*processes*

*Article*

# Anomaly Recognition, Diagnosis and Prediction of Massive Data Flow Based on Time-GAN and DBSCAN for Power Dispatching Automation System

**Wenjie Liu, Pengfei Lei, Dong Xu and Xiaorong Zhu ***

College of Telecommunications and Information Engineering, Nanjing University of Posts and Telecommunications, Nanjing 210003, China; 1221013822@njupt.edu.cn (W.L.); 1222014217@njupt.edu.cn (P.L.); 1022010208@njupt.edu.cn (D.X.)
* Correspondence: xrzhu@njupt.edu.cn

**Abstract:** Existing power anomaly detection is mainly based on analyzing static offline data. But this method takes a long time and has low identification accuracy when detecting timing and frequency anomalies, since this method requires offline screening, classification and preprocessing of the collected data, which is very laborious. Anomaly detection with supervised learning requires a large amount of abnormal data and cannot detect unknown anomalies. So, this paper innovatively proposes the idea of applying Time-series Generative Adversarial Networks (Time-GAN) in a dispatching automation system for the identification, diagnosis and prediction of massive data flow anomalies. First of all, regarding the problem of insufficient abnormal data, we use Time-GAN to generate a large number of reliable datasets for training fault diagnosis models. In addition, Time-GAN can ameliorate the imbalance between various types of data. Secondly, unsupervised learning methods such as Density-Based Spatial Clustering of Applications with Noise (DBSCAN) and K-means are used to detect unknown anomalies that may exist in the power grid. Finally, some supervised learning methods are selected to compare with unsupervised learning methods. Experimental results show that the proposed algorithm has a higher recognition rate of known anomalies than other benchmark algorithms and it can find new unknown anomalies. It lays a good foundation for the safe, stable, high-quality and economical operation of the power grid.

**Keywords:** Time-GAN; DBSCAN; supervised learning; fault diagnosis; fault prediction

**Citation:** Liu, W.; Lei, P.; Xu, D.; Zhu, X. Anomaly Recognition, Diagnosis and Prediction of Massive Data Flow Based on Time-GAN and DBSCAN for Power Dispatching Automation System. *Processes* **2023**, *11*, 2782. https://doi.org/10.3390/pr11092782

Academic Editor: Olympia Roeva

Received: 27 July 2023
Revised: 30 August 2023
Accepted: 11 September 2023
Published: 18 September 2023

## 1. Introduction

With the rapid expansion of the power grid's construction, various dispatching automation systems have successively built upon power dispatching data networks. The era of big data has officially arrived. In order to meet the needs of the smart grid, the intelligent dispatching automation system should have a powerful, intelligent and early-warning function, and we should pay attention to the coordination of system safety and economy in dispatching decisions. When the system fails, it can quickly diagnose faults and provide fault recovery decisions. The current anomaly detection methods for scheduling data mainly include methods such as simple threshold judgment based on a single system and analysis methods based on static offline data. The simple threshold judgment method based on a single system has limitations. On one hand, the utilization rate of equipment information and the accuracy of status evaluation are low. On the other hand, it is difficult to detect the latent faults and fault categories of the equipment.

In recent years, domestic and foreign scholars have actively explored anomaly detection of the smart grid based on machine learning and they have achieved certain results. Reference [1] uses the anomaly detection method based on support vector machine to diagnose the abnormality according to the characteristics of the power system's data and it achieves a higher efficiency compared to traditional methods. Due to the rapid expansion of the power grid's scale, power monitoring data have the characteristics of sequential and rapid and have continuous arrival time series. To address this, some scholars have carried out research on the time series problems of power systems. Yao et al. [2] combine the Convolutional Neural Networks (CNN) and Support Vector Machine (SVM) models to detect abnormal consumption behaviors of users and achieve better results compared to traditional methods. Yang et al. [3] use Long Short Term Memory (LSTM) to extract features based on the traffic anomaly detection method of the Light Gradient Boosting Machine (LightGBM) and LSTM models. Experimental results show that this method has a higher anomaly classification accuracy and a higher anomaly detection accuracy. Liu et al. [4] use LSTM to extract features and adopt the traffic anomaly detection method based on the improved SVM embedding decision tree model. Compared with the traditional method, it has a higher accuracy. Authors in [5] use the self-organizing neural network to quantify the historical data of power transmission and transformation equipment, mine the data changing over time and use the auto regressive model to establish an anomaly model to achieve the goal of high detection ratio and low false warning ratio.

The above-mentioned references mainly detect known abnormal features, but there are still many faults in electric power that cannot be analyzed for specific reasons and there are still many abnormalities that are unknown, so unsupervised learning is required. Unsupervised learning is a learning method that learns patterns from raw data without the help of labels. Unlike supervised learning, which artificially specifies labels of data categories under prior knowledge, unsupervised learning can discover the inherent connections or structures contained in the data. In recent years, domestic and foreign scholars have achieved some research results in the application of clustering algorithms to identify anomalies in network traffic in power systems. Huang et al. [6] use the Canopy-K-means algorithm to cluster the traffic data in the key business system of electric power to identify attacking traffic and business traffic. Attacking traffic refers to network packets sent by illegitimate IP addresses, including attacks on important control systems. The authors in [7] use the DBSCAN algorithm to establish a power transformer fault diagnosis model and take the typical oil chromatography data of various fault types as the input of the model to obtain typical clusters of various faults. Wang et al. [8] apply the method of fuzzy clustering to realize the fusion of multiple expert diagnosis results in the process of power system fault diagnosis and achieve a faster fault diagnosis result. Dong et al. [9] use methods such as hierarchical clustering, K-means and DBSCAN to improve the detection rate of wireless network intrusion detection methods, to reduce false detection rates and to improve the overall performance of intrusion detection systems. Jian et al. [10] use the hierarchical clustering method to construct an abnormal traffic model, which improves the detection efficiency of attacking traffic in the network. Compared with supervised learning, unsupervised learning methods such as DBSCAN and K-means can distinguish abnormal data from normal data and the separated anomalies may contain many unknown anomalies, which are of great help for us to explore the possible unknown anomalies in electric power.

When tackling anomaly detection problems in the power grid, some following challenges should be well met. Firstly, the traditional power grid fault diagnosis methods are not accurate enough. What's more, it is difficult to find hidden dangers and identification of grid faults will also rely on more Key Performance Indicators (KPI). So, one objective for us is to improve the accuracy of the detection of unknown anomalies. Secondly, existing works lack data support and require a large number of manually trained datasets, which is very time-consuming. So, how to expand the sample of a small number of packets so as to obtain a large number of reliable datasets in such a complex network environment needs to be well addressed. Finally, messages in the power grid have various types, like control message, signaling message, telemetry message and call message. And different types of messages have different generation speed. So, we need to classify messages according to their types and ensure the sample balance of various types of power data.

Among all those challenges, how to enlarge training data used in unsupervised learning is fundamental and Generative Adversarial Network (GAN) is to be adopted. As a typical method of artificial intelligence, GAN consists of two independent deep networks [11], namely, generator and discriminator, where generator is used to generate samples and discriminator is used to classify samples when training GAN. Using this method to recognize the Mixed National Institute of Standards and Technology (MNIST) handwriting dataset, experimental results demonstrate the potential of this framework. However, it still has some problems, such as difficulties in training and lack of diverse generated samples. To address these problems, references [12,13] propose Wasserstein GAN (WGAN) and it shows that this framework can ensure the stability in GAN training and guarantee the diversity of generated samples. Further, the method of Time-GAN is introduced in [14], which is specially designed to generate real-time series, and this kind of data are widespread in the power grid. Because WGAN and traditional GAN do not consider the temporal dynamics of the training set, this method of Time-GAN has obvious advantages in generating time series.

Therefore, this paper innovatively proposes the idea of applying Time-GAN to the field of online analysis of data volume and rapid identification of problems in a dispatching automation system and combines the idea of Time-GAN with typical fault diagnosis methods. Using the idea of Time-GAN, based on a small number of datasets, a large number of reliable datasets are obtained for the training of fault diagnosis models. The method above not only reduces the time of manually labeling training datasets but also improves the precision of the fault diagnosis model. Then, the expanded samples are input into the supervised learning detection model and the unsupervised learning detector for comparison. Finally, unsupervised learning is used to detect the unknown anomalies that may exist in the power grid. The simulation results show that this method can achieve accurate and efficient power grid fault diagnosis and prediction results, which lay a good foundation for the safe, stable, high-quality and economical operation of the power grid.

The rest of the paper is organized as follows. Section 2 presents the system model. Section 3 discusses the proposed fault diagnosis method. Section 4 presents the experimental results and analysis. Section 5 concludes this paper.

