IEEE Access
Multidisciplinary ⋮ Rapid Review ⋮ Open Access Journal

Received 10 July 2024, accepted 24 July 2024, date of publication 2 August 2024, date of current version 14 August 2024.

Digital Object Identifier 10.1109/ACCESS.2024.3437368

RESEARCH ARTICLE

Line Loss Anomaly Perception Method Based on MIC-IF Algorithm for Photovoltaic Low-Voltage Transformer Area

PENGHE ZHANG[1], YINING YANG[1], RUNAN SONG[1], BICHENG WANG[1], JINHAO SHENG[2], AND BO ZHAO[2]

[1]China Electric Power Research Institute, Beijing 100192, China
[2]School of Automation, Beijing Information Science and Technology University, Beijing 100192, China

Corresponding author: Bo Zhao (13910889512@126.com)

This work was supported by State Grid Corporation of China Headquarters Technology Project under Grant 5700-202255222A-1-1-ZN.

ABSTRACT The topology and line parameters of low-voltage transformer area are often difficult to obtain, and distributed photovoltaic (PV) access makes the distribution grid's power flow characteristic changes, which leads to high difficulty in recognizing line loss anomalies in transformer areas. To address the above problems, a method for perceiving line loss anomaly in PV transformer area called MIC-IF is proposed. Based on this algorithm, the feature vectors of each transformer area are constructed by combining several operation indicators and line loss rate, and it is considered that the outlier vectors correspond to the transformer areas with abnormal operation states. After completing the judgment of the cause of anomalies, PV energy theft detection is carried out for areas in which the PV power generation is anomalous based on RUSBoost algorithm. Finally, the results of analysis are summarized and the conclusion on anomaly perception is obtained. The effectiveness of the proposed method is verified based on data from 20 simulation transformer areas, and the results show that the accuracy and F1-socre of MIC-IF reach 0.95 and 0.89, respectively, and are higher than the comparison algorithm. The detection framework takes into account the PV access and does not rely on line parameters, with high interpretability and accuracy, providing a certain reference for engineering applications.

INDEX TERMS Anomaly perception, ensemble learning, line loss rate, maximal information coefficient, photovoltaic low-voltage transformer area.

I. INTRODUCTION

Line loss rate can effectively reflect the economic status of the power system [1]. Low-voltage distribution grid lines are relatively complex and change frequently, its management and maintenance are difficult, and its loss occupies a large part of the total loss of the power system [2]. With the rapid development of photovoltaic (PV) industry, the access of distributed photovoltaic causes some changes in the power flow characteristics of transformer areas [3], [4], [5]. In addition, in order to obtain additional PV power generation subsidies, some photovoltaic customers attempted to implement energy theft by tampering with meter readings, resulting in more complex operating conditions and greater difficulty in identifying anomalies in PV transformer areas [6]. The popularity of smart meters provides favorable conditions for the application of data-driven techniques in line loss anomaly analysis [7]. In summary, there is an urgent need to conduct research on data-driven line loss characterization and anomaly perception methods for PV low-voltage transformer area.

At present, directly comparing line loss rate with pre-set threshold is still the main way to recognize line loss anomalies in distribution network, and researchers in various countries have carried out a wealth of studies on related topics. Some of the studies focus on calculating theoretical line loss and consider it as the reference for line loss analysis. Reference [8] calculated the theoretical line loss based on

The associate editor coordinating the review of this manuscript and approving it for publication was Wencong Su.

improved equivalent resistance method, then compared it with the actual measured line loss to obtain the identification results. Reference [9] achieved accurate prediction of theoretical line loss in transformer areas through IBP neural network and optimized the hyper-parameters of the network based on CGA algorithm. Reference [10] proposed a line loss analysis method based on GA-LMBP algorithm, and the related model has good computational accuracy and generalization ability. Reference [11] described a line loss analysis method based on kernel density estimation and BP neural networks. References [12] and [13] both used CNN as the computational model for theoretical line loss and implemented the work of feature extraction before computation based on ReliefF algorithm and marketing customer portrait analysis, respectively. Reference [14] classified the abnormal data based on AP clustering, then found potential anomalies by comparing the characteristics of measured data and abnormal data, and carried out line loss correction by DNN algorithm. Reference [15] improved the Random Forest algorithm and obtained a high accuracy of theoretical line loss calculation based on this method. Reference [16] used Gradient Boosting Decision Tree (GBDT) to effectively identify nonlinear patterns among data and achieve the goal of line loss rate prediction. Reference [17] firstly discussed the factors affecting line loss based on PCA algorithm, and then accurately predicted line loss using improved CHAID Decision Tree.

While another part of the research achieves the goal of line loss anomaly detection through clustering or outlier detection algorithms. Reference [18] analyzed the multidimensional feature similarity between line losses by using improved Hasusdorff distance and identified potential anomalous conditions based on hierarchical clustering algorithm. Reference [19] considered the impact of distributed PV access on transformer areas, selected appropriate indicators based on gray correlation analysis, and then combined with K-means clustering, the anomaly coefficients for each area were calculated. Reference [20] constructed a system of line loss rate influencing factors, and then the indicators and the line loss rate were jointly inputted into VAE model to obtain the anomaly detection result. Reference [21] used neural networks to capture data-driven power flow characteristics of low-voltage distribution networks, thereby effectively identifying loss anomalies and tracking potential suspected users. Reference [22] constructed a semi-supervised hierarchical anomaly detection framework taking into account the problems of sample imbalance and limited labeled samples. Reference [23] implemented the anomaly discrimination function based on SVM, and the paper used Improved Sparrow Search Algorithm (ISSA) to optimize the model parameters. Reference [24] proposed a hybrid cluster detection framework based on K-means and hierarchical clustering, which showed higher detection accuracy and lower false positive rate. Reference [25] clustered the line loss rate data based on K-means as well, and identified the anomalous areas through cluster center distance analysis and anomaly discrete degree analysis. Reference [26] summarized a number of feature indicators for line loss data and inputted them into Boost K-means model for discrimination, which achieved high detection accuracy.

