EAI Endorsed Transactions
on Energy Web Research Article EAI.EU

Risk prediction method for power Internet of Things operation based on ensemble learning

Chao Hong[1,2,*], Xiaoyun Kuang[1,2], Yiwei Yang[1,2], Yixin Jiang[1,2], Yunan Zhang[1,2]

[1]CSG Electric Power Research Institute Co., Ltd, Guangzhou 510663, China
[2]Guangdong Provincial Key Laboratory of Power System Network Security, Guangzhou 510663, China

Abstract

INTRODUCTION: The power Internet of Things is an important strategic support for the State Grid Corporation of China to build an international leading energy internet enterprise. However, the operating environment of the power Internet of Things is complex and varied, which has serious implications for the safe operation of the power Internet of Things.
OBJECTIVES: To timely predict the various risk.
METHODS: A data set is fused based on time series. The training set is over-sampled using an adaptive synthetic oversampling method. Then, by jointly considering the contribution of features to classification and the correlation between features, a risk prediction method ground on ensemble learning is established.
RESULTS: From the results, the accuracy of predicting 5 risk categories increased by 7.00%, 1.10%, 2.20%, 2.30%, and 0.60%, respectively, reducing the features from the original 118 columns to 60 columns and reducing the data dimension by 49.00%. Compared with traditional models, the accuracy was 98.61%, and the overall accuracy was improved by 0.60%.
CONCLUSION: This risk prediction scheme can quickly and accurately predict the risk categories that affect its operation. It has high prediction accuracy and fast speed than other algorithms. This research can provide strong assistance for security decision-making in the power Internet of Things.

Keywords: adaptive synthetic oversampling, ensemble learning, power internet of things, prediction, risk.

Received on 13 05 2024, accepted on 18 02 2025, published on 26 02 2025

Copyright © 2025 Ch. Hong *et al.*, licensed to EAI. This is an open access article distributed under the terms of the CC BY-NC-SA 4.0, which permits copying, redistributing, remixing, transformation, and building upon the material in any medium so long as the original work is properly cited.

doi: 10.4108/ew.6045

1. Introduction

As an important strategic support for State Grid Corporation of China, the Internet of Things (IoT) in electricity essentially combines some socially relevant elements (such as non-electric objects, human activities, natural environment, government policies, etc.) on the basis of information physical integration in the power system, forming a power information and physical social integration system [1-3]. The integration of information, physics, and society has led to a complex and diverse operating environment for the power IoT, with a surge in the types and quantities of terminal devices connected, making them susceptible to external risk interference at all stages [4]. In addition, the coupling characteristics of energy flow, data flow, and business flow in the power IoT are gradually strengthening, which also makes the interaction coupling characteristics of information, physics, and society increasingly complex, thereby affecting the overall safe and reliable operation of the power IoT [5]. In addition, the risks faced by the power IoT are characterized by diversification and expansion of scope. Equipment failures, malicious attacks, human errors, and other risks can affect the stable operation of the power IoT, resulting in a series of cross space chain failures. In severe cases, it can even lead to catastrophic power outages. In current research methods, risk prediction methods based on probability distribution are too complex. The accuracy and efficiency of risk prediction based on machine learning methods for mining data still need to be improved. Therefore, based on the random matrix theory,

[*]Corresponding author. Email: hongchao_2024@163.com

power data is fused. The minority class samples in the training set are over-sampled using the Adaptive Synthetic Oversampling (ADASYN) method to achieve balanced data processing. After that, the redundant features are dropped by jointly considering the contribution value of features to classification and the correlation between features. The ReliefF-S algorithm is applied to select the optimal features for the processed power IoT risk balance sample set. Finally, a risk prediction model for the power IoT operation based on BO-CatBoost is established.

If a risk occurs in a certain part of the power grid and is not detected and cut off timely, the entire power grid may be affected and suffer losses. Therefore, conducting risk prediction is crucial for maintaining the safety of the power grid. Ajayi et al. developed a deep learning model for power infrastructure risk management by using text mining methods to retrieve meaningful terms from accident data to reduce power infrastructure accidents. The overall accuracy was 0.93, with an average absolute error between 0.91 and 0.94, which could minimize project costs and provide effective strategies to reduce risks [6]. He et al. developed a risk warning system for the distribution network operation ground on the IoT. Firstly, the weight of risk indexes was analyzed to determine the detection indicators. Then, a risk warning evaluation model was established to determine the detection requirements in the distribution network. Tests showed that the real-time detection error was below 5% [7]. Qu et al. built a novel method for predicting power risk areas based on correlated Markov chains. By characterizing the load and constraints of non-uniform power coupled networks, a power risk area prediction method ground on correlated Markov chains was proposed. Finally, the adaptive position adjustment strategy and cross optimal solution strategy were used to improve the cross adaptive Grey Wolf optimization algorithm. Simulation results verified its effectiveness and superiority [8]. Li et al. introduced a distributed flow processing mechanism to address the issue that traditional security risk monitoring techniques were not suitable for network physical power systems. A log analysis architecture for power log anomaly detection was proposed. An integrated prediction method ground on time series and asymmetric error cost evaluation criteria were used to predict abnormal features. This method could effectively detect abnormal data [9]. To improve the reliability and continuity of the distribution system, Kong et al. constructed a power IoT analysis and monitoring system for data collection and fault analysis. Based on the measurement information demonstrated by the distributed phasor measurement unit, the phase difference between the positive and negative sequence currents determines the fault section and fault type. Ground on the load symmetry of the distribution network, the fault section was determined. The results showed that this method could achieve high-precision fault localization [10].

The operational status of the power IoT is constantly changing in real-time. Equipment failures, human errors, extreme weather, and network attacks can all be triggers for system risks. The existing methods for predicting time series data in the power IoT are unable to handle inter sequence related information. Therefore, Li et al. proposed a decision fusion architecture. The ensemble learning method was used to make judgments on distributed time series data, and integrated multiple sources of data for real-time prediction. It provided better accuracy while reducing communication burden [11]. Piotrowski et al. proposed an integrated integrator artificial neural network based on hybrid methods to optimize the electricity and promote energy storage. A variety of machine learning solutions were mixed for prediction. Research showed that the proposed integrated method generated the smallest error, which was also suitable for short-term electricity production prediction of other renewable energy sources [12]. Kim proposed an integrated model for optimizing energy management in smart homes ground on ensemble learning and Particle Swarm Optimization (PSO). Five different baseline models were combined to establish a hybrid ensemble model. The PSO was used to optimize the hyper-parameters of each combination mode. Different random samples were trained. The optimized ensemble learning model improved the prediction accuracy by 95.6% [13]. Wang et al. proposed an integrated deep learning framework to classify automatically for power quality interference. The signal was classified using a long short-term memory network. The training results of multiple long short-term memory networks were integrated using Bagging theory to improve the generalization ability. It had better classification performance and computational speed [14]. To improve the performance of modeling and classification problems, Larrea et al. integrated extreme learning machines into time series modeling problems to improve modeling and classification performance. The optimal topology results for each time series problem were statistically analyzed. The PSO was applied to adjust the parameter weights. This strategy had more accurately results [15].

