energies

MDPI

Review

Virtual Collection for Distributed Photovoltaic Data: Challenges, Methodologies, and Applications

Leijiao Ge [1], Tianshuo Du [1], Changlu Li [1,*], Yuanliang Li [2], Jun Yan [2] and Muhammad Umer Rafiq [1]

[1] School of Electrical and Information Engineering, Tianjin University, Tianjin 300072, China
[2] Concordia Institute for Information Systems Engineering, Concordia University, Montreal, QC H3G 1M8, Canada
* Correspondence: changlu@tju.edu.cn

Abstract: In recent years, with the rapid development of distributed photovoltaic systems (DPVS), the shortage of data monitoring devices and the difficulty of comprehensive coverage of measurement equipment has become more significant, bringing great challenges to the efficient management and maintenance of DPVS. Virtual collection is a new DPVS data collection scheme with cost-effectiveness and computational efficiency that meets the needs of distributed energy management but lacks attention and research. To fill the gap in the current research field, this paper provides a comprehensive and systematic review of DPVS virtual collection. We provide a detailed introduction to the process of DPVS virtual collection and identify the challenges faced by virtual collection through problem analogy. Furthermore, in response to the above challenges, this paper summarizes the main methods applicable to virtual collection, including similarity analysis, reference station selection, and PV data inference. Finally, this paper thoroughly discusses the diversified application scenarios of virtual collection, hoping to provide helpful information for the development of the DPVS industry.

Keywords: distributed photovoltaic; virtual collection; similarity analysis; reference station; data inference; artificial intelligence

check for updates

Citation: Ge, L.; Du, T.; Li, C.; Li, Y.; Yan, J.; Rafiq, M.U. Virtual Collection for Distributed Photovoltaic Data: Challenges, Methodologies, and Applications. *Energies* **2022**, *15*, 8783. https://doi.org/10.3390/en15238783

Academic Editor: Francesco Calise

Received: 14 October 2022
Accepted: 18 November 2022
Published: 22 November 2022

Publisher's Note: MDPI stays neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Copyright: © 2022 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

In the context of the current global energy crisis and increasing environmental pollution, photovoltaic (PV) power generation has received strong support from countries worldwide due to its high efficiency and cleanliness, rapidly becoming the third largest renewable energy source after hydropower and wind power [1]. According to the reports of the International Energy Agency (IEA), more than 175 GW of new PV capacity was installed worldwide in 2021, accounting for more than half of the new renewable energy capacity. By the end of 2021, the cumulative installed PV capacity worldwide had exceeded 942 GW. Figure 1 illustrates the changing dynamics of the global PV market and the substantial influence of the Chinese PV market [1].

PV stations are mainly divided into two types: centralized systems and distributed systems. Distributed photovoltaic systems have been rapidly developed due to their flexible installation, outstanding environmental benefits, and coexistence of power generation and consumption [2]. As shown in Figure 2, the newly installed capacity of distributed photovoltaics exceeded that of centralized photovoltaics in 2021 [3]. Therefore, efficient and accurate access to DPVS operational data is becoming increasingly important. High-quality operational data can help assess the output performance index of DPVS to improve the reliability of the PV plant in terms of the operation and maintenance and the accuracy of DPVS output prediction. It can also provide power companies with accurate electricity metering and billing audit indicators, better monitoring of the market, and prolong the service life of DPVS. However, most DPVS are scattered and disorderly, with many points

and wide areas. To effectively manage them, a large number of sensors, collectors, and concentrators need to be deployed to monitor the output of DPVS, as well as dedicated communication channels, servers, databases, and data monitoring software [4]. However, higher implementation costs and personal privacy requirements make a significant portion of PV users reluctant to purchase these data monitoring services, limiting the further development of the PV industry. In addition, as the scale of distributed PV continues to expand and the operating environment becomes more complex and diverse, the collection of its operating data often suffers from transmission blockage and equipment failure. For this reason, it is crucial and beneficial to develop a cost-effective and computationally efficient data collection method for large-scale DPVS clusters with relatively small numbers of sensing devices deployed at strategic locations. If deployed at strategic locations with proper redundancy, the reduced sensing network can still provide low-cost yet highly sufficiently accurate measurements of the DPVS networks for various power operations.

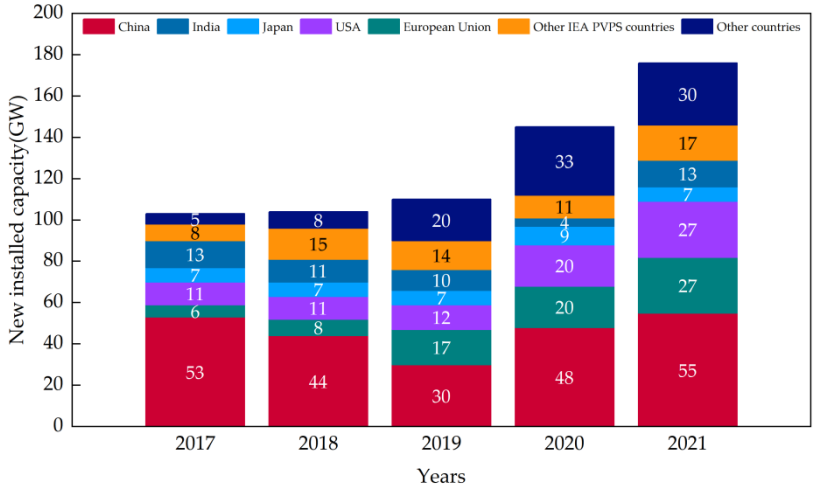

Figure 1. 2017–2021 growth per region.

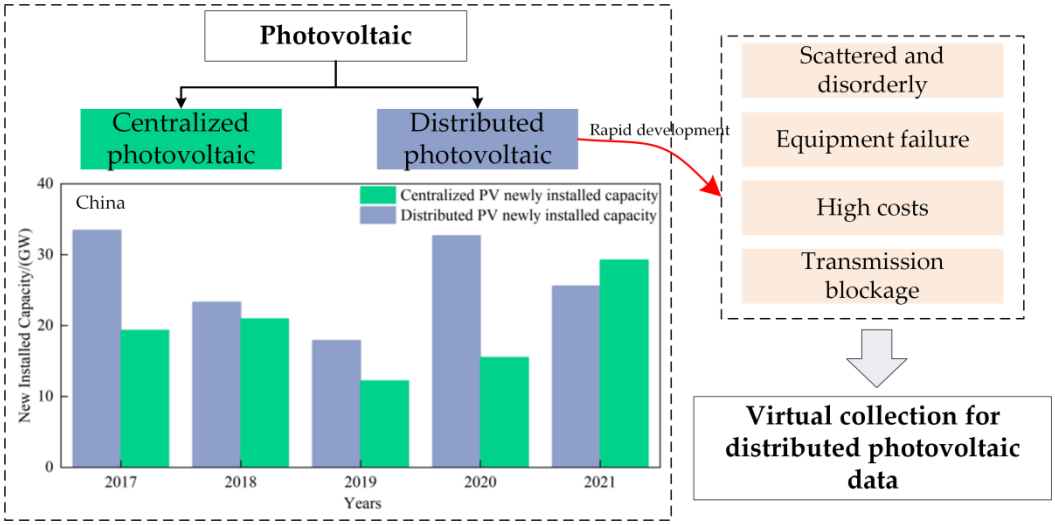

Figure 2. Research motivation of virtual collection technology.

Aware of this need, relevant scholars have been inspired by the virtual collection concept [4] and have researched virtual collection for DPVS. The core idea of virtual collection for DPVS is to use the power data of selected reference power stations (RPSs) in the region as input to infer the power data of other stations through computational intelligence algorithms. To reduce the number of sensors, Ref. [5] fitted all DPVS station operation data in the selected region by a deep recurrent denoising autoencoder and

introduced a bionic artificial neural network to dynamically select the best subset of reference stations in the set of candidate RPSs.

It can be seen that there are relatively few studies combining artificial intelligence algorithms for the virtual collection of DPVS data, an approach which is still in the early exploration stage. Moreover, there is no comprehensive introduction to current knowledge in virtual collection research, resulting in a lack of attention to the virtual collection of DPVS in the industry. Various methods suitable for virtual collection and application scenarios of the virtual collection have not been summarized. This paper aims to rectify this by providing a comprehensive review of virtual collection and bringing the attention of more scholars to this technology. Therefore, we will examine the existing research in this field to achieve a comprehensive overview, generalization, and summary of virtual collection technology, particularly in relation to the challenges in its implementation, with the following specific contributions.

1. This paper explains the specific definition of virtual collection for DPVS, gives the prerequisites for implementing virtual collection and refines virtual collection technology into three key steps. In addition, considering the current lack of research on the virtual collection, this paper illustrates the challenges faced by virtual collection technology through problem analogy.
2. Given the challenges faced by virtual collection for DPVS data, this paper summarizes the methods applicable to DPVS similarity analysis, reference power station selection, and DPVS system data inference in various fields to provide theoretical support for the development of virtual acquisition technology.
3. In this paper, according to the characteristics of virtual collection for DPVS, four application scenarios of DPVS virtual collection technology are proposed, giving diversified practical application values to the virtual collection technology.
4. To the authors' knowledge, this paper is the first comprehensive introduction, generalization, and summary of DPVS virtual collection and can provide a theoretical reference for subsequent research on DPVS virtual collection methods.

The rest of the paper is organized as follows: Section 2 gives an overview of DPVS virtual collection; Section 3 elaborates on the key steps and corresponding challenges of the virtual collection; Section 4 summarizes the specific methods applicable to DPVS virtual collection; Section 5 introduces four application scenarios to which virtual collection technology is adapted; and the conclusions are presented in Section 6.

2. Overview of DPVS Virtual Collection

The word "virtual" in the virtual collection indicates that the technology does not collect PV data through collection equipment such as sensors, collectors, and concentrators in the field. The virtual collection is a new type of inference technology for data that cannot be collected in real-time or is difficult to collect. Its essence is the system identification and state estimation of a large system composed of multiple subsystems [5].

As shown in Figure 3, this paper divides the virtual collection process into three steps according to the implementation conditions of DPVS virtual collection:

1. Similarity analysis of DPVS in the virtual collection region.
2. Selection of reference power stations (RPSs) for virtual collection.
3. Data inference of DPVS, i.e., accurate estimation of the output of all power stations in the region through computational intelligence.

Firstly, the data of all power stations need to be transmitted to the company's PV intelligent operation and maintenance cloud platform through 5G, ZigBee, and other wireless communication means. Then, the similarity between each power station in terms of geography, equipment, and climate is analyzed by the similarity analysis method to obtain the set of power stations that meet the prerequisites for virtual collection. Further, the best RPSs in the whole PV system are selected through clustering or intelligent optimization algorithms to deploy the sensing equipment so as to accurately estimate the operation

data of the whole PV system. In this paper, the stations configured with complete and normal operation PV data collection devices and are similar to the current stations to be collected are called RPSs. Finally, the data inference model is built through the PV intelligent operation and maintenance cloud platform to implement virtual collection.

