IEEE Access
Multidisciplinary ⋮ Rapid Review ⋮ Open Access Journal

Received 21 June 2023, accepted 2 July 2023, date of publication 5 July 2023, date of current version 12 July 2023.

Digital Object Identifier 10.1109/ACCESS.2023.3292516

RESEARCH ARTICLE

Knowledge Extraction From PV Power Generation With Deep Learning Autoencoder and Clustering-Based Algorithms

SEYED MAHDI MIRAFTABZADEH, (Member, IEEE), MICHELA LONGO, (Member, IEEE), AND MORRIS BRENNA, (Member, IEEE)
Department of Energy, Politecnico di Milano, 20156 Milan, Italy

Corresponding author: Seyed Mahdi Miraftabzadeh (seyedmahdi.miraftabzadeh@polimi.it)

This work was supported in part by the Il Centro Nazionale per Mobilità Sostenibile (MOST)—Sustainable Mobility Center, and in part by the European Union Next-Generation EU (Piano Nazionale di Ripresa e Resilienza (PNRR)—-Missione 4 Componente 2, Investimento 1.4—D.D. 1033 17/06/2022) under Grant CN00000023.

ABSTRACT The unpredictable nature of photovoltaic solar power generation, caused by changing weather conditions, creates challenges for grid operators as they work to balance supply and demand. As solar power continues to become a larger part of the energy mix, managing this intermittency will be increasingly important. This paper focuses on identifying daily photovoltaic power production patterns to gain new knowledge of the generation patterns throughout the year based on unsupervised learning algorithms. The proposed data-driven model aims to extract typical daily photovoltaic power generation patterns by transforming the high dimensional temporal features of the daily PV power output into a lower latent feature space, which is learned by a deep learning autoencoder. Subsequently, the Partitioning Around Medoids (PAM) clustering algorithm is employed to identify the six distinct dominant patterns. Finally, a new algorithm is proposed to reconstruct these patterns in their original subspace. The proposed model is applied to two distinct datasets for further analysis. The results indicate that four out of the identified patterns in both datasets exhibit high correlation (over 95%) and temporal trends. These patterns correspond to distinct weather conditions, such as entirely sunny, mostly sunny, cloudy, and negligible power generation days, which were observed approximately 61% of the analyzed period. These typical patterns can be expected to be observed in other locations as well. Identified PV power generation patterns can improve forecasting models, optimize energy management systems, and aid in implementing energy storage or demand response programs and scheduling efficiently.

INDEX TERMS Clustering algorithm, data mining, deep learning autoencoder, pattern extraction and analysis, PV power generation, unsupervised learning.

I. INTRODUCTION

Over the past decade, there has been a significant rise in the production of electrical energy generated from Renewable Energy Sources (RES) to meet the ever-growing demand for electricity while minimizing the environmental impact of production [1], [2]. This trend is driven by the world's increasing awareness of the adverse effects of fossil fuel usage on the environment, making the need for sustainable energy alternatives more pressing. The intermittent and fluctuating nature of Photovoltaic (PV) power production poses significant challenges to the stability and economic operation of the existing energy management system. These challenges include maintaining system stability and ensuring its economic viability [2], [3], [4].

In recent years, various techniques, including physical, statistical, and machine learning models, have been employed to analyze PV datasets and forecast solar power generation for different time scales [1], [5]. To the best of the authors' knowledge, no study has yet analyzed and extracted the

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

This work is licensed under a Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 License.
For more information, see https://creativecommons.org/licenses/by-nc-nd/4.0/

patterns of PV power production. Understanding the patterns of PV power production is crucial as it can provide valuable insights into the intermittent and fluctuating nature of solar power generation. Identifying and analyzing the patterns of PV power production can lead to improving the accuracy of solar energy forecasts, developing more robust energy management systems, and integrating solar energy into the existing energy mix, thereby contributing to a more sustainable and efficient energy system [3], [6], [7].

Given the large volume of daily PV power production data, visually identifying similarities is impractical. Thus, an intelligent algorithm is required to extract dominant patterns from these vast datasets. An intelligent automated approach not only efficiently identifies and analyzes significant patterns in PV power production but also provides a more comprehensive insight into the underlying patterns and trends.

