energies

MDPI

Article

Multimodal Operation Data Mining for Grid Operation Violation Risk Prediction

Lingwen Meng [1], Jingliang Zhong [1], Shasha Luo [1], Xinshan Zhu [2], Yulin Wang [2] and Shumei Zhang [2,*]

[1] Electric Power Scientific Research Institute of Guizhou Power Grid Guizhou, Guiyang 550000, China; menglw000@163.com (L.M.); 13985469075@139.com (J.Z.); luoshasha931@163.com (S.L.)
[2] School of Electrical Automation and Information Engineering, Tianjin University, Tianjin 300072, China; xszhu@tju.edu.cn (X.Z.); yulin_2000@tju.edu.cn (Y.W.)
* Correspondence: shumeizhang@tju.edu.cn

Abstract: With the continuous expansion of the power grid, the issue of operational safety has attracted increasing attention. In power grid operation control, unauthorized operations are one of the primary causes of personal accidents. Therefore, preventing and monitoring unauthorized actions by power grid operators is of critical importance. First, multimodal violation data are integrated through information systems, such as the power grid management platform, to construct a historical case database. Next, word vectors for three types of operation-related factors are generated using natural language processing techniques, and key vectors are selected based on generalized correlation coefficients using mutual information, enabling effective dimensionality reduction. Independent component analysis is then employed for feature extraction and further dimensionality reduction, allowing for the effective characterization of operational scenarios. For each historical case, a risk score is derived from a violation risk prediction model constructed using the Random Forests (RF) algorithm. When a high-risk score is identified, the K-Nearest Neighbor (KNN) algorithm is applied to locate similar scenarios in the historical case database where violations may have occurred. Real-time violation risk assessment is performed for each operation, providing early warnings to operators, thereby reducing the likelihood of violations, and enhancing the safety of power grid operations.

Keywords: violation dictionary construction; operation risk prediction; random forests; independent component analysis; mutual information

check for updates

Citation: Meng, L.; Zhong, J.; Luo, S.; Zhu, X.; Wang, Y.; Zhang, S. Multimodal Operation Data Mining for Grid Operation Violation Risk Prediction. *Energies* **2024**, *17*, 5424. https://doi.org/10.3390/en17215424

Academic Editor: Tek Tjing Lie

Received: 20 September 2024
Revised: 21 October 2024
Accepted: 29 October 2024
Published: 30 October 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

The smart grid plays a crucial role as a foundational platform in the Chinese national strategy to promote energy transformation and drive innovation-led development [1]. With the ongoing expansion of electric power infrastructure, construction safety has garnered significant attention [2]. Enhancing risk resilience is a top priority, and strengthening grid security must be the primary goal of grid development [3]. Grid operations face numerous challenges, including a large number of operators, high personnel turnover, outdoor construction risks, and poor working conditions, all of which contribute to potential safety hazards. To implement the policy of "safety first, prevention foremost, and comprehensive management", power grid enterprises have undertaken extensive safety and anti-violation efforts [4]. Workplace accidents pose a significant threat to employees' health and safety. Despite advancements in safety technology across various industries, non-compliant behavior remains a major contributor to workplace accidents [5]. Memarian et al. [6] argue that the rational use of risk analysis and assessment methods is essential for effectively mitigating risk factors in power safety production. The Southern Power Grid Corporation embraces the safety concept that "all accidents can be prevented". However, the current management of power operation violations primarily focuses on site control, dispatching safety personnel to inspect work sites, or using cameras for video surveillance. Peng [7] proposed an object detection model, CORY-Net, combining contrastive self-supervised

pretraining and YOLOv5 for intelligent safety monitoring on power grid construction sites. Nevertheless, unauthorized operations are still largely identified manually by monitoring personnel, which is inefficient [8]. Additionally, violation prevention mainly relies on improving management systems, with little focus on assessing operational violation risks before grid operations, making effective prevention challenging.

On-site power operations involve numerous risk factors, such as temperature, operating environment, voltage levels, and work modes, which are characterized by complexity, dynamic changes, randomness, and uncontrollability. A scientific analysis of these risk factors is essential for accurately assessing the risks associated with on-site power operations [9]. In the complex power grid system, the diversity of data sources and the complexity of data connotations result in multimodal and cross-modal data characteristics. This means that the data contain various modalities that correspond to or interact with one another, such as text, information tables, and more. For text data, converting textual information into numerical representations is often required. In this context, Qaiser et al. [10] utilized Term Frequency-Inverse Document Frequency (TF-IDF) to comprehensively evaluate document content, increasing the significance of important keywords and reducing the impact of high-frequency words. However, when the keyword "command" is included in text data, the TF-IDF formula assigns it a low weight, despite its critical role in operational information. Midya et al. [11] employed independent component analysis (ICA) to extract response features for automotive fault diagnosis, simplifying computation and aiding in system design for vehicle diagnostics and quality control. In the power sector, Zhou et al. [12] examined the deep relationship between noncompliance and violations using data from electric power companies, paving the way for improved workplace safety. Awasthi et al. [13] proposed a K-Nearest Neighbor (KNN) algorithm for accurately and efficiently classifying fault types in distribution networks with distributed generation (DG), thereby enhancing the reliability and stability of complex distribution systems. Mao [14] introduced a deep multi-task learning approach, addressing issues such as insufficient feature mining and error accumulation in multi-node load forecasting by incorporating multi-modal feature extraction and a soft-shared multi-task learning model with strong practical value. Chang et al. [15] combined dynamic information, such as operator images and locations, with static information derived from historical data to perform a comprehensive risk assessment and estimate safety probabilities. Maria et al. [16] proposed a new algorithm for managing unexpected events in smart grids, evaluating potential impacts to ensure smooth grid operation. Praveen [17] developed a CNN-based model for fault classification in power distribution networks integrated with distributed generation, achieving 99.52% accuracy without the need for pre-processing or feature engineering. While some research has focused on risk prediction and prevention mechanisms for violations in field power operations, these studies remain inadequate. There is currently no system that provides a future operational plan combined with relevant data to predict the likelihood of violations, key violation types, and potential issues prior to the operation. This information, if available, would offer operators valuable insights into preventing violations before they occur.

