

## Research Article

# Intelligent Classification Method of Low Occupancy Big Data Based on Grid Index

Haiyan Zhao <sup>1</sup> and Shuangxi Li<sup>2</sup>

<sup>1</sup>*School of Information Science and Engineering, Tianjin Tianshi College, Tianjin 301700, China*

<sup>2</sup>*School of Computer Science and Engineering, Tianjin University of Technology, Tianjin 300384, China*

Correspondence should be addressed to Haiyan Zhao; [glistner@email.tjut.edu.cn](mailto:glistner@email.tjut.edu.cn)

Received 18 October 2021; Revised 11 November 2021; Accepted 17 November 2021; Published 16 December 2021

Academic Editor: Bai Yuan Ding

Copyright © 2021 Haiyan Zhao and Shuangxi Li. 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.

In order to enhance the load balance in the big data storage process and improve the storage efficiency, an intelligent classification method of low occupancy big data based on grid index is studied. A low occupancy big data classification platform was built, the infrastructure layer was designed using grid technology, grid basic services were provided through grid system management nodes and grid public service nodes, and grid application services were provided using local resource servers and enterprise grid application services. Based on each server node in the infrastructure layer, the basic management layer provides load forecasting, image backup, and other functional services. The application interface layer includes the interfaces required for the connection between the platform and each server node, and the advanced access layer provides the human-computer interaction interface for the operation of the platform. Finally, based on the obtained main structure, the depth confidence network is constructed by stacking several RBM layers, the new samples are expanded by adding adjacent values to obtain the mean value, and the depth confidence network is used to classify them. The experimental results show that the load of different virtual machines in the low occupancy big data storage process is less than 40%, and the load of each virtual machine is basically the same, indicating that this method can enhance the load balance in the data storage process and improve the storage efficiency.

## 1. Introduction

In the process of operation and development, enterprise networks will accumulate a large number of low occupancy big data [1, 2]. Effective analysis of this low occupancy big data can obtain implicit data and knowledge, resulting in data value added and providing and attaching a variety of services [3, 4], which shows that low occupancy big data is very key. These low occupancy big data are stored in the form of message file or database and rise exponentially [5], so high-quality storage methods are required. Intelligent classification technology is a very important big data management technology. At present, it has been successfully used in various fields, but the problems of high occupancy rate and low classification efficiency of traditional big data classification are still difficult to solve. In view of the above phenomena, scholars at home and abroad have put forward the following solutions.

In [6] with the rapid development of sensing and digital technology, network physical system is regarded as the most feasible platform to improve architectural design and management. It investigates the possibility of integrating energy management system with network physical system to form energy network physical system to promote building energy management. However, due to the dynamics of building occupants, minimizing energy consumption while realizing the architectural function of energy network physical system is a challenge. Because occupant behavior is the main source of uncertainty in energy management, ignoring it will usually lead to energy waste caused by overheating and undercooling, as well as discomfort caused by insufficient heat and ventilation services. In order to alleviate this uncertainty, an energy network physical system related to occupancy is proposed, which combines occupancy detection based on WiFi probe. The framework uses integrated classification algorithm to extract three forms of

occupancy information. It creates a data interface connecting the energy management system and the network physical system and realizes automatic occupancy detection and interpretation by assembling multiple weak classifiers for WiFi signals. A validation experiment was conducted in a large office to check the performance of the proposed network physical system occupying linked energy. Experimental and simulation results show that the proposed model can save about 26.4% of refrigeration and ventilation energy consumption with appropriate classifiers and occupancy data types. However, this method does not conduct in-depth research on data classification methods. In [7] in complex image environment, text detection and location from natural images embedded with text is still a challenging problem. Foreground object segmentation and classification are common methods for this task. Therefore, component level target classification in a clutter environment is an important subproblem. Proper extraction of foreground objects can achieve effective classification, so as to improve the performance of text detection. A feature vector based on equidistant pixel area distribution is proposed for text/nontext classification. The generated feature descriptor is script invariant and is very effective in the actual scene. Five different pattern classifiers are used to evaluate the proposed feature set on our data set. The experimental results show that the classification accuracy of the feature set is more than 86% regardless of the script. However, this method only classifies image data and has high limitations. Reference [8] uses three kinds of random forest classifiers to solve the problem of anomaly classification of sensor data. Consider the sensor deployment scenario specifically, where the sensor fields of view may overlap. According to the sensor type, we design signal features in time, frequency, and space-time domain. The results show that the proposed random forest classifier has higher true positive rate and lower false positive rate than unsupervised k-means method and random forest classifier with single signal energy feature. However, the classification effect of this method for low occupancy data is still unknown.

Grid technology can transform different computer resources distributed in a wide range of space into a computing power and data processing power characterized by universality, standardization, accuracy, and economy, so as to realize wide area resource sharing [9]. In essence, grid technology can be understood as maximizing the use of existing resources (software and hardware) in the network, meeting the storage, sharing and calculation of data and resources, improving the problem of resource island [10], and realizing the value-added network resources. Grid technology is widely used in data storage fields such as e-commerce, academic research, and so on. Based on this, this paper studies the low occupancy big data classification method based on grid technology and uses grid technology to integrate low occupancy big data and realize high-quality storage of low occupancy big data.

## 2. Low Occupancy Big Data Processing and Classification

*2.1. Big Data Preprocessing Stage.* Multisource network data is massive and high-dimensional. Cross-source classification

algorithm feature extraction of multisource network data needs dimension reduction function. For multisource heterogeneous data, the traditional dimensionality reduction method can not determine the dimensionality of all kinds of data. The amount of data contained in multisource heterogeneous data is high, and its workload is also high. Feature extraction methods with high computational efficiency are needed.

Incremental orthogonal component analysis (IOCA) is selected to reduce the dimension and extract features of multisource network data, so as to improve the time complexity of massive multisource network data classification. IOCA method does not need to set a fixed target dimension, and the target dimension can be adjusted according to the changes of input data in the learning process [11]. Using this method to preprocess massive data will form better orthogonal components and avoid data redundancy and good compression dimension.

IOCA method can use the prefetched multisource network data to obtain the orthogonal component space  $\{d_1, d_2, \dots, d_n\}$  that can automatically determine the dimension, so as to realize the rapid dimensionality reduction of multisource network data. The dimensionality reduction of multisource network data needs to be realized through the following two steps:

- (1) Let the existing new data be represented by  $\varphi_{k+1}$  and the learned orthogonal component space be represented by  $D = \{d_1, d_2, \dots, d_k\}$ , and calculate the new potential orthogonal component  $d_{k+1}$  possibly generated by  $\varphi_{k+1}$  and the linear independence between them.
- (2) Set the adaptive threshold, and use the set adaptive threshold to judge whether  $d_{k+1}$  can be used as a new orthogonal component added to  $D$ .

