effectiveness. Small sample data acquisition and reasonable learning mechanism setting are the key to improvement. Because of the custom characteristics of the CEMS system, it is more scientific to collect sample data regularly from the site as a training sample. But how to determine the validity of the system model still needs further study. This method provides improvement ideas for quality and concentration measurement of large-caliber or irregular pipeline gas with the single platinum film probe. Using the writing characteristics of a single chip microcomputer, it can be reprogrammed on the basis of the existing hardware so as to improve the detection accuracy and automatic calibration of an analytical instrument in an actual working environment, and it allows performance improvement of the overall CEMS system.

Data Availability

The data used to support the findings of this study are available from the corresponding author.

Conflicts of Interest

The author declares that there are no conflicts of interest.

Acknowledgments

This study was supported by the research project of Chongqing Municipal Education Commission (Grant no. KJ1500639) and the research project of Chongqing Engineering Laboratory for Detection, Control and Integrated System (Grant no. KFJJ2018097).

References

[1] "National standard industry information service network ," 2017, http://www.freebz.net.

[2] National Development and Reform Commission, Beijing, China, 2017, http://www.sdpc.gov.cn/gzdt/201409/t20140919_626240.html.

[3] H. Wang, W. Zhang, L. You, G. Yuan, Y. Zhao, and Z. Jiang, "Back propagation neural network model for temperature and humidity compensation of a non dispersive infrared methane sensor," *Instrumentation Science & Technology*, vol. 41, no. 6, pp. 608–618, 2013.

[4] S. Ding and X. H. Chang, "Application of improved BP neural networks based on LM algorithm in characteristic curve fitting of fiber-optic micro-bend sensor," *Advanced Materials Research*, vol. 889–890, pp. 825–828, 2014.

[5] R. Yang, P. V. Er, Z. Wang, and K. K. Tan, "An RBF neural network approach towards precision motion system with selective sensor fusion," *Neurocomputing*, vol. 199, pp. 31–39, 2016.

[6] M. Dabiri, M. Ghafouri, H. R. R. Raftar, and T. Björk, "Neural network-based assessment of the stress concentration factor in a T-welded joint," *Journal of Constructional Steel Research*, vol. 128, pp. 567–578, 2017.

[7] S. Mohammadzadeh, H. Zargari, and M. A. Ghayyem, "The application of intelligent computation (artificial neural network-ANN) prediction of sweet gas concentration in a gas absorption column," *Energy Sources, Part A: Recovery, Utilization, and Environmental Effects*, vol. 37, no. 5, pp. 485–493, 2015.

[8] J. Roj, "Correction of dynamic errors of a gas sensor based on a parametric method and a neural network technique," *Sensors*, vol. 16, no. 8, p. 1267, 2016.

[9] G.-B. Huang, L. Chen, and C.-K. Siew, "Universal approximation using incremental constructive feedforward networks with random hidden nodes," *IEEE Transactions on Neural Networks*, vol. 17, no. 4, pp. 879–892, 2006.

[10] J. Nedbalek, "RBF neural networks for function approximation in dynamic modelling," *Journal of Konbin*, vol. 8, no. 1, pp. 223–232, 2008.

[11] W. Jia, D. Zhao, and L. Ding, "An optimized RBF neural network algorithm based on partial least squares and genetic algorithm for classification of small sample," *Applied Soft Computing*, vol. 48, pp. 373–384, 2016.

[12] H. Mirinejad and T. Inanc, “An RBF collocation method for solving optimal control problems,” *Robotics and Autonomous Systems*, vol. 87, pp. 219–225, 2017.

[13] C. Dong, L. Dong, and M. Yang, “The application of the BP neural network in the nonlinear optimization,” *Advancesin Intelligent and Soft Computing*, Springer, vol. 78, pp. 727–732, , Berlin, Germany, 2011.

[14] N. Izeboudjen, A. Bouridane, A. Farah, and H. Bessalah, “Application ofdesign reuse to artificial neural networks: case study ofthe back propagation algorithm,” *Neural Computing and Applications*, vol. 21, no. 7, pp. 1–14, 2011.