In summary, the risk prediction method based on probability distribution is too complicated and cannot meet the requirement of real-time prediction. Although the use of machine learning to mine data for risk prediction is the mainstream method at present, the accuracy and efficiency of prediction need to be improved. In addition, most current studies focus on the information side and the physical side for analysis, and it is easy to ignore the impact of social risks on the operation of power iot, resulting in incomplete risk investigation. And the risk is easy to spread and spread, easy to cause an impact that cannot be ignored. Therefore, the study develops a risk prediction method for the operation of the power IoT based on ensemble learning, aiming to quickly and accurately classify the operational risks of the power IoT by mining data information.

The article is divided into four parts. The first section reviews the research status of operational risk prediction methods for power iot. In the second section, the risk prediction model of power iot operation based on ensemble learning is constructed. In the third section, the performance of the designed method is verified. The fourth section is discussion. The fifth section is the conclusion.

2. Construction of risk prediction model for power IoT operation based on ensemble learning

A risk prediction method for the operation of the power IoT is designed based on ensemble learning algorithms, taking into account factors related to power information, physics, and society. This method includes: risk data fusion and balancing, risk prediction optimal feature subset selection, and research on risk prediction methods based on ensemble learning.

2.1. Data fusion and balancing processing based on ADASYN algorithm

The amount of data generated during the operation of the power IoT is very large, with rich attributes, which may have many redundant, irrelevant, and even interfere with the risk prediction of the power IoT operation. This leads to the inability to predict risks well using the original dataset directly, reducing the accuracy of risk prediction models [16]. By combining multiple basic learners together, ensemble learning can reduce the overfitting problem of a single learner, thereby improving the stability and accuracy of the overall prediction. This approach performs well on traditional machine learning tasks such as classification, regression, and clustering [17]. The basic idea is to obtain multiple different weak learners through training data, and then use a certain combination strategy to ultimately form a strong learner.

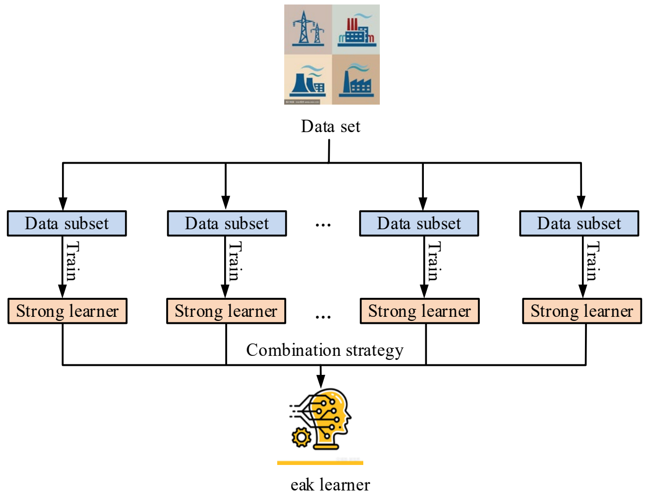


Figure 1. Ensemble learning algorithm framework

Figure 1 displays the framework. Compared with the single model, the ensemble learning algorithm mining data information is more sufficient, and the prediction results are more accurate and reliable.

The study uses random matrix theory to comprehensively analyze the operational risks of the power IoT from three perspectives: power information D_e, physical D_p, and social D_s. The measurement data collected from any feature in the power information, physical, and social aspects during any period of time can form a column vector. The collected data is represented as $\{a_1, a_2, \ldots, a_n\}$, $\{b_1, b_2, \ldots, b_n\}$, and $\{c_1, c_2, \ldots, c_n\}$. Among them, a, b and c represent data information in power information, physical, and social aspects. The measurement data from the information, physical, and social sides are extracted to form the original data set $Dataset$, as shown in formula (1).

$$Dataset = \begin{cases} D_e = \{a_1, a_2, \ldots, a_n\} \\ D_p = \{b_1, b_2, \ldots, b_n\} \\ D_s = \{c_1, c_2, \ldots, c_n\} \end{cases} \tag{1}$$

After constructing a complete data set of electricity information, physical and social aspects, data from the same time but different spaces are fused based on time series. hen data fusion occurs, the time series in one of the data files is used as the benchmark, which is called the benchmark file. The parameters of other data stream files must be unified to this time benchmark. According to the time series, the constructed datasets of information, physics, and society are integrated together to construct a high-dimensional random matrix D, as shown in formula (2).

$$D = \begin{bmatrix} a_{11} & a_{21} & \cdots & a_{n1} & b_{11} & b_{21} & \cdots & b_{n1} & c_{11} & c_{21} & \cdots & c_{n1} \\ a_{12} & a_{22} & \cdots & a_{n2} & b_{12} & b_{22} & \cdots & b_{n2} & c_{12} & c_{22} & \cdots & c_{n2} \\ \cdots & \cdots & \cdots & \cdots & \cdots & \cdots & \cdots & \cdots & \cdots & \cdots & \cdots & \cdots \\ a_{1N} & a_{2N} & \cdots & a_{nN} & b_{1N} & b_{2N} & \cdots & b_{nN} & c_{1N} & c_{2N} & \cdots & c_{nN} \end{bmatrix} \tag{2}$$

After completing the data fusion, the ADASYN method was used to handle the data imbalance in the training set, sampling to generate a specified number of risk class pseudo samples, achieving data balance processing, and overcoming the low training accuracy and unstable performance due to the small samples in certain categories. ADASYN s main idea is to oversample different categories of samples to different degrees according to the distribution density of different samples. The lower the density of the categories, the more composite samples are generated, which balances the number of samples between the different categories while maintaining the diversity of the data [18]. The raw data concludes training and testing sets. The imbalance degree in the training set is calculated. The minority class sample is signified as q_s, and the majority class is signified as q_l. The imbalance degree d is shown in formula (3).

$$d = \frac{q_s}{q_l}, d \in (0,1) \tag{3}$$

Next, the number of synthesized samples G is shown in formula (4).