## 2. System Model

According to the structure of a power dispatching automation system and the abnormal classification characteristics of dispatching business information flow, this paper designs the abnormal characteristics and fault identification process based on massive information flow.

### *2.1. Power Dispatching Automatic System*

The data collection and analysis scene of the power dispatching automation system is shown in Figure 1, including the main network and the distribution network, respectively, collecting datagram from their own networks. Among them, the data collection center mainly collects the message data in the main network; that is, the message data are sent from the remote control device of the factory station through a network switch to the main network data collection center and are then sent to the anomalies detection server for fault diagnosis. The data acquisition center mainly collects the message data in the distribution network, which means the message data are sent by the power distribution automation terminal of the factory station. The power distribution automation terminal of the factory station sends data to the safe access area of the main station, and the message data are collected to the network switch in the safe area and then sent to the distribution network data collection center. The data collection center of the distribution network is connected to the switch in the safe area through the mirror port. Based on the collected big data and AI technology, the anomalies detection and diagnosis server can intelligently identify, diagnose and predict the abnormal faults of the power dispatching automation system.

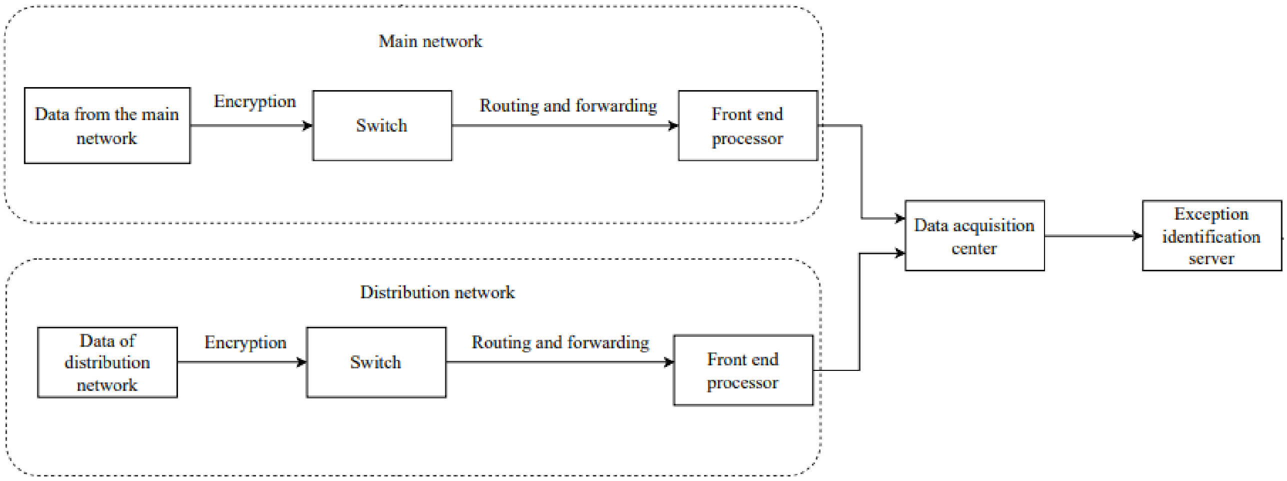

**Figure 1.** Schematic diagram of data acquisition and processing of power dispatching automation system.

### *2.2. Anomaly Classification of Scheduling Service Information Flow*

The real-time performance and reliability of scheduling business information flow directly affect the realization of various business functions and they are specifically reflected in the fact that the transmission delay should be guaranteed within the required time range with no packet loss or retransmission occurrences. Since the channel, equipment and system often have some problems in design, setting and maintenance that cannot be completely eliminated, sometimes there will be some abnormal information. Part of this abnormal information is the response of the real state of the site and the other is caused by various errors. These abnormalities have troubled the power system's personnel for operation and maintenance.

Classified according to the formation mechanism and characteristic quantity of dispatching business information flow anomalies, the data flow anomalies in this paper are divided into functional anomalies, timeliness anomalies, communication anomalies, integrity anomalies and frequency anomalies. Among them, functional anomalies include message disorder and telemetry values not being refreshed. Timing anomalies include irregular message delay and remote signaling jitter. Communication anomalies include abnormal flow and communication retransmission anomaly. Integrality anomalies include incomplete message and intermittent incomplete message. Frequency anomalies include frequent uploading of remote message and collective uploading of telemetry messages.

Because the traditional power grid anomaly detection is mainly based on functional anomalies and timeliness anomalies, the accuracy and real-time performance are poor, so this paper makes correlations based on the experience of experts and mainly analyzes the correlation between some communication anomalies, timeliness anomalies and message parameters. Among them, communication anomalies include device reboot error and device alternate channel error. Timeliness anomalies include telemetry errors and total call error, as shown in Tables 1 and 2.

### *2.3. Abnormal Characteristics and Fault Identification Process Based on Massive Information Flow*

Aiming at the characteristics of the power grid's abnormal data, this paper designs the abnormal characteristics and fault discrimination and counting framework based on massive information flow, as shown in Figure 2. First of all, the data generated by the power dispatching data network are obtained by nondestructive collection to obtain message information flow. For the problem of insufficient abnormal data, we use Time-GAN to obtain a large number of reliable datasets based on a small number of datasets. The dataset is used for training the fault diagnosis model. Then, supervised learning and unsupervised learning are compared for the accuracy of the power network's anomaly recognition. For unknown anomalies that may exist in the power grid, unsupervised learning methods such as DBSCAN and K-means are used to detect them. These methods can achieve an accurate and efficient grid fault diagnosis and prediction effect. It lays a good foundation for the safe, stable, high-quality and economical operation of the power grid.

**Table 1.** Network KPI parameters.

| KPI Parameters | Symbolic | KPI Parameters | Symbolic |
|---|---|---|---|
| 104 message | 104_M | TCP Keep-Alive message explosion | TCP_KA_E |
| Type ID | TY | The number of retransmission packets exploded | RET_E |
| Timestamp | TS | Remote letters explosion | RL_E |
| Telemetry message explosion | TM_E | Status value | SV |

**Table 2.** Correlations between fault cause and message parameters.

| Anomaly Classification | Representational Phenomenon | KPI |
|---|---|---|
| Device reboot error | A large number of devices restart recovery | 104_M, TY, TS, TCP_KA_E |
| Device Alternate Channel Error | Can't connect to backup channel | 104_M, TY, TS, TCP_KA_E, RET_E |
| telemetry error | Collective telemetry upload | 104_M, TY, TS, TM_E |
| Total call error | The total summoning frequency is abnormal | 104_M, TY, TS, TCP_KA_E |

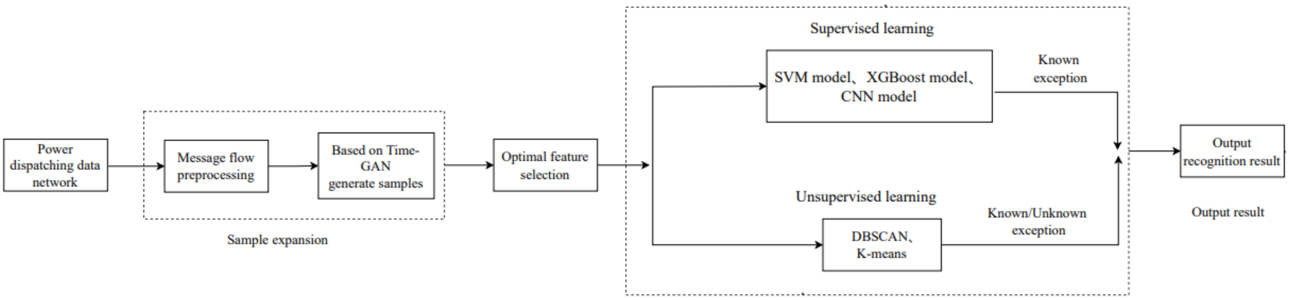
Power dispatching data network
Message flow preprocessing
Based on Time-GAN generate samples
Sample expansion
Optimal feature selection
Supervised learning
SVM model、XGBoost model、CNN model
Known exception
Unsupervised learning
DBSCAN、K-means
Known/Unknown exception
Output recognition result
Output result

**Figure 2.** Anomaly characteristics and fault discrimination technology framework based on massive information flow.

## 3. System Fault Diagnosis and Prediction Based on Time-GAN and DBSCAN

This paper proposes a system fault diagnosis and prediction model based on the generation adversarial network, as shown in Figure 3. Firstly, a small amount of KPI data under different system states are collected by the power dispatching automation system and the correlation between the fault causes and message parameters is sorted out by combining expert knowledge. Secondly, feature selection is carried out on the power data to select the best feature combination and then we input the processed small sample data into the generated adversarial network for data fitting under various network states, so as to obtain a large number of marked simulation data and unmarked data under various network states. Next, the generated adversarial network generates a simulation dataset and the original dataset is processed simultaneously. The finally processed data are divided into a training set and a test set. The marked dataset is input into the unsupervised learning fault detection module and the recognition result of known anomalies is obtained, which is compared with that of supervised learning. Finally, the unmarked data is input into the unsupervised learning anomaly recognition module to output the recognition results of possible unknown anomalies.

