

PAPER • OPEN ACCESS

## An Intelligent Operation and Maintenance System for Power Consumption Based on Deep Learning

To cite this article: Zengrui Huang *et al* 2019 *IOP Conf. Ser.: Mater. Sci. Eng.* **486** 012107

View the [article online](#) for updates and enhancements.

# An Intelligent Operation and Maintenance System for Power Consumption Based on Deep Learning

Zengrui Huang<sup>1</sup>, Wei Mao<sup>1</sup>, Ming Chen<sup>2</sup>, Qiang Wu<sup>2</sup>, Boyue Xiong<sup>2</sup> and Wei Xu<sup>2</sup>

<sup>1</sup>School of Computer Science, Fudan University, Yangpu District, Shanghai 200433, China,

<sup>2</sup>State grid Shanghai municipal electric power, LTD., Minghang District, Shanghai 201199, China,

E-mail address: zruang17@fudan.edu.cn; Telephone number: +86-15216680682

**Abstract.** With the advent of the era of power big data, power companies can easily access more and more user power data, but they are not making full use of such massive datasets to reflect their value. Due to the huge amount of data collected by the power system, the total amount of abnormal data collected cannot be ignored. These abnormal data cause great interference to subsequent data analysis and maintenance; It's also important to find how to identify user's abnormal behaviour based on the power dataset. In this paper, a smart operation and maintenance system based on deep learning is proposed. Based on the composite mathematical statistics method, the data is efficiently reviewed and cleaned. The isolated forest algorithm is used to quickly identify the users with abnormal power consumption and GAN is used to help evaluate this model. This system can help power companies efficiently complete data review and anomaly identification, and can integrate data information from multiple dimensions to guide the actual maintenance and verification.

## 1.Introduction

With the comprehensive promotion and extensive deployment of the smart grid, the grid company has accumulated a large amount of related data in the actual operation and maintenance work, and in the deepening application of those data, whether it is the practical billing and power consumption behaviour characteristic analysis for marketing, or power applications such as anti-stealing and line loss management, both rely on high quality data. Only through accurate and reliable power metering data, can minimize the difference between power generation and power consumption, can prevent the occurrence of power stealing behaviour, can ensure the stable and healthy development of the power market.[1] However, there may be a large number of incorrect values in the collected dataset due to hardware failures and software defects, and the wrong situation is also diverse. It is very important to select high-quality analytical data from a large amount of raw data. Identifying abnormal data from the original dataset can effectively locate and repair hardware failures to improve data quality and the stability of the acquisition system, and also be a basis for subsequent analysis of the user's electricity behaviour [2]. For the data already be cleaned, it is also an urgent problem for the power grid company



to identify the abnormal power users including power stealing. Fully data mining analysis of these datasets can help the power company to better grasp the user's power consumption situation, so as to further allocate power resources more reasonably and improve the allocation efficiency of power resources. At the same time, it can help the power company to enrich the description of the user's power consumption behaviour from multiple dimensions, gradually depict the model of user power usage behaviour and exploit deeper level information, so as to provides decision basis for the next step of operational analysis and other works [3].

