electronics

MDPI

Article

Intelligent Scene-Adaptive Desensitization: A Machine Learning Approach for Dynamic Data Privacy in Virtual Power Plants

Ruxia Yang [1,2,*], Hongchao Gao [3], Fangyuan Si [3] and Jun Wang [4]

[1] State Grid Smart Grid Research Institute Co., Ltd., Nanjing 210003, China
[2] State Grid Laboratory of Power Cyber-Security Protection and Monitoring Technology, Nanjing 210003, China
[3] State Key Laboratory of Power System, Department of Electrical Engineering, Tsinghua University, Beijing 100000, China; hcgaoth@tsinghua.edu.cn (H.G.); sifangyuan@mail.tsinghua.edu.cn (F.S.)
[4] State Grid Shanghai Municipal Electric Power Company, Shanghai 201507, China; wangjun@sh.sgcc.com.cn
* Correspondence: yangruxia@geiri.sgcc.com.cn

Abstract: In the context of virtual power plants (VPPs), the one-size-fits-all approach of traditional static desensitization methods proves inadequate due to the diverse and dynamic operational scenarios encountered. These methods fail to provide the necessary flexibility for varying data privacy requirements across different scenarios. To address this shortcoming, our research introduces a dynamic desensitization method specifically designed for VPPs. Leveraging machine learning for adaptive scene recognition, the method adjusts data privacy levels intelligently according to each unique scenario. A novel similarity utility function and a Gaussian processes-based differential privacy algorithm ensure tailored and efficient privacy protection. Experimental results highlight an 87.5% accuracy in scene recognition, validating our method's capability to adapt to diverse scenarios effectively. This study contributes to the field by providing a nuanced approach to data protection, effectively addressing the specific needs of complex VPP environments.

Keywords: virtual power plant; load identification; support vector machine; differential privacy; dynamic desensitization

check for updates

Citation: Yang, R.; Gao, H.; Si, F.; Wang, J. Intelligent Scene-Adaptive Desensitization: A Machine Learning Approach for Dynamic Data Privacy in Virtual Power Plants. *Electronics* **2024**, *13*, 1051. https://doi.org/10.3390/electronics13061051

Received: 8 January 2024
Revised: 5 March 2024
Accepted: 6 March 2024
Published: 12 March 2024

Copyright: © 2024 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (https://creativecommons.org/licenses/by/4.0/).

1. Introduction

Virtual power plants (VPPs) represent a paradigm shift in electrical energy management, leveraging advanced information and communication technologies to synergize the supply-demand dynamics of electrical loads virtually. Traditional power dispatch methods often grapple with delays due to obsolete information systems, suboptimal emergency response solutions, and resource inefficiencies [1,2]. In contrast, VPPs integrate comprehensive control platforms, enabling real-time, intelligent adjustments to the power resource supply-demand balance.

In conventional power dispatch processes, issues often arise, including delays in processing due to delayed information updates, challenges in deriving optimal solutions during emergencies, and unnecessary resource wastage. Virtual power plants, integrating comprehensive power control across the platform, enable effective, intelligent adjustments in the allocation of power resources at various instances. Managing resource dispatch across multiple scenarios involves the information of diverse users, making data privacy protection [3] a critical aspect of ensuring virtual power plant security. To effectively manage the allocation of power resources while simultaneously protecting user privacy, the technology of data desensitization can be employed. Data desensitization is a method of processing sensitive data, aimed at preventing the disclosure of user identities and private information during the processes of data sharing and analysis.

In practice, VPPs are required to manage a diverse array of business scenarios, ranging from routine power dispatch to emergency response, each presenting distinct requirements for data security and privacy. Applying a uniform security protocol to diverse user data can

lead to either redundant expenses or inadequate protection. Thus, targeted anonymization strategies are vital. These strategies, while ensuring precise privacy needs, also enhance system adaptability to dynamic environments, balancing data utility and privacy constraints effectively. This paper aims to delve deeper into these aspects, proposing a novel approach to dynamic anonymization in VPPs, bolstered by machine learning techniques for adaptive scenario recognition and privacy protection.

In the context of virtual power plants, the necessity for tailored security strategies becomes paramount given the varied confidentiality requirements of user data across different scenarios. Traditional uniform anonymization methods fall short of addressing the diverse data security needs across multiple scenarios. This paper addresses this gap by introducing a dynamic and adaptive anonymization approach, particularly suited to the nuanced needs of virtual power plant user data. The main contributions of our study include:

1. Proposing a dynamically adaptive anonymization method, which identifies and responds to different application scenarios in VPPs based on user characteristics, tailoring anonymization protection to match the specific security requirements of each scenario.
2. Implementing a differential privacy algorithm grounded in Gaussian processes which provides varying intensities of privacy protection across different scenarios. This approach ensures dynamic and effective anonymization of sensitive data, aligning with the specific privacy demands of each recognized scene.
3. Demonstrating through experimental validation that the proposed method, with an appropriately allocated privacy budget, can significantly enhance pattern recognition accuracy up to 87.5%. This improvement underscores the effectiveness of our differential privacy approach in aligning with dynamic anonymization strategies tailored for specific scenarios.

The structure of this paper is meticulously organized to facilitate a comprehensive understanding of our research. Section 2 provides a detailed review of related work, setting the context for our study. Section 3 describes the methodology, including the development of our dynamic anonymization method and the differential privacy algorithm. Section 4 presents the experimental setup and results, demonstrating the efficacy of our approach. Finally, Section 5 discusses the implications of our findings, followed by Section 6, which concludes the paper and outlines future research directions.

2. Related Work

In recent power market demand analysis research, a crucial step involves identifying and distinguishing load electricity usage patterns based on historical data. Extensive research on load identification algorithms has been conducted by numerous scholars globally. Yao, J. [4] applied data mining principles, using K-means and fuzzy C-means to develop a composite model for pattern recognition. This model facilitated dimension reduction, visual clustering, and plane clustering of supply data, leading to a predictive model for residential power grid loads in Shanghai. Chen, C. [5] improved the nearest neighbor classifier by integrating SVM and MC-kNN, enhancing the classification accuracy for identifying electrical appliances. Sun, S. [6] used a cascading network approach with CNN + LSTM, cross-verifying load identification with decomposition results, and formulated a model transfer method adaptable to varying scenarios.

Differential privacy protection has garnered significant interest for its robust, mathematically provable model and customizable protection levels. It has been extensively applied in power system privacy protection. Chen, R. [7] applied this to battery charging, developing a privacy-centric charging and discharging strategy. Xu, G. [8] introduced a distributed differential privacy algorithm for time-related time series data. Xiong P. [9] used differential privacy in electricity pricing, assessing privacy leakage and user electricity costs. Safeguarding users' electricity usage patterns and associated privacy is now a research priority.