$$G = (q_l - q_s) * e, e \in [0,1] \tag{4}$$

In formula (4), when $e = 1$, G is the difference between the minority class and the majority class. The majority class sample and the minority class sample after synthesizing the data exactly reach equilibrium. For each minority class sample x_i, its K nearest neighbors are determined. The distribution of the majority class samples around each minority class sample is represented as r_i, as shown in formula (5).

$$r_i = \frac{\Delta i / K}{Z} \tag{5}$$

In formula (5), Δi signifies the majority class sample number among K nearest neighbor sample points. Z is the normative factor. r_i forms a distribution. If there are more majority class samples around minority class sample x_i, the generated distribution r_i will have a higher value. The sample size g_i that need to be assembled for each minority class sample is shown in formula (6).

$$g_i = r_i \times G \tag{6}$$

The synthesized sample s_i is shown in formula (7).

$$s_i = x_i + \left(x_{zi} - x_i\right)\sigma, \sigma \in [0,1] \tag{7}$$

In formula (7), $\left(x_{zi} - x_i\right)$ represents the difference vectors of different dimensional spaces. σ represents a random number. The above steps are repeated to synthesize minority class samples until the required samples to be synthesized according to formula (6) is met. The ADASYN algorithm uses the density distribution of minority class samples to automatically determine the sample size to be assembled for each minority class sample, which can effectively obtain sufficient pseudo data highly similar to the original data. At the data end, it can effectively solve the data imbalance in the training accuracy of machine learning algorithms. The overall process of integrating and balancing risk data is shown in Figure 2.

The study quantitatively analyzes the imbalance in the training set using the ADASYN method. If the data is imbalanced, oversampling is performed on the minority class samples in the data to increase their quantity and ultimately obtain a balanced training set. A balanced training sample set used for subsequent ensemble learning algorithm training can overcome the low training accuracy and unstable performance caused by a low minority class sample. Finally, the testing set verifies the performance of subsequent ensemble learning algorithms.

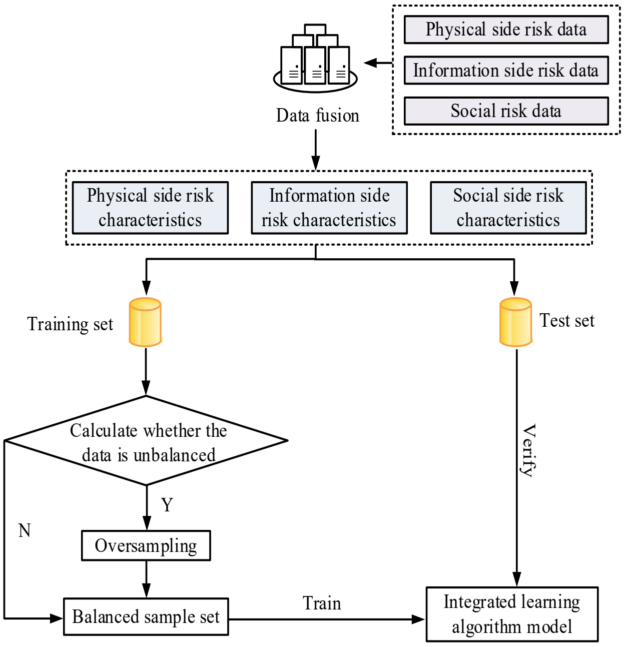

Physical side risk data
Information side risk data
Social risk data
Data fusion
Physical side risk characteristics
Information side risk characteristics
Social side risk characteristics
Training set
Test set
Calculate whether the data is unbalanced
Y
N
Oversampling
Verify
Balanced sample set
Train
Integrated learning algorithm model

Figure 2. Fusion and balanced processing flow of risk data

2.2. Construction of risk prediction model for power IoT operation based on BO-CatBoost

It is crucial to enhance the timeliness and accuracy when studying the operational risks of the power IoT. Otherwise, it will affect the efficiency of investigation and even further spread the risk. However, there are also many redundant and irrelevant features in the operational data. For example, there may be some correlation between the electrical physical quantities on the physical side that can be derived from each other. There may be some meteorological features on the social side that are unrelated to the operation of the power IoT [19]. Excessive redundancy can have a certain impact on data mining, and even increase the training cost and model complexity of machine learning algorithms.

ReliefF algorithm is an effective method for feature selection. Relieff considers the differences between similar and heterogeneous training samples and assigns a score to each feature to reflect the contribution degree of the feature to the classification [20]. ReliefF algorithm is suitable for dealing with multi-class problems, but it only evaluates the contribution value of each feature to the classification, as long as the features that play a positive role in the classification are likely to be retained, while ignoring the correlation between each pair of features, which may cause mutual redundancy among features. Therefore, based on ReliefF algorithm, Spearman correlation coefficient is introduced to analyze the correlation between features to solve this problem. Relief-s algorithm jointly considers the correlation between features and categories and the correlation between features to achieve the de-redundancy operation of risk features, so as to minimize redundancy and finally get the optimal set of risk features. The flow of Relief-S algorithm is shown in Figure 3.

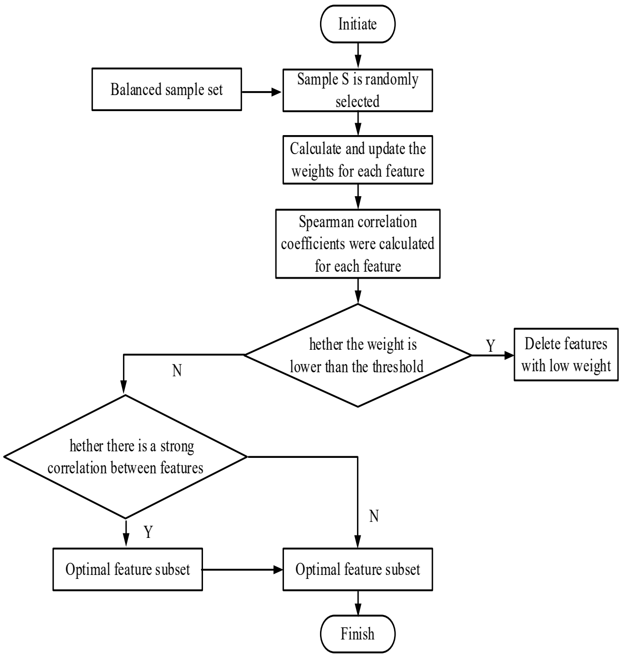

Initiate
Balanced sample set
Sample S is randomly selected
Calculate and update the weights for each feature
Spearman correlation coefficients were calculated for each feature
hether the weight is lower than the threshold
Y
Delete features with low weight
N
hether there is a strong correlation between features
Y
N
Optimal feature subset
Optimal feature subset
Finish

