Research Article

Convolution-LSTM-Based Mechanical Hard Disk Failure Prediction by Sensing S.M.A.R.T. Indicators

Junjie Shi, Jing Du, Yingwen Ren, Boyu Li, Jinwei Zou, and Anyi Zhang

1State Grid Information & Telecommunication Branch, Beijing, China
2Beijing Fibrlink Communications Co., Ltd, Beijing, China
3State Key Laboratory of Networking and Switching Technology, Beijing University of Posts and Telecommunications, Xitucheng Road, Beijing, China

Correspondence should be addressed to Anyi Zhang; zhanganyi@bupt.edu.cn

Received 28 July 2022; Accepted 16 September 2022; Published 4 October 2022

Copyright © 2022 Junjie Shi et al. This is an open access article distributed under the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

The traditional Infrastructure as a Service (IaaS) cloud platform tends to realize high data availability by introducing dedicated storage devices. However, this heterogeneous architecture has high maintenance cost and might reduce the performance of virtual machines. In homogeneous IaaS cloud platform, servers in the platform would uniformly provide computing resources and storage resources, which effectively solve the above problems, although corresponding mechanisms need to be introduced to improve data availability. Efficient storage resource availability management is one of the key methods to improve data availability. As mechanical hard disk is the main way to realize data storage in IaaS cloud platform at present, timely and accurate prediction of mechanical hard disk failure and active data backup and migration before mechanical hard disk failure would effectively improve the data availability of IaaS cloud platform. In this paper, we propose an improved algorithm for early warning of mechanical hard disk failures. We first use the Relief feature selection algorithm to perform parameter selection. Then, we use the zero-sum game idea of Generative Adversarial Networks (GAN) to generate fewer category samples to achieve a balance of sample data. Finally, an improved Long Short-Term Memory (LSTM) model called Convolution-LSTM (C-LSTM) is used to complete accurate detection of hard disk failures and achieve fault warning. We evaluate several models using precision, recall, and Area Under Curve (AUC) value, and extensive experiments show that our proposed algorithm outperforms other algorithms for mechanical hard disk warning.

1. Introduction

At present, Infrastructure as a Service (IaaS) cloud platforms have become the main solution to provide enterprise IT infrastructure. With the development of big data technology and application, more and more enterprises begin to realize the importance of data, so they put forward higher requirements for the availability of data. The traditional IaaS cloud platform generally introduces dedicated storage devices into the platform to achieve high availability of data storage and provides virtual machines in cooperation with dedicated computing devices in Figure 1 [1]. This heterogeneous architecture often leads to two problems: first, it makes the heterogeneity of platform hardware more significant and increase the operation and maintenance cost and scalability cost of the platform; second, when computing resources and storage resources come from different devices, the connection between the computing resources and the storage resource of one virtual machine have to be based on the network connection among devices, which would reduce the performance of the virtual machine. With the proposal of Hyperconverged Infrastructure (HCI), more and more IaaS cloud platforms begin to adopt the homogeneous architecture. Servers in the platform would uniformly provide
computing resources and storage resources that are shown in Figure 2 [1]. This homogeneous architecture could effectively solve the problems encountered by the heterogeneous architecture.

Since there is no dedicated storage hardware for high data availability in the heterogeneous architecture, the cloud platform needs to introduce corresponding mechanisms to guarantee the data availability. Realizing high data availability of IaaS cloud platform mainly involves two aspects: one is data backup and the other is to realize the storage resources availability management. The data backup part mainly introduces backup policy management and backup data management. This part is not the focus of this paper. Furthermore, there are two main types of storage resources in the server: solid-state drive (SSD) hard disk and mechanical hard disk. SSD hard disk could provide higher data reading and writing speed, yet the cost is high, hence it is often used to realize the virtual machine system disk with high performance requirements. The mechanical hard disk, although its data reading and writing speed is relatively low, yet its cost is low, hence the mechanical hard disk is the main way to realize the data storage capacity of IaaS cloud platform. If we can predict the life of mechanical hard disks more accurately and perform operations such as data backup in a timely manner, we can effectively reduce the risk of damage. Existing mechanical hard disks already provide Self-Monitoring Analysis and Reporting Technology (S.M.A.R.T.) that can be used to sensing the operational status of the mechanical hard disk. Furthermore, S.M.A.R.T. provides indicators of the operational status of the various components of the drive, such as heads, platters, motors, and circuits assisting in the prediction of mechanical hard disk status.

Therefore, the key issue to address is how to predict mechanical hard disk life based on S.M.A.R.T. indicators in a timely and accurate manner. In recent years, several researchers have proposed methods for mechanical hard disk failure prediction, mainly divided into mathematics-based methods [2, 3] and machine learning-based failure prediction method [4]. These methods do not adequately consider the problems of removing unnecessary S.M.A.R.T. indicators, small number of failure samples, and making full use of timing data while predicting mechanical hard disk lifetime. In addition to this, some studies [5–7] have focused on assessing the dynamic reliability and fault prediction of the whole system, while we have mainly completed fault prediction for individual component hard disk.