The specific calculation process of extracting data features by IOCA method is as follows:

- (1) Initialize the orthogonal component space  $D = \emptyset$  with the initial dimension  $k = \dim(D) = 0$ .
- (2) Use  $x_j$  to represent the new input data and meet  $j > k$ .
- (3) Use  $r_{i,k+1}$  to represent the eigenvector and calculate  $r_{i,k+1} = x_j^T d_i$ .
- (4) Calculate  $d_{k+1}' = x_j - \sum_{i=1}^k r_{i,k+1} d_i$ .
- (5) Calculate  $r_{k+1,k+1} = \|d_{k+1}'\|_2$ .
- (6) Calculate  $d_{k+1} = d_{k+1}' / r_{k+1,k+1}$ .
- (7)  $g$  represents the original data dimension. When  $r_{k+1,k+1} / \|x_j\| \geq \dim(D) / g$ , it means that  $d_{k+1}$  belongs to the new orthogonal component. At this time,  $d_{k+1}$  is added to  $D$ .
- (8) Let  $k = k + 1$  repeat the above process until all data preprocessing is completed.

The features of multisource network data are extracted by using the corresponding orthogonal space to obtain multiple groups of feature vectors with lower dimensions

[12]. The feature vectors with lower dimensions are sent to the multicore learning support vector machine classifier to realize the cross-source classification of multisource network data.

*2.2. Intelligent Classification Algorithm Selection.* Support vector machine is a machine learning method based on structural risk minimization principle and statistical learning theory [13], and support vector machine has good generalization performance. Support vector machine realizes multisource network data classification by searching the optimal hyperplane [14]. The solution formula of the optimal hyperplane obtained through the quadratic optimization problem is as follows:

$$\begin{cases} \min \frac{1}{2} \|w\|^2 + C \sum_{i=1}^N \xi_i \\ \text{s.t. } y_i [(w, x) + b] \geq 1 - \xi_i \end{cases} \quad (1)$$

In formula (1),  $\|w\|^2$  and  $\xi_i$  represent the regularization term and the relaxation variable measuring the training error of the sample, respectively. The smaller the value of  $\xi_i$ , the lower the training error. At this time, the support vector machine has higher classification accuracy;  $C$  represents the penalty coefficient of the balance adjustment parameter between the training error and the regularization term. The smaller the value, the lower the degree of penalty error classification;  $(w, x) + b = 0$  represents the hyperplane to be solved, and  $w$  and  $b$  represent the normal vector and offset of the hyperplane, respectively.

*2.2.1. Multicore Learning Method.* Multisource heterogeneous data can replace a single kernel with multiple kernels to improve the interpretability of decision function and the classification performance of support vector machine classifier [15, 16]. The convex combination formula based on kernel function  $K(x, x')$  is as follows:

$$K(x, x') = \sum_{m=1}^M s_m K_m(x, x'). \quad (2)$$

In formula (2),  $s_m \geq 0$ ,  $\sum_{m=1}^M s_m = 1$ .  $M$  and  $K_m$  represent the total number of cores and the number of positive definite cores  $M$  in the same input space  $\chi$ , respectively. The classical kernel with different parameters is represented by each basic kernel  $K_m$ .

Through the above process, the weight  $s_m$  is selected to replace the data representation of the kernel.

Based on the gradient descent of the target value of support vector machine, the gradient descent of support vector machine solver is used to determine the combination of different problem kernels, that is, multicore learning method. The multicore learning is realized by clarifying the coefficient  $\{s_m\}$  of the learning process of the

decision function. The multicore learning formula is as follows:

$$\begin{cases} \min_{\{f_m\}, b, \xi, s} \frac{1}{2} \sum_{m=1}^M \frac{1}{s_m} \|f_m\|_{H_m}^2 + C \sum_{i=1}^l \xi_i \\ \text{s.t. } y_i \sum_{m=1}^M f_m(x_i) + y_i b \geq 1 - \xi_i, \forall i \\ \xi_i \geq 0, \forall i \\ \sum_{m=1}^M s_m = 1, s_m \geq 0, \forall m \end{cases} \quad (3)$$

In formula (3),  $C > 0$  and  $\xi_i \geq 0$  represent the penalty coefficient and relaxation variable, respectively, and  $l$  represents the number of samples. When  $s_m = 0$ , the square norm  $\|f_m\|_{H_m}$  of  $f_m$  in each  $s_m$  control objective function needs to be 0 so that the objective value is limited.

Let the Lagrange function exist as follows:

$$\begin{aligned} L = & \frac{1}{2} \sum_{m=1}^M \|f_m\|_{H_m}^2 + C \sum_{i=1}^l \xi_i + \sum_{i=1}^l \alpha_i \\ & \left( 1 - \xi_i - y_i \sum_{m=1}^M f_m(x_i) - y_i b \right) \\ & - \sum_{i=1}^l v_i \xi_i + \lambda \left( \sum_{m=1}^M s_m - 1 \right) - \sum_{m=1}^M \eta_m s_m. \end{aligned} \quad (4)$$

In formula (4), both  $\alpha_i$  and  $v_i$  represent Lagrange multipliers related to support vector machine problems. Both  $\lambda$  and  $\eta_m$  represent Lagrange multipliers related to constraints on  $s_m$ .

Relative to the original variable, set the Lagrange function gradient to 0, substitute the set optimization conditions into the Lagrange function, and obtain the dual problem formula as follows:

$$\begin{cases} \max_{\alpha} -\frac{1}{2} \sum_{i,j=1}^l \alpha_i \alpha_j y_i y_j \sum_{m=1}^M s_m K_m(x_i, x_j) + \sum_{i=1}^l \alpha_i \\ \text{s.t. } \sum_{i=1}^l \alpha_i y_j = 0 \\ C \geq \alpha_i \geq 0, \forall i \end{cases} \quad (5)$$

The formula (5) is transformed into the dual formula of standard support vector machine by using kernel combination  $K(x_i, x_j) = \sum_{m=1}^M s_m K_m(x_i, x_j)$ .  $J(s)$  is used to represent the optimal objective function of cross-source classification of multisource network data. Multicore learning has strong duality, so  $J(s)$  can be used as the

objective function of the dual problem of formula (5) at the same time. Select the descending direction  $S$  of the gradient descent method and update  $s$  through the obtained  $J(s)$  gradient. The updating process is  $s \leftarrow s + \gamma S$ , where  $\gamma$  represents the step size. It is necessary to search the maximum allowable step according to the descent direction  $S$  to judge whether there is a decrease in the target value. When the target value decreases,  $s$  needs to be updated. Repeat the above process until the target value does not decrease.