Figure 3. Flowchart of ReliefF-S algorithm

A sample X is randomly selected in the sample set. Firstly, k nearest neighbor samples $L_j = (j = 1, 2, \ldots, k)$ of the same category as X are searched, and the distance $X(Y)$ between the samples X_i and L_j under the feature Y is calculated, as presented in formula (8).

$$X(Y) = \sum_{j=1}^{k} diff(Y, X_i, L_j) \tag{8}$$

Then, k nearest neighbor samples $M_j = (j = 1, 2, \ldots, k)$ of different categories from X are searched. The distance $D(Y)$ between samples X_i and M_j under the feature Y is calculated, as shown in formula (9).

$$D(Y) = \sum_{j=1}^{k} diff(Y, X_i, M_j) \tag{9}$$

Given the feature Y, the distance between two samples X_1 and X_2 is shown in formula (10).

$$diff(Y, X_1, X_2) = \begin{cases} \dfrac{|X_1(Y) - X_2(Y)|}{\max(Y) - \min(Y)}, \text{ Feature Y is continuous} \\ 0, \text{ Feature Y is discrete and } X_1(Y) = X_2(Y) \\ 1, \text{ Feature Y is discrete and } X_1(Y) \neq X_2(Y) \end{cases} \tag{10}$$

The weight of feature Y is constantly updated. Under the feature Y, if the distance between sample X and samples of the same category is less than the distance between the sample and samples of different categories, it indicates that the feature has strong classification ability. It should be given a larger weight. According to this idea, the weights are updated by iterating n times. The average weight of each feature is used as the final weight. The calculation for updating the weight $W(Y)$ is shown in formula (11).

$$W(Y) = W(Y) - \frac{X(Y)}{nk} + \frac{\sum_{t \in class(X_i)} \left[\frac{P(t)}{1 - P(class(X_i))} \square D(Y) \right]}{nk} \tag{11}$$

In formula (11), $class(X_i)$ represents the category of sample X_i. $P(t)$ signifies the ratio of this category. $P(class(X_i))$ is the ratio of randomly selected sample X_i category. After normalization, the range for each feature weight is [0,1]. The correlation coefficient λ_{Y_i,Y_j} between any two features is calculated to determine the correlation between features, as shown in formula (12).

$$\lambda_{Y_i,Y_j} = \frac{\sum_{f=1}^{k} (Y_{if} - \bar{Y}_i)(Y_{jf} - \bar{Y}_j)}{\sqrt{\sum_{f=1}^{k} (Y_{if} - \bar{Y}_i)^2 (Y_{jf} - \bar{Y}_j)^2}} \tag{12}$$

In formula (12), Y_i and Y_j represent any feature. Y_{if} and Y_{jf} are the observed values of these two features. $\bar{Y}_i$ and $\bar{Y}_j$ represent the average of k observations. According to the set weight threshold, features with low weights are deleted. Then, among the retained features, the features with low contribution to classification in the strongly correlated features are removed [21]. The Categorical Boosting (CatBoost) can improve classification performance by merging multiple learners. However, the performance will be affected by key parameters. Manual parameter tuning requires a certain amount of work and is blind, which can easily lead to the loss of optimal parameter solutions and affect the accuracy of risk prediction models. The Bayesian Optimization (BO) algorithm requires less initial sample points and has high optimization efficiency when searching for the optimal parameters. Compared with grid search, random search, genetic algorithm and other parameter finding methods, it more suitable for model parameter optimization. CatBoost is used as the base classifier to construct a risk prediction model for the operation of the power IoT. The BO is introduced to optimize the parameters of the CatBoost, to obtain the optimal parameter combination of the model and achieve high-precision prediction of operational risks in the power IoT. Assuming that the i-th dimensional feature of the k-th sample is a discrete feature, the numerical feature x_k^i is shown in formula (13).

$$x_k^i = \frac{\sum_{x_j \in D_k} \{x_k^i = x_j^i\} \times y_j + a \times p}{\sum_{x_j \in D_k} \{x_k^i = x_j^i\} + a} \tag{13}$$

In formula (13), D_k represents the portion of the sample set before the k-th sample in the sorting. $\{x_k^i = x_j^i\} = 1$. If samples x_k and x_j belong to different categories in the i-th dimensional feature, then $\{x_k^i = x_j^i\} = 0$. p is the added prior value. a is its weight coefficient. They can reduce the noise problem of low-frequency categories.

On the basis of data imbalance processing and feature selection, a risk prediction model is constructed. The BO-CatBoost model is used as the final risk prediction model for the operation of the power IoT. This model is a multi-classification model. The final output represents a type of risk event or non-risk event. The specific process is displayed in Figure 4. Ground on the random matrix theory, data from power information, physics, and society are fused. The ADASYN is applied to over-sample minority class samples to solve the impact of data imbalance on algorithm accuracy. Then, the ReliefF-S algorithm is used to jointly consider the correlation between features and categories, as well as the correlation between features, to achieve redundant operations on risk features, reduce data dimensions, and improve model training speed. Finally, a CatBoost ensemble learning model is constructed based on a symmetric tree classifier. Combined with the BO, the optimal parameters are searched to better improve its performance.

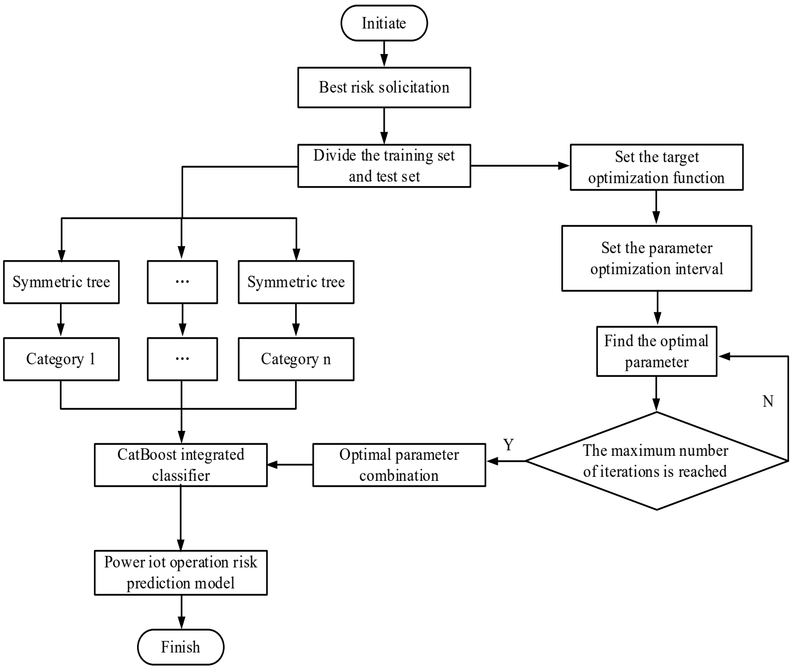

Initiate
Best risk solicitation
Divide the training set and test set
Set the target optimization function
Set the parameter optimization interval
Symmetric tree
···
Symmetric tree
Find the optimal parameter
Category 1
···
Category n
N
Y
The maximum number of iterations is reached
CatBoost integrated classifier
Optimal parameter combination
Power iot operation risk prediction model
Finish

