Intelligent Anomaly Detection Method of Gateway Electrical Energy Metering Devices using Deep Learning

Lihua Zhang[1]*, Xu Chen[2], Chao Zhang[3], Lingxuan Zhang[4], Binghang Zou[5]
Marketing Service Center (Metrology Center), State Grid Ningxia Electric Power Co. Ltd, Yinchuan, China[1,2,3]
College of Electrical Engineering, Sichuan University, Chengdu, China[4,5]

***Abstract*—Accurate anomaly detection of gateway electrical energy metering device is important for maintenance and operations in the power systems. Traditionally, anomaly detection was typically performed manually through the analysis of the collected energy information. However, the manual process is time-consuming and labor-intensive. In this condition, this paper proposes a hybrid deep-learning model, which integrates Stacked Autoencoder (SAE) and Long Short-Term Memory (LSTM), for intelligently detecting the abnormal events of gateway electrical energy metering device. The proposed model named SAE-LSTM model, first uses SAE to extract deep latent features of three-phase voltage data collected from the gateway electrical energy metering device, and then adopts LSTM for separating the abnormal events based on the extracted deep latent features. The SAE-LSTM model, can effectively highlight the temporal information of the electrical data, thereby enhancing the accuracy of anomaly detection. The simulation experiments verify the advantages of the SAE-LSTM model in anomaly detection under different signal-to-noise ratios. The experimental results of real datasets demonstrate that it is suitable for anomaly detection of gateway electrical energy metering devices in practical scenarios.**

***Keywords*—Anomaly detection; gateway electric energy metering device; stacked autoencoder; long short-term memory**

I. Introduction

The importance of anomaly detection in gateway electrical energy metering device lies in ensuring the accuracy and reliability of energy measurement. The gateway electrical energy metering devices play a crucial role in power systems as they are utilized to measure and record energy consumption. The significance of anomaly detection in gateway electrical energy metering devices extends to various aspects such as data accuracy, system safety, and energy management. Anomalies occurring in these devices can result in inaccurate energy consumption data, thereby impacting the billing and settlement processes between energy suppliers and consumers. Moreover, anomalies can serve as indicators of underlying issues or faults within the power system, and their timely detection can unveil potential problems. By promptly identifying and addressing anomalies, it becomes possible to ensure the accuracy and reliability of energy measurement, improve energy management efficiency, and guarantee the safe operation of the power system[1]. Presently, the detection of abnormal operating states in gateway electrical energy metering devices heavily relies on manual on-site inspections, which pose safety risks, have lengthy detection cycles, and may not promptly identify faults[2-4]. With the increasing number of gateway electrical energy metering devices, manual inspections necessitate greater human and material resources, making it challenging to fully meet the current requirements for metering device management. Hence, it is imperative to propose and establish a anomaly detection system specifically designed for gateway metering devices. This system should employ suitable anomaly detection algorithms to promptly identify abnormal states.

Currently, methods for anomaly detection can be categorized into three distinct classes. The first class comprises statistical-based detection methods, including Gaussian distribution[5], probability density functions[6], clustering algorithms[7], and Markov models[8]. Although statistical-based methods are grounded in solid theoretical foundations, the task of selecting an appropriate distribution to effectively discriminate normal instances from anomalous ones poses significant challenges. The second class encompasses rule-based detection methods, involving the establishment of thresholds and the utilization of rule engines, among others. Rule-based methods offer ease of implementation and interpretation, but their ability to detect more intricate anomalies may be constrained. The third class encompasses deep learning-based anomaly detection methods, such as those relying on convolutional neural networks[9,10]. These methods extract robust latent features; however, they necessitate the conversion of input data into images, thereby augmenting the data processing burden, while inadequately considering the influence of network structure information on the accuracy of feature extraction[11]. Although variant models based on the support vector machines (SVM)[12] demonstrate commendable performance in non-temporal data processing, their accuracy in handling complex time series data still needs to be improved.

With the rapid developments in artificial intelligence (AI), numerous approaches using machine learning (ML), especially deep learning (DL), have been proposed to overcome these challenges. For example, Lee S et al. used a self-encoder consisting of a graph convolutional network and a bidirectional long short-term memory network to detect anomalies in smart detection data with higher accuracy than a single LSTM network and with reduced power cost and grid power supply[13]. However, the graph convolution operation of the graph convolutional neural network becomes difficult when processing sequential data and suffers from dimensional mismatch. Wang et al. provided a semi-supervised learning based power anomaly detection strategy[14]. Their proposed framework not only detects anomalous power patterns in real time,

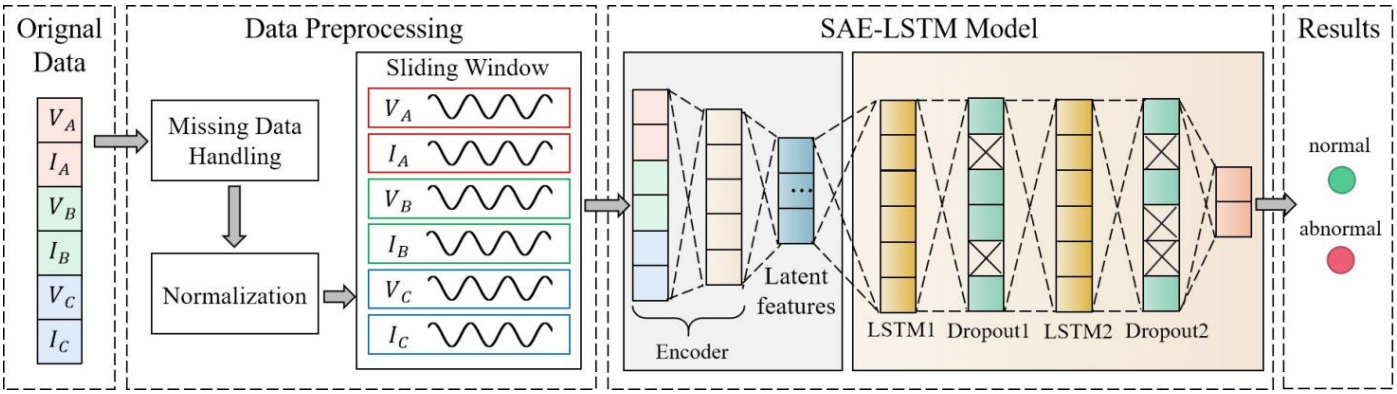

Fig. 1. The flowchart of the intelligent anomaly detection method of gateway electrical energy metering devices using deep learning.

but also identifies suspicious power usage that is inconsistent with customers' lifestyles and typical daily routines, but is not suitable for application to anomaly detection at grid gate metering devices. Hussain et al. have proposed an unsupervised detection approach aimed at identifying power theft behaviors without data labeling costs[15]. Their proposed method is evaluated using accuracy and detection rate.

Although most of the current models perform well, there are several limitations.

