Research on Abnormity Detection based on Big Data Analysis of Smart Meter

JINGXUAN FANG[1,2], FEI LIU[2], LINGTAO SU[2], XIANG FANG[2,*]
[1]Yale University,
New Haven, CT 06520,
USA

[2]Suzhou Haoxing Haizhou Technology Co., Ltd,
Suzhou 215000,
CHINA

*Corresponding Author

Abstract: - There are over five hundred million smart meters in China. The current standard for the use of smart meters is physical inspection of meter dismantling within 8 years. The method leads to many issues including high cost of testing, low sampling rate, unknown meter status huge waste of resources etc. Searching for non-dismantling meter detection solution is necessary. Although the smart grid can be managed much better with the increasing use of smart meters, the current standard brings many issues. To solve the problems like a huge waste of resources, detecting inaccurate smart meters and targeting them for replacement must be done. Based on the big data analysis of smart meters, abnormity can be predicted and diagnosed. For this purpose, the method is based on Long Short-Term Memory (LSTM) and a modified Convolutional Neural Network (CNN) to predict electricity usage patterns based on historical data. In this process, LSTM is used to fit the trend prediction of smart meters, and recurrence plot is used to detect the abnormality of smart meter. Both LSTM and recurrence plot method is the first time to be used in smart meter detection. In actual research, many methods including Elastic Net, GBR, LSTM and etc. are used to predict the trend of smart meters. Through the best method LSTM, the accurate rate of the trend prediction of smart meters can arrive at about 96%. Similarly many methods are used to detect the abnormality of smart meters. In single-input modeling, there are sequence-input and matrix-input methods. In dual-input modeling, there are TS-RP CNN, VGG+BiLSTM, ResNet50+1D-CNN and ResNet50+BiLSTM etc. Eventually based on the most successful method recurrence plot, the abnormity testing and failure recognition can be got at 82% roughly. This is the breakthrough in the electricity power domain. With the success of the solution, the service time of a normal meter can be prolonged by abnormity detection. This will lead to saving a lot of resources on smart meter applications.

Key-Words: - smart meter, big data analysis, abnormity detection, deep learning time series model, LSTM, CNN.

Received: August 7, 2023. Revised: May 16, 2024. Accepted: July 2, 2024. Published: July 17, 2024.

1 Introduction

The foundation of the smart grid is a swift two-way intelligent communication network. The efficient and safe operation of the power grid can be obtained through the sensor measurement and control technology etc. Thus the intelligent construction of a power grid can be ensured, [1]. A smart meter is an important part of the smart grid. It touches many parts including measurement, communication and data processing unit etc., [2]. A smart meter is an intelligent toll equipment. It can effectively measure electric parameters and measure two-way electric energy, [3]. Smart meters can achieve real-time data interaction, monitor the quality of electric energy and implement remote monitoring.

In recent years, the smart meter industry in China has developed rapidly, with a continuously expanding market size and broad development prospects. According to the analysis of the smart meter market size, it is predicted that it will grow at a compound annual growth rate of 9.4% in the next five years, from an estimated 23.1 billion US dollars in 2023 to 36.3 billion US dollars in 2028. According to statistics, as of the end of December 2022, the number of smart meters in China has exceeded 650 million. According to the national requirement of mandatory 8-year periodic rotation, the annual replacement volume of smart meters in China is about 80 million, [4].

A smart meter is the smart terminal and data entry of the smart grid. The smart meter has many application features such as two-way multi-rate metering, user terminal real-time control, multiple data transmission modes intelligent interaction, etc. to adapt to the smart grid. Smart grid construction has brought broad market demand for global smart meters, power consumption information collection and processing system products etc., [5], [6] It is estimated that nearly 1.2 billion smart meters will be installed in the world by 2024 based on Wood Mackenzie report. The penetration rate of smart meters will be 60%, [7]. Currently, China has become the largest consumer market for smart meters in the world. Smart meters will also have broad market demand along with the construction and transformation of the Chinese power grid.

At present, the life span of smart meters stipulated by the state is 8 years. But 8 years after the use of smart meters, most of them can be continued to use normally. It can save huge economic costs at least over 30 billion RMB per year for the state and individuals if only abnormal meters are replaced. The national key R&D plan focused on the research on new electromagnetic measurement standards under big data. The purpose of this study is to collect electricity data from the electricity consumption area. The abnormal information on the user's electricity meter can be excavated based on the collected data. Then the abnormal meters can be discovered, [8].

2 Previous Research and Relevant Methods on Smart Meter

2.1 Overview

The main research objective of the paper is to predict and analyze whether the meter is abnormal based on the data of the meter. The abnormalities include failures or stealing electric leakage etc.

Now several related studies have been in progress in China. The difference in user consumption models under different time types is obtained through analysis of electricity consumption types of different users. The abnormal electricity consumption can be detected by using support vector machine theory. This method requires less training time and does not need to classify artificial abnormal data. It can effectively reduce the application cost of the plan, [9].

Besides that, the detection method based on big data in AMI is proposed based on the abnormal increase in CPU utilization rate and network traffic flow rate of the smart meter because of the attack. The CPU load rate and network traffic flow are recorded by smart meters. Then the data and power data are uploaded to the power management center data server. Furthermore the CPU load rate and the network communication traffic are compared by the abnormal load screening system. The abnormal meter with a high CPU load rate and high traffic flow can be identified by utilizing the statistical characteristics of a large number of meter data, [10].

The research on smart meters ends deeper abroad. The research is focused on electricity theft, malware detection, and so on.

A new protocol to deploy the network layer and application layer were proposed a few years ago. The detection accuracy of malicious code can reach 99.72% to 99.96% by checking the semantics and syntax of messages, [11].

The accuracy of the sample can be realized to touch 98.4% through the classification of power data and using the SVM method to train samples, [12].

2.2 Recent Development

Many specialists demonstrate their related work outcomes.

The smart meter reliability life prediction is an important work. And the reliable life prediction of smart meters is extremely important for design, which will bring significant convenience to the maintenance work of smart meters. Based on the analysis of smart meter operating time data, a two-parameter Weibull distribution model can be established to determine the parameters of the two-parameter Weibull distribution. The corresponding distribution function and reliability function are calculated to determine the preventive maintenance cycle at a reliability of 90% and the preventive maintenance cycle corresponding to the minimum maintenance cost, [13].

Communication load balancing of smart meters is also explored based on artificial intelligence algorithms. By analyzing the importance of communication load balancing for smart energy meters and the current communication load situation of smart energy meters, it is found that there are still problems including communication congestion, load imbalance, data loss and errors, and extended response time. By studying the application value of different artificial intelligence algorithms, improvement strategies for communication load balancing for smart energy meters are proposed, including multi-algorithm combination strategy, reinforcement learning strategy, optimized scheduling algorithm, and prediction and dynamic adjustment, providing a reference for achieving more intelligent and efficient load balancing, improving the operational efficiency and stability of power systems, [14].