In details, there are three challenges for the fault prediction of hard disks:

1. How to filter the S.M.A.R.T. indicators that have the greatest impact on fault warning. The S.M.A.R.T. indicators of mechanical hard disks are the basis for determining faults. Nevertheless, there are also some characteristics that are not relevant to the failure result, excessive characteristics that are useless and may even affect the final analysis result.

2. How to solve the problem of imbalanced sample size of failures. Statistically, the annual mechanical hard disk damage rate in data centers is around 2%-5%. Therefore, in the sensing data of hard disk operation status, the data related to abnormal status is far less than that related to normal status.

3. How to make the most of the timing relationships of mechanical hard disk data. Existing warning models for faulty hard disks first use time series data compression to complete feature extraction, and then pass the extracted data into a classifier for classification. This process has the potential to result in the loss of a large number of valuable features.

Therefore, the key problem to be solved in mechanical hard disk failure prediction in this paper is how to timely...
and accurately predict the service life of mechanical hard disk, so as to actively carry out data backup or migration before mechanical hard disk failure, so as to improve data availability.

To address the above challenges, we first propose the Relief feature selection algorithm to filter indicators and select valuable indicators. And we propose the Generative Adversarial Networks (GAN) model to generate a small number of class samples. Then, we propose the Convolution-Long Short-Term Memory (C-LSTM) to solve the problem of long-term dependence on time-series data and accurately detects faulty hard disk data.

The outline of this paper is listed as follows: Section II Related work reviews and discusses the previous related work; Section III Algorithm presents our algorithm; the experiment setup, results, and analysis are presented in Section IV Experimental results and discussion; in the end, Section V Conclusions makes a conclusion of this paper.

2. Related Work

Mechanical hard disk failure alerts have become increasingly important with the growth of IaaS cloud platforms. The hard disk is one of the most common failed components in today's IT systems, and damage to it can lead to suspension of system services or loss of data. As more and more services run on them, the damage caused by hard disk corruption is increasing every year.

2.1. Anomaly Detection of Mechanical Hard Disks. There are already several methods for detecting anomalies on mechanical hard disks. Yang et al. [8] proposed an evaluation method for comparing feature selection methods and anomaly detection algorithms for predicting hard disks failures. Yu et al. [9] proposed an adaptive error tracking method for hard disks fault prediction. Wang et al. [10] proposed a domain adaptive method to improve fault prediction performance.

With the development of deep learning, combined with its many excellent properties, deep learning is now widely used to solve problems in the prediction domain [11–13]. How to handle time series data needs to be considered when using deep learning methods to accomplish hard disk failure prediction. Several existing studies have been considering how to handle time series data. Hu et al. [14] propose a disk failure prediction system based on a Long Short-Term Memory (LSTM) network. By replacing the input in the LSTM network with the continuous operating records of the disk, the problem of individual variation of the disk can be solved.

2.2. Self-Monitoring Analysis and Reporting Technology (S.M.A.R.T.) Indicators. Self-Monitoring Analysis and Reporting Technology (S.M.A.R.T.) is a monitoring system that collects indicator performance that can be used to infer the actual condition of a hard disk. S.M.A.R.T.-based active fault tolerance uses a threshold approach [15], but traditional S.M.A.R.T.-based fault detection has problems in terms of accuracy [16]. It is no longer sufficient to complete the analysis using S.M.A.R.T. alone. A number of S.M.A.R.T.-based optimisation methods have been proposed. Li et al. [2] explored the ability of Decision Trees (DTs) [17] and Gradient Boosted Regression Trees (GBRT) [18] to predict hard disk faults based on S.M.A.R.T. indicators, and experimentally demonstrated that both prediction models have high fault detection rates and low false alarm rates. Chaves et al. [3] present a failure prediction method using a Bayesian network. The method calculates the deterioration of hard disks over time using S.M.A.R.T. indicators to predict eventual failures. De Santo et al. [19] propose a model based on LSTM, which combines S.M.A.R.T. indicators and temporal analysis for estimating the health of a hard disk based on its failure time.

Li et al. [20] proposed a combination of XGBoost, LSTM, and ensemble learning algorithms to effectively predict hard disk failures based on S.M.A.R.T. indicators. In conjunction with S.M.A.R.T., Shen et al. [21] propose a hard disk failure prediction model based on LSTM recurrent neural networks and a new method for assessing the degree of health. The model exploits the long-term time-dependent characteristics of hard disk health data to improve prediction efficiency and efficiently stores current health details and deterioration.

In addition to selecting all the attributes of S.M.A.R.T., some studies have also taken the approach of selecting some of the attributes. Wu et al. [4] propose the use of information entropy to optimise S.M.A.R.T. indicators to enable the selection of the most relevant attributes for prediction, combined with a Multichannel Convolutional Neural Network-Based Long Short-Term Memory (MCCNN-LSTM) model to complete the prediction of hard disk failures.

2.3. Sample Imbalance. The above study focuses on the use of S.M.A.R.T. indicators to detect anomalies and health states of hard disks. In addition, hard disks are healthy for most of their life cycle with relatively few failures, which creates a problem of sample imbalance.