*2.2.2. Support Vector Machine Classification of Multicore Learning.* Support vector machine needs to use the combination of multiple binary classifiers to solve the multiclass classification problem. Suppose that the category in the massive multisource network data is  $q$ , and the number of binary classification tasks is  $q(q-1)/2$  through pairing the categories with the number of  $q$ . Send the multisource network data to the trained support vector machine classifier, and finally obtain the classification results with the number of  $q(q-1)/2$ .

Use  $J(s)$  to represent the target value of cross-source classification of multisource network data, search for the kernel combination that can jointly optimize all decision functions as an even number [17], and obtain the objective function of optimizing cross-source classification of multisource network data according to kernel weight  $\{s_m\}$  as follows:

$$J(s) = \sum_{q \in Q} J_q(s). \quad (6)$$

In formula (6),  $Q$  and  $J_q(s)$ , respectively, represent all binary classifier sets to be considered and the target value of support vector machine for binary classification problems related to binary classifier. The Lagrange multipliers of each binary classification problem are obtained according to the gradient descent method, and the kernel combination of all binary classification problems is obtained, that is, the sum of maximized intervals, so as to realize the cross-source classification of multisource network data.

### 3. Build a Big Data Classification Platform with Low Occupancy

The low occupancy big data classification platform based on grid technology is composed of four parts: advanced access layer, application interface layer, basic management layer, and infrastructure layer, as shown in Figure 1.

Using the open grid services architecture (OGSA) and GT3 toolkit, build the infrastructure layer of the low occupancy big data private cloud platform on the basis of ensuring the low occupancy big data system structure. Develop the basic management layer, application interface layer, and advanced access layer according to the actual application requirements of low occupancy big data storage.

The infrastructure layer is the foundation of the physical and storage devices in the low occupancy big data system classification platform, which mainly includes four parts: grid system management node, grid public service node, local resource

server, and enterprise grid application service [7, 18]. Each part of the equipment is scattered, and some types of server equipment have a large number and scattered locations. Therefore, WAN connection is adopted between different server devices [19].

The main function of the basic management layer is to use integrated or distributed control to manage the collaborative operation of all storage devices in the low occupancy big data private cloud, complete the optimal utilization of resources, and build a work contradiction mechanism for problems such as workflow contradiction. At the same time, this layer is also responsible for data backup, data encryption, and data disaster recovery in the process of low occupancy big data classification.

The application interface layer includes all interfaces used by the low occupancy big data classification platform to connect with other equipment and servers and provides interfaces and corresponding services to power management institutions at all levels according to the actual application requirements and corresponding levels of different power enterprises. The user can successfully log into the platform by inputting the corresponding account password through the public interface of the low occupancy big data classification platform and collect the corresponding data resources according to the account level authority.

The main function of the advanced access layer is to provide the low occupancy big data classification platform with the interface required for operation. On the basis of ensuring the primary application of low occupancy big data classification, the advanced application is developed according to the actual application requirements of power enterprises to realize the interpersonal interaction function.

#### 3.1. Hardware Design

*3.1.1. Infrastructure Layer.* Grid technology has a wide range of applications and can be effectively deployed on LAN, Wan, and Internet [20]. In order to meet the internal low occupancy big data sharing of power enterprises, a low occupancy big data network including shared areas is established by using grid technology. The application of grid technology can ensure the safe access and sharing of different low occupancy big data resources in the grid system.

The high-speed network inside the power enterprise is used to connect different application system servers, database servers, storage backup servers, and other servers. Set up the grid operating system, set up professional grid scheduling servers, registration servers, server pools, and other devices, and integrate these servers and devices into a grid system to build a low occupancy big data grid. Using the open grid service architecture and GT3 toolkit, build the infrastructure layer of the low occupancy big data private cloud platform on the basis of ensuring the low occupancy big data system structure. The infrastructure layer grid is shown in Figure 2.

The infrastructure layer designed by using grid technology includes grid basic services provided by grid system management node and grid public service node, grid

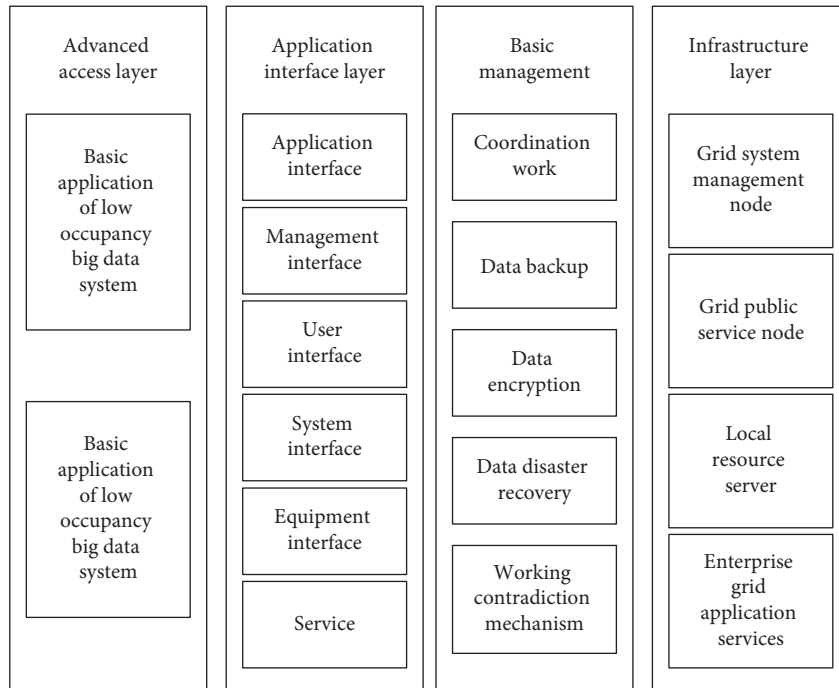


FIGURE 1: Overall structure of low occupancy big data classification platform.

