IEEE Access
Multidisciplinary ⋮ Rapid Review ⋮ Open Access Journal

Received 5 January 2024, accepted 19 January 2024, date of publication 25 January 2024, date of current version 1 February 2024.

Digital Object Identifier 10.1109/ACCESS.2024.3358683

APPLIED RESEARCH

Electricity Theft Detection in Smart Grids Based on Omni-Scale CNN and AutoXGB

SANYUAN ZHU[1,2], ZIWEI XUE[1,2], AND YOUFENG LI[1]
[1]School of Computer and Information Science, Hubei Engineering University, Xiaogan 432000, China
[2]School of Computer Science and Information Engineering, Hubei University, Wuhan, Hubei 430062, China

Corresponding authors: Ziwei Xue (202121116013177@stu.hubu.edu.cn) and Youfeng Li (feng.li@hbeu.edu.cn)

This work was supported in part by the Hubei Province Science and Technology Development Special Project under Grant 2022BGE258, in part by the Science and Technology Research Projects of Hubei Provincial Department of Education under Grant Q20162706, and in part by the Xiaogan Municipal Natural Science Foundation Projects under Grant XGKJ2020010037.

ABSTRACT Electricity theft is a prevalent global issue that has detrimental effects on both utility providers and electricity consumers. This phenomenon undermines the economic stability of utility companies, worsens power hazards, and influences electricity costs for consumers. The advancements in Smart Grid technology play an essential role in Electricity Theft Detection (ETD), as they generate large amounts of data that can be effectively utilized for ETD through the application of Machine Learning (ML) and Deep Learning (DL) methodologies. The present study presents a novel approach for ETD by combining Omni-Scale CNN (OS-CNN) and AutoXGB. Firstly, the Piecewise Cubic Hermite Interpolating Polynomial (PCHIP) is employed as the data interpolation technique to address the limitations and missing data in the dataset. Additionally, a combination of the Synthetic Minority Over-Sampling Technique (SMOTE) and the Edited Nearest Neighbors (ENN), known as SMOTEENN, is utilized for data resampling to tackle the issue of class imbalance in the dataset. Secondly, the multi-layer Omni-Scale block stack is employed to effectively cover the receptive fields of diverse time series scales based on a straightforward rule. This enables the One-dimensional Convolutional Neural Network (1D-CNN) to acquire enhanced learning capabilities for both irregular electricity consumption data anomalies and periodic normal electricity consumption patterns in smart grid datasets, facilitating superior extraction of essential data features. The AutoXGB classifier is then utilized to classify the extracted features. AutoXGB possesses the capability of automatically optimizing the hyperparameters required by the model, ensuring that the classification model maintains optimal accuracy and settings. Finally, the method exhibits superior competitiveness compared to other methods on the same dataset. The experimental results demonstrate that the proposed model achieves an accuracy rate of 99.2%, a precision rate of 97.5%, and an area under the ROC curve of 98.4%. These results establish its significant superiority over alternative models.

INDEX TERMS Electricity theft detection, SMOTEENN, omni-scale CNN, AutoXGB, smart grid.

I. INTRODUCTION

The utilization of electricity is pervasive in everyday life and is continuously consumed worldwide. Meanwhile, there are certain degrees of loss during the transmission and conversion processes of power. In general, power losses can be categorized into two types: Technical Losses (TLs) and Non-Technical Losses (NTLs) [1].

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

The occurrence of TLs is an inevitable consequence of power transmission, primarily arising from the Joule effect in power lines and inefficiencies in transformers. Due to the inherent characteristics of TLs, the calculation of TLs is quite complex, and TLs loss cannot be completely eliminated, only some existing techniques can be used to reduce TLs [2].

The primary causes of NTLs arise from delays and violations in billing processes, instances of energy theft, meter malfunctions, fraudulent activities, and outstanding payments [3]. In recent years, a small proportion of users have

been engaging in meter data tampering as a means to reduce electricity consumption and illicitly acquire electricity, thereby constituting one of the primary factors contributing to NTLs [4]. The act of power theft can also have detrimental effects on the stability of the power grid, potentially disrupting accurate calculation of regional electricity load and impeding proper upgrading and corresponding power supply facilities. During periods of peak electricity consumption, there is a high likelihood that the regional power grid will experience paralysis due to excessive load, resulting in direct or indirect economic losses [5]. According to the survey, various countries have experienced economic losses due to NTLs. For instance, electricity theft in the United States results in an annual economic loss exceeding $6 billion, while the United Kingdom incurs approximately $234 million and China around $560 million [6].

In the present era, the advent of Advanced Meter Infrastructure (AMI) has presented novel challenges and methodologies for detecting theft. The AMI comprises intelligent meters, sensors, computing devices, and advanced communication technologies to facilitate bidirectional communication between electricity generation and consumption points. Additionally, the AMI is responsible for gathering data on electricity consumption, real-time electricity prices, and grid conditions. The AMI encompasses intelligent meters, sensors, computing devices, and advanced communication technologies to facilitate bidirectional communication between electricity generation and consumption points. Moreover, the AMI is responsible for collecting data on electricity consumption, real-time electricity prices, and grid conditions. Although the smart meter is equipped with a tamper-proof detection function, its communication capability renders it susceptible to increased instances of electricity theft attacks, leading to meter damage and subsequent financial losses [7]. Due to the aforementioned factors, ETD has emerged as a crucial concern in the current era of AMI.

In this context, numerous researchers have proposed theft detection technologies based on various perspectives and methodologies to address the issue of NTLs. The following are descriptions of some techniques:

1) Hardware-based: The objective of hardware-based ETD is to develop diverse hardware components that incorporate sensors with distinct functionalities into smart meters, enabling the identification of meter status and prevention of unauthorized serial modifications. However, this approach incurs significant manpower and material costs, as well as long-term maintenance and upgrade expenses [8].
2) Game Theory: In the ETD based on game theory, the ETD problem is formulated as a game between the power company and the thief to achieve an equilibrium state. ETD identifies different distributions of expected billing energy consumption; However, finding the appropriate equilibrium function requires significant computation [9].
3) Data driven: The emergence of data-driven ETD as a novel technology in recent years can be attributed to advancements in big data and ML. With the increasing adoption of smart meters, the smart grid continuously collects vast power data, alongside a plethora of meteorological, economic, and other related data. Currently, researchers have proposed various data-driven techniques for performing ETD. These techniques include ML, meta-learning, ensemble learning, and DL. However, the existing data-driven technology also faces the following challenges: (i) In ETD, there is a significant class imbalance where the proportion of electricity theft users is minuscule compared to the overall user population. The issue of imbalanced data can significantly contribute to overfitting problems and hinder the generalization performance of the electricity theft model. The model's decisions tend to exhibit bias towards the majority class, rendering it incapable of detecting instances of electricity theft. (ii) The power data collected by smart meters spans a substantial amount of time and has extensive dimensions. The presence of high-dimensional data can potentially lead to dimensionality disaster in ETD which compromises the accuracy of models. (iii) A range of non-malicious factors, such as sensor failures in smart meters and fluctuations in communication networks, can lead to abnormal power data records. These factors have also contributed to the model's poor performance, leading to misclassification of abnormal consumers as normal ones.

Addressing certain issues identified in the aforementioned papers, this study presents a novel and efficacious model for ETD. The primary contributions of this study are as follows:

1) Basedased on the literature [10], the present study proposes an OS-CNN that employs multi-layer 1D convolution and utilizes diverse convolution kernel sizes to ensure overlapping receptive fields, thereby encompassing various temporal scales of the input time series data for effective feature extraction.
2) The PCHIP method is employed to impute a substantial amount of missing data in the original dataset, while preserving its inherent distribution.
3) The SMOTE oversampling technique and ENN undersampling technique are integrated to address the issue of imbalanced trainset, thereby enhancing the model's robustness.
4) The OS-CNN network is employed for feature extraction, followed by the utilization of the automatic hyperparameter optimization framework AutoXGB as the classifier.
5) Finally, the proposed model in this paper demonstrates consistent and effective detection performance even when tested on datasets that conform to the original data distribution.

The subsequent sections of this paper are organized as follows. Section II describes the related work conducted in

the literature to address the issue of electricity theft. The techniques employed in this paper are concisely outlined in Section III. In Section IV, the results are presented and discussed. Finally, the paper concludes in Section V.

II. RELATED WORKS

This section provides an overview of existing research and techniques for ETD, primarily focusing on data-driven approaches to address NTLs. In [11], the authors used the Deep Neural Network (DNN) method to ETD using time domain and frequency domain features, and solved the problem of missing data and class imbalance through data interpolation and synthetic data, but the evaluation index of the model was insufficient. In [12], the authors proposed an Inter-week and intra-week convolutional block (IIWCBlock), which employs convolutional layers with varying dilation rates to capture inter-week and intra-week data, while extracting features through multiple sets of convolutional integration to generate a first-order representation. The Self-Dependency Model (SDM) was concurrently employed to acquire the second-order representation from the autocorrelation matrix, subsequently integrating it with the first-order representation for predicting abnormal scores of electricity users. However, the authors fail to address the issue of imbalanced data classes.

In addition, in [13], the authors employed AlexNet to address the issue of high dimensionality, while utilizing Adaptive Boosting (AdaBoost) for classifying electricity stealing users and ordinary users. Furthermore, the undersampling technique is employed to tackle the problem of class imbalance, resulting in favorable experimental performance. However, relying solely on undersampling technology will result in the loss of a substantial number of samples from the normal class, thereby diminishing the model's performance and robustness. In [14], the present study introduces a Bagging Chi-square Automatic Interaction Detection (CHAID) Decision Tree (DT) algorithm for consumer classification and detection, which exhibits superior accuracy compared to conventional detection methods. However, in the absence of balanced dataset classes, the model's performance on samples from the minority class may be suboptimal. In [15], the authors employed a diverse range of ML techniques to train and optimize the dataset through hyperparameter tuning, aiming to identify the most effective model for consumer type detection. However, the authors did not address class imbalance in the dataset and employing a multiple ML techniques may result in diminished training efficacy and substantial time consumption. In [16], the authors proposed a hybrid model, CNN-XGB, which combines Convolutional Neural Network (CNN) and Extreme Gradient Boosting (XGBoost). This model utilizes both the original One-dimensional (1D) power data and the processed Two-dimensional (2D) power data inputs. The proposed model achieves an accuracy of 92%. The authors, however, neglect to tackle the problem of imbalanced data classes.