GAN-based methods are often used to solve the problem of sample imbalance. Lee and Park [22] proposed a GAN-based fusion detection system for imbalanced data. Xu et al. [23] proposed a convergent Wasserstein GAN to solve the problem of class imbalance in network threat detection. Huang and Lei [24] proposed a novel Imbalanced GAN (IGAN) to deal with the problem of the class imbalance.

In addition to the GAN-based approach, a number of others have proposed solutions to the problem of imbalanced hard disk failure samples. Tomer et al. [25] propose to apply machine learning techniques to accurately and proactively predict hard disk failures. Shi et al. [26] proposed a deep generative transfer learning network (DGTL-Net) that integrates a deep generative sample for generating pseudo-failure network to generate pseudofailure samples and a deep transfer network to solve the problem of hard disk distribution discrepancy, enabling intelligent fault diagnosis of new hard disks. Ircio et al. [27] proposed an optimised classifier to solve the problem of imbalance between hard disk failure and normal hard disk height. Wang G. et al. [28] propose a multi-instance long-term data classification method
based on LSTM and attention mechanism to solve the problem of data imbalance.

3. Algorithm

We present an evaluation method for comparing feature selection methods and anomaly detection algorithms for predicting hard drive failures. It enables the rapid selection of the best algorithm for a particular model of hard disk. It includes an evaluation mechanism for assessing feature selection methods from a performance and robustness perspective and for assessing the performance, robustness, efficiency and generalisability of anomaly detection algorithms.

Hard disk failure prediction needs to deal with three important points, indicator selection, timing compression, and imbalanced sample processing. The overall process is shown in Figure 3.

First, \( n \) vectors of characteristic timing parameters are input, and the vector \( A_i(a_{i1}, a_{i2}, \ldots, a_{in}, \ldots, a_{iN}) \) is defined as the timing characteristics of parameter \( i \) input, and correlation analysis is performed using the correlation with the results, and the parameter with the higher correlation is selected. In the time series feature extraction stage, the current mainstream approach is to use single-value compression for continuous time series data over a period, it can be represented as:

\[
S_t = \alpha Y_t + (1 - \alpha) S_{t-1},
\]

where \( S_t \) is the cumulative data, \( Y_t \) is the data of a node, and \( \alpha \) is the coefficient.

Nevertheless, these time series feature extractions are often not enough. The main problem is that the previous data is forgotten faster and faster over time, and the sequence of values is not considered, resulting in the data not playing its due role.

On the other hand, the processing of imbalanced samples is relatively rough, often using oversampling of a few categories of data or undersampling of most categories of data. Nevertheless, oversampling of a few categories of data leads to changes in the probability of data features, which appear to have excellent performance in the training set, and decrease in the effect of the test set, resulting in a low recall rate. Using undersampling algorithms, clustering, and other methods to remove part of the samples to achieve sample balance, often result in loss of important features, or reduced data sample size, resulting in overfitting problems.

This algorithm is divided into offline model implementation and online data analysis. The detailed algorithm flow is shown in Figure 4.

As shown in Figure 4, this algorithm mainly includes offline model generation and online model detection. In the offline model generation stage, historical data is used for parameter selection, then time series features are extracted, samples are equalized, and finally a discriminant model is generated. In the online detection stage, parameter selection is performed, after time sequence features are extracted, model detection is performed, and finally the prediction result is generated.

1. Indicator Selection. Relief feature selection algorithm is used to filter parameters and select valuable indicators

2. Imbalanced Processing. In view of the few samples of mechanical hard disk damage, the GAN [29] model is used to generate a small number of class samples to achieve a sample balance state model is used to generate a small number of class samples to achieve a sample balance state

3. Model Generation. Using the processed data set to train the health status of the mechanical hard disk to generate models such as C-LSTM

3.1. Relief-Based Feature Selection Algorithm Parameter Selection. The S.M.A.R.T. [30] indicators gathered by the sensors installed in the mechanical hard disks for sensing the mechanical hard disks’ status usually have a fault warning characteristic, which are the basis for determining faults [31]. However, there are also some indicators that are not relevant to the failure result—excessive indicators that are useless and may even affect the final analysis result. When performing a hard disk analysis, it is necessary to consider the various complexities faced by hard disks. For instance, the capacity of a hard disk will gradually increase over time. In addition, the hard disk will slowly deteriorate, although the two are not very relevant as the capacity of a hard disk may be adjusted at any time. Therefore, it is essential to select the indicators to remove interfering features.

To address these issues, we select suitable indicators as model inputs based on the Relief feature selection algorithm [32]. The Relief algorithm focuses on the binary classification problem, which in this paper refers to whether a hard disk has been damaged. We propose the “correlation statistic” to measure the importance of a feature. The correlation statistic is a vector, each component of which is the evaluation of one of the initial features, and the importance of a subset of features is the sum of the correlation statistics for each feature in the subset. For the feature measurement problem, Relief borrows the idea of the hypothetical interval, the maximum distance that the decision surface can move while keeping the sample classification constant, which can be expressed as [33]:

\[
\theta = \frac{1}{2} (||x - M(x)|| - ||x - H(x)||),
\]

where \( M(x) \) and \( H(x) \) refer to the nearest neighbors that are of the same kind as \( x \) and those that are not of the same kind as \( x \), respectively.