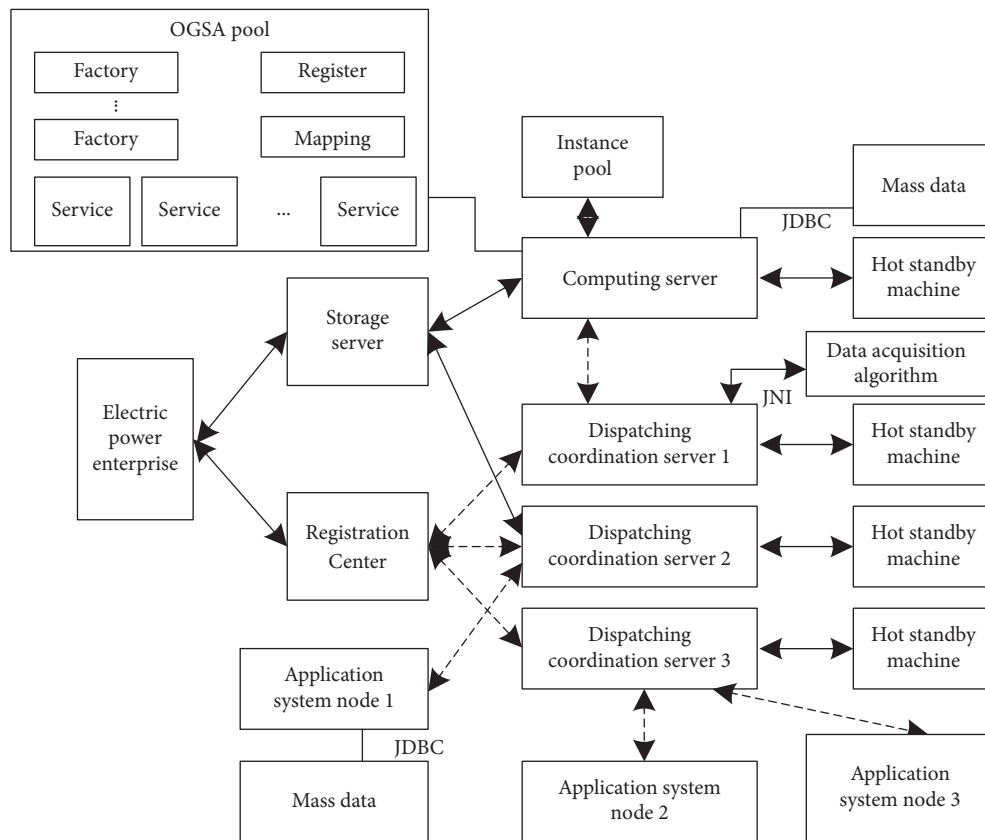


FIGURE 2: Infrastructure layer construction based on grid technology.

application services provided by local resource server, and enterprise grid application services.

Dispatching coordination server, registration server, and other servers and equipment that realize basic grid technology

form a grid system management node, and different servers are equipped with hot standby machines to ensure their uninterrupted operation [21, 22]. The main function of each server in the grid system management node is to provide basic grid

services and ensure the effective work of the grid system. Dispatching coordination server is the main component of grid technology and is responsible for controlling and coordinating all grid operations and services.

Storage server, computing server, and instance pool form a grid common service node [23]. The main function of the storage server is to provide data storage services in the grid, which has the characteristics of high capacity and reliability. The computing server usually adopts high-performance computer, which has high-precision computing performance. The instance pool contains different services in the grid, such as web services (the basis for building grid services) and public services.

As the infrastructure of grid technology, local resource server is mainly composed of heterogeneous systems such as application system and automatic office system [24].

Enterprise grid application service is an advanced service that defines and implements different grid applications according to the needs of power enterprises based on the foundation and public services provided by public service nodes. Based on the concept of grid, different computing resources and storage resources in low occupancy big data grid can be realized through grid services.

**3.1.2. Basic Management.** The basic management layer includes various functional modules applied in the process of low occupancy big data classification, such as load prediction module, image backup module, data security module, physical host module, etc. The structure hierarchy of each functional module is shown in Figure 3.

**3.1.3. Load Forecasting Layer.** Based on the application requirements of low occupancy big data classification platform, it is necessary not only to provide the historical load change trend of virtual machine (storage server) to power enterprises, but also to display the load prediction results of virtual machine to power enterprises, so as to meet the purpose of diversified application of low occupancy big data resources [25, 26]. Using load forecasting, you can apply for an appropriate amount of virtual machines to balance the load before the virtual machine load reaches the upper limit [27], so as to alleviate the access pressure. When the virtual machine load forecast is lower than a certain load value within a certain time, the virtual machine can be recovered to improve the utilization rate of the virtual machine and meet the requirements of environmental protection.

Load forecasting is based on the load monitoring results. After preliminary processing, the load monitoring results are calculated by BP neural network [28], and the prediction results are obtained. The prediction results are stored in the database and displayed through the interface in the advanced access layer. The load forecasting process is shown in Figure 4.

**3.1.4. Image Backup Module.** The main function of the image backup module is to backup each virtual machine in the low occupancy big data classification platform within a fixed time under the control of the timer, so as to restore the virtual machine in case the platform cannot run due to an accident. The image backup process is shown in Figure 5.

In the process of image backup, it is necessary to judge whether the virtual machine has image backup. If so, delete the old image backup and build a new image backup to ensure the uniqueness of image backup.

**3.2. Software Optimization.** Based on the above hardware platform, the low occupancy data after part 2 training and classification are input into the deep confidence network to optimize the software part of the low occupancy big data classification platform.

**3.2.1. RBM Training.** In the process of classifying big data with low occupancy in complex scenes by using depth confidence network, the depth confidence network is obtained by stacking several RBM (restricted Boltzmann machine) layers [29], and each RBM layer is trained separately by contrast divergence method.

RBM includes visible layer and hidden layer, which are represented by  $v$  and  $h$ , respectively. The connection weight only exists in  $v$  and  $h$  nodes, and there is no connection weight between nodes of each layer. Under the condition that  $v$  is known, all  $h$  nodes have conditional independence. When  $v$  is input,  $h$  can be obtained based on conditional probability, and  $v$  can also be obtained according to  $h$ . Under the condition of optimizing the internal parameters of RBM, if the  $v'$  obtained by  $h$  is the same as the initial  $v$ , it indicates that the obtained  $h$  is another description form of  $v$ .

As a standard energy model, the energy function of RBM can be described by

$$E(v, h) = -\sum_{i=1}^m \sum_{j=1}^n w_{ij} v_i h_j - \sum_{i=1}^m a_i v_i - \sum_{j=1}^n b_j h_j. \quad (7)$$

In formula (7),  $w_{ij}$ ,  $a_i$ , and  $b_j$  represent the weight between  $E$  and  $R$  and the offset between them, respectively. Based on the joint configuration energy function shown in formula (7), the joint probability [30, 31] of  $v$  and  $h$  can be obtained. The formula is described as follows:

$$P(v, h) = \frac{e^{-E(v, h)}}{\sum_{v, h} e^{-E(v, h)}}. \quad (8)$$

Since there is no connection between nodes in the layer in RBM, the conditional probability can be obtained according to the joint probability. The formula is described as follows:

$$P(v) = \frac{\sum_h e^{-E(v, h)}}{\sum_{v, h} e^{-E(v, h)}},$$

$$P(h) = \frac{\sum_v e^{-E(v, h)}}{\sum_{v, h} e^{-E(v, h)}}, \quad (9)$$

$$P(v|h) = \frac{e^{-E(v, h)}}{\sum_v e^{-E(v, h)}},$$

$$P(h|v) = \frac{e^{-E(v, h)}}{\sum_h e^{-E(v, h)}}.$$

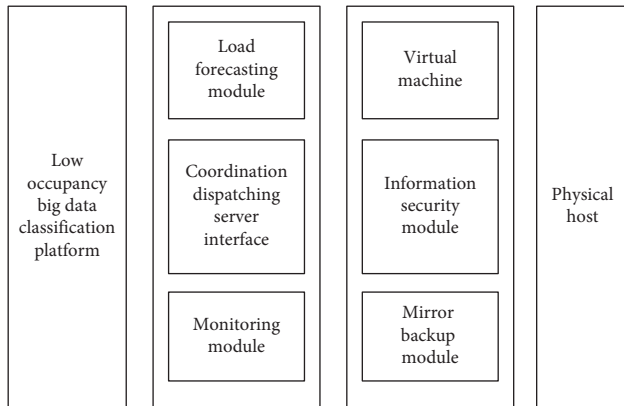


FIGURE 3: Structural hierarchy diagram of each functional module.