This study distinguishes itself from [4–6] by conducting deeper scene recognition on datasets prior to differential privacy computation, enhancing dataset-specific applicability. Unlike the static desensitization methods in [7–9], our approach incorporates dynamic grading, allowing flexible data handling per varying security needs, hence offering greater adaptability and practical relevance. These enhancements render our methodology superior to existing ones. In virtual power plants post-desensitization, it is crucial to assess data accuracy, completeness, and protection. Desensitized data retain consistency with original data in terms of format, type, and range while safeguarding privacy, security, and traceability. Thus, appropriate differential privacy budgets must be chosen based on specific virtual power plant scenarios, ensuring data integrity and accuracy.

3. Basic Background Knowledge

3.1. K-Means Clustering Algorithm

The K-means algorithm is adept at efficiently processing and simplifying complex datasets, making it ideal for exploring and identifying inherent patterns within data. It is particularly effective for feature extraction in virtual power plant user datasets. K-means achieves this by organizing diverse electrical data into smaller, similar-feature clusters, thereby unveiling underlying patterns and structures [10]. The algorithm proceeds through several steps:

1. INITIALIZATION: K input objects are randomly selected as the initial centroids of the clusters.
2. ASSIGNMENT: Each remaining point is assigned to the cluster with the nearest centroid.
3. UPDATE: Calculate the mean of each cluster's points to update its centroid.
4. ITERATION: Repeat the ASSIGNMENT and UPDATE steps until centroids stabilize or a predetermined iteration count is reached.

K-means aims to minimize the within-cluster sum of squared distances, as expressed in the following objective function:

$$J = \sum_{i=1}^{k} \sum_{x \in S_i} |x - \mu_i|^2 \tag{1}$$

Here, k is the number of clusters, S_i is the point set in the i-th cluster, μ_i is the centroid of the i-th cluster, and x is the point in the cluster.

3.2. Support Vector Machine

Support Vector Machines (SVMs) stand out in machine learning for their robustness in nonlinear classification, largely due to their use of kernel functions. These kernels transform the original feature space into a higher-dimensional space where linear separation becomes feasible [11]. This transformation is critical in applications like power load classification, where data complexity often defies linear analysis.

The goal of SVM is encapsulated in the following optimization problem, seeking the minimum of the function:

$$f(w) = \frac{1}{2}\|w\|^2 \tag{2}$$

where w represents the weight vector, used to define the direction of the classification hyperplane; $\|w\|^2$ represents the sum of squares of each component, and the objective function is to minimize this norm, so as to maximize the edge of the hyperplane. By employing the Lagrange multiplier method, the solution can be obtained by satisfying the following conditions: $\alpha_i \geq 0, i = 1, \ldots, l$. Introducing Lagrange multipliers allows us to transform the problem into the following set of equations:

$$L = \frac{1}{2}\|w\|^2 - \sum_{i=1}^{l} \alpha_i y_i (x_i \cdot w + b) + \sum_{i=1}^{l} \alpha_i \tag{3}$$

Here, α_i represents the Lagrange multiplier, with each training sample having a corresponding multiplier for constraints in the optimization problem. y_i denotes the class label of the ith training sample, and x_i is the feature vector of the ith training sample. The term b serves as the bias component. Together with the weight vector w, it defines the classification hyperplane. The objective involves summing the products of all samples' constraints and their associated Lagrange multipliers. This sum reflects how the hyperplane accurately classifies each training sample.

The solution to this dual problem gives the optimal margin classifier. By introducing a kernel function $K(x_i, x_j)$, SVM effectively handles nonlinear data relationships, a key feature for complex power system datasets.

3.3. Differential Privacy

Differential privacy serves as a cornerstone in data privacy, ensuring individual data points remain confidential within aggregated datasets. Its significance is pronounced in smart grid systems, where consumer data sensitivity is high [12]. If there are two data sets D and D', where D is a proper subset of D', and D' has only one more piece of data than D, then D and D' are called adjacent data sets. For any two adjacent data sets D and D', given a certain randomization algorithm F, apply it to the adjacent data sets D and D', respectively.

If it can satisfy Formula (4), it is called algorithm M; it provides differential privacy protection. The core principle of differential privacy is quantified by the privacy budget, ε, and is mathematically represented as:

$$\Pr[M(D) \in S_M] \leq e^{\varepsilon} \times \Pr\left[M\left(D'\right) \in S_M\right] \tag{4}$$

Among them, S is the set of all possible output results of F, and Pr represents the probability that the data in D are inferred. ε is the privacy protection budget, which is the ratio of the probability that the differential privacy algorithm outputs the same result on two adjacent data sets. This equation signifies that the probability of any outcome from a dataset does not substantially increase with the modification of a single data point, thereby ensuring privacy.

An additional aspect of differential privacy is the concept of sensitivity, defined as the maximum change in the function output when a single record in the database is altered. Sensitivity is crucial in determining the amount of noise to be added to preserve privacy. The general formula for calculating sensitivity, Δf is:

$$\Delta f = \max_{D,D'} \|f(D) - f(D')\| \tag{5}$$

In the smart grid context, differential privacy techniques balance the need for data utility in load forecasting against the imperative of preserving individual consumer privacy.

4. System Architecture Design

4.1. Innovative Data Protection Strategy in Power Systems

This study introduces a groundbreaking data privacy framework for virtual power plants, emphasizing a differential privacy approach. This innovative framework addresses the pressing need for robust data protection in the power industry, especially considering the vast and diverse nature of the data involved [13].

Referencing Figure 1, our methodology commences with a comprehensive analysis and feature extraction from power load data, segregating privacy-sensitive attributes based on unique scenarios in electricity generation and consumption. This step is followed by the deployment of state-of-the-art AI algorithms for the scenario-based classification of data, utilizing these classifications in real-time applications through a novel similarity utility function. The framework then employs advanced differential privacy techniques in varying scenarios for the proactive protection of user data.

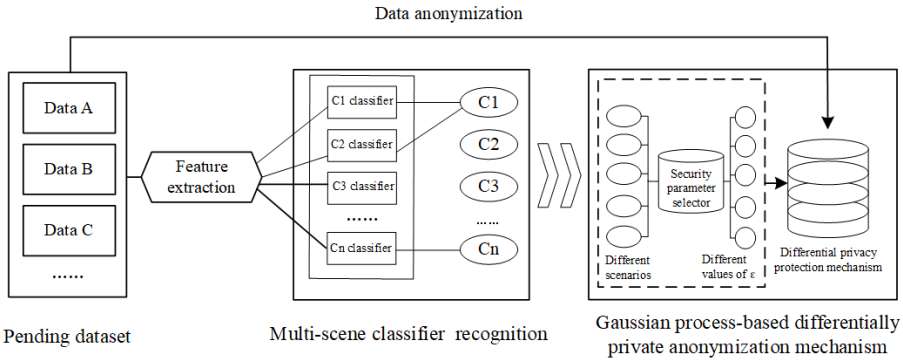

Figure 1. Schematic of advanced user data anonymization in varied operational contexts.

