energies

MDPI

Article

Blending-Based Ensemble Learning Low-Voltage Station Area Theft Detection

Dunchu Chen [1,2,*], Wenwu Li [1,2] and Jie Fang [1]

[1] College of Electrical Engineering and New Energy, China Three Gorges University, Yichang 443002, China
[2] Hubei Key Laboratory of Cascaded Hydropower Stations Operation and Control, China Three Gorges University, Yichang 443002, China
* Correspondence: 2025cdc@ctgu.edu.cn

Abstract: In order to improve the efficiency of electricity theft detection, the power theft detection area and users should be better integrated, we proposed a Blending ensemble learning electricity theft detection model based on the Base Learner Selection Strategy (BLSS). Firstly, the adaptive synthetic (ADASYN) sampling method is used to process the unbalanced power consumption data, and the sample distribution of training data is balanced. Secondly, the BLSS selection method is used to screen the optimal base learner combination and construct the Blending ensemble learning model. Then, based on the historical data, the model makes a short-term prediction of the power consumption of the station area the next day, and focuses on the verification of the suspected energy-stealing station area where the Root Mean Square Percentage Error (RSPE) exceeds the threshold, so as to lock in the potential energy stealing users. Finally, through the comparison and verification of real examples, the search scope for electricity theft inspections was reduced by 79.17%, greatly improving the detection efficiency of the power supply company. At the same time, the model's electricity theft detection and recognition accuracy rate can be as high as 97.50%. The Blending ensemble learning electricity stealing detection model based on the BLSS base learner selection method has strong electricity stealing detection and recognition ability.

Keywords: blending combination strategy; ensemble learning; electricity stealing detection; unbalanced data

Academic Editor: Ahmed Abu-Siada

Received: 9 December 2024
Revised: 23 December 2024
Accepted: 24 December 2024
Published: 25 December 2024

Citation: Chen, D.; Li, W.; Fang, J. Blending-Based Ensemble Learning Low-Voltage Station Area Theft Detection. *Energies* **2025**, *18*, 31. https://doi.org/10.3390/en18010031

1. Introduction

Electricity is an indispensable energy source in social production and life, which is provided to consumers by power companies. In return, the power company gains in the form of money. In the process of scheduling energy from the power generation side to the consumption side, two types of power losses, technical loss (TL) and non-technical loss (NTL), will be generated [1]. Studies show that NTL caused by electricity theft can account for up to 10–40% of total power losses in developing regions, resulting in billions of dollars in economic losses annually. The malicious attack of electricity stealing users on power data is the main cause of NTL, which causes great losses to the economic benefits of power companies. With the continuous transformation of economic structure, social electricity demand and power load will continue to maintain rapid growth in the next few years. In the face of unknown power theft attacks, improving the accuracy of energy theft detection and protecting energy security are urgent problems to be solved.

Traditional theft detection methods rely on electrical models or physical measurements, such as analyzing voltage, current, and power factor anomalies. While these methods

achieve high accuracy under controlled conditions, they require substantial infrastructure investment, limiting their scalability. In contrast, data-driven approaches leverage advanced metering infrastructure (AMI) and historical electricity consumption data to identify abnormal patterns, providing cost-effective solutions. More importantly, current data-driven methods often focus solely on user-level detection, neglecting the broader context of station area-level abnormalities, leading to higher inspection costs and incomplete analysis.

To address the above challenges, this paper combines station-level detection with user-level identification and proposes a two-stage Blending ensemble learning model based on base learner selection strategy (BLSS) optimization. This model combines macro anomaly detection with micro user behavior analysis to improve detection efficiency while taking into account detection accuracy, thereby reducing the operating costs of power supply companies.

1.1. Related Work

At present, the electric larceny detection technology is mainly divided into two kinds: the electric larceny detection technology based on electrical model and measurement equipment and the data-driven electric larceny detection technology [2]. Among them, the electric larceny detection technology based on electrical model and measurement equipment generally relies on detecting the abnormal changes of the user's power consumption and its electrical characteristics such as voltage, current, and power factor [3] to determine whether there is electric larceny, but it puts forward higher accuracy requirements for infrastructure and increases the cost. With the gradual improvement of the advanced measurement infrastructure of big data and the automation process of collecting users' electricity consumption data, the data-driven electricity theft detection technology has become the most widely used electricity theft detection technology in recent years. The use of electricity theft sample set and historical electricity consumption data to train classifier model for abnormal electricity consumption characteristics in distribution network to identify suspected electricity theft [4–7] effectively reduces the inspection cost of power supply companies.

Electricity theft detection is essentially a binary classification problem. In recent years, many scholars have used ensemble learning methods to study data-driven electricity theft detection and achieved good results. Reference [8] proposed a detector based on ensemble learning and deep learning to detect error readings in measurement facilities in real time and lay a good data foundation. Reference [9] proposed a method of stealing electricity detection based on improved rotation forest algorithm, using the method of secondary sampling to improve the difference between base learners. Reference [10] proposed to train the isolated forest after optimizing the isolated features and perform Bagging secondary weighted ensemble, which effectively improved the performance of electricity theft detection without labels. Reference [11] used the information recorded by smart meters to deeply analyze customers' consumption behavior, and used extreme gradient boosting (XGBoost) supervised learning algorithm to detect abnormal electricity consumption behavior. Reference [12] used the Stacking integration strategy to detect electricity stealing users, which improved the detection accuracy of electricity stealing users, but did not consider from the station area, which increased the inspection cost. In [13], a multi-model ensemble method based on deep learning is proposed to capture abnormal electricity consumption patterns in smart grids.

Scholars at home and abroad have made different improvements to the data-driven electricity theft detection technology, but few studies have been conducted from the per-

spective of the distribution station area, and the station area detection and user detection have not been fully combined, as the analysis is not comprehensive enough.

1.2. Contributions

The main contributions of this paper are reflected in the following four aspects:

(1) Proposed Methodology:

We propose a novel Blending ensemble learning model for electricity theft detection, optimized using a Base Learner Selection Strategy (BLSS). The proposed methodology systematically evaluates the classification ability and diversity of base learners to construct an ensemble model with superior generalization performance.

(2) Addressing Data Imbalance:

To handle the inherent imbalance in electricity consumption data, the ADASYN algorithm is applied to generate synthetic samples for minority classes. This ensures a balanced dataset, enhancing the robustness of the proposed model, especially for detecting electricity theft scenarios with low occurrence rates.

(3) Model Validation and Results:

Using a processed Irish electricity user dataset, the Blending ensemble model demonstrates superior accuracy in detecting various theft patterns. Experimental results show that the proposed approach achieves 97.50% theft detection accuracy, outperforming traditional ensemble methods. Moreover, the two-stage detection framework combines station-level anomaly detection with user-level behavior analysis, significantly reducing the search space of suspicious users by 79.17%. This practical contribution improves the operational efficiency of power supply companies and mitigates the economic losses caused by power theft.