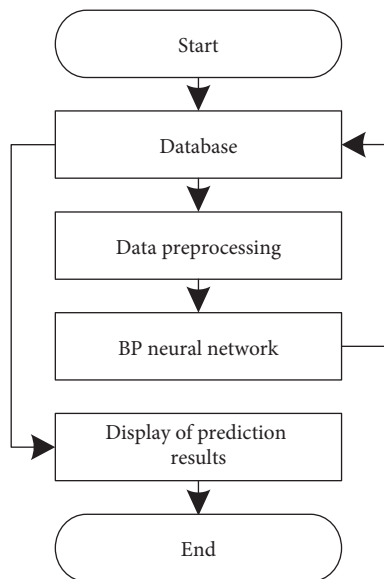


FIGURE 4: Load forecasting process.

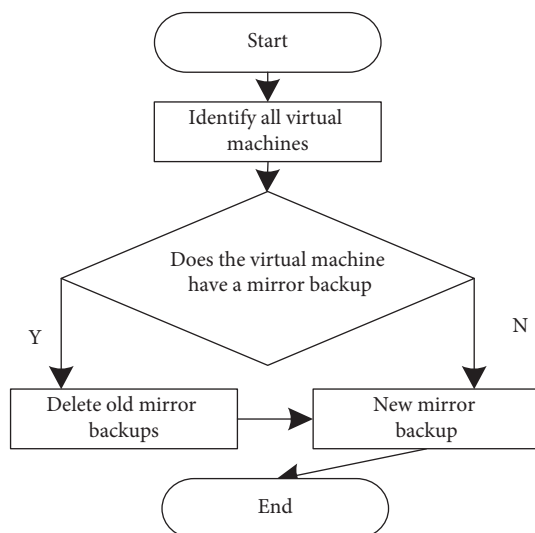


FIGURE 5: Image backup process.

After obtaining the conditional probability of RBM network, in order to make the Gibbs distribution described infinitely approximate to the fitting input data [32], it is necessary to learn  $(w, a, b)$  and other parameters. Generally, the parameter solution is obtained by solving the upper likelihood limit of the input sample.

In view of the long training time, the contrast divergence method can be used to improve the training efficiency. In this method, the gradient of the number likelihood function is solved by two approximations: (1) the average summation in the gradient calculation process is approximately replaced by the samples obtained in the conditional distribution; (2) Gibbs sampling was performed only once. The algorithm based on contrast divergence can obtain the optimization results of  $(w, a, b)$  and other parameters at a faster speed, so as to realize RBM network training.

**3.2.2. Classification Process Based on Deep Confidence Network.** The main feature of the deep confidence network is that it requires a large number of training samples. Therefore, when using the deep confidence network to divide the categories of low occupancy big data, it is necessary to reconstruct the data, expand the samples, and reduce the dimension of the samples [33].

When expanding samples, new samples can be obtained by adding adjacent to each other to find the mean. This method not only improves the number of training samples, but also considers the spatial correlation of samples. The specific process is described as follows:

Calculate the adjacent sum of each and its surrounding four directions, and divide the sum by 3 to obtain 4 times the number of new samples in the original low occupancy big data.

The same category of big data with low occupancy rate is relatively unified. Therefore, optimizing the input data based on spatial combination can greatly improve the classification accuracy of deep confidence network. If the input is simply added to the surrounding neighbors through training, the dimension and redundancy will be improved.

In order to reduce the dimension on the basis of spatial combination, principal component analysis or self-encoder is used to reduce the dimension of object metadata. The former method mainly reduces the dimension of linear data and the latter method mainly reduces the dimension of unstructured data. Compared with the two dimensionality reduction methods, the self-encoder can maximize the protection of the characteristics of low occupancy big data metadata, while the principal component analysis method can describe the data through fewer dimensions [34, 35].

The image after expanding the sample data and dimensionality reduction is used as the input of the first RBM visible layer of the depth confidence network, and several training depth confidence networks are randomly determined.

RBM layer training is to obtain parameter values through iteration, so as to describe low occupancy big data in other ways. The activation probability of the  $j$ -th node of the hidden layer in RBM can be obtained through formulas (8)–(10). The formula is described as follows:

$$P(h_j = 1|v) = \frac{1}{1 + \exp\left(-\sum_i w_{ij}v_i - b_j\right)}. \quad (10)$$

Meanwhile, under the condition that  $h$  is known, the activation probability of the  $i$ -th node can be described by

$$P(v_i = 1|h) = \frac{1}{1 + \exp\left(-\sum_j w_{ij}h_j - a_i\right)}. \quad (11)$$

The training is carried out according to the contrast divergence method, and the  $(w, a, b)$  isoparametric value in the RBM layer can be obtained after several iterations. Under this condition,  $h_1$  is another way to describe big data with low occupancy.

Taking  $h_1$  as the RBM input of the next layer and implementing iterative training in the same way,  $h_2, h_3, \dots$  will be obtained, and the corresponding  $(w, a, b)$  isoparametric values in each layer can be obtained. The pre-training process of the deep confidence network model can be realized through the layer-by-layer training process.

In order to obtain higher classification accuracy, a BP layer optimization parameter is introduced after the last hidden layer. If  $d$  and  $y^l$  are used to represent the expected classification result and the final hidden layer output result, respectively, the residual of the two is determined by

$$\delta^{(l)} = -(d - y^l)y^l(1 - y^l), \quad (12)$$

$\delta^{(l)}$  is transmitted from back to front. In each layer of  $l = n^l - 1, n^l - 2, \dots$ ,  $\delta^{(l)}$  is determined as

$$\delta^{(l)} = (w^{(l)})^T \delta^{(l+1)} y^l (1 - y^l). \quad (13)$$

Determine the partial derivatives of  $w$  and  $b$  according to  $\delta^{(l)}$  of each layer:

$$\begin{aligned} \nabla w^{(l)} J &= y^l \delta^{(l+1)}, \\ \nabla b^{(l)} J &= \delta^{(l+1)}. \end{aligned} \quad (14)$$

After obtaining the partial derivative, the weights of  $w$  and  $b$  can be updated. After several iterations, the depth confidence network model optimization after pretraining can be realized, and the low occupancy big data classification method based on grid technology can be realized.