When dealing with substantial data volumes in virtual power plants, maintaining data privacy becomes complex. Here, our framework recommends employing finely-tuned noise mechanisms that balance privacy with data utility, adapting these mechanisms to suit the diverse data characteristics of different virtual power plant scenarios [14].

4.2. Analysis of Critical Data in Power Generation and Consumption

Our focus extends to the analysis of business service data flow in virtual power plants, segregating it into two pivotal scenarios: power generation and consumption, as shown in Figure 2. The study identifies various sensitive data types, including customer-specific information, business operational data, and location-based energy usage statistics. This classification method helps to gain a deeper understanding of how virtual power plants handle different types of private data and can be classified more clearly when performing feature extraction and scene discrimination.

For the power generation scenario, we consider: (1) Details on the output and capability of decentralized energy sources. (2) Comprehensive transaction data, including participants, timing, and energy quantities. (3) Forecasts related to energy production, market trends, and supply-demand correlations.

In contrast, the power consumption scenario involves: (1) Detailed analysis of user energy consumption patterns. (2) In-depth transactional data, including energy type, volume, and pricing. (3) Predictive analytics focusing on future consumption trends and market price estimations.

4.3. Feature Extraction and Clustering of Virtual Power Plant Data

The virtual power plant platform can be categorized into power generation and electricity usage scenarios. The data features related to these activities are influenced by usage patterns, frequency, and external factors. The model for classifying these behaviors leverages various data types like individual electricity usage and total consumption, focusing on multi-dimensional data analysis and modeling to enhance the precision in categorizing target groups. Literature suggests that using the K-means algorithm for clustering based on characteristics like electricity consumption and usage timing can identify groups with similar consumption behaviors. This approach aids in recognizing different electric load patterns and learning the interrelationships between various scenarios, leading to more accurate scenario classification.

In the scenario classification model's development, features serve to distinguish between the different data attributes of scenarios. For instance, as shown in Table 1, the results from K-means clustering provide valuable inputs for classification and regression models, boosting their predictive accuracy.

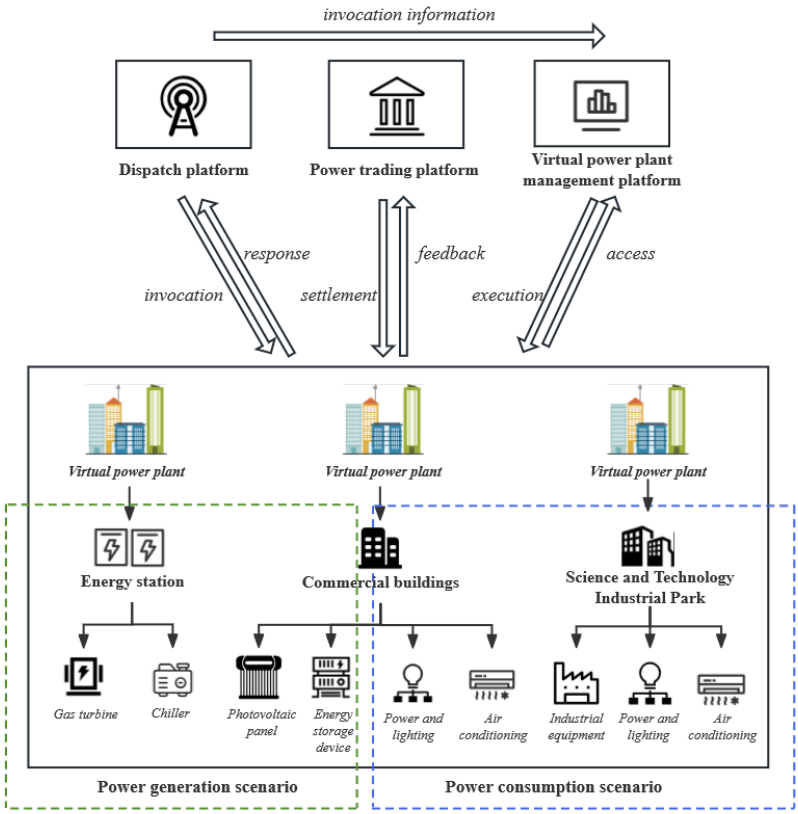

Figure 2. Strategic data segmentation in virtual power plants.

Table 1. Data formats in the electricity context.

Data Category	Example Data Points	Data Type
Customer Personal Data	Customer ID (e.g., 3455****123)	Character
	Name (e.g., Zhang San)	Character
	Address (e.g., xxx Province, xx City)	Character
Electricity Usage Data	Usage Data (e.g., 12.18 kWh)	Numeric
	Total Consumption (e.g., 11,553.65 kWh)	Numeric
Power Generation Data	Monthly Generation (e.g., 5784 kWh)	Numeric
Settlement Information	Electricity Unit Price (e.g., 0.79 $)	Numeric

K-means clustering outcomes can be effectively utilized as input features for classification and regression models, thereby boosting their predictive accuracy and strength. Feature extraction via K-means clustering on datasets offers a streamlined and coherent perspective for subsequent data analyses. In the virtual power plant platform, scenarios are bifurcated into electricity consumption and generation. The characteristics of data pertaining to power generation or usage are shaped by usage norms, frequency, and external factors. In developing classification identification models, the primary features include individual and total electricity consumption, among other data types. This approach is designed to leverage multidimensional data for comprehensive analysis and modeling, enhancing the precise classification of targeted categories. As per the literature [15], clustering user-specific features like electricity usage and time-of-use patterns aids in identifying groups with comparable electricity consumption behaviors. This technique facilitates the recognition of distinct power load patterns, enabling the model to discern patterns and relationships across various scenarios, ultimately leading to accurate scenario classification.

5. Scene Adaptive Recognition Matching Model Design

5.1. Enhancing Recognition in Diverse Power Generation Scenarios

Traditional recognition models predominantly utilize statistical and machine learning techniques. However, these models grapple with accurately distinguishing between scenarios like wind and solar power, which exhibit distinct characteristics and data distributions [16]. To surmount this challenge, the paper advocates for a multi-scenario recognition approach, leveraging a fusion of multiple binary classifiers. This approach entails constructing a multi-classifier system through autoencoder binary classifiers, forming a comprehensive "one-vs-all" multi-layer classification framework.