Online anomaly detection for smart meter data was done a few years ago. The solution includes (1) Data preprocessing: including data cleaning, scaling, and transformation to adapt to model inputs; (2) Anomaly detection model construction: The use of prediction models to learn normal behavior and perform anomaly detection based on prediction errors; (3) Abnormal score calculation: Calculate the abnormal score for each data point based on prediction error and data history; (4) Online learning: When new data arrives, update the model to adapt to the new data and continue learning, [15]. Some research in relevant domains on still image recognition was done, [16]. Some effective online anomaly detection algorithms like the Gaussian Mixture Model were used for vectored area navigation and detecting spectrum access violations etc., [17], [18].

As part of smart grid upgrades, traditional electricity meters are being replaced with smart meters that can improve accuracy, efficiency, and visibility in electrical energy consumption patterns and measurements. However, in most of the deployments, smart meters are only used to digitally measure the energy usage of consumer premises and transmit that data to the utility providers. Despite this, smart meter data can be leveraged into numerous potential applications such as demand side management and energy savings via consumer load identification and abnormality detection. Anyhow, these features are not enabled in most deployments due to high sampling rate requirements, lack of affordable communication bandwidth and resource constraints in analyzing a huge amount of data. The suitability of the embedded edge computing paradigm which not only enriches the functionalities but also overcomes the limitations of smart meters is demonstrated through the relevant study. It achieves significant improvements in accuracy, latency, and bandwidth, [19].

Machine learning can be used in many kinds of industry domains for prediction research. A structural graph-coupled advanced machine learning ensemble model for disease risk prediction is utilized in a tele-healthcare environment, [20].

Some key applications using data analytics, machine learning, and deep learning in health sciences and biomedical data are explored in data analytics in biomedical engineering and healthcare. The areas cover such as predictive health analysis, electronic health records, medical image analysis, computational drug discovery, and genome structure prediction using predictive modeling. Case studies demonstrate big data applications in healthcare using the MapReduce and Hadoop frameworks, [21], [22].

The relevant research method is also used in construction and related industries. For example, heating load and cooling load forecasting are crucial for estimating energy consumption and improvement of energy performance during the design phase of buildings. Since the capacity of cooling ventilation and air-conditioning system of the building contributes to the operation cost, it is ideal to develop accurate models for heating and cooling load forecasting of buildings. A machine-learning technique for the prediction of the heating load and cooling load of residential buildings is proposed. The proposed model is a deep neural network (DNN), which presents a category of learning algorithms that adopt nonlinear extraction of information in several steps within a hierarchical framework, primarily applied for learning and pattern classification. The output of DNN has been compared with other proposed methods such as gradient boosted machine (GBM), Gaussian process regression (GPR) and mini max probability machine regression (MPMR). To develop the DNN model, the energy data set has been divided into training (70%) and testing (30%) sets. The performance of the proposed model was benchmarked by statistical performance metrics such as variance accounted for (VAF), relative average absolute error (RAAE), root means absolute error (RMAE), coefficient of determination (R2), standard deviation ratio (RSR), mean absolute percentage error (MAPE), Nash–Sutcliffe coefficient (NS), root means squared error (RMSE), weighted mean absolute percent error (WMAPE) and mean absolute percentage Error (MAPE). DNN and GPR have produced the best-predicted VAF for cooling load and heating load of 99.76% and 99.84% respectively, [23].

These methods have room to improve though they are valuable. This is because none of them provided good accuracy for large-scale usage. No study has applied any cutting-edge deep learning methods for smart meter malfunction detection, even though deep learning methods have been successfully used for several other malfunction detection problems in recent years.

To avoid huge waste on direct smart meter physical testing or replacement, the new abnormal smart meter detection based on artificial intelligence is necessary for smart grid management in China.

3 Data Preparation and Processing

In this case study, the data used is real historical data of State Grid. Since the data is confidential, here no more discussion is done. Deep learning technology is widely used on analysis and research of data mining on smart meters, [24].

Given the study of the community meter in this project, data processing should be followed as below.

1) Extract the characteristics that affect the meter error as far as possible through various methods of feature engineering;
2) The predicting test is focused on the error of one point/time by building a regression model through the relationship between feature and error. The conclusion is whether the error is within the normal range of the meter. If it is not in the normal range, it is judged whether there is any anomaly in the community meter;
3) If an abnormal meter exists, it will be found according to the electricity characteristic behavior model of the single ammeter.

3.1 Data Format Setting

To the obtained data of the community meter, the relevant parameters are set as below.

U_super_15 is the voltage of the total community meter every 15 minutes;

I_super_15 is the current of the total community meter every 15 minutes;

U_sub_60 is the voltage of the user meter per 60 minutes;

I_sub_60 is the current of the user meter per 60 minutes;

W_sub is the power consumption of user meters per 24 hours;

W_super is the power consumption of the total community meter per 24 hours;

The difference between the total meter and the sum of the user's meter is E as shown in Formula 1.

$$E = W_{super} - \sum_{k=1}^{n} W_{sub_i} \tag{1}$$

The error E can be obtained by UI compared to the electricity meter. The time granularity is one day. There are two kinds of wrong values in data: duplicate value and illegal value.

The current-voltage different accuracy of the total meter and sub-meter in different times is found by data observations. A large number of missing values appear in the sub-meters. The solution is to replace the missing value with integral point current and voltage. Eventually data cleaning is finished by filling in missing values and deleting incorrect values.

3.2 Data Characteristics Analysis

The data used in this study was collected from two residential areas. The smart meter being studied has five technical specifications: 1) a rated power of 1100 W, 2) a rated voltage of 220 V, 3) a rated current of 5 A, 4) a rated frequency of 50 Hz, and 5) an error rate of 2%.

The desensitized data were collected from two residential areas called Hua Yuan (residential area A) and Dong Hui (residential area B). Residential areas A and B collected the voltage readings and the electric current readings of 104 sub-meters every hour from August 2014 to August 2016. In addition, the master meters of residential areas A and B recorded real-time voltage and current every 15 minutes. All the data in this paper were synchronized in corresponding records after data cleaning and preprocessing..

The data of the 2016 Huayuan community and the total data of Donghui garden are ideal after related analysis is done.

Then the distribution of error is analyzed. Time 2016.1.1-3.1 error distribution is shown in Figure 1. Time 2016.3.1-5.1 error distribution is shown in Figure 2. Time 2016.5.1-7.1 error distribution is shown in Figure 3. The time 2016.1.1-2016.7.1 error distribution is shown in Figure 4.