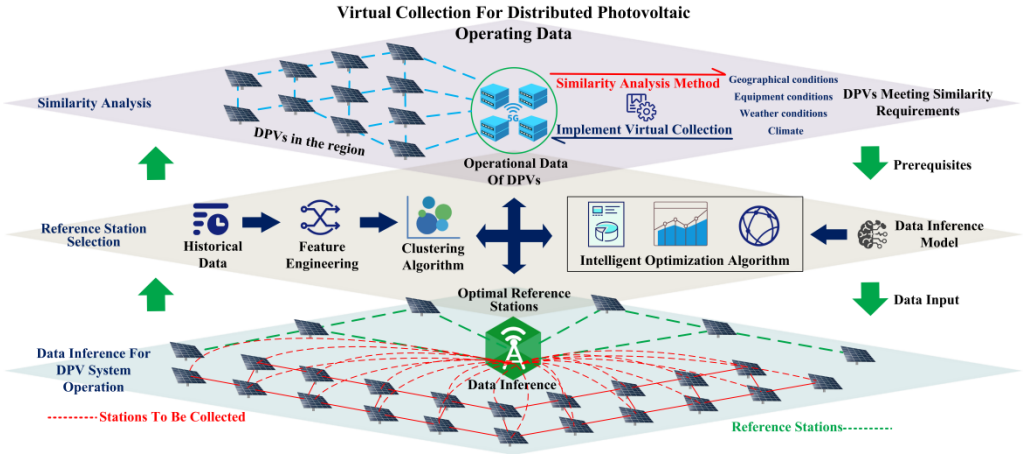

Figure 3. Schematic diagram of the DPVS virtual collection process.

The above analysis shows that the following conditions need to be met to implement DPVS virtual collection: (1) real-time operational data of the RPS is known; (2) the installation information of the power stations used in virtual collection is known, such as installed capacity, installation inclination, etc.; and (3) the meteorological factors, terrain conditions and various parameters of equipment of selected RPSs are similar to those of power stations to be virtually collected. The following sections will elaborate on the detailed virtual collection steps and challenges.

3. Process and Challenges of DPVS Virtual Collection

The virtual collection of DPVS data is a new field that few scholars have studied. Therefore, in this section, to facilitate understanding, we compare and analyze the similarities and differences between the steps of virtual collection and other methods. Focusing on their similarities, we give directions for DPVS virtual collection research, and then focusing on their differences, we summarize the challenges faced by DPVS virtual collection. It is worth noting that this approach to elaboration is novel for the review literature and can help the reader understand the connections and differences clearly between virtual collection and other studies.

3.1. Similarity Analysis of Regional DPVS

PV data is most closely correlated with external conditions, and factors such as the geographic location of the installation site. Environmental factors have a significant impact on the accuracy of the virtual collection model. Therefore, one of the prerequisites assumed for the realization of virtual collection is that the station to be collected and the RPS have similar external factors. From the data point of view, we want the data set to obey similar distribution patterns as much as possible, thus providing higher-quality input data for supervised learning. This similarity can make the virtual collection more robust and ensure the virtual collection data's accuracy even in weather changes.

To illustrate, Badong County in southwestern Hubei Province, China, and Jiangning District in Nanjing City, Jiangsu Province, produce widely different power data due to different terrain, topography, meteorology and other conditions. Figure 4 shows the power output of PV stations in Badong County and Jiangning District on a typical summer day. For the DPVS of Jiangning district, using the DPVS operation data of Badong district for data inference would seriously reduce the accuracy of the virtual collection because of the

extremely low similarity between them. Therefore, it is necessary to define clusters of PV stations that satisfy the similarity requirement by similarity analysis in advance.

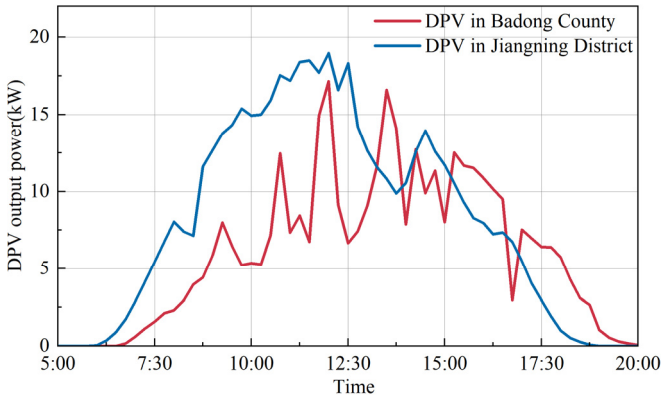

Figure 4. Badong county and Jiangning district DPVS output on a typical day.

Many factors affecting the PV output state are coupled with each other [6]. The main factors influencing the solar energy conversion process are shown in Figure 5. It can be seen that the degree of solar irradiance received by the PV module is significantly influenced by the geographical location and meteorological conditions. The climate is the comprehensive pattern in the general state of the atmosphere and weather processes in a certain area on a long timescale, which is an important factor affecting the level of light resources, and meteorology refers to the physical phenomena of the atmosphere on a short time scale, such as temperature, clouds, etc. Secondly, the link of solar irradiance to power for conversion is closely related to the selection of equipment, the design of the station, and electrical efficiency. After the series of the energy conversion process mentioned above, the final PV power output is obtained. Therefore, similarity analysis can be performed from two perspectives: influencing factors (causes) and power output trends (results). However, from the perspective of influencing factors, it is difficult to analyze the similarity due to the large number of factors affecting PV output, the significant difference between the dimensions, and the complex types of characteristics. From the perspective of the PV output trend, the trend changes are complicated, and the time scale is long, which makes it challenging to analyze the trend characteristics.

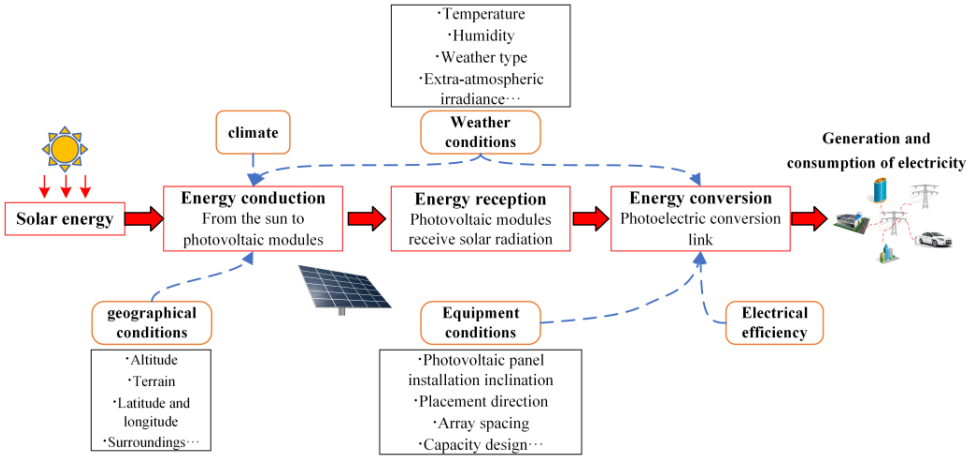

Figure 5. Factors affecting DPVS power output.

The importance of data similarity is also reflected in many areas of research. In Ref. [7], an anomaly identification and reconstruction model based on curve similarity analysis with a BP neural network is proposed for detecting anomalous and compensating missing PV historical data. Similar to the virtual collection, the method also requires the power

of neighboring PV stations. Considering the periodicity of PV power, Ref. [8] proposes a data cleaning method based on approximate periodic time series, effectively improving the quality of PV data. Considering the uncertainty of PV power generation due to the variation in weather conditions, Ref. [9] proposes a prediction framework combining similar day selection techniques. In this framework, the authors first screen external variables that can accurately capture the similarity between different days and select dates with higher similarity based on these external variables for the historical day and the day to be predicted, thus improving the prediction accuracy. Although the research methods of the above studies are different, they all desire to obtain higher-quality data. Therefore, we can mine the similarity analysis methods suitable for virtual collection from these studies.

3.2. RPS Selection for Virtual Collection

Selecting the RPSs is the most crucial step in the virtual collection process. The RPSs' real-time power data will be input into the computational intelligence algorithm as multidimensional features to estimate the output of all regional DPVS. We aim to select the subset of DPVS among the regional DPVS that can estimate the data of other stations with higher accuracy. The key to this step is to identify selective sensor locations where the most important data is collected to monitor the status of all DPVS. Therefore, we will analyze the differences and associations between the RPS selection step and other methods from the perspectives of input data and equipment placement.

As shown in Figure 6, from the perspective of input data, the selection of RPSs can be approximated as the feature selection problem of machine learning. Both aim to improve the accuracy of the results as much as possible by selecting RPSs (features). Therefore, although there are few studies on the selection of reference power stations, the relatively mature feature selection theory can also provide us with inspiration. For data mining techniques, the feature quality of input data seriously affects the model's performance, so many scholars have researched the feature selection problem. Ref. [10] systematically examined the existing sparse learning models for feature selection from the perspective of individual sparse feature selection and group sparse feature selection. It analyzed the differences and connections among various sparse learning models. Ref. [11] proposes a new incremental feature selection that makes the method robust to dynamically ordered data. Ref. [12] proposes a grasshopper optimization algorithm that can solve the binary optimization problem by selecting a subset of features that can better characterize the data attributes from a large set of original features, thus improving the classification accuracy. The above studies proposed effective processing for the feature selection problem, which can provide some theoretical reference for selecting RPSs, such as transforming the RPS selection into a combinatorial optimization problem. However, it is worth noting that if a power station is selected as the RPS, it is used as the input feature, and the remaining power stations are used as the power stations to be collected. It can be seen that the RPS selection problem for virtual collection is similar to the high-dimensional feature selection problem [13] yet different from the feature selection in the traditional regression and classification [14] problems. Therefore, choosing reasonable RPSs is more challenging than feature selection.

From a sensor placement perspective, the selection of RPSs can also be inspired by the data aggregation point (DAP) selection problem in smart meters. As shown in Figure 7, both DAP selection and RPS selection can be regarded as the optimal configuration of transmission nodes in the system. DAPs are selected to reduce data redundancy and bandwidth requirements by aggregating data locally at the sensor or intermediate nodes to form high-quality information and reduce the quality of packets sent to the base station, thus saving energy and bandwidth. Ref. [15] treats DAP placement as a mixed integer programming problem and proposes a new heuristic algorithm to minimize installation, transmission, and delay costs to select the optimal DAP placement location. Ref. [16] proposes an improved k-means clustering algorithm to assign DAPs, significantly reducing the number of DAPs installed.

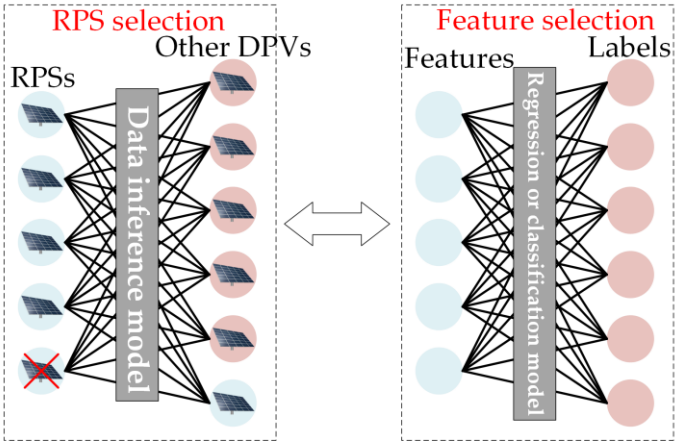

Figure 6. RPS selection and feature selection process.