Moreover, in [17], the authors proposed a novel method for ETD based on the Wide and Deep CNN (Wide&Deep CNN) model. The width CNN component captures the global characteristics of 1D user data, while the depth CNN component accurately identifies non-periodic stealing data and periodic normal electricity data from 2D electricity data. The performance of the proposed model was evaluated using area under the ROC curve (AUC) and Mean Average Precision (MAP). The issue of imbalanced data classes distribution, however, remains unaddressed. In [18], the authors proposed a hybrid model comprising of a Multilayer Perceptron (MLP) and a Long Short Term Memory Network (LSTM), wherein the LSTM is employed for processing daily power consumption data while the MLP is utilized for handling non-sequential data other than power data. The experimental results demonstrate the superiority of the hybrid model over the baseline model. However, the authors have overlooked the issue of imbalance data classes, which compromises the model's generalization capability. In [19], the authors utilized ensemble learning models, including XGBoost, Random Forests (RF), AdaBoost, Light Gradient Boosting (LGB), Extra Trees, and Categorical Boosting (CatBoost) for ETD purposes. Data preprocessing techniques were applied to enhance the detection performance of the models. Additionally, SMOTE was employed to address class imbalance issues. However, the training and testing process incurs a significant computational cost. Furthermore, SMOTE alone fails to capture the probability distribution curve inherent in complex power data. As a result, this leads to class overlap issues in the synthesized data and ultimately diminishes the generalization performance of the classifier.

Besides, in [5], the authors proposed a novel ETD model by combining CNN and LSTM, where CNN is employed for automated feature extraction, while LSTM is utilized for feature classification. Additionally, this study implemented a novel data preprocessing algorithm to estimate missing values in the dataset based on local values. To address the issue of data imbalance, the oversampling technique SMOTE was also employed, resulting in favorable outcomes when applied to the power data from Multan Electric Power Company (MEPCO). However, employing SMOTE alone for synthetic data generation gives rise to the issue of class overlap and may lead to model overfitting phenomenon. In [20], the authors proposed a hybrid model, namely CNN-GRU-PSO, for ETD by integrating CNN, Gated Recurrent Unit (GRU), and Particle Swarm Optimization (PSO). The CNN is employed to automate feature extraction, the GRU is utilized to classify the extracted features, and the PSO algorithm is applied to optimize the hyperparameters. SMOTE is employed to address the issue of imbalanced data; however, generating synthetic data through SMOTE may result in class overlap and subsequently lead to overfitting of the model. In [21], the authors proposed a classification framework that combines the techniques of Visual Geometry Group (VGG-16) and Firefly Algorithm-based Extreme Gradient

Boosting (FA-XGBoost). The VGG-16 model is used for data processing to identify anomalous power consumption patterns, followed by the utilization of FA-XGBoost for data classification. To address the issue of data class imbalance, the authors employed Adaptive Synthetic Sampling (Adasyn) oversampling technique for minority classes as a means to mitigate this problem. However, utilizing Adasyn in isolation entails significant computational overhead and may not yield desirable outcomes when dealing with high-dimensional datasets, potentially impacting the model's performance.

And, in [22], the authors employed a hybrid approach that combines a stacked autoencoder bagged ensemble RF for ETD, while utilizing the stacked autoencoder to extract salient features in order to enhance the classifier's performance in detecting theft incidents. The proposed model's performance is assessed using both the Irish power dataset and the Chinese power dataset. Additionally, to address the issue of imbalanced data, Random Under-Sampling (RUS) was used by the authors to balance the class distribution. However, relying solely on RUS would result in a significant loss of data samples, leading to a small sample size for model training and potentially causing underfitting issues. In [23], the authors employed feature engineering techniques to reduce data dimensionality and utilized advanced ensemble technology CatBoost for the purpose of ETD. In addition, the K-Nearest Neighbor (KNN) interpolation method was employed for missing value imputation, while the SMOTE-Tomek technique was utilized to address data imbalance issues. The hyperparameter optimization of the model, however, has been overlooked, leading to the classifier being trapped in a local optimum and consequently compromising the performance of the model. In [24], the authors proposed an ensemble DL detector comprising multiple DL-GRU. The outputs of these diverse DL models are subsequently fed into a majority voting classifier to determine the final classification outcome. However, the model overlooks the issue of data class balance. In [25], the authors proposed a semi-supervised DL model that leverages a substantial amount of high-dimensional unlabeled data and incorporates adversarial modules to mitigate the risk of overfitting. Experimental results demonstrate that the proposed model exhibits remarkable performance even when trained on limited samples. However, the absence of hyperparameter optimization may lead to the model converging towards a suboptimal solution.

III. THE PROPOSED SYSTEM MODEL

The proposed model primarily consists of three main units and several sub-units. The main units are (1) missing values handling unit (2) outlier handling unit (3) data normalization unit (4) data class balancing unit (5) proposed electricity theft detection model unit. The subsequent sections provide comprehensive coverage of the units and their associated subunits. Figure1 shows the flow chart of the system and a diagram of its important units.

TABLE 1. Dataset detail.

Dataset Description	Values
Dataset acquisition intervals	1/01/2014-10/31/2016
Total abnormal users count before the dataset balancing	3615
Total normal users count before the dataset balancing	38757
Total abnormal users count of the trainset before the data balancing	2857
Total normal users count of the trainset before the data balancing	31040
Total abnormal users count of the testset	758
Total normal users count of the testset	7717
Total abnormal users count of the trainset after the data balancing	20990
Total normal users count of the trainset after the data balancing	29986
Total users count before the initial preprocessing of raw data	42372
Total users count of the trainset before the data balancing	33897
Total user count of the testset	8475
Total users count of the trainset after the data balancing	50976

A. DATA DESCRIPTION

The present study utilizes the authentic electricity consumption data of users, which has been publicly released by the State Grid Corporation of China (SGCC). The dataset consists of data collected from January 2014 to October 2016, with a daily sampling frequency. Each row represents the electricity consumption data of an individual user, while each column corresponds to a specific sampling time [17]. Among them, there are 42,372 power data records, comprising 38,757 normal power data records and 3,615 abnormal power data records. The distribution of normal data and abnormal data exhibits a significant imbalance. The data contains a significant number of NaN values, outliers, and discrete data due to uncontrollable factors. Therefore, it is imperative to address these issues during the ETD. To address the issue of class imbalance in the dataset, this study employs a sampling technique that combines oversampling and undersampling to achieve a balanced trainset, while reserving 20% of the imbalanced data as an independent testset. The detailed description of this series of processing will be the subsequent section on data processing. The details of all datasets in this paper are presented in Table1. Additionally, figure2 illustrates the distribution of abnormal and normal users across various datasets: the original dataset, the testset, the trainset prior to data balancing, and the training number set post data balancing.

B. MISSING VALUES HANDLING UNIT

Missing cases in the collected data samples may arise due to staff errors, collector failures, or network fluctuations of

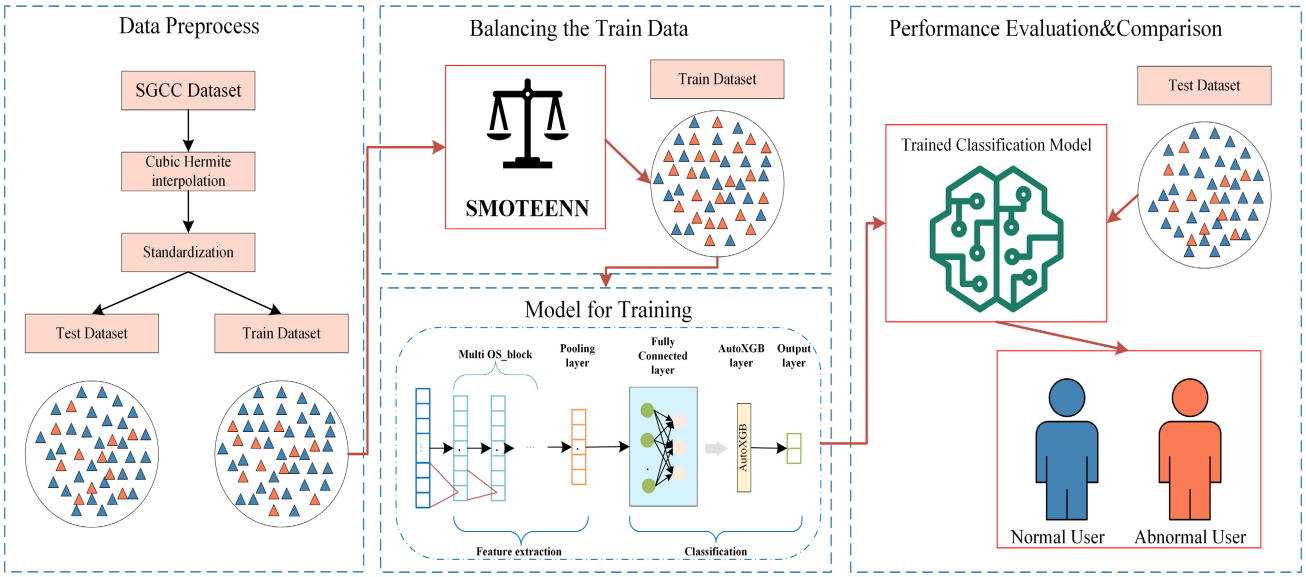


FIGURE 1. Electricity theft detection workflow diagram.