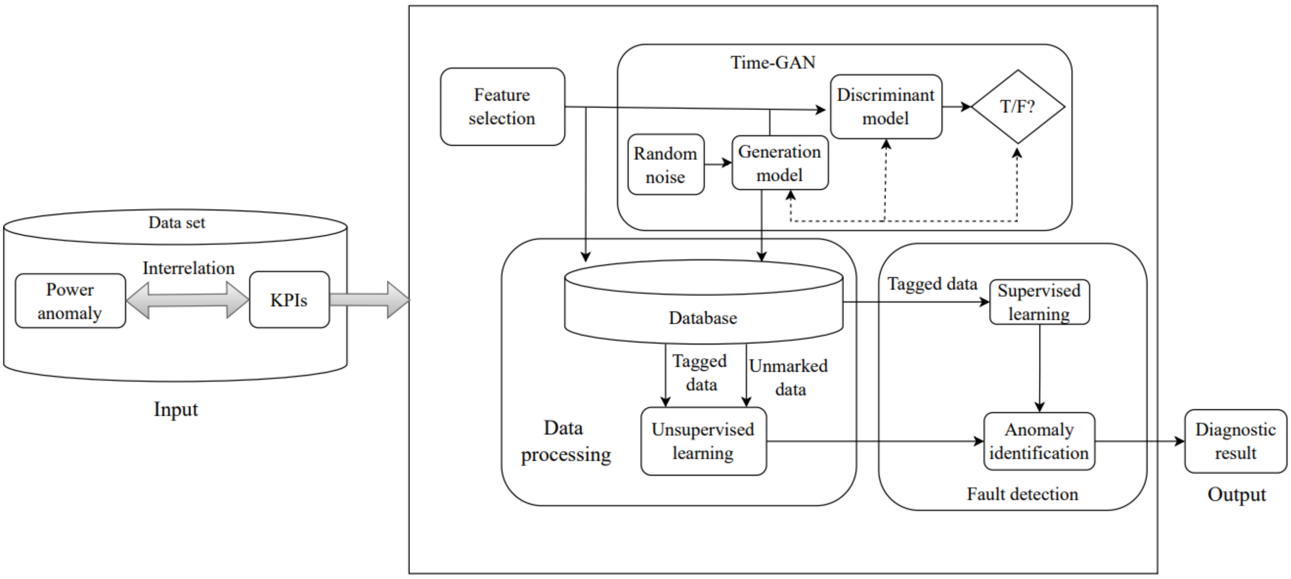

**Figure 3.** Power grid fault detection and diagnosis model based on generative confrontation network.

### 3.1. Data Generation by Time-GAN

Time-GAN is a time-series data generation model. Its main idea is to combine the versatility of the unsupervised GAN method with the conditional probability principle provided by the supervised autoregressive model; it contains a generator, a discriminator and a time aligner. The generator uses random noise to generate fake time series data, the discriminator distinguishes real and fake data, and the time aligner aligns the data to handle different time scales. The generator uses convolutional and deconvolutional layers to generate data, and the discriminator is a convolutional neural network classifier. The time aligner uses a self-attention mechanism to align data on the time axis. The model optimizes the generator and discriminator through adversarial training to generate more realistic time series data. To generate a time series with dynamic retention time, Time-GAN is used in this paper to expand the grid data.

This paper uses the data generation model of Time-GAN proposed by Jinsung Yoon et al. [14], which is specially designed for the generation of real Time series. Firstly, the generation model not only introduces the unsupervised antagonistic loss of real data and synthetic data, but also introduces the gradual supervisory loss of original data as supervision. Secondly, the method introduces an embedded network to provide reversible mapping between feature space and latent space, which reduces the dimension of adversarial learning space. Finally, supervision losses are minimized by training the embedded network and the generator network jointly, so that the latent space not only helps to improve parametric efficiency, but also helps the generator to learn time relationships.

After the IEC104 packet is obtained, the packet is segmented according to the packet's structure. Then, the features selected by Principal Component Analysis (PCA) are extracted and taken as the sample set to be expanded, in which the timestamp features of the message are retained. Then, after the extended sample set is normalized, the batch is processed and converted into the data form suitable for Time-GAN processing. Finally, according to the common generation ratio of 1:1, through the game between generator and discriminator in the Algorithm 1, the extended sample set of power dispatching network is obtained after 50 iterations of output.

The hyperparameter $\lambda$ is used to balance the target loss function $\mathcal{L}_S$ and $\mathcal{L}_R$, while $\eta$ is used to balance the target loss function $\mathcal{L}_S$ and $\mathcal{L}_U$. In practice, we find that Time-GAN is insensitive to $\lambda$ and $\eta$, so for all of our experiments, we set $\lambda = 1$ and $\eta = 10$.

**Algorithm 1** Time-GAN Algorithm [14]

```
Input: $\lambda = 1, \eta = 10$, training set $\mathcal{D}$, input size per batch $n_{mb}$, learning rate $\gamma$
Initialization: $\theta_e, \theta_r, \theta_g, \theta_d$
while generator does not converge do
Transformation between feature space and latent space
  Sample $(s_1, x_{1,1:T_n}), \ldots, \left(s_{n_{mb}}, x_{n_{mb},1:T_{n_{mb}}}\right) \overset{i.i.d}{\sim} \mathcal{D}$
  for $n = 1, \ldots, n_{mb}, t = 1, \ldots, T_n$ do
    $(h_{n,S}, h_{n,t}) = (e_S(s_n), e_{\mathcal{X}}(h_{n,S}, h_{n,t-1}, x_{n,t}))$
    $(\tilde{s}_n, \tilde{x}_{n,t}) = (r_S(h_{n,S}), r_X(h_{n,t}))$
Generate latent space codes
  Sample $(z_{S,1}, z_{1,1:T_n}), \ldots, \left(z_{S,n_{mb}}, z_{n_{mb},1:T_{n_{mb}}}\right) \overset{i.i.d}{\sim} p_{\mathcal{Z}_{s\times x}}$
  for $n = 1, \ldots, n_{mb}, t = 1, \ldots, T_n$ do
    $\left(\hat{h}_{n,S}, \hat{h}_{n,t}\right) = \left(g_S(z_{S,n}), g_{\mathcal{X}}\left(\hat{h}_{n,S}, \hat{h}_{n,t-1}, z_{n,t}\right)\right)$
Discrimination between real data and synthetic data
  for $n = 1, \ldots, n_{mb}, t = 1, \ldots, T_n$ do
    $\left(\mathrm{y}_{\mathrm{n,S}}, \mathrm{y}_{\mathrm{n,t}}\right) = \left(\mathrm{d}_{\mathrm{S}}(\mathrm{h}_{\mathrm{n,S}}), \mathrm{d}_{\mathcal{X}}\left(\overleftarrow{\mathrm{u}}_{\mathrm{n,t}}, \overrightarrow{\mathrm{u}}_{\mathrm{n,t}}\right)\right)$
    $\left(\hat{\mathrm{y}}_{\mathrm{n,S}}, \hat{\mathrm{y}}_{\mathrm{n,t}}\right) = \left(\mathrm{d}_{\mathrm{S}}\left(\hat{\mathrm{h}}_{\mathrm{n,S}}\right), \mathrm{d}_{\mathcal{X}}\left(\overleftarrow{\mathrm{u}}_{\mathrm{n,t}}, \overrightarrow{\mathrm{u}}_{\mathrm{n,t}}\right)\right)$
Calculate reconstruction loss, unsupervised and supervised loss
  $\hat{\mathcal{L}}_R = \frac{1}{n_{mb}} \sum_{n=1}^{n_{mb}} [\|s_n - \tilde{s}_n\|_2 + \sum_t \|x_n - \tilde{x}_{n,t}\|_2]$
  $\hat{\mathcal{L}}_U = \frac{1}{n_{mb}} \sum_{n=1}^{n_{mb}} \left[[log y_{n,S} + \sum_t log y_{n,t}] + \left[log(1 - \hat{y}_{n,S)} + \sum_t log(1 - \hat{y}_{n,t)}\right]\right]$
  $\hat{\mathcal{L}}_S = \frac{1}{n_{mb}} \sum_{n=1}^{n_{mb}} [\sum_t \|h_t - g_X(h_{n,S}, h_{n,t-1}, z_{n,t})\|_2]$
Update by gradient operator $\theta_e, \theta_r, \theta_g, \theta_d$
  $\theta_e = \theta_e - \gamma \nabla_{\theta_e} - \left[\lambda \hat{\mathcal{L}}_S + \hat{\mathcal{L}}_R\right]$
  $\theta_r = \theta_r - \gamma \nabla_{\theta_r} - \left[\lambda \hat{\mathcal{L}}_S + \hat{\mathcal{L}}_R\right]$
  $\theta_g = \theta_g - \gamma \nabla_{\theta_g} - \left[\eta \hat{\mathcal{L}}_S + \hat{\mathcal{L}}_U\right]$
  $\theta_d = \theta_d + \gamma \nabla_{\theta_d} - \hat{\mathcal{L}}_U$
Generate synthetic data
  Sample $(z_{S,1}, z_{1,1:T_n}), \ldots, (z_{S,N}, z_{N,1:T_N}) \overset{i.i.d}{\sim} p_{\mathcal{Z}_{s\times x}}$
Generate synthetic hidden space codes
  for $n = 1, \ldots, N, t = 1, \ldots, T_n$ do
    $\left(\hat{h}_{n,S}, \hat{h}_{n,t}\right) = \left(g_S(z_{S,n}), g_{\mathcal{X}}\left(\hat{h}_{n,S}, \hat{h}_{n,t-1}, z_{n,t}\right)\right)$
Convert latent space code to feature space
  for $n = 1, \ldots, N, t = 1, \ldots, T_n$ do
    $(\hat{s}_n, \hat{x}_{1,T_n}) = (r_S(h_{n,S}), r_X(h_{n,t}))$
end while
output: $\hat{\mathcal{D}} = \{\hat{s}_n, \hat{x}_{1:T_n}\}_{n=1}^{N}$
```