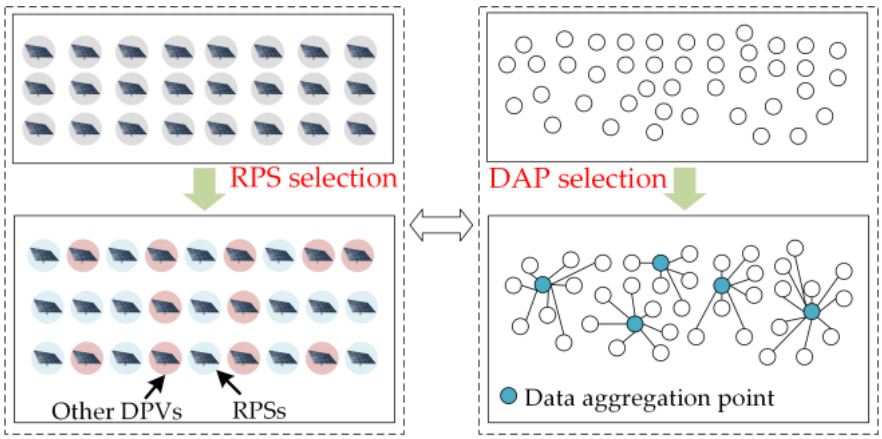

Figure 7. RPS selection and DAP selection process.

Although there are certain commonalities between the selection of DAPs and RPSs, there are still many challenges that need to be studied. Data aggregation points are obtained with the objective of determining the lowest transmission and delay cost among all SM layout points to achieve aggregation and transmission of data for the whole system. The RPS is selected by selecting a subset of PVs among the regional PV systems to achieve a data estimation of the whole system. Therefore, the elements considered in the selection of RPSs are more diversified. In addition to communication and equipment costs, the accuracy of data estimation for the whole system from different RPS sets needs to be considered, as well as the time and space coupling characteristics, which are lacking in the current study.

3.3. Data Inference for Regional DPVS

The final step of the virtual collection technique is to infer the operational data of the whole DPVS through an artificial intelligence algorithm. This step maps the relationship between the RPS and the whole system by building a computational intelligence model between the RPSs and the power stations to be collected in the region, using the data from the RPSs selected in the second step as the input. This step is similar to the method used in PV prediction techniques, both of which require certain historical data as a driver to obtain the unknown PV output power.

There is relatively little research in the industry on DPVS virtual data inference, with most studies focusing only on PV power prediction, using historical data, real-time weather, and other environmental information to predict PV power output. Thankfully, the current DPVS power prediction algorithms are relatively mature and can provide some theoretical references for virtual DPVS data collection. However, it is worth noting that data inference in virtual collection differs from traditional PV prediction in model construction and use.

Virtual data collection estimates the current PV power output in real time through a data inference model, whereas the PV predictor estimates the future power output. The input to the virtual collection model is real-time PV data from the RPSs, and the input to the PV predictor is historical operational data and environmental information. This real-time nature makes it necessary that the data inference model for virtual collection has better robustness and higher accuracy requirements than that for PV prediction.

4. Methods for DPVS Virtual Collection

The previous section introduces the specific implementation steps of virtual collection and its purpose, and pinpoints the urgent need to provide solutions to the challenges faced by the above steps. Therefore, this section provides theoretical support for the development of virtual collection technology by summarizing the methods applicable to DPVS similarity analysis, RPS selection, and DPVS data inference in various fields. Various methods for DPVS virtual collection are summarized in Figure 8.

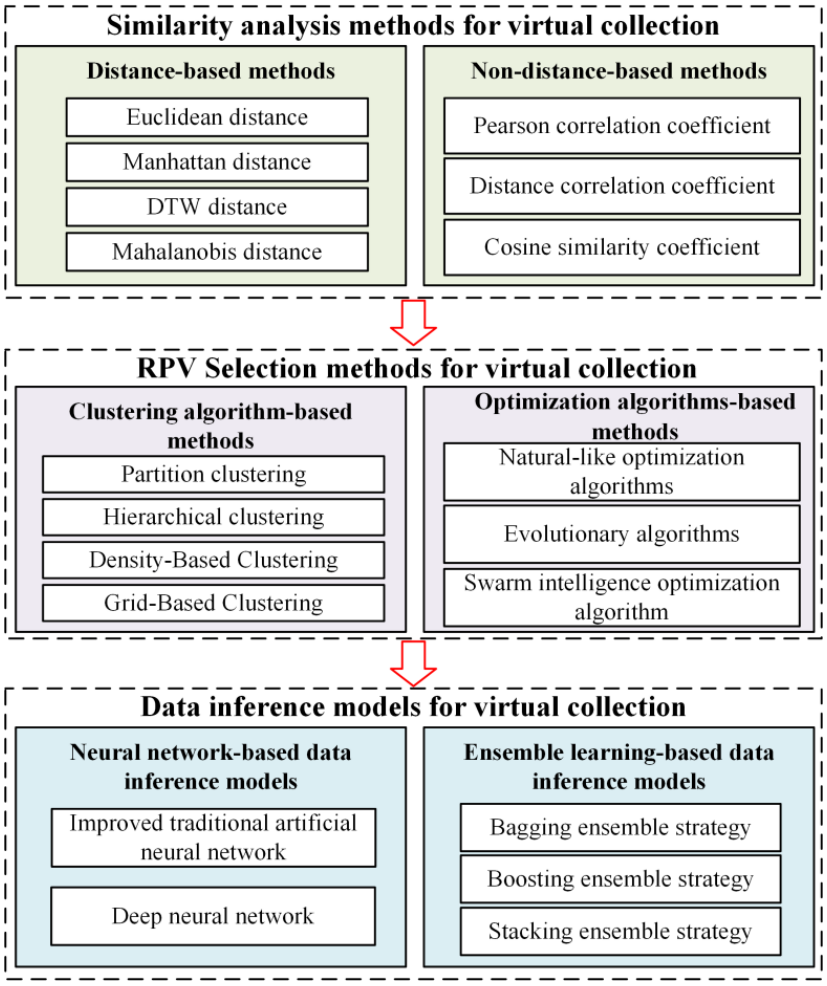

Figure 8. Summary of methods for virtual collection.

4.1. Similarity Analysis for Virtual Collection

Similarity analysis refers to mining the association relationship between different objects based on the attribute values of the data. Many similarity analyses have been widely used in multivariate statistics, machine learning, and other fields. For the characteristics of DPVS operation data, this section divides similarity analysis into two categories:

- Distance-based similarity analysis;
- Non-distance-based similarity analysis.

Table 1 summarizes various similarity metrics and the corresponding references.

Table 1. Summary of similarity analysis and references.

Methods		References
Distance-based similarity analysis	Minkowski distance	[17–19]
	Mahalanobis distance	[20,21]
	DTW distance	[22–24]
Non-distance-based similarity analysis	Pearson correlation coefficient	[25,26]
	Distance correlation coefficient	[27–29]
	Cosine similarity coefficient	[30,31]
	Other similarity coefficients	[32–34]

4.1.1. Distance-Based Similarity Analysis

The distance-based similarity analysis evaluates the similarity based on the distance between samples, and the closer the distance, the higher the similarity. According to different criteria, the commonly used metrics can be classified as Euclidean distance [17], Manhattan distance [18], Mahalanobis distance [21], and DTW distance [22].

In this paper, the Euclidean distance and Manhattan distance are uniformly defined by the Minkowski distance measure. The difference between the two in two-dimensional space is shown in Figure 9. For two given m-dimensional samples, $X^i = \left\{x_1^i, x_1^i, \cdots, x_m^i\right\}$ and $X^j = \left\{x_1^j, x_1^j, \cdots, x_m^j\right\}$, the Minkowski distance can be calculated by the following equation:

$$d(X^i, X^j) = \sqrt[p]{\sum_{k=1}^{m} \left|x_k^i - x_k^j\right|^p} \quad (1)$$

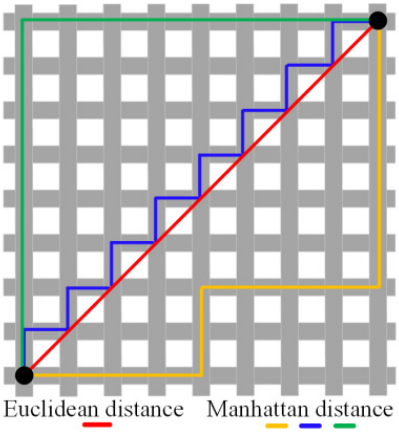

Figure 9. Schematic diagram of Minkowski distance when p is different.

When p = 1, Equation (1) is the Manhattan distance. As shown in Figure 8, the Manhattan distance represents the sum of the absolute axis distances of two points on the standard coordinate system, so it is also called the city block distance or cab distance and is widely used as a similarity measure for various samples. When p = 2, Equation (1) is the Euclidean distance. The Euclidean distance is one of the most easily understood distance algorithms representing the straight-line distance between two points in Euclidean space. Euclidean distance is also one of the most commonly used similarity analyses, and many scholars have conducted research based on it. For example, Ref. [19] clusters the historical PV generation data by Euclidean distance to obtain the similarity day matrix and corrects the predicted values, thus improving the prediction accuracy.

The Mahalanobis distance represents the covariance distance of the data and is an effective method to calculate the similarity of two sample sets at two locations. The Mahalanobis distance between X and Y samples can be calculated by Equation (2). It is noteworthy that it takes into account the connection between various characteristics and is scale-independent, overcoming the influence of the magnitude and feature distribution on the distance measurement in the Mahalanobis distance mentioned above. Ref. [20] applied the Mahalanobis distance to cluster the prediction errors and meteorological factors to improve the accuracy of PV power interval prediction, which is the only study that the authors could retrieve so far in the field of PV.

$$d(X,Y) = \sqrt{(X-Y)^T S^{-1}(X-Y)} \tag{2}$$

Dynamic time warping (DTW), known as the "time-warped" distance measure, is one of the most popular time series similarity metrics and is calculated as shown in Figure 10. For distance measurement of two series, Minkowski distance requires strict alignment, whereas DTW relaxes this alignment restriction and can capture the trend similarity between different series very well in time series similarity analysis. Ref. [23] used weighted DTW to measure the similarity of meteorological factors between the days to be predicted and the historical days. Ref. [24] provides a standardized approach for PV system design and analysis through the DTW similarity analysis method. Moreover, it has been proved experimentally that the DTW algorithm is a simple and effective time series similarity analysis. Therefore, in virtual collection technology, DTW can analyze the time series similarity of different DPVS.

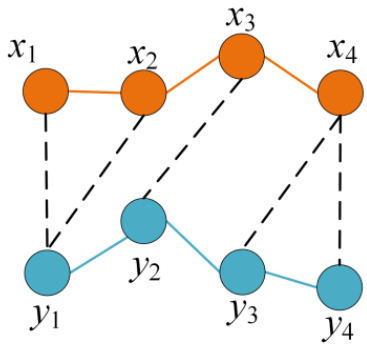

Figure 10. The calculation principle of DTW.

4.1.2. Non-Distance-Based Similarity Analysis

The similarity analysis based on non-distance methods is carried out by the similarity measure function, and generally speaking, the larger the value of the similarity function, the higher the similarity. Common similarity metric functions include the Pearson correlation coefficient, distance correlation coefficient, and cosine similarity.

The Pearson correlation coefficient is used to measure the degree of linear correlation between two variables. It has been widely used in various research fields such as environment, biology, chemistry, agronomy, and electricity, with as many as 30,703 search results in WoS. The Pearson correlation coefficient between two variables can be calculated by Equation (3). Ref. [25] applied the Pearson correlation coefficient to select data samples similar to the target prediction date as the training samples of the extreme learning machine, which effectively improved the prediction accuracy of PV power. Ref. [26] used the Pearson coefficient to extract the main features that affect photovoltaic power generation output the next time and divided the training set into different groups according to the similarity of each feature to improve the prediction accuracy. The Pearson correlation coefficient is calculated as:

$$\rho_{X,Y} = \frac{\mathrm{cov}(X,Y)}{\sigma_X \sigma_Y} = \frac{\sum_{i=1}^{T}(X_i-\overline{X})(Y_i-\overline{Y})}{\sqrt{\sum_{i=1}^{T}(X_i-\overline{X})^2}\sqrt{\sum_{i=1}^{T}(Y_i-\overline{Y})^2}} \tag{3}$$

where cov(*X*,*Y*) represents the covariance coefficient between *X* and *Y*, and σ represents the standard deviation of the sample.