## 4. Experimental Analysis

In order to verify the effectiveness of the low occupancy big data classification method based on grid technology, taking the low occupancy data of an enterprise network in China as the research object, the low occupancy big data classification test was carried out using this method. The test results are as follows.

**4.1. Experimental Environment and Parameter Setting.** The experimental environment is the hardware environment of Intel Celeron Turing 1 GHz CPU and 384 mb SD memory

TABLE 1: Parameter setting.

Parameter name	Numerical value
Threshold	—
Initial data	One
Node communication radius	11 m
Number of iterations	19 times
Test duration	50 $\mu$ m
Transmission speed	15 Gb/s
Data overhead size	7984B

and the software environment of Matlab6.1. The simulation system includes data interference module, resource scheduling module, and task generation module. Based on the composition of the above parts, ms-coco data set is used as the data acquisition source, and 1000 task data are arbitrarily taken in the data set. At the same time, the task data is proposed to be stored in the simulation system, and the data bytes are controlled between [256~568 kb]. The experiment termination condition was set to 300 times. The simulation algorithm parameters in this paper are shown in Table 1.

**4.2. Virtual Machine Load Prediction Results.** Using this method, according to the load of the research object in the past 30 days, the load trend in the next 5 days is predicted and compared with the current load. The results are shown in Figure 6.

According to the analysis of Figure 6, the load prediction results of this method in the storage platform during the implementation of low occupancy big data classification are basically consistent with the actual load, which shows that this method can accurately predict the load in the low occupancy big data storage process, make use of load balancing control, and meet the actual application requirements of the research object.

**4.3. Load Balancing Test.** The experiment verifies the load balancing performance of this method in the process of low occupancy big data classification with two indicators of load balancing and response time. The results are shown in Figure 7.

By analyzing Figure 7(a), it can be seen that the load of different virtual machines is less than 40% and the load of each virtual machine is basically the same in the process of low occupancy big data classification of the research object by using the method in this paper. By analyzing Figure 7(b), it can be seen that the response time of different virtual machines in the process of big data classification with low occupancy is controlled between 0.5 s and 0.7 s. This shows that the load balancing degree is high in the process of low occupancy big data classification, and the classification efficiency can be significantly improved through load balancing.

**4.4. Storage Synchronization Test.** Storage synchronization is one of the key indexes to evaluate the performance of storage methods. The storage performance of this method is verified



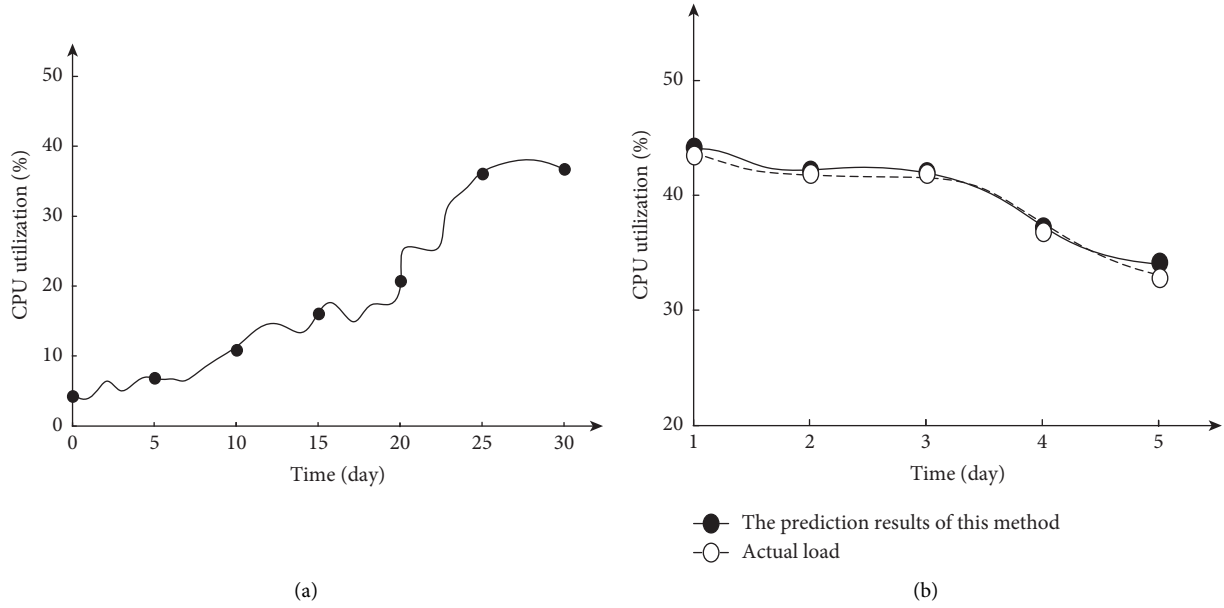


FIGURE 6: Load prediction results of this method. (a) Load in the last 30 days. (b) Forecast results for the next five days.

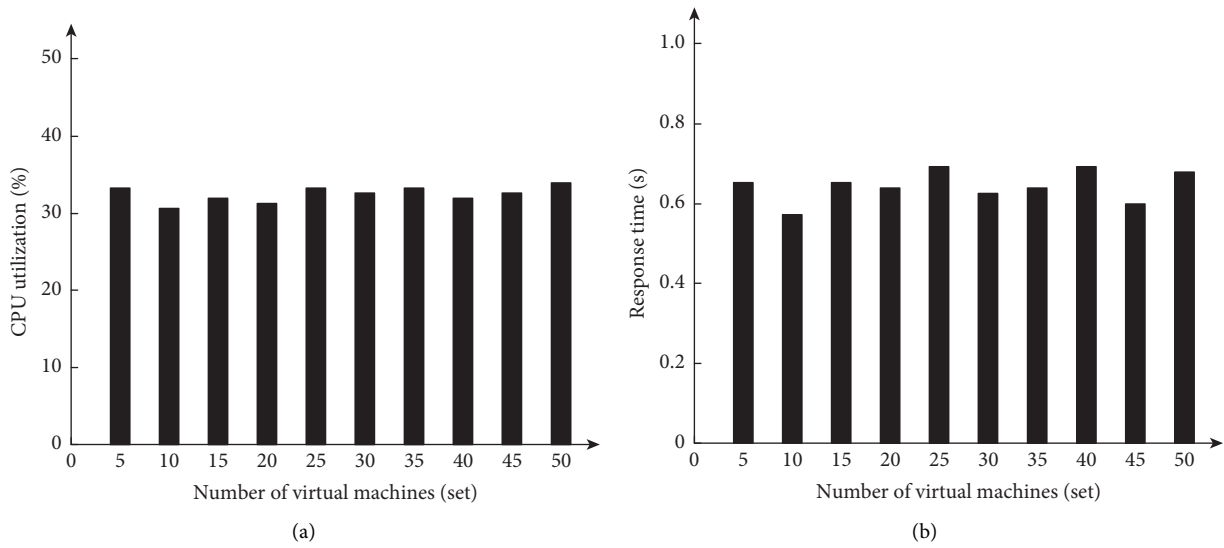


FIGURE 7: Load balancing test results. (a) Load balancing. (b) Response time.

by taking the storage synchronization as the index, and the results are shown in Table 2.