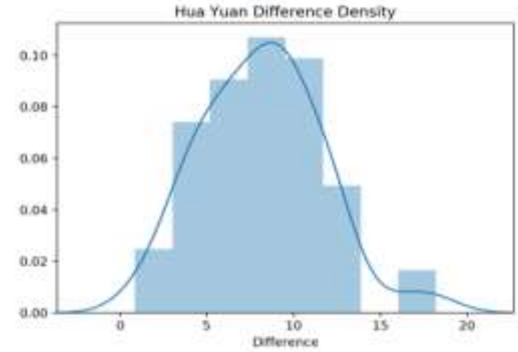


Fig. 1: 2016.1.1-3.1 Error Distribution

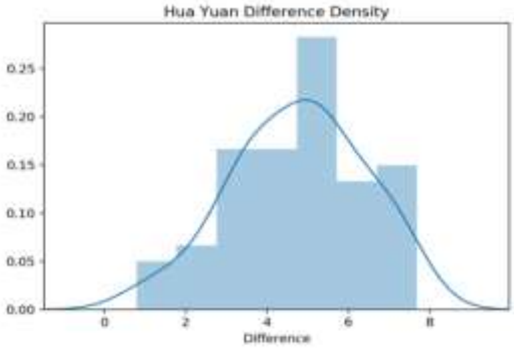


Fig. 2: 2016.3.1-5.1 Error Distribution

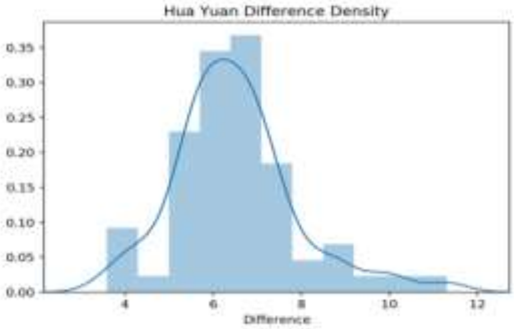


Fig. 3: 2016.5.1-7.1 Error Distribution

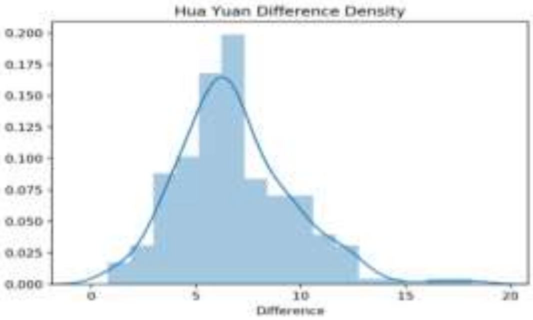


Fig. 4: 2016.1.1-7.1 Error Distribution

3.3 Further Data Processing

To facilitate machine learning analysis, each date of the data adds the following characteristics as shown in Table 1.

To give machine learning more directional suggestions, the correlation coefficient between each dimension is analyzed, [25].

5-fold cross-validation is adopted for data. 1/5 of the data is used for testing. The other data is used for training. Thus five groups of different training sets and test sets can be obtained, [26].

Table 1. Add Data Characteristics

Characteristics	Explanation
Total meter data(super)	Current value of total meter
Difference(error)	Total meter-Sum of sub meter
Relative date(com_date)	The number of days difference between 0 and the base date
Week(week)	Normalization into 7 dimensional vectors
Month(month)	Normalization into 12 dimensional vectors
Year(year)	Normalization into 3 dimensional vectors
Logarithm(log)	The total meter is based on the logarithm value of 2
Number of users(numbers)	Total number of households on that day
Average current(A_mean)	Mean daily current
Average voltage(V_mean)	Mean daily voltage

4 Feature Engineering and Fitting

4.1 Data Reliability Analysis

The reliability needs to be analyzed after getting the data. Firstly different users at the community are selected to analyze the voltage at the same time in the same period. At the same time, the voltage of different users is roughly the same based on the analysis. The trend of voltage change is also similar according to different dates.

Then the change of current is analyzed. A user's meter is selected randomly. The broken line diagram of its current at different times is drawn as shown in Figure 5. The blue line represents 3 a.m. The orange line represents 6 a.m. The green line stands for 6 p.m. It is found that the current value at 3 a.m. is the lowest among the three. The value at 6 a.m. is higher. The value at 6 p.m. is the highest. It can be understood that the electricity consumption is very low at 3 a.m. Most people fall asleep at that time. However, electricity consumption increased at 6 p.m.

The three lines become higher in August. And the gap is smaller. It may be that August is the hottest season in summer. As temperatures rise, users begin to use air-conditioning frequently which result in electricity consumption increasing even at night. It can explain why the current situation is similar at three-time points in August.

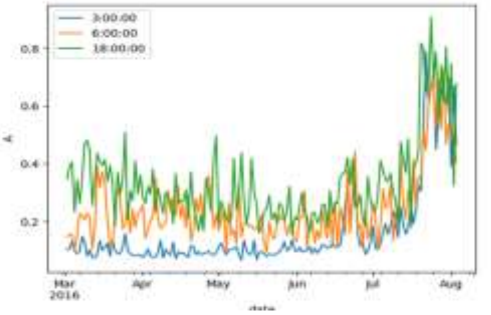


Fig. 5: Electricity Consumption of Same User

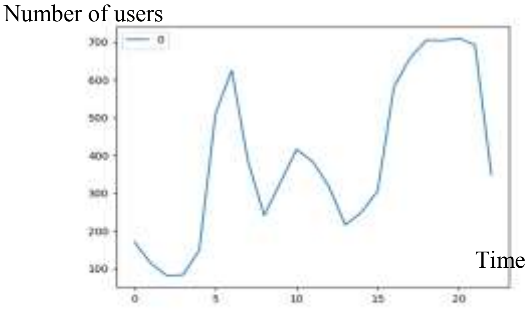


Fig. 6: Number of Users at Different Time of Same Date

Then 0.2A is as a standard. If 0.2A is exceeded, the relevant user can be considered in a state of electricity consumption. Number of users who are in the state of electricity consumption from 0 a.m. to 24 p.m. can be drawn as Figure 6. According to Figure 6, more users are more in the state of electricity consumption at 6 a.m. and 6 p.m. At 3 a.m., there are few users. This indicates that the former hypothesis is correct.

4.2 Outlier Analysis

Outliers in data require additional attention. It often has adverse effects on the result if excluding abnormal values for calculation and analysis. The box diagram can be used to identify outliers in data batches. The box diagram of the user about voltage can be drawn with the Python method, [27]. The box diagram of the voltage for all users within 5 months is shown in Figure 7. It can be found that there are very few exceptions between 0 a.m. and 6 a.m. Then it is difficult to judge whether the abnormal points are excessive because there may be overlapping problems.