Through the analysis of references, it can be found that a large number of studies focus mainly on the subject of anomaly detection, however, in actual engineering applications, it is also crucial to diagnose the causes of line loss anomalies, a step that can provide guidance for the maintenance and management of transformer areas. Secondly, the existing references are more oriented to traditional single power structure distribution network, however, the access of distributed PV has changed the power flow and line loss characteristics of the grid, and there is an urgent need for a more comprehensive analysis for transformer areas containing PV [27]. In addition, some models are implemented based on power flow analysis, which requires access to the line topology and parameters of the grid, but in practice these data are often difficult to obtain in a comprehensive manner, which reduces the applicability of related methods.

Therefore, a line loss anomaly perception method based on MIC-IF is proposed for photovoltaic low-voltage transformer area. The main contributions of the proposed method to the literature are as follows.

- The proposed method does not depend on the line topology and parameters of the grid and is data-driven, which provides better practical applicability compared to the method of calculating line loss by power flow analysis.
- The detection framework takes into account the access of distributed PV and is able to realize intelligent diagnosis, its function is relatively comprehensive.
- The model extracts the operating characteristics of areas through correlation analysis, which is with good interpretability, and it is also convenient to adjust the indicators and parameters appropriately when applied to different scenarios.

The rest of the paper is organized as follows. In Chapter II, the construction of the anomaly perception framework is introduced, including the parts of transformer area classification, feature extraction, anomaly detection and diagnosis, and PV energy theft detection. Then in Chapter III, the validity of the proposed method is verified by experiments.

II. IDENTIFICATION AND DIAGNOSIS OF LINE LOSS ANOMALY BASED ON MIC-IF ALGORITHM

A. CLASSIFICATION OF TRANSFORMER AREAS

The characteristics of line loss rate are closely related to many factors such as the type of electricity consumption, line distribution, distributed PV access, if there is no differentiation of transformer areas before anomaly detection, it is easy to cause the occurrence of misjudgment and omission.

The classification is based on several inherent attributes of transformer areas, combined with the available information, transformer areas are divided from four dimensions: main type of electricity consumption, power supply radius, number of consumers and distributed photovoltaic access level.

TABLE 1. Rules for the classification of transformer areas.

Category	Attribute			
	Main type of electricity consumption	Power supply radius (m)	Number of consumers	Distributed photovoltaic access level
Category 1	Residential living	0~100	1~100	0~0.5
Category 2	General industry and commerce	100~200	101~250	0.5~1
Category 3	Agricultural production	200~300	251~400	1-1.5
Category 4	Others	300~500	⩾401	⩾1.5

1) MAIN TYPE OF ELECTRICITY CONSUMPTION

Since electricity is used for a variety of purposes, consumers may differ in terms of the scale of electricity consumption and the demand for electricity at different times of a day. Therefore, from the view of the main type of electricity consumption, transformer areas are categorized into four categories: residential living, general industry and commerce, agricultural production and others.

2) POWER SUPPLY RADIUS

This variable captures the length of lines between the substation and consumers in transformer areas, which portrays the level of line loss rate. In general, urban areas have a higher concentration of consumers and a smaller supply radius than suburban and rural areas.

3) NUMBER OF CONSUMERS

The combination of this variable and power supply radius can effectively describe the consumer density of transformer areas, which helps to realize the clustering of similar areas.

4) DISTRIBUTED PHOTOVOLTAIC ACCESS LEVEL

Appropriate photovoltaic power generation can shorten the power supply distance for some consumers and effectively reduce line loss, at this time, the line loss rate of anomalous transformer area may not be high [27]. However, excessive PV generation can also lead to power being sent back to the substation and higher line loss [19]. The distributed PV access level is calculated by synthesizing the installed PV capacity of transformer areas and the general meter information, as shown in Equation (1).

$$\alpha = \frac{T \sum_{i=1}^{m} S_{\mathrm{pv},i}}{W_{\mathrm{z,T}} + T \sum_{i=1}^{m} S_{\mathrm{pv},i}} \tag{1}$$

In Equation (1), α is the distributed photovoltaic access level of the transformer area. $S_{\mathrm{pv},i}$, $W_{\mathrm{z,T}}$ are the installed capacity of the i-th PV customer in the transformer area and the power supply of the general meter in the time period to be analyzed, respectively. T is the time span of the data.

Based on the above analysis and considering the actual situation, the transformer area classification rules are obtained as shown in Table 1.

B. EXTRACTION OF OPERATION INDICATORS

The causes of line loss anomaly can be categorized into technical and non-technical reasons. Among them, technical reasons mainly include severe three-phase imbalance, no-load or overload, and low power factor while non-technical reasons mainly include energy theft and meter malfunction.

Therefore, several operation indicators are selected and the correlations between them and line loss rate are calculated to portray the operation status of transformer areas. In turn, it is able to perform comparative analysis between areas and identify potential abnormal line loss situation.

1) PHOTOVOLTAIC PENETRATION RATE

Photovoltaic penetration rate is the percentage of total electricity supplied by distributed PV over a period of time, reflecting the extent to which distributed PV supports electricity in the transformer area. It is calculated as shown in Equation (2).

$$\lambda = \frac{\sum_{i=1}^{m} W_{\mathrm{pv},i}}{W_{\mathrm{z}} + \sum_{i=1}^{m} W_{\mathrm{pv},i}} \tag{2}$$

In Equation (2), $W_{\mathrm{pv},i}$ is the amount of electricity supplied per unit of time by the i-th PV customer, and W_{z} is the amount of electricity supplied by the general meter during the same period.