The quintessence of the "one-vs-all" strategy lies in its training methodology, where it systematically designates one category of samples as positive and the rest as negative. During the testing phase, each binary classifier adjudicates based on the training data. Classifiers yielding a positive result demarcate the sample's classification. With k categories, the model necessitates k binary classifiers. However, the conventional approach falters when facing classification overlap, where multiple classifiers yield positive outcomes simultaneously, obfuscating further distinction between types.

To mitigate this conundrum, the paper introduces a novel scenario-adaptive recognition matching technique. This method astutely resolves classification overlaps, ensuring precise categorization even in instances of concurrent positive outcomes from multiple classifiers. The adaptive nature of this approach tailors recognition processes to specific scenario characteristics, enhancing the overall classification accuracy. At its core, the adaptive method is characterized by its flexibility and intelligence, enabling the autonomous optimization of algorithm parameters and strategies across different datasets and application scenarios without the need for human intervention. This method, through the analysis of the structure and features of virtual power grid user data, automatically adjusts data preprocessing, feature extraction, and model training processes, culminating in the achievement of multi-scenario recognition. Capable of effectively adapting to both large-scale dataset processing and pattern identification in dynamic environments, this method ensures the accuracy and efficiency of analysis and recognition processes. By harnessing this adaptability, our method not only elevates the flexibility and intelligence of data processing but also significantly amplifies the model's application value and performance across a spectrum of data scenarios.

5.2. Enhanced Scene Adaptive Recognition Matching

This research introduces an enhanced algorithm that merges cosine similarity and the "one-vs-rest" classification strategy to adeptly distinguish and adaptively recognize various business scenes, as depicted in Figure 3.

Initially, the feature vector F, extracted from the dataset is processed using a one-vs-rest model composed of k binary classifiers. Each classifier is uniquely trained for a specific business scenario in a k-class classification problem. The positive output of a classifier, $type(i)$, suggests that F aligns with scenario i. In scenarios where multiple classifiers indicate positive results, the algorithm advances to a cosine similarity analysis. This stage involves computing the cosine similarity between F and the representative feature vectors of positively indicated types. The algorithm calculates the mean cosine similarity across all vectors to determine the overall similarity. For a sample feature vector set in power system scenarios represented as $(S_{i,1}, S_{i,2}, \ldots, S_{i,m})$, the average cosine similarity is given by:

$$\text{similarity} = \frac{1}{m} \sum_{j=1}^{m} \frac{F \cdot S_{i,j}}{\|F\| \cdot \|S_{i,j}\|} \tag{6}$$

where $S_{i,j}$ is the j-th sample in the power system scenario set, and m is the total number of samples. The scenario with the highest average cosine similarity is then selected as the recognized outcome.

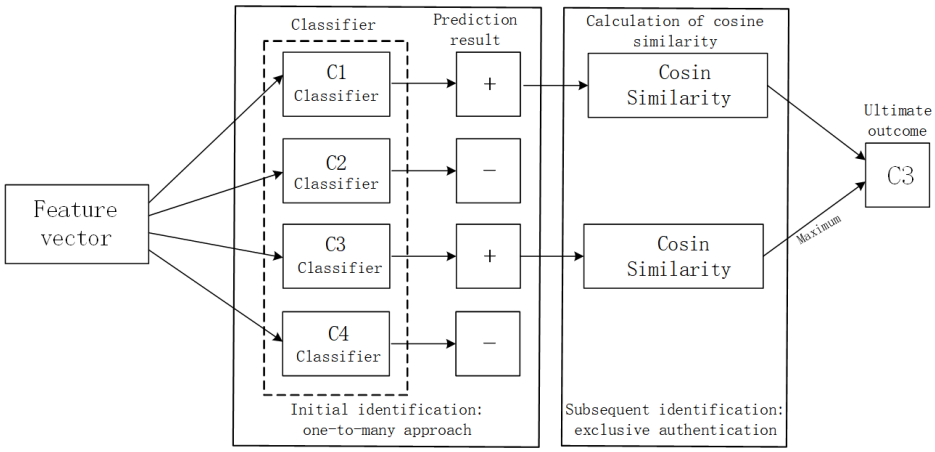

Figure 3. Enhanced cosine similarity-based multi-layer classifier-support vector machine model.

In this context, the study employs a machine-learning-based adaptive recognition method for data classification, specifically focusing on virtual power plants. This approach prioritizes computational efficiency in a limited range of scenarios over handling complex ones, thereby optimizing resource usage while maintaining accuracy. The adaptation of machine learning in this simplified context underlines the balance between simplicity and computational efficacy, ensuring the precise recognition of standard power system scenarios.

6. Enhanced Differential Privacy for Data Anonymization

To address the varying privacy requirements in diverse data scenarios, a dynamic data anonymization approach is necessary [17]. We introduce a differential privacy-based mechanism, adept at preserving individual data privacy and ensuring data utility across multiple scenarios. This mechanism anonymizes data by severing the link between data content and its source based on scenario-specific requirements, thus achieving adaptive desensitization.

6.1. Utility Function and Parameter Optimization

For differential privacy, the utility function incorporates cosine similarity, a measure of vector alignment, to evaluate output strategies [18]. Denote the utility of a strategy θ as $u(\theta)$, with θ as the feature vector. The cosine similarity between θ and a target vector θ_* is given by:

$$\cos(\theta, \theta_*) = \frac{\theta \cdot \theta_*}{\|\theta\| \|\theta_*\|} \tag{7}$$

Incorporating this into the utility function, we have:

$$u(\theta) = \cos(\theta, \theta_*) + C \tag{8}$$

where C includes additional terms. This refined utility function is crucial in the exponential mechanism:

$$p_\theta \propto \exp\left(\frac{u(\theta)}{2\Delta u}\right) \tag{9}$$

Here, p_θ is the selection probability, and Δu is the utility sensitivity. Adjusting cosine similarity for compatibility with other utility components is vital.

Considering differential privacy, the perturbation magnitude is bounded by parameter d. For a perturbed data item j, the change in the sum of squared errors is:

$$(\mathrm{y}_{*j} + d - \mathrm{y}_{tj})^2 - (\mathrm{y}_{*j} - \mathrm{y}_{tj})^2 = d^2 + 2d(\mathrm{y}_{*j} - \mathrm{y}_{tj}) \tag{10}$$

To manage global utility sensitivity (Δu), we enforce a range within $4d$. Exceeding this range triggers automatic thresholding to maintain sensitivity within acceptable limits.

6.2. Refined Algorithm Design for Information Security

This subsection introduces a refined approach to algorithm design in information security, particularly focusing on the effective application of k-fold cross-validation. The method involves accumulating the sum of squared errors (SSE) from each fold to assess overall performance.