The box diagram of the current is shown in Figure 8. However, the outliers are found above the box. Because most electrical appliances are closed at most times, this leads to the average low. It is easier to generate abnormal points above. It is in Figure 8 that 0.2A is used to determine whether users are using electricity.

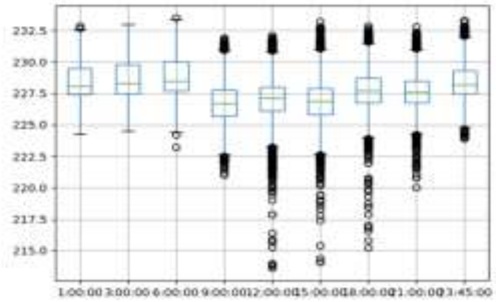


Fig. 7: Box Diagram of Voltage

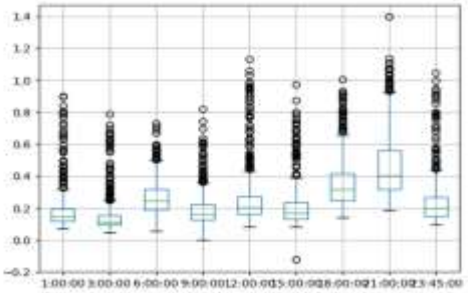


Fig. 8: Box Diagram of Current

4.3 Polynomial Fitting

By using the formula W=U*I*t and then summation can be obtained. Thus the electricity value can be obtained through the calculation by households. Then the electricity value measured by the total meter can deduct. Then the error value of the electric quantity measurement can be obtained. Poly0fit is used to fit the error values previously obtained. The curve is drawn, [28], [29].

The data selection interval is 7 consecutive months or 209 days. Different fitting figures can be made by changing the number of the highest number of polynomials. The following two graphs are the fitted graphs of the highest power of 4 and 8 respectively as shown in Figure 9 and Figure 10. From 4 to 8, the fitting degree increased obviously. The experiment proved that the effect of the highest power from 8 to 10 is not obvious.

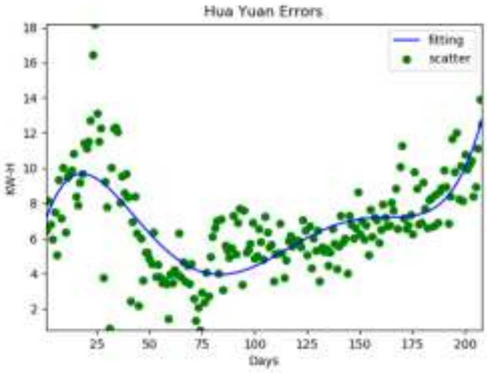


Fig. 9: Fitting the Highest Power to 4 Fitting Curve

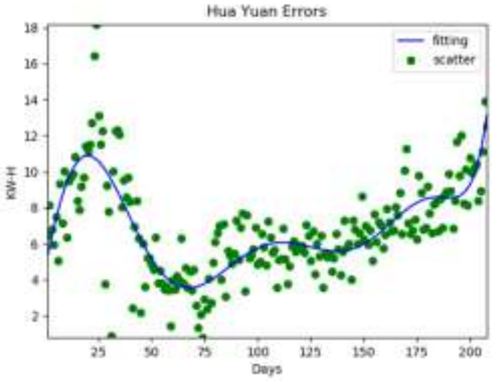


Fig. 10 Fitting the Highest Power to 8 Fitting Curve

5 Deep Learning Time Series Model

To realize smart grid management or abnormal smart meters being replaced only, many deep learning algorithms and models need to be analyzed. Because nobody has done similar work before.

Some models are used to reflect the fluctuations of data in the year and seasons to find the timing of the error between the total meter and the sum of sub-meters based on the preceding analysis. There are two main sources of information for models. They are local features and global features. Long and short-term memory (LSTM) recurrent neural network is an important variant of RNN. It can almost be modeled seamlessly with multiple-input variables. This brings great benefits to time series prediction. This is because classical linear methods are difficult to adapt to multi-variables or multiple-input prediction problems. LSTM models for multivariable time series prediction can be built in Tensorflow and Keras deep learning base.

Recently studies have found that the LSTM model has some advantages in dealing with timing problems. It is because the prediction of future values is based on past values and past values' predictions. It is not just using discrete and uncorrelated features such as season etc. Using the predicted values of past values can make the model more stable. Every step of the training process is cumulative. When a certain extreme error occurs, it may destroy the prediction quality of all subsequent steps, [30], [31].

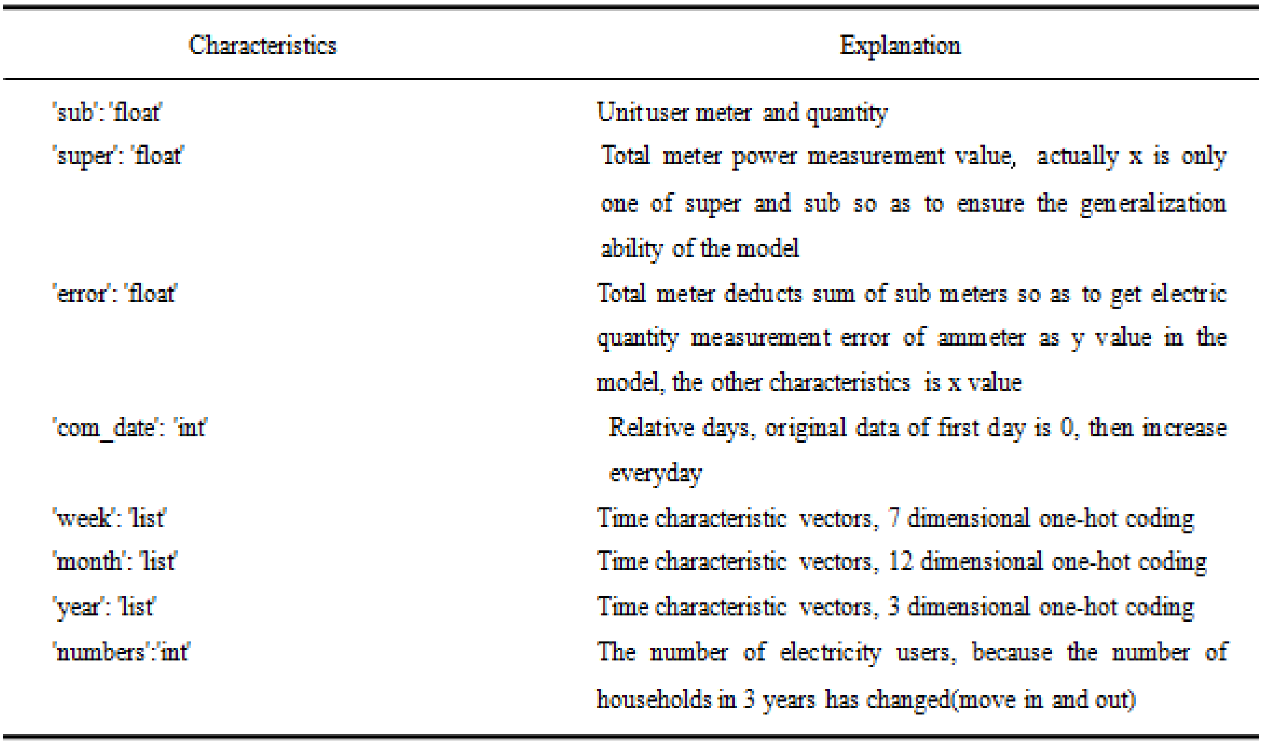
Deep learning time series models include: 1) Transforming original data sets into data sets suitable for time series prediction; 2) Processing data and adapting it to the LSTM model for multivariate time series prediction problems; 3) Making predictions and analyzing the results.