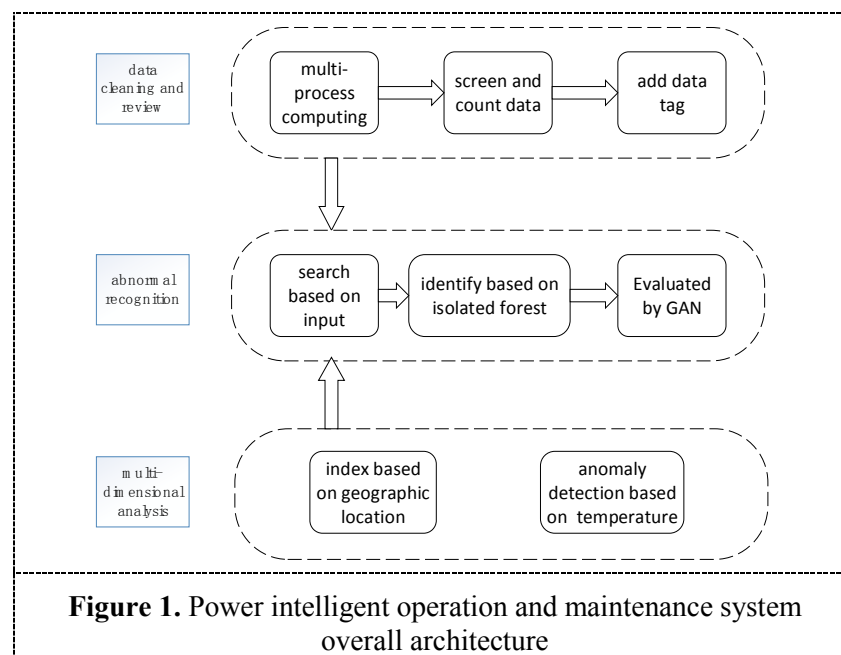
At present, some experts and scholars have carried out related research on the issue of cleaning and screening of power data. As described in literature [4], a time series-based big data cleaning method is proposed to analyse the influence of various outliers on the time series model, then construct a data cleaning model. The literature [5] calculates the load change rate of the electricity consumption data before and after, and statistically removes the data whose rate of change is out of range. The literature [6] uses the Korhonen neural network to dynamically clean the dirty data of power load, and optimizes the algorithm by the idea of fuzzy soft clustering. Most of the methods that have been proposed so far are applied to the data set of a specific situation, and cannot cope with the data screening under normal circumstances. Combined with the subsequent data analysis requirements, we hope that the data cleaning process can filter the noise data and reduce false detection rate as much as possible.

For the problem of abnormal identification, the traditional statistical analysis method cannot meet the requirements in terms of accuracy and efficiency in the face of massive and dynamic measurement data [7]. Many scholars have also conducted research on this issue. Literature [8] proposed an abnormal power detection method based on particle swarm optimization. Based on the user's historical power load data, this method uses the particle swarm algorithm to extract the load pattern curve of the user and the load pattern curve of historical data, and uses the pattern matching method to evaluate the user. Literature [9] proposed the power big data outlier detection and power behaviour based on density peak clustering algorithm. Literature [10] proposed to build an abnormal power detection algorithm model based on neural network. Most of these methods use tagged data and supervised learning. Now the pain point of company to analysis data is the lack of tagged tag data in the dataset. It is difficult to achieve completely supervised learning.

Aiming at the above problems and difficulties, this paper proposes an intelligent operation and maintenance system based on deep learning. In this system, with the composite mathematical statistics method, the data set can be efficiently reviewed, and the isolated forest algorithm is used to quickly identify users with abnormal power usage behaviour and GAN is used to help evaluate this model. This system can also help power companies to integrate multiple dimensions of data information to guide the actual operation and verification of the grid company.

## 2. Overall structure

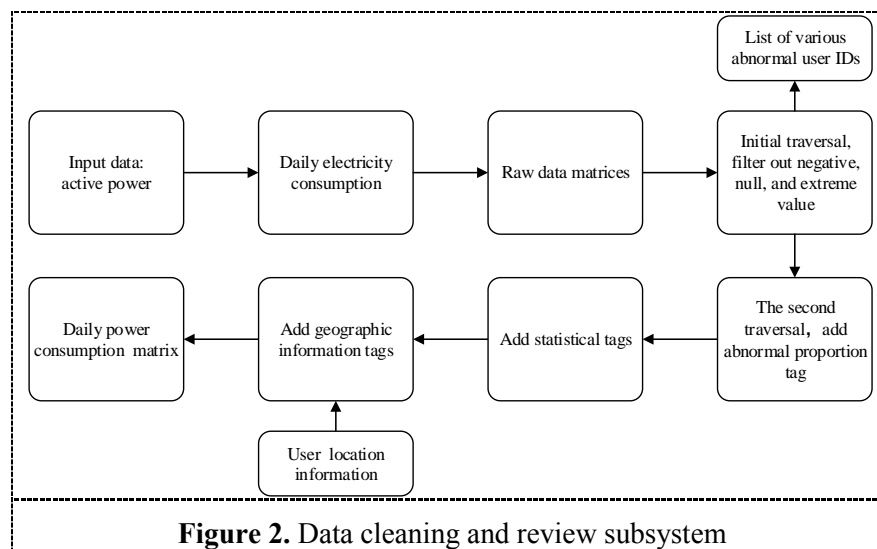
The overall architecture of the system is shown in Figure 1. According to the different stages of data process, it can be further subdivided into three parts which are data cleaning and review subsystem, abnormal recognition subsystem, and multi-dimensional analysis subsystem. The cleaning and review subsystem comprehensively uses the quartile detect [11] and Z score [12] methods in statistics to pre-process and screen the data. This can filter out the noise data as much as possible, such as null, negative, abnormal zero and extreme large values. The data multi-dimensional analysis subsystem can make multi-dimensional correlation analysis of power measurement data based on geography, time and other dimensions, thus to guide the actual operation and maintenance of the power grid company. Finally, the anomaly recognition subsystem performs feature extraction and dimensionality reduction based on the processed data and tags, then through the isolated forest [13] algorithm screens the abnormal users and uses the GAN to evaluate this anomaly detection model. In summary, this system can identify the problem or weak points of the acquisition system, and automatically find out the abnormal power consumption behaviour such as user stealing. Combined with user geographic information, weather temperature and other information, different risk levels and response plans can be given based on the analysis results.