Sensitivity Analysis in k-fold Cross-Validation: For a given perturbed data item j, the sensitivity within a test fold is calculated by summing sensitivities from the largest $k-1$ folds, including an additional term of $9d^2$. This assumes perturbation in the least sensitive training fold. Thus, the sensitivity equation for k-fold cross-validation is:

$$\Delta u^{\theta_i} = 9d^2 + \sum_{k=1}^{k-1} d^2 \max j|cjk|_2^2, \theta_i \in \Theta \tag{11}$$

where Θ is a feature vector in the set of parameter combinations to be examined, and all k folds of cross-validation are sorted in descending order of sensitivity.

Maximizing Sensitivity in Parameter Selection: In parameter selection, the highest sensitivity (Δu) from the potential parameter combinations (Θ) is crucial:

$$\Delta u \triangleq \max_{\theta_i \in \Theta} \Delta u^{\theta_i} \tag{12}$$

This maximization ensures robustness in the presence of perturbation noise.

In the subsequent steps, parameter combinations are selected based on calculated sensitivities aligned with the exponential mechanism. Within a predetermined total differential privacy budget θ, a portion is allocated to the selection of parameters, while the remainder is utilized during actual model training.

Parameter combinations that are highly sensitive to variance in individual training data might also exhibit heightened sensitivity in predictive accuracy. In such scenarios, differential privacy demands the use of more substantial and impactful perturbation noise, highlighting the high variance sensitivity of these parameters. Consequently, these combinations may not represent the most effective configurations. Therefore, the required differential privacy noise's influence is factored into the variance estimation for each parameter set. To mitigate the effects of differential privacy noise on the data, extensive sampling from the differential privacy noise distribution is conducted, ensuring its incorporation into the overall variance calculation.

Algorithm Construction Process: The algorithm construction process, outlined in Algorithm 1, integrates multiple samples from differential privacy perturbation noise to mitigate its effects. Key steps include setting parameters, training the model, and calculating the SSE.

Algorithm 1 Parameter selection and algorithm construction process

Require: Θ: set of parameter combinations to be examined; X: model training input; y: model training output; d: sensitivity threshold; k: number of folds for cross-validation;

for $\theta \in \Theta$ **do**
  $SSE^{(\theta)} \leftarrow 0$;
  for $i \in \{1, 2, \ldots, k\}$ **do**
    $C_i \leftarrow C\left(X^{(i)}, X_*^{(i)}, \theta\right)$;
    $y_*^{(i)} \leftarrow C_k y^{(i)}$;
    $SSE^{(\theta)} \leftarrow SSE^{(\theta)} + \sum_{j=1}^{\frac{N}{k}} \left(y_{*j}^{(i)} - y_{tj}^{(i)}\right)^2$;
    $\alpha_i \leftarrow \max_j \left|c_{ji}\right|_2^2$;
  end for
  $\Delta u^{(\theta)} \leftarrow 9d^2 + \sum_{i=1}^{k-1} d^2 \alpha_i$;
end for

7. System Design and Evaluation

7.1. Experimental Environment and Data Preparation

This experiment utilizes real user electricity consumption load data from the power generation and consumption scenarios of a provincial State Grid Corporation's virtual power plant as the validation data for this paper's proposed solution. Marketing data were collected from October to December 2022, spanning 157 users across six industries within the data collection system. The load information is presented in Table 2.

Table 2. Load information by industry.

Number	Industry	Quantity
1	Light Industry	29
2	Instrumentation	16
3	Pharmaceuticals	7
4	Automotive	39
5	Chemical	21
6	Non-Industrial	45

7.1.1. Dataset Preparation

Our model's training and evaluation are grounded in data from the State Grid Corporation of China, a major player in China's power system, serving a vast population and overseeing extensive power operations. The diversity and breadth of the State Grid's dataset enrich our research, offering a comprehensive view of user behaviors, electricity patterns, and generation scenarios. Despite the potential limitations of a single data source, the unique scope and variety of this dataset substantially offset these concerns, enhancing the model's applicability and generalizability.

After the data cleaning process, this experiment selected a portion of the content as real test data. Due to the limited amount of actual collected data, it was insufficient to train relevant models adequately. The electrical data for the expanded dataset are primarily synthesized through interpolation, a method used to estimate the values of unknown points within a given set of data points. Interpolation is utilized for dataset expansion to generate new samples between known data points, making it suitable for scenarios involving the statistical analysis of electricity usage. To address this limitation, the experiment also employed the method of generating a virtual dataset to expand the test samples. By comparing the training results of the models using both real and virtual datasets, the obtained training results from these two datasets were essentially consistent, as shown in Figure 4.

The figure displays three curves, each representing the recognition accuracy of model training after K-means clustering with real data, synthetic data, and mixed data, respectively. All curves exhibit an upward trend in recognition accuracy as the volume of data increases, stabilizing after a certain point and ultimately approaching a 97% accuracy rate. This indicates that the model performs consistently across different datasets, and the use of synthetic data to augment the training set is effective in this scenario. This indicates that the model established in this experiment still performs well in situations with limited data and demonstrates high generalization ability. Overall, by fully utilizing limited real data and generated virtual data, the experiment successfully validated the effectiveness of the model, laying a solid foundation for subsequent research and applications.

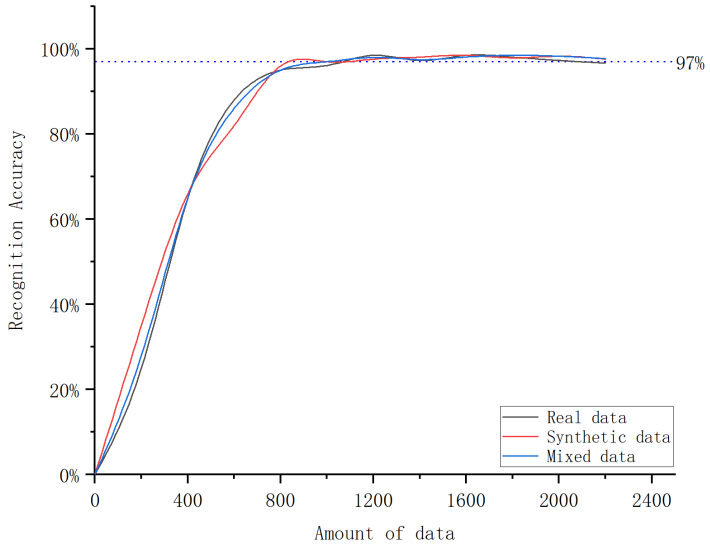

Figure 4. Comparison of Accuracy of K-means Model Based on Real and Virtual Datasets.