The calculation formula of bit error rate  $W$  is as follows:

$$W = \frac{M}{M_z} \times 100\%. \quad (15)$$

In the above formula,  $M$  and  $M_z$ , respectively, represent the error code and the total number of codes input in the process of power application data input.

Analysis of Table 2 shows that when the low occupancy big data input frequency gradually increases, the synchronization bit error rate of this method under the condition of different number of virtual machines shows an upward trend with the increase of data input frequency. When the number

of virtual machines is 10, the synchronization error rate of this method basically maintains a linear upward trend in the low occupancy big data input frequency. When the input data frequency reaches 100 Hz, the synchronization error rate of this method is 0.013%. When the number of virtual machines rises to 30, the synchronization error rate of the method in this paper is basically the same as that under the condition of 10 virtual machines. When the number of virtual machines increases to 50, the synchronization error rate of this method increases obviously, and when the data input frequency is less than 70 Hz, the synchronization error rate fluctuation of this method is small. When the data input frequency reaches more than 80 Hz, the synchronization error rate of this method increases rapidly. When the input

TABLE 2: Storage synchronization test results.

Low occupancy big data input frequency/Hz	Synchronous bit error rate/%		
	10 virtual machines	30 virtual machines	50 virtual machines
10	0.004	0.004	0.013
20	0.006	0.005	0.013
30	0.007	0.005	0.01
40	0.006	0.007	0.012
50	0.008	0.008	0.014
60	0.008	0.009	0.013
70	0.008	0.01	0.014
80	0.01	0.12	0.063
90	0.012	0.012	0.078
100	0.013	0.014	0.089

TABLE 3: Comparison results of storage capacity.

Storage time (s)	Storage capacity (W)	
	Before using this method	After adopting the method in this paper
10	372	456
20	349	461
30	361	459
40	377	452
50	360	448
60	355	451
70	373	450
80	348	446
90	380	453
100	370	460
110	352	452
120	389	457

data frequency reaches 100 Hz, the synchronization error rate of this method is 0.089%. This shows that the number of virtual machines and low occupancy big data input frequency have a significant impact on the synchronization bit error rate of this method. When the number of virtual machines is 30, it can ensure that the synchronization bit error rate is low and meets the application needs of power enterprises.

*4.5. Storage Comparison.* Storage capacity is one of the main indicators to verify the data storage method. Under the same network environment and hardware facility environment, compare the storage capacity of the research object in the low occupancy big data storage process before and after using this method. The results are shown in Table 3.

According to the analysis in Table 3, from the perspective of average storage capacity, the average storage capacity of the research object data before using this method is 365.50 W/s, and after using this method, the average storage capacity increases by 88.25 W/s. From the perspective of storage volume fluctuation, the data storage volume in the low occupancy big data storage process has obvious fluctuation before using this method. After using this method, the data storage volume in the low occupancy big data storage process is relatively stable. The experimental results

show that this method can effectively improve the storage capacity of large data with low occupancy and optimize the storage process.

## 5. Conclusion

High-precision classification of low occupancy data helps to achieve efficient network transmission and improve data utilization. Therefore, aiming at the problems of low efficiency and large space of intelligent classification of low occupancy big data, this paper proposes an intelligent classification method of low occupancy big data based on grid index. The main contents and process of this paper are as follows:

- (1) Feature selection of big data is done by constructing intelligent classification model.
- (2) Build a low occupancy big data intelligent classification platform based on grid index technology.
- (3) The constructed depth confidence network model is trained, and the selected features are intelligently classified.
- (4) Simulation results show that this method has the advantages of high classification efficiency and low occupancy in the intelligent classification of low occupancy big data based on grid index.

## Data Availability

The dataset used in this paper can be accessed upon request.

## Conflicts of Interest

The authors declare no conflicts of interest.

## Acknowledgments

This work was supported by Tianjin Major Scientific and Technological Research Plan under Grant 16ZXHLSF00160; Industry University Cooperative Collaborative Education Project of Higher Education Department of the Ministry of Education under Grant 202101012003; Tianjin Education Commission Scientific Research Project of Mental Health Education Special Task under Grant 2020ZXXL-GX45.