To address these challenges, this paper first constructs a historical case database by leveraging multimodal historical operation data from information systems such as power grid management platforms. Natural language processing techniques are then applied to generate word vectors for three types of operation-related factors. Generalized correlation coefficients based on mutual information are used to select key word vectors, enabling efficient approximation and dimensionality reduction of the word vectors. ICA is employed for feature extraction and dimensionality reduction, facilitating accurate representation of operational scenarios. For each historical case, a violation risk prediction model, built using the Random Forest (RF) algorithm, is utilized to calculate a risk score. If the risk score is high, the KNN algorithm is applied to identify potential violations under similar operational conditions from the historical case database. Real-time violation risk assessment is conducted with each issued operation, providing early warnings to operators.

This approach effectively mitigates the occurrence of violations and enhances the safety and security of power grid operations.

1.1. Jieba Segmentation

Segmentation is a crucial task in natural language processing, involving the division of continuous text into meaningful words or phrases based on specific rules. Unlike English, where spaces naturally separate words, Chinese word segmentation is more complex. The challenges arise from differing tasks, perspectives, and interpretations of what constitutes a "word", making it difficult to establish a universal standard for segmentation.

The Jieba segmentation tool in Python employs a comprehensive general-purpose lexicon, making it both powerful and broadly applicable across various domains. It also allows for the integration of a custom lexicon with the general-purpose lexicon, facilitating efficient segmentation in specialized fields [18]. Jieba's segmentation algorithm is based on Hidden Markov Models (HMMs), which encompasses three distinct segmentation models: the exact model, the complete model, and the search engine model. HMMs are statistical models that describe a Markov process with parameters that are not directly observable. The complete model scans all possible words in the text according to a Chinese thesaurus, which, while fast, does not effectively resolve ambiguities. In contrast, the exact model aims for precise segmentation, ensuring that no redundant words appear in the output. The search engine model builds upon the exact model by re-segmenting long words to enhance recall, albeit at the expense of introducing some redundancy in the results.

1.2. Independent Component Analysis

Independent Component Analysis (ICA) is a signal processing technique grounded in the principle of statistical independence of mixed signals [19]. It operates by separating observed signals into mutually independent source signals through a linear transformation. ICA assumes that the mixed signals are linear combinations of the underlying source signals and aims to maximize the statistical independence of the output signals by adaptively adjusting the transformation matrix. This method has extensive applications in addressing the challenge of separating independent components within multidimensional data.

FastICA, developed by Hyvärinen et al. [20], is a widely recognized and efficient blind source separation algorithm in the field of multidimensional signal processing, noted for its rapid convergence and superior performance. It aims to recover independent source signals from mixed signals. The fundamental principles of FastICA encompass several key processes: centering, whitening, initialization of the separation matrix, iterative updating, and termination condition testing. By iteratively updating the separation matrix, the non-Gaussianity of the output signal is progressively enhanced, allowing for the identification of a transformation matrix that facilitates the blind source separation of the signals. FastICA is particularly effective in processing non-Gaussian signals.

1.3. K-Nearest Neighbor Algorithm

The KNN algorithm is a basic machine learning algorithm for classification and regression problems [21]. It works on the principle of classification based on the votes of the K nearest neighbors in the feature space or regression based on the mean of these neighbors. For a given data point, the KNN determines the class to which it belongs by calculating the distance to its nearest neighbor training samples by majority voting. Specifically, it uses KNN in the feature space to perform classification or regression based on their labels. Among them, the sample distance is measured by Euclidean distance, Manhattan distance, Minkowski distance, and cosine similarity; the choice of K value is also particularly critical [22]. When the K value is too small, the model will be particularly sensitive to noise and is easily affected by local noise, which leads to poor generalization of the model to the overall data, and when the K value is too large, it easily leads to the phenomenon of overfitting of the model.

1.4. Random Forests Algorithm

The RF algorithm was introduced by Leo Breiman and Adele Cutler [23], who elaborated on the principles and applications of the Random Forests algorithm in their 2001 paper. The RF algorithm is an ensemble learning method that addresses classification and regression tasks by aggregating the outputs of multiple decision trees. Each decision tree is constructed using a randomly selected subset of features and a randomly selected subset of samples. The final prediction is then determined through voting or averaging the results from all the trees [24].

2. Construction of Grid Operation Violation Risk Prediction Model for Operation Data Mining

This paper presents a grid operation violation risk prediction method based on operational data mining, enabling real-time violation risk assessment for each issued operation and providing early warnings for operators. Specifically, as illustrated in Figure 1, this violation assessment method constructs a predictive model through the processing of multimodal unstructured data. The modeling process primarily involves the following steps: preprocessing of multimodal unstructured data, construction of an operational lexicon and word vector space, keyword vector filtering and dimensionality reduction, feature extraction via ICA, violation score prediction via RF, and subsequent violation risk assessment.

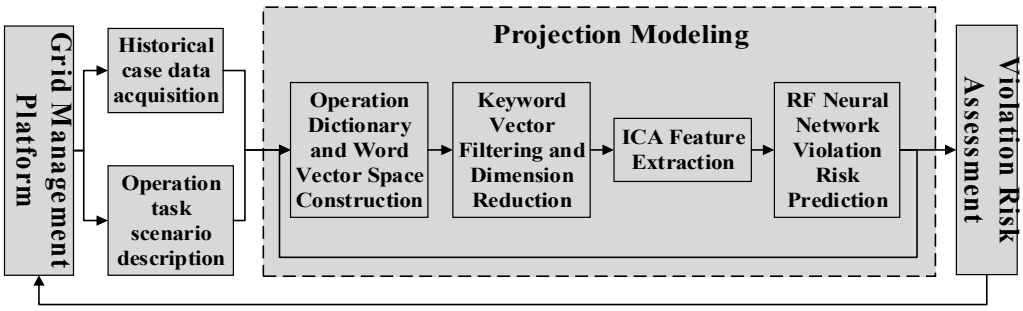


Figure 1. Flowchart for the proposed method.

2.1. Data Acquisition and Case Base Construction

Relying on the grid management platform and other information systems to obtain multimodal data on grid operation violations, recorded as $\widetilde{X} = [\widetilde{X}_1, \widetilde{X}_2, \widetilde{X}_3] \in \mathbb{R}^{n\times m}$, its corresponding violation label is noted as $Y \in \mathbb{R}^{n\times 1}$, where $\widetilde{X}_1 \in \mathbb{R}^{n\times m_1}$ denotes m_1 factors whose attributes are in numerical form—examples include weather temperature, latitude and longitude of the workplace, etc.; $\widetilde{X}_2 \in \mathbb{R}^{n\times m_2}$ denotes m_2 factors in the form of fixed text phrases; $Y \in \mathbb{R}^{n\times 1}$ is the violation label. n denotes the number of historical cases, and $m = m_1 + m_2 + m_3$ denotes the number of operational factors.