5.1 Time Series Data Preprocessing

Before feature engineering and data processing, many variables related to error have been extracted from the data. This is because the RNN itself is powerful enough for feature extraction. The features and data types used in this model are shown in Table 2.

All characteristics including one-hot coding, x and y etc. are regularized into zero mean and unit variance data. Each characteristic sequence is regularized to this column individually. Commonly used regularization methods are standardization and maximizing regularization. Standardized standardization refers to adjusting the distribution of characteristic data to standard positive distribution, or the mean value of the data is 0 and the variance is 1 called Gauss distribution.

Table 2. Characteristics and Data Types of Deep Learning Time Series Models



Characteristics	Explanation
'sub': 'float'	Unit user meter and quantity
'super': 'float'	Total meter power measurement value, actually x is only one of super and sub so as to ensure the generalization ability of the model
'error': 'float'	Total meter deducts sum of sub meters so as to get electric quantity measurement error of ammeter as y value in the model, the other characteristics is x value
'com_date': 'int'	Relative days, original data of first day is 0, then increase everyday
'week': 'list'	Time characteristic vectors, 7 dimensional one-hot coding
'month': 'list'	Time characteristic vectors, 12 dimensional one-hot coding
'year': 'list'	Time characteristic vectors, 3 dimensional one-hot coding
'numbers':'int'	The number of electricity users, because the number of households in 3 years has changed(move in and out)

The reason for standardization is if the variance of some features is too large, it will lead the objective function so that the parameter estimator cannot learn other features accurately.

Maximizing regularization-MaxAbsScaler makes characteristic distribution within a range between minimum and maximum. In general, it is between [0, 1]. Maximum regularization is specially designed for the scale of sparse data. Standardization is also verified in subsequent experiments without regularization and maximum value regularization. It is found in the experiments that standardized standardization sometimes produces negative numbers. Additional attention should be paid to training and data analysis, [32], [33].

Fixed-length samples are selected randomly for training from the original time series in the model. For example, if the original time series is 600 days in length, then the training sample's time step can be set to 200 days. Thus there will be 400 different starting points. This sampling method is equivalent to an effective data enhancement mechanism. In each step of training, the training program randomly selects the starting point of the timing. It is equivalent to generating infinitely long training data with almost no repetition. The time step is an important super-parameter in this model. When learning sequence predicting problems, LSTM propagates backward through time steps. Then the dedicated time series data sets can be prepared for the LSTM model. This involves the use of data sets as supervised learning problems. Supervised learning problems can be set to include: 1) Electric meter error for the current time (T) is predicted according to the total meter and other inputs of the last period; 2) Forecast of the next hour's meter error is done based on the past day's electricity meter and the forecast for the next hour.

With all 5 years of data, only first-year data is used to fit the model. Then use the remaining 4 years of data to evaluate.

1) The data set is divided into a training set and a test set;
2) The training set and the test set are divided into input and output variables respectively;
3) Reconfigure input (X) to LSTM expected 3D format, or [sample size, time step, characteristics].

5.2 Time Series Data Training

There are two ways to divide training sets and validation sets in time series problems as shown in Figure 11.

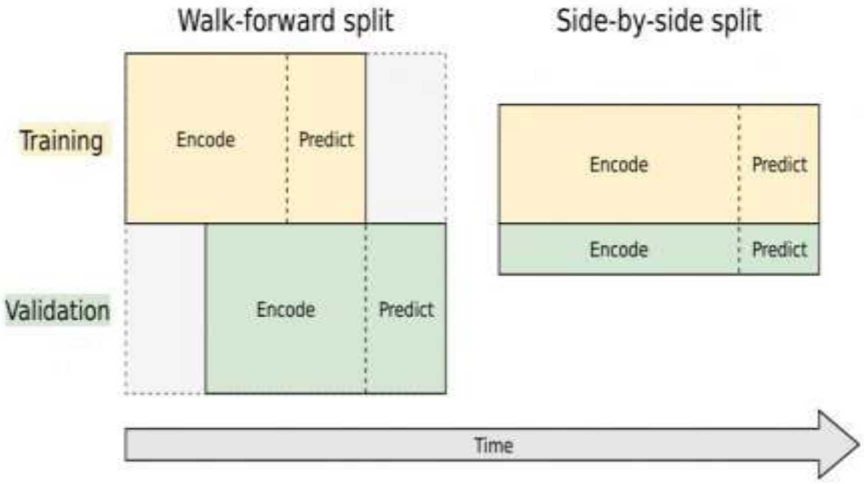


Fig. 11: Training Sets and Validation Sets

5.2.1 Walk-forward Split Method

The walk-forward split method is not dividing data. The complete set of its data sets is also used as training set and validation set. But the different timetables are used for the validation set. Compared to the timetable of the training set, the timetable of the validation set is adjusted to a forth prediction interval.

5.2.2 Side-by-side Split Method

The side-by-side split method is a mainstream way of partitioning. The data set is segmented into independent subsets. One part is for training and the other is for verification.

The result of the walk-forward split method is more impressive. It is more consistent with the final goal of the study. The future value is predicted with historical value. However, this segmentation method has its drawbacks. Because it needs to use data points that are completely used for prediction only at the end of time series. In this way, the trained data points and the predicted data points are longer in time series. It will be difficult to accurately predict future data. If there are 300 days of historical data, the prediction of the next 100 days is wanted. If the walk-forward split partition method is chosen, the first 100 days will be used as training data. The next 100 days are the prediction data in the training process. Then the next 100 days are used as validation sets. So in fact, 1/3 of data points are used in training. There is a 200-day interval between the last training data point and the first prediction data point. This interval is longer. So once the training scenario is left, the quality of prediction will decrease significantly.