Without prior knowledge about potential patterns, unsupervised machine learning algorithms can be employed to discover them. These algorithms can autonomously identify hidden structures and patterns within the data, providing valuable insights and discoveries. By leveraging unsupervised clustering learning, we can uncover novel and unexpected patterns that may have otherwise gone unnoticed, enhancing our understanding of daily PV power production.

However, the high dimensionality of daily PV output poses a challenge for clustering algorithms, as it can decrease their efficiency and performance. Dimensionality reduction techniques such as unsupervised autoencoders can be applied to overcome this issue. By reducing the dimensionality of the input data, autoencoders help improve the efficiency and effectiveness of the learning algorithm. Consequently, more accurate and meaningful clustering results can be achieved, leading to enhanced insights and better decision-making in the analysis of PV power production data.

This paper presents a hybrid data-driven model based on deep learning and Partitioning Around Medoids (PAM) clustering algorithms to identify daily PV power generation patterns without prior knowledge. This article seeks to enhance our knowledge of PV power generation and its effective utilization, with potential applications in energy management systems, PV power predictions, and anomaly detection [8]. The unsupervised model presented in this article can help identify hidden patterns and relationships within PV generation data, enabling better decision-making for system optimization and efficiency improvement. The main contributions of this paper are as follows:

1) Identification of typical PV power generation patterns using deep learning autoencoder and PAM clustering algorithms.
2) Proposing a new algorithm for reconstructing the extracted patterns from latent (reduced) dimensions to the original sub-space.
3) In-depth analysis of the identified PV patterns through diverse statistical and visualization methods to acquire novel knowledge and insights into the production of PV power.

The rest of this paper is organized as follows: Section II is dedicated to the related works in the literature. Section III details the proposed framework that utilizes unsupervised machine learning techniques such as autoencoder for dimensionality reduction and PAM for clustering analysis. Section IV introduces the case studies and used PV datasets. Section V showcases the outcomes achieved through the implementation of the proposed methodology. Section VI discusses and deliberates on the findings, including additional analysis. Finally, Section VII concludes the paper.

II. RELATED WORK

Clustering algorithms have become a popular technique for knowledge extraction from time series datasets in the literature [9], [10]. Time series datasets are commonly found in many fields, such as finance, engineering, and ecology, and can be challenging to analyze due to their dynamic nature and high dimensionality. Clustering algorithms effectively group similar time series data together, allowing recognize patterns and trends that might not be immediately apparent. For example, in anomaly detection, clusters of time series data that deviate from the norm can indicate potential issues or anomalies [11]. Moreover, these algorithms can help researchers to gain a deeper understanding of complex time series data by identifying underlying structures and relationships between different variables and providing valuable insights into the factors that influence their behavior.

K-means is a famous and widely used clustering algorithm for its simplicity, versatility, efficiency, and interpretability [12]. Once the time-series data has been clustered, one can utilize various metrics to link particular patterns to each cluster, such as the mean or median of all samples in each group or the sample that is closer to the center of the clusters [10], [13]. Reference [14] used the k-means algorithm to cluster the residential heating consumption and presented the centroid of each cluster as representative patterns. The average of samples in each K-means cluster has been used to identify various patterns for characterizing electricity load profiles [15]. Electrical load patterns were identified and categorized using centroid clustering models like ant colony algorithms and K-means [16]. The study in [17] utilizes the Gaussian mixture clustering method to identify typical daily electricity usage patterns in buildings, selecting the average of each cluster as representative.

High-dimensional datasets such as time series can negatively impact the performance of clustering algorithms [18]. In order to tackle this issue and enhance the efficiency of algorithms, dimensionality reduction techniques such as PCA, Kernel Principal Component Analysis (KPCA), or Deep Learning Autoencoders (DAEs) are utilized before clustering. Utilizing techniques such as PCA and Sammon Map for dimensionality reduction prior to clustering, [19] determined load patterns by examining the center of each cluster, which is represented by the average of samples associated with it. Singular Value Decomposition (SVD) and K-means