In order to construct a standardized database, the multimodal historical operation data are first preprocessed to construct word vectors of operation-related factors.

Step 2.1-1 Pre-processing of numerical form factors

For operation factors whose attributes are in numerical form $\widetilde{\mathbf{X}}_1^i (i = 1, 2, \ldots, m_1)$, normalization is performed to remove the effects of different magnitudes:

$$\mathbf{X}_1^{i,j} = \frac{\widetilde{\mathbf{X}}_1^{i,j} - \min\left(\widetilde{\mathbf{X}}_1^i\right)}{\max\left(\widetilde{\mathbf{X}}_1^i\right) - \min\left(\widetilde{\mathbf{X}}_1^i\right)} \tag{1}$$

where $\widetilde{\mathbf{X}}_1^{i,j}$ denotes the value of the attribute of the operational factor $\widetilde{\mathbf{X}}_1^i (i = 1, 2, \ldots, m_1)$ in the $jth (j = 1, 2, \ldots n)$ historical case; $\max\left(\widetilde{\mathbf{X}}_1^i\right)$ and $\min\left(\widetilde{\mathbf{X}}_1^i\right)$ denote the maximum and minimum values of the attributes of the operational factor $\widetilde{\mathbf{X}}_1^i (i = 1, 2, \ldots, m_1)$, respectively.

Step 2.1-2 Pre-processing of fixed text phrase form assignment factors

For factors that fix the form of textual phrases $\widetilde{\mathbf{X}}_2^i (i = 1, 2, \ldots, m_2)$, their subfactor is used as a keyword in the operational violation dictionary. The attribute values in the factor are counted and coded sequentially, and these values are converted to numeric vector form.

Step 2.1-3 Pre-processing of descriptive linguistic form assignment factors

For descriptive linguistic forms with the factor $\widetilde{\mathbf{X}}_3^i (i = 1, 2, \ldots, m_3)$, the Jieba segmentation engine is firstly employed for text segmentation. After removing low-frequency partitions by filtering out deactivated words with low information content, a lexicon is constructed. Subsequently, the bag-of-words (BoW) model is applied based on this filtered lexicon, resulting in a total of J words. The frequency of each word from the lexicon appearing in the segmented results of each operational case is recorded as the corresponding position in the J-dimensional vector. This process generates a J-dimensional vector for each operational case, thereby establishing the word vector space by applying the previously described procedures across all operational cases.

With the above steps, the word vector space of grid operations can be constructed, and n operation cases are represented as a structured data matrix $X \in \mathbb{R}^{n \times J}$ with variable dimension J and number of samples n:

$$X = \begin{pmatrix} x_1(1) & x_2(1) & \cdots & x_J(1) \\ x_1(2) & x_2(2) & \cdots & x_J(2) \\ \vdots & \vdots & \ddots & \vdots \\ x_1(n) & x_2(n) & \cdots & x_J(n) \end{pmatrix} \in \mathbb{R}^{n \times J} \tag{2}$$

where the value of $x_j(i)$ denotes the value of the attribute corresponding to the *jth* factor or sub-phrase in the *ith* historical case.

Step 2.1-4 Pre-processing of violation label

For the violation label, there are two elements: "violation" and "no violation". "No violation" is coded as 0, while "violation" is coded as 1.

2.2. Key Segmentation Extraction and Classical Word Vector Space Construction

To address the challenges posed by the high-dimensional sparse word vector matrix, this paper introduces a keyword vector extraction method grounded in mutual information. First, the mutual information between each keyword in the dictionary and the violation label is calculated. Subsequently, correlation coefficients based on this mutual information are computed, enabling the selection of word vectors with higher correlation coefficients as keywords. The specific steps involved in this process are outlined below.

Step 2.2-1 Mutual information between segmentation and violation label

For the high-dimensional sparse word vector matrices, the mutual information between each keyword in the dictionary and the violation label is first calculated:

$$I(X_j, Y) = H(X_j) + H(Y) - H(X_j, Y) \tag{3}$$

where $\mathbf{X}_j = \left[x_j(1), x_j(2), \ldots, x_j(n)\right]^T$ denotes the $jth (j = 1, 2, \ldots, J)$ segmentation, $H(X_j)$ is the information entropy of the segmentation X_j:

$$H(X_j) = -\int_x p(X_j) \log p(X_j) dx \tag{4}$$

$H(Y)$ is the information entropy of the violation label Y:

$$H(Y) = -\int_y p(Y) \log p(Y) dx \tag{5}$$

$H(X_j, Y)$ is the joint information entropy of the partition X_j and Y:

$$H(X_j, Y) = -\iint_{x_j, y} p(X_j, Y) \log p(X_j, Y) dx_j dy \tag{6}$$

where $p(X_j)$ and $p(Y)$ denote the probability density functions of the variables X_j and Y, and $p(X_j, Y) = \iint_{x_j, y} p(x, y) dx_j dy$ is the joint probability density function.

Step 2.2-2 Generalized correlation coefficients for segmentation and violation Label

The generalized correlation coefficients of the segmentation X_j and segmentation Y are obtained based on the mutual information derived from Equation (6):

$$r_{X_j, Y} = r(X_j, Y) = \frac{I(X_j, Y)}{\min\big(H(X_j), H(Y)\big)} \tag{7}$$

where $r_{X_j, Y} \in [0, 1]$.

Step 2.2-3 Filtering of keywords

If the correlation coefficient $r_{X_j, Y}$ of the *jth* segmentation with the label Y is greater than the threshold δ, it means that the segmentation is a keyword:

$$r_{X_j, Y} = r(X_j, Y) > \delta \tag{8}$$

Otherwise, the segmentation is considered to contain less information and discarded, where the threshold δ is selected in practice based on the mean value of the mutual information values between the calculated segmentation and the violation label.