### 3. Cleaning and review subsystem

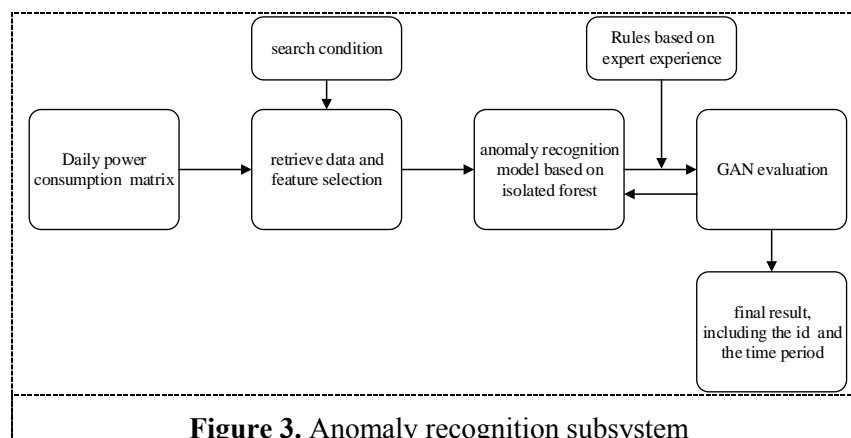
Among the power consumption data collected by the power system, there are many redundant data that are not related to the subsequent analysis work. In addition, data anomalies and data missing are also common in actual power metering data set. The electricity data cleaning and review subsystem is based on the composite mathematical statistics model to process the obvious data errors, perform basic cleaning and review of the original power metering data.

The basic workflow of the power data cleaning and review subsystem is shown in Figure 2. Firstly, extract the data items needed for our analysis from the massive raw data set, then structure and preprocess the data so that the original data can meet the data requirements of the next data analysis module. For the first phase of the user's electricity data processing requirements and objectives, the data review module uses some simple and efficient judgment logic and some methods in mathematical statistics. Most of these rules or logics are universal, such as whether the data content meets the content requirements of the collection, whether the collected data is correct, whether the collected quantities meet the basic quantity relationship, and whether the power data conforms to the basic logic (like the daily usage of electricity collected should be non-negative) etc. As for the statistical method, we mainly use the quartile detection method and the Z score detection method. The screening logic of the two methods is detecting the outliers in the daily electricity data, these data are usually some extreme large values and violent data jitter. Finally, we combined some of the actual operating experience to get some new features and add the features to the data, then output the matrix to the downstream for further analysis.



#### 4. Anomaly recognition subsystem

The original data is without any labels, so we use the isolated forest algorithm as a preliminary anomaly detection tool. It's an unsupervised anomaly detection algorithm and is has no requirement for the dataset. The algorithm can output a subset of the data set that is most likely to be abnormal based on the parameters. After obtaining an approximate positive sample set by the isolated forest algorithm, we can train the GAN, and The resulting generator can be used to help us identify the anomaly. For the data point to be tested, we enter it into the generation network obtained above, and then we get the "normal" version of the data point. By comparing this normal version with the original data point, we can quantify the degree of abnormality of the data points.