were utilized by [20] to cluster solar energy production. Also, [21] proposed a hybrid model based on principal component analysis and K-means for extracting residential electricity consumption patterns by averaging the samples in clusters using smart meters data. Inspired by the K-means algorithm, [22] proposed a new method to cluster multivariate time series via Common Principal Component Analysis (CPCA). The typical patterns of voltage variations at the sub-10 min time scale are identified by a hybrid model based on KPCA and K-means [23], [24]. A hybrid model based on convolutional Deep learning autoencoder with K-means algorithms was proposed in [25] to drive gene expression of high dimensional time series datasets. Reference [26] used deep convolutional autoencoders and K-means to cluster seismic signals. Reference [27] introduces anomaly detection techniques that rely on deep autoencoder and K-means methods; the method employs cosine similarity to identify normal and anomalous time series signals. RNA sequence data is clustered and analyzed with deep learning autoencoder as dimensionality reduction for improving the clustering efficiency [28], [29]. Reference [30] proposed a hybrid model to learn representations considering deep autoencoder for dimensionality reduction and K-means for clustering.

Clustering algorithms have been extensively employed in the literature for feature extraction to uncover meaningful patterns and groupings in data. These techniques have been utilized in various domains, including bioinformatics, image processing, energy management systems, and natural language processing, to improve our understanding of complex phenomena and facilitate decision-making processes [18], [31], [32], [33], [34], [35], [36]. Reference [35] used K-means clustering to identify the key segments in speech signals and then determined the emotion associated with an input speech signal through a hybrid deep learning model. The benefits of different clustering algorithms for energy system optimization have been investigated in [31] and [37]. Clustering time series data is utilized in [32] to extract new knowledge into energy consumption patterns, which ultimately improved the energy efficiency of the system. Reference [38] utilized Dynamic Time Warping (DTW) and K-means clustering algorithms to identify the four working conditions of photovoltaic array systems based on the time series of voltage and current signals without prior knowledge of the system.

Clustering techniques can improve the accuracy of prediction models and anomaly detection by providing insights into the underlying patterns and relationships within the data, which can inform the selection and optimization of these models. Moreover, clustering can also help to identify the outliers and anomalies in the data, which can be used to develop more effective anomaly detection methods. References [11] and [39] proposed anomaly detection models based on fuzzy C-mean clustering. The accuracy of solar power generation prediction models was enhanced through the clustering of the input time series [40], [41], [42], [43]. To enhance the precision of cloud motion speed calculation, the classification of different types of clouds is performed using K-means clustering [44]. In [45], a hybrid model was proposed for day-ahead electricity price forecasting, where a new feature was created by applying K-means clustering on the daily electricity price data.

Unsupervised machine learning algorithms such as autoencoder and clustering can be utilized to identify and analyze the patterns of PV power production. These algorithms play a crucial role in uncovering hidden structures and relationships within the data, complementing the analysis conducted in previous studies.

Despite the increasing utilization of machine learning techniques in analyzing PV datasets and forecasting solar power generation, there remains a gap in the literature regarding the analysis and extraction of patterns specific to PV power production. Addressing this research gap is therefore crucial for advancing our understanding of solar energy and achieving a more sustainable energy future. The main motivation of this paper is to fill this research gap by analyzing and extracting patterns of PV power production using unsupervised learning algorithms. By understanding these patterns, valuable insights can be gained into the intermittent and fluctuating nature of solar power generation.

III. METHODOLOGY

This study proposes a framework based on the standard Knowledge-Data-Driven methodology (KDD) introduced by Fayyad et al. [46], shown in Fig. 1, to obtain new knowledge of PV power production patterns. The framework described here comprises five distinct steps: preprocessing, data transformation, data mining, post-processing, and knowledge extraction. This section explains each step in details, providing a comprehensive overview of the entire process. With the help of this framework, complex time-series data can be efficiently analyzed and transformed into meaningful insights, providing new valuable knowledge to promote PV systems.