smart meters. The absence of data processing for missing values can result in a substantial loss of valuable information and a decline in data quality, ultimately leading to the failure of the model to achieve the anticipated outcome. In handling missing values, a conventional approach is to directly remove the row or column containing the missing value. Although this method is straightforward and easy to implement, it may lead to significant data loss. Another alternative involves employing ML algorithms or DL networks to predict missing values, which can yield more realistic data; however, this approach requires significant time and resource investments for computation, and its predictive accuracy may be influenced by the trainset. Therefore, considering its comprehensive performance, interpolation methods are widely acknowledged for effectively handling missing power data samples. The following interpolation algorithms are commonly employed:

1) SIMPLE IMPUTER WITH MEAN METHOD

The Simple Imputer with Mean Method (SIMM) is a widely utilized technique, with the following formula:

$$f(x_i) = \begin{cases} \dfrac{x_{i-1} + x_{i+1}}{2}, & x_i \in NaN, x_{i-1}, \\ & x_{i+1} \notin NaN \\ 0, & x_i \in NaN, x_{i-1} \text{ or} \\ & x_{i+1} \in NaN, \\ x_i, & \text{otherwise}, \end{cases} \tag{1}$$

where, the vector x represents the daily electricity consumption data, x_i, x_{i-1} and x_{i+1} denote the data values of day i, $i-1$ and $i+1$ in x. SIMM, although easy to implement and consuming fewer resources, only considers a fraction of the daily electricity data within the current sample when handling missing values. It disregards the overall electricity consumption pattern of the user. Furthermore, in cases where consecutive missing values occur in the sample, employing SIMM may result in a significant number of successive zero values within the sampled data.

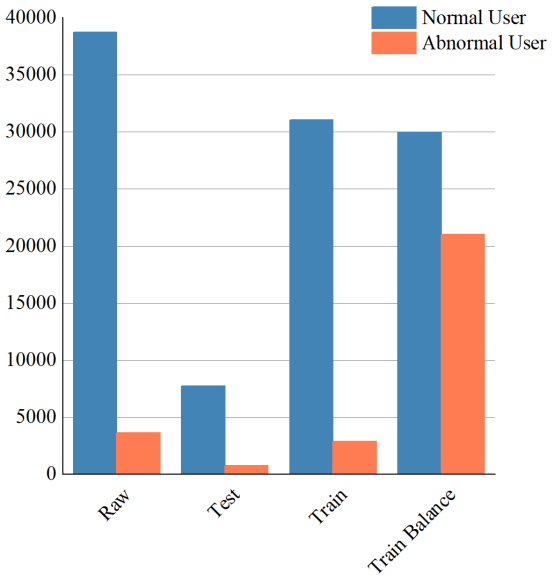


FIGURE 2. Data distribution.

2) PIECEWISE CUBIC HERMITE INTERPOLATING POLYNOMIAL

The PHCIP represent a form of interpolation that effectively preserves the inherent shape and characteristics of function [26]. Hermite, a French mathematician, used the function value and derivative value of the unknown function $f(x)$ at the interpolation point to construct the PCHIP. This type

of polynomial has characteristics such as C^1 continuity and monotonicity in subintervals, making it suitable for fitting differences in electricity data. The mathematical principle of PCHIP is presented herein:

The value of function $f(x)$ at node $a = x_0 < x_1 < \cdots < x_{k-1} < \cdots < x_n = b$ denoted as $y_0, y_1, \cdots, y_{k-1}, y_k, \cdots, y_n$, and the corresponding function interval for subinterval $[x_{k-1}, x_k]$ in partition k is represented by $[y_{k-1}, y_k]$. Therefore, the PCHIP function $P_k(x)$ is defined on the interval as follows:

$$\begin{cases} P_k(x) = y_{k-1} + a_{k,1}(x - x_{k-1}) \\ \qquad +a_{k,2}(x - x_{k-1})^2 + a_{k,3}(x - x_{k-1})^3, \\ a_{k,1} = d_{k-1}, \\ a_{k,2} = \dfrac{3(y_k - y_{k-1})}{(x_k - x_{k-1})^2} - \dfrac{2d_{k-1} + d_k}{x_k - x_{k-1}}, \\ a_{k,3} = -\dfrac{2(y_k - y_{k-1})}{(x_k - x_{k-1})^3} + \dfrac{d_k + d_{k-1}}{(x_k - x_{k-1})^2}, \end{cases} \tag{2}$$

the formula incorporates d_{k-1}, d_k, which represent the function value and first derivative of the interpolation function $P_k(x)$ at the left and right endpoints of the subinterval, respectively. The determination of the derivative of the interpolation function at each node, based on known interval endpoint function values, is crucial for constructing a monotonic PCHIP function within an interval.

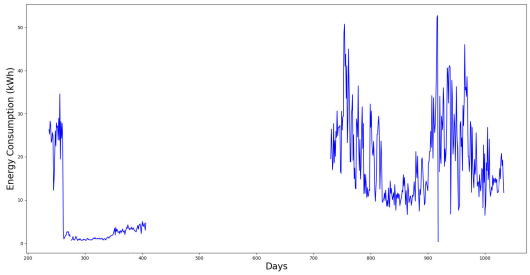
The derivative for each intermediate node $k = 1, 2, \cdots, n-1$ is approximately calculated by weighting the first-order difference quotient of the adjacent intervals on both sides.

$$\begin{aligned} \delta_k &= \frac{y_k - y_{k-1}}{x_k - x_{k-1}}, \quad w_1 = \frac{1}{3}\left(1 + \frac{x_k - x_{k-1}}{x_{k+1} - x_{k-1}}\right), \\ w_2 &= \frac{1}{3}\left(1 + \frac{x_{k+1} - x_k}{x_{k+1} - x_{k-1}}\right), \\ d_k &= \begin{cases} \dfrac{\delta_k \cdot \delta_{k+1}}{w_1\delta_k + w_2\delta_{k+1}}, & \delta_k \cdot \delta_{k+1} > 0 \\ 0, & \delta_k \cdot \delta_{k+1} \leqslant 0 \end{cases} \\ d_0 &= \delta_1, \quad d_n = \delta_n. \end{aligned} \tag{3}$$

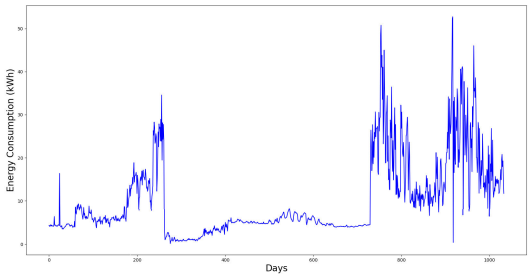
The polynomial coefficients are calculated using Equation (2), and the first derivatives of the interpolation function at each node are computed using Equation (3). Finally, the interpolation function is determined $P_k(x)$. Furthermore, it can be demonstrated that $P_k(x)$ satisfies C^1 continuity on the interval $[x_0, x_n]$ and remains monotonic in the subinterval $[x_{k-1}, x_k]$.

The PHCIP method was selected for handling missing values in the data sample during the experiment, after evaluating two interpolation approaches.

The interpolated data generates a smooth curve between the maximum and minimum values of adjacent points, while preserving the consumption pattern, as illustrated in Figure3 for a random sample.



(a)Consumption data before interpolation



(b)Consumption data after interpolation

FIGURE 3. Plots of consumption data before and after interpolation.

C. OUTLIER HANDLING UNIT

The processing of outliers is a crucial step in data preprocessing as it serves to mitigate or eliminate errors and biases that arise from including anomalous observations in the dataset. The generation of outliers primarily arises from non-resistant factors that induce deviations in smart meter records, resulting in the emergence of discrete data points. Unprocessed outliers in the data can lead to reduced model performance and compromise its robustness. Therefore, appropriate handling of outliers is imperative. Common methods for outlier processing include deletion, replacement, and interpolation techniques. Deletion represents the most straightforward and cost-effective approach; however, it may result in the loss of valuable information. Replacement can preserve the data sample, but it is crucial to select an appropriate method based on specific circumstances. The process of interpolation involves estimating outliers based on existing data points. In practical applications, selecting an appropriate method for processing outliers requires considering factors such as data distribution and feature analysis to ensure accuracy and reliability of the data. This paper employs the Three Sigma Rule (TSR) to address the presence of outliers in the data. The formula for this rule is presented as follows:

$$f(x_i) = \begin{cases} \text{avg}(x) + 3 \cdot \text{std}(x), & \text{if } x_i > \text{avg}(x) + 3 \cdot \text{std}(x), \\ x_i & \text{otherwise}, \end{cases} \tag{4}$$

the vector x represents the daily electricity consumption data, while x_i denotes the data value of x on the i day. Additionally, $avg(x)$ corresponds to the mean value of sample x, and finally,

$std(x)$ signifies the standard deviation of this sample. The value that surpasses $\text{avg}(x) + 3 \cdot \text{std}(x)$ in TSR is recognized as an outlier and attributed with the value of $\text{avg}(x) + 3 \cdot \text{std}(x)$. This approach not only mitigates the impact of outliers but also partially alleviates the potential reduction in model effectiveness caused by misjudgment of outliers.

D. DATA NORMALIZATION UNIT

The normalization of raw data is essential prior to training, as DL techniques exhibit heightened sensitivity towards sparse, diverse, and unscaled data. If the data range is excessively large, it may adversely impact the convergence performance of the model or even prevent it from converging, so that the model fails to attain the intended outcome. The two most prevalent standardization approaches are as follows:

1) MIN-MAX SCALING

The Min-Max Scaling function is employed to transform the raw data, mapping it onto a specified interval, typically set as the default range of [0, 1]. The expression for Min-Max Scaling is as follows:

$$f(x) = \frac{x_i - \min(x)}{\max(x) - \min(x)}, x_i \in x, \tag{5}$$

the vector x represents the daily electricity consumption data, where x_i denotes the data value of day i in x. Additionally, $min(x)$ corresponds to the minimum value in sample x, while $max(x)$ represents the maximum value in sample.