Table 1 illustrates the relationship between the contributions of an article and its sections. The contents of this paper are as follows:

Table 1. Article Organization.

Key Contribution	Corresponding Section
Proposed Methodology	Section 4: Blending ensemble learning model establishment
Addressing Data Imbalance	Section 2.1: ADASYN adaptive synthetic sampling
Model Validation and Results	Section 5: Experimentation and Results

In Section 2, we present the treatment of the data imbalance problem in this paper. The Blending ensemble learning model for selecting base learners based on BLSS method is proposed. Section 3 describes the data used in this paper and the proposed strategy. In particular, in Section 3.2, we describe how to detect energy theft in buses and customers. In Section 4, we build the Blending model for energy theft detection. The results are discussed and compared in Section 5. Finally, in Section 6, we give the main conclusions and discuss future work.

2. Introduction to Related Theories

2.1. ADASYN Adaptive Synthetic Sampling

In practice, the percentage of electricity theft users is relatively small, and the vast majority of electricity consumption datasets suffer from data imbalance. In order to achieve the balance between the two types of samples and improve the utilization of the dataset of electricity theft users, this paper adopts the ADASYN algorithm. Based on the sample situation using weight distribution, different numbers of minority class samples are generated

for different minority class samples, which improves the comprehensive performance of the model in the classification edge region [14].

The specific process is as follows:

Step 1: Calculate the number of samples to be synthesized as shown in Equation (1):

$$G = (m_l - m_s) \times \beta \tag{1}$$

where G is the number of samples to be synthesized, m_l is the number of samples in the majority category, and m_s is the number of samples in the minority category. β is a random coefficient between [0, 1]. If β = 1, the two categories are in equilibrium.

Step 2: Calculate the ratio r_i, i.e., the ratio of the number of near-neighboring points of normal users to the number of near-neighboring points of energy theft users among the K near-neighboring points of energy theft users, as shown in Equation (2):

$$r_i = \frac{\Delta_i}{K} \tag{2}$$

where Δ_i is the number of normal customers in the K neighborhoods of the electricity theft customer. i = 1, 2, 3, ..., m_s

Step 3: Normalize r_i as shown in Equation (3):

$$\hat{r}_i = \frac{r_i}{\sum_{i=1}^{m_s} r_i} \tag{3}$$

Step 4: Calculate the number of additional electricity theft samples that need to be added for each electricity theft sample, as shown in Equation (4):

$$g = \hat{r}_i \times G \tag{4}$$

Step 5: Randomly select one of the few categories of the K nearest neighbor points of the energy theft sample and synthesize the sample based on g as shown in Equation (5):

$$s_i = x_i + (x_{zi} - x_i) \times \lambda \tag{5}$$

where s_i is a synthetic sample, x_i is the ith sample in the electricity theft category sample. x_{zi} is a randomly selected minority class sample from the K nearest neighbors of x_i. λ is a random coefficient between [0, 1].

In the actual distribution bus, the proportion of normal users is much larger than the proportion of electricity theft users, and the imbalance of user data categories will make the machine learning detection biased. In this paper, we choose the adaptive synthetic sampling method, so that the two types of users' electricity consumption data will achieve balance.

2.2. Blending Ensemble Learning

Ensemble learning accomplishes learning tasks by constructing and combining multiple learners, often resulting in significantly superior generalization performance over a single learner [15]. The combining strategies for models are broadly categorized into averaging, voting, and learning methods. Among them, the learning method is a more powerful combining strategy in ensemble learning, where a meta-learner is chosen to fuse the results of different base learners. Stacking ensemble learning method and Blending ensemble learning method are typical representatives of the learning method. The base learner and meta-learner of Stacking ensemble learning use the same training set during training, which is prone to raising the risk of information leakage. The Blending ensemble

learning will set aside a set for the primary data, which is exclusively used to train the meta-learner, and can effectively protect privacy security [16].

Blending consists of multiple independent classifiers at two levels (level 0 and level 1), the base learner and the meta-learner, respectively. As an efficient combining strategy, the basic principle is that the original data is first divided into training set, validation set, and test set according to the proportion. The validation set is predicted using the trained base learner model, and the prediction results are used as new training data to train the meta-learner model. The test set is predicted using the base learner model and the prediction results are used as the meta-learner test data. Finally, the final Blending model classification results are output by the meta-learner as shown in Figure 1.

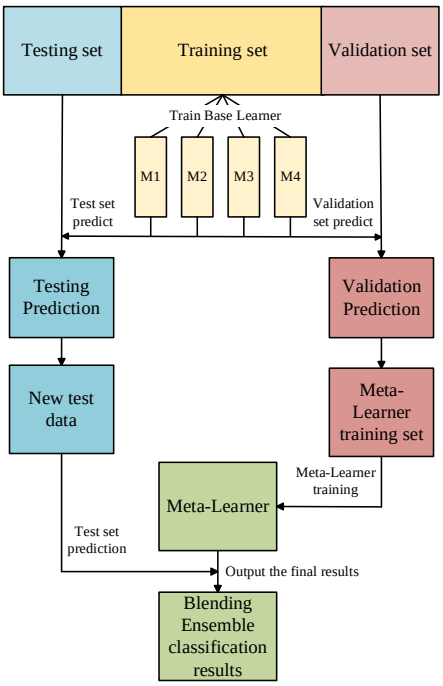


Figure 1. Blending combination strategy schematic diagram.

2.3. BLSS Selection Metric Method

When selecting base learners, individual learners should follow the principle of "good but different", i.e., base learners should not only have strong comprehensive classification ability, but also have certain differences among learners. The more accurate and diverse the individual learners are, the better the integration effect will be. In order to select the primary learner combinations with excellent performance and large differences as the basis of the proposed antitheft model in this paper, this paper proposes a base learner selection strategy, which initially evaluates the impact of various primary learner combinations on the performance of the Blending strategy, and selects the optimal combinations of base learners. The specific steps of this strategy are described as follows:

Step 1: Train the base learner to obtain preliminary results.

The base learner is trained by the processed Irish user electricity usage dataset, and the evaluation indexes of the training results include AUC, ACC, F1_score, AUC-PR, and training time T.

Step 2: Evaluate the comprehensive classification ability of the base learner.

