International Journal of Engineering and Technology Innovation, vol. 15, no. 4, 2025, pp. 401-416

Enhanced Sample-Based Online Fault Identification for Electric Energy Meter Verification Devices

Lin Cong, Zhu Ge*, Zhao Jing, Zhao-Lei He, He Ao

Metering Center, Yunnan Power Grid Co., Ltd., Yunnan, China/ Key Laboratory of Green Energy and Digital Electric Power Measurement, Control and Protection of Yunnan, Yunnan, China

Received 20 February 2025; received in revised form 23 July 2025; accepted 28 July 2025

DOI: https://doi.org/10.46604/ijeti.2025.14869

Abstract

To solve the issues of low accuracy and difficulty of online fault identification for Automatic Verification Devices (AVDs) of Electric Energy Meters (EEMs), a method based on Installed Standard Energy Meters (ISEMs) is proposed. ISEMs are tested concurrently with EEMs undergoing verification, and test data from meter positions are collected online without disrupting AVD operation. The features of the meter positions are constructed, and their principal components are extracted to reduce feature dimensionality. Unlabeled samples are categorized into typical fault categories using the K-means clustering algorithm. A Multi-Class Support Vector Machine model is trained and optimized by Bayesian optimization based on the labeled samples. The model is then employed for AVD online fault identification. Enhanced with Monte Carlo samples augmentation, the proposed approach achieves a 0.35% error rate, a 94.40% accuracy improvement compared to the model without sample enhancement. This method provides a reliable and cost-effective solution for online fault identification of AVDs.

Keywords: automatic verification device (AVD), installed standard electric meter (ISEM), online fault identification, Monte Carlo sample enhancement, multi-class support vector machine (MSVM)

1. Introduction

Electric Energy Meters (EEMs) serve as a foundation for billing both users and power supply companies, ensuring fairness in transactions, and are a crucial part of the power system [1]. EEMs must undergo appearance inspections, pressure tests, and error tests before they can be used [2-3]. With the significant increase in the demand for installing or replacing EEMs [4], Automatic Verification Devices (AVDs) for EEMs have been widely adopted. The AVD is an integrated system that combines automatic transmission facilities and fully automatic verification functions for EEMs, ensuring that the EEMs meet the required metering accuracy standards. The AVD provides electric energy to EEMs under verification and measures the supplied electric energy. A typical AVD consists of multiple verification units, each containing several meter positions for the verification of EEMs. However, during long-term operation, AVDs may suffer from terminal deformation or corrosion, supply deterioration, output instability, and repeatability loss, causing higher verification errors and lower reliability, so online fault identification is urgent. [5-6].

Traditional periodic verification or intermediate checks for AVDs require suspending the verification tasks. They involve testing through a standard electric meter connected to the AVD's verification interface, with identifying faults based on the measured electric energy data and collected electrical signals. This method relies heavily on the experience and subjective

* Corresponding author. E-mail address: zhugeygp@163.com

judgment of the operators, lacking objectivity and accuracy, and cannot effectively identify faults during the device's operation [7]. Researchers pursue online anomaly detection via EEM error analysis. In [8], potential anomalies in AVDs are evaluated from critical points. In [9], unqualified meters are continuously spotted, while thresholds are set empirically. In [10], a semi-supervised SVM with labeled/unlabeled samples is employed to locate abnormal meters. In [11], AVD verification performance is linked to error distributions for online detection of parallel devices. However, due to the lack of data from the AVDs themselves, these methods are unable to pinpoint the root causes of anomalies, necessitating additional manual inspection for fault identification.

To achieve online fault identification, numerous scholars have conducted extensive research. In [12-13], a fault knowledge base and a diagnosis expert system are built, respectively, and fault location is realized. This method uses traditional methods for fault diagnosis, relying on the experience of experts or engineers. In [14], two machine learning methods are integrated to achieve fault classification and location on power transmission lines, providing a novel perspective on fault identification of AVDs. In [15], a novel fault early warning model is developed based on a residual neural network to address the issue of relatively low model accuracy within AVDs. It integrates fault early warning and diagnosis through feature extraction, data training, and classification methods, capable of identifying carrier deformation. These methods generally rely on a large amount of historical error data of electric energy and mainly focus on identifying mechanical faults. In actual verification processes, the AVDs may also experience electrical faults and reliability issues, which, if not promptly detected and repaired, will severely affect the accuracy and reliability of AVDs.

To identify various types of faults, it is necessary to collect more test data from AVDs. AVDs are capable of gathering error test data corresponding to each meter position. However, they are unable to acquire data that is related to electrical aspects and reliability, which can be achieved by Installing Standard Energy Meters (ISEMs). The ISEM is a new type of miniaturized standard electric energy meter, which can be conveniently installed on the meter positions of AVDs. Conforming to the requirements of GB/T 17215.211-2021 Alternating Current Measuring Equipment, it meets the measurement range and accuracy requirements for parameters such as voltage, current, and frequency. Meanwhile, it also complies with the electrical performance requirements specified in GB/T 17215.701-2011 Standard Electric Energy Meters and satisfies the electric energy measurement error limits for the specified accuracy class. Therefore, it can operate stably and reliably for a long time and be used to measure the electrical parameters and electric energy data of AVDs. Different types of faults exhibit distinct characteristics.

When an AVD experiences mechanical faults such as terminal deformation or corrosion, the basic error will deviate significantly; when an electrical fault occurs, it usually manifests as poor power quality, with phase symmetry, current, or voltage distortion exceeding the specified range; reliability faults are directly reflected in abnormal device output stability and measurement repeatability. To gather these data, ISEMs with a higher accuracy class than the AVD are tested simultaneously with the EEMs undergoing verification, collecting current, voltage, phase, power, electric energy, and other data for each meter position, used for fault diagnosis. Given the substantial number of meter positions within AVDs and the restricted quantity of ISEMs, the samples and their corresponding labels gathered for each meter position are likewise constrained. When the number of samples is too small, the diagnosis model cannot sufficiently learn the distribution characteristics of the samples. This deficiency gives rise to underfitting, thereby limiting the practical efficacy of identification [16].

To solve the problem of "small samples," it is usually necessary to enhance the samples. In [17], a data enhancement method is proposed. It combines compressed sensing, dictionary learning, and transfer learning to generate a large number of new samples with target domain characteristics, but this method is computationally intensive and algorithmically complex. In [18-19], Generative Adversarial Networks (GANs) are utilized to generate new samples to expand the training dataset, improving model generalization and partially alleviating data insufficiency issues. However, the GAN method is susceptible to

gradient disappearance and mode collapse, leading to low similarity in generated data. In [20], the Monte Carlo method for data enhancement is introduced to generate a series of credible samples, which helps increase the model's robustness to data disturbances and improve model accuracy and reliability. The Monte Carlo method is highly adaptable, retains the characteristics of the original data well, and is easy to implement for complex data expansion.

To address the aforementioned issues of the inability to perform online fault identification in AVDs, as well as the limitation posed by the limited number of available samples, this study proposes an online fault identification method utilizing ISEMs. After constructing the meter position features using online test data from ISEMs, Principal Component Analysis (PCA) is employed to extract the principal components of the feature data, thereby enhancing the efficiency of data reduction and feature extraction. Afterwards, the *K*-means algorithm is applied to classify and label faults, and a Multi-class Support Vector Machine (MSVM) dedicated to fault identification is trained using the labeled data. Concurrently, the Bayesian optimization algorithm is harnessed to optimize the hyperparameters of the MSVM. Finally, the Monte Carlo method is employed to augment the samples, further improving the model's accuracy in online fault identification.