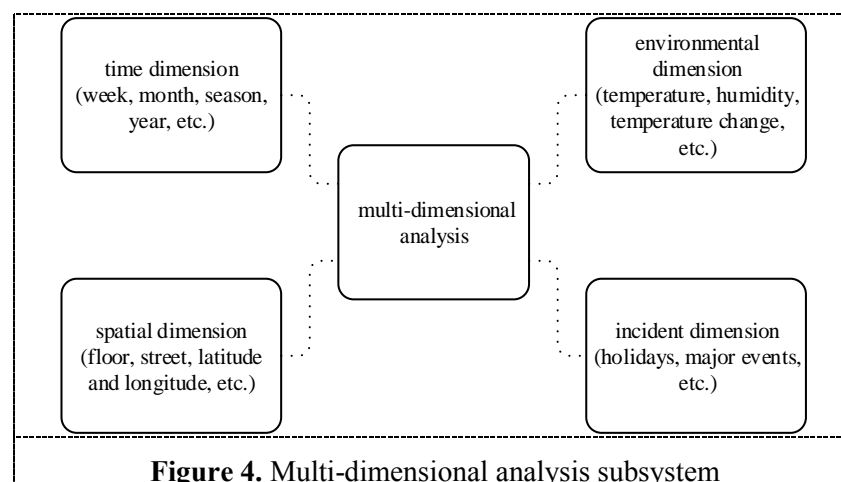
The working flow chart of the anomaly recognition subsystem is shown in Figure 3. The model proposed in this paper includes modules such as feature extraction, feature dimension reduction, and isolated forest calculation and GAN evaluation. Firstly, the user's power consumption data is extracted from the original data set as an initial feature set, then add some self-defined features by statistics and practical experience to enrich the dataset. Then we make feature selection and process the dataset to make it dimensionless. Finally, the isolated forest algorithm is used to calculate the abnormal score of each user to determine whether the user data is abnormal, and the model can be evaluated by the GAN. After obtaining the preliminary judgment result, the obtained abnormal data is fed back to the relevant staff of the power company. Depending on the rules and feedback given by their experience, the data can be used as the labeled data for secondary screening, and get the result which we need. Through the

application of actual data, the system is proved to be efficient and accurate, and has good practical application potential.

### 5. Multi-dimensional analysis subsystem

Multidimensional analysis is based on the statistical principle of data analysis after correlation between different dimensions. When multi-dimensional analysis of power measurement data is carried out, the factors affecting the change of measurement data can be divided into two categories: One is the factor that has long-term effects on the load, and the impact on the load is the long-term trend of the load change, such as economic development, industrial structure changes, etc. The other type is the influencing factors of the short-term effects on the load, such as temperature, rainfall and other climatic factors. Studying the influence of various relevant factors on the changes of electricity load in various industries or regions will help to better identify anomalies and better understand the causes of anomalies, so as to better propose more reasonable operation and maintenance counter measures.

Multi-dimensional analysis subsystem considers measurement data in multiple dimensions. Specifically, it can analyse by time (year, quarter, month, week, day) and area (district, street, floor), and also considers common influencing factors such as temperature and holiday, as shown in the figure. 4.



**Figure 4.** Multi-dimensional analysis subsystem

## 6. Experiment implementation

### 6.1 Data overview

The data set of the experiment comes from the electricity consumption data of about 1.6 million users in a certain area of Shanghai in 2017. The data set also contains data such as address and weather, which is the original feature that describes the user's power usage pattern. Considering the insufficient data and expert experience rules, our experiments only complete the core part of the above system, and there is room for further improvement.