First, three-phase voltage data is often noisy and complex, making it difficult to distinguish normal fluctuations from actual anomalies. Second, power systems are dynamic and their operating conditions may change rapidly. Anomaly detection methods must be adaptable and able to evolve with these changes. Finally, many sophisticated machine learning-based techniques can effectively detect anomalies, but often lack interpretability.

To address the above problem, in this paper, we present a novel anomaly detection model that combines the Stacked Autoencoder (SAE) and Long Short-Term Memory (LSTM) networks with elastic network regularization. Initially, the model extracts latent feature representations of the data using SAE and subsequently employs the LSTM algorithm for classification purposes. Following the training of the SAE-LSTM network, elastic network regularization is applied to fine-tune the model using a composite loss function. Bayesian optimization techniques are then utilized to determine optimal hyperparameter values. Lastly, the model is evaluated using a performance assessment metric. Experimental results demonstrate that the proposed model effectively considers the temporal dependencies of the data, leading to improved detection accuracy. It is well-suited for detecting anomalies in gateway power metering devices and effectively assessing their operational state.

The contributions in this paper can be summarized as follows.

(1) This paper presents a novel anomaly detection model that combines the Stacked Autoencoder (SAE) and Long Short-Term Memory (LSTM) networks with elastic network regularization.

(2) Apply the SAE layer for feature extraction. The application of the encode layer allows the proposed model to extract high-level temporal features more efficiently while reducing model performance's dependence on data processing, thus improving the accuracy and efficiency of anomaly detection.

(3) Apply LSTM layer for classification. In the voltage anomaly detection task, the LSTM model can capture the dynamic characteristics of voltage and current signals and is suitable for processing time series data.

(4) The short-term patterns and dependencies in the three-phase voltage and current time series data can be captured using the sliding window technique, and at the same time, by analyzing the sequence within a smaller window, it can help reduce noise in the simulated data and improve the SNR in the data. This further improves the feature extraction efficiency of the proposed model, enabling more accurate anomaly detection.

(5) A real dataset collected from a power grid is applied to evaluate the effectiveness and applicability of the proposed model.

The rest of this paper is organized as follows. Section II describes the Methods. Section III applies the method in a real case and analyzes the results.Section IV discusses the advantages and shortcomings of the proposed method, and the potential future work and concludes the paper.

II. Methods

Accurate anomaly detection of gateway electrical energy metering device is important for maintenance and operations in the power systems. In this paper, a hybrid deep-learning model is proposed to intelligently detecting the abnormal events of gateway electrical energy metering device. The overall process flowchart of the proposed method is shown in Fig. 1.

A. Data Preprocessing

During the data preprocessing stage, three key procedures are utilized to increase the validity of the electrical energy data. These procedures handle missing values, normalize the data, and implement sliding windows on the data.

1) Missing value handling: The data collected from gateway electrical energy metering devices may partial be lost due to various factors, such as human error or equipment issues[16]. In order to address the problem of missing values,

several methods exist depending on the data nature and desired outcome. Common strategies involve deletion, imputation, and machine learning techniques. Although direct deletion is a straightforward approach, it may lead to the loss of valuable information, reduction in sample size, and decreased efficiency. Conversely, the utilization of machine learning methods can be excessively intricate. Therefore, Lagrange interpolation is employed to estimate the missing values. The formula for Lagrange interpolation is specified as follows:

$$l_j(x) = \prod_{\substack{i=0 \\ i \neq j}}^{n} \frac{x - x_i}{x_j - x_i} = \frac{x - x_0}{x_j - x_0} \cdots \frac{x - x_{j-1}}{x_j - x_{j-1}} \frac{x - x_{j+1}}{x_j - x_{j+1}} \cdots \frac{x - x_n}{x_j - x_n} \tag{1}$$

$$L(x) = \sum_{j=0}^{n} y_j l_j(x) \tag{2}$$

Equation (1) represents the Lagrange basic polynomial for n+1 data points, while Equation (2) represents the Lagrange interpolation polynomial for n+1 data points[17]. x_i and y_i represent the x-coordinate and y-coordinate, respectively, of the known data points. $l_j\,(x)$ denotes the Lagrange basis function, where the index i ranges from 0 to n, representing the i-th basis function. Consequently, $L\,(x)$ signifies the estimated value at a given point x, acquired through the employment of the Lagrange interpolation method.

2) Normalization: There are several methods available for data normalization, including min-max scaling, Z-score standardization, and mean normalization. For the purposes of this study, the dataset is normalized using the Z-score standardization method. This approach is chosen due to its simplicity and ease of computation, as well as its ability to effectively normalize data regardless of the scale or presence of extremely large or small values. The Z-score standardization formula utilized is as follows:

$$z = \frac{x - \mu}{\sigma} \tag{3}$$

where x represents the data mean, μ represents the standard deviation, and z represents the standardized score.

3) Sliding window: The sliding window is widely utilized for time series analysis and sequence data processing. It serves as a valuable tool for feature extraction and performing computations on data subsets[18]. With the application of the sliding window, it becomes possible to capture short-term patterns and dependencies presented in the three-phase voltage and current. Moreover, by analyzing sequences within smaller windows, the impact of noise in analog data can be minimized, subsequently improving the signal-to-noise ratio. This aspect proves advantageous for tasks such as anomaly detection and experimental evaluation. Moreover, the mean, median, and variance of the data within the window are extracted as input features.

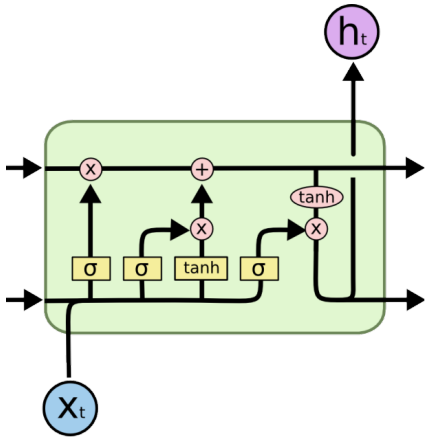

Fig. 2. The structure of the "cell" in the LSTM model[23].

B. The Anomaly Detection Model using Deep Learning

1) Basis of SAE and LSTM: The Stacked Autoencoder (SAE) is a hierarchical neural network comprised of multiple encoders connected layer by layer. It is an unsupervised learning method that follows a layer-wise greedy approach[19]. The Autoencoder (AE), which is a constituent of SAE, employs the backpropagation algorithm to ensure the output values match the input values. Initially, it compresses the input into a latent space representation and then reconstructs the output based on this representation. SAE possesses several advantages, including powerful expressive capability, a straightforward training process, and the ability to construct multiple layers of stacked architecture. It effectively mitigates challenges like "vanishing gradients" and "exploding gradients" that arise with increased depth in autoencoders. Consequently, SAE finds extensive application in target recognition, anomaly detection, anomaly diagnosis, and other domains.