The main contributions of this paper are as follows:

(1) Online Fault Identification: The method involves testing the ISEMs concurrently with the EEMs under verification. This allows for the real-time collection of verification data from the meter positions, enabling online fault identification without interrupting the normal operation of the AVDs.
(2) Sample Labeling: A machine learning approach is employed to automatically classify and label the samples. This reduces the subjective bias associated with manual labeling and significantly lowers the associated labor costs.
(3) Data Augmentation for Improved Accuracy: The Monte Carlo method is utilized to augment the sample data. This technique enhances the diversity and volume of the training data, thereby improving the accuracy of fault identification even when working with a limited initial sample size.

The remainder of this paper is organized as follows: The second part introduces the testing process and feature data construction for the meter positions. The third part analyzes data characteristics and labels samples. The fourth part constructs and optimizes the online fault identification model. The fifth part presents the results analysis and data augmentation. The final part is the conclusions.

2. Test Procedure and Feature Construction

In this section, ISEMs are first installed at chosen positions and then cyclically relocated across successive batches while the meters are under verification, during which electrical parameters are continuously recorded. From each position, multiple diagnostic features are extracted and statistically summarized to generate high-dimensional feature vectors that collectively constitute a comprehensive dataset. This test procedure enables seamless online fault identification without ever interrupting the ongoing operation of the verification device.

2.1. Test procedure with ISEMs

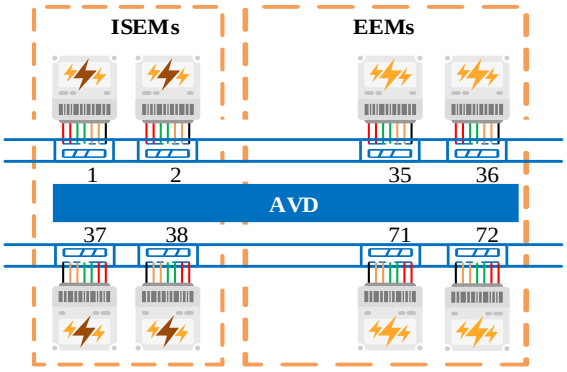
Considering two verification units of an AVD designed for a three-phase EEM, each unit comprises 36-meter positions for verification purposes. To ensure comprehensive coverage and continuous monitoring, four high-precision ISEMs are introduced into the system and systematically relocated across these positions in a cyclic manner while routine verifications proceed uninterrupted. As illustrated in Fig. 1(a) and further clarified by the flow chart in Fig. 1(b), the test procedure proceeds as follows:

First Test Round: Four ISEMs with an accuracy class of 0.02 are selected and installed in meter positions 1, 2, 37, and 38. These ISEMs operate synchronously with the other EEMs undergoing verification. After completing the verification of one

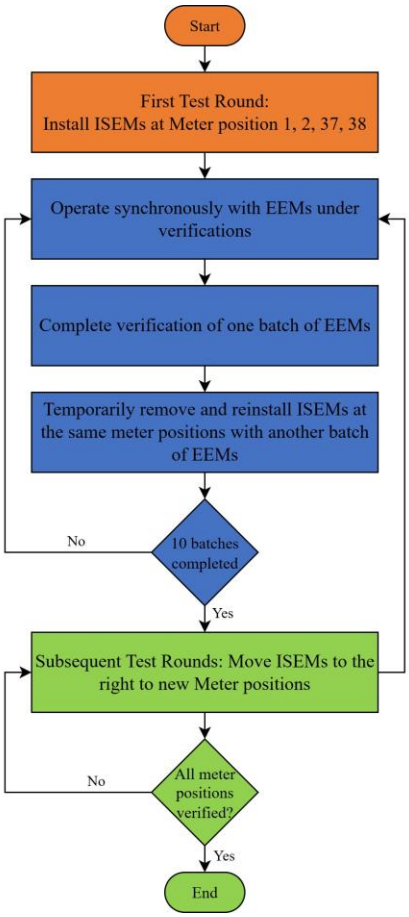
batch of EEMs, the four ISEMs are temporarily removed and then reinstalled to verify the next batch. This process is repeated for a total of 10 batches.

Subsequent Test Rounds: In the second test round, the four ISEMs are moved to the right and installed in meter positions 3, 4, 39, and 40. The verification steps are then repeated. This process of moving the ISEMs and repeating the verification continues until all meter positions have been verified by the ISEMs.

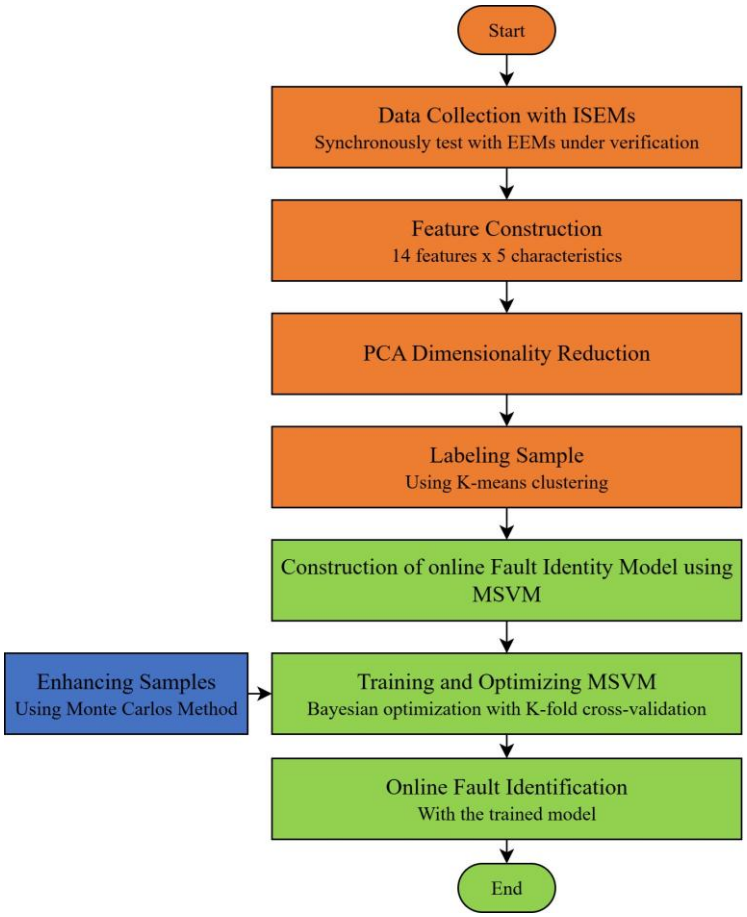
In general, in accordance with JJG 596 (Electrical Meters for Measuring Alternating-current Electric Energy), each verification of an EEM necessitates the completion of 50 load points, with each load point taking approximately 40 seconds. During the verification process of the EEMs, the in-service ISEMs perform data sampling simultaneously.