Figure 4. Power IoT operation risk prediction model based on BO-CatBoost

3. Simulation experiment and result analysis of risk prediction for power IoT operation

A topology model is constructed through simulation experiments and a partial data set is generated. The data generated by the operation of the power IoT in 8 different states is simulated and collected. Performance indicators such as Accuracy, Precision, Recall, and F1-Score are applied to analyze the performance of the proposed risk prediction model.

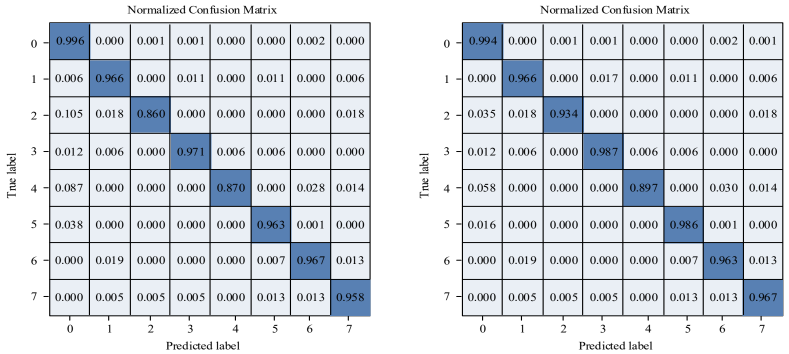
3.1. Data balance analysis

The fused training set is over-sampled and entered in the CatBoost for learning. The model is validated using the testing set. Power information, physical and social risk data are obtained through joint simulation using RT-LAB and OPNET. The compilation tool Python 3.7 platform is used for implementation. The study collects over 20000 pieces of data within 210s at intervals of 0.01s. Different nodes are selected for simulation. The single-phase short circuit risk is set to 15.0-16.5s. Within 45.0s to 45.5s, a two-stage short-circuit hazard is set. Within 75.0s to 76.5s, there is a danger of two-phase short circuit. The danger time for three-phase short circuit is 120-120s. The risk of incorrect instruction injection is between 150s and 15s. 195.0-195.5s, the risk of human error is simulated by manually pulling the gate in violation of regulations. Within 200.0-200.5s, by adjusting the line parameters under specific weather conditions, weather risks in the social aspect are simulated. A complete data set of operational risks in the power IoT is obtained, as shown in Table 1. This data set contains a total of 118 features, including physical parameters such as three-phase current and three-phase voltage of 16 physical nodes, as well as attack instructions, switching states, weather, and other factors.

To analyze the improvement effect of oversampling on risk prediction models, the CatBoost is used to train the datasets before and after data balancing. The risk prediction performance is verified using a testing set. The confusion matrix of the risk prediction results before and after data balancing processing is shown in Figure 5. In the figure, the accuracy of most categories that were balanced was improved. In Figure 5 (a), the prediction accuracy trained on raw data for categories 2, 4, and 7 was 86.00%, 87.00%, and 95.80%, respectively. From Figure 5 (b), the prediction accuracy of risk categories 2, 4, and 7 after balanced processing was increased to 93.40%, 89.70%, and 96.70%, respectively. The misjudgment rate was significantly decreased. The above results indicate that data imbalance processing exerts a crucial role in improving the prediction accuracy of small category samples and reducing the false alarm rate of risk prediction.

Table 1. Complete data set of operational risks of power IoT

Time (s)	Physical side				Information side		Social side		
	Node1-V1 (V)	...	Node14-V2 (V)	...	Attack signal	On-off state	Air pressure (hPa)	Ice thickness (mm)	...
0	-1.06	...	0.01	...	0.00	0.00	976.20	0.00	...
0.01	-0.30	...	-1965.32	...	0.00	0.00	976.00	0.00	...
...	...	...	...	...	...	...	...	...	...
150.31	-5039.08	...	-0.19	...	1.00	0.00	970.50	0.00	...
150.32	8154.50	...	-0.19	...	1.00	0.00	970.40	0.00	...
...	...	...	...	...	...	...	...	...	...



(a) Original data+CatBoost

(b) ADASYN+CatBoost

Figure 5. Confusion matrix of risk prediction before and after data balancing treatment

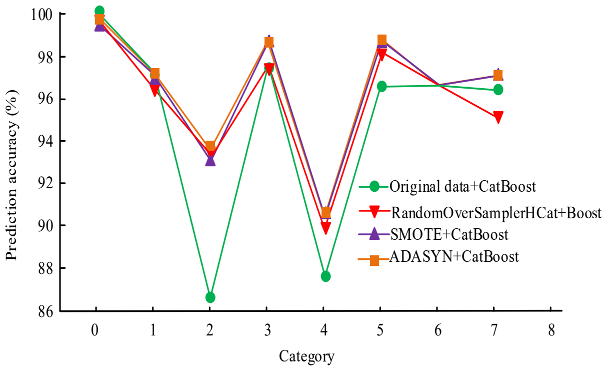


Figure 6. Comparison of oversampling algorithms

Common oversampling methods include SMOTE, ADASYN, and Random Over Sampler algorithm. The CatBoost is used as a standard, dividing the training and testing sets in a 7:3 ratio. The data in the training set is over-sampled. Four oversampling algorithms, including raw data+CatBoost, SMOTE+CatBoost, Random Over Sampler+CatBoost, and ADASYN+CatBoost are analyzed and compared. The data processed by different oversampling algorithms are input into the CatBoost for training. The accuracy comparison curves for different risk classifications are obtained, as shown in Figure 6. From the graph, compared with the original data+CatBoost, the accuracy of the other two oversampling methods +CatBoost method was improved by about 1.50% in most categories. The ADASYN was improved by about 2%, which was more superior. This means that the data must be over-sampled before establishing risk forecasts.