The training process of SAE involves training one layer at a time[20]. Initially, a network with a single hidden layer is trained. After completing the training of this layer, training of a network with two hidden layers is initiated, and so on. Once all the layers have been trained, the encoder weights of each layer are combined to form a complete deep neural network. Subsequently, fine-tuning is performed, where the entire network is fine-tuned using supervised learning methods to optimize network performance. This training approach is known as the greedy layer-wise training algorithm.

Due to its deep structure and layer-wise training strategy, stacked autoencoders can learn higher-level and more abstract feature representations, thereby achieving superior performance across various tasks. In contrast, single-layer autoencoders are limited by their shallow structure and can only capture relatively simple features.

Long Short-Term Memory (LSTM) is a special type of recurrent neural network (RNN) structure[21]. The neurons in an LSTM model consist of four main components: the memory cell, the input gate, the forget gate, and the output gate. The internal structure of an LSTM neuron is illustrated in Fig 2.

$$i_t = \sigma(W_{xi}x_t + W_{hi}h_{t-1} + b_i) \tag{4}$$

$$f_t = \sigma(W_{xf}x_t + W_{hf}h_{t-1} + b_f) \tag{5}$$

$$\tilde{C}t = \tanh(Wxcx_t + W_{hc}h_{t-1} + b_c) \tag{6}$$

$$C_t = f_t \odot C_{t-1} + i_t \odot \tilde{C}_t \tag{7}$$

$$o_t = \sigma(W_{xo}x_t + W_{ho}h_{t-1} + b_o) \tag{8}$$

$$h_t = o_t \odot \tanh(C_t) \tag{9}$$

The input gate decides whether to include the input feature x_t of the current time step into the state update of the LSTM. W_{xi} and W_{hi} are weight matrices for the linear transformation of the input and the hidden state h_{t-1} of the previous time step, respectively. b_i is the bias term and σ is the sigmoid activation function.The forget gate decides whether to keep the previous state information. W_{xf} and W_{hf} are weight matrices, and b_f is a bias term. The candidate memory unit computes the candidate value at the current time step by applying the hyperbolic tangent activation function. Cell states are updated through multiplication and addition operations. The forget gate determines

the proportion of the previous cell state C_{t-1} retained, and the input gate determines the contribution of the candidate memory unit $\tilde{C}_t$ to the cell state. The output gate determines the output of the LSTM cell. W_{xo} and W_{ho} are weight matrices, and b_o is a bias term. The hidden state is the output of the LSTM cell, computed by element-wise multiplication of the cell state C_t with the output of the output gate and application of the hyperbolic tangent activation function.

2) The hybrid model of SAE-LSTM: The architecture of the proposed anomaly detection model, which jointly using SAE-LSTM, is illustrated in Fig. 1. The main steps of the model are described below, with a focus on the SAE-LSTM component.

a) The SAE component is employed to extract the deep latent features from the preprocessed three-phase voltage and current data. To ensure compatibility with the LSTM model, the latent features are further converted into an appropriate format through data type conversion.

b) The SAE network is utilized to extract the hidden features from the data, which are then transformed into time series data suitable for input to the LSTM model. By configuring the hyperparameters of the LSTM model, the model is trained and fine-tuned. Additionally, Bayesian optimization techniques are applied to optimize the model.

c) The optimized SAE-LSTM anomaly detection model is then evaluated using the test set. Various performance metrics, including accuracy and F1 score, are employed to assess the model's predictive capabilities.

3) Fine-tuning and optimization of network parameters: In this study, a composite loss function is employed, incorporating elastic net regression to mitigate overfitting. This technique applies both L1 and L2 regularization to penalize the coefficients in the regression model. By integrating L1 regularization's sparsity and L2 regularization's weight shrinkage, elastic net regression achieves a harmonious balance, leading to improved generalization performance. The combined loss function is calculated using the following formula:

$$LOSS_{com} = \alpha LOSS_{SAE} + (1-\alpha)\, LOSS_{LSTM} + LOSS_1 + LOSS_2 \tag{10}$$

$$LOSS_1 = \lambda\beta\,(W_{en} + W_{de} + W_{ih} + W_{hh} + W_{fc}) \tag{11}$$

$$LOSS_2 = \lambda\,(1-\beta)\,\left({W_{en}}^2 + {W_{de}}^2 + {W_{ih}}^2 + {W_{hh}}^2 + {W_{fc}}^2\right) \tag{12}$$

The weight between $LOSS_{SAE}$ and $LOSS_{LSTM}$ is controlled by α. When α is set to 1, only the reconstruction loss is active, and the classification loss has no impact. This means that the model focuses primarily on reconstructing the input data and minimizing reconstruction errors[22]. When α is set to 0, only the classification loss is active, and the reconstruction loss has no impact. This means that the model primarily focuses on the classification task and strives to optimize classification accuracy. When takes an intermediate value, both the reconstruction loss and the classification loss are considered, and the model optimizes between balancing the reconstruction and classification tasks. By adjusting the value of α, the optimal trade-off between reconstruction and classification tasks can be found to meet specific problem requirements and performance demands. λ determines the importance of L1 and L2 regularization, while β controls the weight between L1 and L2 regularization. $LOSS_1$ calculates the L1 regularization term, which penalizes the sum of the absolute values of the model parameters. It measures the sparsity of the parameters by computing the L1 norm of different weight matrices. W_{en} and W_{en} represent the weight matrices of the encoder and decoder in the SAE network, respectively. W_{ih} includes the weights connecting the LSTM layer input to its hidden state, responsible for transforming the input features into hidden state representations within the LSTM unit. W_{hh} is the weight matrix associated with the connections between hidden-to-hidden in the LSTM network, encompassing the weights that link the previous hidden state to the current hidden state within the LSTM layer. W_{fc} is responsible for propagating hidden state information, enabling the LSTM to maintain dependencies across the input time series. It represents the weight matrix of the fully connected layer in the LSTM network. The L1 norm of each weight matrix is the sum of the absolute values of its elements. These norms are summed together, multiplied by λ and β , to weight the L1 regularization term. The purpose of this is to encourage the model to produce sparse parameters, reducing redundancy and improving the model's generalization ability.

$LOSS_2$ calculates the L2 regularization term, which penalizes the sum of squared model parameters. It measures the smoothness of the parameters by computing the squared L2 norm of different weight matrices. The squared L2 norm of each weight matrix is the sum of the squares of its elements. These sums are added together, multiplied by λ and $1-\beta$, to weight the L2 regularization term. The purpose of this is to encourage the model to produce smooth parameters, reducing overfitting and improving the model's generalization ability.

By using λ and β multiplied by the regularization terms, an appropriate balance can be found between model complexity, reconstruction task, and classification task. This helps optimize the overall loss function to achieve better model performance and generalization ability.