Step 2.2-4 Construction of key segmentation lexicon and word vector matrices

After filtering out the keyword segments through the previous step, the key lexicon of historical cases is built and the word vector matrix is constructed as

$$\widehat{X} = \begin{pmatrix} \widehat{x}_1(1) & \widehat{x}_2(1) & \cdots & \widehat{x}_J(1) \\ \widehat{x}_1(2) & \widehat{x}_2(2) & \cdots & \widehat{x}_J(2) \\ \vdots & \vdots & \ddots & \vdots \\ \widehat{x}_1(n) & \widehat{x}_2(n) & \cdots & \widehat{x}_J(n) \end{pmatrix} \in \mathbb{R}^{n \times J} \tag{9}$$

where the value of $\widehat{x}_k(i)$ denotes the attribute value corresponding to the *kth* key segmentation in the *ith* historical case, and J is the number of key segmentations.

2.3. Feature Extraction and Operational Scenario Characterization

To further extract the underlying information from the word vector matrix, this paper employs ICA for feature extraction and dimensionality reduction. The identified potential features serve as characterizations of the operational scenarios for each historical case in the subsequent steps.

Step 2.3-1 Whitening process

Whiten the sparse high-dimensional word vector matrix $\widehat{X}$ obtained in Equation (9):

$$Z = \Lambda^{-1/2} U^T \widehat{X} = Q\widehat{X} \in \mathbb{R}^{n \times J} \tag{10}$$

where matrix Z is the whitened data, $Q = \Lambda^{-1/2} U^T$ is the whitened matrix, and U and Λ are the orthogonal eigenvector array and eigenvalue diagonal array obtained by eigenvalue decomposition of the covariance matrix $E(\widehat{X}\widehat{X}^T)$, respectively.

Step 2.3-2 Low-dimensional feature extraction

Mapping the high-dimensional word vector matrices to the low-dimensional space using FastICA yields the independent principal element matrices:

$$S = WZ^T = \begin{bmatrix} W_{1:\mathrm{p}} \\ W_{p+1:J} \end{bmatrix} Z^T \in \mathbb{R}^{J \times n} \tag{11}$$

where S denotes the independent principal elements and p is the number of retained principal elements; W is the decomposition matrix, the solution of which can be solved iteratively according to the non-Gaussian maximization principle using the FastICA algorithm proposed by Hyvarine [17].

In this way, with minimal loss of information, the operational scenarios for each historical case are characterized as a few of the most representative independent principal elements:

$$\widetilde{S} = ZW_{1:p}^T = \begin{bmatrix} s_1(1) & s_2(1) & \cdots & s_p(1) \\ s_1(2) & s_2(2) & \cdots & s_p(2) \\ \vdots & \vdots & \ddots & \vdots \\ s_1(n) & s_2(n) & \cdots & s_p(n) \end{bmatrix} \tag{12}$$

2.4. Modeling of Violation Risk Estimation

To facilitate the estimation of violation risk within a specific operational scenario, it is essential to develop a prediction model that incorporates both the operational scenario representation vector and the violation risk indicator. The key steps involved in this process are described below.

Step 2.4-1 Violation risk indicators

For the *ith* historical case, the KNN method is used to find the operational cases under similar operational scenarios in the historical case database, and the distance between the two cases is calculated:

$$d_{ij} = \sqrt{\sum_{t=1}^{p} \left(s_i(t) - s_j(t)\right)^2} \tag{13}$$

Select the K cases closest to $\widetilde{S}_i$, noted as $N(\widetilde{S}_i) = \left\{\widetilde{S}_1^i, \widetilde{S}_2^i, \cdots, \widetilde{S}_k^i\right\}$, and find the corresponding violation label $y(\widetilde{S}_i) = \left\{y_i^1, y_i^2 \cdots, y_i^k\right\}$.

The probability of violation is calculated as an indicator of the risk of violation under the operational scenario of the *ith* historical case:

$$p(\widetilde{S}_i) = \frac{sum(y(\widetilde{S}_i))}{k} = \frac{\sum_{t=1}^{k} y_i^t}{k} \tag{14}$$

where k can be determined using a cross-validation algorithm.

Step 2.4-2 Violation risk estimation modeling

The RF algorithm is utilized to construct the prediction model of operational key factors and violation risk indicators, the structure of which is shown in Figure 2. $\widetilde{S} = (\widetilde{S}_1, \widetilde{S}_2, \ldots, \widetilde{S}_n)$ is the input layer data, $Y_p = (p(\widetilde{S}_1), p(\widetilde{S}_2), \ldots, p(\widetilde{S}_n))$ is the output layer data, and the RF model consists of multiple decision trees, each of which is a type of base learner. The input layer data is transmitted to each decision tree, where each node represents a decision based on a feature of the input data. The RF algorithm aggregates the outputs of all decision trees through voting or averaging to ultimately produce a prediction of the target variable. The construction of each decision tree introduces diversity into the model by randomly selecting subsets of samples and features. The overall RF model enhances its predictive capability by adjusting model parameters and the number of trees while minimizing loss functions, such as the root mean square error, to achieve an accurate prediction model.

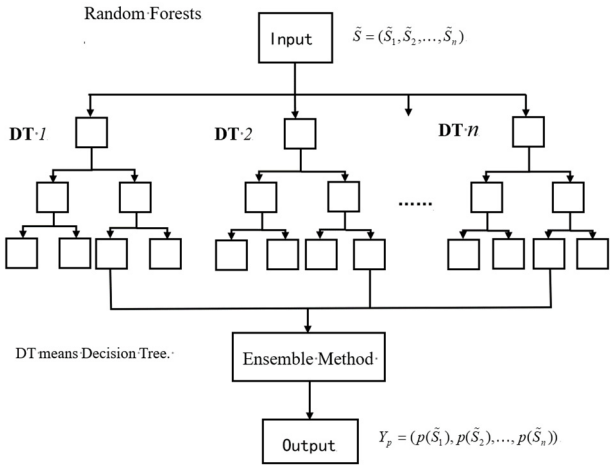


Figure 2. Model structure of RF.

2.5. Risk Prediction of Behavioral Violations in Operational Tasks

In order to realize the violation risk assessment of the grid operation, it is necessary to predict the violation risk in advance according to the operation scenario when the operation is issued, the specific steps for which are described below.

Step 2.5-1 Pre-processing of data for the online operations work plan

When a grid operation is planned, its operation factors can be represented as $\widetilde{x} = [\widetilde{x}_1, \widetilde{x}_2, \widetilde{x}_3] \in \mathbb{R}^{1\times m}$, where $\widetilde{x}_1 \in \mathbb{R}^{1\times m_1}$ denotes m_1 factors whose attributes are in numerical form; $\widetilde{x}_2 \in \mathbb{R}^{1\times m_2}$ denotes m_2 factors in the form of fixed text phrases; and $\widetilde{x}_3 \in \mathbb{R}^{1\times m_3}$ denotes m_3 groups of factors in descriptive language.