The Min-Max Scaling method exhibits high sensitivity to outliers in the dataset and should only be applied to data that has a well-defined range without any outliers. In the power data, a significant number of values have high magnitudes alongside a substantial proportion of values approaching zero. Consequently, using Min-Max Scaling may result in an abundance of near-zero values within the sample, leading to excessive elimination of data features and adversely affecting model performance.

2) ZERO-SCORE STANDARDIZATION

The function of Zero-Score Standardization is to let raw data follow Gaussian distribution. The expression for Zero-Score Standardization is as follows:

$$f(\boldsymbol{x}) = \frac{x_i - \text{avg}(\boldsymbol{x})}{\text{std}(\boldsymbol{x})}, x_i \in \boldsymbol{x}, \tag{6}$$

the vector x represents the daily electricity consumption data, where x_i denotes the data value of day i in x, additionally, $avg(x)$ corresponds to the average value in sample x, while $std(x)$ represents the standard deviation value in sample.

The Zero-Score Standardization method is less influenced by outliers in the sample compared to Min-Max Scaling, making it suitable for datasets lacking a distinct range. In the experiment, while controlling for other variables, both standardization methods were examined as shown in Figure4. The results indicate that the Zero-Score Standardization model demonstrates superior performance in this specific experiment.

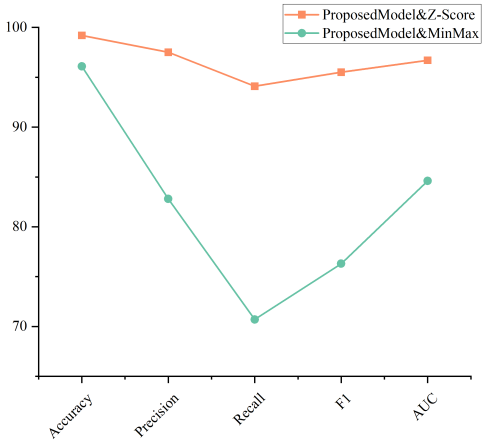


FIGURE 4. Data distribution.

E. DATA CLASS BALANCE UNIT

The data in ETD exhibits a significant imbalance, with the proportion of electricity theft users being considerably smaller compared to the total number of users. Imbalanced data can lead to overfitting issues and hinder the generalization performance of the theft detection model. Without data balancing, models that demonstrate satisfactory performance on the trainset may arise; however, on the testset, the model tends to be biased towards the majority class and performs poorly on the minority class.

The most commonly employed processing technique at the data level involves utilizing oversampling technology to enhance minority samples and achieve class balance in the sample data. Within oversampling technology, the widely adopted approach is SMOTE, which operates by identifying KNN samples surrounding the minority class and generating new samples through linear interpolation operations. The SMOTE sampling technique [27], however, faces the challenge of generating minority class samples is difficult to distinguish due to their overlap with majority class samples. Therefore, some researchers have proposed a hybrid sampling technology that combines over-sampling and down-sampling techniques, which is SMOTEENN [28]. Firstly, the minority class samples were generated using the SMOTE algorithm to obtain newly synthesized instances. Subsequently, a clustering algorithm KNN was employed to cluster these newly generated samples. If the classification result at a specific point coincided with the clustering outcome of its K-nearest neighbor samples, the synthesized instances were retained; otherwise, they were discarded iteratively until achieving balance between minority and majority class data samples. This approach addresses the issue of duplication in both minority and majority class data within the SMOTE algorithm. The steps for implementing the SMOTEENN algorithm are presented in Algorithm1:

The trainset in this paper is balanced by 80% based on the algorithm section above. The number of normal users and abnormal users in the training samples after

Algorithm 1 SMOTEENN Based Data Augmentation

Input: Training set(X_train) that consists of majority class data(X_train$_{maj}$) and minority class data (X_train$_{min}$), Number of minority class samples (T),Sample Rate(N)%, count of neighbors(K), New minority class sample(x_s)

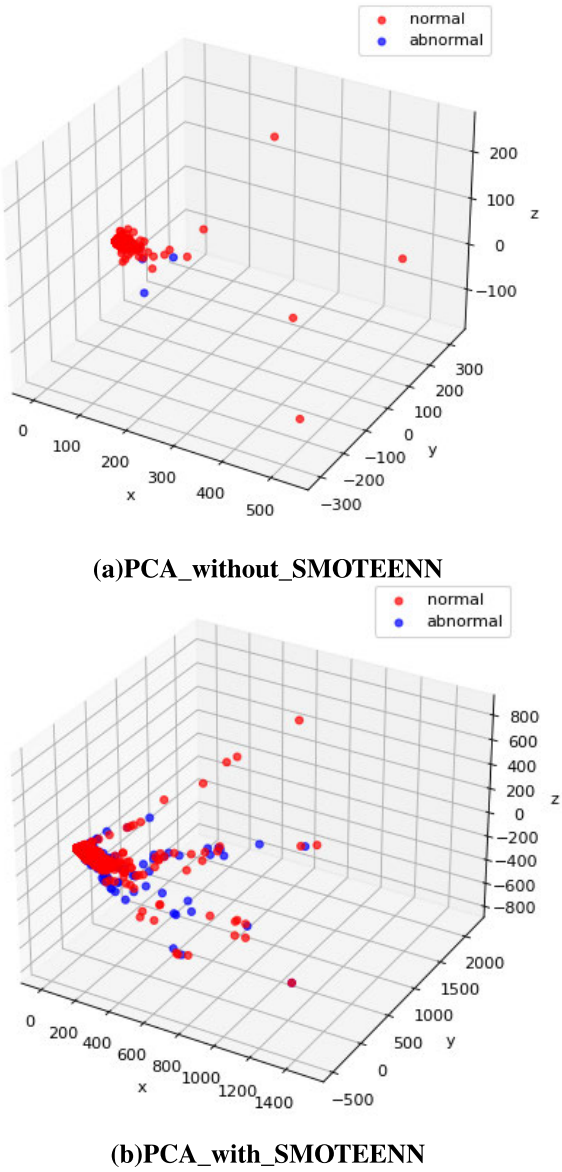
Output: Augmented Data

1: **if** $N < 100$ **then**
2: Randomly select $T^*N\%$ samples from the (X_train $_{\min}$)
3: Find the K nearest neighbor points $Xinn, nn \in \{1, 2, \cdots, K\}$ of X_train $_{\min}$ in X_train
4: **while** $x_s < (N/100) * T$ **do**
5: Select a sample $Xiab$ arbitrarily from $Xinn$
6: Calculate the vector difference between the remaining current traversed sample $Xicd$
7: New minority class sample $x_s = x_{icd} + (x_{iab} - x_{icd}) * \text{rand}(0, 1)$
8: According to the KNN algorithm, predict and classify the newly generated data samples. If it has same type as most of its K nearest neighbor samples, save the data, otherwise delete it.
9: Add new samples to original samples
10: **end while**
11: **end if**

generating synthetic data are presented in columns 7 and 8 of Table1. Figure5 illustrates the data distribution before and after sampling, achieved by employing Principal Component Analysis (PCA) to reduce the dimensionality of the data.

F. PROPOSED ELECTRICITY THEFT DETECTION MODEL UNIT

The subsequent steps involving feature extraction and anomaly classification can be initiated after completing the data processing task. The SGCC dataset encompasses a substantial volume of high-dimensional feature data. To enhance the process of anomaly classification, it is imperative to perform dimensionality reduction as an initial step, thereby alleviating the curse of dimensionality. This will prevent the model from being overwhelmed by excessive noise and enable it to demonstrate stronger generalization capabilities towards novel data. The proposed model for ETD in this paper is depicted in Figure6, comprising two structures: the OS-CNNs feature extraction structure, which consists of multiple Omni-Scale block (OS-block) layers to extract dimensional data features in SGCC; and the anomaly classification structure composed of AutoXGB. Following the passage through the pooling layer and fully connected layer, the features extracted by the OS-CNNs network are fed into the AutoXGB classifier for performing anomaly classification. The subsequent subsections will provide a comprehensive account of the detailed process involved in feature extraction and anomaly classification.



(a)PCA_without_SMOTEENN

(b)PCA_with_SMOTEENN

FIGURE 5. PCA visualization of data.

1) OS-CNNS FEATURE EXTRACTION BASED ON OS-BLOCK STACKING

The primary challenge in time series data lies in selecting an appropriate time window scale for effective feature extraction. Conventional ML methods exert significant efforts in capturing crucial time scales, however, as the length of the time series increases, computational resources grow exponentially, as exemplified by Shapelet. CNNs have demonstrated effective feature extraction capabilities for time series analysis, with the size of the Receptive Field consistently recognized as a crucial factor that influences the performance of 1D-CNNs in time series classification tasks.