However, the Pearson correlation coefficient has a significant drawback in that it only captures the degree of linear correlation between different variables. Therefore, Szekely [27] and others defined the concept of the distance correlation coefficient, which can reflect the degree of nonlinear correlation between different variables and compensate for the shortcomings of the Pearson correlation coefficient. The distance correlation coefficient, which can be calculated by Equations (4)–(7), has been widely used to mine the degree of nonlinear correlation between different samples. Ref. [28] applied distance correlation coefficients to analyze the nonlinear correlation between wind power and numerical weather forecast data. Ref. [29] extracted feature vectors by distance correlation coefficients and verified the validity by two industrial cases. The distance correlation coefficient is calculated as:

$$R^2(X,Y) = \frac{v^2(X,Y)}{\sqrt{v^2(X,X)v^2(Y,Y)}} \tag{4}$$

$$v^2(X,Y) = \frac{1}{n^2}\sum_{i,j=1}^{n} A_{i,j}B_{i,j} \tag{5}$$

$$A_{i,j} = \|X_i - X_j\|_2 - \frac{1}{n}\sum_{k=1}^{n}\|X_k - X_j\|_2 - \frac{1}{n}\sum_{l=1}^{n}\|X_i - X_l\|_2 + \frac{1}{n^2}\sum_{k,l=1}^{n}\|X_k - X_l\|_2 \tag{6}$$

$$B_{i,j} = \|Y_i - Y_j\|_2 - \frac{1}{n}\sum_{k=1}^{n}\|Y_k - Y_j\|_2 - \frac{1}{n}\sum_{l=1}^{n}\|Y_i - Y_l\|_2 + \frac{1}{n^2}\sum_{k,l=1}^{n}\|Y_k - Y_l\|_2 \tag{7}$$

where $R^2(X,Y)$ is the distance correlation coefficient of two variables *X* and *Y*. X_i, X_j, X_k, X_l are the *i*-th, *j*-th, *k*-th, and *l*-th samples of variable *X*. Y_i, Y_j, Y_k, Y_l are the *i*-th, *j*-th, *k*-th, and *l*-th samples of variable *Y*. Both $v^2(X,X)$ and $v^2(Y,Y)$ can be calculated by the above equations.

The cosine similarity function measures the magnitude of the difference between two individuals using the cosine of the angle between two vectors in the vector space. The calculation formula is shown in Equation (8). Unlike the traditional distance calculation, the cosine similarity function focuses more on the difference in the direction between two vectors. It is an effective method to measure the similarity between the trends of data series changes. Ref. [30] used cosine similarity to measure the correlation between reactive power optimization-related factors and historical load data to assist in finding reactive power optimization solutions. In Ref. [31], the importance of pixels in the image features of photovoltaic cells was measured by cosine similarity to achieve defect detection in photovoltaic cells.

$$CS(X,Y) = \frac{X \cdot Y}{\|X\|\|Y\|} = \frac{\sum_{i=1} X_i Y_i}{\sqrt{\sum_{i=1} X_i^2}\sqrt{\sum_{i=1} Y_i^2}} \tag{8}$$

With the extensive research on similarity analysis methods, many scholars use other effective similarity analysis functions in the PV field. Ref. [32] combined gray correlation analysis with k-means to determine the similarity day and the optimal similarity day for the forecast day. Ref. [33] proposed a radiometric coordinate analysis method for analyzing the correlation between various influencing factors and PV power output in different periods and weather conditions. In Ref. [34], the nonlinear effect of copula function and propensity correlation measurement is used to extract the critical meteorological factors that affect wind and PV power generation to improve the interval prediction accuracy of PV power generation.

In summary, the similarity analysis methods applied to virtual collection are innovatively grouped into two categories in this paper: distance-based similarity analysis and non-distance-based similarity analysis. Combined with the analysis in Section 3.1, the Minkowski distance can roughly analyze the similarity between different power stations.

Considering the complexity of the dimensions and types of influencing factors, we can use the Mahalanobis distance to overcome the influence brought by different magnitudes. Furthermore, we can also perform the analysis from the perspective of power output trends, such as using DTW to extract the time characteristics of different PVs output power or using the Pearson correlation coefficient and cosine similarity function to obtain the correlation degree between different PVs. In the future, it is necessary to verify these approaches using physical models.

4.2. RPS Selection Methods for Virtual Collection

Based on the above overview of virtual collection, it can be learned that RPS selection is similar to the feature selection problem in machine learning and the sensor DAP optimization problem. The essence is to select the most representative DPVS among DPVS so that the operational data of other DPVS can be well estimated. Therefore, this paper summarizes the methods that can be applied to RPS selection in two categories:

- Clustering-based algorithms;
- Optimization-based algorithms.

Table 2 summarizes the RPS selection methods and the corresponding references.

Table 2. Summary of algorithms applicable to RPS selection.

Methods		References
RPS selection methods based on clustering algorithms	Partition clustering	[35–40]
	Hierarchical clustering	[41–43]
	Density-based clustering	[44–48]
	Grid-based clustering	[49–51]
RPS selection methods based on optimization algorithms	Natural-like optimization algorithms	[52–58]
	Evolutionary algorithms	[59–64]
	Swarm intelligence optimization algorithm	[65–71]

4.2.1. RPS Selection Based on Clustering Algorithm

The clustering algorithm is an unsupervised learning algorithm in machine learning, and it is used in a wide range of scenarios in data science. Its principle is to partition a data set into different classes or clusters according to specific criteria (e.g., distance). Effective clustering methods have been proposed by many scholars, which can be classified as partition clustering [35], hierarchical clustering [41], density-based clustering [44], and grid-based clustering [49], depending on the principle. Among them, cluster centroids, which can be considered representatives of clusters, can well characterize the data in the clusters. Therefore, in the selection of RPSs, the historical operation data of DPVS can be clustered by clustering algorithms so that the DPVS within each cluster with the closest distance to the cluster centroid can be selected as the RPS.

Partition clustering is one of the most commonly used methods in the study of time series clustering algorithms and is usually achieved with the help of metric similarity. Among them, the similarity metric function can use various distance metric functions as described in Section 3.1, and the mainstream methods include k-means [36], k-medoids [37], and fuzzy c-means [38]. Ref. [39] selected the best DAP placement through a k-means clustering algorithm, effectively reducing the overhead of the data aggregator. Ref. [40] aggregated wireless sensors through k-means clustering to reduce the number of transmission nodes.

The basic idea of hierarchical clustering is to iteratively merge or split a given set of data into nested class hierarchy results or class spectral graphs. It has the advantage that the number of clusters does not have to be specified in advance, and the results of clustering at different levels can be obtained. Ref. [42] selects the aggregation node with the lowest cost and highest efficiency through the hierarchical clustering algorithm. Ref. [43]

proposes an extended hierarchical clustering method for energy-harvesting mobile wireless sensing networks, thus extending the lifetime of the network sensors. It can be seen that if the RPS is selected by the hierarchical clustering algorithm without specifying the number of RPSs.

Density-based and grid-based clustering are often used in combination to divide the data space into grid cells, map the set of data objects into grid cells, and cluster them according to the density of each cell. When oriented to time series data, the time series data is first transformed into another data form [45,46], making it applicable to traditional clustering algorithms in data mining, including DBSCAN [47], OPTICS [48], STING [50], and Wave Cluster [51]. It is worth noting that the distance metric has failed in many cases in a high-dimensional space, so density-based and grid-based clustering algorithms are suitable for a large number of DPVS clusters in a region. However, it is difficult to control the performance of virtual collection for a large number of DPVS clusters, so the density- and grid-based clustering algorithms can be used to divide a large region into multiple sub-regions. Therefore, the PV clusters within each region satisfy the similarity requirement, and then the RPV selection is performed by partition and hierarchical clustering as described above.

4.2.2. RPS Selection Based on Intelligent Optimization Algorithm

The advantage of RPS selection based on a clustering algorithm is that the RPSs can be selected directly based on the data characteristics by an unsupervised algorithm without the need to be combined with the data inference model. Although this method is simple and fast, it is difficult to guarantee the accuracy of virtual collection in practical applications. Therefore, we can also consider the selection of the reference power station as a binary optimization problem, which is solved based on the intelligent optimization algorithm. This approach must be combined with the data inference model to optimize the virtual collection performance evaluation index as a fitness function. It has the advantage of having a solid search capability, thus finding the best combination of RPSs. In addition, constraints can be set to simultaneously optimize the parameters of the data inference model, thus ensuring the performance of the virtual collection.

The intelligent optimization algorithm is a kind of stochastic search algorithm based on biological intelligence or natural phenomena, and the current research on intelligent optimization algorithms is very extensive. In this paper, these intelligent optimization algorithms are classified into three categories:

- Naturalistic optimization algorithms;
- Evolutionary algorithms;
- Swarm intelligence optimization algorithms.

Naturalistic optimization algorithms are intelligent optimization algorithms that simulate various natural phenomena and various laws of physics. The most typical representative is the simulated annealing optimization algorithm proposed by the American physicist Metropolis based on the annealing process of solids [52]. Subsequently, Erol, OK proposed the Big Bang Grand Convergence algorithm in 2006 [53], Rashedi E proposed the Gravitational Search algorithm in 2009 [54], H. Shareef proposed the Light Search algorithm in 2015 [55], and Biyanto, TR proposed the Rainwater algorithm in 2019 [56]. Moreover, this class of algorithms is still increasing and is widely used in various configuration optimization problems. Ref. [57] proposed a feature selection method based on a biogeography-based optimization algorithm and added binary coding to enhance the algorithm performance. Ref. [58] proposes a feature selection algorithm based on a binary gravity search algorithm and incorporates mutual information to improve the search efficiency. These improved strategies can provide inspiration for the selection of RPSs in virtual collection.

Evolutionary algorithms are intelligent optimization algorithms that simulate the evolution of natural organisms in the process of reproduction through genetic variation and the natural law of "survival of the fittest." Four typical evolutionary algorithms have been

studied, namely, genetic algorithms based on natural selection and evolutionary mechanisms [59], evolutionary strategy algorithms [60], genetic programming algorithms [61], and differential evolutionary algorithms [62]. Evolutionary algorithms are robust methods widely used in different industrial problems, and many scholars have improved on the original algorithms. A genetic algorithm-based feature selection method is proposed in [63]. Ref. [64] introduced a hybrid mechanism in the differential evolution algorithm and further enhanced the feature subset obtained from the binary differential evolution (BDE) algorithm by using the local search method. It is worth noting that since the RPS selection is a binary optimization problem, while most evolutionary algorithms are applicable to both binary and decimal problems, no conversion of variables is required to complete the RPS selection. Therefore, evolutionary algorithms should be focused on in the RPS selection problem.