The variation of prediction accuracy of various methods with test cases under different oversampling algorithms is shown in Table 2. From the table, compared with the other two methods, the random sampling method had the worst performance, while the performance of SMOTE and ADASYN was roughly equivalent. The accuracy of most classifications was improved. The ADASYN performed slightly better, with accuracy improvements of 2.3% for Class 2 (bipolar short circuit) and Class 5 (pseudo instruction attack), respectively. However, the prediction accuracy of Class 0 (normal operating data) was slightly decreased. This means that oversampling algorithms are more sensitive to improving the accuracy of small class samples, and the prediction accuracy may slightly decrease for most class samples. Overall, this is significant for improving the model stability.

The balanced data may also contain some redundant and irrelevant features. Therefore, to improve the prediction accuracy, feature selection is performed on it. The study uses a balanced sample set to test the ReliefF and ReliefF-S, with F1-Score as the objective function. Based on the proportion of each feature in the sample, a contribution threshold is set, and each indicator is screened. CatBoost is used to evaluate each indicator. From Table 3, the performance comparison of ReliefF and ReliefF-S was different at various thresholds. When there are fewer features, although this method shortened learning time, the effect was not ideal. Compared with the ReliefF, the ReliefF-S was more suitable for processing high-dimensional data. Compared to ReliefF, the method takes an average of 211 seconds, respectively. From this point, the ReliefF-S is superior to the ReliefF, which can effectively reduce the data dimensionality.

Table 2. Prediction accuracy of various categories under various oversampling algorithms

Risk category	Original data + CatBoost (%)	Random Over Sampler + CatBoost (%)	SMOTE + CatBoost (%)	ADASYN + CatBoost (%)
0	99.6	99.3	99.0	99.4
1	96.6	99.0	96.6	96.6
2	86.0	92.7	92.5	93.0
3	97.1	97.1	98.2	98.2
4	87.0	89.7	89.2	89.2
5	96.1	97.7	98.3	98.4
6	96.1	96.1	96.1	96.1
7	95.9	94.7	96.4	96.5

Table 3. Performance comparison of ReliefF and ReliefF-S under different contribution thresholds

Contribution threshold	ReliefF + CatBoost			ReliefF-S + CatBoos		
	Characteristic number	F1-Score (%)	Training time (s)	Characteristic number	F1-Score (%)	Training time (s)
1	21	88.64	103	14	87.63	81
0.60	39	93.05	165	29	92.65	130
0.55	41	94.68	184	37	94.63	156
0.52	52	98.46	200	48	95.35	184
0.50	67	95.45	266	60	95.67	219
0.48	79	95.23	294	60	95.36	264

0.40	94	95.36	326	83	95.48	318
0.10	107	94.86	357	96	95.01	337

3.2. Evaluation of risk prediction methods

To further prove the performance of the CatBoost, the BO is combined to find the optimal parameters. There is a 7:3 split between training and test sets. The former is applied to train the model, and the accuracy is used as the objective function. The maximum iterations are 30. Figure 7 displays the visualization effect of BO processing. From the graph, when the iteration reached the 7th time, the prediction accuracy was 99.77%, the maximum number of trees was 815, the L2 regularization coefficient was 1.3, and the learning rate was 0.27. At this point, the parameter combination is the optimal parameter combination.

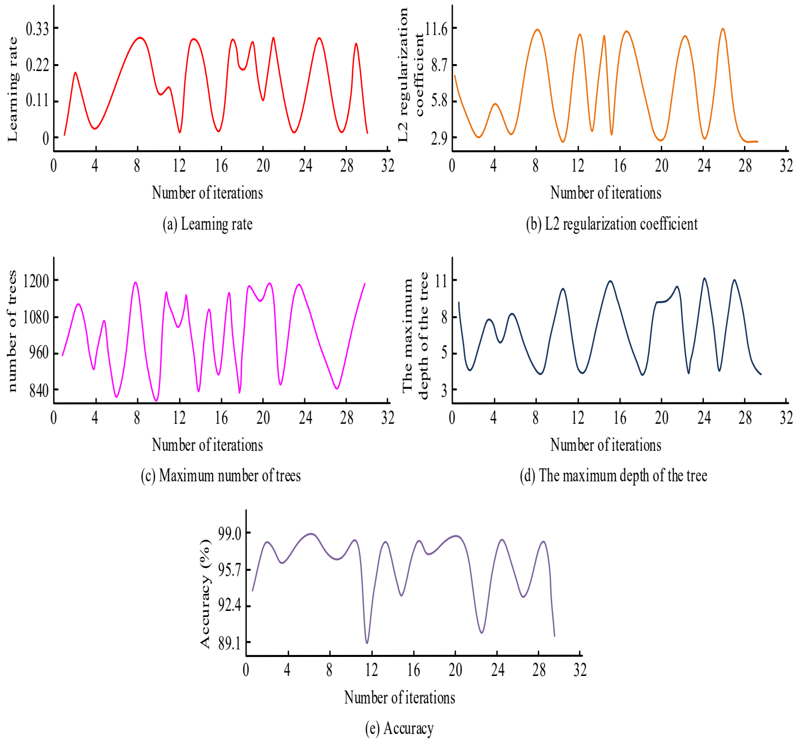


Figure 7. Parameter finding process of Bayesian optimization algorithm

To further evaluate the effectiveness of the algorithm, a comparison is made between Multi-layer Perceptron (MLP), K-nearest Neighbor Classification (KNN), Gradient Boosting Decision Tree (GBDT), and some mainstream ensemble learning algorithms: XGBoost, LightBoost, CatBoost, and BO-CatBoost. Figure 8 displays the results. From the graph, the accuracy of the BO-CatBoost was 98.61%, the recall was 98.97%, and the F1-Score was 98.82%. MLP is the weakest,which is not suitable for solving such prediction problems. Compared with GBDT, the average F1-Score of BO-CatBoost risk prediction was 15.12% higher. Compared with KNN, which was 6.73% higher. Meanwhile, ensemble learning methods such as XGBoost and LightBoost also achieved good results. Compared with other machine learning methods, the BO-CatBoost algorithm has higher accuracy in risk prediction, which can more accurately predict various risks that may occur during the operation of the power IoT.