7.1.2. Computational Cost

The primary cost of the dynamic anonymization method originated from iterative training in the differential privacy process, averaging 1.19 s per iteration. The desensitization strategy in this study effectively reduced computational costs by using adaptive recognition results to guide parameter selection. Additionally, the clustering of datasets and multi-scenario recognition, employing a simplified structure as shown in Figure 5, further reduced overall computational expense.

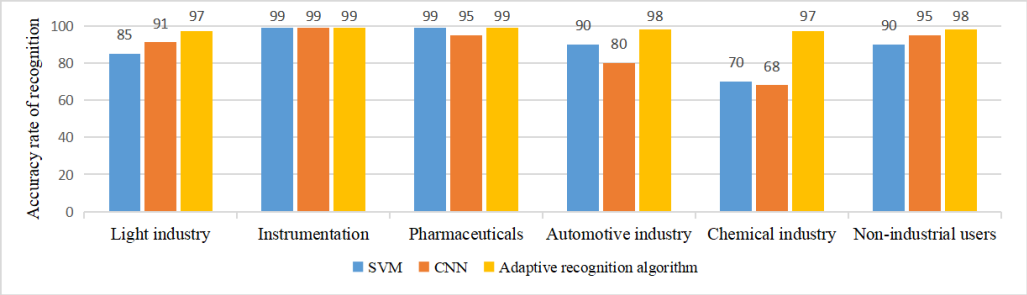

Figure 5. Comparison of load recognition accuracy under different algorithms.

7.2. Experimental Evaluation and Analysis

7.2.1. Clustering Algorithm Selection

The experiment involved selecting an appropriate algorithm for dataset training to facilitate the extraction of practical data features. A comparison of various algorithm combinations, presented in Table 3, highlighted the superior performance of the K-means and adaptive algorithm combination. This scheme is designed as an ablation study, with a particular focus on the performance of the K-means algorithm combined with adaptive algorithms. The purpose is to demonstrate how and why this combination excels in data preprocessing and feature extraction and maintains high adaptability across various data recognition scenarios. The choice of K-means plus adaptive algorithms is motivated by the combination's ability to effectively process diverse data scenarios, optimize the data processing flow, enhance recognition accuracy, and reduce computational costs and time consumption. This aspect is especially critical in practical applications, as it directly relates to the model's efficiency and feasibility.

Table 3. Comparison of scenario classification model performance.

Model	Accuracy	Precision	Recall	F1
K-means + SVM	0.896	0.886	0.702	0.783
K-means + CNN	0.876	0.883	0.889	0.886
Fuzzy Clustering + Adaptive Algorithm	0.916	0.908	0.914	0.911
Multi-layer Clustering + Adaptive Algorithm	0.933	0.901	0.893	0.897
K-means + Adaptive Algorithm	0.957	0.932	0.871	0.901

This combination excelled in recognition accuracy, demonstrating its strengths in data preprocessing and feature extraction. It stands out for its adaptability in recognizing data across various scenarios, effectively balancing high accuracy with reduced computational demands, making it a valuable asset for practical application in complex data environments.

7.2.2. Scenario Recognition Capability

This experiment employed feature classification recognition in the dataset for business scenario identification and classification. It compared the accuracy of scene recognition using the adaptive recognition algorithm based on machine learning with the original SVM algorithm and CNN algorithm. In this process, CNN demonstrated optimal performance when configured with a learning rate of 0.01, a convolutional kernel size of 3×3, and a fully connected layer consisting of 256 neurons. For the original SVM, the linear kernel function achieved optimal performance with a C value of 1.0, and non-linear kernel functions (such as radial basis kernel) also performed well with an appropriate bandwidth. Additionally, the experiment considered setting category weights to balance the importance of different categories. The results are shown in Figure 5.

Our enhanced adaptive algorithm demonstrated superior accuracy, ranging from 97% to 99% in various virtual power plant scenarios. Unlike the conventional SVM and CNN, our approach leverages diverse data types and fine-tuned parameters, offering more nuanced feature extraction and robust performance in scenario recognition. It effectively balances dataset generalization with high accuracy, contributing to user privacy and data security in virtual power plants.

7.2.3. Effectiveness of Differential Privacy Protection

The efficacy of differential privacy in our experiment was assessed in two dimensions: the accuracy of anonymized user data and the effectiveness of desensitization across multiple scenarios.

Figure 6 illustrates the interplay between the privacy protection budget ε and model accuracy within the differential privacy framework. The privacy protection budget ε, serves as a metric for assessing the risk of information leakage. A lower ε value entails adding more noise to the data, enhancing privacy at the potential cost of reduced accuracy and usability. In contrast, a higher ε value introduces less noise, preserving data utility but potentially diminishing privacy protection.

Figure 6 demonstrates that with $\varepsilon = 2$, model accuracy reaches 87.5%, signifying that substantial privacy protection can coexist with high accuracy. As ε increases, there is a corresponding improvement in model accuracy, indicating a relaxation in privacy protection intensity. For instance, at ε values of 10 and 20, model accuracy is higher, albeit at a reduced level of privacy protection compared to $\varepsilon = 2$. Therefore, implementing differential privacy strategies necessitates balancing privacy protection with data utility, a balance influenced by specific application scenarios, each requiring tailored ε settings.

Figure 7 presents data desensitization in four distinct business scenarios using differential privacy algorithms. The figures highlight the influence of varying scale parameters n (100, 200, 500, 1000) on desensitization efficacy. An increase in n correlates with a marked reduction in the average relative error, showcasing the differential privacy budget's effect on data error in various scenarios, thereby enhancing data usability.

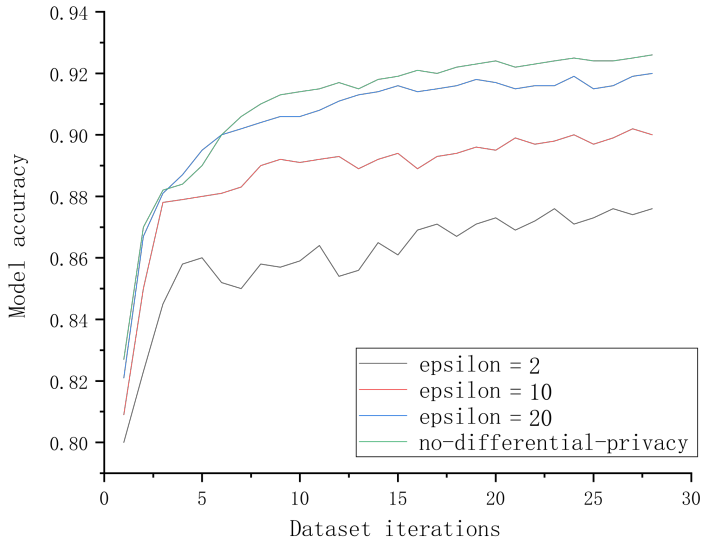

Figure 6. Privacy protection budget values and model accuracy.