(a) Test Procedure with ISEMs



(b) Test flow chart



(c) Framework of this method

Fig. 1 Test procedure and framework

The proposed method for online fault identification of AVDs involves synchronously testing ISEMs with EEMs under verification to collect real-time meter position data. After *Z*-score normalization, the PCA method is applied to reduce dimensionality by selecting the top primary components. Unlabeled samples are then categorized into four classes (normal, mechanical, electrical, and reliability faults) using *K*-means clustering. An MSVM with a One-vs-One strategy is trained on labeled data, optimized via Bayesian optimization for hyperparameters. To address small-sample limitations, the Monte Carlo method generates synthetic samples based on the mean and standard deviation of original data, enhancing model training and reducing the error rate for accurate online fault identification without interrupting AVD operations. The framework of this method is illustrated in Fig. 1(c).

2.2. Feature construction

During each test, data collected from the ISEMs at the meter positions are used to calculate 14 distinct features. These features include basic error, phase voltage symmetry, line voltage symmetry, current symmetry, phase-to-phase phase symmetry, line-to-line phase symmetry, A-phase voltage distortion, A-phase current distortion, B-phase voltage distortion, B-phase current distortion, C-phase voltage distortion, C-phase current distortion, device output power stability, and device measurement repeatability. These features are crucial as they capture various aspects of the meter's performance and are instrumental in identifying potential faults.

For each of these 14 features, five characteristic values are derived: maximum (MAX), minimum (MIN), expected value (E), standard deviation (S), and median (M). This results in a comprehensive set of metrics that provide a detailed profile of each feature's behavior. After completing 10 batches of verification for each meter position, each meter position yields 10 rows of feature data, with each row containing 70 columns (14 features × 5 characteristic values). Consequently, with 72-meter positions in total, the entire dataset comprises 720 rows of 70-column feature data. Table 1 presents an example of the feature data for a specific meter position, illustrating the structure and content of the dataset.

Table 1 Feature data for a representative meter position

Features \ Characteristics	Max	Min	E	S	M
Basic error	0.0696	-0.0405	0.0594	0.0364	0.0584
Phase voltage symmetry	0.4276	0.0273	0.4002	0.0131	0.4004
Line voltage symmetry	0.4266	0.0355	0.3994	0.0086	0.3993
Current symmetry	0.8559	-0.0511	0.7970	0.0500	0.8002
Phase-to-phase symmetry	1.7044	0.2029	1.6033	0.0078	1.6034
Line-to-line symmetry	1.7071	0.2057	1.6040	0.0101	1.6039
A-phase voltage distortion	0.9116	0.2197	0.8568	0.1136	0.8665
A-phase current distortion	0.8746	0.2161	0.8271	0.0084	0.8268
B-phase voltage distortion	0.9153	0.2265	0.8601	0.1215	0.8745
B-phase current distortion	0.8742	0.2158	0.8270	0.0080	0.8264
C-phase voltage distortion	0.9181	0.2237	0.8599	0.1196	0.8758
C-phase current distortion	0.8739	0.2116	0.8271	0.0082	0.8269
Device output power stability	0.0464	0.0055	0.0418	0.0084	0.0415
Device measurement repeatability	0.0065	0.0011	0.0058	0.0010	0.0057

2.3. Normalizing features

To ensure that all subsequent analyses and model training are both reliable and unbiased, the raw feature data collected from the ISEMs undergo Z-score normalization. Because the fourteen diagnostic features span different physical domains—ranging from percentage errors to voltage distortion indices—they carry disparate units and dynamic ranges. Leaving them

unscaled would allow high-magnitude variables to dominate distance-based algorithms while masking subtle but informative variations in lower-magnitude ones. By subtracting each column's mean and dividing by its standard deviation, the procedure transforms every feature into a distribution with zero mean and unit variance, thereby eliminating dimensional effects, mitigating the influence of outliers, and creating a homogeneous numerical space. The Z-score normalization can be represented as:

$$z = \frac{x - \mu}{\sigma} \tag{1}$$

where x is the original feature, μ is the mean of x, and σ is the standard deviation of x.

3. Data Feature Analysis and Sample Labeling

In this section, PCA is first applied to the normalized feature matrix to condense its dimensionality while preserving the dominant variance that carries fault-relevant information, after which the *K*-means clustering algorithm partitions the resulting components into four interpretable groups that correspond to normal operation, mechanical faults, electrical faults and reliability faults, thereby automatically generating the labeled training set required by the subsequent supervised classifier without any manual intervention.

3.1. Principal component analysis

When analyzing raw data, the existence of a multitude of features gives rise to complex analysis procedures and exorbitant computational expenses. Therefore, PCA is first used to reduce dimensionality and extract features from the data. When the cumulative variance contribution rate ρ of m principal components approaches the set threshold (usually taken as $\rho \geqslant 85\%$) [21], these m principal components are used to replace the original high-dimensional feature data. The variance contribution rate and cumulative variance contribution rate of the principal components are shown in Fig. 2. As can be seen, the cumulative contribution of the first 56 principal components has already exceeded 85%. Therefore, this paper selects the first 56 principal components to represent the original high-dimensional data features.

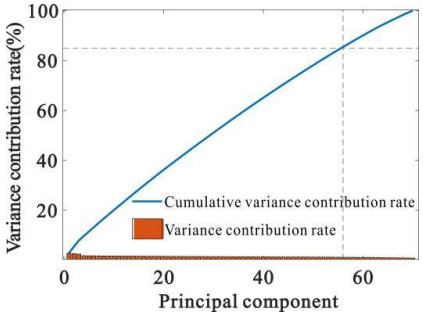


Fig. 2 Variance contributions of the principal components

3.2. K-means clustering for labeling samples

The features obtained from the previous tests are unlabeled. By means of clustering algorithms, the unlabeled data is partitioned into four categories: normal, mechanical faults, electrical faults, and reliability faults. The *K*-means clustering algorithm is a simple and widely used method of cluster analysis, designed to partition data points into k clusters based on their features so that the members of each cluster are more similar to each other than to members of other clusters.

The main process of the *K*-means clustering algorithm is as follows:

(1) Determine the number of clusters, k. Initially, randomly select k data points as the initial cluster centers, denoted as a = a_1, a_2, ..., a_k.

(2) Calculate the shortest distance from each data point x_i within the dataset to the k cluster centers and assign the data points to the nearest cluster.

In this algorithm, Euclidean distance is used as the metric for gauging the similarity between data points can be illustrated as:

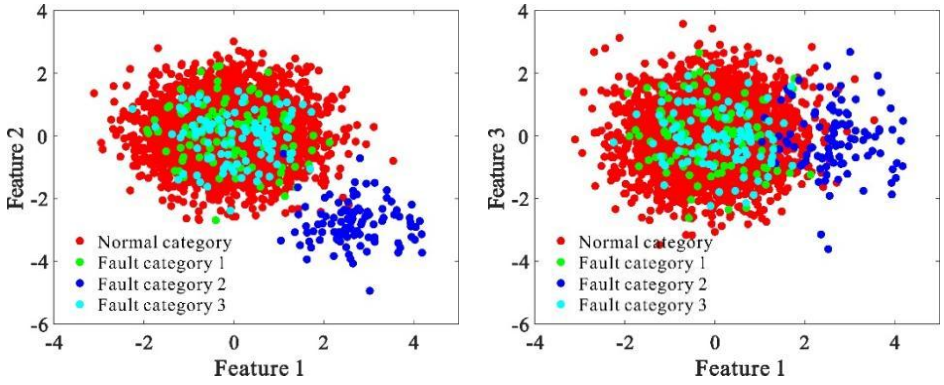
$$D_{i,k} = \sqrt{\sum_{i=1}^{n} (x_i - a_k)^2} \tag{2}$$

where D represents the shortest distance between a sample and the cluster center, x_i represents the i-th sample, and n is the total number of samples.