In [10], the authors proposed the concept of an OS-block, which automatically sets the kernel selection of a 1D-CNN through a simple and general rule that can cover Receptive

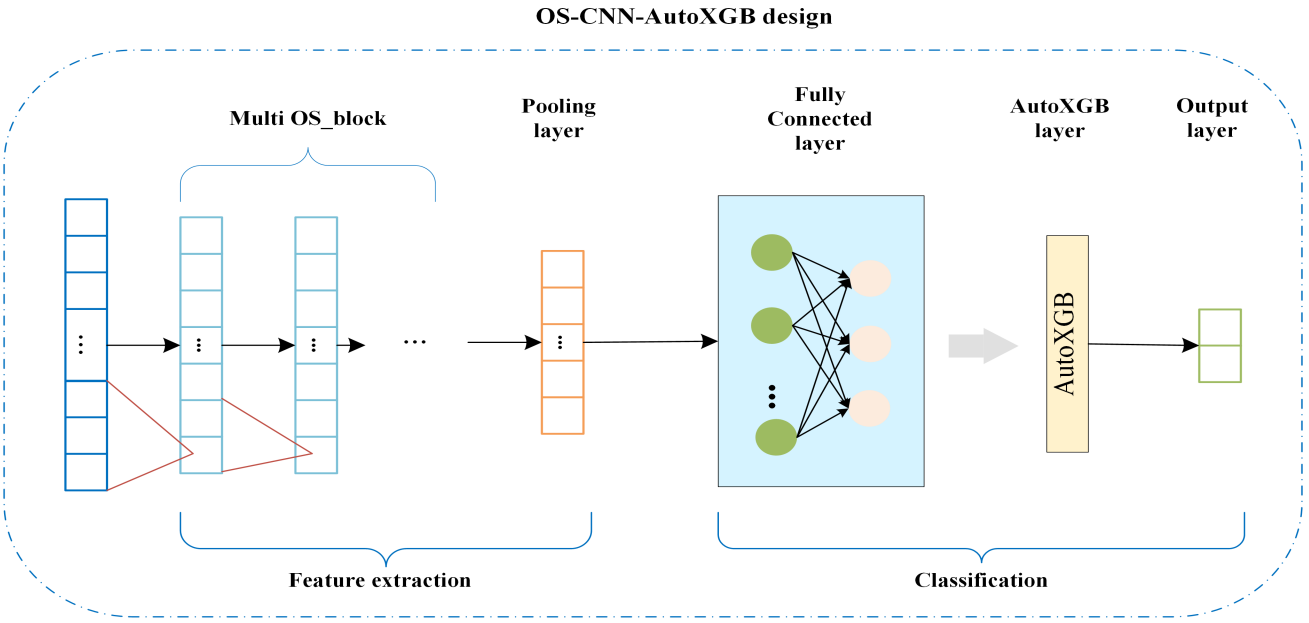


FIGURE 6. Proposed model.

Field at different time series scales. The rule is inspired by the Goldbach Conjecture [29], which states that every positive even integer can be expressed as the sum of two prime numbers. Hence, OS-block employs a set of numbers as the kernel size and exclusively utilizes 1 and 2 as the kernel size for the final layer in each block. In this manner, a 1D-CNN utilizing prime size kernels can effectively transform the time series by employing various combinations of these prime size kernels to encompass receptive fields at multiple scales. More significantly, OS-block can execute the processing of diverse time series datasets by selecting the maximum prime number based on the length of each time series.

As shown in Figure7, each even number from 2 to 38 can be composed of two prime numbers from 1 to 19. According to Goldbach Conjecture, this phenomenon can be extended to all even numbers. The OS-block structure is proposed as depicted in Figure7, based on this conjecture. It constitutes a-layer multi-kernel 1D-CNN architecture. Specifically, the first two layers employ prime-sized kernels ranging from 1 to 19 to encompass all even receptive field sizes, while the third layer utilizes kernels of size 1 and 2. Consequently, diverse p_k are employed to cover all receptive field sizes within the specified range. The set of kernel sizes at layer i can be denoted as $P^{(i)}$.

$$P^{(i)} = \begin{cases} \{1, 2, 3, 5, \ldots, p_k\} & , i \in (1, 2), \\ \{1, 2\} & , i = 3. \end{cases} \tag{7}$$

The exceptional feature extraction properties of OS-block on time series render it a highly favorable choice. The present study employs an OS-CNNs network composed of stacked multi-layer OS-blocks to extract features from high-dimensional data obtained from SGCC. The mathematical principle of OS-block design, as illustrated in Figure7, demonstrates its ability to cover receptive fields of all scales based on the length of time series.

2) ANOMALY CLASSIFICATION BASED ON AutoXGB

The hyperparameter tuning is a pivotal aspect in ML, and appropriately hyperparameter tuning can significantly enhance the model's performance. In the realm of hyperparameter tuning, researchers commonly employ network search and random search techniques. Grid search involves exhaustively exploring all possible combinations of hyperparameters within a given search space to identify the optimal setting based on evaluation metrics. However, grid search exhibits evident limitations. Firstly, it necessitates enumerating all potential values for each hyperparameter within a given range. This can pose challenges when dealing with continuous hyperparameters as determining their value range may be arduous. Furthermore, when the search scope is excessively broad, there will be a significant increase in both the time and resources required for web searching, thereby adversely impacting optimization efficiency. Therefore, to address the limitations of grid search, random search employs random sampling within the specified hyperparameter search range to generate candidate hyperparameters and subsequently selects the optimal combination. This approach partially mitigates the resource consumption issue associated with grid searching. However, both random searching and grid searching suffer from artificially setting the range for hyperparameter searches which prevents obtaining optimal hyperparameter settings.

The AutoXGB [30], [31], [32] tool is an open-source, user-friendly, and highly efficient Automated Machine Learning

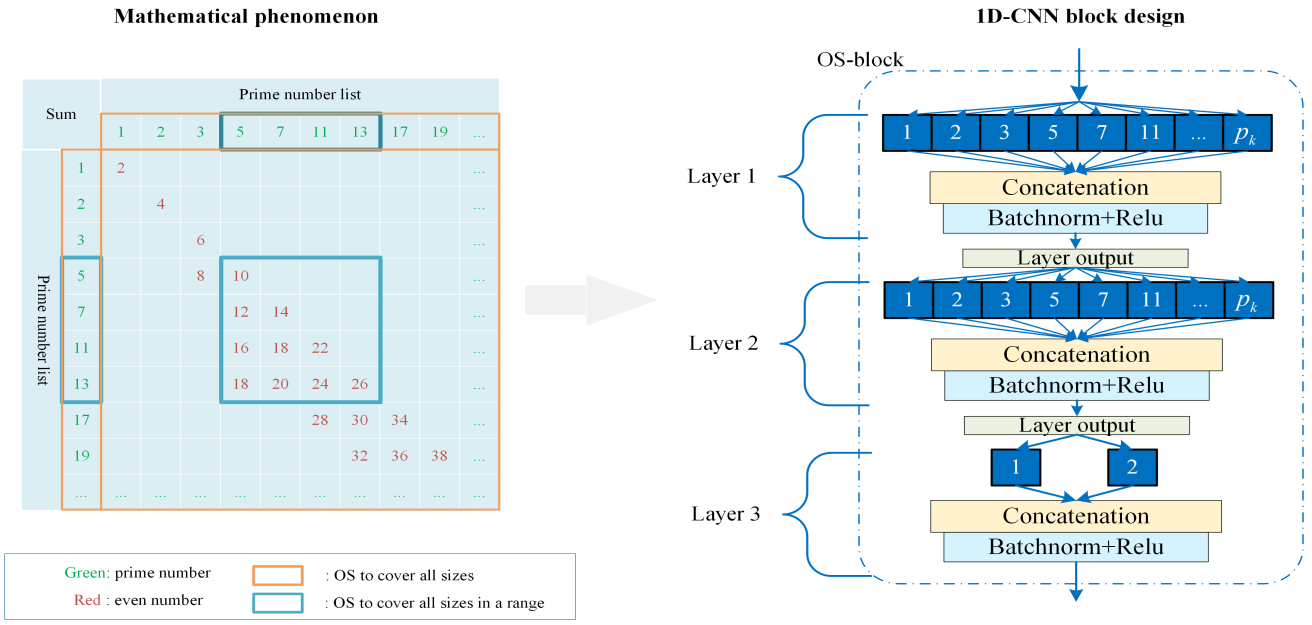


FIGURE 7. OS-block design.

(AutoML) development tool. The AutoXGB framework leverages the optimized Distributed Gradient Boosting library XGBoost for model training, while employing Optuna, an automatic hyperparameter optimization framework tailored for ML and DL, to fine-tune the hyper of XGBoost. The need for engineers to re-optimize XGBoost hyperparameters is eliminated, resulting in saved optimization steps and time. Optuna, unlike manual experience-based hyperparameter optimization, offers a range of advanced parameter tuning algorithms that can terminate underperforming sampling points early to expedite the search process. Additionally, it dynamically constructs the hyperparameter search space to adapt to various optimization problems. This enables the discovery of more suitable for enhancing model performance. The optimal combination of hyperparameters for XGBoost in ETD, as determined by the Optuna automatic hyperparameter optimization framework, is presented in Table2.

XGBoost is an enhanced version of the Gradient Boosting Decision Tree (GBDT) algorithm. The idea behind XGBoost is to add one tree at a time in order to fit the residual of the previous prediction, thereby continuously reducing the loss by adding new trees. The proposed approach amalgamates multiple weak classifiers into a robust classifier, thereby yielding an ML model with exceptional accuracy. The algorithm utilizes the second-order Taylor expansion of the loss function and incorporates a regularization term to mitigate overfitting, rendering it an efficient and high-precision Boosting ensemble learning algorithm.

For the dataset $D = \{(x_i, y_i)(i = 1, 2, \ldots, n)\}$, x_i represents the i sample, while y_i represents the true value corresponding to the i sample. XGBoost utilizes CART

TABLE 2. Hyperparameters of XGBoost.

Hyperparameter	Detail	Best Value
learning_rate	learning rate	0.1817
reg_lambda	L2 regularization	0.0041
reg_alpha	L1 regularization	0.0013
subsample	proportion of samples randomly selected from each individual tree	0.5705
colsample_bytree	proportion of features randomly selected from each individual tree	0.6200
max_depth	maximal depth of the tree	1
n_eatimators	Boost times	7000

regression tree as its weak classifier, and the prediction output of XGBoost for the input x_i is:

$$\hat{y}_i = \theta(x_i) = \sum_{k=1}^{K} f_k(x_i), f_k \in F, \tag{8}$$

where:K represents the number of sub-models;F denotes all regression trees, while $F = \{f(x) = \boldsymbol{\omega}\}$, $\boldsymbol{\omega}$ refer to weight vectors comprising weights assigned to each leaf node of the regression trees. Meanwhile, $\hat{y}_i$ signifies the predicted value of the model's output;x_i represents the sample input; and finally, f_k corresponds to the regression tree indexed as 'k'.