We know that when an attribute is favorable for classification, then samples of that kind are closer to that attribute, while samples of the opposite kind are further apart from that attribute.

Suppose the training set \( D = \{(x_1, y_1), (x_2, y_2), \ldots, (x_m, y_m)\}, \) for each sample \( x_i \), the nearest neighbour \( x_{i, nth} \) of the
same category as \( x_i \) is calculated, which is called “guessed nearest neighbor” (near - heat). Then the nearest neighbor \( x_{i_{nn}} \), which is not the same as \( x_i \), is called the “wrong nearest neighbor” (near - miss), and the relevant statistic for attribute \( j \) is [33]:

\[
\delta^j = \sum_i -\text{diff}(x^j_i, x^j_{i_{nn}})^2 + \text{diff}(x^j_i, x^j_{i_{nm}})^2,
\]

where \( x^j_i \) represents the value of sample \( x_i \) on attribute \( j \). The calculation of \( \text{diff}(x^j_a, x^j_b)^2 \) depends on the type of attribute \( j \).

Discrete attributes:

\[
\text{diff}(x^j_a, x^j_b) = \begin{cases} 
0, & x^j_a = x^j_b \\
1, & \text{otherwise}
\end{cases}
\]

Continuous attributes:

\[
\text{diff}(x^j_a, x^j_b) = |x^j_a - x^j_b|.
\]
3.2. GAN-Based Imbalance Data Processing. In the daily operation of an IaaS cloud platform system, the number of failed hard disks is relatively small, while the number of normal samples is always large. Statistically, the annual mechanical hard disk damage rate in data centers is around 2%-5%, and a hard disk is normal for most of the time, which results in the raw positive and negative sample data always being imbalanced. Using machine learning methods for failure prediction on imbalanced data sets requires either oversampling a smaller number of data categories to achieve data balance or undersampling a larger portion of the data. Conventional oversampling algorithms, however, can lead to changes in the probability of data for a few classes, undersampling leading to loss of important features in most

![Figure 6: Experimental results based on Relief screening algorithm.](image1)

![Figure 7: WGAN network generates samples.](image2)
classes, or overfitting problems due to insufficient training data. Examples include the use of Synthetic Minority Oversampling Technique (SMOTE) oversampling algorithms [34], which synthesize new samples for a few classes based on interpolation, and the use of clustering algorithms to implement undersampling and discard some samples to alleviate class imbalance.

Considering the problems of the original algorithm in dealing with imbalanced data, the innovation of this algorithm is to use the zero-sum game idea of GAN to generate less category samples. The GAN continuously plays a game through the generative network $G$ and the discriminative
network \(D\), which in turn enables \(G\) to learn the distribution of the data. Using the GAN method, the generative network \(G\) and the discriminative network \(D\) are continuously played by using the zero-sum game idea in game theory, which in turn enables \(G\) to learn the distribution of the data.

Define the distribution \(P_{\text{data}}(x)\) of the set of real images, \(x\) being a real image. Now it is necessary to generate some pictures that also fall within this distribution. Define the distribution generated by the generator \(G\) as \(P_G(x; \theta)\), with \(\theta\) being the distribution parameter. Now, compute the likelihood function in the generative model as [35]:

\[
\mathcal{L} = \prod_{i=1}^{m} P_G(x^i; \theta). \tag{6}
\]

To implement the generator \(G\) to generate the real picture maximum, the likelihood function needs to be maximized. That is, it is necessary to find a \(\theta^*\) to maximize the likelihood [36]:

\[
\theta^* = \arg \min_{\theta} \text{KL}(P_{\text{data}}(x) \| P_G(x; \theta)). \tag{7}
\]

The generator randomly generates a vector \(Z\) and generates a picture \(X\) through the generator \(G(Z) = X\) network, that is, the generator sample interval. Then the discriminator \(D(X) \in [0, 1]\) is used to distinguish the samples generated by the generator from those in the original sample space. And GAN is computed as follows [36]:

\[
V(G, D) = E_{x \sim P_{\text{data}}} \left[ \log D(x) \right] + E_{x \sim P_G} \left[ \log (1 - D(x)) \right]. \tag{8}
\]

Through \(k\) rounds of training, the discriminator can accurately distinguish between the original data and the data generated by the generator \(G\). Next, train the generator so that the generator can confuse the discriminator and make it indistinguishable. Through multiple rounds of training and adjusting the discriminator and generator network results, a better model effect can be achieved. However, the stability of GAN training is not good, and it is difficult to achieve the desired effect in this experiment. By improving GAN, there are currently better algorithms such as Deep Convolutional GAN (DCGAN) [37], Wasserstein GAN (WGAN) [38], and Wasserstein GAN with Gradient Penalty (WGAN-GP) [39].