2) THREE-PHASE IMBALANCE RATE

Excessive three-phase imbalance rate can lead to reduced equipment operating efficiency and increased heating of devices and lines, posing a serious hazard to the operation of transformer areas. According to the enterprise standard of State Grid Corporation of China (Q/GD W519-2010) "Procedure for distribution network" [28], this indicator is calculated as shown in Equation (3).

$$\varepsilon = \frac{I_{\max} - I_{\min}}{I_{\max}} \times 100\% \tag{3}$$

In Equation (3), I_{max} and I_{min} are the maximum and minimum current RMS values in the three phases, respectively.

3) LOAD FACTOR

This indicator reflects the current status of load carrying for transformer and PV systems. Typically, the inherent loss in the lines and devices makes the line loss rate of the transformer area high when the area is unloaded, and overloading causes the equipment to operate less efficiently, which may likewise lead to an increase in line loss rate.

4) POWER FACTOR

Power factor is the ratio of active power to apparent power and reflects the efficiency of energy utilization of electrical equipment in the distribution network. Excessive inductive load access or improper configuration of equipment may result in increased reactive power, lower power factor, and higher line loss rate in transformer areas.

The symbols corresponding to the above four operation indicators are shown in Table 2.

TABLE 2. Symbols of operation indicators.

Operation indicator	Symbol
Photovoltaic penetration rate	λ
Three-phase imbalance rate	ε
Load factor	η
Power factor	p

C. FEATURE VECTOR CONSTRUCTION AND OUTLIER DETECTION

Based on the analysis in the previous section, the line loss rate of a normal transformer area can be expressed as Equation (4).

$$L_r = f(\lambda, \varepsilon, \eta, p) \tag{4}$$

In Equation (4), L_r is the line loss rate of the area.

If an area has line loss anomaly caused by an abnormal technical indicator, for example, a high three-phase imbalance rate, the line loss rate function can be expressed as Equation (5).

$$L_r = f(\lambda, \varepsilon, \eta, p, \varphi(\varepsilon)) \tag{5}$$

In Equation (5), consider $\varphi(\varepsilon)$ as the additional loss due to severe three-phase imbalance, which is variable with ε. The correlation between the sequence of indicator ε and L_r for this area under this condition should be higher than the normal range.

If there is energy theft by electricity consumers under a transformer area, the line loss rate function can be expressed as Equation (6).

$$L_r = f(\lambda, \varepsilon, \eta, p, W_{theft}) \tag{6}$$

In Equation (6), W_{theft} is the amount of electricity stolen by the electricity consumers, which is determined by the consumer's behavior and not affected by the operation indicators. Therefore, the independent variables of the function increase in this case, and the correlation between each operation indicator and the line loss rate sequence should be lower than the normal range.

The construction of feature vectors of transformer areas is realized through the correlation analysis between the operation indicators and line loss rate. Maximal information coefficient (MIC) can effectively evaluate the linear and nonlinear correlation between two time series with high robustness and low computational complexity [29], [30], so it is based on it to realize the time series correlation analysis.

Assume that the time series of photovoltaic penetration rate and line loss rate for transformer area k are $\Lambda = \{\lambda_t, t = 1, 2 \ldots J\}$ and $L = \{l_t, t = 1, 2 \ldots J\}$, respectively. Where J is the length of the time series. The mutual information (MI) and MIC of the two sequences can be obtained as shown in Equations (7) and (8), respectively.

$$M(\lambda, l) = \sum_{\lambda \in \Lambda} \sum_{l \in L} p(\lambda, l) \log_2 \frac{p(\lambda, l)}{p(\lambda)\, p(l)} \tag{7}$$

$$c(\lambda, l) = \max_{ab<D} \frac{M(\lambda, l)}{log_2 \min(a, b)} \tag{8}$$

In Equations (7) and (8), $M(\lambda,l)$ and $c(\lambda,l)$ each represent the mutual information and MIC. $p(\lambda,l)$ is the joint probability density, $p(\lambda)$ and $p(l)$ are the edge probability densities of variables λ and l, respectively, and a and b are the number of intervals into which each dimension is divided, $D = J^{0.6}$ [29].

The correlations between the other three operation indicators and line loss rate time series can be calculated in the same way as for the photovoltaic penetration rate. Then, the feature vector of transformer area k can be obtained as shown in Equation (9).

$$C_k = \left[c_k^{\lambda}, c_k^{\varepsilon}, c_k^{\eta}, c_k^{p}\right]^T \tag{9}$$

Vector C_k shown in Equation (9) reflects the operation status of transformer area k during the time period being analyzed. Assuming that a total of K areas is involved in the analysis, summarizing the feature vectors of each transformer area gives the total matrix as shown in Equation (10).

$$C = [C_1, C_2, \ldots, C_K] \tag{10}$$

Since the analyzed transformer areas have similar attributes, their feature vectors are also considered to be not very different, and the transformer areas corresponding to outlier vectors can be regarded as areas with abnormal line loss rate.

Isolation forest (IF) algorithm has good outlier detection performance. The principle is to randomly select features and segmentation points, so that the sample falls to the leaf nodes in the binary tree to form an isolated tree. In turn, multiple isolated trees are constructed, and the abnormal sample identification function can be realized by observing

the path length of the samples from the root node in trees [31]. Typically, outlier samples are more likely to be segmented into leaf nodes earlier compared to normal samples, and the paths they travel through are relatively short. The degree of sample outliers is calculated as shown in Equations (11) and (12).

$$s(c, K) = 2^{-\frac{E(u(c))}{v(K)}} \tag{11}$$

$$v(K) = 2[\ln(K-1) + \gamma] - \frac{2(K-1)}{K} \tag{12}$$

In Equations (11) and (12), $s(c, K)$ is the anomaly score of sample c, $E(u(c))$ is the average path length of c in the isolated trees, and γ is Euler's constant.