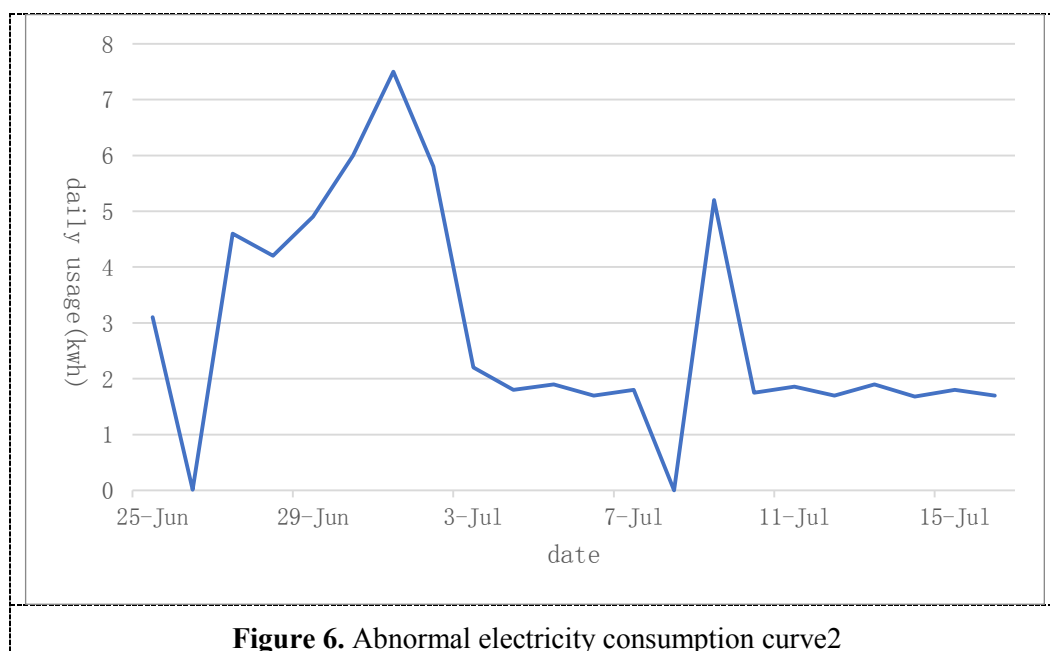
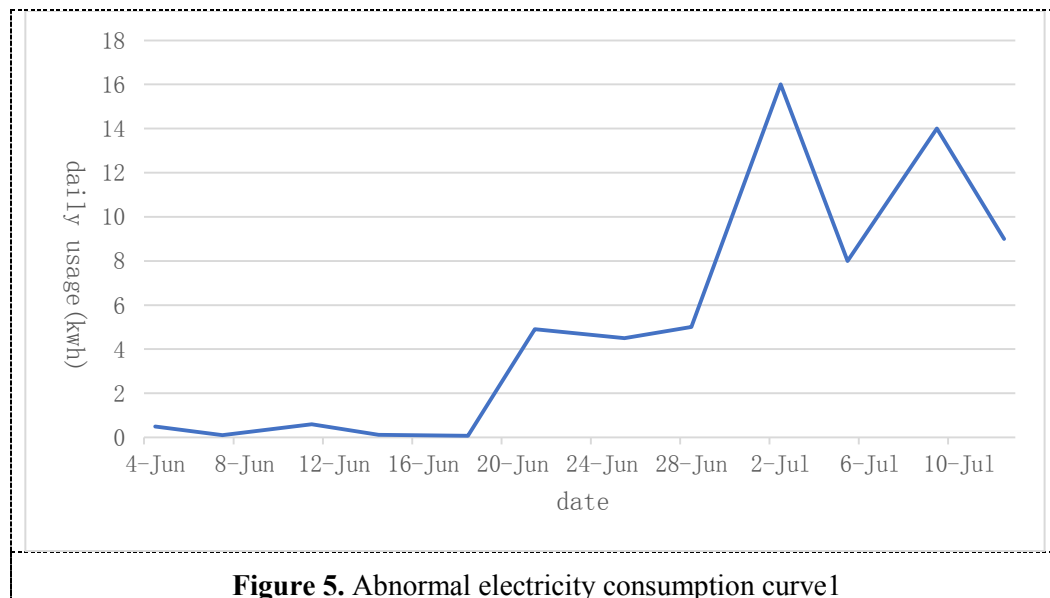
### 6.2 Data preprocessing

Through the processing of the cleaning and review subsystems, we extract and calculate the data items we need, then we filter and correct the following data anomalies: missing data, data jitter and numerical anomalies. After the above two steps, we can get multiple  $M \times N$  data matrices, each of which contains the power consumption data of all users in a certain month. The first column of the matrix is the user's number, and each of the remaining rows represents the user's power usage data for each day of the month. Therefore, the number  $M$  of rows of the matrix is determined by the total number of users in the monthly data, and the number  $N$  of columns of the matrix is determined by the number of days in the month.