In order to comprehensively evaluate the classification ability of the base learner, the evaluation metrics of the base learner in Step 1 are multiplied together as a criterion to measure the classification performance of the base learner, as shown in Equation (6):

$$C(m) = \mu(m) \times \varphi(m) \times \delta(m) \times \gamma(m) \tag{6}$$

where $\mu(m)$, $\varphi(m)$, $\delta(m)$, and $\gamma(m)$ denote the AUC, ACC, F1_score, and AUC-PR of the base learner, respectively. $C(m)$ is the product of the four larger values which indicate better classification ability of the base learner.

Therefore, multiplying the evaluation metrics can more clearly represent the classification ability of the base learner.

Step 3: Quantify the diversity of base learners.

In this paper, we use the pairwise Cohen's Kappa metric [17] to quantify the degree of diversity between base learners i and j through K, as shown in Equation (7):

$$K(i,j) = \left| \frac{Po - Pe}{1 - Pe} \right| \tag{7}$$

where i and j denote the intermediate values of the number of samples predicted correctly or incorrectly by calculating the base learner and the number of samples predicted correctly or incorrectly, as shown in Equations (8) and (9):

$$Po = \frac{x_1 + x_4}{x_1 + x_2 + x_3 + x_4} \tag{8}$$

$$Pe = \frac{(x_1 + x_3)(x_1 + x_2) + (x_2 + x_4)(x_3 + x_4)}{(x_1 + x_2 + x_3 + x_4)^2} \tag{9}$$

where x_1, x_2, x_3, x_4 represents the two classifier results for i and j, respectively, as shown in Table 2.

Table 2. Confusion matrix.

Learner Classification Results	*i* Correct Classification	*i* Error Classification
j Correct classification	x_1	x_3
j Error classification	x_2	x_4

The Kappa value is positively related to the diversity between pairs of base learners. Through quantitative comparison, selecting base learners with high degree of diversity is conducive to improving the performance of the Blending algorithm.

Step 4: Calculate the BLSS value of the combination of base learners.

Following the principle of "good but different", better classification performance, shorter training time, and greater diversity are the characteristics of a good combination of base learners. Among them, the training time used directly in the denominator of the BLSS formula may overly reduce the scores of the base learners and thus be too sensitive to the training time. By taking the natural logarithm of the training time, the effect of the training time is smoothed and transformed so that each base learner receives an appropriate tradeoff. Thus, the factors mentioned in steps 1, 2, and 3 are combined to calculate the degree of impact on the performance of the Blending algorithm as shown in Equation (10):

$$BLSS(m1, m2) = \frac{[C(m1)gC(m2)]gK(m1, m2)}{\ln[T(m1) + T(m2)]} \tag{10}$$

where $m1$ and $m2$ are different base learners; BLSS denotes the degree of influence of the combination of base learners on the performance of the Blending algorithm. The larger the

value of BLSS, the greater the contribution of the combination of base learners to improve the performance of the Blending algorithm. Here, we will give an example, such as KNN and LR. First, we need to train the base learners KNN and LR, and record their AUC, ACC, F1_score, AUC-PR, and training time T. Multiply the first four indicators to represent the comprehensive classification ability of the two base learners. Then calculate the Kappa index between the two base learners, and in the last step, calculate it through the BLSS index calculation formula to get the final answer.

3. Energy Theft Bus Detection and User Identification

3.1. Data

The experimental dataset was released by the Irish Electricity and Sustainable Energy Authority in January 2012, including the electricity usage report of more than 6000 Irish consumers for 535 days from 2009 to 2010 [18]. The sampling time scale was 30 min and the unit was kWh. After the dataset was processed with missing values and outliers, the electricity data of 1000 residential users were selected for experiments. Once all consumers in the dataset agree to participate in this study, it is assumed that they are all honest consumers, that is, no electricity stealing behavior is carried out.

Since data usually comes from actual business, there may be missing or noisy data due to measurement device failure, line maintenance, network interruption, etc., which may lead to unreliable data analysis results. In order to carry out subsequent work more effectively, it is necessary to clean, denoise, pre-analyze, and carry out other preliminary processing of the original data.

For missing records in the data, the processing methods usually include deleting missing data, completing data, and not processing. If there is a large amount of relevant data in the sample to be mined, directly deleting the missing data can quickly meet the established requirements without affecting the internal structure of the data and the final results; but when the sample data is small, it may cause insufficient sample size and even change the original distribution, resulting in inaccurate analysis results. When the amount of missing data is small, it can be manually filled by business experts based on rich experience; however, when the amount of missing data is large, the workload of manual filling is huge, hence the automatic filling method should be used at this time.

For outliers in the dataset, we use three methods: replacing outliers with the average of the observations before and after the data, directly deleting outlier data, and treating outliers as missing values and using the missing value processing method.

In order to ensure the sufficient amount of electricity stealing data, we randomly selected 10% of the electricity usage records of Irish users and modified them as electricity theft samples. Six common electricity theft modes are as follows:

(1) Reducing the amount of electricity x^t in period t according to a fixed ratio α can be achieved through undervoltage method, undercurrent method, and differential expansion method.

$$h_1(x^t) = \alpha x^t, \alpha \epsilon (0.2, 0.8) \tag{11}$$

(2) According to the random threshold γ, the power x^t is reduced, and the power above the threshold γ is fixed to γ, which can be achieved by the expansion method.

$$h_2(x^t) = \begin{cases} x^t, x^t \leq \gamma \\ \gamma, x^t > \gamma \end{cases} \tag{12}$$

(3) Setting all electric quantities x^t in a random time period (t_1, t_2) to 0 can be achieved through the undervoltage method, undercurrent method, and meter less method.

$$h_3(x^t) = \begin{cases} 0, t_1 < t < t_2, t_2 \geq t_1 + 4 \\ x^t, \text{else} \end{cases} \quad (13)$$

(4) According to the random threshold γ, the power x^t at any time can be set to 0, which can be achieved by undervoltage method, undercurrent method, and no meter method.

$$h_4(x^t) = max\{x^t - \gamma, 0\} \quad (14)$$

(5) This can be achieved by reversing the electricity usage sequence and placing the electricity x^t in the low-price period through the phase shift method.

$$h_5(x^t) = x^{49-t} \quad (15)$$

(6) Taking the average value of power consumption can be achieved through undercurrent method, undervoltage method, and phase shift method.

$$h_6(x^t) = mean(x) \quad (16)$$

We randomly generate six kinds of electricity stealing data according to the six kinds of electricity stealing modes found so far, and form a sample database of electricity stealing. The generated six kinds of electricity stealing data are mixed with normal data respectively, and a total of six mixed datasets of ETD1, ETD2, ETD3, ETD4, ETD5, and ETD6 are obtained. At the same time, the stealing data and the normal data are randomly selected from the stealing sample library and the normal sample library for mixing, and the MIX mixed dataset is obtained (that is, it contains six kinds of stealing data). The above seven datasets, in which all the data of each dataset are divided, are according to 60% of the training set, 20% of the validation set, and 20% of the test set. ADASYN resampling strategy is used to generate samples of electricity theft category, so that the sample size of positive and negative categories is balanced. In this paper, the effectiveness of the proposed energy stealing detection method is verified by the dataset containing six kinds of energy stealing modes.