A. PREPROCESSING

Initially, the time-series dataset $X = \left[x^{(1)}, x^{(2)}, \cdots, x^{(mn)}\right]$ is preprocessed by reorganizing it into an *mxn* matrix that serves as input for machine learning algorithms by (1). To mitigate the impact of missing records on X, timestamps of the original dataset are utilized to construct $\hat{\boldsymbol{X}}$, ensuring that each time point has a corresponding input value in $\hat{\boldsymbol{X}}$. This step is crucial in optimizing the performance of the models and facilitating the extraction of valuable insights and patterns from the data.

$$\hat{\boldsymbol{X}} = \begin{bmatrix} x_1^{(1)} & x_2^{(1)} & \cdots & x_n^{(1)} \\ x_1^{(2)} & x_2^{(2)} & \cdots & x_n^{(2)} \\ \vdots & \vdots & \ddots & \vdots \\ x_1^{(m)} & x_2^{(m)} & \cdots & x_n^{(m)} \end{bmatrix} \tag{1}$$

where indices in superscripts (m) indicate the sample points corresponding to each day, and subscript (n) indices indicate

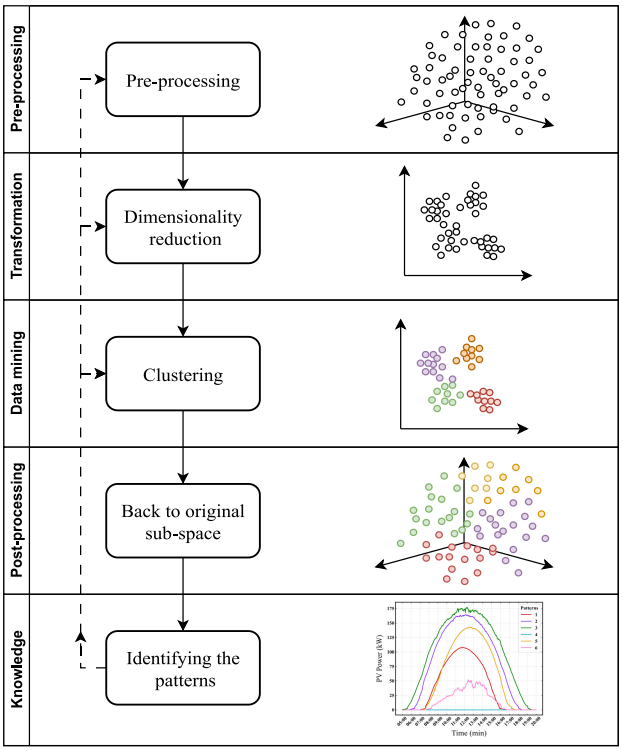


FIGURE 1. Proposed data-driven framework for extracting PV power production patterns.

the features, which in this study, n equals 1440 for daily PV power output (one sample for each min). The days that contain missing values are excluded from the dataset to ensure the integrity of the analysis.

Then, the features in dataset $\hat{\boldsymbol{X}}$ are standardized by removing their mean and scaling to unit variance (2).

$$\hat{\boldsymbol{X}}_s = \frac{\hat{\boldsymbol{X}} - \boldsymbol{\mu}}{\sqrt{\boldsymbol{\sigma}^2 + \varepsilon}} \tag{2}$$

where, $\boldsymbol{\mu}$ is the mean vector, $\boldsymbol{\sigma}$ is the standard deviation vector, and ε is introduced to avoid division by zero. Standardization ensures that all features contribute equally, thus preventing any biases towards a particular feature during the training process.

B. DATA TRANSFORMATION

The data transformation or data reduction and projection in the KDD process helps prepare the data for analysis by converting it into a form that the machine learning algorithm can interpret more efficiently [9], [46], [47]. Various techniques have been proposed and used in the literature to improve clustering algorithm performance and address the "Curse of dimensionality" issue, which refers to the difficulty in analyzing data with a high number of features. Generally, dimensionality reduction methods such as PCA, Independent Component Analysis (ICA), and Non-negative Matrix Factorization (NMF) are used before clustering algorithms [19], [21]. Furthermore, nonlinear dimensionality