The matrix C is partitioned in terms of the dimensions of the operation indicators and four rounds of outlier detection are performed. If the anomaly score of an element is higher than the threshold θ_1 =0.6, the transformer area corresponding to the vector in which the element is located is considered to be anomalous.

D. DIAGNOSIS OF LINE LOSS ANOMALY

Oriented towards the anomalous transformer areas detected in the previous section, the causes of their anomalies are discussed. The mean values of the anomaly scores for each of the operation indicators in normal transformer areas are calculated separately, and they are used as the basis for dividing the outlier elements into upper and lower outlier elements. Using the photovoltaic penetration rate as an example, the mean value of the score is calculated as shown in Equation (13).

$$\overline{\lambda} = \frac{1}{K^*} \sum_{k^*=1}^{K^*} \lambda_{k^*} \tag{13}$$

In Equation (13), K^* is the number of normal transformer areas. The calculation of the mean values of the other three indicators is similar to $\overline{\lambda}$. If the value of outlier element is higher than the corresponding mean value, name it as upper outlier element, otherwise name it as lower outlier element. The diagnosis of line loss anomaly is based on the following rules.

1) RULE 1

If there is an upper outlier element in the outlier vector, the anomaly is considered to be caused by the operation indicator corresponding to that element.

2) RULE 2

If there is no upper outlier element in the outlier vector, the anomaly is considered to originate from the electricity consumer side, in other words, there may be some energy theft electricity consumers in the transformer area.

3) RULE 3

If there are multiple upper outlier elements in the outlier vector, the indicator corresponding to the element with the highest anomaly score among them is considered as the cause of line loss anomaly.

E. DETECTION OF PHOTOVOLTAIC ENERGY THEFT

Considering that line loss anomaly of transformer areas may be caused by a combination of PV customer energy theft and electricity consumer energy theft, the PV energy theft detection is carried out for transformer areas diagnosed with abnormal photovoltaic penetration rate indicator.

1) SELECTION OF METEOROLOGICAL FACTORS

Photovoltaic power generation is closely related to the meteorological conditions of the region. Therefore, the correlation analysis between numerous meteorological indicators and PV power generation is carried out using historical data, and the variables with strong correlation with PV power generation are selected for subsequent analysis. This step is implemented based on Pearson Correlation Coefficient (PCC), which can accurately measure the linear correlation between two sequences. PCC is able to be obtained by Equation (14).

$$P(x, y) = \frac{\sum_{t=1}^{J} (x_t - \overline{x})(y_t - \overline{y})}{\sqrt{\sum_{t=1}^{J} (x_t - \overline{x})^2} \sqrt{\sum_{t=1}^{J} (y_t - \overline{y})^2}} \tag{14}$$

In Equation (14), x_t and y_t are the t-th number in the time series.

2) CONSTRUCTION OF PHOTOVOLTAIC ENERGY THEFT DETECTION MODEL

Since energy theft data is much less than normal data, the sample set available for model training is usually unbalanced. Therefore, the photovoltaic energy theft detection model is constructed based on RUSBoost algorithm. RUSBoost is combined by Random Under-sampling and AdaBoost. Its basic principle is that randomly selected majority class samples and minority class samples form a balanced data set used to train the weak learners, and then the output of weak learners are weighted and combined to obtain the final result. This algorithm is able to show good training results when the sample set is unbalanced [32].

Variables required for the model include meteorological indicators and the amount of electricity generated per unit of time and capacity by PV customers. After completing the discrimination, the density of anomalies for each PV customer in each time period is counted using sliding windows. Let the sequence of discrimination result for a PV customer be $R = \{r_t, t = 1,2,\ldots, J\}$, and each sliding window contains T_{win} metering points, if the number of anomalies in the window exceeds the threshold θ_2 =0.5T_{win}, the customer is considered to have PV energy theft in that time period. For such PV customers, with reference to the generation of normal customers in the same region during the same period, the amount of electricity stolen by them is estimated and the time series of line loss rate of the transformer areas to which

they belong are corrected. The new line loss rate time series can be calculated as shown in Equation (15).

$$L_{\text{new}} = \frac{W_{\text{loss}} - \sum_{n=1}^{N^*} S_n \left(W_{\text{pv,unit},n} - W_{\text{pv,unit,average}}\right)}{W_{\text{z}} + \sum_{i=1}^{m} W_{\text{pv},i}} \quad (15)$$

In Equation (15), W_{loss} is the original line loss of the transformer area, $W_{\text{pv,unit},n}$ and $W_{\text{pv,unit,average}}$ are the measured value of the unit capacity generation of PV energy theft customer n and the average value of the generation of normal PV customers in the same region, respectively. S_n is the installed capacity of PV customer n. N^* is the number of PV energy theft customer in the area. It is worth noting that when calculating the denominator, if user i is an energy theft customer, its generation $W_{\text{pv},i}$ needs to be substituted in combination with $W_{\text{pv,unit,average}}$ and its own installed capacity.

The corrected time series are utilized for MIC-IF algorithm re-analysis to determine whether there is any anomaly on the power consumption side of the transformer area, and the detection results are summarized to generate the final diagnostic conclusion.

F. TIME SERIES PREPROCESSING

Since both MIC-IF and RUSBoost require the use of multiple time series data, and the algorithms require the sequences to be of equal granularity and equal length. Therefore, it is recommended that the data collection period for time series is 15 minutes/30 minutes/1 hour and the sequences are guaranteed to be of equal length. In addition, missing values in the sequences are supplemented using Lagrange interpolation.

Based on the analysis in chapter II, the process of line loss anomaly perception in PV low-voltage transformer area can be summarized as shown in Fig. 1.