The swarm intelligence optimization algorithm is an intelligent optimization algorithm that simulates the survival behavior of gregarious species in nature. In 1996, M. Dorigo [65] proposed the ant colony optimization (ACO) algorithm by simulating an ant colony to choose the shortest path from an anthill to a food source to avoid obstacles. In 1995, Kennedy, an American psychologist, proposed the particle swarm optimization (PSO) algorithm inspired by the predatory behavior of bird flocks [66]. Later, other scholars proposed the bat algorithm [67], the whale optimization algorithm (WOA) [68], and the Harris hawk optimization (HHO) algorithm [69]. These algorithms exhibit advanced and complex functions through mechanisms such as cooperation, competition, interaction, and learning. Their potential parallelism and distributed characteristics give them significant advantages in handling big data. Ref. [70] proposes a whale optimization feature selection method considering a stability index. Ref. [71] proposes a binary version of a hybrid Gray Wolf Optimization (GWO) and Particle Swarm Optimization to solve the feature selection problem. These studies provide inspiration for improvements in swarm intelligence optimization algorithms to solve the RPS selection problem.

From the above analysis, it can be seen that there are many naturalistic and swarm intelligence optimization algorithms with high efficiency and global search capability. Since the type of DPVS is divided into RPS and non-RPS, the selection of RPSs can be regarded as a combinatorial optimization problem. However, most optimization algorithms are oriented toward continuous values and are not directly applicable to the selection of RPSs. In order to apply the above intelligence optimization algorithms to the RPS selection problem, this paper proposes a typical binary conversion scheme that converts the positions of search agents into binary values through the conversion function shown in Equations (9) and (10). More binary optimization schemes can be found in [71–73]. The conversion is calculated as follows:

$$T_s(X^d(t)) = \frac{1}{1 + e^{-X^d(t)}} \tag{9}$$

$$\widetilde{X}^d(t) = \begin{cases} 1 & rd < T_s(X^d(t)) \\ 0 & rd \geq T_s(X^d(t)) \end{cases} \tag{10}$$

where $X^d(t)$ denotes the position of the d-th variable updated according to the original equation, $\widetilde{X}^d(t)$ denotes the position of the d-th binary variable updated by the transfer function, and rd is a random number of [0, 1]. When $\widetilde{X}^d(t)$ is 0, the d-th station is the station to be collected, and when $\widetilde{X}^d(t)$ is 1, the d-th station is the RPS.

In summary, two types of schemes suitable for selecting virtual collection RPSs are summarized in this paper. The use of the optimization algorithm is superior in terms of accuracy because it incorporates the data inference model but accordingly results in a substantial amount of computation. Moreover, using optimization algorithms may lead to overfitting problems because they rely too much on the data involved in the training. Therefore, another key to improving virtual collection accuracy is using training data with the largest possible time scale when selecting RPSs. For a larger number of DPVS, the clustering algorithm can be combined with the intelligent optimization algorithm to first

divide the DPVS into different sub-regions by the clustering algorithm and then select the best RPSs in each region by the intelligent optimization algorithm.

4.3. Data Inference Models for Virtual Collection

Data inference for DPVS virtual collection essentially takes operation data of the RPSs as input and estimates the data of the whole DPVS in real-time by regression models. With their excellent fitting performance, models such as long short-term memory (LSTM), random forests (RF), and Light Gradient Boosting Machine (LightGBM) have become the most widely used methods at present [74]. Therefore, this paper classifies the current mainstream algorithms of data inference models applicable to DPVS virtual collection into two categories:

- Neural network-based models;
- Ensemble learning-based models.

It is worth noting that most of the studies surveyed in this paper are devoted to PV power prediction due to the lack of data inference models for virtual collection. However, they both belong to different application scenarios of regression models, and thus can provide references to each other. Table 3 summarizes data inference models and the corresponding references.

Table 3. Summary of data inference models and references.

Methods		References
Neural network-based data inference models	Improved traditional artificial neural network	[75–78]
	Deep neural network	[79–87]
Ensemble learning-based data inference models	Bagging ensemble strategy	[88–90]
	Boosting ensemble strategy	[91–96]
	Stacking ensemble strategy	[97–99]

4.3.1. Data Inference Based on Neural network

A classical neural network usually consists of an input layer, a hidden layer, and an output layer, which can store complex mapping relationships through learning without prior knowledge of the specific mathematical expressions of the inputs and outputs. The learning parameters in its network usually adopt a back-propagation strategy to find the combination of parameters that minimizes the network error with the help of the most rapid gradient information. The structure of a typical neural network model is shown in Figure 11, and the specific calculation process is shown in Equations (11)–(15):

$$h = W_1 \cdot x + b_1 \tag{11}$$

$$y = W_2 \cdot h + b_2 \tag{12}$$

$$Loss = \frac{1}{2}\sum_{i=1}^{N}(y - \hat{y})^2 \tag{13}$$

$$W \leftarrow \eta \frac{\partial Loss}{\partial W} \tag{14}$$

$$b \leftarrow \eta \frac{\partial Loss}{\partial b} \tag{15}$$

where x denotes the input data, h denotes the data of the hidden layer, and η denotes the step size of gradient descent. y and $\hat{y}$ represent the virtual collection data and actual data, respectively. W_1 represents the weight between the input and hidden layer, and W_2 represents the weight between the hidden and output layers. b represents the bias.

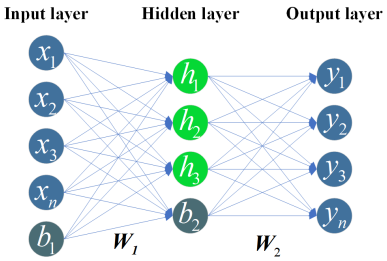

Figure 11. Typical neural network structure diagram.

With the increase in training data, the traditional neural network can hardly meet the demand for efficiency and accuracy, so scholars have made many improvements to traditional neural networks. Broomhead and Lowe introduced the radial basis function into the neural network in 1988 to create the radial basis function (RBF) neural network [75].

Subsequently, Zhang et al. proposed wavelet neural networks (WNN) [76], Jiao et al. proposed multiwavelet neural networks [77], and Yang et al. proposed ridge wave neural networks [78]. These models showed good potential and value for applications in non-smooth, nonlinear, non-Gaussian signal and image processing.

With the improvement of computer processing speed and storage capacity, the design and implementation of deep neural networks have gradually become possible, and the field of machine learning has entered the era of "deep learning" [79]. Hinton et al. proposed a two-stage strategy based on "layer-by-layer pre-training" and "fine-tuning" to solve the problem of training network parameters in deep learning [80]. Subsequently, LeCun and Geoffrey Hinton proposed self-encoder structures [81], deep belief networks [82], and convolutional neural networks [83]. To overcome the "gradient explosion problem" in traditional recurrent neural networks, Hochreiter et al. proposed long- and short-term memory networks [84].

The above methods have been widely applied to data prediction and inference. Ref. [85] combined deep belief networks with a gray theory-based data preprocessor for predicting the power generation of PV plants on the same day. Ref. [86] mines the critical features of DPVS data by convolutional graph neural networks and captures the time-dependent features by long short-term memory networks to effectively improve the solar irradiance prediction accuracy. In Ref. [87], a combined wavelet neural network model combining improved particle swarm and chaotic optimization algorithms is proposed for short-term load prediction of integrated energy systems, which significantly improves prediction accuracy when comparing the traditional artificial neural network model with wavelet neural network model. For virtual collection, Ge et al. proposed an artificial neural network with affine optimization [5], which endows the neural network with the ability to output PV power with uncertainty and thus reduces the impact of the overfitting phenomenon of an artificial neural network on systems with uncertain solid variables. It can be seen that the good performance of neural networks in PV power prediction has been verified by many scholars. Pioneering research has already been performed in the application of neural networks in virtual collection. Therefore, it is worth exploring how to further improve neural networks for virtual collection.

4.3.2. Data Inference Based on Ensemble Learning

In addition to neural network-related algorithms, another class of regression models including support vector machines and decision and the regression tree are also suitable for virtual collection. However, the generalization ability or robustness of a single learner of this type is often poor, so some studies have combined multiple learners with certain strategies to form integrated models to improve the problem-solving ability of the learners. The commonly used integration strategies are mainly classified as bagging [88], boosting [91], and stacking [97].

The bagging algorithm was one of the earliest integrated learning algorithms widely used due to its simple structure and good performance. The algorithm randomly draws

multiple training subsets from the original training set by self-sampling, then trains individual learners with different training subsets separately and in parallel, and finally integrates the results by a voting method. Many kinds of research and improvements on the bagging algorithm have emerged in recent years, thus effectively improving the performance of regression models. Among them, the well-known random forest algorithm introduces random attribute selection in training based on the decision tree as the base learner to construct bagging integration. Ref. [89] proposed integration of the random forest, XGBoost, and LightGBM models based on the bagging strategy, thus improving the prediction accuracy of DPVS power. Ref. [90] compared seven different prediction methods, including neural network, bagging decision tree, etc. The experimental results show that bagging decision tree has the best effect.

Unlike the bagging algorithm, the training mechanism of boosting is serial. First, a weak learner is trained from the training set with initial weights, and the weights of the training samples are updated according to the learning error of the weak learner. In other words, the training sample points that have a high learning error rate for the previous weak learner become higher weights and more important in the next weak learner. Algorithms of this class include the most classical Adaboost proposed by Yoav Freund [93], XGBoost proposed by Chen et al. [94], and LightGBM proposed by Ke, GL et al. [95]. In Ref. [96], an XGBoost-based transient stability prediction was proposed for discovering the relationship between power system characteristics and transient stability. Ref. [92] combines the new data filtering program with LightGBM and proposes three levels of prediction models according to the different adaptability of data sources.

The core idea of the stacking ensemble learning method is that after the training of several base learners from the initial training data, the prediction results of these learners are used as a new training set to obtain a new learner. Unlike the bagging and boosting integration methods, the stacking method can integrate heterogeneous learners. Therefore, the stacking method has broader room for improvement, and researchers can integrate different well-performing models by the stacking method to achieve complementary advantages between different models. Ref. [98] proposed a PV power prediction model based on the stacking integrated learning method which fully combines the advantages of different models by stacking random forest, XGBoost, and SVR models, and effectively improves the prediction accuracy. In Ref. [99], two deep learning algorithms, an artificial neural network and long short-term memory, are used as the base models, and an improved stacking integration algorithm is used to integrate the two methods. The results show that this proposed method outperforms individual ANN and LSTM.

The analysis above shows that the relevant algorithms of neural networks have been researched in-depth and have a wide range of application scenarios. Ensemble learning can combine different models and give full play to the advantages of different algorithms, rather than simply weighted fusion, to effectively improve the accuracy of virtual collection. Moreover, it has been demonstrated that improvement based on the classical neural network algorithm makes it more applicable to the virtual acquisition of DPVS data and can effectively improve the virtual acquisition performance [5]. In addition, since the real-time virtual collection accuracy is affected by multiple factors such as climate, environment, and irradiation, data derivation models with robustness are an area for potential improvement and worthy of attention. Therefore, traditional neural networks and integrated algorithms, combined with advanced improvement schemes, offer important supports for DPVS virtual collection.

5. Application Scenarios of Virtual Collection Technology

With the scale expansion of DPVS, the DPVS application scenarios are more and more complex and variable. The acquisition of operation and maintenance information often suffers from incomplete data collection, transmission blockage, and high collection and transmission costs. Therefore, to bring more scholars' attention to the practical application value of virtual collection, this paper innovatively summarizes a variety of application

scenarios for virtual collection based on multi-source information, including but not limited to the following:

- DPVS operation data anomaly detection;
- DPVS fault diagnosis;
- DPVS missing data recovery;
- DPVS real-time operation data collection;

Figure 12 summarizes the four application scenarios and the significance of DPVS virtual collection technology.