3.2. Energy Theft Bus Detection and User Identification

In the power system, a bus is the power supply range or area of a transformer. As the core of power supply to power users and the "brain" controlling the low-voltage power, the abnormal power consumption of users in the bus will lead to changes in the overall power consumption of the bus. Therefore, in this paper, we first start from the distribution system bus, and utilize the processed Irish user electricity consumption dataset in Section 2.1 to assign the corresponding number of users to each bus. Through Blending ensemble learning to predict the power consumption of the station area, the root mean square error between the predicted value and the true value is calculated to narrow down the scope of abnormal power consumption detection buses. The specific process is shown in Figure 2.

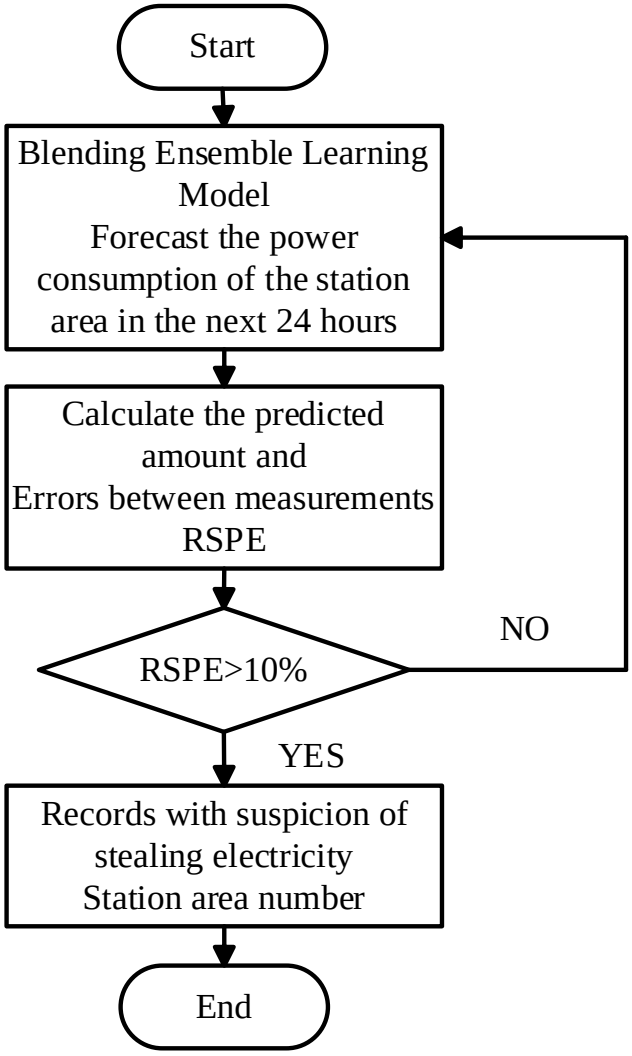


Figure 2. The detection flow chart of stealing radio area.

In order to accurately measure the gap between predicted and measured values, RSPE is defined in Equation (11). For the transformers with error values greater than 10% (value defined from analyses in different simulations of energy thefts) were detected as suspected energy theft customers.

$$RSPEk = \frac{\sum_{t=1}^{48} \sqrt{\left(P_{k,t}^{pred} - P_{k,t}^{true}\right)^2}}{P_{k,t}^{pred}} * 100 \quad (17)$$

where k denotes the number of the bus, $P_{k,t}^{pred}$ denotes the predicted active power at time t in bus k, $P_{k,t}^{true}$ denotes the actual active power at time t in bus k.

In the case of a given transformer, the results obtained from Blending prediction are compared with the total energy consumption of the smart meter, and after the detection of energy theft at the transformer level, the next work focuses on identifying the users attached to the suspected energy theft transformer. In this paper, we propose an anti-stealing early warning model based on Blending ensemble learning, and its model flowchart is shown in Figure 3. Firstly, the ADASYN method is used to deal with the problem of uneven ratio between the data of energy theft users and normal users, and then, this paper utilizes the BLSS metric proposed in Section 2.3 to select the best combination of base learners,

which is combined with the best-performing meta-learner for the identification of energy theft users.

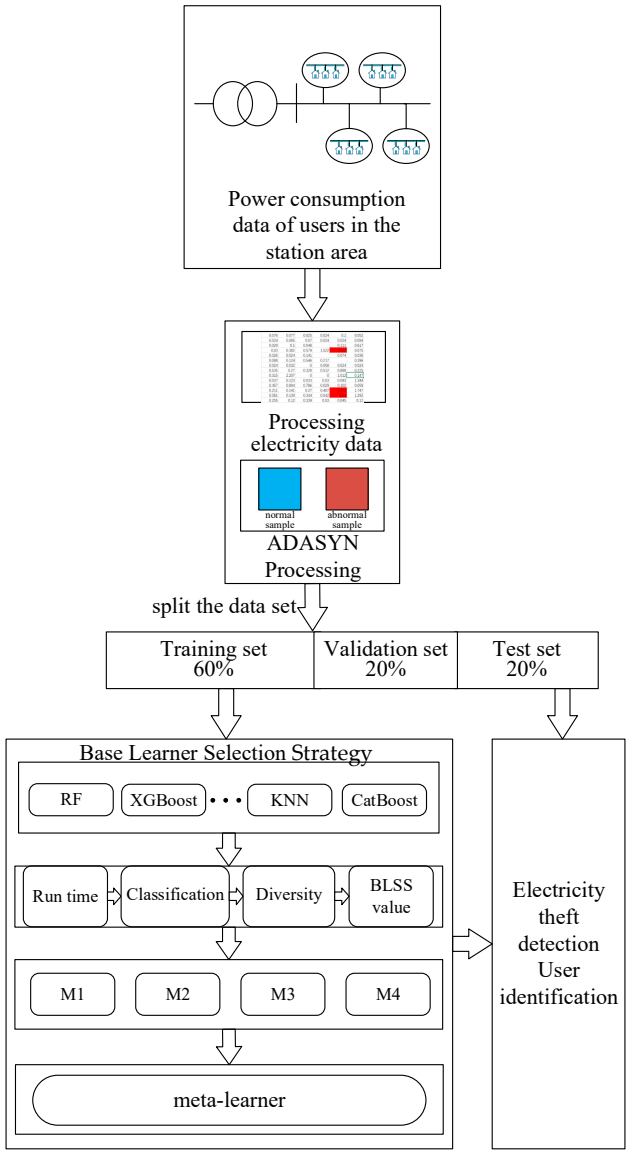


Figure 3. Blending model architecture diagram.