(3) For each cluster, the cluster center is recalculated based on the positions of all sample points within that cluster. The new cluster center is the mean position of all its member data points. For each dimension, the coordinates of the new cluster center are the average of the coordinates of all points in that dimension within the cluster. All points are then reassigned to clusters based on this updated similarity.

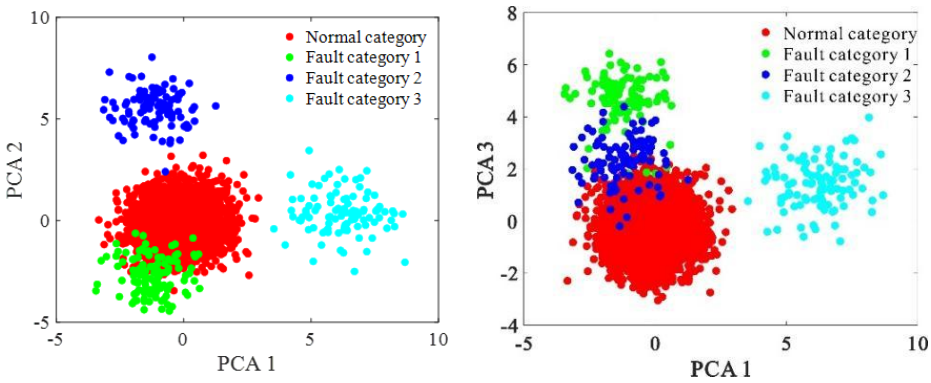
(4) Repeat the above steps (2) and (3) until either the positions of the cluster centers converge, the clustering criterion function converges, or the maximum number of iterations is reached.

The *K*-means clustering algorithm is applied to classify the original features and principal components. Figs. 3(a) and 3(b) respectively show the classification scenarios involving original Features 1 and 2, and Feature 3, while Figs. 4(a) and 4(b) show the classification scenarios involving Principal Component 1 and 2, and Principal Component 3. From the figures, it can be seen that compared to the classification with original features, the classification boundaries between the principal components are wider, indicating good classification performance.



(a) Feature 1 and Feature 2 (b) Feature 1 and Feature 3

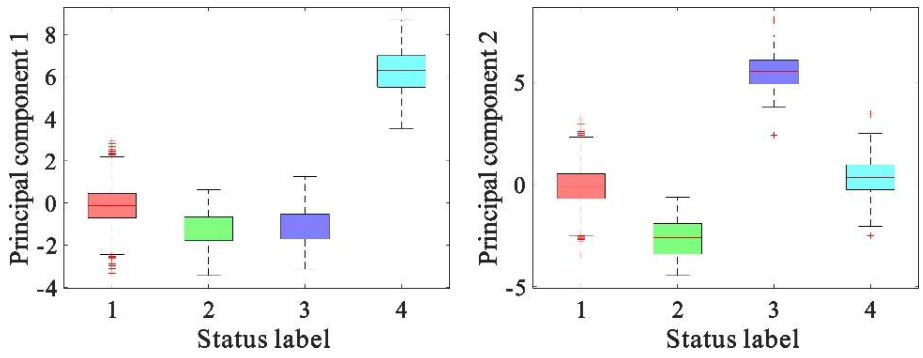
Fig. 3 Classification relationships of original feature data



(a) PCA1 and PCA2 (b) PCA1 and PCA3

Fig. 4 Classification relationships of PCA data

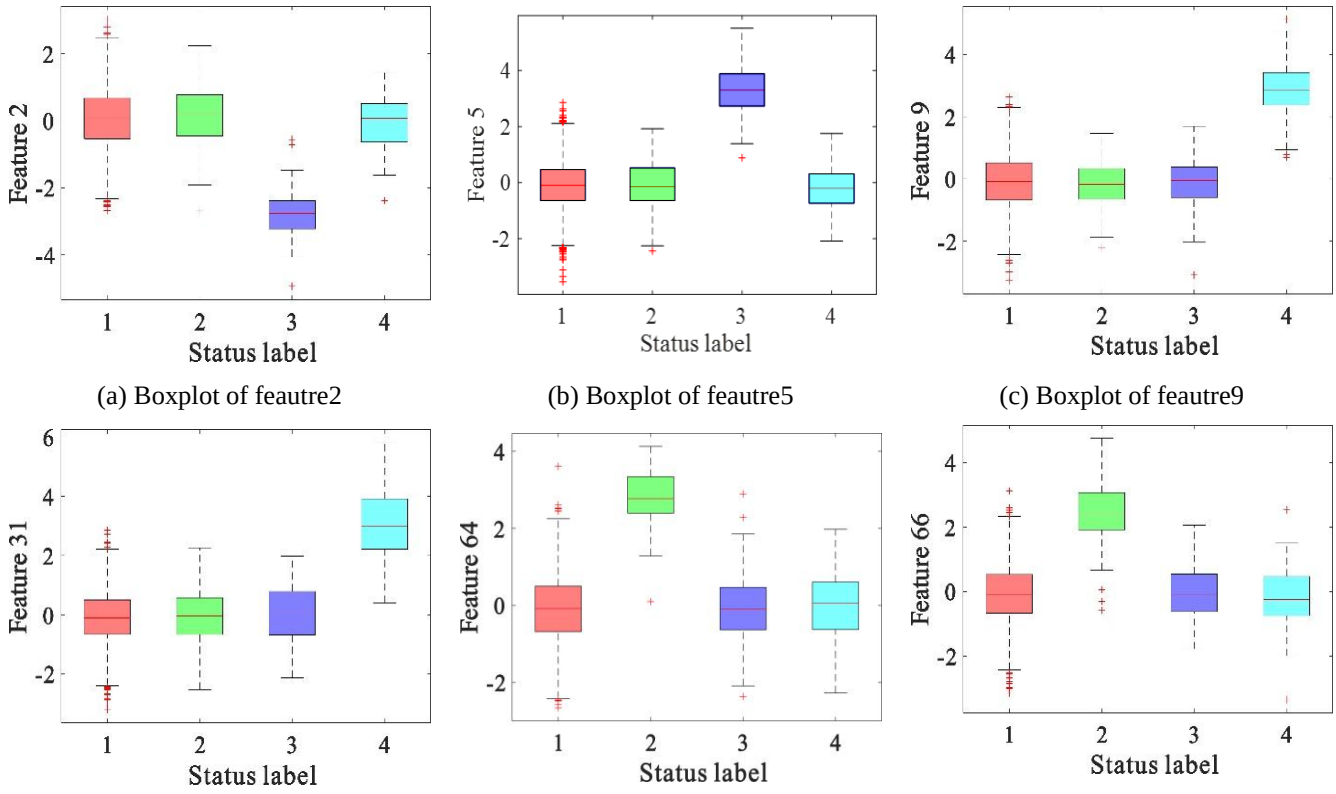
Figs. 5(a) and 5(b) respectively show the box plots for Principal Components 1 and 2, while Figs. 6(a)-(f) respectively show the box plots for original Features 2, 5, 9, 31, 64, and 66. The real-world meanings of these features are respectively the minimum basic error, median basic error, standard deviation of phase voltage symmetry, maximum distortion of phase A voltage, standard deviation of device output power stability, and maximum repeatability of the device's measurements.