Step 2.5-2 Pre-processing of numerical type factors for the online operations work plan

For numerical type factors, normalization is performed using the minimum and maximum values of the operational factors in Step 2.1-1:

$$x_1^i = \frac{\widetilde{x}_1^i - \min(\widetilde{X}_1^i)}{\max(\widetilde{X}_1^i) - \min(\widetilde{X}_1^i)} \tag{15}$$

Step 2.5-3 Constructing keyword vectors for the online operations work plan

Using the key dictionary developed in Step 2.2-4, a keyword vector is constructed to represent the current grid operation task scenario.:

$$\widehat{x} = [\widehat{x}(1) \quad \widehat{x}(2) \quad \cdots \quad \widehat{x}(J)] \in \mathbb{R}^{1\times J} \tag{16}$$

Step 2.5-4 The whitening vector of the online operations work plan

Using the whitening transformation matrix obtained in Step 2.3-1, the whitening vector for the current grid operation task scenario is derived:

$$z = Q\widehat{x} \in \mathbb{R}^{1\times J} \tag{17}$$

Step 2.5-5 Independent principal element representation of the online operations work plan

Using the decomposition matrix obtained in Step 2.3-2, the independent principal element representation for the current grid operation task scenario is derived:

$$\widetilde{s} = zW_{1:p}^T = [s_1 \quad s_2 \quad \cdots \quad s_p] \in \mathbb{R}^{1\times p} \tag{18}$$

Step 2.5-6 Violation risk indicators for the online operations work plan

The independent principal element representation under the current grid operation task scenario is input into the RF prediction model obtained in Step 2.4-2 to calculate the

violation risk indicator $p(\widetilde{s})$ for the current grid operation task scenario. When this indicator is relatively large, the violation situation of cases under similar operation scenarios in the historical case database is searched, and the corresponding violation dictionary is queried to provide a reference for the grid operator to effectively reduce the occurrence of violation behaviors and provide safety for the grid operation.

3. Application and Analysis of Results

3.1. Data Presentation and Pre-Processing

A total of 24,001 operational data points from January–July 2022 were obtained from a power grid company, and 21,715 historical cases were derived by removing data cases with incomplete information. A total of 90% of the data were randomly selected as historical cases to construct a case database. Therefore, 19,544 data points were used for model training and the remaining 2171 data points were used as testing data. As shown in Table 1, each piece of operational data contains 17 operational factors, $\widetilde{X} = [\widetilde{X}_1, \widetilde{X}_2, \widetilde{X}_3] \in \mathbb{R}^{19544\times17}$ Among them, $\widetilde{X}_1 \in \mathbb{R}^{19544\times5}$ denotes five factors whose attributes are in the form of numerical values, including { Longitude of the operating site, Latitude of the operating site, Maximum temperature on working days, Minimum temperature on working days, Voltage level}; $\widetilde{X}_2 \in \mathbb{R}^{19544\times11}$ represents 11 factors in the form of fixed text phrases, including { municipal bureau, Supervisory levels, systematic classification, Project management units, Operating units (shifts), Main body of operation, Task origin, Type of operation, Professional fields, Risk levels and Operation methods}; $\widetilde{X}_3 \in \mathbb{R}^{19544\times1}$ represents the descriptive language factor " Operational plan implementation contents"; and $Y \in \mathbb{R}^{19544\times1}$ is the violation label, which is transformed into a numeric form with the attribute value {1, 0}, where 1 means "violation" and 0 means " no violation".

Table 1. Example of a numerical factor in an operational plan.

Operational Influencing Factors	Examples of Contents	Factors Format
Municipal bureaus	XX Power Supply Bureau	Fixed texts
Supervisory levels	Prefectural level	
systematic classification	Supervision system	
Project management units	XX Power Supply Bureau	
Operating units (shifts)	XX XX Development Ltd.	
Main body of operation	Home unit, outside unit	
Task origin	Infrastructure	
Type of operation	Infrastructure type	
Professional fields	Power transmission	
Risk levels	Medium	
Operation methods	Non-stop power supply	
Operating site longitude (°)	104.95764	Numeric data
Operating site latitude (°)	26.54658	
Maximum temperature (°C)	14	
Minimum temperature (°C)	5	
Voltage level	35 kV	
Operational plan implementation contents	35 kV Guan Steepening Line to 110 kV Stroke Field Substation 35 kV Project (P53 Pylon Establishment)	Descriptive Language Texts

In this case, relying on the grid management platform and other information systems to obtain the data from the database of grid operation violation behavior is noted as $\widetilde{X} = [\widetilde{X}_1, \widetilde{X}_2, \widetilde{X}_3] \in \mathbb{R}^{19544\times17}$, and its corresponding violation label is noted as $Y \in \mathbb{R}^{19544\times1}$. Following the data preprocessing method outlined in Section 2.1, a word vector space for grid operations is constructed. Attributes such as "temperature" and "longitude", represented as numerical data, are processed for their magnitudes. A factor dictionary is created to quantify these attribute values and convert them into numerical vector form; for

instance, the factor "Task Origin" encompasses three textual attributes—{Infrastructure, Marketing, Production}—which are encoded as {1, 2, 3}. For the descriptive language of the factor "Operational Plan Implementation Contents", the Jieba segmentation engine is employed to segment the text. Words with low information content are filtered out by setting deactivation words, resulting in a total of 8675 filtered segments that include key descriptors relevant to grid operations, such as "Operations", "Site", "Switch", and "Ticket". After removing less frequent segments, 113 segmentation entries are retained. The bag-of-words (BoW) model is then utilized to establish the word vector space, leading to the formation of a structured data matrix that represents 19,544 operation cases across 113 variable dimensions and 19,544 samples, with the initial word cloud illustrated in Figure 3.



Figure 3. Word cloud map of initial segmentation.

3.2. Keyword Dictionary Construction and Feature Extraction Results

To extract fault-related key segments, the mutual information between each segment and the violation label in the dictionary is first calculated. Subsequently, the correlation coefficients based on the mutual information are computed, allowing for the selection of word vectors with larger correlation coefficients as key segments. Through practical analysis, it has been determined that a correlation coefficient $r_{X_j,Y} > 0.22$ between a segment and the violation label qualifies that segment as a key segment. Figure 4 illustrates the mutual information of each segment in the lexicon relative to the violation label, with the horizontal axis representing the lexicon keywords (such as "violation", "operation", etc.) and the vertical axis indicating the mutual information value between each corresponding keyword and the violation occurrence.