The model's objective function is as follows:

$$O = l\left(y_i, \hat{y}_i\right) + \sum_{k=1}^{K} \Omega\left(f_k\right), \tag{9}$$

	Positive	Negative
Positive	**True Positive (TP)**	**False Negative (FN)**
Negative	**False Positive (FP)**	**True Negative (TN)**

FIGURE 8. Confusion matrix.

where, the loss function, denoted as $l\left(y_i, \hat{y}_i\right)$, quantifies the discrepancy between the predicted value $\hat{y}_i$and the actual value y_i. Meanwhile, the regularization term Ω plays a crucial role in mitigating overfitting of the model. In general, the performance of the XGBoost model is influenced by the size of the trainset. When confronted with a high-dimensional feature space and limited training samples, XGBoost may struggle to capture all relevant information from the data, leading to overfitting issues that hinder its applicability in real-world testing scenarios. Therfore, the paper initially employs OS-CNN for feature extraction, followed by training the XGBoost model using the extracted features, thereby significantly enhancing the classification efficacy of the model.

IV. MODEL PERFORMANCE EVALUATION

A. PERFORMANCE METRICS

The selection of a suitable performance metric is imperative when dealing with class imbalance [33]. The performance of the model is assessed by employing various evaluation metrics, including accuracy, precision, recall, F-1 score, and AUC score. The confusion matrix serves as a fundamental tool for assessing the performance of a classifier. The abnormal power consumption data sample is categorized as the positive class, while the normal power consumption data sample is classified as the negative class in this experiment. The distribution of the confusion matrix is illustrated in Figure8.

TP: indicates that the predicted abnormal user is actually an abnormal user;

FN: indicates that the predicted normal user is actually an abnormal user;

FP: indicates that the predicted abnormal user is actually a normal user;

TN: indicates that the predicted normal user is actually a normal user;

The detection effect is enhanced with higher values of TP and TN.

The following section provides a detailed explanation of specific metrics:

1) ACCURACY

The accuracy rate is defined as the ratio of correctly classified samples to the total number of samples, providing a measure of classification performance. The formula for calculating accuracy is as follows:

$$ACC(\%) = \frac{TP + TN}{TP + TN + FP + FN} \times 100. \tag{10}$$

In the case of datasets that are generally balanced, accuracy is commonly employed to assess model performance. However, for imbalanced datasets, accuracy becomes an inadequate measure of a model's true predictive power due to its susceptibility to sample size variations across different classes. The testset in this experiment is characterized by an extreme imbalance, with normal users accounting for 91.0% of the testset. Consequently, if the model were to predict all instances as normal users, it would achieve a correct classification rate of up to 91.0%. However, such performance lacks significance within the context of this study. Hence, it is not advisable to solely rely on accuracy as the sole metric for assessing the efficacy of a model.

2) PRECISION

It denotes the ratio of true positive samples to the sum of true positive and false positive samples, as defined by the following formula:

$$\text{Precision}(\%) = \frac{TP}{TP + FP} \times 100. \tag{11}$$

The higher the precision, the greater the model's predictive capacity for normal users. In the context of ETD, a high level of precision can alleviate the later-stage workload for workers by reducing misclassifications of normal users as abnormal.

3) RECALL

It denotes the proportion of correctly identified positive samples to the total number of samples classified as positive. The formula is defined as follows:

$$\text{Recall}(\%) = \frac{TP}{TP + FN} \times 100. \tag{12}$$

The higher the recall, the greater the predictive capacity of the representation model for identifying abnormal users. In the ETD, a high recall enables accurate identification of electricity theft by users, thereby mitigating potential harm to smart grid companies caused by undetected instances of electricity theft.

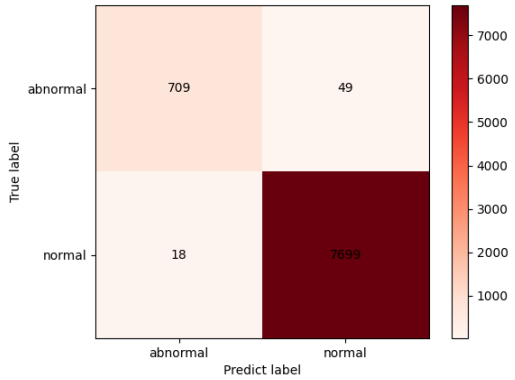


FIGURE 9. Confusion matrix of the proposed model.

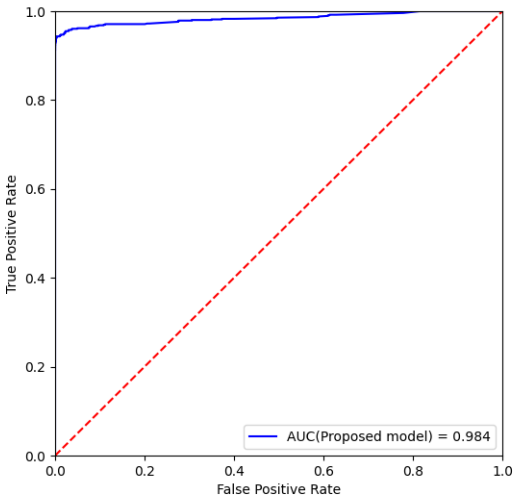


FIGURE 10. ROC curve of the proposed model.

4) F1 SCORE

It represents the harmonic mean between precision and recall, which is mathematically defined by the following formula:

$$F1(\%) = \frac{2 * \text{Precision} * \text{Recall}}{\text{Precision} + \text{Recall}} \times 100. \quad (13)$$

In the context of ETD, it is not advisable to solely prioritize higher precision or recall due to the unique nature of this task. To ensure accurate identification of electricity theft users and avoid manual misjudgment troubleshooting in later stages, a comprehensive consideration of both indicators is necessary. Hence, the F1 score can be considered as a well-balanced metric, indicating that the model's robustness increases with higher values.

5) AUC SCORE

The calculation involves determining the disparity between positive and negative samples, where AUC values closer to 1 indicate superior classification performance. The formula is defined as follows:

$$\text{AUC} = \sum \frac{\text{Rank}_i - \frac{M(1+M)}{2}}{M \times N}, \quad (14)$$

where, the positive class is represented by i, with the number of positive samples denoted as M and the number of negative samples denoted as N.

The AUC score serves as a robust metric that quantifies the ability to accurately distinguish between positive and negative classes. In ETD, a higher AUC score indicates stronger discriminatory power in distinguishing normal users from those who engage in power theft.

For example, the confusion matrix of the proposed model is depicted in Figure9, while Figure10 illustrates the ROC curve and the corresponding area under the curve for our model.

B. COMPARING MODELS

To demonstrate the superiority of our proposed method, we compared the performance of the OS-CNN-AutoXGB model with five other well-performing ML algorithms for ETD on a given trainset. The open-source frameworks PyTorch and Scikit-Learn were utilized to construct five comparative models. The optimal hyperparameters of the model were obtained through random search. Detailed descriptions of these five models, along with their corresponding specific parameters, are provided below.

1) RF [34]: The algorithm in question is an ML technique that falls under the category of bagging, which itself belongs to the realm of ensemble learning. The weak classifiers in RF are constructed in parallel, and their predictions are integrated through a voting mechanism to determine the final prediction. Compared to a single decision number, random forests employ multiple decision trees trained on different subsets of the trainset, thereby ensuring diversity among the decision tree models. The RF algorithm can simultaneously assign different weights to various categories in order to address the issue of dataset imbalance.
2) CNN [35]: It is a deep feedforward neural network that exhibits the characteristics of local connectivity and weight sharing, making it one of the prominent algorithms in the field of DL. The CNN composed of 1D convolutional layers has been widely used in ETD, and different CNNs are used in many stealing models.
3) Wide&Deep CNN [17]: The architecture comprises two components: a width CNN and a depth CNN. The width CNN effectively captures the characteristics of 1D electricity data, while the depth CNN accurately discerns the periodicity of aperiodic electricity theft and normal electricity consumption based on 2D electricity data. Thus, The Wide&Deep CNN model has demonstrated exceptional performance in the ETD.
4) CNN-LSTM [5]: The proposed system integrates a CNN and a LSTM for ETD. Specifically, the CNN is employed to extract discriminative features, while the LSTM is utilized for feature classification. The model has also demonstrated promising outcomes in the ETD.

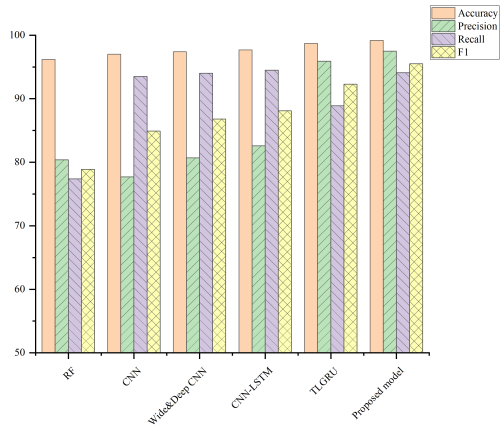


FIGURE 11. Comparison results.

5) TLGRU [36]: The architectural framework consists of two integral components, namely LSTM and GRU. The LSTM network is utilized to capture the features of electricity data for extraction, thereby addressing the issue of dimensionality catastrophe. Subsequently, GRU is employed for feature classification. The TLGRU model demonstrates outstanding performance in the field of ETD.

The specific hyperparameters and descriptions of the compared methods are presented in Table3.

C. MODEL COMPARISON RESULTS AND DISCUSSION

The metrics for each contrasting model in the given dataset are presented in Table4. The proposed model's performance is comprehensively summarized in comparison to other related models. The proposed model achieves the highest accuracy of 99.2%, precision of 97.5%, F1 score of 95.5%, and AUC score of 98.4% according to Table4, while maintaining a recall rate as high as 94.1%. Therefore, the proposed model demonstrates its outstanding capability in detecting theft.