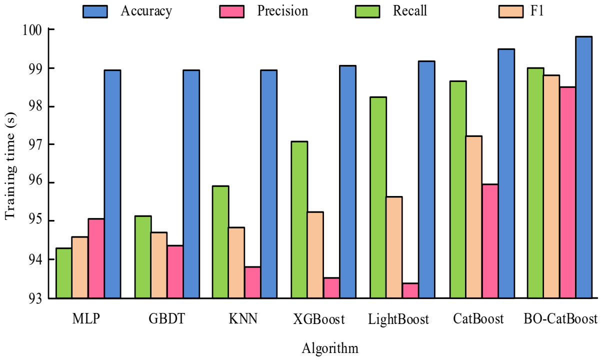


Figure 8. Comparison of performance indicators of each algorithm

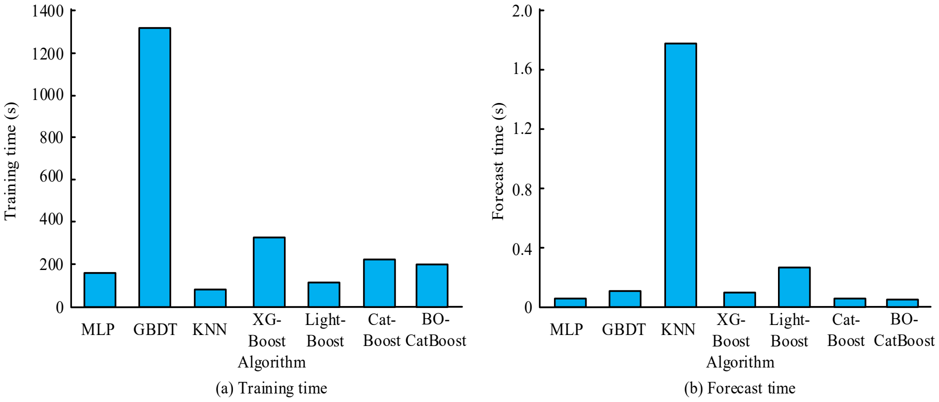
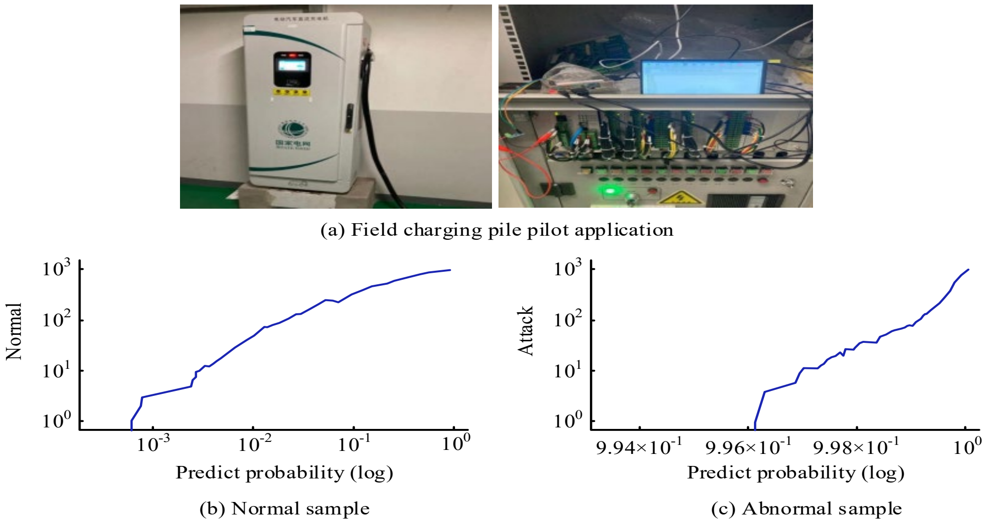


Figure 9. Comparison of time performance of each algorithm

Figure 9 displays the time performance of various algorithms. Figure 9 (a) shows the training time, and Figure 9 (b) displays a comparison of the predicted time. From the graph, the training time of KNN was 136s, which was the fastest among all models, but its prediction time (1.78s) was the largest. Although the LightBoost method had better training speed, compared with other ensemble learning methods, the prediction time reached 0.38s, which was relatively insufficient to predict danger in real-time. The MLP method had a faster prediction speed, but its performance was poor. The training and prediction time of the CatBoost was 203s and 0.21s, respectively. The BO-CatBoost was 187s and 0.18s, respectively, which had good prediction ability and real-time performance.

In order to evaluate the effect of the risk prediction model in the actual system, the pilot application is carried out on the charging pile of the State Grid Corporation. Figure 10 (a) shows the pilot application on the charging pile in the actual scenario. In actual scenarios, the firmware of running devices cannot be changed. Therefore, malicious software configuration changes and hardware cable connections are used to simulate attacks. The risk prediction algorithm based on BO-CatBoost is run on the local embedded device or board computer, and the low-sample rate and low-power INA219 current sampling module is used as the power acquisition device with the sampling frequency of 200/s. Figure 10 (b) and Figure 10 (c) show the prediction probability curves of normal samples and abnormal samples of charging pile terminal under the actual scenario. When the prediction probability is 0, it means normal samples; when the prediction probability is 1, it means abnormal samples. It can be seen that the prediction probability of 80% normal samples is less than 0.3, while that of 100% abnormal samples is greater than 0.9. Therefore, relatively high accuracy can be achieved in practical applications.



(a) Field charging pile pilot application

(b) Normal sample

(c) Abnormal sample

Figure 10. Pilot application of the risk prediction algorithm based on BO-CatBoost in the field charging pile

4. Discussion

The operating state of the power Internet of Things changes in real time, and equipment failure, human error, extreme weather, and cyber attacks can all be incentives for system risks. Tommarello et al. designed regression algorithms and predicted risks by analyzing a large number of historical measurement data. Based on the consistency analysis results of the measurement data, they judged the deviation between the predicted value and the actual value. Although the accuracy rate of the model was significantly improved, the mathematical principle was complicated and the calculation time was long. It cannot meet the requirements of real-time prediction well [22]. From the perspective of data mining, this paper presents a set of operational risk prediction methods for power Internet of Things, uses Bayesian optimization algorithm to improve CatBoost algorithm, and builds an integrated operational risk prediction model for power Internet of Things. The experimental results show that the prediction accuracy of risk categories 2, 3, 4, 5 and 7 increased by 7%, 1.1%, 2.2%, 2.3% and 0.6% respectively, and the false positive rate also decreased significantly, indicating that data balancing processing can effectively solve the problem of excessive false positive rate of risk prediction, and improve the stability of risk prediction model as a whole.

Power iot systems need to be able to process large amounts of data in a short period of time to ensure timely detection and early warning of risks. Sharma et al. designed regression algorithm and predicted risk by analyzing a large number of historical measurement data, and judged the deviation between predicted value and actual value based on the consistency analysis results of the measurement data [23]. Guo et al. carried out correlation analysis and fault risk level prediction of distribution network data through improved machine learning algorithm, which played a certain role in reducing the probability of distribution network faults and risks [24]. The above two methods are analyzed from the information side, ignoring the impact of real-time risk prediction on the operation of power Internet of Things. The study uses CatBoost model training to learn the selected key features. The training time of BO-CatBoost model changes from 440s to 217s, and the prediction time changes from 0.24s to 0.04s. Compared with the CatBoost model, the overall accuracy rate is increased by 0.6%. It shows that the proposed method can meet the requirement of speed performance of risk prediction.