4. Blending Ensemble Learning Model Establishment

4.1. Selection of Base Learners

The selection of base learner combinations refers to the application of BLSS method, based on the comprehensive classification ability C, diversity K, and training time T of different base learner combinations. The BLSS values between different base learners are calculated, and the base learner combinations that are most beneficial to improve the performance of Blending ensemble learning are selected by pairwise comparison. This paper plans to select a total of 10 base learners from four single learners KNN, LR, DT, and SVM, and six ensemble learners RF, AdaBoost, GBDT, XGBoost, LightGBM, and CatBoost. Firstly, the Irish resampling dataset is used to train the above different base learners, and the average value is taken on the basis of the training result index of each base learner to balance the possible impact of different electricity stealing modes on each base learner. The training results included AUC, ACC, AUC-PR, and F1_score, as shown in Table 3.

Table 3. The training results of 10 base learners on seven mixed datasets.

Base Learner	ACC	AUC	AUC-PR	F1_Score
KNN	0.9411	0.8193	0.9609	0.9681
LR	0.9092	0.7830	0.9671	0.9523
SVM	0.6264	0.5036	0.9091	0.7608
DT	0.9574	0.9574	0.9523	0.9769
RF	0.9608	0.9569	0.9735	0.9789
AdaBoost	0.9172	0.9254	0.9706	0.9557
GBDT	0.9628	0.9661	0.9756	0.9798
XGBoost	0.9694	0.9753	0.9817	0.9811
CatBoost	0.9748	0.9776	0.9873	0.9864
LightGBM	0.9721	0.9730	0.9825	0.9842

Through the performance of each learner in Table 3, the base learner with better performance is selected as the object of BLSS selection strategy. In Table 3, AdaBoost, XGBoost, LightGBM, and CatBoost all belong to Boosting lifting algorithm, and CatBoost performs best in this experiment. Therefore, this paper determines CatBoost as a base learner candidate. In addition, AdaBoost, GBDT, XGBoost, LightGBM, and CatBoost all use DT as their base learner. Therefore, considering that the diversity between heterogeneous classifiers is often higher than that between homogeneous classifiers [19], the BLSS candidate base learner objects are determined to be KNN, LR, SVM, RF, and CatBoost. The BLSS values between each base learner are calculated according to the determined base learner, and the result matrix is generated, as shown in Table 4.

Table 4. The BLSS Result Matrix.

BLSS	KNN	LR	SVM	RF	CatBoost
KNN	1	1.2×10^{-4}	8.3×10^{-7}	8.8×10^{-4}	6.0×10^{-4}
LR	1.2×10^{-4}	1	1.7×10^{-5}	1.5×10^{-3}	6.1×10^{-4}
SVM	8.3×10^{-7}	1.7×10^{-5}	1	7.1×10^{-6}	7.4×10^{-7}
RF	8.8×10^{-4}	1.5×10^{-3}	7.1×10^{-6}	1	7.6×10^{-3}
CatBoost	6.0×10^{-4}	6.1×10^{-4}	7.4×10^{-7}	7.6×10^{-3}	1

From Table 4, it can be seen that the BLSS value calculated by the combination of SVM and any other base learner is lower than that of other base learner combinations, indicating that it does not have the ability to significantly improve the performance of the Blending algorithm; other base learner combinations can significantly improve the performance of the Blending algorithm. Therefore, the best combination of base learners selected is KNN + LR + RF + CatBoost.

4.2. Meta-Learner Selection

Traditional ensemble learning methods usually ignore the consideration of bias and variance, and only classify them according to simple base learners' serial or parallel integration. The composition of base learners is different, and the prediction results are different, each having its own advantages and disadvantages. The ensemble application has a certain improvement in prediction performance [20,21]. However, traditional ensemble learning ignores the correlation between base learners, resulting in limited performance of the model. At this time, a meta-learner with excellent performance will optimize the final Blending ensemble learning classification results and improve the performance of the algorithm [22]. After the base learner has completed the selective integration, this paper still trains the above-mentioned 10 learners through the Irish user electricity dataset to select a secure meta-learner to obtain the final average ACC value, AUC value, and F1_score value of the model based on the Blending integration strategy. The results are shown in Table 5.

Table 5. The training results of different learners as meta-learners on seven mixed datasets.

Meta-Learner	ACC	AUC	F1_Score
Blending-KNN	0.9504	0.8904	0.9743
Blending-LR	0.9442	0.8965	0.9858
Blending-DT	0.9638	0.9060	0.9856
Blending-SVM	0.8433	0.9031	0.9058
Blending-RF	0.9561	0.9596	0.9759
Blending-GBDT	0.9537	0.9315	0.9746
Blending-AdaBoost	0.9616	0.9439	0.9744
Blending-XGBoost	0.9647	0.9511	0.9761
Blending-CatBoost	0.9730	0.9540	0.9857
Blending-LightGBM	0.9750	0.9606	0.9862

After selective integration of base learners, KNN, LR, DT, SVM, RF, AdaBoost, GBDT, XGBoost, CatBoost, and LightGBM are tested as meta-learners, respectively. Except for SVM, the ACC and F1_score values of the overall model after Blending fusion exceed 0.94, which shows that the meta-learner can integrate the advantages of each learner. The goal of electricity theft detection is to automatically detect and identify abnormal parts from the dataset that are different from most of the data [23]. As a learner that finally integrates the classification results of the base learner, the meta-learner should make the base learner fully demonstrate its excellent performance [24]. Table 5 shows that when LightGBM is used as a meta-learner, its ACC value, F1_score value, and AUC value are the highest, and the comprehensive performance of the Blending ensemble model is the best, which can fully improve the detection success rate.

5. Results

The data provided by the Irish Energy Agency are all normal electricity usage data and do not include electricity theft data. The data used in the following tests include normal electricity usage data and electricity theft data after processing the normal electricity usage data.

5.1. Detection and Analysis of Energy Theft Bus

In the context of energy theft risk detection in a bus, energy theft by low voltage users can cause changes in the power consumption of that bus, and this section presents the results obtained from the step-by-step simulation of energy theft detection and identification.

From Figures 4 and 5, it can be seen that the difference between the predicted daily load curve of the station area and the daily load curve of the normal station area is not large. If the root mean square percentage error between the predicted electricity consumption and the actual electricity consumption exceeds the threshold, this indicates that the user connected to the back of the transformer may have abnormal conditions including electricity theft. By comparing the root mean square error percentage between the predicted power consumption and the actual power consumption of the station area with the threshold, the station area with the risk of stealing electricity is determined.