WGAN uses the Wasserstein distance, which has superior smoothing properties compared to Jense-Shannon (JS) and solves the gradient disappearance problem [23]. In addition, WGAN addresses not only the problem of GAN training instability but also provides a reliable indicator of the training process, and the indicator is highly correlated with the quality of the generated samples. Therefore, we choose WGAN as a method to solve the data imbalance problem.

3.3. Based on LSTM Network Anomaly Detection and Recognition. Our proposed LSTM network-based model solves the problem of long-term dependence on time-series data and accurately detects faulty hard disk data. The traditional faulty hard disk early warning model uses time series data compression to first extract features, and then transfer
the extracted data to the classifier for classification, resulting in the loss of many valuable features. To extract the temporal relationships of mechanical hard disk data, LSTM networks are added to the model training in this paper.

3.3.1. The Improved Network Structure of LSTM. The original LSTM network structure only takes into account the temporal sequence of data in time. Nevertheless, for hard disks, changes in certain parameters will affect the data of other parameters. Compared to the common LSTM structure, this algorithm borrows from the Convolutional LSTM, which means that convolutional computation is added to the input layer, local perception and pooling are introduced, spatial features are added and input to the LSTM structure together with the original data. The structure of C-LSTM is shown in Figure 5.

Considering that this model is a multicategory model, the output should be the probability of each category. The values obtained from the neural network are normalized using the Softmax function, placing the results between [0,1], with larger values corresponding to larger probabilities. The Softmax function category $i$ probability $y_i = P(C_i|\mathbf{x})$ is calculated as follows [40]:

$$y_i = \frac{e^{C_i}}{\sum_{j=1}^{n} e^{C_j}},$$

where $C_i : w_i, b_i, z_i = w_i \cdot \mathbf{x} + b_i$.

After the Softmax function has processed the results, our model uses cross entropy as the loss function. The loss function formula for Softmax is as follows [41):

$$L = \sum -\hat{y}_i \ln y_i,$$

where $y_i$ is the probability of category $i$.

When calculating the loss function, the problem of gradient explosion arises, our model uses clip gradients method [42] to keep the weights within a certain range.

4. Experimental Results and Discussion

To verify the predictive effectiveness of the algorithm, fault warning experiments were conducted on mechanical hard
Figure 15: AUC and precision of LSTM.

Figure 16: Recall of LSTM.
Figure 17: Accuracy and loss of C-LSTM.

Figure 18: AUC and precision of C-LSTM.
disk data from data centers and compared with traditional undersampling-based equalization and binary classification methods (the demo of experiments: https://github.com/Eva0417/HardDisk).

4.1. Data Description. The experimental data is from Backblaze, and it consists mainly of data which gathered by sensors from nearly 30,000,000 mechanical hard disks over a 1-year period in 2017 (the dataset: https://github.com/1210882202/data). The data is mainly S.M.A.R.T. indicators gathered once a day, with some of the disks not sensoring S.M.A.R.T. indicators over time, indicating that the mechanical disk has been damaged. The objective of the experiment was to predict whether the disk would become damaged in the future based on the last sixty days of data for these disks. As mechanical hard disks generally deteriorate slowly as the components age, this experiment assumes that the mechanical hard disk is not damaged within fifteen days and this data is marked as healthy, and if the disk is damaged within fifteen days this data is marked as faulty. Based on the sample data, the experiment hopes to design a fault warning model with excellent performance in terms of accuracy, recall, and Area Under Curve (AUC) value.

4.2. Baseline. To evaluate the performance of C-LSTM against traditional models, traditional classifiers were added to the experiments. Details are as follows:

(1) Logistic Regression (LR) [43]. LR is a supervised learning method often used in anomaly detection. For one variable or multiple independent variables, find the best fitting model to describe the set of independent variables, and complete the anomaly detection in this way

(2) Random Forest (RF) [44]. RF is a common method of anomaly detection by bringing together multiple decision trees. The basic unit is a tree-structured decision tree. With this structure, normal instances can be learned and instances that are not classified as normal are considered as anomalies

4.3. Experimental Setup

4.3.1. The Settings of LSTM

(1) Input and Output. For the input data, the data relevance is first judged using the Relief algorithm to obtain valid 16-dimensional data, and the data sample map is obtained based on the faulty sample generation method. The specific input is a None * Seq * 16-dimensional tensor, and the output is a None * 2-dimensional tensor

(2) Network Structure. The LSTM network used in the experiments uses a network containing two layers of LSTM hidden layers, with a dropout layer added after each hidden layer, followed by a fully connected layer connecting the LSTM and the output, and finally a SoftMax layer

(3) Network Parameters. The key parameters of the neural network used for the experiments were set as shown in Table 1

4.3.2. The Settings of C-LSTM

(1) Input and Output Settings. The input data is the same as the original LSTM

(2) Network Structure. The network used in the experiment adds a layer of Convolutional network after the input layer, which is combined with the original input data and fed into the LSTM hidden layer network

4.4. The Results of Experiment. In applying the Relief screening algorithm, attribute-related statistical components are calculated for the indicators which gathered by sensors in hard disks, the larger the score, the greater the classification power. The statistical components are ranked by size and scaled to take the key indicators needed. First, we analyzed data collected from 26,339 disks over a six-month period
in the first half of 2017. The results based on the Relief filtering algorithm are shown in Figure 6.