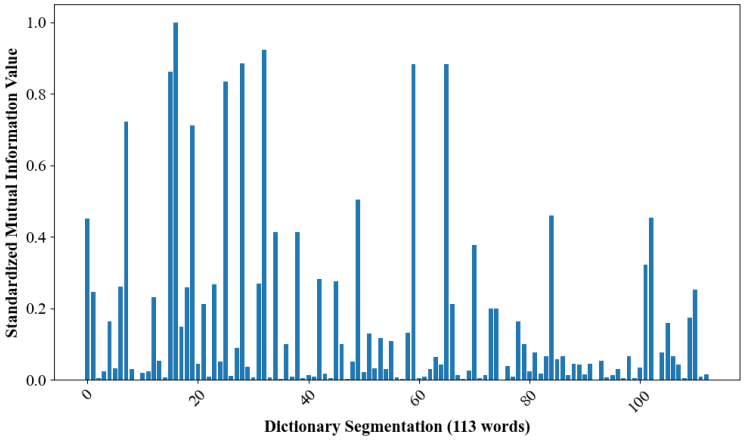


Figure 4. Mutual information between key segmentations and violation tags.

Finally, a total of 26 key sub-words, such as "Violation", "Personnel", "Site", " Work", "Violate", "Insulation", "Ticket", etc., are screened as shown in Figure 5 to establish the keyword dictionary for the historical cases and construct the word vector matrix.



Figure 5. Word cloud map of key segmentations (26 words.).

Subsequently, the 26-dimensional sparse high-dimensional word vector matrix is processed using the ICA algorithm. Through whitening and FastICA, the high-dimensional word vector matrix is mapped to a low-dimensional space, resulting in a 3-dimensional independent principal matrix. Figure 6 illustrates the independent principal components of the historical cases. The ICA algorithm is employed for feature extraction, effectively mapping the high-dimensional word vector matrix to the low-dimensional space, where the most representative features are selected as the operational scene representations for each historical case, ensuring minimal loss of information.

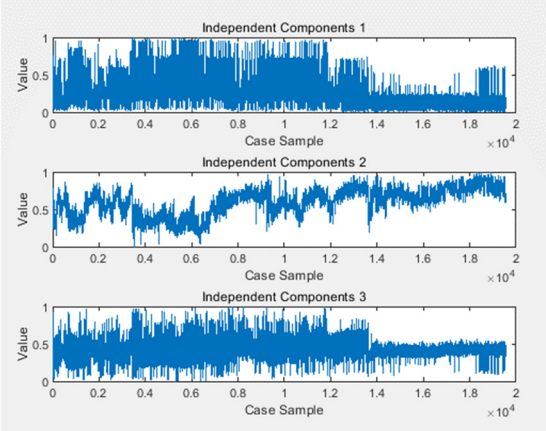


Figure 6. Independent feature representation of historical casework scenarios.

3.3. Operational Risk Prediction Modeling and Online Assessment Results

When establishing the operational risk prediction model, for each historical case, the KNN algorithm is used to find operational cases under similar operational scenarios in the historical case database, and the probability of violation of the k historical operational cases in the nearest neighbor is statistically analyzed and used as the violation risk indicator under the operational scenario. The RF algorithm is used to construct the estimation model of operation key factors and violation risk index, $\widetilde{S} = (\widetilde{S}_1, \widetilde{S}_2, \ldots, \widetilde{S}_{19544})$ is the input data, $Y_p = \left(p(\widetilde{S}_1), p(\widetilde{S}_2), \ldots, p(\widetilde{S}_{19544})\right)$ is the output data, the number of decision trees is set to 100, the minimum number of leaf nodes is set to 5, the out-of-bag error estimation is turned on, and the importance of features is calculated, which is used for evaluating the performance of the model.

To validate the effectiveness of the proposed grid operation violation risk prediction algorithm, a total of 2171 grid operation tasks were evaluated. The RF algorithm employed in this study was compared against six other algorithms: Extreme Learning Machine (ELM), Radial Basis Function (RBF) neural network, Back Propagation (BP) neural network along with its variants—Genetic Algorithm-Back Propagation (GA-BP) and Particle Swarm Optimization-Back Propagation (PSO-BP)—as well as a Convolutional Neural Network (CNN). The differences between the predicted and actual values of violation risk for all models on the testing set are illustrated in Figure 7.

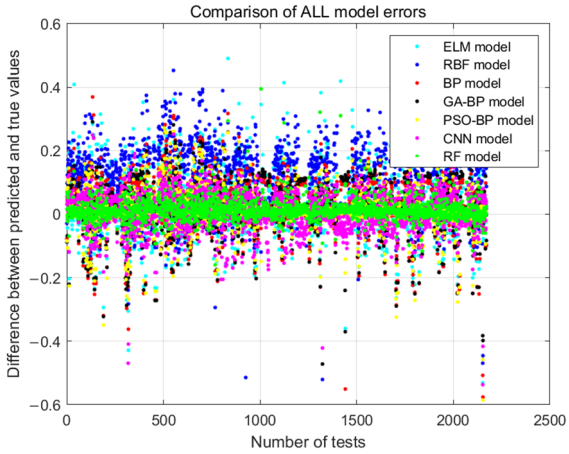


Figure 7. Predicted error on the testing set for different models.

To gain a more comprehensive understanding of the advantages of the proposed method, a comparative analysis was conducted with other algorithms based on three key metrics: R2 (i.e., coefficient of determination), MAE (mean absolute error), and MBE (mean bias error). The comparison of R^2 values for the training and testing sets across different algorithmic models is presented in Figure 8, while the comparisons of the MAE and MBE metrics are summarized in Tables 2–4.

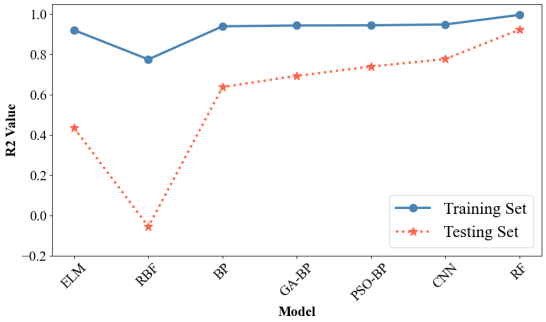


Figure 8. Comparison of R2 values for different models.