### 6.3 Data structuring and feature extraction

Data structuring and feature extraction is done in the multi-dimensional analysis subsystem, the first step is to aggregate the temperature and geographic location information of the original dataset with the electricity consumption data, which can provide an effective data interface for the model to implement more functions in the future. In addition to the daily electricity usage of users, we have proposed some new features. Since the user's power consumption usually has a trend according to the working day cycle, we will count the average power consumption of each user's average power consumption from Monday to Sunday, so that we can get the new seven features.

Moreover, by studying the power consumption curves of some abnormal users who have been manually identified, we found several obvious features as follows in Figure 5 and Figure 6:



The horizontal axis of the above two curves corresponds to time, and the vertical axis is the user's daily power consumption (unit: kWh). From the above two typical abnormal curves, we can use the regular pattern, the sudden rise and sudden drop of power consumption are very important information, we can design features according to this pattern.

Using the idea of a sliding window, a user's power consumption data is treated as a time series and divided into three adjacent windows, denoted as  $w_1$ ,  $w_2$  and  $w_3$ .

The length of  $w_2$  is fixed. We regard it as the main sliding window. Then we calculate the average and standard deviation of the data in  $w_1$  and  $w_3$ , which are recorded as  $avg_1$ ,  $avg_3$ ,  $std_1$ ,  $std_3$ .

The formula for how we calculate the downward trend and the upward trend is as follows:

$$d = \frac{avg_1/(avg_3+0.001)}{std_1 + std_3 + 0.001} \quad (1)$$

$$r = \frac{avg_3/(avg_1+0.001)}{std_1 + std_3 + 0.001} \quad (2)$$

We calculate the maximum values of  $d$  and  $r$  during the sliding process of window  $w_3$  and use it as the downtrend indicator and the uptrend indicator for this time series.

#### 6.4 Abnormal point detection and evaluation

The isolated forest algorithm used in this paper can process large-scale data quickly. This paper applies it to identify abnormal patterns in power consumption data.

The isolated forest algorithm consists of several trees. The tree in the algorithm is called an isolated tree (iTree). When building an iTree, the anomaly is more likely to be isolated, so the number of edges it has experienced from the child to the root is less. Normal points are not easily isolated, so their corresponding paths are longer.

After obtaining several iTrees, the algorithm training ends, and then the generated isolated forest is used to detect the training data. For a training data  $x$ , it traverses each iTree according to the above rules, and calculates the height of the child node where  $x$  is finally located. We can calculate the height average of  $x$  in each tree, denoted as  $E(h(x))$ , and  $h(x)$  is the total number of edges passing from the root node of iTree to the node where  $x$  is isolated. The degree of abnormality can be expressed as:

$$S(x, n) = 2^{-E(h(x))c(n)} \quad (3)$$

Where  $c(n)=2H(n-1) - (2(n-1)/n)$  is the average path length of iTree for normalized output, where  $H(k)=\ln(k)+\xi$ ,  $\xi$  is the Euler constant.

The power side anomaly data has the following two characteristics:

- Abnormal data is only a small amount,
- Abnormal data differs greatly from normal data.

The above two characteristics are the theoretical basis of the isolated forest algorithm. Therefore, when building iTree, the abnormal data is close to the root node, and the normal data is far from the root node. For efficiency reasons, the algorithm defines the depth of the tree:

$$\text{ceil}(\log_2(n)) \quad (4)$$

Unlike other algorithms, isolated forests are an unsupervised algorithm, so excessive sampling of data sometimes reduces the ability of isolated forests to identify anomalous data. Therefore, the usual number of samples is 256. The number of algorithm samples and the depth of the tree are well limited, so even when dealing with large-scale data, there is no great requirement for memory, and it has a



faster processing speed. Practice has proved that it takes only about 25 seconds to process 100,000 user data at a batch.

This paper uses GAN to further evaluate anomalies. A normal data set is used to train generator X. This normal data set should reduce the proportion of abnormal data as much as possible; an artificially defined noise distribution Z can produce the input z of the generator; a potential abnormal data set A, as an input to the trained generator, this preliminary anomaly data is a data set that is filtered from the original data set and has potential exceptions. We mainly train two networks: generator G and discriminator D.

Once we have the data, we start training the model. The training of the model can be divided into two parts: the first part is to train a normal confrontation generation network. The training method is the same as the ordinary GAN. This paper uses wgan-gp based on pytorch. The training process of the network here will not be expanded in detail.