Similarly, figure11 presents a graphical representation that illustrates the performance comparison between the proposed model and other comparative models. To ensure a more precise visualization of the AUC gap, this paper provides a separate graphical representation as depicted in Figure12. The proposed model exhibits significantly enhanced accuracy and precision, as evident from the observations.

To further demonstrate the model's exceptional performance in detecting abnormal data, its effectiveness can also be effectively evaluated under conditions of class imbalance. The ROC curves and PR curves were generated based on the test outcomes. The ROC and PR curves of the various methods are depicted in Figures13 and 14. The results depicted in Figure13 demonstrate the exceptional performance of the proposed classification model when compared to other models, exhibiting a remarkable True Positive Rate (TPR) and an impressively low False Positive Rate (FPR). The preliminary evaluation based on the ROC curve demonstrates that, with a limited number of input samples, the proposed model exhibits effective classification performance by achieving the lowest FPR recorded. Moreover, the model's curve ensures reliable sample classification across various FPR levels. Thus, In order to avoid the imbalanced category, the change in TPR is not readily discernible due to the abundance of positive examples, which inadequately represent the model's efficacy in detecting theft users, resulting in an overly optimistic effect on the ROC curve. Therefore, the PR curve can also serve as evidence of the model's performance. As depicted in Figure14, the proposed

TABLE 3. Description of comparison method and Hyper-parameters selection.

Compare method	Brief Description and Hyper-Parameters Selection
RF [34]	The RF algorithm constructs an ensemble of decision trees using the bagging technique and employs majority voting to obtain the final prediction. In our comparative experiment, we set the number of base estimators to n_estimators=100, while also ensuring that criterion='entropy' for measuring impurity and max_features='auto' for determining the number of features considered at each branching node.
CNN [35]	Two stacked 1D convolutional kernel Max pooling layers are employed, with a kernel size of 7 and an output dimension of 64 and 32 respectively. A window size of 2 is applied for the pooling layer, followed by a linear layer for classification.
Wide&Deep CNN [17]	The original data is utilized as the input for the Wide CNN, where two 1D convolutional layers and Max pooling layers are stacked. The convolutional layer has a kernel size of 7 with an output of 64 and 32 respectively, while the pooling layer has a window size of 2 to extract features. The data is divided into periodic segments with a period of 7 days, serving as input for the Deep CNN. Here, two convolutional layers and Max pooling layers are also stacked. The convolutional layer has a kernel size(7x3) with an output of 64 and 32 respectively, while the pooling layer has a window size of (3x3). The features extracted from the Deep CNN were expanded and integrated with those obtained from the Wide CNN, followed by classification using a linear layer.
CNN-LSTM [5]	The model architecture consists of three consecutive 1D convolutional layers and max pooling layers with a kernel size of 2 for the convolutional layers, resulting in outputs of 128, 64, and 32 respectively. A window size of 2 is used for the pooling layers. These are followed by three LSTM units and dropout layer with a rate of 20% to prevent overfitting, resulting in outputs of size 32 each. Finally, a linear layer is employed for classification.
TLGRU [36]	The original data is transformed into 3D data to ensure compatibility with the input requirements of LSTM. Subsequently, the input data passes through a series of network layers consisting of multi-layer LSTM network layers, LeakyReLU activation functions, Dropout layers, and multi-layer GRU network layers. Finally, a fully connected layer is utilized for classification purposes.

TABLE 4. Comparison of performance metrics of various methods.

Methods	Metrics(%)				
	Accuracy	Precision	Recall	F1	AUC
RF [34]	96.2	80.4	77.4	78.9	87.8
CNN [35]	97.0	77.7	93.5	84.9	95.4
Wide&DeepCNN [17]	97.4	80.7	94.0	86.8	98.3
CNN-LSTM [5]	97.7	82.6	**94.5**	88.1	97.6
TLGRU [36]	98.7	95.9	88.9	92.3	94.3
Proposed model	**99.2**	**97.5**	94.1	**95.5**	**98.4**

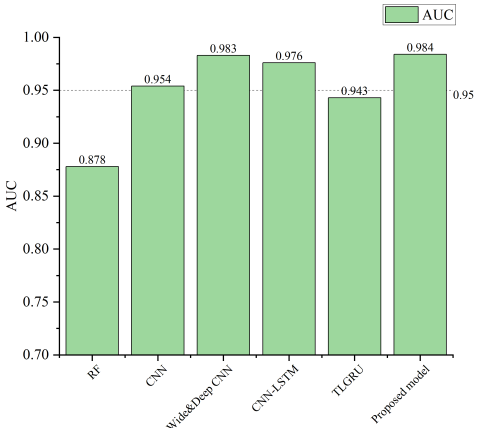


FIGURE 12. Comparison results of AUC.

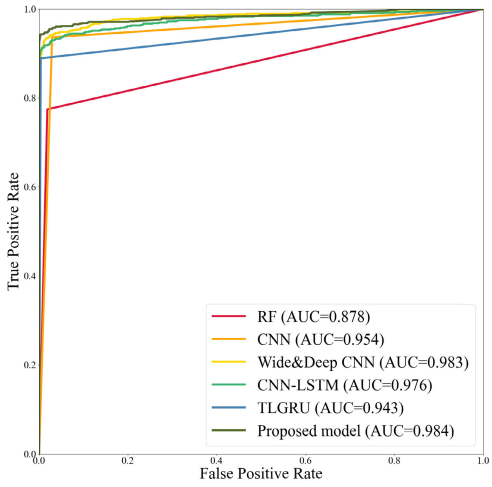


FIGURE 13. Comparison results of ROC curve.

model exhibits commendable performance on the PR curve, indicating its proficiency in classifying positive and negative samples. When applied to real-world abnormal electricity consumption detection, it accurately identifies abnormal electricity users while minimizing interference with normal electricity users. The model not only demonstrates stability

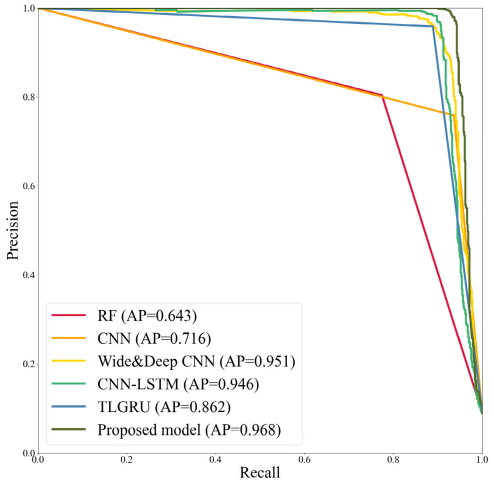


FIGURE 14. Comparison results of PR curve.

but also significantly alleviates the burden on staff in rectifying errors.

V. CONCLUSION AND FUTURE WORK

The present study introduces the OS-CNN-AutoXGB model for the ETD behavior in smart grid systems. In particular, The OS-CNN is specifically employed to extract features from smart grid data at its full scale, surpassing the feature extraction capabilities of both conventional CNN and Wide&Deep CNN models. This advanced approach effectively captures the periodic patterns in normal electricity consumption as well as the irregularities in abnormal electricity consumption, thereby significantly enhancing the performance of our proposed model. In addition, the utilization of AutoXGB as the classifier following feature extraction offers enhanced convenience and efficiency compared to employing ML alone, followed by a series of hyperparameter methods. Moreover, the impact of hyperparameter optimization becomes more pronounced, eliminating the need for manual expertise in this regard. The aforementioned experiments collectively demonstrate that the proposed model exhibits superior equilibrium and stability in comparison to existing models.

Naturally, there are areas for potential improvement in the proposed model due to its limited dataset. Specifically,

within the national smart grid dataset, malicious consumers constitute only 8.5% of the total consumers, presenting a significant issue of data imbalance. Despite using the advanced sampling method SMOTEENN to address class imbalance and augmenting the trainset with additional negative samples resulting in improved model performance, 49 out of the 758 malicious samples in the testset were still misclassified as normal instances. This discrepancy may be attributed to the limited size of the original negative class sample and the single distribution learned by the model from negative class samples. Subsequent research can employ stealing attack methods to simulate a malicious user who modifies consumption data and generates new malicious consumers in order to balance the dataset. Moreover, this model utilizes a hybrid approach that combines DL and ML, thereby retaining the robust feature extraction capabilities of DL while leveraging the classification abilities of ML to reduce resource overhead. However, it is important to note that this method requires more rigorous training compared to conventional approaches. The further research can optimize the network structure and employ a lightweight architecture to achieve an equivalent level of effectiveness in detecting electricity theft. Furthermore, future research will explore the comparison of different proportions of trainset and testset.

REFERENCES

[1] W. Liao, Z. Yang, K. Liu, B. Zhang, X. Chen, and R. Song, "Electricity theft detection using Euclidean and graph convolutional neural networks," *IEEE Trans. Power Syst.*, vol. 38, no. 4, pp. 3514–3527, Jul. 2022.

[2] A. M. Atiku, S. Ismail, F. Roslan, and A. U. Ahmad, "The effect of electricity distribution loos, electricity power consumption, electricity intensity on energy consumption in West Africa," *Int. J. Energy Econ. Policy*, vol. 12, no. 5, pp. 361–369, Sep. 2022.

[3] R. Kaur and G. Saini, "Electricity theft detection methods and analysis using machine learning: Overview," in *Proc. ICSEAPT*, 2022, pp. 527–546.

[4] K. Fei, Q. Li, C. Zhu, M. Dong, and Y. Li, "Electricity frauds detection in low-voltage networks with contrastive predictive coding," *Int. J. Electr. Power Energy Syst.*, vol. 137, May 2022, Art. no. 107715.

[5] M. N. Hasan, R. N. Toma, A.-A. Nahid, M. M. M. Islam, and J.-M. Kim, "Electricity theft detection in smart grid systems: A CNN-LSTM based approach," *Energies*, vol. 12, no. 17, p. 3310, Aug. 2019.