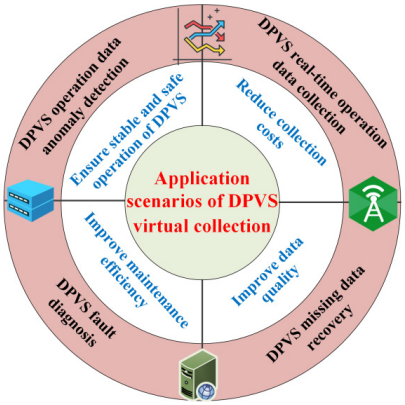

Figure 12. Application scenarios of DPVS virtual collection technology.

With DPVS modules operating in a complex environment for long periods, some failure phenomena will inevitably occur under the influence of thermal power and other external factors [100]. Failure of any component (PV module, converter, inverter, connection line, etc.) can seriously affect the safety and stability of the entire PV system and may lead to greater risks if not detected and maintained in a timely fashion. Therefore, it is essential that the detection of abnormal output status of DPVS is quick and timely, which can be achieved through online abnormality detection methods to ensure the reliable and stable operation of DPVS. The first scenario of the virtual collection technology is the anomaly detection of the real-time operation status of DPVS through the comparison between the virtual collection data and the actual data. Furthermore, Ref. [101] pointed out that anomaly monitoring can be effectively achieved by calculating the spatial performance ratio, similarity coefficient, and characteristic distance using the correlation between different DPVS in the region. Notably, the method is unsupervised and does not require anomaly samples for learning so it can be adapted to an intelligent operation and maintenance platform to monitor the operation status of DPVS in real time.

Artificial intelligence (AI) methods, which do not require additional sensors or infrared testers and do not rely on mathematical models, are widely used in the field of DPVS fault diagnosis [102]. Existing AI methods generally use classification models to obtain the mapping relationship between DPVS array measurement information and fault types. This process requires a large number of features for training, a large data acquisition workload, and there is difficulty in guaranteeing fault diagnosis accuracy. Therefore, it is crucial to effectively acquire feature data that can characterize the fault conditions of DPVS. When different components are damaged, the timing waveforms of PV arrays show different characteristics, while the virtual collection data show different deviation trends from the actual DPVS operation data. In addition, with continuous research on signal feature extraction, scholars have been able to extract a variety of effective features from the curves [103]. Therefore, key features can be extracted from the deviation curves as input for fault diagnosis under different fault situations, thus improving the accuracy of DPVS fault diagnosis.

The complete operational data of DPVS is crucial for calculating various output performance indicators and reliability analysis of PV systems [104]. However, it is common

to lose operational data due to power outages, partial module failures, communication failures, and maintenance at actual field measurement terminals [105]. Missing data makes it difficult to analyze and predict the operational data of DPVS, and most machine learning-based data mining algorithms do not support incomplete data as input. In short, high-precision data filling is important to improve the quality of PV data and reveal the unknown characteristics of DPVS operation data. Many current studies rely on interpolation or prediction models to fill in the missing data [105,106] that require a large amount of historical data for training. Such methods do not fully utilize the spatio-temporal correlation characteristics between different PV plants. Therefore, the second application scenario of virtual acquisition techniques is to provide high-quality infill data for DPVS with missing data. Moreover, according to virtual collection accuracy achieved in Ref. [5], it is known that virtual collection-based missing data filling results in values closer to the actual values than linear interpolation, regression models, and other filling methods.

The first three application scenarios can help improve the quality of collected data and ensure stable operation but cannot reduce the cost of data collection. The O&M system needs to monitor many data points, and an acquisition scheme that only relies on increasing the number of sensors and sampling frequency leads to the cost of data acquisition, transmission, and storage becoming too high. Therefore, the third application scenario is to replace the data collection equipment with virtual collection and realize real-time data collection of the whole system through high-precision data inference models based on similar power output characteristics between power stations in the region. This application scenario only requires the long-term deployment of n collection devices (n is less than the total number of power stations), thus effectively reducing the cost of DPVS data collection.

To sum up, virtual collection technology has diversified application scenarios in the PV field. Among them, the first three scenarios do not have strict requirements on accuracy and have good implementation conditions and theoretical support for long-term coordination with the acquisition equipment. However, the fourth scenario requires real-time estimation of the operational data of the DPVS to be collected, and the complex operational environment imposes higher requirements on the robustness and accuracy of the model. Therefore, it requires researchers to accumulate sufficient historical data while fully considering the uncertainty factors in the operation of DPVS, which is relatively more difficult to implement.

6. Conclusions

Virtual collection is a new, cost-effective and computationally efficient approach for DPVS data collection. This paper provides a comprehensive review of DPVS virtual collection challenges, methodologies, and applications, and the following conclusions can be drawn:

1. The virtual collection process can be subdivided into three steps: similarity analysis, RPS selection, and PV data inference. Considering that there is little research on virtual collection at present, this paper reveals the types of problems similar to virtual acquisition in various fields and analyzes them by analogy so that readers can easily understand the meaning of virtual collection.
2. The system analysis of this paper shows that the virtual collection technique requires strict prerequisites and complex data preprocessing. One of the most critical steps is selecting the reference power station, which determines the quality of the inference model input.
3. This paper summarizes the methods that can be applied to DPVS virtual collection in various fields. As can be seen, DPVS virtual collection is a novel and comprehensive application of artificial intelligence in the PV industry, involving various machine learning methods, including clustering, optimization, regression, etc.
4. This paper proposes a diversified application scenario of virtual collection in the field of PV, hoping to contribute to the needs of distributed energy data management in

the context of carbon peaking and carbon neutrality. Our subsequent research will be directed at further explaining virtual collection from the perspective of physical models and providing more examples of virtual collection.

Author Contributions: Conceptualization, L.G. and T.D.; methodology, L.G. and T.D.; formal analysis, L.G. and T.D.; investigation, L.G., M.U.R., C.L. and T.D.; resources, L.G.; writing—original draft preparation, L.G., J.Y. and T.D.; writing—review and editing, L.G., M.U.R., J.Y., Y.L., and T.D. All authors have read and agreed to the published version of the manuscript.

Funding: This research was funded by the National Natural Science Foundation of China, grant number 52277118, the State Key Laboratory of Power System and Generation Equipment, grant number SKLD21KM10, and the 2022 independent Innovation Fund of Tianjin University-Research and Application of Sensing Traceability and Governance Decision-making Technology for Power Quality Issues in Action Distribution Network, grant number 2022XJS-0066.

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

Abbreviations

The following abbreviations are used in this manuscript:

DPVS	Distributed Photovoltaic Systems
IEA	International Energy Agency
RPS	Reference Power Station
DAP	Data Aggregation Point
DTW	Dynamic Time Warping
BDE	Binary Differential Evolution
ACO	Ant Colony Optimization
PSO	Particle Swarm Optimization
WOA	Whale Optimization Algorithm
HHO	Harris Hawk Optimization
GWO	Gray Wolf Optimization
LSTM	Long Short-Term Memory
RF	Random Forests
LightGBM	Light Gradient Boosting Machine
RBF	Radial Basis Function
AI	Artificial Intelligence
WNN	Wavelet Neural Networks

Nomenclature

$d(X,Y)$	Distance between two samples
S	Covariance matrix
$\mathrm{Cov}(X,Y)$	Covariance coefficient
σ	Standard deviation
$R^2(X,Y)$	Distance correlation coefficient
$v^2(X,Y)$	Distance covariance coefficient
$CS(X,Y)$	Cosine similarity
T_s	Transfer function
$X^d(t)$	Position of the d-th variable updated according to the original equation
$\widetilde{X}^d(t)$	Position of the d-th binary variable updated by the transfer function
h	Output of the hidden layer
η	Step size of gradient descent
y	Virtual collection data, kW
$\hat{y}$	Actual data, kW
W_1	The weight between the input and hidden layer
W_2	The weight between the hidden and output layer
b	Bias of the neural network

References

1. Masson, G.; Bosch, E.; Kaizuka, I.; Jäger-Waldau, A.; Donoso, J. *Snapshot of Global PV Markets 2022 Task 1 Strategic PV Analysis and Outreach PVPS*; IEA PVPS: Paris, France, 2022.
2. Allouhi, A.; Rehman, S.; Buker, M.S.; Said, Z. Up-to-date literature review on Solar PV systems: Technology progress market status and R&D. *J. Clean. Prod.* **2022**, *362*, 132339.
3. National Energy Administration (NEA) China. Available online: http://www.nea.gov.cn/2022-01/20/c_1310432517.htm (accessed on 25 September 2022).
4. Zhu, C.; Long, X.H.; Han, G.J.; Jiang, J.F.; Zhang, S. A virtual grid-based real-time data collection algorithm for industrial wireless sensor networks. *Eurasip J. Wirel. Commun. Netw.* **2018**, *2018*, 134. [CrossRef]
5. Ge, L.; Liu, H.; Yan, J.; Li, Y.; Zhang, J. A Virtual Data Collection Model of DPVs considering Spatio-Temporal Coupling and Affine Optimization Reference. *IEEE Trans. Power Syst.* **2022**, 1–12. [CrossRef]
6. Sobri, S.; Koohi-Kamali, S.; Abd Rahim, N. Solar photovoltaic generation forecasting methods: A review. *Energy Convers. Manag.* **2018**, *156*, 459–497. [CrossRef]
7. Lin, S.M.; Li, P.Q.; Xue, W.Q.; Tang, X.X.; Wang, J.F. Recognition and Reconstruction of Photovoltaic Output Abnormal Data Based on Geographic Correlation. In Proceedings of the 2021 3rd Asia Energy and Electrical Engineering Symposium, Chengdu, China, 26–29 March 2021; pp. 942–948.
8. Zhang, J.; Zhang, S.; Liang, J.; Tian, B.; Hou, Z.; Liu, B.Z. Photovoltaic Generation Data Cleaning Method Based on Approximately Periodic Time Series. *IOP Conf. Ser. Earth Environ.* **2017**, *63*, 12008. [CrossRef]
9. Zhang, Y.; Beaudin, M.; Taheri, R.; Zareipour, H.; Wood, D. Day-Ahead Power Output Forecasting for Small-Scale Solar Photovoltaic Electricity Generators. *IEEE Trans. Smart Grid* **2015**, *6*, 2253–2262. [CrossRef]
10. Li, X.P.; Wang, Y.D.; Ruiz, R. A Survey on Sparse Learning Models for Feature Selection. *IEEE Trans. Cybern.* **2022**, *52*, 1642–1660. [CrossRef]
11. Sang, B.B.; Chen, H.M.; Yang, L.; Li, T.R.; Xu, W.H. Incremental Feature Selection Using a Conditional Entropy Based on Fuzzy Dominance Neighborhood Rough Sets. *IEEE Trans. Fuzzy Syst.* **2022**, *30*, 1683–1697. [CrossRef]
12. Hichem, H.; Elkamel, M.; Rafik, M.; Mesaaoud, M.T.; Ouahiba, C. A new binary grasshopper optimization algorithm for feature selection problem. *J. King Saud. Univ.-Com.* **2022**, *34*, 316–328. [CrossRef]
13. Liang, J.N.; Yang, S.; Winstanley, A. Invariant optimal feature selection: A distance discriminant and feature ranking based solution. *Recognition* **2008**, *41*, 1429–1439. [CrossRef]
14. Zhang, L.; Mistry, K.; Lim, C.P.; Neoh, S.C. Feature selection using firefly optimization for classification and regression models. *Decis. Support Syst.* **2018**, *106*, 64–85. [CrossRef]
15. Lang, A.; Wang, Y.; Feng, C.; Stai, E.; Hug, G. Data Aggregation Point Placement for Smart Meters in the Smart Grid. *IEEE Trans. Smart Grid* **2022**, *13*, 541–554. [CrossRef]
16. Wang, G.D.; Zhao, Y.X.; Huang, J.; Winter, R.M. On the Data Aggregation Point Placement in Smart Meter Networks. In Proceedings of the 2017 26th International Conference on Computer Communication and Networks (Icccn 2017), Vancouver, BC, Canada, 31 July–3 August 2017.
17. Schouten, T.E.; van den Broek, E.L. Fast Exact Euclidean Distance (FEED): A New Class of Adaptable Distance Transforms. *IEEE Trans. Pattern Anal. Mach. Intell.* **2014**, *36*, 2159–2172. [CrossRef]
18. Jiang, W.; Wang, M.J.; Deng, X.Y.; Gou, L.F. Fault diagnosis based on TOPSIS method with Manhattan distance. *Adv. Mech. Eng.* **2019**, *11*, 168781401983327. [CrossRef]
19. Yang, X.Y.; Xu, M.L.; Xu, S.C.; Han, X.J. Day-ahead forecasting of photovoltaic output power with similar cloud space fusion based on incomplete historical data mining. *Appl. Energy* **2017**, *206*, 683–696. [CrossRef]
20. Aik, L.E.; Choon, T.W. An Incremental Clustering Algorithm Based on Mahalanobis Distance. *Aip. Conf. Proc.* **2014**, *1635*, 788–793.
21. Kim, S.; Ham, B.; Kim, B.; Sohn, K. Mahalanobis Distance Cross-Correlation for Illumination-Invariant Stereo Matching. *IEEE Trans. Circuits Syst. Video Technol.* **2014**, *24*, 1844–1859.
22. Zhang, H.W.; Dong, Y.B.; Li, J.; Xu, D.Q. Dynamic Time Warping Under Product Quantization, With Applications to Time-Series Data Similarity Search. *IEEE Internet Things J.* **2021**, *9*, 11814–11826. [CrossRef]
23. Guo, J.; Li, H.; Wang, L.J.; Wang, Z.; Lin, Y.; Huang, D.S. The Model of Photovoltaic Power Short-Term Prediction Based on Dynamic Time Warping Algorithm of Partial Least Squares. In Proceedings of the 3rd IEEE International Electrical and Energy Conference (CIEEC), Beijing, China, 7–9 September 2019; pp. 606–611.
24. Hu, Y.H.; Jia, G.S.; Ai, J.L.; Zhang, Y.; Hou, M.T.; Li, Y.P. Urban heat island estimation from improved selection of urban and rural stations by DTW algorithm. *Appl. Clim.* **2021**, *146*, 443–455. [CrossRef]
25. Zhou, Y.; Zhou, N.R.; Gong, L.H.; Jiang, M.L. Prediction of photovoltaic power output based on similar day analysis, genetic algorithm and extreme learning machine. *Energy* **2020**, *204*, 117894. [CrossRef]
26. Wang, Y.S.; Liao, W.L.; Chang, Y.Q. Gated Recurrent Unit Network-Based Short-Term Photovoltaic Forecasting. *Energies* **2018**, *11*, 2163. [CrossRef]
27. Szekely, G.J.; Rizzo, M.L. Brownian Distance Covariance. *Ann. Appl. Stat.* **2009**, *3*, 1236–1265. [CrossRef] [PubMed]
28. Lu, P.; Ye, L.; Tang, Y.; Zhao, Y.N.; Zhong, W.Z.; Qu, Y.; Zhai, B.X. Ultra-short-term combined prediction approach based on kernel function switch mechanism. *Renew. Energy* **2021**, *164*, 842–866. [CrossRef]