To obtain the optimal performance of the model, a grid search method is used to search for the optimal model parameters. Grid search (GridSearchCV) is a search technique used to find the optimal parameters of a model. Grid search is a brute force algorithm. This makes a complete search for a given subset of the hyperparameter space[23]. Due to the exhaustive search, grid search consumes significant training time and resources, making its search performance inefficient.

Using Bayesian optimization technique can reduce computational costs and improve efficiency. It is more effective than grid search. It consists of two main components: a Bayesian statistical model for modeling the objective function, and an acquisition function for deciding where to sample next[24]. By using Bayesian optimization technique to find the optimal learning rate and weight in the combined loss function, the accuracy of the model in voltage anomaly detection task can be maximized.

TABLE I. DISTRIBUTION OF NORMAL DATA AND ABNORMAL DATA

Data type	Substation A	Substation B	Substation C	Substation D
Normal data	2881	375	157	2881
Abnormal data	0	732	1284	0

III. EXPERIMENTS AND RESULTS

A. Data acquisition

1) Real Dataset: The real dataset was obtained from the national power grid and consists of secondary load data collected from four 330 kV substations situated in distinct geographical regions. Each substation's data includes measurements of active power, reactive power, three-phase voltage and current, and the total power factor. Specifically, Substations A, C, and D were monitored at 15-minute intervals, while Substation B was recorded at 60-minute intervals. Notably, Substations A and D contain normal data, whereas Substations B and C contain abnormal data. Table I presents the distribution of normal and abnormal data in the dataset.

From Table I, it can be observed that substations B and C have a majority of abnormal data, while substations A and D do not contain any abnormal data. The ratio of normal data to abnormal data in the overall dataset is approximately 4:1. When comparing

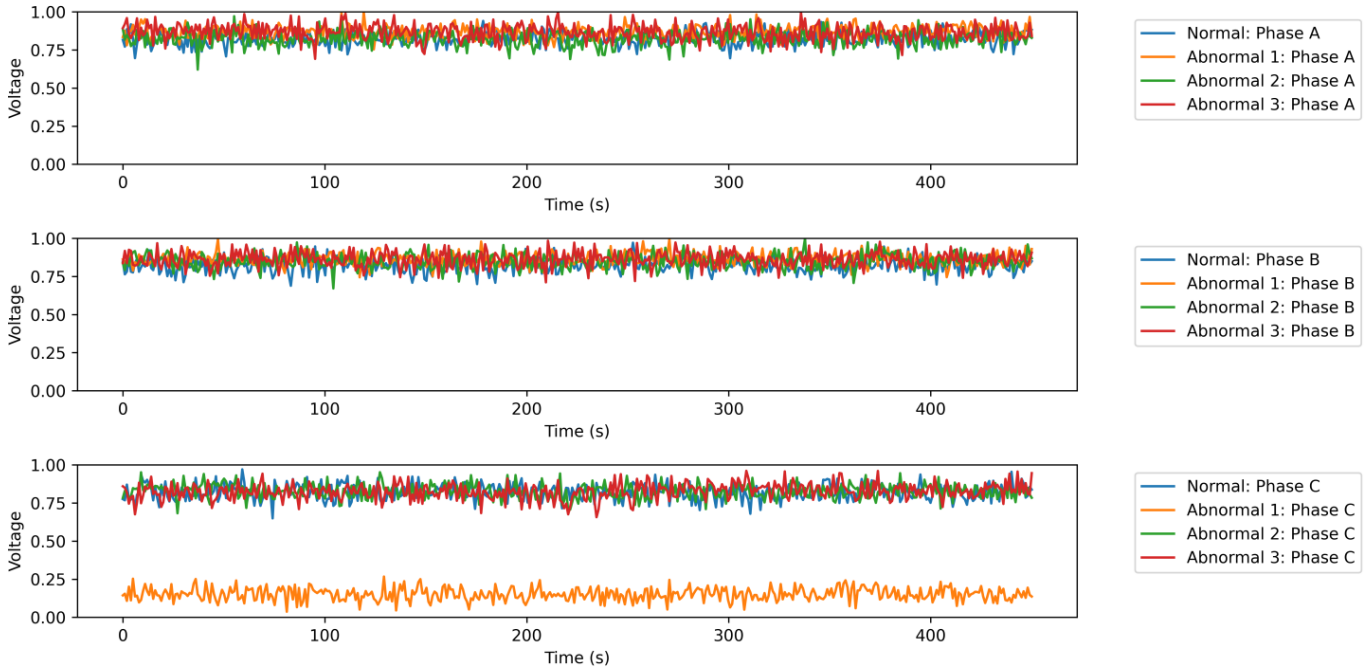

Fig. 3. The SNR is 20dB analog data diagram.

the features of normal and abnormal data, it is found that voltage changes are most pronounced during abnormal occurrences. Although there may be changes in current corresponding to the phase where voltage abnormalities occur, they are not as significant as voltage changes. Therefore, the abnormal data types in this study include voltage overvoltage, undervoltage, and other voltage-related anomalies. Three-phase current is considered as an auxiliary feature to help detect voltage anomalies. Hence, the features extracted for data processing and analysis in this study include three-phase voltage and three-phase current.

2) Computer simulation: The simplicity of the original data does not showcase the advantages of the constructed SAE-LSTM network. Therefore, to facilitate a comparison between the SAE-LSTM network and other networks, a method of generating simulated data by adding noise can be utilized. Signal-to-noise ratio(SNR) generically means the dimensionless ratio of the signal power to the noise power contained in a recording[25]. The parameter settings involve a SNR ranging from -20 dB to 20 dB, with an increment of 4 dB. Consequently, simulated data is generated at 4 dB intervals within the -20 dB to 20 dB range. The simulated data comprises normal and abnormal data, encompassing three types of abnormalities: voltage overvoltage, voltage loss, and low voltage. The figure below depicts the generated simulated data, illustrating SNR of -20 dB and 20 dB.

The Fig. 3 and Fig. 4 illustrate the simulation data with the SNR of 20dB and -20dB respectively, where blue indicates normal data and other colors represent abnormal data. The red color corresponds to phase C undervoltage, green indicates phase B overvoltage, and orange signifies phase C undervoltage. Analysis of the figure reveals that at a SNR of -20 dB, the abnormal data closely overlaps with the original data, indicating a limited ability to distinguish between the signal and noise. Consequently, the detection task becomes highly challenging for the model. However, when the SNR is 20 dB, the distribution of the generated simulated data closely resembles that of the collected system data. Furthermore, the discrimination between the signal and noise significantly improves compared to the -20 dB SNR. This distinction is particularly evident in the case of phase C voltage loss anomalies, where a substantial difference exists between anomaly type 3 and other data types in terms of phase C voltage. Such differentiation is absent when the SNR is -20 dB. Therefore, the detection task becomes relatively simpler when the SNR is 20 dB.