Table 6 shows that RSPE > 10% in buses 9 and 24. The final test results show that the electricity customers connected to the above area are suspected of stealing electricity. The search space was reduced from the original 24 transformers to 5 transformers, and the search range was reduced by 79.17%, which greatly improved the detection efficiency of power supply companies. The reduction in search space translates into tangible economic and operational benefits, resulting in lower inspection costs, allowing utilities to allocate resources more efficiently and reduce costs for field and equipment inspections by focusing inspections on high-risk areas. In addition, reducing the search space for potential power theft significantly reduces the time and manpower required for detection, enabling faster identification and mitigation of theft cases. Assuming an average site area loses

USD20,000 per year due to power theft, reducing the search space by 79.17% for 100 site areas could save up to USD1.58 million per year in inspection and operational costs. The savings help increase utility revenues, which can be reinvested in grid modernization and customer service enhancements.

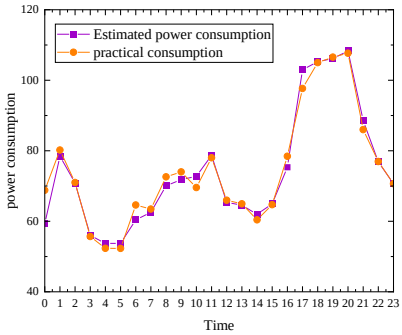


Figure 4. The predicted and the actual load curve of the normal station area.

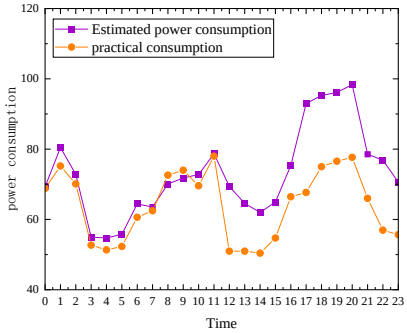


Figure 5. The predicted and the actual load curve of the stealing station area.

Table 6. RSPE values of different buses.

Bus Number	RSPE	Bus Number	RSPE	Bus Number	RSPE
1	3.89	9	0.87	17	11.72
2	2.41	10	2.78	18	4.09
3	1.32	11	2.86	19	15.63
4	5.47	12	3.23	20	0.59
5	0.95	13	30.54	21	2.53
6	0.19	14	7.69	22	8.79
7	4.11	15	4.27	23	0.49
8	13.18	16	1.22	24	12.14

Therefore, there is a suspicion of stealing electricity by locking the consumers connected to the five transformers. Aiming at the consumers connected behind the transformer, the Blending ensemble learning method is used for specific electricity theft detection.

5.2. Comparative Analysis

5.2.1. Robustness to Data Imbalance

To evaluate the robustness of the proposed Blending ensemble learning model under data imbalance, experiments were conducted using a dataset with different proportions of power theft data. For each case, the total number of samples was kept constant by applying oversampling (ADASYN) or not using sampling techniques to maintain consistency.

Table 7 summarizes the performance of the Blending electricity theft detection model under different data balances. The evaluation indicators include true negative rate (TNR) and false negative rate (FNR).

Table 7. Comparison of model applications before and after using the ADASYN algorithm.

Evaluation Indicators	TNR	FNR
Imbalanced dataset (ADASYN algorithm not used)	0.782	0.019
Balancing the dataset (using the ADASYN algorithm)	0.838	0.009

The results show that the proposed Blending model maintains excellent performance at all levels of data imbalance. It is worth noting that after the dataset is processed by the ADASYN sampling algorithm, the true negative rate of the Blending electricity theft detection model is significantly improved because the balanced dataset can prevent the model's classification results from being biased. This can be attributed to the effectiveness of the BLSS strategy in selecting diverse and complementary base learners and the adaptability of ADASYN to electricity theft detection.

These findings show that the proposed Blending model is robust and scalable, making it suitable for real-world scenarios where energy theft data often exhibits high imbalance.

5.2.2. Comparative Analysis with Ensemble Learners

Ensemble learners are usually composed of multiple weak learners, often able to outperform a single model in performance. Moreover, compared with some single models that need to adjust complex parameter sets, the ensemble learner is more robust to parameter selection. In practical engineering problems, ensemble learners usually have good adaptability to different types and dimensions of data. Therefore, in order to verify the effectiveness of the BLSS base learner metric method, the base learner combination selected by the BLSS metric method is compared with the average values of the six ensemble learners RF, AdaBoost, GBDT, XGBoost, LightGBM, and CatBoost mentioned above on the test sets of seven datasets. Following the principle of single variable, all models selected LightGBM as the meta-learner and used the Blending integration strategy. The experimental results are shown in Figure 6. Compared with the ensemble base learner, the base learner combination selected by BLSS has better performance.

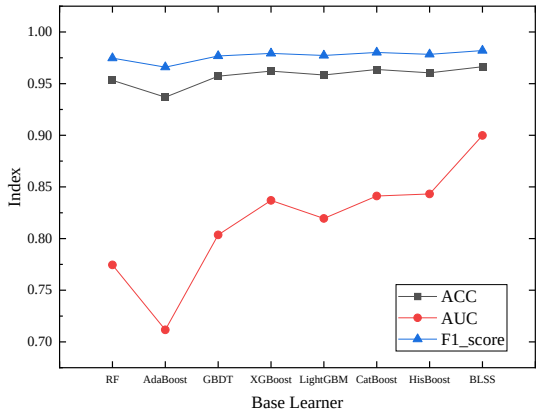
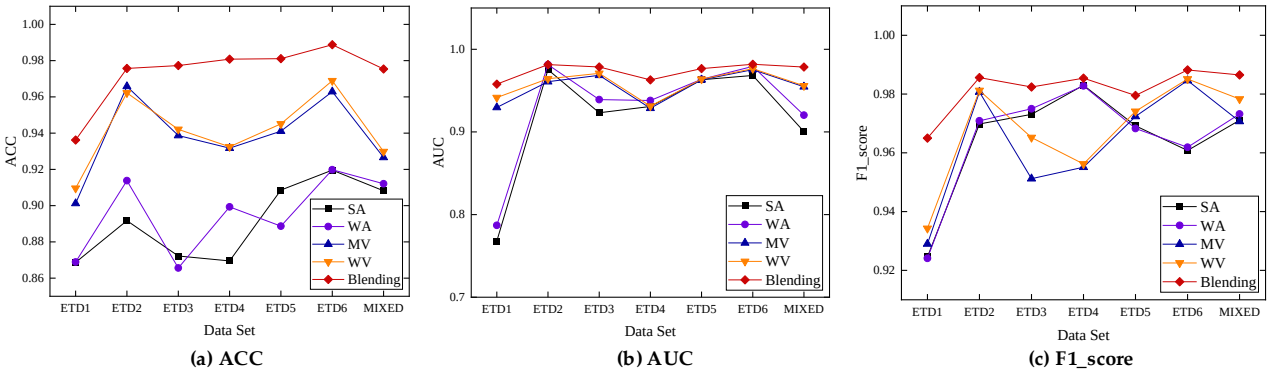


Figure 6. Comparison with ensemble learners.