(a) Boxplot of PCA1 (b) Boxplot of PCA2

Fig. 5 Boxplot of PCA data

From Figs. 5-6, it can be observed that, compared to the original features, the principal components exhibit significant differences in their median and quartile distributions across different states, demonstrating good classification results. This highlights the effectiveness of PCA in extracting significant data components.



(a) Boxplot of feautre2 (b) Boxplot of feautre5 (c) Boxplot of feautre9

(d) Boxplot of feautre31 (e) Boxplot of feature64 (f) Boxplot of feautre66

Fig. 6 Boxplot of original feature data

In Fig. (6), label 1, representing normal operation, is observed to be clustered uniformly near zero across all features, and stable performance is thereby indicated. Figs. 6(a) and 6(b) show that under status label 3, the minimum basic error values are smaller and the median basic error values are larger compared to other labels, which is typically indicative of mechanical faults such as loosened terminal crimping or corrosion aging. Consequently, label 3 corresponds to a mechanical fault at the measurement position. From Figs. 6(c) and 6(d), it is evident that under label 4, the maximum values for phase voltage

symmetry and phase A voltage distortion are higher than those under other labels, indicating poor quality of the output power, and therefore label 4 corresponds to an electrical fault at the measurement position. Figs. 6(e) and 6(f) reveal that under state label 2, the standard deviation of the device's output power stability and the maximum values for the device's measurement repeatability are higher compared to other labels, suggesting increased fluctuations in power output during operation. Therefore, label 2 corresponds to a reliability fault at the measurement position.

By leveraging the K-means algorithm, the unlabelled samples are systematically partitioned into distinct fault-related clusters, and corresponding labels are automatically assigned to each cluster. Consequently, the essential training labels required for the subsequent online fault detection model are reliably furnished without manual intervention.

4. Construction and Optimization of Online Fault Identification Model

In this section, a multi-class support vector machine is trained on the labeled data to discriminate normal, mechanical, electrical, and reliability faults, while Bayesian optimization with *k*-fold cross-validation is employed to automatically tune the kernel width and penalty factor, ensuring that the resulting model attains robust classification accuracy without succumbing to overfitting when deployed for real-time online identification.

4.1. Multiclass support vector machine

During this stage, fault identification is performed with an MSVM. This technology demonstrates highly prominent adaptability advantages when confronting small-sample datasets and other related scenarios, effectively ensuring the accuracy and high efficiency of fault identification. A One-vs-One (OvO) strategy is commonly employed for multiclassification. This OvO strategy involves training an SVM for each pair of categories, requiring the training of $k(k-1)/2$ classifiers for k categories. When predicting a new data point, the model feeds it to all classifiers, each of which outputs a category based on its prediction probability. The final classification result is determined by the voting results furnished by all of the classifiers.

The OvO strategy for MSVM proceeds as follows [22]:

1. Training Phase:

(1) For k distinct classes, the OvO strategy constructs $k(k-1)/2$ binary SVMs, one for every unordered pair (i, j) with $1 \leq i < j \leq k$;
(2) Each binary model is trained only on the subset of samples that belong to classes i or j, ignoring all other classes.

Because every binary problem is smaller than the original dataset, training each individual SVM is fast and memory-efficient.

2. Prediction Phase:

(1) A test sample is submitted to all $k(k-1)/2$ classifiers;
(2) Each classifier casts a single vote for either class i or class j;
(3) The class that accumulates the largest number of votes is declared the final prediction (majority voting);
(4) Ties can be resolved by confidence scores (decision values) or by additional binary comparisons.

The OvO strategy delivers several advantages. By confining each binary classifier to only two classes at a time, it simplifies within-class distributions so that SVM's margin-maximization property is exploited more effectively, yielding high accuracy even on small-sample data. Every sub-problem is small, so the quadratic-programming solver scales with the square of the subset size rather than the full dataset, ensuring computational efficiency. All $k(k-1)/2$ model can be trained independently, making the approach embarrassingly parallel and highly scalable. Limiting each classifier to two classes also mitigates global imbalance, providing robustness against skewed class sizes. Finally, OvO avoids the "dominant class" bias

when one class greatly outnumbers the others. Therefore, in this work, the OvO approach is adopted because it usually yields higher accuracy for small-sample, multi-class problems.

For each nonlinear SVM classification problem, an optimal classification hyperplane is sought in a high-dimensional space by a nonlinear transformation, which can be represented by $\omega^{T}\phi(x) + b = 0$. Given n samples, the distance of a sample $x_i=(x_1, x_2, x_3 \ldots x_n)$ to the hyperplane can be expressed as:

$$\frac{\left|\omega^T \phi\left(x_i\right)+b\right|}{\sqrt{\omega_1^2+\omega_2^2+\cdots+\omega_n^2}}=\frac{y_i(\omega^T \phi\left(x_i\right)+b)}{\sqrt{\omega_1^2+\omega_2^2+\cdots+\omega_n^2}}=\frac{y_i(\omega^T \phi\left(x_i\right)+b)}{\|\omega\|} \tag{3}$$

where ω is the weights, b represents the bias, and $\phi(x)$ indicates the nonlinear transformation applied to the feature vector x. This problem is transformed into an optimization problem that seeks to maximize the classification margin, which involves minimizing $\frac{2}{\|\omega\|}$. The goal is to find a pair (ω, b) that satisfies the following conditions for all samples:

$$\begin{cases}\omega^T \phi\left(x_i\right)+b \geq 1 \quad, y=1 \\ \omega^T \phi\left(x_i\right)+b \leq-1, y=-1\end{cases} \tag{4}$$

In practical scenarios, most data do not satisfy the above separation conditions. For such non-separable problems, a slack variable $\xi_i > 0$ is introduced for each sample (x_i, y_i), where $i = 1, 2, 3, \ldots, n$. Additionally, a penalty factor C is included to prevent the slack variables ξ_i from becoming too large. Consequently, the learning problem for the non-linear non-separable SVM is transformed into a convex quadratic programming problem can be presented as:

$$\begin{cases}\min\limits_{\omega, b} & \frac{1}{2}\|\omega\|^2+C \sum\limits_{i=1}^n \xi_i \\ s.t. & y_i(\omega^T \phi\left(x_i\right)+b)-1+\xi_i \geq 0, i=1,2, \cdots, n\end{cases} \tag{5}$$

By applying the Karush-Kuhn-Tucker (KKT) conditions, this can be converted into the dual optimization problem, which can be calculated by:

$$\begin{cases}\max & Q(\alpha)=\sum\limits_{i=1}^n \alpha_i-\frac{1}{2} \sum\limits_{i=1}^n \sum\limits_{j=1}^n \alpha_i \alpha_j y_i y_j \phi(x_i)^T \phi(x_j) \\ s.t. & \sum\limits_{i=1}^n \alpha_i y_i=0, 0 \leq a_i<C, i=1,2, \cdots, n\end{cases} \tag{6}$$

where α represents the Lagrange multipliers.