The selection of a SNR ranging from -20 to 20 dB serves specific purposes. Extremely low SNR result in noise intensity surpassing signal intensity, hindering accurate anomaly detection by the model, with consistent accuracy below 50%. Conversely, very high SNR yield simulated data that closely resembles the original data, containing minimal noise. Consequently, the distinction between the signal and noise diminishes. Given the relatively simple and 6-dimensional nature of the original data, different models achieve accuracies exceeding 99%, making it impossible to discern performance differences among them. By setting the SNR between -20 and 20 dB, a range of conditions spanning from high-noise environments to strong-signal environments is encompassed. Conducting experiments within this range enables a more comprehensive analysis of the SAE-LSTM model's performance.

B. Evaluation Index

For model evaluation, more advanced classification metrics from the confusion matrix such as accuracy and F1 score are utilized. As the problem involves multi-label classification with imbalanced data, weighted F1 score is used, assigning different weights to different classes. The formulas for calculating accuracy and weighted F1 score are as follows:

$$Accuary = \frac{TP + TN}{TP + TN + FP + FN} \tag{13}$$

$$Precision = \frac{TP}{TP + FP} \tag{14}$$

$$Recall = \frac{TP}{TP + FN} \tag{15}$$

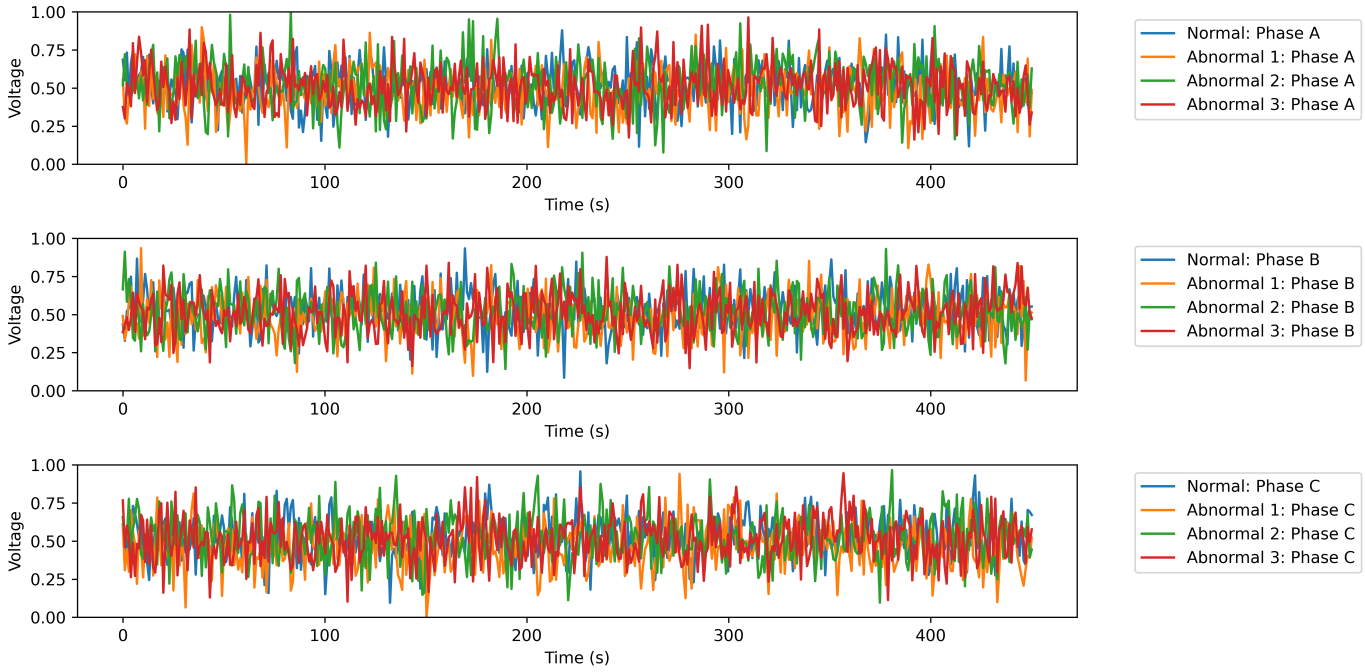

Fig. 4. The SNR is -20dB analog data diagram.

$$F1 = \frac{2PrecisionRecall}{Precision + Recall} \tag{16}$$

where TP represents True Positive. TN represents True Negative. FP represents False Positive. FN represents False Negative.. Recall represents the recall rate, Precision denotes precision.

C. Network Structure Selection

1) Unbalanced data processing: After conducting measurements, it was observed that the dataset exhibits a ratio of approximately 4:1 between normal data and abnormal data[26]. Due to the limited availability of original abnormal data, the SMOTEBoost method is employed to generate synthetic data points in close proximity to the existing abnormal data.

SMOTEBoost is an ensemble method that combines the advantages of SMOTE and Boosting techniques. By integrating multiple weak learners into a robust classifier, it surpasses the performance of the standalone SMOTE method, offering improved accuracy and robustness when dealing with imbalanced datasets.

The fundamental concept of the SMOTEBoost algorithm revolves around augmenting the weight of minority class samples during each iteration of the classification learning process. This emphasis on the minority class allows the weak learners to focus more on these samples. To address the severe class imbalance in the original data, new artificial samples are generated and incorporated into the dataset[27]. The SMOTE algorithm and the update of weak learner weights are described by the following formulas:

$$x_{new} = x + rand\,(0,1) * (\tilde{x} - x) \tag{17}$$

$$\alpha = 0.5 \ln\left(\frac{1-\varepsilon}{\varepsilon}\right) \tag{18}$$

$$\omega_n = \omega_o e^{\alpha} \tag{19}$$

Equation 17 represents the performance calculation of the weak learner, while Equation 19 represents the weight calculation of the weak learner. x represents a selected minority class sample. By applying the principle of nearest neighbors, one sample, denoted as $\tilde{x}$, is randomly chosen from the k nearest neighbor samples of x, where represents a random number ranging from 0 to 1. The newly synthesized sample is denoted as x_{new}. ε denotes the error rate, which represents the cumulative weight of misclassified samples. α represents the weight of the weak learner, while ω_o signifies the weight of each individual sample. Finally, ω_n refers to the updated weights.

2) Ablation experiments with different layers: The neural network comprises an input layer, hidden layers, and an output layer, with the number of hidden layers and hidden units per layer playing a pivotal role in determining the neural network's capacity and complexity. The selection of these hyperparameters significantly influences the model's ability to learn intricate patterns and generalize to unseen data.