The superior performance of the BLSS approach over other strategies can be attributed to its balanced consideration of classification ability, diversity, and training time in the base learner selection process. Metrics such as AUC, ACC, F1 score, and AUC-PR are used to ensure that the selected learner has strong classification ability. Meanwhile, the Kappa metric quantifies the diversity among learners, which is crucial for improving detection performance in scenarios with multiple power theft patterns. In addition, the logarithmic transformation of training time avoids over-penalizing complex but effective learners, achieving a balance between efficiency and performance.

5.2.3. Comparative Analysis with Ensemble Learning Combination Strategy

Since each energy stealing mode has different effects on model detection, in order to verify the overall performance of the Blending ensemble combination strategy in energy stealing detection, in this paper, the Blending integration strategy is compared with the ensemble learning methods using Simple Averaging (SA), Weighted Averaging (WA), Majority Voting (MV), and Weighted Voting (WV) as a combination strategy based on the datasets constructed by different energy stealing modes. In the experiment, the above four comparison models maintain the same stealing dataset as the Blending model for training and the same resampling strategy. The results are shown in Figure 7.



(a) ACC **(b) AUC** **(c) F1_score**

Figure 7. Comparison of Different Combination Strategies.

It can be seen from Figure 7 that for the dataset generated by each method of stealing electricity, the simple average method and the weighted average method are used as the traditional combination strategies of ensemble learning, which are in ACC, AUC, and F1_score: WV > MV > WA > SA, and the classification performance is slightly inferior to the voting method. In fact, in practical engineering applications, the application range of the voting method is also wider than that of the average method.

At the same time, Blending, as a typical representative of the learning method, on the basis of selective integration of base learners, uses the advantages of meta-learners to integrate different base learners, and has strong robustness to daily energy stealing methods. In the ACC, AUC, and F1_score, the performance of the three evaluation indexes is the best compared with the average method and the voting method. Therefore, this paper believes that the selected Blending combination strategy can also have good generalization ability in the face of unknown datasets, which is helpful to the power supply company in the detection of electricity theft.

In summary, aiming at the problem that the stealing station area and the stealing users are not fully combined in the process of stealing detection, the Blending ensemble learning stealing detection model based on BLSS base learner selection method proposed in this paper has the following advantages:

Firstly, in terms of data processing, taking into account the small proportion of energy-stealing users in practical applications, the ADASYN method is used to balance the power consumption data in actual projects to avoid affecting the classification results.

Secondly, in terms of model construction, it is the top priority of accurate model prediction to select good and different base learners for fusion through BLSS base learner selection method.

Thirdly, in terms of integration strategy, Blending ensemble learning fusion strategy will separate the set for the training of meta-learners, avoid the risk of information leakage, and improve the detection performance of the model.

6. Conclusions

This paper proposes a Blending ensemble learning electricity theft detection model based on the BLSS-based learner selection method. The Blending model makes a short-term prediction of the next day's electricity consumption in the station area based on historical data, and focuses on verifying the station area where the RSPE exceeds the threshold. The station area electricity theft detection step reduces the number of suspicious users by 79.17%, effectively improving the efficiency of electricity theft detection. By reducing the user-level detection workload by 79.17%, the approach provides significant economic and operational benefits, including reduced inspection costs, improved resource allocation efficiency, and increased grid stability. The experimental results show that the Blending ensemble learning electricity theft detection model based on the BLSS-based learner selection method can accurately identify electricity theft users with different electricity theft modes, which helps to reduce the economic losses of power supply companies. Previous studies often started directly from the algorithm, using data-driven technology to directly detect the user's electricity consumption, ignoring the relationship between the substation and the user, and consuming a lot of resources. Unlike traditional electricity theft detection methods that only rely on user-level anomaly analysis, this paper starts from the substation, first conducts a preliminary screening of the substation, and focuses on checking the users connected to the suspicious substation. The innovative use of the Blending ensemble learning method based on BLSS ensures higher accuracy and efficiency, especially when dealing with unbalanced datasets and diverse theft scenarios.

The main contributions of this study are as follows:

(1) Two-Stage Detection Framework:

The model employs a two-stage strategy, starting with station area-level detection to identify high-risk areas based on deviations between predicted and actual electricity consumption. This approach narrows the scope for user-level detection, reducing the workload by 79.17%. Compared to conventional user-level methods, this significantly improves detection efficiency and operational feasibility.

(2) Innovative Use of BLSS:

The BLSS method systematically evaluates and selects base learners based on classification performance, diversity, and efficiency. This ensures the optimal combination of base learners, enhancing the robustness and generalization of the Blending ensemble model, which outperforms traditional ensemble strategies.

In order to implement the proposed model in a realistic power system, some factors are suggested to be considered. First, ensure access to high-quality data from the Advanced Metering Infrastructure (AMI), including detailed electricity consumption records and system-level measurement data. Second, consider the calibration of thresholds, and customize RSPE thresholds and base learner configurations according to the specific characteristics of the regional power grid to maximize detection accuracy while minimizing false alarms. Finally, consider expanding the electricity theft detection model with other functions and combining it with non-intrusive load monitoring technology to improve detection accuracy.

The subsequent research team considers incorporating grid parameters (such as voltage, current, and power flow data) into the feature set for anomaly detection, and exploring graph-based models to analyze spatial and topological relationships within the grid, and evaluate the performance of the model under different network configurations and load scenarios. This will be a very meaningful research topic. At the same time, it is also necessary to explore whether the model can be extended to medium-voltage power systems to ensure that the model is applicable to different grid infrastructures and voltage levels.

Author Contributions: Conceptualization, W.L.; methodology, D.C.; investigation, J.F.; resources, W.L.; data curation, D.C.; writing—original draft preparation, D.C.; writing—review and editing, D.C.; visualization, J.F.; supervision, W.L. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Data Availability Statement: Data is contained within the article. The data presented in this study are available in the cited references.

Conflicts of Interest: The authors declare no conflicts of interest.