III. EXPERIMENTS

In order to verify the effectiveness of the above method, 20 simulation PV low-voltage transformer areas are constructed in combination with the operation data of a PV power station in western China and the local meteorological data of the same period. The simulation and programming platform is MATLAB/Simulink and the processor specification is Intel(R) Core(TM) i7-1260P. Refer to Table 1, the main type of electricity consumption, power supply radius, number of consumers and distributed PV access level of the simulation transformer areas are residential living, 0-100 m, 1-100 and 0.5-1. The time span of the time series for each variable is 1 week and the data collection intervals are all 15 minutes.

Then, some of these transformer areas are selected for line loss anomaly transformation, obtaining the electricity consumer energy theft area (Area 5), the PV customer energy theft area (Area 10), the area where both types of energy theft exist (Area 15), and the power factor anomaly area (Area 20).

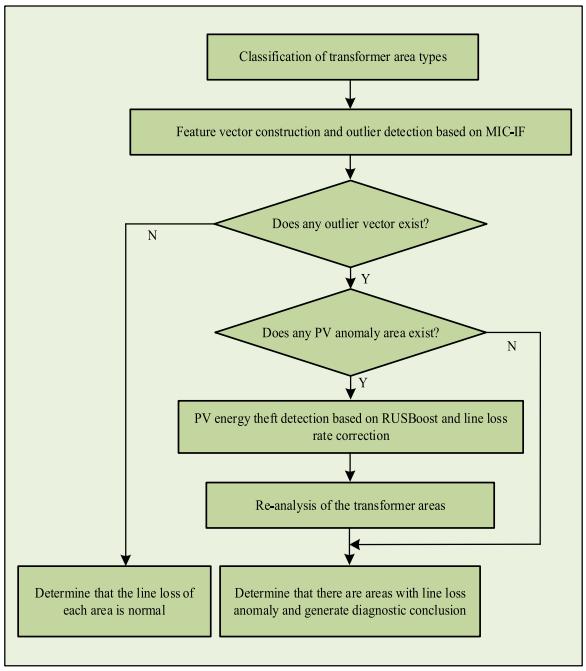

FIGURE 1. Process of line loss anomaly perception in photovoltaic low-voltage transformer area.

Based on MIC-IF algorithm, the result of feature vector extraction for these 20 areas is shown in Table 3, and the result is plotted as a four-dimensional visualization figure as shown in Fig. 2. In addition, the result of anomaly scoring is shown in Fig. 3.

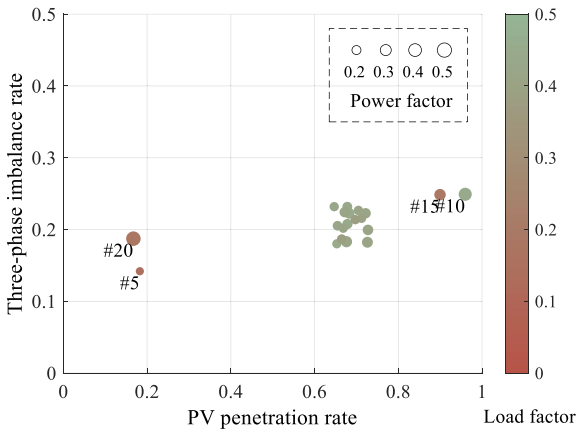

FIGURE 2. Feature vector visualization figure for the transformer areas.

As can be seen from Fig. 3, there are outlier elements in the feature vectors of area 5, 10, 15, and 20. Based on Table 3, Equation (13) and related rules, anomaly diagnosis is made for the above areas. Since the feature vector elements of area 5 are all lower outliers, it is determined that there is an anomaly on the electricity consumer side of this transformer area. In the vectors of area 10 and 15, PV penetration rate is the highest upper outlier element, so it is determined that there are anomalies on the PV customer side of these two areas,

TABLE 3. Result of feature vector extraction.

Area number	Feature vector			
	PV pen. rate	Three-phase imba. rate	Load factor	Power factor
1	0.65	0.18	0.45	0.21
2	0.68	0.21	0.44	0.26
3	0.70	0.23	0.42	0.22
4	0.68	0.22	0.44	0.29
5	0.18	0.14	0.15	0.17
6	0.73	0.18	0.42	0.29
7	0.65	0.21	0.43	0.23
8	0.68	0.18	0.41	0.31
9	0.67	0.20	0.41	0.21
10	0.96	0.25	0.45	0.42
11	0.65	0.23	0.43	0.21
12	0.72	0.22	0.41	0.26
13	0.73	0.20	0.42	0.28
14	0.71	0.22	0.40	0.21
15	0.90	0.25	0.19	0.32
16	0.68	0.23	0.44	0.24
17	0.67	0.22	0.41	0.26
18	0.70	0.21	0.38	0.22
19	0.66	0.19	0.38	0.22
20	0.17	0.19	0.18	0.52

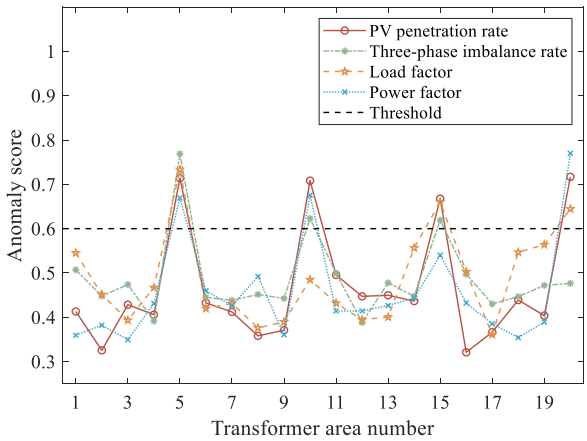

FIGURE 3. Result of anomaly scoring.

and it is necessary to carry out PV energy theft detection and re-analysis. In addition, the power factor is the highest upper outlier element in the vector of area 20, so the line loss anomaly of this transformer area is diagnosed as power factor anomaly.