This dual problem is solved using the Sequential Minimal Optimization (SMO) method, yielding the following equation:

$$\begin{cases}\omega=\sum\limits_{i=1}^q \alpha_i y_i \phi(x_i) \\ b=\frac{1}{q}\left\{\sum\limits_{i=1}^q y_i-\sum\limits_{i, j=0}^q \alpha_i y_i \phi(x_i)^T \phi(x_j)\right\}\end{cases} \tag{7}$$

where q represents the number of support vectors, corresponding to α greater than zero.

Introducing the kernel function $K(x_i \cdot x_j)$, which represents the inner product $\phi(x_i)^T \phi(y_i)$ in the high-dimensional feature space. Then, the classified categories can be written as:

$$\begin{aligned} f(x) &= \omega^T \cdot \phi(x) + b \\ &= \sum_{i=1}^{q} \alpha_i y_i \phi^T(x_i) \phi(x) + b \\ &= \sum_{i=1}^{q} \alpha_i y_i K(x_i, x) + b \end{aligned} \tag{8}$$

Among the kernel functions, the Radial Basis Function (RBF) is the most widely used. Its general form can be represented as follows:

$$K(x_i \cdot x_j) = \exp(-g \left\| x_i - x_j \right\|^2), g > 0 \tag{9}$$

where g represents the width of the RBF.

For per binary SVM classifier, each classifier is trained on samples from two classes. Training a standard SVM involves solving a QP problem to find the optimal hyperplane. The complexity of this process is theoretically $O(l^3)$, where l is the number of support vectors. This study uses the SMO algorithm, which decomposes the large QP problem into smaller subproblems. It reduces the computational complexity from $O(l^3)$ to $O(l^2)$ per binary classifier. The $O(l^2)$ complexity makes the MSVM feasible for practical deployment, even with the 6 binary classifiers required by the OvO strategy. This optimization ensures efficient training without sacrificing accuracy.

4.2. Bayesian optimization with K-fold cross-validation

During this stage, fault identification is performed with an MSVM. This technology demonstrates highly prominent adaptability advantages when confronting small-sample datasets and other related scenarios, effectively ensuring the accuracy and high efficiency of fault identification. An OvO strategy is commonly employed for multiclassification. This OvO strategy involves training an SVM for each.

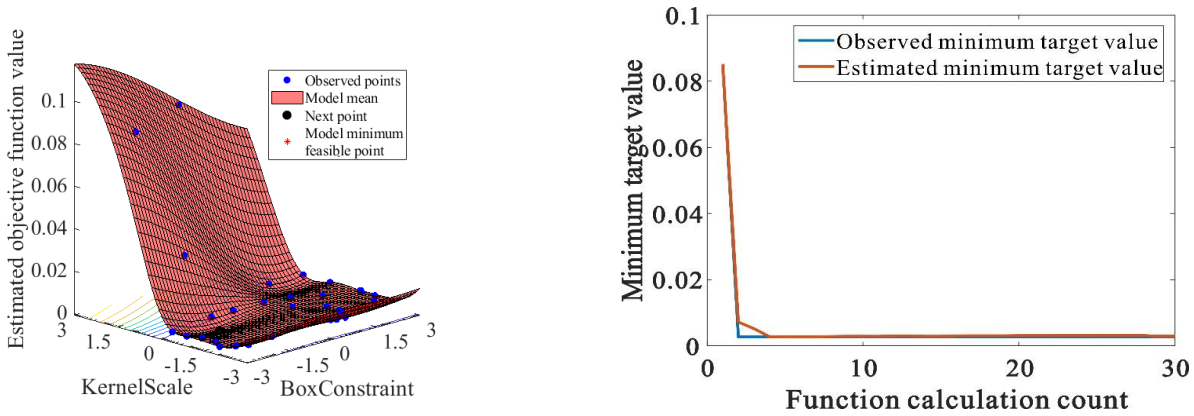
In the non-linearly separable SVM model mentioned above, C and g are hyperparameters that need to be optimized, as these two parameters significantly impact the model's accuracy. A large radial basis width g can enhance the model's classification performance but may lead to overfitting, compromising model robustness. Conversely, too small a g causes excessive smoothing, leading to low classification accuracy. The penalty factor C indicates the tolerance for errors. A higher C means less tolerance, making the classification criteria stricter and potentially leading to overfitting. Therefore, both excessively high and low combinations of C and g can result in poor anomaly detection performance.

A diverse array of sophisticated techniques exists for systematically refining hyperparameters to achieve optimal model performance. Common optimization algorithms include grid search, genetic algorithms, and particle swarm optimization. Besides being time-consuming, these algorithms do not consider previous parameter information when iterating to the next value of the discrete parameter, making them prone to local optima in non-convex problems [23]. Bayesian optimization, however, focuses on reducing evaluation costs. It requires fewer iterations and exhibits greater speed. Moreover, considering prior parameter information, it is less likely to fall into local optima when dealing with non-convex issues. This research selects k-fold cross-validation Bayesian optimization to fine-tune the parameters of the SVM model. The basic idea of cross-validation is to split the training data into k smaller subsets. During each iteration, one subset is used as the test set, and the remaining k-1 subsets serve as the training set. K-fold cross-validation effectively selects the optimal kernel parameters and penalty coefficients for the SVM while preventing overfitting.

The steps of the Bayesian optimization algorithm are as follows:

1. Randomly select sampling points within the predefined search range for C and g. Employ the k-fold cross-validation loss value as the objective function, with different parameter combinations of the model as the independent variables, thereby constructing a surrogate model. This provides the initial distribution of the objective function and the set of sampling points.
2. Choose the next sampling point x_t by maximizing the acquisition function and acquiring the function value $f(x_t)$.
3. Add the new sampling point $[x_t, f(x_t)]$ to the set of sampling points and update the Gaussian process surrogate model to better fit the distribution of the objective function.
4. Set a maximum number of iterations. Once this predetermined number is reached, stop the algorithm iteration. Subsequently, output the optimal sampling point along with the corresponding optimal solutions for C and g.

Fig. 7 shows the result of cross-validation-based Bayesian optimization, with optimized hyperparameters C and g being 0.2805 and 0.0246, respectively.



(a) Hyperparameters distribution (b) Iteration of the target value

Fig. 7 Selection results based on the Bayesian optimization algorithm

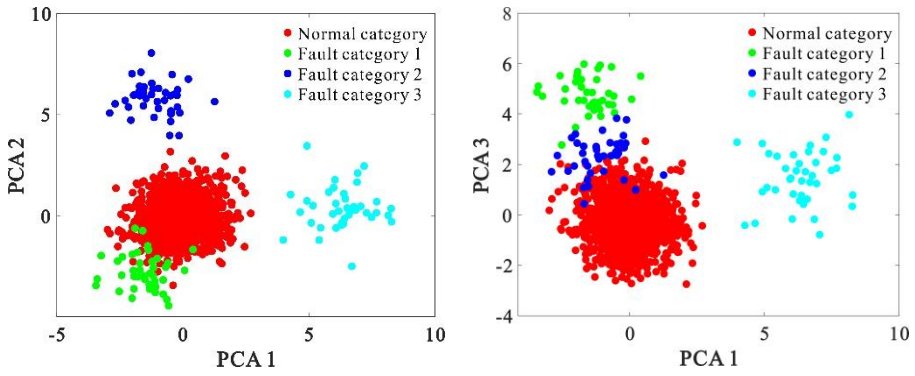
5. Result Analysis and Samples Enhancement

After training, the MSVM is validated on the original data via its confusion matrix, which exposes misclassifications attributable to limited samples. To address this, the Monte Carlo method is employed to expand the dataset by generating normally distributed synthetic samples that faithfully preserve the original statistical properties. These augmented data are then merged with the authentic samples to retrain the model, markedly lowering the error rate and markedly enhancing the reliability of online fault identification.