If there is only a 100-days interval, the forecast quality will be significantly improved. The side-by-side split method does not consume data points as the predicted data set on the end sequence. However the performance of the model on the validation set will be strongly related to the performance of the training set. And there is no correlation with the real data to be predicted in the future. Therefore, dividing data in this way has no substantive effect. It only repeats the model loss observed on the training set.

The validation set that is divided by the walk-forward split method is used to tune parameters in this model only. The final prediction model must be run without any correlation with the training set and validation set.

The actual research is to define LSTM with 40 neurons in the first hidden layer. In the output layer, 1 neuron for error prediction is defined. The input data dimension will be 1 sample with 22 characteristic time steps of 40 days.

The root mean square error (RMSE) loss function and the efficient random gradient descent version of Adam are used in practical research. The model will be suitable to apply to 1000 epochs and with size 128 training. When choosing the number of epochs, it is not clear which step of the model training is the most suitable for predicting the future(Because the validation set based on current data is very weak about future data.). So the training cannot stop too early. After the experiment, the number of epochs was selected as 1000.

After the model fitting, the whole test data set can be predicted. Combining prediction with test data sets, the scale of test data sets is adjusted. The expected error is also used to adjust the size of the test data set. The error fraction of the model can be calculated through the initial and actual values. In this case, the root mean square error (RMSE) of the unit error with the same variable can be calculated. RMSE is the loss function of a commonly used regression problem. It is different from cross entropy loss suitable for classification problems. In this research, it can get losses quickly and the results are very smooth everywhere.

5.3 Result Analysis and Anomaly Detection

The prediction results of double-layer LSTM and single-layer LSTM are tested. The number of hidden layer units is 20 and 40 respectively. From Table 3, among several time series models, LSTM is the best. The trend prediction accuracy of smart meters is nearly 96%. The result is shown in Figure 12 and Figure 13.

Table 3. Smart Meter Trend Prediction on Deep Learning Time Series Models

Threshold	**Classical Methods**		**LSTM**
	Elastic Net	GBR	
0.5	1 (1.4%)	5 (6.9%)	5 (6.9%)
1	1 (1.4%)	13 (18.1%)	17 (23.6%)
4	13 (18.1%)	52 (72.2%)	42 (58.3%)
6	52 (72.2%)	58 (80.6%)	65 (90.3%)
8	66 (91.7%)	59 (81.9%)	69 (**95.8%**)

Comparing the predicted values with the real values in the time dimension, the first few predictions can be found more accurate. At the same time, the predicted value can be found to have a certain lag. In the scatter plot, the closer the y=x line is, the more accurate the prediction is.

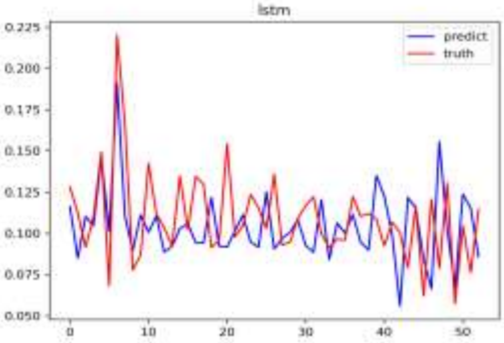


Fig. 12: Prediction and Real Value in LSTM Model

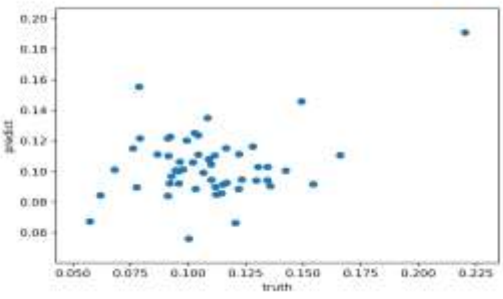


Fig. 13: Scatter Plot of LSTM Model

The sliding window is used to detect the anomaly. A sliding window is set with a width of 1 and the threshold of t.

To each sliding, if the difference between each predicted and actual value in the window exceeds the threshold value, it is believed that errors appear on this day. When the window width is 4 and the threshold is 0.5, the result is shown in Figure 14. From the sixty-fifth day, data can be detected anomalies. By adjusting the width and threshold, higher precision results can be obtained.

From Table 4, many kinds of methods are tested and compared. The result of the time series recurrence plot CNN(TS-RP CNN) is best, [34].

When the data is not abnormal, the result is shown in Figure 15. For smart grid operators, it is easy for them to find abnormal smart meters on monitoring screen based on LSTM algorithm analysis to voltage, current, and electricity consumption. After finding the abnormal meter, the relevant replacement or maintenance can be done.

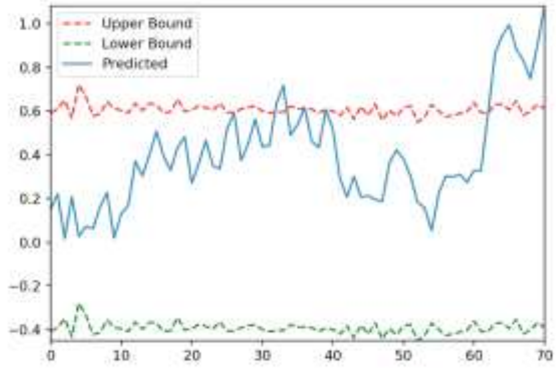


Fig. 14: Related Abnormal Date by Calculation

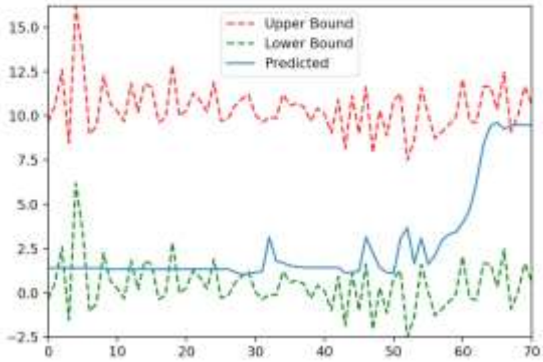


Fig. 15: Predicted Value within the Range

Table 4. Smart Meter Abnormity Detection on Deep Learning Time Series Models

AUC of ROC curve	**Single-input Modeling**		**Dual-input Modeling**			
Fold number	Sequence-input	Matrix-input*	TS-RP CNN (VGG*+1D-CNN)	VGG* + BiLSTM*	ResNet50* + 1D-CNN	ResNet50*+BiLSTM
Fold 1	0.29	0.47	0.83	0.91	0.65	0.37
Fold 2	0.62	0.71	0.80	0.75	0.58	0.45
Fold 3	0.46	0.17	0.74	0.71	0.43	0.54
Fold 4	0.68	0.41	0.80	0.80	0.57	0.52
Fold 5	0.55	0.79	0.94	0.87	0.30	0.48
Mean (±1 std.)	**0.52±0.14**	**0.51±0.22**	**0.82±0.07**	**0.81±0.07**	**0.51±0.13**	**0.60±0.13**

6 Discussion and Conclusion

There are many time series forecast applications in different domains. For example, it is an R package ForecastTB that can be used to compare the accuracy of different forecasting methods as related to the characteristics of a time series dataset, [35]. However, LSTM is the first time to be used in power grid smart meter analysis.