In Figure 6, the horizontal axis represents each dimension number, and the vertical axis represents the relevance of each dimension to the results, taking values in the range [0, 1], with closer to zero indicating less relevance to the results. Based on the results of the statistical components obtained in Figure 6, the parameters whose results are greater than the threshold are selected, and the final hard disk correlation is obtained.

According to the above analysis, we use the WGAN network to conduct experiments, and the sample generation is shown in Figure 7.

The horizontal axis of Figure 7 represents each indicator gathered by sensors of the mechanical hard disk, and the vertical axis represents time. Darker colors in Figure 7 represent lower indicator values, and lighter colors represent higher indicator values. As can be seen from Figure 7, the WGAN network uses the principles of game theory to generate samples that are relatively similar and can simulate a large amount of information, while at the same time differing from direct replication. The experimental results show that the use of WGAN for feature extraction and regeneration of the overall fault sample solves the problem of sample imbalance and expands the fault sample.

In addition to experimenting on our proposed C-LSTM model, we have also experimented on the comparison algorithms.

According to the specific experimental setup, the results of this experiment for the comparison model of LR are as shown in Figures 8–10. In these figures, we use different color to show the different scores of LR.

According to the specific experimental setup, the results of this experiment for the comparison model of RF are as shown in Figures 11–13. In these figures, we also use different color to show the different scores of RF.

According to the specific experimental setup, the results of this experiment for training the network model of LSTM are as shown in Figures 14–16.

The horizontal axis of graphs in Figures 14–16 represents the number of training epochs, the vertical axis of the first graph in Figure 14 represents the accuracy rate during training, and the vertical axis of the second graph represents the training loss data. According to the graphs in Figure 14, we can see that after 3 epochs, the training gradually leveled off (in this paper, we define that a loss reduction of no more than 0.1 after 1 epochs is considered smooth), where the loss converged at around 0.05. Based on the Figures 15 and 16, we can see that the accuracy of training is around 0.91.

The results of this experiment on the training of the C-LSTM network model are shown in Figures 17–19.

The horizontal axis of graphs in Figures 17–19 represents the number of training epochs, the first graph’s vertical axis represents the accuracy rate during training in Figure 17, and the second graph’s vertical axis represents the training loss data. After 4.0 epochs, the training gradually leveled off, with the loss converging at around 0.05 and the accuracy at training at around 0.93.

Comparing the training results of the LSTM in Figures 14–16 with the C-LSTM network model in Figures 17–19, we can conclude that the C-LSTM has a faster convergence speed, lower loss drop, and higher accuracy. Therefore, from a training perspective, the C-LSTM performs better.

The individual classification models were evaluated according to precision, recall and AUC value, and the results are shown in Table 2. In terms of each metric, the C-LSTM model has the best result.

### 5. Conclusions

Firstly, the mechanical hard disk is installed with sensors for sensing the status of the mechanical hard disk and the S.M.A.R.T. indicators gathered by these sensors on the operational status of the various components of the disk can be used to predict the life of the mechanical hard disk. Based on this, we focus on how to accurately predict mechanical hard disk failure and achieve effective improvement of data availability in the IaaS cloud platform.

The model proposed in this paper includes three parts: Relief feature selection algorithm, WGAN, and LSTM models. To remove features from the S.M.A.R.T. indicators of mechanical hard disks that are irrelevant to the failure results, we use the Relief feature selection algorithm to remove interfering features and complete the parameter screening. As the number of failed hard disks is small in the IaaS cloud platform system, we use the zero-sum game idea of WGAN to generate fewer category samples to solve the sample data imbalance problem. Finally, we use the improved C-LSTM model to complete hard disk failure detection and early warning.

Through extensive experiments, we constructed our own model and evaluated the model we designed and other methods using precision, recall, and AUC value. The experiments demonstrate that our proposed algorithm outperforms other algorithms for mechanical hard disk warning.

As our future work, we aim to extend our approach to SSD-based IaaS cloud platforms. In our proposed approach, we mainly implement mechanical hard disk S.M.A.R.T.-based fault warning through WGAN and LSTM to achieve effective improvement of data availability in IaaS cloud platforms. However, as more and more IaaS cloud platform systems gradually adopt SSD to pursue significant performance improvement. Based on this, how to better realize the automation of repair in SSD-based IaaS cloud platform and study the automatic adaptation of parameters are our future goals to achieve.
Data Availability
All data, models, and code generated or used during the study appear in the submitted article.

Conflicts of Interest
The authors declare that they have no conflicts of interests.

Acknowledgments
This work is supported by Science and Technology Project from State Grid Information and Telecommunication Branch of China: Research on Key Technologies of Operation Oriented Cloud Network Integration Platform (52993920002P).