5.1. Original data validation

In this study, 432 out of 720 sample data points (60%) were selected as model training data, and 288 (40%) were used for validation. Figs. 8(a) and 8(b) show the classification of principal components 1 to 3 based on the predictions made by the trained MSVM. The figures show that the classification performance on the test data is good, with clear boundaries between categories and few misclassifications.

Fig. 9 shows the confusion matrix of the MSVM identification results based on 288 validation samples. The identification results include 18 misclassifications, resulting in an error rate of 6.25%. It was found that the majority arose from positions subjected to pronounced electromagnetic interference or subtle drifts in crimp impedance that lay close to the decision boundaries.



(a) PCA2 and PCA1 (b) PCA3 and PCA1

Fig. 8 Classification relationships among PCA

It can be observed that despite the favorable results of the *K*-means clustering with clear boundaries between the principal components, there are some recognition errors due to insufficient model training data. Therefore, it is considered to enhance the original data. The Monte Carlo method has high applicability, preserves the original features of the data well, and is suitable for processing complex data.

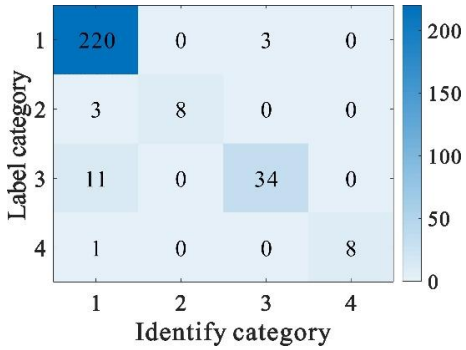


Fig. 9 Confusion matrix of MSVM identification results for original data

5.2. Enhancing samples based on Monte Carlo

To address the challenge of limited sample size and improve the accuracy of our fault identification model, the Monte Carlo method is applied for data augmentation. This method leverages the statistical properties of the existing data to generate additional synthetic samples, thereby increasing the diversity and volume of the training dataset. By doing so, the model is enhanced to generalize and improve its accuracy in identifying faults. The steps of the Monte Carlo data augmentation are outlined below and illustrated in Fig. 10

Steps of Monte Carlo data augmentation are as follows:

1. Merge the data from 10 batches corresponding to each meter position into a matrix possessing 10 rows and 70 columns, thereby constructing a dataset for 72-meter positions.
2. Calculate the mean and standard deviation for each column of data for each meter position.
3. Based on the mean and standard deviation from step 2, generate 50 rows of data adhering to a normal distribution for every meter position, thereby resulting in a dataset of 50 rows and 70 columns for each respective position.
4. Combine the data from all 72-meter positions to obtain a final feature dataset comprising 3600 rows and 70 columns.

To verify the effectiveness of the Monte Carlo-based data augmentation method on model training and identification accuracy, Fig. 11 shows the confusion matrix of the identification results obtained from the MSVM trained on augmented data. After data enhancement, the identification results encompassed merely 5 misclassifications. This corresponds to an error rate of 0.35%, signifying a remarkable 93.3% reduction in the error rate. Therefore, the data augmentation enhances model training effectiveness and refines the accuracy of fault identification.

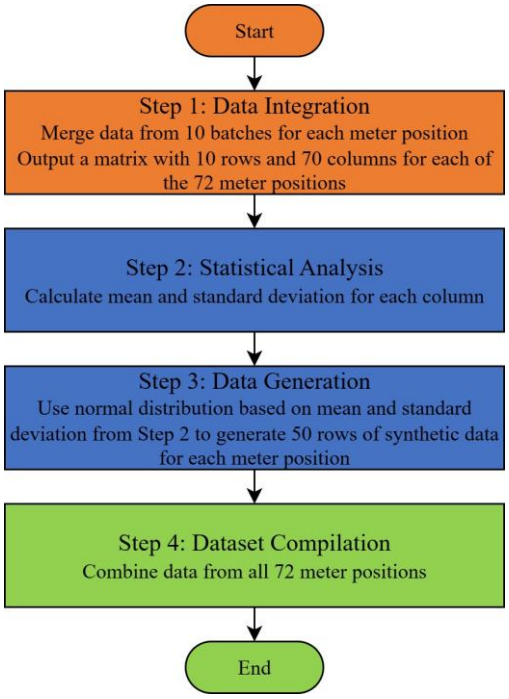


Fig. 10 Flow diagram of data augmentation

Further, in Fig. 11, 123 faults were detected, including 61 mechanical faults (26 from meter position 31 and 35 from meter position 34), 49 electrical faults (22 from meter position 11 and 27 from meter position 48), and 13 reliability faults (8 from meter position 69 and 5 from meter position 54)". Of the 5 misclassification errors, 4 originated from meter position 69, and 1 originated from meter position 34. Upon manual inspection, it was found that meter position 31 had corroded crimp terminals, meter position 34 had deformed port crimps, meter position 11 had loose terminal crimps, meter position 48 had poor connector contacts, and meter positions 69 and 54 had significant electromagnetic noise interference.

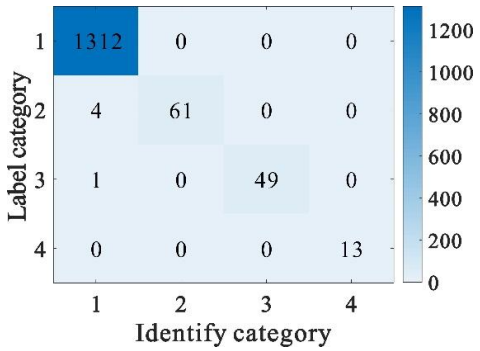


Fig. 11 Confusion matrix of MSVM identification results after sample enhancement.

The comparative analysis of the fault identification methods is presented in Table 2. Method in [8] shows good performance in accuracy, precision, and recall rates, with 95.15% accuracy, 94.80% precision, and 95.92% recall, and it includes anomaly identification capability but lacks real-time processing capability. Method in [11], while achieving 100% accuracy, precision, and recall, but does not offer fault identification capabilities. This method proposed in this study exhibits the highest comprehensive performance across accuracy, precision, and recall rates, achieving 99.65% accuracy, 99.91% precision, and 97.96% recall, while also possessing the capabilities for anomaly detection and online processing.

Finally, the training and testing durations are presented as follows. The 4-class MSVM model was trained and tested on an Intel Core i7 laptop with 32GB RAM. The dataset was split such that 60% was used for model training, with the remaining 40% reserved for testing. Specifically, the training durations for the 720 pre-enhancement samples and 3,600 post-enhancement samples were approximately 2 minutes and 11 minutes, respectively, while their corresponding testing times were less than 1

second and nearly 3 seconds. The execution time (i.e., inference/prediction time) of the trained MSVM model on the same laptop is extremely fast, as it involves only forward calculations using precomputed support vectors, with a duration on the order of 100 microseconds per sample, making it suitable for online fault detection for AVDs.