5. Conclusion

From the perspective of data mining, this paper presents a set of operational risk prediction methods for power Internet of Things, including fusion and balance processing of risk data. Bayes optimization algorithm is used to improve CatBoost algorithm, and key parameters in CatBoost model are optimized to build an integrated operational risk prediction model for power Internet of Things. The results show that compared with ReliefF, the time required by ReliefF-S algorithm is 211 seconds, which is reduced by 25 seconds, indicating that ReliefF-S algorithm is superior to ReliefF algorithm and can effectively reduce the data dimension. Compared with CatBoost model, the overall accuracy of BO-CatBoost model is increased by 0.6%, and the accuracy rate, recall rate and F1-Score are increased by 5.18%, 0.72% and 3.09%, respectively, indicating that parameter optimization further improves the performance of the model. This study mainly forecasts the operational risks of power iot from the perspective of data mining. In the future, the transmission path and scope of risks in the power Internet of things can be further considered, and the consistency check of business rules can be carried out on the abnormal data set to find business anomalies, so as to more comprehensively solve the security problems faced in the operation process of the power Internet of things.

References

[1] Wu Z, Zhao Y, Zhang N. A literature survey of green and low-carbon economics using natural experiment approaches in top field journal. Green and Low-Carbon Economy. 2023; 1(1): 2-14.

[2] Serat Z, Fatemi SAZ, Shirzad S. Design and economic analysis of on-grid solar rooftop PV system using PVSYST software. Archives of Advanced Engineering Science. 2023; 1(1): 63-76.

[3] Gu Y, Zhang S, Wang W, Pan L, Zhang D, Bao Z. Lithofacies prediction driven by logging-based Bayesian-optimized ensemble learning: A case study of lacustrine carbonate reservoirs. Geophysical Prospecting. 2023; 71(9): 1835-1872.

[4] Liu K, Sun Y, Yang D. The administrative center or economic center: Which dominates the regional green development pattern. A case study of shandong peninsula urban agglomeration, China. Green and Low-Carbon Economy. 2023; 1(3): 110-120.

[5] Wu L, Ye X, Zhang Y, Gao J, Lin Z, Sui B, Wen Y, Wu Q, Liu K, He S. A genetic algorithm-based ensemble learning framework for drug combination prediction. Journal of Chemical Information and Modeling. 2023; 63(12): 3941-3954.

[6] Ajayi A, Oyedele L, Owolabi H, Akinade O, Bilal M, Delgado JMD, Akanbi L. Deep learning models for health and safety risk prediction in power infrastructure projects. Risk Analysis. 2020; 40(10): 2019-2039.

[7] He J, Cai B, Yan W, Zhang B, Zhang RK. Internet of things-based risk warning system for distribution grid operation state. Journal of Interconnection Networks. 2022; 22(3): 5007-5028.

[8] Qu Z, Xie Q, Liu Y, Li Y, Cui M. Power cyber-physical system risk area prediction using dependent markov chain and improved grey wolf optimization. IEEE Access. 2020; 8(75): 82844-82854.

[9] Li Q, Meng S, Zhang S. Safety Risk monitoring of cyber-physical power systems based on ensemble learning algorithm. IEEE Access. 2019; 7(12): 24788-24805.

[10] Kong X, Xu Y, Jiao Z, Dong D, Yuan X, Li S. Fault location technology for power system based on information about the power internet of things. IEEE Transactions on Industrial Informatics. 2020; 16(10): 6682-6692.

[11] Li W, Zhang N, Liu Z, Ma S, Ke H, Wang J, Chen T. MLfus: A real-time forecasting architecture for low communication costs in electricity IoT based on ensemble learning. IET communications. 2023; 17(2): 145-161.

[12] Piotrowski P, Kopyt M, Baczyński D, Robak S, Gulczyński T. Hybrid and ensemble methods of two days ahead forecasts of electric energy production in a small wind turbine. Energies. 2021; 14(5): 1225-150.

[13] Kim DH. PSO based optimized ensemble learning and feature selection approach for efficient energy forecast. Electronics. 2021; 10(18): 2188-2205.

[14] Wang J, Zhang D, Zhou Y. Ensemble deep learning for automated classification of power quality disturbances signals. Electric Power Systems Research. 2022; 213(12): 78-85.

[15] Larrea M, Porto A, Irigoyen E, Barragan AJ, Manuel Andujar J. Extreme learning machine ensemble model for time series forecasting boosted by PSO: Application to an electric consumption problem. Neurocomputing. 2021; 452: 465-472.

[16] Cao H, Wu Y, Bao Y, Feng X, Wan S, Qian C. UTrans-Net: A model for short-term precipitation prediction. Artificial Intelligence and Applications. 2023; 1(2): 106-113.

[17] Sun X, Hu J, Zhang Z, Cao D, Huang Q, Chen Z, et al. Electricity theft detection method based on ensemble learning and prototype learning. Journal of Modern Power Systems and Clean Energy, 2023, 12(1): 213-224.

[18] Tripathi A K, Pandey A C, Sharma N. A new electricity theft detection method using hybrid adaptive sampling and pipeline machine learning. Multimedia Tools and Applications, 2024, 83(18): 54521-54544.

[19] Ly A, El-Sayegh Z. Tire wear and pollutants: An overview of research. Archives of Advanced Engineering Science. 2023; 1(1): 2-10.

[20] Lin Q, Gong Y, Shi Y, Feng C, Zhang Y. A Fault Risk Warning Method of Integrated Energy Systems Based on RelieF-Softmax Algorithm. CMES-Computer Modeling in Engineering & Sciences, 2022, 132(3): 929-944.

[21] Kumar VTRP, Arulselvi M, Sastry KBS. Comparative assessment of colon cancer classification using diverse deep learning approaches. Journal of Data Science and Intelligent Systems. 2023; 1(2): 128-135.

[22] Tommarello N D F, Deek R A. The Convergence of the Internet of Things and Artificial Intelligence in Medicine: Assessing the Benefits, Challenges, and Risks. Computer, 2024, 57(2): 95-99.

[23] Sharma R, Villányi B. Safe and secure oil and gas pipeline transportation system based on Industrial Internet of Things. IEEE Sensors Journal, 2024, 24(5): 6834-6845.

[24] Guo P, Xiao K, Wang X, Li D. Multi-source heterogeneous data access management framework and key technologies for electric power Internet of Things. Global Energy Interconnection, 2024, 7(1): 94-105.