In smart meter trend prediction LSTM demonstrates its advantage and meets nearly 96% accuracy. TS-RP CNN is the best in finding the abnormity of smart meters. Since the research is based on State Grid historical data, it is implicated that the relevant method is effective for actual smart grid management. The efficiency of smart grid operators could be improved greatly.

From the case study, the research method based on the deep learning model is effective. With the popularity and application of smart meters in China, the prediction and detection of abnormal meters will be more accurate based on more big data support. It will increase the service life of our normal meters in the future, thus saving a lot of resources. The research method also can be used in relevant industries like water meters and gas meters for reference.

Acknowledgement:
This research was got data and finance support by the National Institute of Metrology.

References:

[1] Qilin Li, Mingtian Zhou. Research on dependable distributed systems for smart grid[J]. *Journal of Software*, Vol. 7, No. 6, June 2012, pp.1250-1257. DOI: 10.4304/jsw.7.6.

[2] Yun Li, Ben Jones. The Use of Extreme Value Theory for Forecasting Long-Term Substation Maximum Electricity Demand[J]. *IEEE Transactions on Power Systems*, Vol. 35, Issue 1, January 2020, pp. 128-139. DOI: 10.1109/TPWRS.2019.2930113.

[3] Yaxian Zheng, Zhenglin Yang, Guangyao Zhang, Xian Zhang. The pattern comparison and optimization model of inter-regional transactions in Smart Grid[C]. *2012 IEEE Innovative Smart Grid Technologies - Asia (ISGT Asia)*, Tianjin, China. DOI: 10.1109/ISGT-Asia.2012.6303342.

[4] Chinabgao. 2024 smart meter market size analysis: The number of China smart meter install base has surpassed 0.65 billion (报告大厅(www.chinabgao.com) 2024年智能电表市场规模分析：国内智能电表保有量已超过6.5亿只), [Online]. https://www.chinabgao.com/info/1248955.html (Accessed Date: March 5, 2024).

[5] Kerry D. McBee, Marcelo G. Simoes. Utilizing a Smart Grid Monitoring System to Improve Voltage Quality of Customers[J]. *IEEE Transactions on Smart Grid*, Vol. 3, Issue 2, June 2012, pp. 738-743. DOI: 10.1109/TSG.2012.2185857.

[6] Gert Rietveld, Jean-Pierre Braun, Ricardo Martin, Paul Wright, Wiebke Heins, Nikola Ell, Paul Clarkson, Norbert Zisky. Measurement Infrastructure to Support the Reliable Operation of Smart Electrical Grids[J]. *IEEE Transactions on Instrumentation and Measurement*, Vol. 64, Issue: 6, June 2015, pp.1355-1363. DOI: 10.1109/TIM.2015.2406056.

[7] Fangxing Liu, Chengbin Liang, Qing He. A Data-Based Approach for Smart Meter Online Calibration[J]. Acta IMEKO. Vol. 9, No. 2(2020): 32-37. DOI: 10.21014/acta_imeko.v9i2.777.

[8] B. Qu, Z Wang, B Shen, H Dong. Decentralized dynamic state estimation for multi-machine power systems with non-

Gaussian noises: Outlier detection and localization[J]. *Automatica*, Vol. 153, July 2023. DOI: 10.1016/j.automatica.2023.111010.

[9] Q. He, F Liu, L Wang, H Huang, Z Jia. Smart Meter Working Status Evaluation Method Based on Evidence Theory[C]. 20*18 International Conference on Precision Electromagnetic Measurements*, July 8-13, 2018, Paris, France. DOI: 10.1109/CPEM.2018.8501081.

[10] Reza Zamani, Mohsen Parsa Moghaddam, Mahmoud-Reza Haghifam. Evaluating the Impact of Connectivity on Transactive Energy in Smart Grid[J]. *IEEE Transactions on Smart Grid*, Vol. 13, Issue 3, May 2022, pp. 2491-2494. DOI: 10.1109/TSG.2021.3136776.

[11] Babu V, Nicol D M. Detection of x86 malware in AMI data payloads[C]. *2015 IEEE International Conference on Smart Grid Communications*, Miami, FL, USA, 2015: 617-622. DOI: 10.1109/SmartGridComm.2015.7436369.

[12] Soma Shekara Sreenadh Reddy Depuru, L Wang, V Devabhaktuni. Electricity theft: Overview, issues, prevention and a smart meter based approach to control theft[J]. *Energy Policy*, 2011. Vol. 39, Issue 2, pp.1007-1015. DOI: 10.1016/j.enpol.2010.11.037.

[13] Li Weibo, Su Wenbin, Xu Chenghu, Zhang Maojie, Fang Hualiang. Maintenance cycle prediction method for smart electricity meters based on Weibull distribution with economy and high reliability[J]. *Electrical Engineering*, 2023, Vol. 24, Issue(1):17-22. (李维波,苏文斌,徐成虎,张茂杰,方华亮.基于威布尔分布的经济性与高可靠度智能电表维修周期预估算法[J]. 电气技术，2023, 24(1):17-22.)

[14] B. Qu, Z Wang, B Shen, H Dong, X Zhang. Secure Particle Filtering With Paillier Encryption–Decryption Scheme: Application to Multi-Machine Power Grids[J]. *IEEE Trans. Smart Grid*, 15(1): 863-873 (2024). DOI: 10.1109/TSG.2023.3271949.

[15] S. Nielsen, Scalable prediction-based online anomaly detection for smart meter data[J], *Information Systems*, 77 (2018) 34 – 47. DOI: 10.1016/j.is.2018.05.007

[16] Dey A, Biswas S, Le DN. Recognition of Human Interactions in Still Images using AdaptiveDRNet with Multi-level Attention[J], *International Journal of Advanced Computer Science and Applications*, 2023, Vol.14, No.10:984-994. DOI: 10.14569/IJACSA.2023.01410103.