Table 2. Comparison of R2, MAE, and MBE metrics for initial word vector space.

Models / Indicators		ELM	RBF	BP	GA-BP	PSO-BP	CNN	RF
R2	Training set	0.3165	0.4182	0.4861	0.5353	0.5390	0.6165	0.8232
	Testing set	0.2414	0.3851	0.4692	0.4791	0.4797	0.4847	0.5728
MAE	Training set	0.3501	0.3104	0.2691	0.2443	0.2314	0.2251	0.1228
	Testing set	0.3924	0.3472	0.2949	0.2768	0.2722	0.2741	0.2374
MBE	Training set	−0.0170	−0.0195	−0.0051	0.0259	−0.0192	0.0036	0.0019
	Testing set	0.0505	−0.0142	0.0296	0.0679	0.00378	0.0165	0.0130

First, the performance metrics considered are R^2, MAE, and MBE. The R^2 metric is a statistical measure used to evaluate the fit of a regression model to the data, assessing how well the predicted values align with the actual observed values. Its range is from 0 to 1, where a value of 0 indicates poor predictive accuracy, while values closer to 1 suggest higher accuracy, with 1 signifying perfect prediction. MAE is commonly employed to assess the accuracy of predictive models; it represents the average of the absolute differences between predicted and actual values. Similarly, MBE measures predictive accuracy and is calculated as the average of the differences between predicted and actual values, divided

by the total number of samples. A more accurate model will yield a smaller absolute value of MBE, ideally approaching zero.

Table 3. Comparison of R2, MAE, and MBE metrics for Keywords vector space.

Models / Indicators		ELM	RBF	BP	GA-BP	PSO-BP	CNN	RF
R2	Training set	0.3376	0.4031	0.5839	0.5745	0.5586	0.5888	0.8173
	Testing set	0.2955	0.3492	0.4797	0.4810	0.4816	0.5006	0.5757
MAE	Training set	0.3461	0.3199	0.2084	0.2088	0.2176	0.2368	0.1231
	Testing set	0.3830	0.3554	0.2561	0.2519	0.2510	0.2794	0.2272
MBE	Training set	−0.0168	−0.0182	0.0013	−0.0021	−0.0028	0.0028	0.0014
	Testing set	0.0185	−0.0191	0.0141	0.0087	−0.0030	0.0089	−0.0125

Table 4. Comparison of R2, MAE, and MBE metrics for dimensional vector space.

Models / Indicators		ELM	RBF	BP	GA-BP	PSO-BP	CNN	RF
R2	Training set	0.9210	0.7758	0.9404	0.9442	0.9453	0.9494	0.9971
	Testing set	0.4348	−0.0551	0.6386	0.6941	0.7403	0.7776	0.9237
MAE	Training set	0.0653	0.1204	0.0548	0.0518	0.0497	0.0511	0.0107
	Testing set	0.0661	0.1062	0.0545	0.0499	0.0412	0.0492	0.0227
MBE	Training set	−0.0052	−0.0234	0.0005	−0.0025	−0.0008	−0.0021	0.0001
	Testing set	0.0230	0.0897	0.0206	0.0168	0.012	0.0087	0.0132

The analysis reveals that, for the performance indicator R2, the RF algorithm achieves values closer to 1 in both the training and testing sets. It indicates that the RF algorithm exhibits stronger explanatory power for the target variables compared to other algorithms, effectively capturing the relationship between independent and dependent variables. Notably, in the testing set, the RBF algorithm yields a negative R2 value when it processes 3-dimensionality-reduced data. This may be attributed to the limited information present in low-dimensional data, which restricts the model's fitting capability and hinders its generalization. Additionally, for the performance metric MAE, the RF algorithm model demonstrates relatively lower values in both the training and testing sets, indicating that the prediction results obtained through the RF algorithm are more accurate than those of the comparative algorithms. Finally, regarding the performance metric MBE, the RF algorithm shows smaller values that are closer to zero in both training and testing sets, signifying that the RF algorithm predictions are largely consistent with the actual values.

Furthermore, the comparative analysis presented in Tables 2–4 demonstrates the significance of processing steps such as removing low-frequency words, keyword extraction, and feature extraction. This approach not only helps avoid the pitfalls associated with high-dimensional data but also allows for the selection of key participles that effectively characterize job construction scenarios, as evidenced by the comparison between Tables 2 and 3. The resulting low-dimensional vector representation of the field operation plan text is achieved while maintaining prediction accuracy. Additionally, FastICA effectively extracts the latent information from the keyword vector matrix, enabling further feature extraction and dimensionality reduction. A comparison of Tables 3 and 4 indicates that this dimensionality reduction operation produces a vector representation that significantly enhances the probability of violation risk prediction.

In summary, a comprehensive comparison of various algorithms reveals that the RF algorithm developed in this study achieves an R2 value of 92.37% on the testing set. It indicates that the model effectively leverages information from current operational scenarios to predict violation risks in real-time. Furthermore, it provides relevant early warning

information to operational staff when issuing commands, specifically regarding common violations observed in similar historical cases. This proactive approach significantly reduces the likelihood of violations and ensures the safe operation of the power grid. In comparison to other algorithms, the proposed method demonstrates superior performance in terms of both accuracy and robustness.

4. Conclusions

This paper analyzes the multimodal grid historical violation operation data from January to July 2022 of a power company and constructs a historical case database. Based on the Jieba partition and bag-of-words model, the different modal data have been processed to construct the initial word vector space, which realizes the numerical structural description of the historical grid violation operation cases. In order to remove the redundancy of the initial lexicon, a novel key feature extraction strategy has been designed to extract the keywords based on mutual information, construct the keyword vector space of historical cases, and introduce the ICA algorithm to realize the dimensionality reduction characterization of the word vector space. After that, a RF algorithm is utilized to construct a grid operation violation risk prediction model. The proposed method can realize real-time assessment of violation risk, so as to effectively reduce the occurrence of violations and provide a safety guarantee for power grid operation.

Author Contributions: Conceptualization, L.M. and J.Z.; methodology, S.Z.; software, X.Z.; validation, Y.W., L.M. and S.L.; formal analysis, Y.W.; investigation, S.L.; resources, J.Z.; data curation, X.Z.; writing—original draft preparation, L.M. and J.Z.; writing—review and editing, S.Z. and S.L.; supervision, L.M.; project administration, X.Z. funding acquisition, L.M. All authors have read and agreed to the published version of the manuscript.