References

1. Yan, Z.; Wen, H. Performance Analysis of Electricity Theft Detection for the Smart Grid: An Overview. *IEEE Trans. Instrum. Meas.* **2022**, *71*, 1–28. [CrossRef]
2. Kim, S.; Sun, Y.; Lee, S.; Seon, J.; Hwang, B.; Kim, J.; Kim, J.; Kim, K.; Kim, J. Data-Driven Approaches for Energy Theft Detection: A Comprehensive Review. *Energies* **2024**, *17*, 3057. [CrossRef]
3. Kolade, A.O.; Adetokun, B.B.; Oghorada, O. Energy Theft Detection in Power System Network: Reviews of Studies on Machine Learning Based Solutions. In Proceedings of the 2023 2nd International Conference on Multidisciplinary Engineering and Applied Science (ICMEAS), Abuja, Nigeria, 1–3 November 2023.
4. Qin, Y.; Xiang, S.; Chai, Y.; Chen, H. Macroscopic-microscopic attention in LSTM networks based on fusion features for gear remaining life prediction. *IEEE Trans. Ind. Electron.* **2020**, *67*, 10865–10875. [CrossRef]
5. Althobaiti, A.; Rotsos, C.; Marnerides, A. Adaptive Energy Theft Detection in Smart Grids Using Self-Learning with Dual Neural Network. *IEEE Trans. Ind. Inform.* **2024**, *20*, 2776–2786. [CrossRef]
6. Punmiya, R.; Choe, S. Energy Theft Detection Using Gradient Boosting Theft Detector with Feature Engineering-Based Preprocessing. *IEEE Trans. Smart Grid* **2019**, *10*, 2326–2329. [CrossRef]
7. Akram, R.; Ayub, N.; Khan, I.; Albogamy, F.R.; Rukh, G.; Khan, S.; Shiraz, M.; Rizwan, K. Towards Big Data Electricity Theft Detection Based on Improved RUSBoost Classifiers in Smart Grid. *Energies* **2021**, *14*, 8029. [CrossRef]
8. Abdulaal, M.J.; Ibrahem, M.I.; Mahmoud, M.M.E.A.; Khalid, J.; Aljohani, A.J.; Milyani, A.H.; Abusorrah, A.M. Real-Time Detection of False Readings in Smart Grid AMI Using Deep and Ensemble Learning. *IEEE Access* **2022**, *10*, 47541–47556. [CrossRef]
9. Liu, J.; Mei, Z.; Liu, M.; Zhou, H.; Dong, Q. Research on Electricity Theft Detection Based on Improved Rotation Forest Algorithm. *J. Electr. Power Sci. Technol.* **2024**, *39*, 93–104.
10. Li, G.; Lu, J.; Wang, Y.; Huang, R.; Lou, M. Isolated-Forest Electricity Theft Detection Algorithm Based on Bagging Secondary Weighted Ensemble. *Autom. Electr. Power Syst.* **2022**, *46*, 92–100.
11. Buzau, M.M.; Tejedor-Aguilera, J.; Cruz-Romero, P.; Gómez-Expósito, A. Detection of Non-Technical Losses Using Smart Meter Data and Supervised Learning. *IEEE Trans. Smart Grid* **2019**, *10*, 2661–2670. [CrossRef]
12. Pamir; Javaid, N.; Akbar, M.; Aldegheishem, A.; Alrajeh, N.; Mohammed, E.A. Employing a Machine Learning Boosting Classifiers Based Stacking Ensemble Model for Detecting Non Technical Losses in Smart Grids. *IEEE Access* **2022**, *10*, 121886–121899. [CrossRef]
13. Khan, I.U.; Javeid, N.; Taylor, C.J.; Gamage, A.A.K.; Ma, X. A Stacked Machine and Deep Learning-Based Approach for Analysing Electricity Theft in Smart Grids. *IEEE Trans. Smart Grid* **2022**, *13*, 1633–1644. [CrossRef]
14. He, H.; Bai, Y.; Garcia, E.; Li, S. ADASYN: Adaptive synthetic sampling approach for imbalanced learning. In Proceedings of the IEEE International Joint Conference on Neural Networks (IEEE World Congress on Computational Intelligence), Hong Kong, China, 1–8 June 2008.
15. Zhou, Z. *Machine Learning*, 8th ed.; Tsinghua University Press: Beijing, China, 2016; pp. 172–185.
16. Huan, J.; Li, D.; Du, Y.; Shen, X.; Zhang, X.; Qiao, B.; He, C.; Lan, X.; Luo, P. Mid-Term Forecasting of Real-Time Load Based on Prophet and Blending Ensemble Learning. *Electr. Power Autom. Equip.* **2024**, *44*, 178–183.
17. Wang, J.; Yang, Y.; Xia, B. A Simplified Cohen's Kappa for Use in Binary Classification Data Annotation Tasks. *IEEE Access* **2019**, *7*, 164386–164397. [CrossRef]
18. ISSDA. Data from the Commission for Energy Regulation. Available online: http://www.ucd.ie/issda/data/commissionforenergyregulationcer/ (accessed on 1 February 2020).
19. Kolárik, M.; Sarnovský, M.; Paralič, J. Diversity in Ensemble Model for Classification of Data Streams with Concept Drift. In Proceedings of the 2021 IEEE 19th World Symposium on Applied Machine Intelligence and Informatics (SAMI), Herl'any, Slovakia, 21–23 January 2021.
20. Cho, S.; Kim, S.; Choi, J. Transfer learning-based fault diagnosis under data deficiency. *Appl. Sci.* **2020**, *10*, 7768. [CrossRef]

21. Sowmya, C.S.; Vibin, R.; Mannam, P.; Mounika, L.; Kabat, S.R.; Patra, J.P. Enhancing Smart Grid Security: Detecting Electricity Theft through Ensemble Deep Learning. In Proceedings of the 2023 8th International Conference on Communication and Electronics Systems (ICCES), Coimbatore, India, 1–3 June 2023.
22. Sun, X.; Hu, J.; Zhang, Z.; Cao, D.; Huang, Q.; Chen, Z.; Hu, W. Electricity Theft Detection Method Based on Ensemble Learning and Prototype Learning. *J. Mod. Power Syst. Clean Energy* **2024**, *12*, 213–224. [CrossRef]
23. Hu, W.; Guo, Q.; Wang, W.; Wang, W.; Song, S. Research on User Loss Contribution Calculation of High-Loss Distribution Area Based on Transfer Entropy. In Proceedings of the 2022 China International Conference on Electricity Distribution (CICED), Changsha, China, 7–8 September 2022.
24. Hou, H.; Liu, C.; Wang, Q.; Zhao, B.; Zhang, L.; Wu, X.; Xie, C. Load Forecasting Combining Phase Space Reconstruction and Stacking Ensemble Learning. *IEEE Trans. Ind. Appl.* **2023**, *59*, 2296–2304. [CrossRef]