[17] Choi HC, Deng C, Park H, Hwang I. Gaussian Mixture Model-Based online anomaly detection for vectored area navigation arrivals[J], *Journal of Aerospace Information Systems*, 2023, 20(1):37-52. DOI: 10.2514/1.I011128.

[18] Pramitha Fernando, Keshawa Dadallage, Tharindu Gamage, Chathura Seneviratne, An Braeken, Arjuna Madanayake, Madhusanka Liyanag. Distributed-Proof-of-Sense: Blockchain Consensus Mechanisms for Detecting Spectrum Access Violations of the Radio Spectrum[J], *IEEE Transactions on Cognitive Communications and Networking*, 2023, Vol. 9, Issue 5: 1110-1125. DOI: 10.1109/TCCN.2023.3291366.

[19] Sirojan, T., Lu, S., Phung, B. T., & Ambikairajah, E. (2019, September). Embedded edge computing for real-time smart meter data analytics[C]. *Proceedings of 2019 International Conference on Smart Energy Systems and Technologies (SEST)*, Portugal (pp. 1-5). IEEE. DOI: 10.1109/SEST.2019.8849012.

[20] Roy, S. S., Samui, P., Deo, R., & Ntalampiras, S. (Eds.). *Big data in engineering applications* [M]. October 1, 2018, Berlin/Heidelberg, Germany: Springer. DOI: 10.1007/978-981-10-8476-8.

[21] Lee, K. C., Roy, S. S., Samui, P., & Kumar, V. (Eds.). [M]. 1st Edition, October 16, 2020, Academic Press, ELSEVIER. DOI: 10.1016/C2018-0-05371-2.

[22] Rachna Kulhare, S. Veenadhari. QLGWONM: Quantum Leaping GWO for Feature Selection in Big Data Analytics[J]. Harbin Gongye Daxue Xuebao, *Journal of Harbin Institute of Technology*, Vol. 30, Issue 4, 2023, pp.85-98. DOI: 10.11916/j.issn.1005-9113.2022026.

[23] Roy, S. S., Samui, P., Nagtode, I., Jain, H., Shivaramakrishnan, V., & Mohammadi-Ivatloo, B. (2020). Forecasting heating and cooling loads of buildings: A comparative performance analysis[J]. *Journal of Ambient Intelligence and Humanized Computing*, 11(3), 1253-1264. DOI: 10.1007/s12652-019-01317-y.

[24] Chen Liang, Huang Youpeng, Lu Tao, Dang Sanlei, Zhang Jie, Zhao Wen, Kong Zhengmin. Remote error estimation of smart meter based on clustering and adaptive gradient descent method [J]. *Journal of*

Computational Methods in Sciences & Engineering, Vol. 22, No. 1(2022): 207-217. DOI：10.3233/JCM-215901.

[25] Chen Liang, Huang Youpeng, Lu Tao, Dang Sanlei, Kong Zhengmin. Metering equipment running error estimation model based on genetic optimized LM algorithm[J]. *Journal of Computational Methods in Sciences & Engineering*, 2022, 22(1): 197-205. DOI：10.3233/JCM-215896.

[26] Lyu Z., Yu Y., Samali B., Rashidi M., Mohammadi M., Nguyen T.N., Nguyen A. Back-Propagation Neural Network Optimized by K-Fold Cross-Validation for Prediction of Torsional Strength of Reinforced Concrete Beam[J]. *Materials* 2022, 15(4), 1477. DOI: 10.3390/ma15041477.

[27] Brett Slatkin. Effective Python[M]. Publishing House of Electronics Industry, Beijing, 2016.

[28] Sulaiman S. M., Aruna Jeyanthy P., Devaraj D. Smart Meter Data Analysis Using Big Data Tools[J]. *Journal of Computational and Theoretical Nanoscience*, 2019, 16(8): 3629-3636. DOI：10.1166/jctn.2019.8338.

[29] Ji Fengxian, Yao Weixing. Weighted Least Square Method for S-N Curve Fitting [J]. *Transactions of Nanjing University of Aeronautics and Astronautics*, 2004, No. 1:53-57.

[30] Wang Yi, Chen Qixin, Hong Tao, Kang Chongqing. Review of Smart Meter Data Analytics: Applications, Methodologies, and Challenges[J]. *IEEE Transactions on Smart Grid*, 2019, 10(3): 3125-3148. DOI: 10.1109/TSG.2018.2818167.

[31] Yikuai Wang, Huadong Qiu, Ying Tu. A Review of Smart Metering for Future Chinese Grids[C]. *2018 Applied Energy Symposium and Forum*, 2018-06-05, Shanghai, China. DOI：10.1016/j.egypro.2018.09.158.

[32] Ibrahim Yasser, Mohamed A. Mohamed, Ahmed S. Samra, Fahmi Khalifa. A Chaotic-Based Encryption/Decryption Framework for Secure Multimedia Communications [J]. *Entropy*, 2020(22), 11: 1253-1276. DOI: 10.3390/e22111253.

[33] Miller Clayton, Meggers Forrest. Mining electrical meter data to predict principal building use, performance class, and operations strategy for hundreds of non-residential buildings [J]. *Energy and buildings*, 2017(156), 12: 360-373. DOI: 10.1016/j.enbuild.2017.09.056.

[34] Ming Liu, Dongpeng Liu, Guangyu Sun, Yi Zhao, Duolin Wang, Fangxing Liu, Xiang Fang, Qing He, Dong Xu. Deep Learning Detection of Inaccurate Smart Electricity Meters: A Case Study[J]. *IEEE Industrial Electronics Magazine*, 2020, Issue 12:79-90. DOI：10.1109/MIE.2020.3026197.

[35] Neeraj Dhanraj Bokde, Zaher Mundher Yaseen and Gorm Bruun Andersen. ForecastTB—An R Package as a Test-Bench for Time Series Forecasting—Application of Wind Speed and Solar Radiation Modeling[J]. *Energies*, 2020, Issue 13, pp.2578-24. DOI: 10.3390/en13102578.

Contribution of Individual Authors to the Creation of a Scientific Article (Ghostwriting Policy)

- Jingxuan Fang, Xiang Fang carried out the structure design of paper and the optimization.
- Fei Liu has organized and executed the experiments of data handling relevant to the paper.
- Lingtao Su has implemented the Algorithm research and the figure/table organization.

The authors equally contributed in the present research, at all stages from the formulation of the problem to the final findings and solution.

Sources of Funding for Research Presented in a Scientific Article or Scientific Article Itself

The research was got funding support of National Institute of Metrology and Suzhou Haoxing Haizhou Technology Co., Ltd.

Conflict of Interest

The authors have no conflicts of interest to declare.