29. Yu, H.Y.; Khan, F.; Garaniya, V. An Alternative Formulation of PCA for Process Monitoring Using Distance Correlation. *Ind. Eng. Chem. Res.* **2016**, *55*, 656–669. [CrossRef]
30. Lin, R.H.; Ye, Z.Z.; Wu, B.D. The application of hydrogen and photovoltaic for reactive power optimization. *Int. J. Hydrog. Energy* **2020**, *45*, 10280–10291. [CrossRef]
31. Su, B.Y.; Chen, H.Y.; Zhou, Z. BAF-Detector: An Efficient CNN-Based Detector for Photovoltaic Cell Defect Detection. *IEEE T Ind. Electron.* **2022**, *69*, 3161–3171. [CrossRef]
32. Lin, P.J.; Peng, Z.N.; Lai, Y.F.; Cheng, S.Y.; Chen, Z.C.; Wu, L.J. Short-term power prediction for photovoltaic power plants using a hybrid improved Kmeans-GRA-Elman model based on multivariate meteorological factors and historical power datasets. *Energy Convers. Manag.* **2018**, *177*, 704–717. [CrossRef]
33. Chen, B.W.; Lin, P.J.; Lai, Y.F.; Cheng, S.Y.; Chen, Z.C.; Wu, L.J. Very-Short-Term Power Prediction for PV Power Plants Using a Simple and Effective RCC-LSTM Model Based on Short Term Multivariate Historical Datasets. *Electronics* **2020**, *9*, 289. [CrossRef]
34. Han, S.; Qiao, Y.H.; Yan, J.; Liu, Y.Q.; Li, L.; Wang, Z. Mid-to-long term wind and photovoltaic power generation prediction based on copula function and long short term memory network. *Appl. Energ.* **2019**, *239*, 181–191. [CrossRef]
35. Pinel, D. Clustering methods assessment for investment in zero emission neighborhoods' energy system. *Int. J. Elec. Power* **2020**, *121*, 106088. [CrossRef]
36. Jain, A.K. Data clustering: 50 years beyond K-means. *Pattern Recognit. Lett.* **2010**, *31*, 651–666. [CrossRef]
37. Park, H.S.; Jun, C.H. A simple and fast algorithm for K-medoids clustering. *Expert. Syst. Appl.* **2009**, *36*, 3336–3341. [CrossRef]
38. Pal, N.R.; Pal, K.; Keller, J.M.; Bezdek, J.C. A possibilistic fuzzy c-means clustering algorithm. *IEEE Trans. Fuzzy Syst.* **2005**, *13*, 517–530. [CrossRef]
39. Tavasoli, M.; Yaghmaee, M.H.; Mohajerzadeh, A.H. Optimal Placement of Data Aggregators in Smart Grid on Hybrid Wireless and Wired Communication. In Proceedings of the 4th IEEE International Conference on Smart Energy Grid Engineering (SEGE), Oshawa, ON, Canada, 21–24 August 2016; pp. 332–336.
40. Rida, M.; Makhoul, A.; Harb, H.; Laiymani, D.; Barharrigi, M. EK-means: A new clustering approach for datasets classification in sensor networks. *Ad. Hoc. Netw.* **2019**, *84*, 158–169. [CrossRef]
41. Wu, J.J.; Xiong, H.; Chen, J. Towards understanding hierarchical clustering: A data distribution perspective. *Neurocomputing* **2009**, *72*, 2319–2330. [CrossRef]
42. Qiu, X.; Lin, Y.; Shao, S.; Guo, S.; Yu, J. Sensor Aggregation Distribution Construction Algorithm for Smart Grid Data Collection System. *J. Electron. Inf. Technol.* **2015**, *37*, 2411–2417.
43. Lee, J.S.; Jiang, H.T. An Extended Hierarchical Clustering Approach to Energy-Harvesting Mobile Wireless Sensor Networks. *IEEE Internet Things J.* **2021**, *8*, 7105–7114. [CrossRef]
44. Cook, E.; Saleem, M.B.; Weng, Y.; Abate, S.; Kelly-Pitou, K.; Grainger, B. Density-based clustering algorithm for associating transformers with smart meters via GPS-AMI data. *Int. J. Electr. Power* **2022**, *142*, 108291. [CrossRef]
45. Corduas, M.; Piccolo, D. Time series clustering and classification by the autoregressive metric. *Comput. Stat. Data* **2008**, *52*, 1860–1872. [CrossRef]
46. Chandrakala, S.; Sekhar, C.C.I. A Density based Method for Multivariate Time Series Clustering in Kernel Feature Space. In Proceedings of the 2008 IEEE International Joint Conference on Neural Networks (IEEE World Congress on Computational Intelligence), Hong Kong, China, 1–8 June 2008; pp. 1885–1890.
47. Sander, J.; Ester, M.; Kriegel, H.P.; Xu, X.W. Density-based clustering in spatial databases: The algorithm GDBSCAN and its applications. *Data Min. Knowl. Discov.* **1998**, *2*, 169–194. [CrossRef]
48. Kim, J.H.; Choi, J.H.; Yoo, K.H.; Loh, W.K.; Nasridinov, A. A Fast Algorithm for Identifying Density-Based Clustering Structures Using a Constraint Graph. *Electronics* **2019**, *8*, 1094. [CrossRef]
49. Tareq, M.; Sundararajan, E.A.; Mohd, M.; Sani, N.S. Online Clustering of Evolving Data Streams Using a Density Grid-Based Method. *IEEE Access* **2020**, *8*, 166472–166490. [CrossRef]
50. Bureva, V.; Sotirova, E.; Popov, S.; Mavrov, D.; Traneva, V. Generalized Net of Cluster Analysis Process Using STING: A Statistical Information Grid Approach to Spatial Data Mining. *Lect. Notes Artif. Int.* **2017**, *10333*, 239–248.
51. Sheikholeslami, G.; Chatterjee, S.; Zhang, A.D. WaveCluster: A wavelet-based clustering approach for spatial data in very large databases. *Vldb J.* **2000**, *8*, 289–304. [CrossRef]
52. Kirkpatrick, S.; Gelatt, C.D., Jr.; Vecchi, M.P. Optimization by simulated annealing. *Science* **1983**, *220*, 671–680. [CrossRef]
53. Erol, O.K.; Eksin, I. A new optimization method: Big Bang Big Crunch. *Adv. Eng. Softw.* **2006**, *37*, 106–111. [CrossRef]
54. Rashedi, E.; Nezamabadi-Pour, H.; Saryazdi, S. GSA: A Gravitational Search Algorithm. *Inf. Sci.* **2009**, *179*, 2232–2248. [CrossRef]
55. Shareef, H.; Ibrahim, A.A.; Mutlag, A.H. Lightning search algorithm. *Appl. Soft Comput.* **2015**, *36*, 315–333. [CrossRef]
56. Biyanto, T.R.; Matradji; Febrianto, H.Y.; Afdanny, N.; Rahman, A.H.; Gunawan, K.S. Rain Water Algorithm: Newton's Law of Rain Water Movements During Free Fall and Uniformly Accelerated Motion Utilization. *Aip. Conf. Proc.* **2019**, *2088*, 020053.
57. Yazdani, S.; Shanbehzadeh, J.; Aminian, E. Feature subset selection using constrained binary/integer biogeography-based optimization. *Isa Trans.* **2013**, *52*, 383–390. [CrossRef]
58. Bostani, H.; Sheikhan, M. Hybrid of binary gravitational search algorithm and mutual information for feature selection in intrusion detection systems. *Soft. Comput.* **2017**, *21*, 2307–2324. [CrossRef]
59. Katoch, S.; Chauhan, S.S.; Kumar, V. A review on genetic algorithm: Past, present, and future. *Multimed. Tools Appl.* **2021**, *80*, 8091–8126. [CrossRef]