In the case of the SAE network, the number of neuron nodes in the input and output layers depends on the dimensionality of the input data. In this study, the extracted and preprocessed dataset exhibits a feature dimensionality of 6. Thus, the SAE network consists of 6 neuron units in both the input and output layers. The purpose of the hidden layers in the neural network is to grasp the complex features inherent in the input data. Augmenting the number of hidden units empowers the network with greater representational capacity, enabling it to capture more intricate and nuanced characteristics of the input data. Nonetheless, an excessive number of hidden layers can lead to prolonged training time, overfitting, and vanishing gradients. In this study, preliminary experiments revealed that exceeding 3 hidden layers in the SAE network resulted in overfitting. Consequently, two optimal combinations of hidden layer units, specifically 64 and 16, were chosen.

Concerning the LSTM network, the extracted hidden features derived from the SAE serve as input features, while the output layer comprises 4 units representing the data labels. Regarding the selection of the number of hidden layers and hidden units, it has been observed that surpassing 3 hidden layers exponentially escalates

TABLE II. MODEL PARAMETER CONFIGURATION UNDER THE SNR FROM -20DB TO 20DB

SNR(dB)	SAE(16) LSTM(128)		SAE(16) LSTM(128,128)		SAE(16) LSTM(128,128,128)		SAE(16,64) LSTM(128)		SAE(16,64) LSTM(128,128)		SAE(16,64) LSTM(128,128,128)	
	Accuracy	F1	Accuracy	F1	Accuracy	F1	Accuracy	F1	Accuracy	F1	Accuracy	F1
-20	0.488	0.543	0.493	0.547	0.529	0.571	0.497	0.541	0.501	0.549	**0.531**	**0.559**
-16	0.519	0.554	0.527	0.563	0.572	0.558	0.529	0.561	0.543	**0.573**	**0.574**	0.568
-12	0.571	0.559	0.588	0.569	0.620	0.566	0.576	0.567	0.581	0.555	**0.622**	**0.573**
-8	0.592	0.564	0.613	0.567	0.636	0.569	0.581	0.567	0.602	0.567	**0.637**	**0.579**
-4	0.632	0.567	0.639	0.572	0.654	0.575	0.636	0.573	0.636	0.573	**0.658**	**0.576**
0	0.663	0.570	0.676	0.576	0.674	0.574	0.673	0.574	0.667	0.576	**0.679**	**0.579**
4	0.680	0.613	0.682	0.618	0.681	0.619	0.679	0.618	0.680	0.621	**0.683**	**0.639**
8	0.740	0.649	0.748	0.657	0.750	0.652	0.740	0.652	0.743	0.662	**0.751**	**0.663**
12	0.743	0.673	0.770	0.677	0.772	0.687	0.746	0.676	0.753	0.687	**0.775**	**0.690**
16	0.779	0.693	0.776	0.694	0.780	0.694	0.776	0.690	0.777	0.694	**0.781**	**0.699**
20	0.774	0.694	0.781	0.697	0.786	0.702	0.779	0.681	0.780	0.682	**0.788**	**0.711**

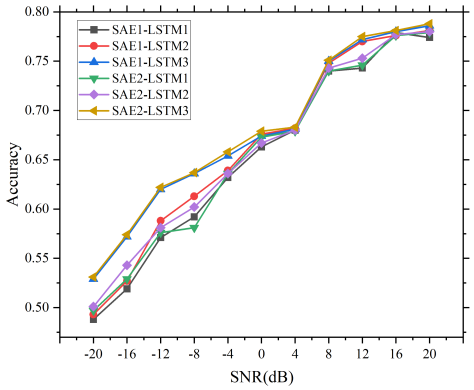

Fig. 5. The accuracy diagram of the proposed model with varying neurons and hidden layers under the SNR from -20dB to 20dB.

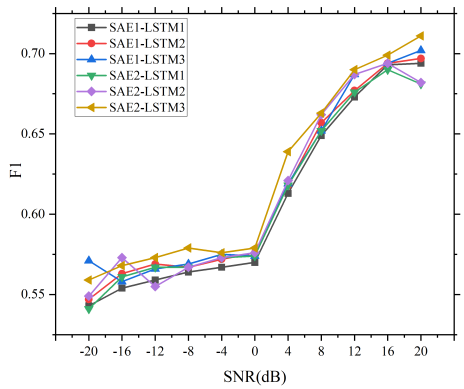

Fig. 6. The F1 diagram of the proposed model with varying neurons and hidden layers under the SNR from -20dB to 20dB.

the network's computational complexity and elevates the risk of overfitting. Consequently, this study concentrates on investigating the range of hidden layers from one to three. To enhance time efficiency, the training iterations were initially set to 100. Given that computers store and process data in binary format, an initial value of 16 was selected for the number of neuron nodes. Subsequently, the program was executed with a progressive increment of neuron nodes by powers of 2 to obtain accuracy values on the test set. The number of hidden units for each of the 1 to 3 hidden layers can be configured from 16 to 256, resulting in numerous possible combinations. Through experimentation, two combinations displaying optimal performance were identified, with the hidden layer units set as 128, 128, and 128, and 128, 128, 128, respectively.

In this study, the optimal number of model layers is initially examined, and the corresponding results are listed in Table II. The comparisons of the results for different models are illustrated in Fig. 5 and Fig. 6.

From the results, it can be observed that the SAE2-LSTM2 and SAE2-LSTM3 models have similar accuracy and F1 scores, but the SAE2-LSTM3 model slightly outperforms the SAE2-LSTM2 model in terms of accuracy and F1 score. This suggests that increasing the number of LSTM hidden layers can improve accuracy. Comparing the SAE1-LSTM2 and SAE2-LSTM2 models, it can be noted that increasing the number of SAE layers does not significantly affect accuracy, indicating that increasing the number of SAE layers has a limited impact on improving model performance. Comparing the four models, it is evident that the SAE2-LSTM3 model achieves the highest accuracy and F1 score, indicating the best model performance. Therefore, the SAE2-LSTM3 anomaly detection model with 2 SAE hidden layers (16, 64 units) and 3 LSTM hidden layers (128, 128, 128 units) exhibits better performance.