After generating the anti-network training, the generator we got can be used to help us evaluate the anomaly. We convert the preliminary abnormal data obtained in the previous step into a grayscale image, and then generate a "normal map" through the generator. By comparing the difference between the original image and the "normal image", we can intuitively evaluate the abnormality of the original image. We note that the detected image is x, the output of the discriminator is D(x), and the output of the generator is G(x). The abnormal score of the user data can be defined as follow:

$$A(x) = (1 - \alpha) \cdot \sum |x - G(x)| + \alpha \cdot (1 - D(x)) \quad (5)$$

When the actual model is running, selecting different abnormal score thresholds as the evaluation criteria of the anomalies will have a greater impact on the accuracy of the model. Ideally, we can estimate the proportion of abnormal samples to the total sample a% in advance, and then calculate the abnormal scores of each sample according to the model. After sorting the abnormal scores, select the top b% of the samples as anomalies. Here b chooses a value slightly larger than a to bring better results. In the best case, the accuracy of the model is 73.5%.

## 7. Conclusions

The intelligent operation and maintenance system based on deep learning proposed in this paper can be combined with the measurement data acquisition system to improve the data quality and provide a better data foundation for power marketing applications. On the other hand, it can be combined with the existing power metering system operation and maintenance system to guide the actual operation and maintenance work, thereby achieving intelligent operation and maintenance. According to some problems encountered by the power supply company in the process of collecting operation and maintenance, and based on the pain points and difficulties in the actual work, through the efficient cleaning and review, rapid anomaly identification and multi-dimensional data analysis, it solves many problems such as low efficiency of manual inspection and high false positive rate in traditional power company operation and maintenance work, and saves a lot of human resources and costs for power grid operators, which provides a strong guarantee for the healthy development of energy company.

## References

- [1] Tao Wu. Discussion on management of electric energy metering equipment in power grid enterprises [J]. Central China Electric Power 2006(03):34-35.
- [2] Calegari F. Electric power/energy measurements for residential single-phase networks[C]//Industrial Electronics, 2005. ISIE 2005. Proceedings of the IEEE International Symposium on. IEEE, 2005, 3: 1087-1092.
- [3] Haini Qu, Pengfei Zhang, Pin Lin. Research on User Demand Analysis of Power Demand Side Based on Massive Data [C]// East China Six Provinces and One City Transmission and Distribution Technology Seminar. 2015.

- [4] Yingjie Yan, Gehao Sheng, Yufeng Chen. Big data cleaning method for power transmission and transformation equipment state based on time series analysis [J]. Automation of Electric Power Systems, 2015(7):138-144.
- [5] Weiren Mo, Boming Zhang, Hongbing Sun. Application of extended short-term load forecasting method [J]. Power System Technology, 2003, 27(5):6-9.
- [6] Guojiang Zhang, Jiaju Qiu, Jihong Li. Identification and Adjustment of Bad Data of Power Load Based on Artificial Neural Network [J]. Proceedings of the CSEE, 2001, 21(8):104-107.
- [7] Luming Fan. Analysis of the Causes of Abnormal Electric Energy Measurement Data in Electricity Information Acquisition System [J]. Engineering Technology: Citation Edition: 00304-00305.
- [8] Haiming Lu, Zhuangzhi Guo. Abnormal electricity detection method based on particle swarm optimization [J]. Northeast Electric Power Technology, 2016, 37(5):56-59.
- [9] Fengkui Liu. Power Big Data Outlier Detection and Power Behavior Analysis Based on Density Peak Clustering Algorithm [D]. China Electric Power Research Institute, 2017.
- [10] Guang Yang. Research on Common Electric Detection Algorithm Model [J]. Distribution & Utilization, 2016, 33(10):56-59.
- [11] Bin Gong, Haipin Xu. Quartile: A simple exploratory data description statistical tool [J]. Liaoning Economic Statistics, 2004(9):29-30.
- [12] Cheadle C, Vawter M P, Freed W J, et al. Analysis of microarray data using Z score transformation[J]. The Journal of molecular diagnostics, 2003, 5(2): 73-81.
- [13] Liu F T, Ting K M, Zhou Z H. Isolation forest[C]//Data Mining, 2008. ICDM'08. Eighth IEEE International Conference on. IEEE, 2008: 413-422.