### 3.2. DBSCAN Algorithm

DBSCAN is a density-based spatial clustering algorithm. The algorithm divides regions with sufficient density into clusters and finds clusters of arbitrary shape in the noisy spatial database. The cluster is defined as the maximum set of density-connected points. It types the samples according to the density of the sample space and separates the sample points that do not belong to the density region. A DBSCAN algorithm can learn the data distribution from a group of power grid datasets, classify the message samples with similar patterns into a cluster and mark the samples that are out of the cluster in the dataset as abnormal samples.

However, a DBSCAN algorithm itself also has limitations. In its working process, it needs a large amount of data to participate in clustering to achieve better results. This is because a relatively dense data cluster cannot be formed when the amount of data involved in clustering is too small, so the cluster cannot be formed. At the same time, when processing anomaly detection tasks, a too small amount of data can not help distinguish abnormal samples from normal samples, resulting in the algorithm failure in anomaly detection. In this paper, the limitations of a DBSCAN algorithm can be well circumvented because there is a large amount of normal message data in the power dispatching network environment.

Here's how it works. A DBSCAN algorithm firstly collects a dataset $D$ containing $n$ data samples, determines the parameter ε by elbow point method and then determines the parameter μ by empirical method. The function of the parameter epsilon is to describe the radius of the core neighborhood, and μ is to describe the minimum sample number in the core neighborhood. Only when these two parameters are determined, the algorithm can carry out the next calculation. In this paper, the values of $\varepsilon$ is 2.5 and $\mu$ is 14. The specific value process is described in detail in Section 4.4. Then, the $D$. Cluster of all samples in dataset $D$ is set to the initial state of unclustered. The $D$. Cluster denotes the cluster classification tag of the samples. For each sample $D_{\mathrm{i}}$, we first judge whether there are at least $\mu$ number of samples in the neighborhood radius of $\varepsilon$. If so, a new cluster $C$ is created and is classified into cluster $C$. Then, all samples $N_{\mathrm{i}}$ in the $\varepsilon$ radius of $D_{\mathrm{i}}$ are extracted to determine whether there are at least $\mu$ number of samples in the neighborhood radius of $\varepsilon$. If so, $N_{\mathrm{i}}$ will be classified into cluster $C$. Otherwise, the classification cluster marker $N$. cluster of $N_{\mathrm{i}}$ will be tagged unclustered. In the end, the samples tagged unclustered will be considered outliers. Compared with the K-means algorithm, this algorithm does not need to specify the cluster number in advance. The specific pseudocode is shown as follows (Algorithm 2).

**Algorithm 2** DBSCAN algorithm [15]

```
Input: dataset $\boldsymbol{D}$ containing $\boldsymbol{n}$ objects, radius parameter $\boldsymbol{\varepsilon}$, minimum number of samples $\boldsymbol{\mu}$
Initialization: $\boldsymbol{\varepsilon = 2.5, \mu = 14}$, Cluster list[ ]
Set the cluster classification tag D.cluster of $D_{\mathrm{i}}$ data in the dataset as unclustered
For i $= 1, \ldots, N$, do
If there are at least $\boldsymbol{\mu}$ samples within the domain radius $\boldsymbol{\varepsilon}$ of $D_{\mathrm{i}}$ (whether the sample is a core instance)
    Create a new cluster $\boldsymbol{C}$, add $\boldsymbol{C}$ to the Cluster list[ ], and add $D_{\mathrm{i}}$ to $\boldsymbol{C}$
        Take all samples in the $\boldsymbol{\varepsilon}$-neighborhood radius of $D_{\mathrm{i}}$ to form a set N (N is consisted of $N_{\mathrm{i}}$)
    for each sample $N_{\mathrm{i}}$ in $\boldsymbol{N}$
        mark $N_{\mathrm{i}}$ as clustered
        If there are at least $\boldsymbol{\mu}$ samples within the neighborhood radius $\boldsymbol{\varepsilon}$
            Add sample $N_{\mathrm{i}}$ to C
        If $N_{\mathrm{i}}$ does not belong to C
            Set the cluster classification tag D.cluster of $N_{\mathrm{i}}$ data in the dataset as unclustered
End while
Data that are still marked as unclustered are classified as outliers, marked as $-\mathbf{1}$ and placed in the Cluster
list[]
Output: the samples tagged as unclustered
```

## 4. Performance Analysis

The dataset used in this paper was provided by the Nanjing Power Supply Branch of Jiangsu Electric Power Co., Ltd. (Nanjing, China) during 2021–2022. There are nearly 12 million pieces of data in total. Among them, 9868 pieces are faulty data; it is also a data set; the ratio of training set and test set is 0.7:0.3, respectively. The parameter settings of the Time-GAN algorithm and the parameter settings of DBSCAN are shown in Section 3. The model parameters of other algorithms are set by default, such as PCA, GAN, WGAN, SVM, XGBoost and CNN.

In this paper, the collected dataset is saved as K12 text file format. In the K12 text file format, the hexadecimal bytes of information are stored as '｜' character segmentation. Each piece of data contains layer headers and 104 protocol information. The packet information starts with the MAC address. The first 12 bytes correspond to the MAC source address and MAC destination address. Protocol 104 starts with 68 and the next byte represents the length of the Application Protocol Data Unit (APDU).

### *4.1. Optimal Feature Selection*

The timing exception and frequency exception correlate the characteristics of protocol header information at each layer of the dataset and the characteristics of 104 packets. The feature dimension of the dataset is greatly increased, so it is necessary to reduce the feature dimension of the entire dataset first.

We perform feature dimensionality reduction on the entire dataset by means of combining expert knowledge and PCA. Expert knowledge refers to the classification table established by experts based on previous abnormal diagnosis, as shown in Tables 1 and 2. PCA is a data dimension reduction algorithm. The main idea of PCA is to reconstruct k−dimensional features on the basis of original n−dimensional features. To be specific, we first find a set of mutually orthogonal axes in sequence from the original space and the selection of new axes is closely related to the data itself. The first new axis selected has the direction with the largest difference in the original data. The second new axis selected has the plane orthogonal to the first axis with the largest variance. The third axis selected has the plane orthogonal to the first and second axes with the largest difference. By analogy, we find that most of the variance is contained in the first k axes and the variance of the later axes is almost zero. So, we can ignore the rest of the axes and just keep the first k axes with most of the variance. In fact, this is equivalent to retaining only the dimensional features containing most of the variance, while ignoring the dimensional features containing almost zero variance, so as to achieve dimensionality reduction of the dataset.