[6] R. Yao, N. Wang, W. Ke, P. Chen, and X. Sheng, "Electricity theft detection in unbalanced sample distribution: A novel approach including a mechanism of sample augmentation," *Appl. Intell.*, vol. 53, no. 9, pp. 11162–11181, May 2023.

[7] S.-C. Yip, W.-N. Tan, C. Tan, M.-T. Gan, and K. Wong, "An anomaly detection framework for identifying energy theft and defective meters in smart grids," *Int. J. Electr. Power Energy Syst.*, vol. 101, pp. 189–203, Oct. 2018.

[8] A. Ullah, N. Javaid, O. Samuel, M. Imran, and M. Shoaib, "CNN and GRU based deep neural network for electricity theft detection to secure smart grid," in *Proc. Int. Wireless Commun. Mobile Comput. (IWCMC)*, 2020, pp. 1598–1602.

[9] R. Xia, Y. Gao, Y. Zhu, D. Gu, and J. Wang, "An efficient method combined data-driven for detecting electricity theft with stacking structure based on grey relation analysis," *Energies*, vol. 15, no. 19, p. 7423, Oct. 2022.

[10] W. Tang, G. Long, L. Liu, T. Zhou, M. Blumenstein, and J. Jiang, "Omni-scale CNNs: A simple and effective kernel size configuration for time series classification," 2020, *arXiv:2002.10061*.

[11] L. J. Lepolesa, S. Achari, and L. Cheng, "Electricity theft detection in smart grids based on deep neural network," *IEEE Access*, vol. 10, pp. 39638–39655, 2022.

[12] Y. Zhu, Y. Zhang, L. Liu, Y. Liu, G. Li, M. Mao, and L. Lin, "Hybrid-order representation learning for electricity theft detection," *IEEE Trans. Ind. Informat.*, vol. 19, no. 2, pp. 1248–1259, Feb. 2023.

[13] A. Ullah, N. Javaid, M. Asif, M. U. Javed, and A. S. Yahaya, "AlexNet, AdaBoost and artificial bee colony based hybrid model for electricity theft detection in smart grids," *IEEE Access*, vol. 10, pp. 18681–18694, 2022.

[14] M. S. Saeed, M. W. B. Mustafa, U. U. Sheikh, A. Khidrani, and M. N. H. Mohd, "Electricity theft detection in power utilities using bagged chaid-based classification trees," *J. Optim. Ind. Eng.*, vol. 15, no. 2, pp. 67–73, 2022.

[15] P. Ghosh, T. T. B. Audry, S. Rahman, F. Bhuiyan, S. T. Rifat, M. N. K. Hredoy, T. Ghosh, and D. M. Farid, "Electricity theft detection employing machine learning algorithms," in *Proc. IEEE 8th Int. Conf. Converg. Technol. (I2CT)*, Apr. 2023, pp. 1–6.

[16] A. Nawaz, T. Ali, G. Mustafa, S. U. Rehman, and M. R. Rashid, "A novel technique for detecting electricity theft in secure smart grids using CNN and XG-boost," *Intell. Syst. Appl.*, vol. 17, Feb. 2023, Art. no. 200168.

[17] Z. Zheng, Y. Yang, X. Niu, H.-N. Dai, and Y. Zhou, "Wide and deep convolutional neural networks for electricity-theft detection to secure smart grids," *IEEE Trans. Ind. Informat.*, vol. 14, no. 4, pp. 1606–1615, Apr. 2018.

[18] M.-M. Buzau, J. Tejedor-Aguilera, P. Cruz-Romero, and A. Gomez-Exposito, "Hybrid deep neural networks for detection of non-technical losses in electricity smart meters," *IEEE Trans. Power Syst.*, vol. 35, no 4, pp. 1254–1263, Sep. 2019.

[19] S. K. Gunturi and D. Sarkar, "Ensemble machine learning models for the detection of energy theft," *Electr. Power Syst. Res.*, vol. 192, Mar. 2021, Art. no. 106904.

[20] A. Ullah, N. Javaid, A. S. Yahaya, T. Sultana, F. A. Al-Zahrani, and F. Zaman, "A hybrid deep neural network for electricity theft detection using intelligent antenna-based smart meters," *Wirel Commun. Mobile Comput.*, vol. 2021, pp. 1–19, Aug. 2021.

[21] F. Shehzad, N. Javaid, A. Almogren, A. Ahmed, S. M. Gulfam, and A. Radwan, "A robust hybrid deep learning model for detection of non-technical losses to secure smart grids," *IEEE Access*, vol. 9, pp. 128663–128678, 2021.

[22] G. Lin, X. Feng, W. Guo, X. Cui, S. Liu, W. Jin, Z. Lin, and Y. Ding, "Electricity theft detection based on stacked autoencoder and the undersampling and resampling based random forest algorithm," *IEEE Access*, vol. 9, pp. 124044–124058, 2021.

[23] S. Hussain, M. W. Mustafa, T. A. Jumani, S. K. Baloch, H. Alotaibi, I. Khan, and A. Khan, "A novel feature engineered-CatBoost-based supervised machine learning framework for electricity theft detection," *Energy Rep.*, vol. 7, pp. 4425–4436, Nov. 2021.

[24] M. J. Abdulaal, M. I. Ibrahem, M. M. E. A. Mahmoud, J. Khalid, A. J. Aljohani, A. H. Milyani, and A. M. Abusorrah, "Real-time detection of false readings in smart grid AMI using deep and ensemble learning," *IEEE Access*, vol. 10, pp. 47541–47556, 2022.

[25] X. Lu, Y. Zhou, Z. Wang, Y. Yi, L. Feng, and F. Wang, "Knowledge embedded semi-supervised deep learning for detecting non-technical losses in the smart grid," *Energies*, vol. 12, no. 18, p. 3452, Sep. 2019.

[26] A. Kumari and V. K. Kukreja, "Survey of Hermite interpolating polynomials for the solution of differential equations," *Mathematics*, vol. 11, no. 14, p. 3157, Jul. 2023.

[27] R. Yang, C. Zhang, R. Gao, and L. Zhang, "A novel feature extraction method with feature selection to identify golgi-resident protein types from imbalanced data," *Int. J. Mol. Sci.*, vol. 17, no. 2, p. 218, Feb. 2016.

[28] M. M. Nishat, F. Faisal, I. J. Ratul, A. Al-Monsur, A. M. Ar-Rafi, S. M. Nasrullah, M. T. Reza, and M. R. H. Khan, "A comprehensive investigation of the performances of different machine learning classifiers with SMOTE-ENN oversampling technique and hyperparameter optimization for imbalanced heart failure dataset," *Scientific Program.*, vol. 2022, pp. 1–17, Mar. 2022.

[29] Y. Wang, "A proof of goldbach conjecture by mirror prime decomposition," *WSEAS Trans. Math.*, vol. 21, pp. 563–571, Jul. 2022.

[30] T. Akiba, S. Sano, T. Yanase, T. Ohta, and M. Koyama, "Optuna: A next-generation hyperparameter optimization framework," in *Proc. 25th ACM SIGKDD Int. Conf. Knowl. Disc. Data Mining*, 2019, pp. 2623–2631.

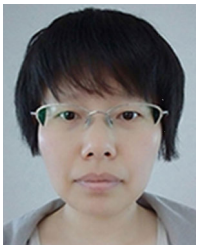
[31] X. Chen, X. Qiu, Y. Ma, L. Wang, and L. Fang, "Boruta-XGBoost electricity theft detection based on features of electric energy parameters," *J. Phys., Conf. Ser.*, vol. 2290, no. 1, Jun. 2022, Art. no. 012121.

[32] T. Chen and C. Guestrin, "XGBoost: A scalable tree boosting system," in *Proc. 22nd ACM SIGKDD Int. Conf. Knowl. Discovery Data Mining*, Aug. 2016, pp. 785–794.

[33] N. Javaid, A. Almogren, M. Adil, M. U. Javed, and M. Zuair, "RFE based feature selection and KNNOR based data balancing for electricity theft detection using BiLSTM-LogitBoost stacking ensemble model," *IEEE Access*, vol. 10, pp. 112948–112963, 2022.
[34] Z. Qu, H. Li, Y. Wang, J. Zhang, A. Abu-Siada, and Y. Yao, "Detection of electricity theft behavior based on improved synthetic minority oversampling technique and random forest classifier," *Energies*, vol. 13, no. 8, p. 2039, Apr. 2020.
[35] E. U. Haq, C. Pei, R. Zhang, H. Jianjun, and F. Ahmad, "Electricity-theft detection for smart grid security using smart meter data: A deep-CNN based approach," *Energy Rep.*, vol. 9, pp. 634–643, Mar. 2023.
[36] N. Javaid, S. Javaid, M. Asif, M. U. Javed, A. S. Yahaya, and S. Aslam, "Synthetic theft attacks and long short term memory-based preprocessing for electricity theft detection using gated recurrent unit," *Energies*, vol. 15, no. 8, p. 2778, Apr. 2022.

SANYUAN ZHU received the bachelor's degree in management information system from Beihang University, in 1993, and the master's degree in computer application from Wuhan University, in 2004. He is currently an Associate Professor with Hubei Engineering University. His research interests include artificial intelligent and big data analysis and processing.

ZIWEI XUE was born in 1998. He received the B.S. degree in software engineering from the College of Technology, Hubei Engineering University, Xiaogan, China, in 2020. He is currently pursuing the M.S. degree in computer technology with Hubei University, Wuhan, China. He is committed to the research of artificial intelligence in electricity theft detection.



YOUFENG LI received the Ph.D. degree in control science and engineering from Zhejiang University, Hangzhou, China, in 2010. She has been a Lecturer with Hubei Engineering University, since 2015. Her research interests include modeling and control of the industrial process, computing intelligence, and ML.