Based on RUSBoost algorithm, PV energy theft detection is carried out for area 10 and 15. The selected meteorological variables include temperature, humidity, and radiation intensity. The length of the sliding window is set to 6 hours, that is, 24 metering points, and the distance of the sliding window moving each time is also 24 points. The detection results are shown in Fig. 4.

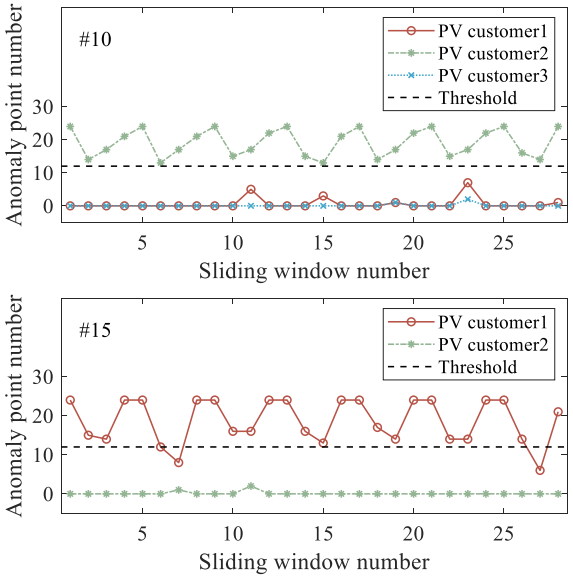

FIGURE 4. Result of photovoltaic energy theft detection.

The result shows that one PV customer in each of the two areas had indeed committed energy theft. The amount of electricity stolen by abnormal customers is evaluated and the time series of line loss rate are corrected, then the transformer areas are re-analyzed. The result is shown in Table 4, Fig. 5 and Fig. 6.

TABLE 4. Result of feature vector extraction of re-analysis.

Area number	Feature vector			
	PV pen. rate	Three-phase imba. rate	Load factor	Power factor
5	0.18	0.14	0.15	0.17
10	0.69	0.18	0.43	0.26
15	0.29	0.18	0.17	0.20
20	0.17	0.19	0.18	0.52

According to Table 4, Fig. 5 and Fig. 6, it can be seen that the feature vector of area 10 are no longer anomalous, while PV penetration rate and load factor element in the vector of area 15 are still outliers and they are both lower outlier elements. Therefore, it is determined that an anomaly also existed on the electricity consumer side of area 15.

In summary, the overall results of the proposed method are as expected, although there are cases where a normal area (Area 8) is mistakenly detected as abnormal in the re-analysis.

In order to verify the effectiveness of the proposed method, the model in reference [25] is used to analyze the above areas

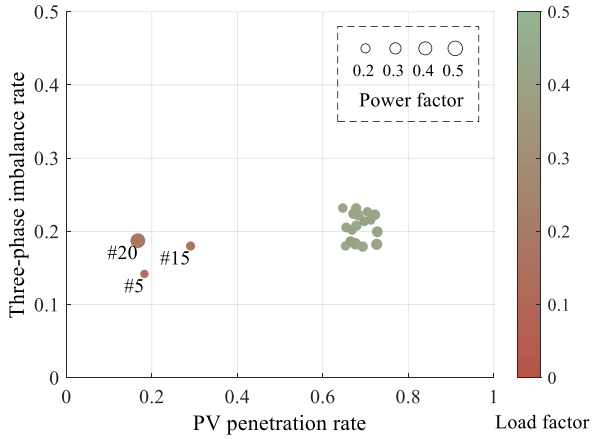

FIGURE 5. Feature vector visualization figure for re-analysis.

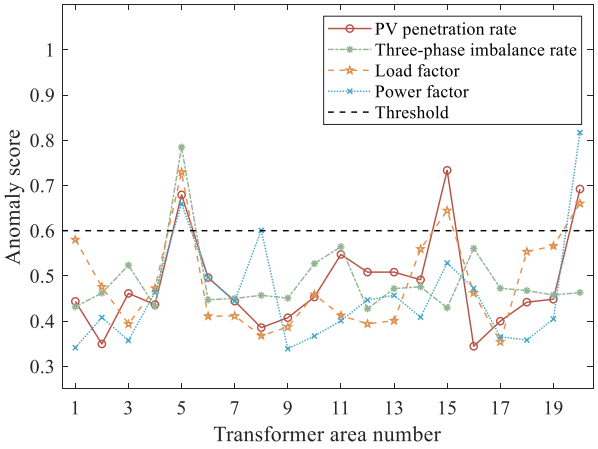

FIGURE 6. Re-analysis result of anomaly scoring.

as well, and the detection result is evaluated using Accuracy and F1-score, as shown in Table 5.

TABLE 5. Result of model comparison.

Method	Accuracy	F1-score
MIC-IF	0.95	0.89
K-means [25]	0.90	0.67

As shown in Table 5, the Accuracy and F1-score of the proposed method reach 0.95 and 0.89, respectively, which are higher than the comparison algorithm, and the model can provide a reference for practical engineering applications to a certain extent.

IV. CONCLUSION

In this paper, a line loss anomaly perception method based on MIC-IF is proposed for PV low-voltage transformer area, and the main work is as follows.

Firstly, the transformer areas to be analyzed are classified based on several intrinsic attributes.

Secondly, MIC-IF algorithm is used to construct transformer area feature vectors to identify and diagnose potential abnormal areas.

For areas with abnormal PV penetration rate, based on RUSBoost, PV energy theft detection is carried out by combining PV power generation and regional meteorological data, then the amount of stolen energy is evaluated and the line loss rate time series is corrected for re-analysis.

Finally, the analysis results are summarized and the final conclusion is generated.