The dataset's features mainly consist of the header information of each layer protocol and 104 packets. The header information of each layer includes Media Access Control (MAC) source address, MAC destination address, IP type, IP header length, protocol, timestamp, timestamp echo reply, etc. The features of the packet contain start character, APDU length, control field, type flag, transmission reason, etc.

The process of feature selection is shown in Figure 4. Firstly, we preprocess the dataset; preprocessing of the dataset mainly includes format conversion, truncation and filling, base conversion and annotation. The format conversion is to convert the dataset's K12 text file format into comma-separated values (CSV) format. Truncation and filling are to unify the number of bytes of the packet into 40. When the number of bytes of the packet exceeds 40 bits, the excess bytes will be truncated. When the number of bytes of the packet is less than 40 bits, −9999 will be used to supplement the vacancy. Base conversion is to convert hexadecimal byte information to decimal, because the hexadecimal notation contains more letters. Marking is to mark data according to existing exception categories. For example, device reboot error is marked as 0. Secondly, the importance of transport layer features and 104 message features are sorted according to the expert knowledge and PCA. Then, the model complexity and model accuracy are comprehensively considered to select the appropriate number of features to obtain the best feature combination in the order of importance. Finally, the dataset dimension was reduced from 40 to 15 dimensions.

In this paper, a PCA algorithm is first used to screen features. We use the PCA algorithm model to sort the importance of the features, which reflects the influence of each feature in the network fault diagnosis process. We mark the 40 features with the numbers 1 through 40, for example, the type of flag feature is labeled 19. We use a PCA algorithm to obtain their weight values f1 through f40 successively and arrange the obtained weight values in descending order. We delete the features with low feature weight value successively and select the feature combination with a high special weight value for model training. A different number of feature selection will obtain a different model accuracy. The importance ranking of features and the influence of different feature combinations on model accuracy are shown in Figure 5a,b.

According to Figure 5a,b, only the top 14 features have a high impact on the accuracy of the model. Among them, f19 has the highest weight value, which represents the weight value of the type of flag. When the number of KPIs reaches 15, the accuracy reaches the highest. We first use the PCA method to sort the importance of features in the dataset, and we select the top 14 features. When the number of features is 15, the model accuracy reaches the highest and adding new features again would not improve the model's accuracy. This suggests that our feature selection does not screen out important features. Based on the PCA's feature selection and expert knowledge, the best feature combination is selected, including IP length, protocol, source address, destination address, TCP serial number, acknowledgement number, time stamp, 104 packet start character, APDU length, type flag, control field, transmission reason, application service data unit, information element and information object address.

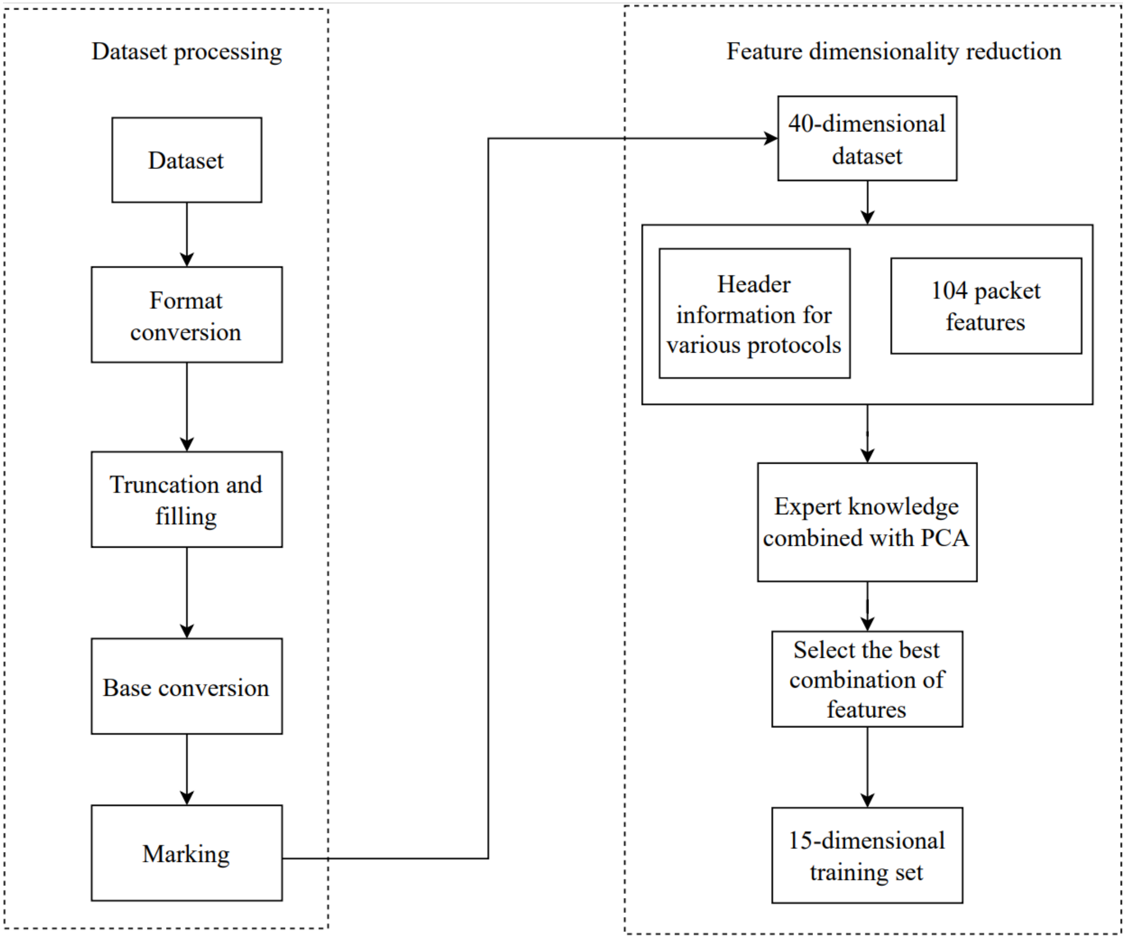
Dataset processing
Dataset
Format conversion
Truncation and filling
Base conversion
Marking
Feature dimensionality reduction
40-dimensional dataset
Header information for various protocols
104 packet features
Expert knowledge combined with PCA
Select the best combination of features
15-dimensional training set

**Figure 4.** Best feature selection process.

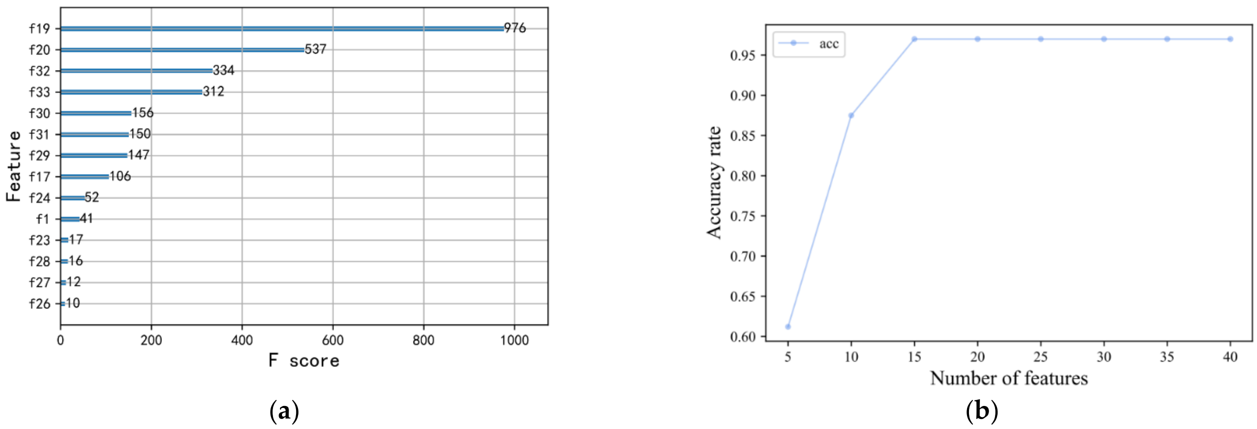

**Figure 5.** The effect of the best feature selection process. (**a**) The weight values of various features. (**b**) The influence of different feature combinations on the accuracy of the model.

As shown in Figure 5a,b, the original training set contains all features, and its accuracy is the accuracy when the number of features is 40. However, when we selected the most important 15 features as the training set, the accuracy reached the highest. Moreover, the dimension of the training set decreases after feature selection, so the training time of the model can be reduced.