References
[1] S. Bhaumik, R. Bansal, R. Karmakar et al., “Netstor: Network and storage traffic management for ensuring application qos in a hyperconverged data-center,” IEEE Transactions on Cloud Computing, vol. 10, no. 2, 2020.
[2] J. Li, R. J. Stones, G. Wang, X. Liu, Z. Li, and M. Xu, "Hard drive failure prediction using decision trees," Reliability Engineering and System Safety, vol. 164, pp. 55–65, 2017.
[3] I. C. Chaves, M. Paula, L. Leite, L. P. Queiroz, and J. C. Machado, "Banahp: a Bayesian network based failure prediction approach for hard disk drives," in 2016 5th Brazilian Conference on Intelligent Systems (BRACIS), pp. 427–432, Recife, Brazil, 2017.
[4] J. Wu, H. Yu, Z. Yang, and R. Yin, "Disk failure prediction with multiple channel convolutional neural network," in 2021 International Joint Conference on Neural Networks (IJCNN), pp. 1–8, Shenzhen, China, 2021.
[5] J. Lie and E. Zio, "System dynamic reliability assessment and failure prognostics," Reliability Engineering and System Safety, vol. 160, pp. 21–36, 2017.
[6] H. Khorasgani, G. Biswas, and S. Sankararaman, "Methodologies for system-level remaining useful life prediction," Reliability Engineering and System Safety, vol. 154, pp. 8–18, 2016.
[7] K. L. Son, M. Fouladirad, and A. Barros, "Remaining useful lifetime estimation and noisy gamma deterioration process," Reliability Engineering and System Safety, vol. 149, pp. 76–87, 2016.
[8] Q. Yang, X. Jia, X. Li, J. Feng, W. Li, and J. Lee, "Evaluating feature selection and anomaly detection methods of hard drive failure prediction," IEEE Transactions on Reliability, vol. 70, no. 2, pp. 749–760, 2021.
[9] W. Yu, H. Long, J. Shan, and T. W. Chow, "Failure prediction of hard disk drives based on adaptive rao–blackwellized particle filter error tracking method," IEEE Transactions on Industrial Informatics, vol. 17, no. 2, pp. 913–921, 2021.
[10] J. Wang, R. Zhang, G. Qi, and L. Hong, "A heuristicRMI method on hard disk failure prediction in out-of-distribution environments," in 2021 IEEE International Conference on Industrial Engineering and Engineering Management (IEEM), pp. 1661–1664, Singapore, Singapore, 2021.
[11] W. Jiang, "Applications of deep learning in stock market prediction: recent progress," Expert Systems with Applications, vol. 184, article 115537, 2021.
[12] A. Appelt, B. Elhaminia, A. Gooya, A. Gilbert, and M. Nix, "Deep learning for radiotherapy outcome prediction using dose data – a review," Clinical Oncology, vol. 34, no. 2, pp. e87–e96, 2021.
[13] J. R. Jiang, J. E. Lee, and Y. M. Zeng, "Time series multiple channel convolutional neural network with attention-based long short-term memory for predicting bearing remaining useful life," Sensors (Basel, Switzerland), vol. 20, no. 1, 2020.
[14] L. Hu, L. Han, Z. Xu, T. Jiang, and H. Qi, "A disk failure prediction method based on LSTM network due to its individual specificity," Procedia Computer Science, vol. 176, pp. 791–799, 2020.
[15] H. Wang and H. Zhang, "AIOPS prediction for hard drive failures based on stacking ensemble model," in 2020 10th Annual Computing and Communication Workshop and Conference (CCWC), pp. 0417–0423, Las Vegas, NV, USA, 2020.
[16] J. Zhao, Y. He, H. Liu et al., "Disk failure early warning based on the characteristics of customized smart," in 2020 19th IEEE International Conference on Thermal and Thermomechanical Phenomena in Electronic Systems (ITherm), pp. 1282–1288, Orlando, FL, USA, 2020.
[17] J. Guo, H. Wang, X. Li, and L. Zhang, "Stream classification algorithm based on decision tree," Mobile Information Systems, vol. 2021, Article ID 3103053, 11 pages, 2021.
[18] J. H. Friedman, "Greedy function approximation: a gradient boosting machine," Annals of Statistics, vol. 29, no. 5, pp. 1189–1232, 2001.
[19] A. De Santo, A. Galli, M. Gravina, V. Moscato, and G. Sperli, "Deep learning for HDD health assessment: an application based on LSTM," IEEE Transactions on Computers, vol. 71, no. 1, pp. 69–80, 2022.
[20] Q. Li, H. Li, and K. Zhang, "Prediction of HDD failures by ensemble learning," in 2019 IEEE 10th International Conference on Software Engineering and Service Science (ICSESS), pp. 237–240, Beijing, China, 2019.
[21] J. Shen, Y. Ren, J. Wan, and Y. Lan, "Hard disk drive failure prediction for mobile edge computing based on an LSTM recurrent neural network," Mobile Information Systems, vol. 2021, Article ID 8878364, 12 pages, 2021.
[22] J. H. Lee and K. H. Park, "Gan-based imbalanced data intrusion detection system," Personal and Ubiquitous Computing, vol. 25, no. 1, pp. 121–128, 2019.
[23] Y. Xu, X. Zhang, Z. Qiu, X. Zhang, J. Qiu, and Z. Hua, "Over-sampling imbalanced data based on convergent WGAN for network threat detection," Security and Communication Networks, vol. 2021, Article ID 9206440, 19 pages, 2021.
[24] S. Huang and K. Lei, "IGAN-IDS: an imbalanced generative adversarial network towards intrusion detection system in ad-hoc networks," Ad Hoc Networks, vol. 105, article 102177, 2020.
[25] V. Tomer, V. Sharma, S. Gupta, and D. P. Singh, "Hard disk drive failure prediction using smart attribute," Materials Today: Proceedings, vol. 46, pp. 11258–11262, 2021.
[26] C. Shi, Z. Wu, X. Lv, and Y. Ji, "DGTL-Net: a deep generative transfer learning network for fault diagnostics on new hard disks," Expert Systems with Applications, vol. 169, no. 16, article 114379, 2020.
[27] J. Ircio, A. Lojo, J. A. Lozano, U. Mori, and J. A. Lozano, "A multivariate time series streaming classifier for predicting hard drive failures [application notes]," IEEE Computational Intelligence Magazine, vol. 17, no. 1, pp. 102–114, 2022.
[28] G. Wang, Y. Wang, and X. Sun, “Multi-instance deep learning based on attention mechanism for failure prediction of unlabelled hard disk drives,” IEEE Transactions on Instrumentation and Measurement, vol. 70, pp. 1–9, 2021.