Of course, the proposed method still has some limitations and shortcomings. The application of the method presupposes that transformer areas basically do not have multiple causes of anomalies in the same time period, and such cases may lead to a decrease in the detection accuracy of the model. In addition, the method is implemented based on the horizontal comparison between transformer areas, and is applied on the assumption that the vast majority of the areas to be analyzed are normal, if the percentage of abnormal areas is too high, the accuracy of the method may be reduced due to anomaly contamination, and it is necessary to combine with other methods to carry out a comprehensive analysis.

REFERENCES

[1] W. Zhou, Y. Li, Y. Guo, X. Qiao, Y. Mei, and W. Deng, "Daily line loss rate forecasting of a distribution network based on DAE-LSTM," *Power Syst. Protection Control*, vol. 49, no. 17, pp. 48–56, Sep. 2021.

[2] H. Zhang, J. Zhao, X. Wang, and Y. Xuan, "Low-voltage distribution grid topology identification with latent tree model," *IEEE Trans. Smart Grid*, vol. 13, no. 3, pp. 2158–2169, May 2022.

[3] M. Markovic, A. Sajadi, A. Florita, R. Cruickshank III, and B.-M. Hodge, "Voltage estimation in low-voltage distribution grids with distributed energy resources," *IEEE Trans. Sustain. Energy*, vol. 12, no. 3, pp. 1640–1650, Jul. 2021.

[4] Y. Z. Gerdroodbari, R. Razzaghi, and F. Shahnia, "Decentralized control strategy to improve fairness in active power curtailment of PV inverters in low-voltage distribution networks," *IEEE Trans. Sustain. Energy*, vol. 12, no. 4, pp. 2282–2292, Oct. 2021.

[5] Y. He, M. Wang, Z. Xu, and Y. Jia, "A novel control for enhancing voltage regulation of electric springs in low-voltage distribution networks," *IEEE Trans. Power Electron.*, vol. 38, no. 3, pp. 3739–3751, Mar. 2023.

[6] M. Shaaban, U. Tariq, M. Ismail, N. A. Almadani, and M. Mokhtar, "Data-driven detection of electricity theft cyberattacks in PV generation," *IEEE Syst. J.*, vol. 16, no. 2, pp. 3349–3359, Jun. 2022.

[7] R. Huang, C. Lu, S. Hu, D. Tang, and B. Pan, "Research on characteristics of electricity theft behavior and risk early warning technology based on big data information of the Internet of Things," in *Proc. IEEE Int. Conf. Electr. Eng., Big Data Algorithms (EEBDA)*, Changchun, China, Feb. 2022, pp. 123–127.

[8] D. Tang, Y. Liu, Z. Xiong, T. Ma, and T. Su, "Early warning method of electricity anti-theft in distribution station area based on spatiotemporal correlation matrix," *Autom. Electr. Power Syst.*, vol. 44, no. 19, pp. 168–176, Oct. 2020.

[9] Y. Liu, Y. Li, and C. Li, "Research on line loss calculation of urban cable line," in *Proc. IEEE Int. Conf. Mechatronics Autom. (ICMA)*, Aug. 2021, pp. 397–402.

[10] Z. Jiang, G. Li, Y. Cai, J. Li, and K. Zhang, "Design of line loss rate calculation method for low-voltage desk area based on GA-LMBP neural network model," *IEEE Access*, vol. 11, pp. 144394–144407, 2023.

[11] X. Miao, Z. Ou, M. Yang, J. Yuan, Y. Cao, S. Huang, and W. Liu, "A transformer district line loss calculation method based on data mining and machine learning," in *Proc. 4th Asia Energy Electr. Eng. Symp. (AEEES)*, Chengdu, China, Mar. 2022, pp. 909–915.