### *4.2. Comparison of Accuracy of Various Algorithms on Known Anomalies*

We hope that the generated samples can be used not only as a test set to detect the model, but also as a training set to solve the problem of insufficient training samples. Therefore, this paper sets up two methods to test the generated samples. One method is to use the model trained by the original data to test the generated samples, while the other method is to use the model trained by generated samples to test the original data. The samples generated by the three algorithms of classic GAN, WGAN and Time-GAN are compared. The test set used in the second method is the same as the original dataset used for generating samples. Taking the anomalies of device restart error as an example, the number of original samples with anomalies is 1008.

Firstly, we test the accuracy of generated data and the results are shown in Figure 6. It can be seen that the optimal generation ratio of Time-GAN is much larger than the other two algorithms of classic GAN and WGAN. Moreover, the accuracy rate of Time-GAN is 100% when the generation ratio is a number between 1 and 4.

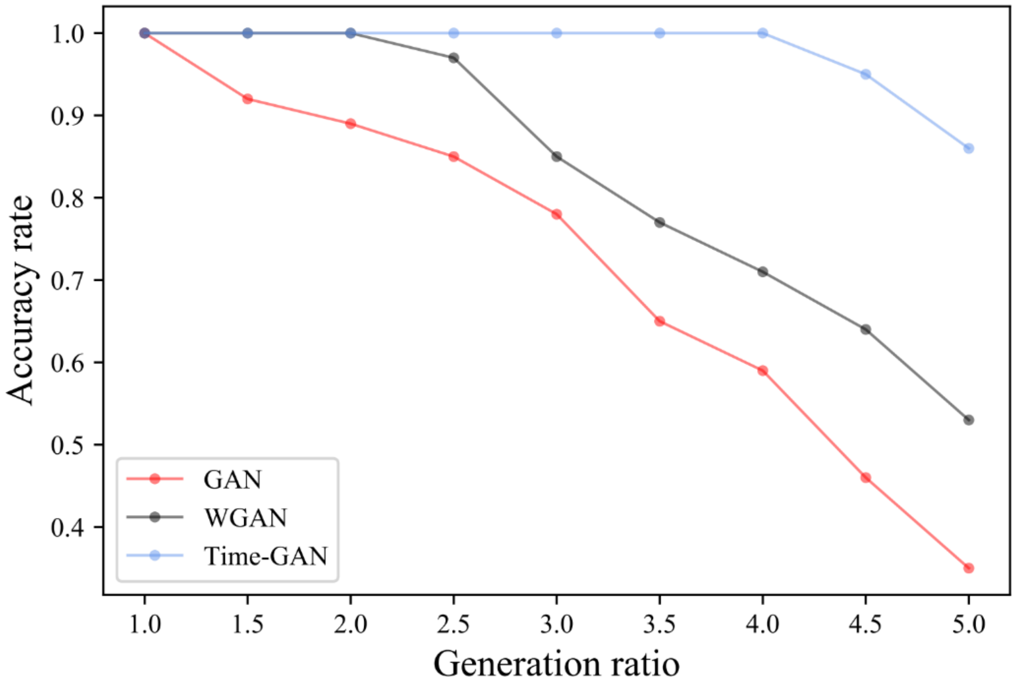

**Figure 6.** Test set accuracy under different generation ratios of various algorithms.

Next, we test the accuracy of the generated model. Because a CNN algorithm has the highest accuracy in anomaly identification compared with other supervised learning algorithms, the data generated by selecting the optimal generation ratio of each algorithm are input into the CNN algorithm to train the model and the same original abnormal samples are used to test the generated model. The results are shown in Table 3 and it can be seen that the accuracy of the generation model of Time-GAN is slightly higher than that of the classic GAN and WGAN generation algorithms.

**Table 3.** Best generating scale model accuracy.

| Generating Algorithm | Classic GAN | WGAN | Time-GAN |
|---|---|---|---|
| Optimal generation ratio | 1:1 | 1:2.1 | 1:4 |
| Generating model accuracy | 0.93 | 0.96 | 1 |
| Generating model precision | 0.92 | 0.96 | 1 |
| Generating model recall | 0.94 | 0.96 | 1 |

As can be seen from Figure 6 and Table 3, Time-GAN is the algorithm with the highest expansion ratio and more abnormal samples are helpful to our model training. Because the Time-GAN algorithm considers the time characteristics between datasets, the accuracy of the model is the highest. State Grid tried to use classical GAN to enlarge the samples before, but the enlarged samples had serious distortion and could not be used as a dataset. Therefore, this paper chooses Time-GAN as the generation method.

### *4.3. Comparison of Accuracy of Various Algorithms on Unknown Anomalies*

It can be seen from Section 4.2 that Time-GAN has the best performance in enlarging the power grid’s samples, so in this section we would adopt Time-GAN to enlarge abnormal samples whether marked or unmarked with the optimal generation ratio of 1:4. The supervised learning methods commonly used include SVM, XGBoost and CNN, while unsupervised learning methods include DBSCAN and K-means.

We select four kinds of anomalies for comparison, including device restart error, communication interruption, telemetry collective uplink and total call error. In the test dataset, there are four types of abnormal message data that are the most critical. In addition to collecting the small number of abnormal message data from the real power grid’s environment, we enlarge the abnormal sample data with a generation ratio of 1:4 through the Time-GAN method. Apart from abnormal message data, the test set also includes a large number of normal message data collected from the real power grid’s automated dispatching system. This implies that the sample set obtained is unbalanced, since the proportion of abnormal message data is very small. As to the supervised learning methods, namely SVM, XGBoost and CNN+LSTM methods, we mark the messages during training, while in terms of the unsupervised learning methods which include DBSCAN and K-means, we do not mark any data during training. In the final test, each algorithm classifies each message using the above test set. The category labels are named normal, device restart error, communication interruption, telemetry collective uplink and total call error. The accuracy of the test set is shown in the following table.

As can be seen from Table 4, the DBSCAN method has the highest accuracy. In supervised learning methods, the model test set of a CNN combined algorithm has the highest accuracy, but it is far less than DBSCAN, the unsupervised learning method. Considering that DBSCAN can effectively isolate abnormal data, we intend to use DBSCAN to isolate possible unknown exceptions from the dataset and let experts further analyze whether it is a new exception.

**Table 4.** Various abnormal accuracy.

| | SVM | XGBoost | CNN | DBSCAN | K-Means |
|---|---|---|---|---|---|
| Device restart error | 0.12 | 1 | 0.92 | 1 | 0.37 |
| Communication interruption | 0.74 | 0.65 | 0.72 | 1 | 0.45 |
| Collective telemetry upload | 0.23 | 0.52 | 0.82 | 1 | 0.76 |
| Total call error | 0.46 | 0.92 | 1 | 1 | 0.49 |

### *4.4. Unknown Anomaly Detection*

Grid data are constantly being transmitted and there could exist various kinds of unknown anomalies in the existing electric power grid in addition to the ones we've already found and documented. However, supervised learning requires a large number of marked anomaly samples as a training set to train the model for identifying anomalies. Under this circumstance, we propose applying the unsupervised learning method to detect unknown anomalies in the power grid. According to the characteristics of power data, K-means and DBSCAN are selected for comparison with the unsupervised learning methods commonly used. The sample adopted is the dataset with known anomalies removed and the dataset may contain other unknown anomalies, which is called the preliminary screening dataset.

#### 4.4.1. Applying DBSCAN Algorithm to Detect Unknown Anomalies

In this paper, 85,033 preliminary screening datasets are taken as input and a DBSCAN clustering algorithm is performed on them. Firstly, we should determine the parameters of the algorithm, including the neighborhood radius $\varepsilon$ and the minimum number of points within the domain radius that become the core object $\mu$. The minimum number of points $\mu$ is determined empirically, which is usually twice the number of selected features, and it is in this paper determined as 30. The common way to determine the neighborhood radius $\varepsilon$ is the elbow point method and it is determined here as 2.5. The result is shown in Figure 7, where 18,000 samples are selected to draw the k-distance diagram.