D. Comparisons in the Simulation Dataset

To comprehensively evaluate the performance of the SAE-LSTM model, the outcomes achieved through the proposed approach are contrasted with those of SAE, LSTM, as well as other fundamental machine learning models, including Convolutional Neural Network (CNN) and Support Vector Machine (SVM)[28]. The signal-to-noise ratio ranges from -20 to 20 dB, with a step size of 4, enabling an extensive assessment of these models on multi-classification data. score of each model are visualized in the following graph:

From Fig. 7, Fig. 8 and Table IV, it can be concluded that it is evident that the proposed method presented in this study achieves superior evaluation metrics in the task of three-phase voltage anomaly detection, surpassing the other four methods. The LSTM model exhibits comparatively lower recognition performance, with an average accuracy in multi-classification detection that is approximately 10% lower than the other four methods, and a correspondingly lower weighted-F1 score. This discrepancy can be attributed to the introduction of noise in the original data, which affects the temporal nature of the data and consequently hampers the LSTM model's recognition capabilities. The CNN model, primarily designed to capture local spatial correlations, demonstrates inferior performance compared to the SAE and LSTM models. Furthermore, in the case of non-image data, CNN may not fully comprehend the intricate relationship between input features and output predictions[29]. In contrast, both SAE and SVM exhibit superior recognition performance relative

TABLE III. THE RESULTS OF ACCURACY AND F1 FOR THE COMPARISON MODELS UNDER THE SNR FROM -20DB TO 20DB

SNR(dB)	SAE		LSTM		CNN		SVM		SAE-LSTM	
	Accuracy	F1	Accuracy	F1	Accuracy	F1	Accuracy	F1	Accuracy	F1
-20	0.515	0.568	0.512	0.552	0.529	0.549	**0.553**	0.548	0.531	**0.571**
-16	0.547	0.558	0.533	0.561	0.544	**0.572**	0.557	0.549	**0.574**	0.568
-12	0.581	**0.566**	0.529	0.553	0.576	0.555	0.561	0.553	**0.622**	0.556
-8	0.592	0.568	0.533	0.547	0.590	0.564	0.587	0.559	**0.637**	**0.578**
-4	0.635	0.576	0.573	0.564	0.638	0.573	0.593	0.571	**0.658**	**0.577**
0	0.662	0.568	0.618	0.504	0.657	0.570	0.624	0.570	**0.679**	**0.572**
4	0.668	0.613	0.624	0.512	0.686	0.605	0.664	0.603	**0.683**	**0.629**
8	0.722	0.651	0.641	0.593	0.718	0.645	0.722	0.644	**0.756**	**0.663**
12	0.766	0.687	0.650	0.669	0.727	0.687	0.732	0.689	**0.775**	**0.690**
16	0.771	0.694	0.676	0.682	0.740	0.693	0.751	0.694	**0.781**	**0.699**
20	0.779	0.694	0.689	0.692	0.776	0.694	0.777	0.694	**0.788**	**0.710**

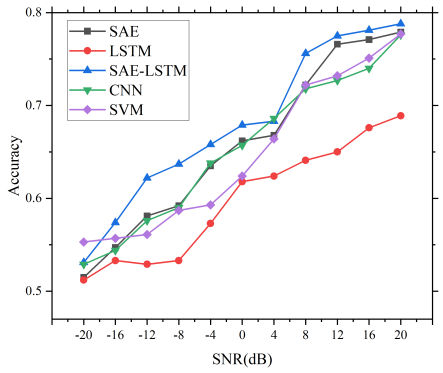

Fig. 7. The accuracy diagram of the comparison models under the SNR from -20dB to 20dB.

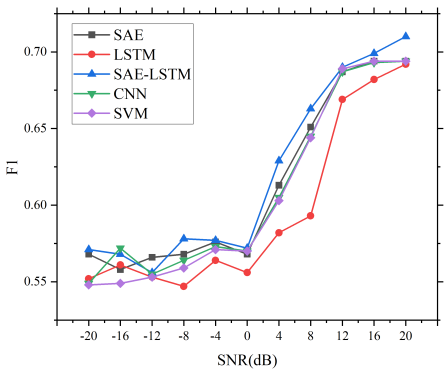

Fig. 8. The F1 diagram of the comparison models under the SNR from -20dB to 20dB.

to CNN and LSTM methodologies. SAE and SVM possess robust feature extraction capabilities and excel in classification tasks.

In addition, the SAE-LSTM hybrid model proposed in this paper has the best detection accuracy with F1 values in all signal-to-noise environments. This can be attributed to the SAE component within the SAE-LSTM model, which compensates for LSTM's limitations in feature extraction, allowing the LSTM component to leverage its strengths in handling sequential problems.

E. Comparisons in the Real Dataset

In order to evaluate the model more effectively and reduce experimental bias, k-fold cross-validation is employed. The KCV consists in splitting a dataset into k subsets; then, iteratively, some of them are used to learn the model, while the others are exploited to assess its performance[30].

From the collected data of the substation monitoring system, three-phase voltage and current data along with the label column were extracted. The original data was subjected to anomaly detection, and the experimental results are shown in the Table III.

The analysis of the results indicates that all models exhibit favorable performance when applied to the initial three-phase voltage and current data. The reason for this is that the real data used in this paper is very simple. Not only there are few types of anomalies, but also there is almost no interference, which is very easy to identify. Therefore, traditional machine learning algorithms such as SVM, CNN, etc. can also achieve very good results.

Moreover, although SAE and LSTM do not work perfectly when used alone. However, the SAE-LSTM network proposed in this paper combines the advantages of the two algorithms, which is both very powerful in feature extraction and well adapted to handle such time-series data as power grids. The simulation experiments in the previous paper show that the detection effect is still very good when facing the low signal-to-noise ratio data with great interference.

IV. CONCLUSION

This paper presents a method that addresses the challenge of delayed anomaly detection in current gateway metering devices by combining Stacked Autoencoders (SAE) and Long Short-Term Memory Neural Networks (LSTM) with elastic network regularization. The advantages of the proposed model are verified by real data experiments and simulated data experiments.

In the experiments utilizing real data, all examined models exhibited satisfactory performance, largely attributed to the dataset's simplicity. To further enrich our investigation, we expanded our research scope to include experiments using simulated data.

These simulated data experiments introduced noise to increase data complexity before comparing the effectiveness of various models.

TABLE IV. THE RESULTS OF ACCURACY AND F1 FOR THE COMPARISON MODELS IN REAL DATASETS

Model	SVM	CNN	SAE	LSTM	SAE-LSTM
Accuracy	1	1	0.999	0.997	1
F1	1	1	0.999	0.997	1

The findings underscored that the model proposed within this paper demonstrates superior performance in managing complex data.

As delineated in Fig. 8, the F1 score exhibits a gradual ascent when the Signal-to-Noise Ratio (SNR) spans -20dB to 0dB and 12dB to 20dB. In contrast, a brisk rise is observed between 0dB and 12dB. This observation can be attributed to the reduction in noise level as SNR increases, thus facilitating the model's anomaly detection capabilities and consequently leading to the swift enhancement in F1 scores. The sluggish elevation in the F1 score with an increase in SNR might be indicative of the model nearing its maximum performance potential, unable to capitalize fully on the added clarity from higher SNR values. Alternatively, it could imply that the data lacks further valuable information to aid in more distinct anomaly detection, thereby causing the measured F1 score to ascend more slowly.

Power grid data often exhibits specific patterns of variation, making it suitable for analysis using the temporal nature of LSTM networks and the robust feature extraction capabilities of SAE networks. Experimental results have confirmed the high effectiveness of this method for anomaly detection. But fully applying this model to the anomaly detection of the actual substation gateway metering device will have certain shortcomings. In practical applications, there is no corresponding label for the data measured by the metering device. In order to realize the abnormality detection of the metering device more conveniently, it is necessary to increase the learning of unsupervised algorithms; It can also display some parameters measured by the metering device in real time.