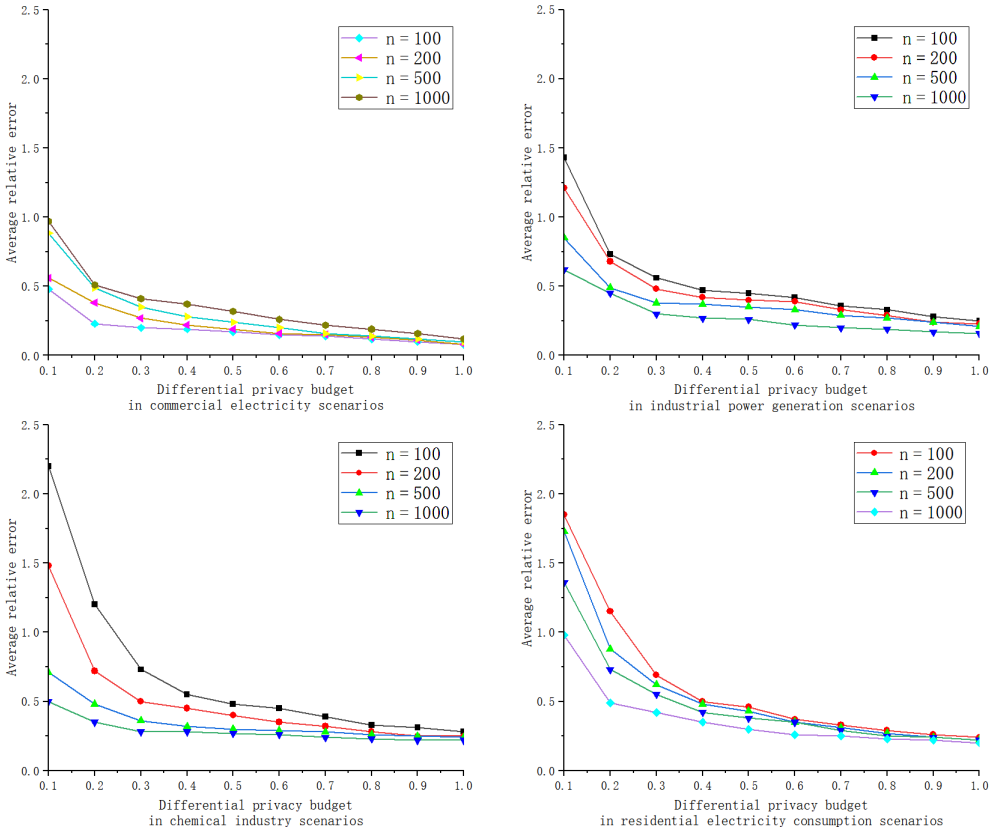

Figure 7. Differential privacy anonymization capability under different virtual power plant business scenarios.

Distinct trends are observable in Figure 7. The top left and right figures show a flattening of curves with increasing n, signaling a decline in error. In contrast, the bottom figures exhibit a more pronounced trend, indicating that higher n values significantly improve the average relative error in those scenarios.

These findings underscore the effectiveness of Gaussian process-based differential privacy mechanisms in safeguarding data privacy while maintaining usability [19]. Practically, this suggests that adjusting differential privacy algorithm parameters can optimize the balance between privacy protection and data usability, maximizing data value retention while securing user privacy.

7.3. Analysis of Advantages and Disadvantages

The dynamic desensitization protection method introduced in this study demonstrates both potential strengths in comparison with current advanced technologies.

1. Scene Adaptability: This method leverages machine learning for adaptive scene recognition, offering a significant advantage over traditional uniform privacy protection methods. It enables tailored privacy protection levels based on varying business scenarios, thus preserving greater data utility while safeguarding privacy.
2. Dynamic Privacy Protection Intensity: Unlike existing technologies, this approach dynamically adjusts privacy protection intensity in response to scenario shifts, enhancing model flexibility and adaptability for diverse business needs.
3. Similarity Utility Function: The innovative introduction of a similarity concept to evaluate scenario similarities provides a quantitative basis for dynamic desensitization across various scenarios, a feature not commonly found in existing technologies.

Similarly, this system has limitations when compared to other existing systems.

1. Model Complexity: Integrating machine learning with Gaussian processes, this method may entail greater complexity in training and parameter optimization than some existing methods, demanding more computational resources and expertise.
2. Data Dependency: The method's effectiveness is largely contingent on the quality and representativeness of the dataset. Limited generalizability may arise if the dataset fails to encompass all potential business scenarios.
3. Algorithm Transparency: For decision-makers and users, the internal workings of differential privacy algorithms based on Gaussian processes could be less intuitive compared to simpler algorithms, thus impacting their transparency and interpretability.

7.4. Future Research Directions

Given the analysis above, the proposed method exhibits advantages in terms of flexibility, scene adaptability, and data utility, yet faces challenges regarding model complexity, data dependency, and algorithm transparency. Recognizing and addressing these limitations and biases is vital for understanding the scope of the research outcomes and guiding future research.

Focusing on virtual power grid models, the method's similarity utility function for assessing inter-scenario similarity may be affected by data feature distribution, possibly leading to inaccurate utility assessments in certain scenarios. Future research should aim to overcome these limitations, possibly through dataset expansion, algorithm optimization for enhanced computational efficiency, and developing novel approaches to adapt to dynamic scenarios and improve model robustness. Additionally, given the computational demands of Gaussian processes, further exploration is warranted into utilizing modern computing technologies, like sparse Gaussian processes, to augment the scalability and efficiency of algorithms, tailoring parameters to dataset characteristics and privacy budgets for an optimal balance between user privacy and utility.

8. Conclusions

In the intricate operating environment of virtual power plants, ensuring the security of user privacy data is of paramount importance. Conventional data desensitization methods, typically static in approach, fall short in the multifaceted context of virtual power plants operating under multiple scenarios simultaneously. These traditional approaches lack intelligent classification for processing data across various scenarios and are inefficient in

desensitization processing amidst differing security demands. The dynamic desensitization protection method developed in this study employing cutting-edge machine learning techniques, accomplishes the intelligent and adaptive recognition of operational scenarios in virtual power plants. This recognition goes beyond mere static characteristics of scenarios, incorporating the dynamic relationships and similarities between them. By conducting comprehensive machine learning analyses of scenario features, the algorithm autonomously identifies and adapts to diverse business scenarios, selecting suitable privacy protection levels for handling sensitive data.