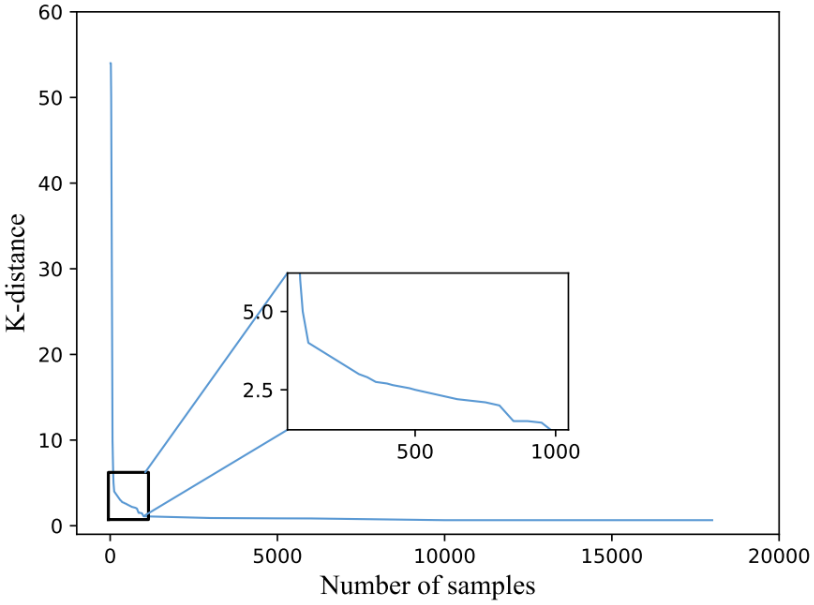

**Figure 7.** Elbow point method k-distance diagram.

We find that only telemetry messages may have unknown anomalies. So, we analyze the classification results of telemetry messages as well, and we only show the analysis results of the telemetry information. The result of clustering by DBSCAN algorithm is shown in Figure 8. After the preliminary screening of 85,033 datasets is processed by DBSCAN, the number of transmitted telemetry packets with cause 3 is 67,947, the number of transmitted S messages with cause 0 is 17,013, the number of transmitted telemetry message with cause 20 is 48 and the number of transmitted telemetry messages with cause 6, 7, 10 is 11, 11, 4, respectively. With the analysis of expert knowledge, we know that both type 0 and type 2 are composed of telemetry messages with cause 3, type 1 is composed of the S message with cause 0 and type 3 is composed of telemetry messages with cause 20. The result is in line with our expectations, which are that similar packets are classified into one type. Next, we will focus on analyzing packets with type −1, which are abnormal samples that do not belong to any type.

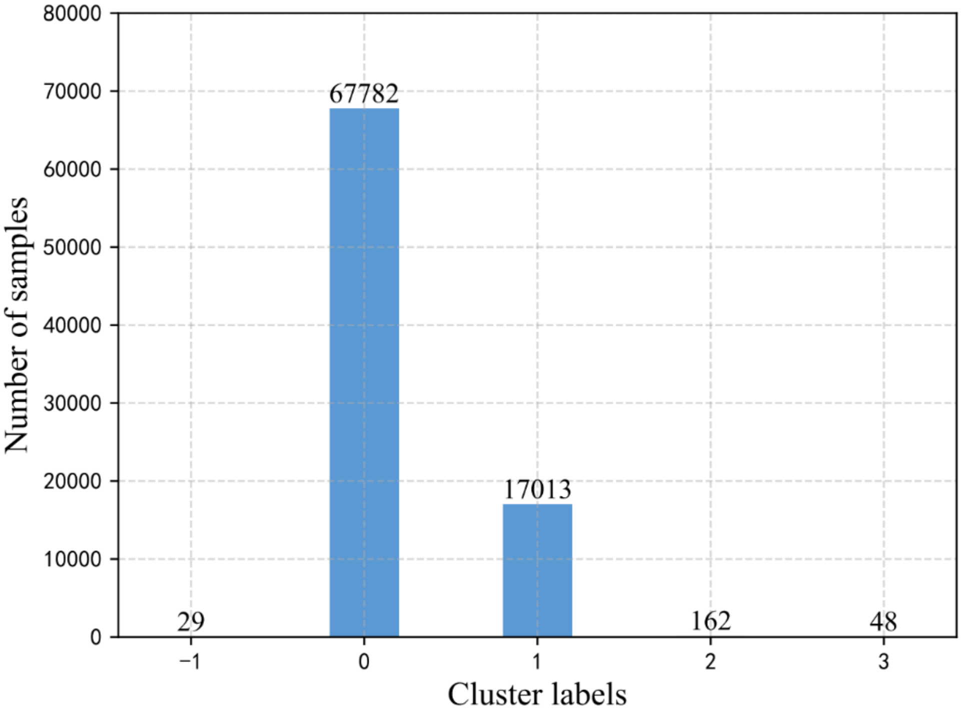

**Figure 8.** DBSCAN clustering results.

As mentioned before, type −1 includes telemetry packets with cause 6, 7 and 10. Because few telemetry packets with cause 6, 7 and 10 are sent, they are misjudged as outliers. However, if the number of these telemetry packets increases appropriately, these packets can establish their own type and will not be regarded as abnormal samples. Therefore, telemetry packets with cause 6, 7 and 10 are not outliers. After removing the 26 telemetry packets, there are only three abnormal packets actually, as shown in Table 5.

**Table 5.** Unknown exceptions isolated by DBSCAN algorithm.

| Length | Control Bit 1 | Control Bit 2 | Control Bit 3 | Control Bit 4 | Send Reason | Address 1 | Address 2 | Telemetry Value 1 | Telemetry Value 2 | Cluster Class |
|---|---|---|---|---|---|---|---|---|---|---|
| 70 | 60 | 58 | 238 | 9 | 3 | 12 | 64 | 99 | 122 | −1 |
| 154 | 22 | 124 | 242 | 9 | 3 | 7 | 64 | 0 | 0 | −1 |
| 124 | 88 | 235 | 244 | 9 | 3 | 145 | 64 | 46 | 9 | −1 |

As to the first abnormal message, the length of the packet sent by the device with an address of 64, 12 (64 and 12 represent two parts of a device address, separated by commas for better distinction) should not be larger than 30. However, the length of this telemetry packet is 70. As to the third abnormal packet sent by the device with an address of 64,145, it has a length of 124. However, we cannot simply conclude that the first and third abnormal messages are not misjudged. We also need to see the messages associated with this message before and after and their original packets. The sudden increase in the telemetry message's value is probably due to the fact that the telemetry value of a single address is sent in the previous packet. In this section, the telemetry value of multiple addresses is sent. Therefore, the sudden increase in the length of the packet is allowed.

As to the second abnormal message, it is the only one packet with the telemetry value of 0 sent by the device with address of 64, 7. Normally, if the telemetry value suddenly drops to 0, it may be caused by a sudden failure of the power protection device. If a fault really occurs, the telemetry value should be 0 many times. And we also need to check the duration of the telemetry value of 0. By analyzing the messages before and after the potential anomaly message, we find that it is indeed an unknown anomaly which means the sudden failure of the power protection device and it should be recorded in the scheduling automation system's data flow anomaly classification of functional anomaly.

#### 4.4.2. Applying K-Means to Detect Unknown Anomalies

There are two main algorithm parameters of the K-means algorithm. The first is the number of categories after clustering, namely $k$. The second is the $k$ initial cluster centers. When the number of clustered sample categories is determined, the value of $k$ is also determined. When the number of sample categories is uncertain, the $k$ value is usually determined by the elbow method or the silhouette coefficient method. The initial cluster centers are generally randomly assigned. The clustering dataset selected in this paper is the same as the DBSCAN dataset in Section 4.4.1, because there are only normal samples and samples that may be abnormal, so the $k$ value is determined to be 1 and then the cluster center c is calculated. The criterion for judging outliers is the Euclidean distance, namely $d_i$ between the input data, namely $I_i$ and the cluster center c. Then, the threshold $x = \mu + 3\delta$ is given for judging outliers, where $\mu$ refers to the mean value of the Euclidean distance from the input data to the cluster center and $\delta$ refers to the standard deviation of the Euclidean distance from the input data to the cluster center. Finally, input data satisfying $d_i > x$ are classified as outliers, that is, abnormal data.

It can be seen from Table 6 that the APDU length of the abnormal message is relatively large and the last few digits of the characteristic value are nonzero, while the last few digits of the packet with a small APDU length is padded by zero. Since the number of the last few digits accounted for an excessively large proportion in calculating the Euclidean distance, the packets with large APDU lengths would be judged as outliers and thus misjudged as abnormal packets.

**Table 6.** Unknown abnormal fields identified by K-means.

| APDU Length | Type ID | Transmission Reason | ASDU Public Address | Information Object Address | Telemetry Value |
|---|---|---|---|---|---|
| 58 | 9 | 3 | 1 | 0x4001 | 10,441 |
| 58 | 9 | 3 | 1 | 0x4039 | 10,462 |
| 64 | 9 | 3 | 1 | 0x4037 | 10,442 |
| 64 | 9 | 3 | 1 | 0x4037 | 10,458 |
| 64 | 9 | 3 | 1 | 0x4037 | 10,431 |

Because Time-GAN divides categories based on the density of the dataset, it does not need to determine the number of clusters in advance. However, a K-means algorithm is greatly influenced by the number of clusters. In reality, we do not know how many unknown anomalies exist in the dataset, so it is impossible to determine the number of clusters. As a result, the effect of a K-means algorithm is relatively poor. In addition, experimental results in Section 4.4 show that a DBSCAN algorithm can detect unknown anomalies, while K-means cannot do it. In general, this paper chooses a DBSCAN algorithm for anomaly identification.