## References

- [1] X. Q. Yang and Y. Zhu, "Research on-line monitoring method of noise pollution in robot master-slave control system," *Computer Simulation*, vol. 38, no. 6, pp. 301–305, 2021.
- [2] S. Salimi and A. Hammad, "Sensitivity analysis of probabilistic occupancy prediction model using big data," *Building and Environment*, vol. 172, Article ID 106729, 2020.
- [3] S. A. Petersen, Z. Pourzolfaghar, I. Alloush, D. Ahlers, and M. Helfert, "Value-added services, virtual enterprises and data spaces inspired enterprise architecture for smart cities," in *Proceedings of the Collaborative Networks and Digital Transformation-20th IFIP WG 5.5 Working Conference on Virtual Enterprises, PRO-VE 2019*, Turin, Italy, September 2019.
- [4] A. S. Cabezuelo, "Using open data repositories and geo-location to create value-added services for tourism," in *Proceedings of the 2020 24th International Conference Information Visualisation (IV)*, Melbourne, Australia, September 2020.
- [5] V. Yesin, M. Karpinski, M. Yesina, V. Vilihura, and K. Warwas, "Hiding the source code of stored database programs," *Information*, vol. 11, no. 12, p. 576, 2020.
- [6] W. Wang, T. Hong, N. Li, R. Q. Wang, and J. Chen, "Linking energy-cyber-physical systems with occupancy prediction and interpretation through wifi probe-based ensemble classification," *Applied Energy*, vol. 236, no. FEB.15, pp. 55–69, 2019.
- [7] T. Khan and A. F. Mollah, "Text non-text classification based on area occupancy of equidistant pixels," *Procedia Computer Science*, vol. 167, 2020.
- [8] G. Violatto and A. Pandharipande, "Anomaly classification in people counting and occupancy sensor systems," *IEEE Sensors Journal*, vol. 20, p. 1, 2020.
- [9] G. Ravikumar, M. Govindarasu, S. Misra, S. Brahma, and R. Tourani, "Anomaly detection and mitigation for wide-area damping control using machine learning," *IEEE Transactions on Smart Grid*, vol. 18, no. 5, p. 1, 2020.
- [10] M. Mohamma Dd Oust, A. Harounabadi, and M. Neizari, "Evaluation of grid computing environment using topsi," *The International arab Journal of Information Technology*, vol. 16, no. 3, pp. 323–331, 2019.
- [11] D. Toshniwal and M. Dipasquale, "Counting the dimension of splines of mixed smoothness: a general recipe, and its application to planar meshes of arbitrary topologies," *Advances in Computational Mathematics*, vol. 47, no. 1, 2021.
- [12] S. Zhu, D. Wang, L. Liu, W. Yan, and D. Guo, "Inferring region significance by using multi-source spatial data," *Neural Computing & Applications*, vol. 32, no. 22, 2020.
- [13] Z. Rustam, J. Pandelaki, D. A. Utami, R. Hidayat, and A. A. Ramli, "Comparison support vector machine and fuzzy possibilistic c-means based on the kernel for knee osteoarthritis data classification," *International Journal of Advanced Science, Engineering and Information Technology*, vol. 9, no. 6, pp. 2142–2146, 2019.
- [14] A. A. Ribeiro and M. Sachine, "On the optimal separating hyperplane for arbitrary sets: a generalization of the svm formulation and a convex hull approach," *Optimization*, vol. 8, no. 10, pp. 1–14, 2020.
- [15] N. Ali, D. Neagu, and P. Trundle, "Evaluation of k-nearest neighbour classifier performance for heterogeneous data sets," *SN Applied Sciences*, vol. 1, no. 12, pp. 1–15, 2019.
- [16] R. P. Monteiro, M. Cerrada, D. R. Cabrera, R. V. Sánchez, and C. J. A. Bastos-Filho, "Using a support vector machine based decision stage to improve the fault diagnosis on gearboxes," *Computational Intelligence and Neuroscience*, vol. 2019, Article ID 1383752, 13 pages, 2019.
- [17] M. T. Sattari, H. Feizi, M. . Sevba olak et al., "Estimation of sodium adsorption ratio in a river with kernel-based and decision-tree models," *Environmental Monitoring and Assessment*, vol. 192, no. 9, p. 575, 2020.
- [18] E. Hugues-Salas, F. Ntavou, D. Gkounis, G. T. Kanellos, R. Nejabati, and D. Simeonidou, "Monitoring and physical-layer attack mitigation in sdn-controlled quantum key distribution networks," *Journal of Optical Communications and Networking*, vol. 11, no. 2, p. A209, 2019.
- [19] W. Huang, J. Fang, S. Wan, C. Xie, and X. He, "Design and evaluation of a risk-aware failure identification scheme for improved ras in erasure-coded data centers," *IEEE Transactions on Parallel and Distributed Systems*, vol. 32, 2020.
- [20] S. S. Hameed, W. H. Hassan, L. Abdul Latiff, and F. Ghabban, "A systematic review of security and privacy issues in the internet of medical things; the role of machine learning approaches," *PeerJ Computer Science*, vol. 7, no. 4, p. e414, 2021.
- [21] N. Sridhar and M. Kowsalya, "Enhancement of power management in micro grid system using adaptive alo technique," *Journal of Ambient Intelligence and Humanized Computing*, vol. 12, no. 2, pp. 2163–2182, 2021.
- [22] J. Kiljander, D. Gabrijelcic, O. Werner-Kytola, A. Krpic, and M. Taumberger, "Residential flexibility management: a case study in distribution networks," *IEEE Access*, vol. 7, 2019.
- [23] O. Khutsoane, B. Isong, N. Gasela, and A. M. Abu-Mahfouz, "Watergrid-sense: a lora-based sensor node for industrial iot applications," *IEEE Sensors Journal*, vol. 20, 2019.
- [24] I. González, A. José Calderón, and J. María Portalo, "Innovative multi-layered architecture for heterogeneous automation and monitoring systems: application case of a photovoltaic smart microgrid," *Sustainability*, vol. 13, no. 4, p. 2234, 2021.
- [25] Y. Kim, U. Shahab, and S. H. Bae, "Local augment: utilizing local bias property of convolutional neural networks for data augmentation," *IEEE Access*, vol. 9, 2021.
- [26] G. S. Krishna and N. Prakash, "Deep learning for efficient and multi-labelled classification of synthetic aperture radar images," *Evolving Systems*, vol. 12, no. 1, pp. 1–14, 2021.
- [27] W. Guo, W. Ge, X. Lu, and H. Li, "Short-term load forecasting of virtual machines based on improved neural network," *IEEE Access*, vol. 7, 2019.
- [28] Q. Liu, S. Liu, G. Wang, and S. Xia, "Social relationship prediction across networks using tri-training bp neural networks," *Neurocomputing*, vol. 401, pp. 377–391, 2020.

- [29] G. Wang, J. Qiao, J. Bi, Q. S. Jia, and M. C. Zhou, "An adaptive deep belief network with sparse restricted Boltzmann machines," *IEEE Transactions on Neural Networks and Learning Systems*, vol. 31, pp. 1–12, 2019.
- [30] M. Kakimoto, Y. Endoh, H. Shin, R. Ikeda, and H. Kusaka, "Probabilistic solar irradiance forecasting by conditioning joint probability method and its application to electric power trading," *IEEE Transactions on Sustainable Energy*, vol. 10, no. 2, pp. 983–993, 2019.
- [31] K. Qiu and Q. Zhang, "The m-dug methodology to calculate the joint probability distribution of directed cycle graph with local data and domain causal knowledge," *IEEE Access*, vol. 9, 2021.
- [32] J. Cabessa and O. Finkel, "Computational capabilities of analog and evolving neural networks over infinite input streams," *Journal of Computer and System Sciences*, vol. 101, no. MAY, pp. 86–99, 2019.
- [33] K. R. M. Fernando and C. P. Tsokos, "Dynamically weighted balanced loss: class imbalanced learning and confidence calibration of deep neural networks," *IEEE Transactions on Neural Networks and Learning Systems*, vol. 14, no. 1, pp. 1–12, 2021.
- [34] A. K. Seghouane, N. Shokouhi, and I. Koch, "Sparse principal component analysis with preserved sparsity pattern," *IEEE Transactions on Image Processing*, vol. 28, 2019.
- [35] N. Gilbert, R. E. Mewis, and O. B. Sutcliffe, "Classification of fentanyl analogues through principal component analysis (PCA) and hierarchical clustering of GC-MS data," *Forensic Chemistry*, vol. 21, Article ID 100287, 2020.