ACKNOWLEDGMENT

This work was supported by the Technology Research Project of National Grid of China (5700-202155204A-0-0-00).

REFERENCES

[1] Himeur, Y., Ghanem K., Alsalemi A., et al. Artificial intelligence based anomaly detection of energy consumption in buildings: A review, current trends and new perspectives[J].*Applied Energy*, 287. DOI:10.1016/j.apenergy.2021.116601.

[2] Jayachandran, M., Reddy, C.R., Padmanaban, S. et al. Operational planning steps in smart electric power delivery system. *Sci Rep 11*, 17250 (2021). https://doi.org/10.1038/s41598-021-96769-8

[3] Himeur Y , Alsalemi A , Bensaali F ,et al.Smart power consumption abnormality detection in buildings using micromoments and improved K-nearest neighbors[J].*International Journal of Intelligent Systems*, 2021.DOI:10.1002/int.22404.

[4] Li Dan, Chiu Wei-Yu, Sun Hongjian, Poor H Vincent. Multiobjective optimization for demand side management program in smart grid. *IEEE Trans Ind Inf* 2018;14(4):1482–90. http://dx.doi.org/10.1109/TII.2017.2776104.

[5] Fadlullah Z M, Fouda M M, Kato N, et al. An early warning system against malicious activities for smart grid communications[J]. *IEEE Network*, 2011, 25(5): 50-55.

[6] Gans W, Alberini A, Longo A. Smart meter devices and the effect of feedback on residential electricity consumption: Evidence from a natural experiment in Northern Ireland[J]. *Energy Economics*, 2013, 36: 729-743.

[7] Jan B, Farman H, Javed H, et al. Energy efficient hierarchical clustering approaches in wireless sensor networks: A survey[J]. *Wireless Communications and Mobile Computing*, 2017.

[8] Zendehboudi A, Baseer M A, Saidur R. Application of support vector machine models for forecasting solar and wind energy resources: A review[J]. *Journal of cleaner production*, 2018, 199: 272-285.

[9] WANG Wei, ZHU Ming, ZENG Xuewen, et al. Malwaretraffic classification using convolutional neural network forrepresentation learning[C]. *2017 International Conference onInformation Networking, Da Nang, Vietnam*, 2017: 712–717.doi: 10.1109/ICOIN.2017.7899588.

[10] Himeur Y, Alsalemi A, Bensaali F, Amira A. A novel approach for detecting anomalous energy consumption based on micro-moments and deep neural networks. *Cogn Comput* 2020;12(6):1381–401.

[11] Wang, F., Zhou, Y., Yan, H. et al. Enhancing the generalization ability of deep learning model for radio signal modulation recognition. *Appl Intell* (2023). https://doi.org/10.1007/s10489-022-04374-7

[12] Shang W L, Zhang S S, Wan M, et al. Modbus/TCP communication anomaly detection algorithm based on PSOSVM[J]. *Acta Electronica Sinica*, 2014, 42(11): 2314-2320.

[13] Lee S, Nengroo S H, Jin H, et al.Anomaly detection of smart metering system for power management with battery storage system/electric vehicle[J].*ETRI Journal*, 2022.

[14] Wang Xinlin, Yang Insoon, Ahn Sung-Hoon. Sample efficient home power anomaly detection in real time using semi-supervised learning. *IEEE Access 2019*;7:139712–25. http://dx.doi.org/10.1109/ACCESS.2019.2943667

[15] Hussain Saddam, Mustafa Mohd Wazir, Jumani Touqeer Ahmed, Baloch Shadi Khan, Saeed Muhammad Salman. A novel unsupervised featurebased approach for electricity theft detection using robust PCA and outlier removal clustering algorithm. *Int Trans Electr Energy Syst 2020*;30(11):e12572.

[16] Ghori K, Imran M, Nawaz A, Abbasi R, Ullah A, Szathmary L. Performance analysis of machine learning classifiers for non-technical loss detection. *J Ambient Intell Humaniz Comput 2020*;1–16. http://dx.doi.org/10.1007/s12652-019- 01649-9.

[17] Kudo T, Morita T, Matsuda T, Takine T. Pca-based robust anomaly detection using periodic traffic behavior. In: *2013 IEEE international conference on communications workshops (ICC)*. 2013, p. 1330–4. http://dx.doi.org/10.1109/ ICCW.2013.6649443.

[18] C,I,Podilchuk,et al.Three-dimensional subband coding of video.[J].*IEEE transactions on image processing : a publication of the IEEE Signal Processing Society*, 1995.

[19] Weng Y, Zhang N, Xia C. Multi-agent-based unsupervised detection of energy consumption anomalies on smart campus. *IEEE Access 2019*;7:2169–78.

[20] Vafaeipour M, Rahbari O, Rosen M A, et al. Application of sliding window technique for prediction of wind velocity time series[J]. *International Journal of Energy and Environmental Engineering*, 2014, 5: 1-7.

[21] Shi Z, Li P, Sun Y. An outlier generation approach for one-class random forests: An example in one-class classification of remote sensing imagery. In: *2016 IEEE international geoscience and remote sensing symposium (IGARSS)*. 2016, p. 5107–10.

[22] Ghanbari M, Kinsner W, Ferens K. Anomaly detection in a smart grid using wavelet transform, variance fractal dimension and an artificial neural network. In: *2016 IEEE electrical power and energy conference (EPEC)*. 2016, p. 1–6

[23] Hochreiter S , Schmidhuber J .Long Short-Term Memory[J].*Neural Computation*, 1997, 9(8):1735-1780.DOI:10.1162/neco.1997.9.8.1735.

[24] Xu X, Liu H, Yao M. Recent progress of anomaly detection.*Complexity 2019*; 2019:1–11

[25] Liashchynskyi P, Liashchynskyi P. Grid search, random search, genetic algorithm: a big comparison for NAS[J]. *arXiv preprint* arXiv:1912.06059, 2019.

[26] Frazier P I. A tutorial on Bayesian optimization[J]. *arXiv preprint* arXiv:1807.02811, 2018.

[27] Anguita D, Ghelardoni L, Ghio A, et al. The'K'in K-fold Cross Validation[C]//*ESANN*. 2012: 441-446.

[28] Linda O, Wijayasekara D, Manic M, Rieger C. Computational intelligence based anomaly detection for building energy management systems. In: *2012 5th international symposium on resilient control systems*. 2012, p. 77–82.

[29] Cao N, Lin C, Zhu Q, Lin Y, Teng X, Wen X. Voila: Visual anomaly detection and monitoring with streaming spatiotemporal data. *IEEE Trans Vis Comput Graphics* 2018;24(1):23–33.

[30] Johnson D H. Signal-to-noise ratio[J]. *Scholarpedia*, 2006, 1(12): 2088.