#### 4.4.3. The Practical Application of the Algorithm

The method proposed in this paper has been applied to the main station platform of a distribution automation dispatching system in Nanjing, Jiangsu Province. The main station is connected to tens of thousands of distribution slave stations and the main station receives an average of about 2000 packets per second. We start by selecting the targeted static power grid data to define parameters for various DBSCAN algorithms. Additionally, we process stream data using a sliding window approach, where every 20,000 data points constitute a fixed-size window that is stored in a database. A DBSCAN algorithm is then applied to each window using the previously defined parameters. In a real-world operation, the average time taken for detecting 20,000 data points is 10.6 s. By analyzing and processing each message, all already known anomalies in the scheduling network can be analyzed quickly and efficiently, like device restart errors, communication interruptions, telemetry collective uplink and total call errors. Moreover, the method can also analyze the possible unknown anomalies because a DBSCAN algorithm is used to separate possible anomalies and determine whether they are new anomalies after further analysis by experts. And, we have successfully discovered a new unknown anomaly called failure of the relay protection device. So, the algorithm proposed in this paper is helpful to the safe and stable operation of power system.

## 5. Conclusions

In this paper, we focus on identification, diagnosis and prediction of massive data flow in a dispatching automation system. Considering the problem of insufficient abnormal data or training dataset, Time-GAN is firstly used to enlarge the dataset and we can thus obtain a large number of reliable data which conform to the characteristics of the power grid's actual data. By introducing Time-GAN, we can reduce the time of manually labeling training data and balance various types of power data. Secondly, the enlarged data are input into the fault diagnosis model which consists of CNN and LSTM. Finally, the unsupervised learning method is used to detect the possible unknown anomalies in the power grid. As to the proposed method, we first compare the unsupervised learning method and supervised learning method in terms of accuracy of known anomalies and it verifies that a DBSCAN algorithm outperforms other algorithms and it can identify the abnormality of the power grid. Then, we conduct some other experiments and numerical results show that the proposed algorithm can realize efficient and reliable fault diagnosis. In the actual deployment, we find that samples' slow generation limits the rapid detection of samples and lots of data are needed to participate in the clustering process. This paper mainly studies the problem of identification, diagnosis and prediction of massive data flow in the scheduling automation system. Considering the problem of insufficient abnormal data or training data, Time-GAN is first used to expand the data set, so as to obtain a large number of reliable data conforming to the characteristics of actual power grid data. By introducing Time-GAN, we can reduce the time it takes to manually label training data and balance the various types of power data. Secondly, the amplified data is input into the fault diagnosis model composed of CNN and LSTM. Finally, the unsupervised learning method is used to detect the possible unknown anomalies in the power grid. For the proposed method, we first compare the accuracy of the unsupervised learning method and supervised learning method for known anomalies, and verify that a DBSCAN algorithm is superior to other algorithms and can identify power grid anomalies. Experimental results show that this algorithm can realize efficient and reliable fault diagnosis. In the actual deployment, we found that slow sample generation limited the rapid detection of samples and required a large amount of data to participate in the clustering process. In further research, we intend to address these issues. We will improve the Time-GAN algorithm or choose a more appropriate Time series generation algorithm, such as a Transformer-based Time-Series Generative Adversarial Network (TTS-GAN) or a Progressive Self Attention GANs (PSA-GAN) algorithm. In the further research, we would intend to solve the above problems.

**Author Contributions:** Writing—original draft, W.L.; Writing—review & editing, P.L., D.X. and X.Z. All authors have read and agreed to the published version of the manuscript.

**Funding:** This work was supported in part by Natural Science Foundation of China (92067101) and in part by the Key R&D plan of Jiangsu Province (BE2021013-3).

**Data Availability Statement:** The data set involved in this paper is provided by the Jiangsu Nanjing Branch of the State Grid, and the power grid data is related to the security and privacy of the State grid, so the data set in this paper cannot be disclosed.

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

## References

1. Jian, F.; Cao, M.; Wang, L.; Sun, Z.; Zhang, J.; Wang, H. Research on Electrical anomaly Detection in AMI environment based on SV [基于SVM的AMI环境下用电异常检测研究]. *Electr. Meas. Instrum.* **2014**, *51*, 64–69.
2. Xu, Y.; Li, S.; Han, Y. Abnormal power consumption behavior detection based on CNN-GS-SVM [基于CNN-GS-SVM的用户异常用电行为检测]. *Control. Eng. China* **2021**, *28*, 1989–1997.
3. Yang, Z.; Ding, J.; Chen, G.; Kang, X.; Sheng, M. Research on Abnormal Power Consumption Detection Method of Power Big Data based on LightGBM and LSTM model [基于 LightGBM 和LSTM模型的电力大数据异常用电检测方法研究]. *Electr. Instrum. Meas.* **2023**, 1–7. Available online: http://kns.cnki.net/kcms/detail/23.1202.TH.20220713.1958.004.html (accessed on 15 May 2023).
4. Liu, D.; Jiang, Z.; Zhu, Y.; Huang, Y.; Xiao, Y. Abnormal detection of power monitoring system network traffic based on LDSAD [基于LDSAD的电力监控系统网络流量异常检测]. *J. Zhejiang Electr. Power* **2022**, *9*, 87–92. [CrossRef]
5. Yan, Y.; Sheng, G.; Chen, Y.; Jiang, X.; Guo, Z.; Du, X. Abnormal detection method of state data of power transmission and transformation equipment based on big data analysis [基于大数据分析的输变电设备状态数据异常检测方法]. *Proc. CSEE* **2015**, *35*, 52–59.
6. Huang, G.J. Research on Flow Data Analysis of power Enterprises based on Canopy-Kmeans algorithm [基于Canopy-Kmeans算法的电力企业流量数据分析研究]. *Inf. Technol. Netw. Secur.* **2022**, *41*, 5. [CrossRef]
7. Wu, G.; Yao, J.; Guan, M.; Zhu, X.; Wu, K.; Li, H.; Song, S. Power transformer fault diagnosis based on DBSCAN [基于DBSCAN的电力变压器故障诊断]. *J. Wuhan Univ. (Eng. Sci.)* **2021**, *54*, 1172–1179.
8. Wang, P.; Zhang, Q. Power expert fault diagnosis information fusion method based on fuzzy clustering [基于模糊聚类的电力专家故障诊断信息融合方法]. *J. Heilongjiang Electr. Power* **2020**, *2*, 109–112+118. [CrossRef]
9. Dong, X.-Y. Research on Intrusion Detection Method of Wireless Network based on Clustering Algorithm [基于聚类算法的无线网络入侵检测方法研究]. Master's Thesis, Hebei Normal University, Shijiazhuang, China, 2022. [CrossRef]
10. Jian, S.; Lu, Z.; Jiang, B.; Liu, Y.; Liu, B. Traffic anomaly detection based on hierarchical clustering method. *Inf. Secur. Res.* **2020**, *6*, 474–481.
11. Goodfellow, I.J.; Pouget-Abadie, J.; Mirza, M.; Xu, B.; Warde-Farley, D.; Ozair, S.; Courville, A.; Bengio, Y. Generative adversarial nets. In Proceedings of the International Conference on Neural Information Processing System/s, Montreal, QC, Canada, 8–13 December 2014; MIT Press: Cambridge, MA, USA, 2014; pp. 2672–2680.
12. Arjovsky, M.; Bottou, L. Towards principled methods for training generative adversarial networks. *arXiv* **2017**, arXiv:1701.04862.
13. Liu, H.; Gu, X.; Samaras, D. Wasserstein GAN With Quadratic Transport Cost. In Proceedings of the 2019 IEEE/CVF International Conference on Computer Vision (ICCV), Seoul, Republic of Korea, 27 October–2 November 2019; pp. 4831–4840. [CrossRef]
14. Yoon, J.; Jarrett, D.; Van der Schaar, M. Time-series generative adversarial networks. In Proceedings of the Neural Information Processing Systems (NeurIPS), Vancouver, BC, Canada, 8–14 December 2019.
15. Ester, M.; Kriegel, H.P.; Sander, J.; Xu, X. A Density-Based Algorithm for Discovering Clusters in Large Spatial Databases with Noise. In Proceedings of the 2nd International Conference on Knowledge Discovery and Data Mining, Portland, OR, USA, 2–4 August 1996.

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