A pivotal innovation of this method is the introduction of the similarity utility function. This function not only incorporates the concept of scenario similarity but also adaptively modifies the desensitization strategy based on the assessed similarity among scenarios. This implies that the desensitization process can deliver precise privacy protection without compromising data utility. Furthermore, the study employs a differential privacy algorithm based on Gaussian processes, refining the selection of privacy security parameters.

Experimental outcomes reveal that the model achieves a recognition accuracy of up to 87.5% under an appropriate privacy budget. This highlights the mechanism's capability to respond to varied privacy protection requirements across different scenarios while maintaining data processing efficiency and accuracy.

In conclusion, the methodologies and experimental findings of this study establish that dynamic desensitization protection is not just viable but also highly efficient and adaptable in the complex, multi-scenario, multi-objective, and dynamic operational milieu of virtual power plants. This approach provides a novel solution for privacy data protection in virtual power plants, imbued with significant theoretical and practical implications. Employing this dynamic desensitization approach, virtual power plants can safeguard user privacy while enhancing operational efficiency, and fostering intelligent and personalized data management and protection.

Author Contributions: Conceptualization, R.Y.; methodology, R.Y.; software, J.W.; validation, R.Y. and F.S.; formal analysis, J.W.; investigation, R.Y.; resources, H.G.; data curation, R.Y.; writing—original draft preparation, F.S.; writing—review and editing, R.Y. and H.G.; visualization, R.Y.; supervision, H.G.; project administration, R.Y. All authors have read and agreed to the published version of the manuscript.

Funding: This work was supported by the National Key R&D Program of China (No. 2021YFB2401200).

Data Availability Statement: We confirm that the data supporting the findings of this study are available within the article. Additional data that support the findings of this study are available from the corresponding author upon reasonable request.

Conflicts of Interest: Author Ruxia Yang was employed by the company State Grid Smart Grid Research Institute Co., Ltd. Author Jun Wang was employed by the State Grid Shanghai Municipal Electric Power Company. The remaining authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

References

1. Natalia, N.; Yusta, J.M. Virtual Power Plant Models and Electricity Markets—A Review. *Electr. Meas. Instrum.* **2021**, *149*, 111393.
2. Singh, A.K.; Kumar, J. A Privacy-preserving Multidimensional Data Aggregation Scheme with Secure Query Processing for Smart Grid. *J. Supercomput.* **2023**, *79*, 3750–3770. [CrossRef]
3. Zhan, Y.; Zhou, L.; Wang, B.; Duan, P.; Zhang, B. Efficient Function Queryable and Privacy Preserving Data Aggregation Scheme in Smart Grid. *IEEE Trans. Parallel Distrib. Syst.* **2022**, *33*, 3430–3441. [CrossRef]
4. Yao, J. Combinational Recognition Model for Demand Side Load Profile in Shanghai Power Grid. *Power Syst. Technol.* **2010**, *34*, 145–151.
5. Chen, C.; Gao, P.; Jiang, J. A Deep Learning Based Non-intrusive Household Load Identification for Smart Grid in China. *Comput. Commun.* **2021**, *177*, 176–184. [CrossRef]
6. Sun, S.; Zhang, K.; Feng, J.; Li, B.; Zhu, S.; Chen, S. Research on Non-intrusive Load Identification Technology Based on Deep Learning. In Proceedings of the 2019 IEEE 3rd Conference on Energy Internet and Energy System Integration, Changsha, China, 8–10 November 2019; pp. 462–467.

7. Chen, R.; Sun, H.; Guo, Q. A Generation-interval-based Mechanism for Managing the Power Generation Uncertainties of Variable Generation. *IEEE Trans. Sustain. Energy* **2016**, *7*, 1060–1070. [CrossRef]
8. Xu, G.; Qi, C.; Yu, H.; Xu, S.; Zhao, C.; Yuan, J. Detecting Sensitive Information of Unstructured Text Using Convolutional Neural Network. In Proceedings of the 2019 International Conference on Cyber-Enabled Distributed Computing and Knowledge Discovery, Guilin, China, 17–19 October 2019; pp. 474–479.
9. Xiong, P.; Zhu, T.; Wang, X. A Survey on Differential Privacy and Applications. *Acta Comput. Sci.* **2014**, *37*, 101–122.
10. Kanungo, T.; Mount, D.M.; Netanyahu, N.S. An Efficient K-means Clustering Algorithm: Analysis and Implementation. *IEEE Trans. Pattern Anal. Mach. Intell.* **2002**, *24*, 881–892. [CrossRef]
11. Vapnik, V.N. *Estimation of Dependencies Based on Empirical Data*; Springer: Berlin, Germany, 1982.
12. Smith, M.; Lopez, M.A.A.; Zwiessele, M. Differentially Private Regression with Gaussian Processes. In Proceedings of the International Conference on Artificial Intelligence and Statistics, Lanzarote, Spain, 9–11 April 2018; pp. 1195–1203.
13. Tang, Z.; Zhao, W.; Wang, C.; Yang, Z.; Xu, Y.; Cui, S. A Data Desensitization Algorithm for Privacy Protection Electric Power Industry. *IOP Conf. Ser. Mater. Sci. Eng.* **2020**, *768*, 52–59. [CrossRef]
14. Wang, J.; Xu, M.; Lu, K. The Research of Adaptive Data Desensitization Method Based on Middle Platform. *Wirel. Commun. Mob. Comput.* **2022**, *2022*, 5348637. [CrossRef]
15. Xia, D.; Song, D.; Dong, W. A Short-term Power Load Forecasting Method Based on K-means and SVM. *J. Ambient Intell. Humaniz. Comput.* **2021**, *13*, 5253–5267.
16. Atef, S.; Nakata, K.; Eltawil, A.B. A Deep Bi-directional Long-short Term Memory Neural Network-based Methodology to Enhance Short-term Electricity Load Forecasting for Residential Applications. *Comput. Sci.* **2022**, *170*, 108364. [CrossRef]
17. Yang, H.; Ji, Y.; Pan, Y. Differentially Private Distributed Logistic Regression with the Objective Function Perturbation. *Int. J. Wavelets* **2022**, *21*, 2250043. [CrossRef]
18. Zhang, T.; Hu, Z. Optimal Scheduling Strategy of Virtual Power Plant with Power-to-gas in Dual Energy Markets. *IEEE Trans. Ind. Appl.* **2021**, *2*, 58. [CrossRef]
19. Pan, K.; Gong, M.; Feng, K. Differentially Private Regression Analysis with Dynamic Privacy Allocation. *Knowl.-Based Syst.* **2021**, *217*, 106795. [CrossRef]

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.