60. Barukcic, M.; Hederic, Z.; Spoljaric, Z. The estimation of I-V curves of PV panel using manufacturers' I-V curves and evolutionary strategy. *Energy Convers. Manag.* **2014**, *88*, 447–458. [CrossRef]
61. De La Iglesia, B. Evolutionary computation for feature selection in classification problems. *Wiley Interdiscip. Rev.-Data Min. Knowl. Discov.* **2013**, *3*, 381–407. [CrossRef]
62. Li, T.; Dong, H.B. Unsupervised Feature Selection and Clustering Optimization Based on Improved Differential Evolution. *IEEE Access* **2019**, *7*, 140438–140450. [CrossRef]
63. Yu, Y.; Wang, Y.L. Feature Selection for Multi-label Learning Using Mutual Information and GA. *Lect. Notes Artif. Int.* **2014**, *8818*, 454–463.
64. Varghese, N.V.; Singh, A.; Suresh, A.; Rahnamayan, S. Binary Hybrid Differential Evolution Algorithm for Multi-label Feature Selection. In Proceedings of the IEEE International Conference on Systems, Man, and Cybernetics (SMC), Electr Network, Toronto, ON, Canada, 11–14 October 2020; pp. 4386–4391.
65. Dorigo, M.; Maniezzo, V.; Colorni, A. Ant system: Optimization by a colony of cooperating agents. IEEE transactions on systems, man, and cybernetics. *Part B Cybern. A Publ. IEEE Syst. Man Cybern. Soc.* **1996**, *26*, 29–41. [CrossRef]
66. Eberhart, R.C.; Shi, Y.H. Particle swarm optimization: Developments, applications and resources. *IEEE C Evol. Comput.* **2001**, *1*, 81–86.
67. Yang, X.S. A New Metaheuristic Bat-Inspired Algorithm. *Stud. Comput. Intell.* **2010**, *284*, 65–74.
68. Mirjalili, S.; Lewis, A. The Whale Optimization Algorithm. *Adv. Eng. Softw.* **2016**, *95*, 51–67. [CrossRef]
69. Heidari, A.A.; Mirjalili, S.; Faris, H.; Aljarah, I.; Mafarja, M.; Chen, H.L. Harris hawks optimization: Algorithm and applications. *Future Gener. Comp. Syst.* **2019**, *97*, 849–872. [CrossRef]
70. Khaire, U.M.; Dhanalakshmi, R. Stability Investigation of Improved Whale Optimization Algorithm in the Process of Feature Selection. *Iete Tech. Rev.* **2022**, *39*, 286–300. [CrossRef]
71. Al-Tashi, Q.; Kadir, S.J.A.; Rais, H.M.; Mirjalili, S.; Alhussian, H. Binary Optimization Using Hybrid Grey Wolf Optimization for Feature Selection. *IEEE Access* **2019**, *7*, 39496–39508. [CrossRef]
72. Wang, J.H.; Khishe, M.; Kaveh, M.; Mohammadi, H. Binary Chimp Optimization Algorithm (BChOA): A New Binary Meta-heuristic for Solving Optimization Problems. *Cogn. Comput.* **2021**, *13*, 1297–1316. [CrossRef]
73. Nguyen, B.H.; Xue, B.; Andreae, P.; Zhang, M.J. A New Binary Particle Swarm Optimization Approach: Momentum and Dynamic Balance Between Exploration and Exploitation. *IEEE T Cybern.* **2021**, *51*, 589–603. [CrossRef]
74. Fernandez-Delgado, M.; Sirsat, M.S.; Cernadas, E.; Alawadi, S.; Barro, S.; Febrero-Bande, M. An extensive experimental survey of regression methods. *Neural Netw.* **2019**, *111*, 11–34. [CrossRef]
75. Buchtala, O.; Klimek, M.; Sick, B. Evolutionary optimization of radial basis function classifiers for data mining applications. *IEEE. Trans. Syst. Man. Cy. B* **2005**, *35*, 928–947. [CrossRef]
76. Zhang, Q.; Benveniste, A. Wavelet networks. *IEEE Trans. Neural Netw.* **1992**, *3*, 889–898. [CrossRef]
77. Jiao, L.C.; Pan, J.; Fang, Y.W. Multiwavelet neural network and its approximation properties. *IEEE Trans. Neural Netw.* **2001**, *12*, 1060–1066. [CrossRef]
78. Yang, S.Y.; Wang, M.; Jiao, L.C. A new adaptive ridgelet neural network. *Lect. Notes Comput. Sc.* **2005**, *3496*, 385–390.
79. Vargas-Hakim, G.A.; Mezura-Montes, E.; Acosta-Mesa, H.G. A Review on Convolutional Neural Network Encodings for Neuroevolution. *IEEE Trans. Evol. Comput.* **2022**, *26*, 12–27. [CrossRef]
80. Hinton, G.E.; Osindero, S.; Teh, Y.W. A fast learning algorithm for deep belief nets. *Neural Comput.* **2006**, *18*, 1527–1554. [CrossRef]
81. Yang, Z.; Xu, B.B.; Luo, W.; Chen, F. Autoencoder-based representation learning and its application in intelligent fault diagnosis: A review. *Measurement* **2022**, *189*, 110460. [CrossRef]
82. Massaoudi, M.; Abu-Rub, H.; Refaat, S.S.; Trabelsi, M.; Chihi, I.; Oueslati, F.S. Enhanced Deep Belief Network Based on Ensemble Learning and Tree-Structured of Parzen Estimators: An Optimal Photovoltaic Power Forecasting Method. *IEEE Access* **2021**, *9*, 150330–150344. [CrossRef]
83. Jain, L.C.; Seera, M.; Lim, C.P.; Balasubramaniam, P. A review of online learning in supervised neural networks. *Neural Comput. Appl.* **2014**, *25*, 491–509. [CrossRef]
84. Yu, Y.; Si, X.S.; Hu, C.H.; Zhang, J.X. A Review of Recurrent Neural Networks: LSTM Cells and Network Architectures. *Neural Comput.* **2019**, *31*, 1235–1270. [CrossRef]
85. Chang, G.W.; Lu, H.J. Integrating Gray Data Preprocessor and Deep Belief Network for Day-Ahead PV Power Output Forecast. *IEEE Trans. Sustain. Energy* **2020**, *11*, 185–194. [CrossRef]
86. Jiao, X.; Li, X.S.; Lin, D.Y.; Xiao, W.D. A Graph Neural Network Based Deep Learning Predictor for Spatio-Temporal Group Solar Irradiance Forecasting. *IEEE Trans. Ind. Inform.* **2022**, *18*, 6142–6149. [CrossRef]
87. Ge, L.J.; Li, Y.L.; Yan, J.; Wang, Y.Q.; Zhang, N. Short-term Load Prediction of Integrated Energy System with Wavelet Neural Network Model Based on Improved Particle Swarm Optimization and Chaos Optimization Algorithm. *J. Mod. Power Syst. Clean Energy* **2021**, *9*, 1490–1499. [CrossRef]
88. Strobl, C.; Malley, J.; Tutz, G. An Introduction to Recursive Partitioning: Rationale, Application, and Characteristics of Classification and Regression Trees, Bagging, and Random Forests. *Psychol. Methods* **2009**, *14*, 323–348. [CrossRef]
89. Choi, S.; Hur, J. An Ensemble Learner-Based Bagging Model Using Past Output Data for Photovoltaic Forecasting. *Energies* **2020**, *13*, 1438. [CrossRef]

90. Jahan, I.S.; Blazek, V.; Misak, S.; Snasel, V.; Prokop, L. Forecasting of Power Quality Parameters Based on Meteorological Data in Small-Scale Household Off-Grid Systems. *Energies* **2022**, *15*, 5251. [CrossRef]
91. Friedman, J.H. Greedy function approximation: A gradient boosting machine. *Ann. Stat.* **2001**, *29*, 1189–1232. [CrossRef]
92. Ascencio-Vasquez, J.; Bevc, J.; Reba, K.; Brecl, K.; Jankovec, M.; Topic, M. Advanced PV Performance Modelling Based on Different Levels of Irradiance Data Accuracy. *Energies* **2020**, *13*, 2166. [CrossRef]
93. Ratsch, G.; Onoda, T.; Muller, K.R. Soft margins for AdaBoost. *Mach. Learn.* **2001**, *42*, 287–320. [CrossRef]
94. Chen, T.Q.; Guestrin, C. XGBoost: A Scalable Tree Boosting System. In Proceedings of the 22nd Acm Sigkdd International Conference on Knowledge Discovery and Data Mining, San Francisco, CA, USA, 13–17 August 2016; pp. 785–794.
95. Ke, G.L.; Meng, Q.; Finley, T.; Wang, T.F.; Chen, W.; Ma, W.D.; Ye, Q.W.; Liu, T.Y. LightGBM: A Highly Efficient Gradient Boosting Decision Tree. *Adv. Neural Inf. Process. Syst.* **2017**, *30*, 3149–3157.
96. Chen, M.H.; Liu, Q.Y.; Chen, S.H.; Liu, Y.C.; Zhang, C.H.; Liu, R.H. XGBoost-Based Algorithm Interpretation and Application on Post-Fault Transient Stability Status Prediction of Power System. *IEEE Access* **2019**, *7*, 13149–13158. [CrossRef]
97. Divina, F.; Gilson, A.; Gomez-Vela, F.; Torres, M.G.; Torres, J.E. Stacking Ensemble Learning for Short-Term Electricity Consumption Forecasting. *Energies* **2018**, *11*, 949. [CrossRef]
98. Guo, X.F.; Gao, Y.; Zheng, D.; Ning, Y.; Zhao, Q.N. Study on short-term photovoltaic power prediction model based on the Stacking ensemble learning. *Energy Rep.* **2020**, *6*, 1424–1431. [CrossRef]
99. Khan, W.; Walker, S.; Zeiler, W. Improved solar photovoltaic energy generation forecast using deep learning-based ensemble stacking approach. *Energy* **2022**, *240*, 122812. [CrossRef]
100. Zhao, Y.Y.; Liu, Q.; Li, D.S.; Kang, D.H.; Lv, Q.; Shang, L. Hierarchical Anomaly Detection and Multimodal Classification in Large-Scale Photovoltaic Systems. *IEEE Trans. Sustain. Energy* **2019**, *10*, 1351–1361. [CrossRef]
101. Shi, Y.C.; He, W.G.; Zhao, J.; Hu, A.Y.; Pan, J.N.; Wang, H.Z.; Zhu, H.L. Expected output calculation based on inverse distance weighting and its application in anomaly detection of distributed photovoltaic power stations. *J. Clean Prod.* **2020**, *253*, 119965. [CrossRef]
102. Mansouri, M.; Trabelsi, M.; Nounou, H.; Nounou, M. Deep Learning-Based Fault Diagnosis of Photovoltaic Systems: A Comprehensive Review and Enhancement Prospects. *IEEE Access* **2021**, *9*, 126286–126306. [CrossRef]
103. Zhu, H.L.; Shi, Y.C.; Wang, H.Z.; Lu, L.X. New Feature Extraction Method for Photovoltaic Array Output Time Series and Its Application in Fault Diagnosis. *IEEE J. Photovolt.* **2020**, *10*, 1133–1141. [CrossRef]
104. Koubli, E.; Palmer, D.; Rowley, P.; Gottschalg, R. Inference of missing data in photovoltaic monitoring datasets. *IET Renew. Power Gener.* **2016**, *10*, 434–439. [CrossRef]
105. Livera, A.; Theristis, M.; Koumpli, E.; Theocharides, S.; Makrides, G.; Sutterlueti, J.; Stein, J.S.; Georghiou, G.E. Data processing and quality verification for improved photovoltaic performance and reliability analytics. *Prog. Photovolt.* **2021**, *29*, 143–158. [CrossRef]
106. Lei, Z.; Wang, B.; Wang, K.X.; Pei, Y.; Huang, Z.Y. Photovoltaic power missing data filling based on multiple matching and long- and short-term memory network. *Int. Trans. Electr. Energy Syst.* **2021**, *31*, e128291.1–e128291.20. [CrossRef]