[12] R. Liu, F. Pan, Y. Yang, W. Hong, Q. Li, K. Lin, and S. Liu, "Theoretical line loss calculation method for low-voltage distribution network via matrix completion and ReliefF-CNN," *Energy Rep.*, vol. 9, pp. 1908–1916, Sep. 2023.
[13] S. Sun, X. Gao, D. Yang, H. Ma, and M. Li, "Calculation method of benchmark value of line loss rate in transformer district considering marketing customer portrait," *Electr. Power Syst. Res.*, vol. 223, Oct. 2023, Art. no. 109618.
[14] Z. Sicheng, X. Jiguang, F. Zhibo, D. Sitong, and Q. Junji, "Abnormal line loss data detection and correction method," in *Proc. 4th Asia Energy Electr. Eng. Symp. (AEEES)*, Chengdu, China, Mar. 2022, pp. 832–837.
[15] L. Huang, G. Zhou, J. Zhang, Y. Zeng, and L. Li, "Calculation method of theoretical line loss in low-voltage grids based on improved random forest algorithm," *Energies*, vol. 16, no. 7, p. 2971, Mar. 2023.
[16] M. Yao, Y. Zhu, J. Li, H. Wei, and P. He, "Research on predicting line loss rate in low voltage distribution network based on gradient boosting decision tree," *Energies*, vol. 12, no. 13, p. 2522, Jun. 2019.
[17] X. Tao, M. Yu, J. Zhu, C. Yu, and S. Zhang, "Prediction of zone area line loss anomalies based on PCA and improved CHAID decision tree," in *Proc. 34rd Youth Academic Annu. Conf. Chin. Assoc. Autom. (YAC)*, Jinzhou, China, Jun. 2019, pp. 221–226.
[18] B. Lin and Z. Yang, "Anomaly recognition of line loss data in power grid stations based on multi-dimensional features," *Power Syst. Protection Control*, vol. 50, no. 9, pp. 172–178, May 2022.
[19] P. Han, S. Chen, N. Zhang, H. Wu, R. Qiu, and Z. Zhang, "Line loss anomaly identification method for low-voltage station area considering distributed PV," *Power Syst. Protection Control*, vol. 51, no. 8, pp. 140–148, Apr. 2023.
[20] Z. Ma, W. Chen, C. Liu, H. Zhu, and L. Wei, "Detection of abnormal line loss rate in low-voltage transformer district based on VAE," in *Proc. IEEE Sustain. Power Energy Conf. (iSPEC)*, Nanjing, China, Dec. 2021, pp. 3739–3744.
[21] Z. Sun, Y. Xuan, Y. Huang, Z. Cao, and J. Zhang, "Traceability analysis for low-voltage distribution network abnormal line loss using a data-driven power flow model," *Frontiers Energy Res.*, vol. 11, Sep. 2023, Art. no. 1272095, doi: 10.3389/fenrg.2023.1272095.
[22] W. Li, W. Zhao, J. Li, J. Li, and Y. Zhao, "Abnormal line loss identification and category classification of distribution networks based on semi-supervised learning and hierarchical classification," *Frontiers Energy Res.*, vol. 12, Mar. 2024, Art. no. 1378722, doi: 10.3389/fenrg.2024.1378722.
[23] Y. Liao, W. En, B. Li, M. Zhu, B. Li, Z. Li, and Z. Gu, "Research on line loss analysis and intelligent diagnosis of abnormal causes in distribution networks: Artificial intelligence based method," *PeerJ Comput. Sci.*, vol. 9, Dec. 2023, Art. no. 1753, doi: 10.7717/peerj-cs.1753.
[24] L. Chen, S. Xiaolu, C. Weimin, W. Jiaju, B. Tai, Z. Li, and F. Jianquan, "Distribution network line loss assessment method based on data clustering," in *Proc. 2nd Asian Conf. Frontiers Power Energy (ACFPE)*, Chengdu, China, Oct. 2023, pp. 178–182.
[25] H. Chen, H. Cai, X. Li, Y. Wang, and E. Zheng, "Abnormal line loss identification method for low-voltage substation area based on K-means clustering algorithm," *Southern Power Syst. Technol.*, vol. 13, no. 2, pp. 2–6, Feb. 2019.
[26] J. Chen, A. Zeb, Y. Sun, and D. Zhang, "A power line loss analysis method based on boost clustering," *J. Supercomput.*, vol. 79, no. 3, pp. 3210–3226, Sep. 2022.
[27] F. Karakuş, A. Çiçek, and O. Erdinç, "Integration of electric vehicle parking lots into railway network considering line losses: A case study of Istanbul M1 metro line," *J. Energy Storage*, vol. 63, Jul. 2023, Art. no. 107101.
[28] J. Chen, F. Zhou, L. Gu, H. Yin, L. Zhang, and C. Gao, "Evaluation index and evaluation method of three-phase imbalance treatment effect based on commutation," *IEEE Access*, vol. 10, pp. 101913–101921, 2022.
[29] Q. She, G. Jin, R. Zhu, M. Houston, O. Xu, and Y. Zhang, "Upper limb cortical-muscular coupling analysis based on time-delayed back maximum information coefficient model," *IEEE Trans. Neural Syst. Rehabil. Eng.*, vol. 31, pp. 4635–4643, 2023.
[30] X. Qi, J. Song, H. Qi, and Y. Shi, "Maximum information coefficient feature selection method for interval-valued data," *IEEE Access*, vol. 12, pp. 53752–53766, 2024.
[31] H. Gao, D. Yang, G. Cai, Z. Chen, J. Ma, L. Wang, F. Duan, and B. Wang, "Machine learning-based reliability improvement of ambient mode extraction for smart grid utilizing isolation forest," *IEEE Trans. Power Syst.*, vol. 38, no. 5, pp. 4752–4760, Sep. 2023.
[32] S. Liu, S. You, Z. Lin, C. Zeng, H. Li, W. Wang, X. Hu, and Y. Liu, "Data-driven event identification in the U.S. power systems based on 2D-OLPP and RUSBoosted trees," *IEEE Trans. Power Syst.*, vol. 37, no. 1, pp. 94–105, Jan. 2022.

PENGHE ZHANG received the Ph.D. degree in high voltage and insulation technology from the School of Electrical and Electronic Engineering, Huazhong University of Science and Technology, Wuhan, China, in 2012. She is currently a Professor-Level Senior Engineer with China Electric Power Research Institute. Her research interests include measurement instrumentation fault diagnosis and energy theft detection.

YINING YANG received the B.E. and M.E. degrees in electrical and automation from the School of Electrical and Electronics Engineering, North China Electric Power University, Beijing, China, in 2016 and 2019, respectively. She is currently an Engineer with China Electric Power Research Institute. Her research interests include energy theft detection and measurement instrumentation.

RUNAN SONG received the M.E. degree in electrical and automation from the School of Electrical and Electronics Engineering, Tianjin University, Tianjin, China, in 2020. She is currently an Engineer with China Electric Power Research Institute. Her research interests include energy theft detection and measurement instrumentation.

BICHENG WANG received the M.E. degree from Harbin Normal University, Harbin, China. He is currently an Engineer with China Electric Power Research Institute. His research interests include energy theft detection and measurement instrumentation.

JINHAO SHENG received the B.E. degree from Chongqing Jiaotong University, Chongqing, China, in 2022. He is currently pursuing the M.E. degree with Beijing Information Science and Technology University. His current research interests include energy theft detection and machine learning techniques.

BO ZHAO was born in Qingdao, China, in 1977. He received the B.E. and M.E. degrees in electrical engineering from Beijing University of Aeronautics and Astronautics, in 2000 and 2003, respectively, and the Ph.D. degree in electrical engineering from China Electric Power Research Institute, in 2013. He was an Electrical Engineer at China Electric Power Research Institute, State Grid of China. Since 2018, he has been with Beijing Information Science and Technology University, where he is currently a Researcher and a Professor-Level Senior Engineer. His current research interests include the analysis and control of new energy and energy storage and the protection and control of microgrid.