[29] M. Loey, F. Smarandache, and N. Khalifa, “Within the lack of chest covid-19 x-ray dataset: a novel detection model based on GAN and deep transfer learning,” Symmetry, vol. 12, no. 4, p. 651, 2020.

[30] A. Hudson-Smith, S. Hügel, and F. Roumpani, “Self-monitoring, analysis and reporting technologies. The Routledge companion to smart,” Cities, 2020.

[31] T. Jiang, J. Zeng, K. Zhou, P. Huang, and T. Yang, “Lifelong disk failure prediction via GAN-based anomaly detection,” in 2019 IEEE 37th International Conference on Computer Design (ICCD), pp. 199–207, 2019.

[32] “Relief-based feature selection: introduction and review,” Journal of Biomedical Informatics, vol. 85, pp. 189–203, 2018.

[33] G. Wang, J. Xu, L. Xu, D. Liang, and Y. Xu, “Method of extracting characteristic parameters of medium-speed maglev train levitation controller based on relief algorithm,” in 2020 Chinese Control And Decision Conference (CCDC), pp. 3366–3370, Hefei, China, 2020.

[34] S. Maldonado, J. López, and C. Vairetti, “An alternative smote oversampling strategy for high-dimensional datasets,” Applied Soft Computing, vol. 76, pp. 380–389, 2018.

[35] D. P. Kingma and M. Welling, “Auto-encoding variational Bayes,” 2013, https://arxiv.org/abs/1312.6114.

[36] J. Gui, Z. Sun, Y. Wen, D. Tao, and J. Ye, “A review on generative adversarial networks: algorithms, theory, and applications,” IEEE Transactions on Knowledge and Data Engineering, pp. 1–1, 2022.

[37] A. Radford, L. Metz, and S. Chintala, “Unsupervised representation learning with deep convolutional generative adversarial networks,” 2015, https://arxiv.org/abs/1511.06434.

[38] M. Arjovsky, S. Chintala, and L. Bottou, Wasserstein GAN, 2017.

[39] I. Gulrajani, F. Ahmed, M. Arjovsky, V. Dumoulin, and A. Courville, “Improved training of Wasserstein GANs,” Advances in Neural Information Processing Systems, vol. 30, pp. 5769–5779, 2017.

[40] M. A. Hussain and T.-H. Tsai, “An efficient and fast softmax hardware architecture (EFSHA) for deep neural networks,” in 2021 IEEE 3rd International Conference on Artificial Intelligence Circuits and Systems (AICAS), pp. 1–4, Washington DC, DC, USA, 2021.

[41] K. Lohia and J. Singh, “LP constrained softmax loss for handwritten image recognition,” in 2018 4th International Conference for Convergence in Technology (I2CT), pp. 1–6, Mangalore, India, 2018.

[42] R. Pascanu, T. Mikolov, and Y. Bengio, “Understanding the exploding gradient problem,” 2012, https://arxiv.org/abs/1211.5063.

[43] S. S. Noureen, S. B. Bayne, E. Shaffer, D. Porschet, and M. Berman, “Anomaly detection in cyber-physical system using logistic regression analysis,” in 2019 IEEE Texas Power and Energy Conference (TPEC), pp. 1–6, College Station, TX, USA, 2019.

[44] Y. Guo, Y. Zhou, X. Hu, and W. Cheng, “Research on recommendation of insurance products based on random forest,” in 2019 International Conference on Machine Learning, Big Data and Business Intelligence (MLBDII), pp. 308–311, Taiyuan, China, 2019.