Funding: This research was funded by Guizhou Power Grid Limited Liability Company, grant number GZKJXM20222320.

Data Availability Statement: Data is available upon request from the corresponding authors or works cited in the article.

Conflicts of Interest: Author Lingwen Meng were employed by the Electric Power Scientific Research Institute of Guizhou Power Grid Guizhou. 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. Marot, A.; Kelly, A.; Naglic, M.; Barbesant, V.; Cremer, J.; Stefanov, A.; Viebahn, J. Perspectives on Future Power System Control Centers for Energy Transition. *J. Mod. Power Syst. Clean Energy* **2022**, *10*, 328–344. [CrossRef]
2. Sun, Y.; Zhu, J.; Yu, X.; Ye, C.; Fan, G. Research on early warning of power grid construction safety based on PSO-SVM model. *J. Phys. Conf. Ser.* **2020**, *1449*, 012037. [CrossRef]
3. AlHaddad, U.; Basuhail, A.; Khemakhem, M.; Eassa, F.E.; Jambi, K. Towards sustainable energy grids: A machine learning-based ensemble methods approach for outages estimation in extreme weather events. *Sustainability* **2023**, *15*, 12622. [CrossRef]
4. Zhang, Z.; Lin, S.; Ye, Y.; Xu, Z.; Zhao, Y.; Zhao, H.; Sun, J. A hybrid MCDM model for evaluating the market-oriented business regulatory risk of power grid enterprises based on the bayesian best-worst method and MARCOS approach. *Energies* **2022**, *15*, 2978. [CrossRef]
5. Reason, J. *The Human Contribution: Unsafe Acts, Accidents and Heroic Recoveries*; CRC Press: Boca Raton, FL, USA, 2017.
6. Memarian, B.; Mitropoulos, P. Safety incidents and high-risk activities of masonry construction. In Proceedings of the Construction Research Congress 2012: Construction Challenges in a Flat World, West Lafayette, Indiana, 21–23 May 2012; pp. 2510–2519.
7. Peng, G.; Lei, Y.; Li, H.; Wu, D.; Wang, J.; Liu, F. CORY-Net: Contrastive res-YOLOv5 network for intelligent safety monitoring on power grid construction sites. *IEEE Access* **2021**, *9*, 160461–160470. [CrossRef]
8. Chen, B.; Zhang, H.; Li, Y.; Fang, X.; Chen, H. A lightweight detection scheme for the safety of human behaviors in power construction sites. *J. Circuits Syst. Comput.* **2023**, *32*, 2350230. [CrossRef]
9. Peng, B.; Zhang, W.; Lin, X.; Qiu, H. A novel risk assessment method and application for the on-site power operations. *Vibroengineering Procedia* **2021**, *37*, 47–53. [CrossRef]
10. Qaiser, S.; Ali, R. Text mining: Use of TF-IDF to examine the relevance of words to documents. *Int. J. Comput. Appl.* **2018**, *181*, 25–29. [CrossRef]

11. Midya, M.; Ganguly, P.; Datta, T.; Chattopadhyay, S. ICA-Feature-Extraction-Based Fault Identification of Vehicular Starter Motor. *IEEE Sens. Lett.* **2023**, *7*, 1–4. [CrossRef]
12. Zhou, F.; Lu, H.; Jiang, C. Violation with concerns of safety: A study on non-compliant behavior and the antecedent and consequent effects in power grid construction. *Saf. Sci.* **2024**, *170*, 106353. [CrossRef]
13. Awasthi, S.; Singh, G.; Ahamad, N. Classifying electrical faults in a distribution system using k-nearest neighbor (knn) model in presence of multiple distributed generators. *J. Inst. Eng. Ser. B* **2024**, *105*, 621–634. [CrossRef]
14. Tan, M.; Hu, C.; Chen, J.; Wang, L.; Li, Z. Multi-node load forecasting based on multi-task learning with modal feature extraction. *Eng. Appl. Artif. Intell.* **2022**, *112*, 104856. [CrossRef]
15. Chang, Z.; Deng, Y.; Wu, J.; Xiong, X.; Chen, M.; Wang, H.; Xie, X. Safety risk assessment of electric power operation site based on variable precision rough set. *J. Circuits Syst. Comput.* **2022**, *31*, 2250254. [CrossRef]
16. Skydt, M.R.; Bang, M.; Shaker, H.R. A probabilistic sequence classification approach for early fault prediction in distribution grids using long short-term memory neural networks. *Measurement* **2021**, *170*, 108691. [CrossRef]
17. Rai, P.; Londhe, N.D.; Raj, R. Fault classification in power system distribution network integrated with distributed generators using CNN. *Electr. Power Syst. Res.* **2021**, *192*, 106914. [CrossRef]
18. Ding, Y.; Teng, F.; Zhang, P.; Huo, X.; Sun, Q.; Qi, Y. Research on text information mining technology of substation inspection based on improved Jieba. In Proceedings of the 2021 International Conference on Wireless Communications and Smart Grid (ICWCSG), Aachen, Germany, 25–28 October 2021; IEEE: Piscataway, NJ, USA; pp. 561–564.
19. Hyvärinen, A.; Oja, E. Independent component analysis: Algorithms and applications. *Neural Netw.* **2000**, *13*, 411–430. [CrossRef]
20. Hyvarinen, A. Fast and robust fixed-point algorithms for independent component analysis. *IEEE Trans. Neural Netw.* **1999**, *10*, 626–634. [CrossRef] [PubMed]
21. Zhang, S. Challenges in KNN classification. *IEEE Trans. Knowl. Data Eng.* **2021**, *34*, 4663–4675. [CrossRef]
22. Bian, Z.; Vong, C.M.; Wong, P.K.; Wang, S. Fuzzy KNN method with adaptive nearest neighbors. *IEEE Trans. Cybern.* **2020**, *52*, 5380–5393. [CrossRef]
23. Breiman, L. Random forests. *Mach. Learn.* **2001**, *45*, 5–32. [CrossRef]
24. Yu, M.; Niu, D.; Gao, T.; Wang, K.; Sun, L.; Li, M.; Xu, X. A novel framework for ultra-short-term interval wind power prediction based on RF-WOA-VMD and BiGRU optimized by the attention mechanism. *Energy* **2023**, *269*, 126738. [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.