Table 2 Comparative analysis of fault identification methods

Methodology	Accuracy	(Micro-) Precision	(Micro-) Recall	Anomaly Identification	Fault Identity	Online Capability
[8]	95.15%	94.80%	95.92%	√	×	Online, but time-consuming
[11]	100%	100%	100%	√	×	√
A	93.75%	96.38%	83.96%	√	√	√
B	99.65%	99.91%	97.96%	√	√	√

Note: A represents the method used in the study without sample augmentation; B indicates the method used in this study.

6. Conclusion

This research has successfully developed an online fault identification method for AVDs that eliminates the need for manual sample labeling and achieves high identification accuracy. The primary conclusions are as follows:

(1) The method integrates real-time data collection and sample labeling processes. By verifying ISEMs concurrently with the tested electric meters, it efficiently gathers relevant data. The *k*-means clustering algorithm is then used to categorize samples, significantly enhancing the overall efficiency of online fault identification for AVDs.

(2) Fault identification accuracy is significantly improved with limited samples. The Monte Carlo method enhances data representation, enabling the model to learn more effectively and reducing the error rate to 0.35%—a 94.40% improvement over pre-enhancement levels.

This research provides a reliable and efficient solution for online fault identification in AVDs. Future research may focus on advanced data augmentation techniques to enhance model robustness, real-time monitoring, integration with IoT, and big data for scalability. Addressing these areas could significantly expand the impact of online fault detection and drive the development of smarter, more efficient monitoring systems.

Acknowledgments

This work was supported by the science and technology project of China Southern Power Grid Co., Ltd. under Grant YNKJXM20230134.

Conflicts of Interest

The authors declare no conflict of interest.

References

[1] J. Yang, H. Li, H. Chen, Z. Zhu, and C. Zhang, “Data-Driven Evaluation for Error States of Standard Electricity Meters on Automatic Verification Assembly Line,” IEEE Transactions on Industrial Informatics, vol. 15, no. 9, pp. 4999-5010, 2019.

[2] H. Carstens, X. Xia, and S. Yadavalli, “Low-Cost Energy Meter Calibration Method for Measurement and Verification,” Applied Energy, vol. 188, pp. 563-575, 2017.

[3] A. Ferrero, M. Faifer, and S. Salicone, “On Testing the Electronic Revenue Energy Meters,” IEEE Transactions on Instrumentation and Measurement, vol. 58, no. 9, pp. 3042-3049, 2009.

[4] J. S. Choi, S. Lee, and S. J. Chun, “A Queueing Network Analysis of a Hierarchical Communication Architecture for

Advanced Metering Infrastructure," IEEE Transactions on Smart Grid, vol. 12, no. 5, pp. 4318-4326, 2021.
[5] L. B.Wang, H. Wang, and C. Zhang, "Research on the Optimal Maintenance Frequency of Automatic Verification Pipeline Equipment in Electric Energy Meter, Electric Energy Meter Automatic Verification Pipeline Equipment," Electrical Measurement & Instrumentation, vol. 54, no. 08, pp. 89-92, 2017.
[6] L. Zhang, Z. Fu, L. Li, and J. Ding, "Research on Intelligent Monitoring of the Measurement Property in Electric Energy Meter Automatic Verification System," Instrument Standardization & Metrology, no. 6, pp. 38-40 and 46, 2016.
[7] Q.Chen, et al., "Research on Automated Verification Method for Temperature-Related Tests of Electrical Energy Meters," Journal of Physics: Conference Series, vol. 2853, article no. 012020, 2024.
[8] Z.Qu, et al., "An Online Accuracy Evaluation Method for Smart Electricity Meter Calibration Devices Based on Mutual Correlations of Calibration Data," IEEE Transactions on Instrumentation and Measurement, vol. 73, pp. 1-13, 2024.
[9] L. Wang, B. Gao, P. Zhao, C. Zhang, and A. Sida, "Design of Meter Position Fault Location and Alarm System for Automatic Verification Pipeline of Electric Energy Meter," Hebei Electric Power, no. 1, pp. 33-34 and 48, 2018.
[10] T. Śmiałkowski and A. Czyżewski, "Detection of Anomalies in the Operation of a Road Lighting System Based on Data from Smart Electricity Meters," Energies, vol. 15, no. 24, article no. 9438, 2022.
[11] Y. Jiao, Q. Chen, Z. Bao,L. Pan, and H. Li, "An On-line Anomaly Identifying Method for Calibration Devices in an Automatic Verification System for Electricity Smart Meters," Measurement, vol. 180, article no. 109606, 2021
[12] Y. Junfeng, "Research on Intelligent Operation and Maintenance Platform Based on Automatic Verification System of Electric Energy Meter," Industrial Instrumentation & Automation, no. 6, pp. 104-107 and 116, 2019.
[13] Y. Dongsheng, L. Guanna, D. Hengchun, Y. Ruiming, and L. Yanguo, "An Expert System for Fuzzy Fault Diagnosis of Auto-verification Line for Electric Energy Meters," Electrical Measurement & Instrumentation, no. 7, pp. 94-96 and 102, 2017.
[14] N. N. Bon and L. V. Dai, "Fault Identification, Classification, and Location on Transmission Lines Using Combined Machine Learning Methods," International Journal of Engineering and Technology Innovation, vol. 12, no. 2, pp. 91-109, 2022.
[15] W. Liu, et al., "Power Grid Vulnerability Identification Methods Based on Random Matrix Theory and Entropy Theory," Proceedings of the CSEE, no. 20, article no. 5893, 2017.
[16] G. Lu, et al., "Research on Meter Position Detection Accuracy for Automatic Verification Pipeline of Intelligent Watt-hour Meter," Electrical Measurement & Instrumentation, vol. 60, no. 10, pp. 148-154, 2023.
[17] L. Zhang, X. Jia, D. Xiao, H. Luo, and Z. Xu, "Machinery Fault Diagnosis with Imbalanced Data using Deep Generative Adversarial Networks," Measurement, vol. 152, article no. 107377, 2020.
[18] W.Wan, S.Chen, J.Chen, A.Li and Y.Feng, "QSCGAN: An Un-Supervised Quick Self-Attention Convolutional GAN for LRE Bearing Fault Diagnosis Under Limited Label-Lacked Data," IEEE Transactions on Instrumentation and Measurement, vol. 70, pp. 1-16, 2021.
[19] Y. Hang and L. Chen, "Rolling Bearing Fault Diagnosis under Fluctuant Conditions Based on Compressed Sensing," Structural Control and Health Monitoring, vol. 24, no. 5, article no. e1918, 2016.
[20] G. Kiar, Y. Chatelain, A. Salari, A. C. Evans, and T. Glatard, "Data Augmentation Through Monte Carlo Arithmetic Leads to More Generalizable Classification in Connectomics," Neurons, Behavior, Data Analysis, and Theory 1, vol. 1, pp. 1-20, 2021.
[21] Y. Liu, J. Wang, P. Zhao, D. Qin, and Z. Chen, "Research on Classification and Recognition of Driving Styles Based on Feature Engineering," IEEE Access, vol. 7, pp. 89245-89255, 2019.
[22] H. Guo, W. Wang, "An Active Learning-Based SVM Multi-Class Classification Model," Pattern Recognition, vol. 48, no.5, pp. 1577-1597, 2015.
[23] G. N. Kouziokas, "SVM Kernel Based on Particle Swarm Optimized Vector and Bayesian Optimized SVM in Atmospheric Particulate Matter Forecasting," Applied Soft Computing, vol. 93, article no. 106410, 2020.