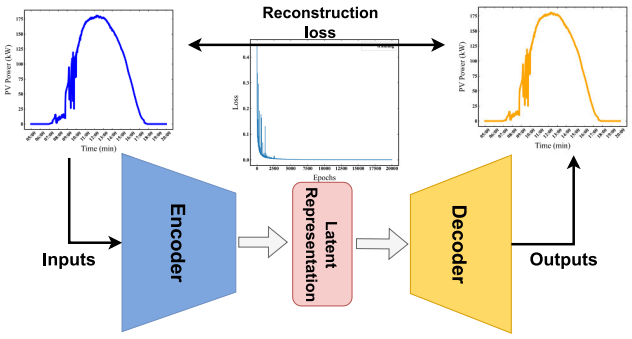


FIGURE 2. Schematic presentation of deep neural network autoencoder.

reduction techniques such as KPCA and DAEs can map high-dimensional input data to a lower-dimensional subspace for clustering [24], [30]. These techniques are particularly useful for capturing highly complex and nonlinear relationships between input variables.

In this study, a deep-learning autoencoder is designed and optimized with Bayesian optimization [48] to learn a latent (compressed) representation of the input data by encoding it into a lower-dimensional space and then decoding it back to the original input space, as shown in Fig. 2.

An autoencoder consists of an encoder $\boldsymbol{z} = f(\boldsymbol{x})$ that maps the input data $\boldsymbol{x} \in \mathbb{R}^n$ to a lower-dimensional subspace $\boldsymbol{z} \in \mathbb{R}^r$ and a decoder $\boldsymbol{h} = g(\boldsymbol{z})$ ($g : \mathbb{R}^r \rightarrow \mathbb{R}^n$) that takes the encoded data and maps it back to the original subspace, $\hat{\boldsymbol{X}} = g(f(\boldsymbol{x}))$ [49], [50]. A trained autoencoder found the best mapping functions $f(\cdot)$ and $g(\cdot)$ that satisfy (3):

$$\arg\min_{f,g} \sum_{i=1}^{m} \mathcal{L}(\boldsymbol{x}^{(i)}, g(f(\boldsymbol{x}^{(i)})) \tag{3}$$

where $\mathcal{L}$ is a loss function that measures the reconstruction error between the original input and reconstructed data that the autoencoder aims to minimize.

C. DATA MINING

Clustering algorithms extract knowledge from time series datasets, enabling new insights and discoveries in various fields [51]. This study uses the PAM as a clustering algorithm in data mining step to cluster the PV power production in order to extract the new knowledge about them. PAM is a variant of the well-known k-medoids algorithm, which uses medoids instead of using means as the center of the cluster. Medoids $\boldsymbol{m}^{(j)}$ are data points that best represent their respective cluster by having the lowest average dissimilarity to all other points in the cluster. The PAM algorithm works as follows [52]:

- Select the K medoids $C = \{C^1, C^2, \ldots, C^k\}$.
- Calculate the dissimilarity (distance) $dist(\cdot)$ for each data point $\boldsymbol{x}^{(i)}$.
- Assign each data point to the nearest medoid C^j.
- Perform pairwise swapping of medoids and non-medoids to minimize the loss function $J(C)$ by (4).

$$J(C) = \sum_{j=1}^{k} \sum_{\boldsymbol{x}^{(i)} \in C^j} dist(\boldsymbol{x}^{(i)}, \boldsymbol{m}^{(j)}) \tag{4}$$

where $\boldsymbol{m}^{(j)}$ represents each cluster medoid. The algorithm continues to iterate to minimize $J(C)$, the sum of dissimilarities between $\boldsymbol{m}^{(j)}$ and all other points in C^j. The PAM algorithm minimizes the loss function through a series of swaps between medoids and non-medoid data points within each cluster. The algorithm evaluates different configurations of medoids and selects the one with the lowest dissimilarity to minimize the loss function, typically measured as the sum of dissimilarities between data points and their assigned medoids. By iteratively updating the medoids through swaps, PAM converges towards a configuration that improves the clustering quality and minimizes the loss function. The process continues until no further improvements in dissimilarity can be achieved [53], [54].

Different distance measures such as Euclidean distance, Manhattan distance, cosine distance, and Minkowski distance can be used as function $dist(\cdot)$. Compared to other clustering algorithms like K-means, PAM offers several advantages, such as its ability to handle non-Euclidean distance metrics, robustness to outliers, and the option to specify the number of clusters beforehand. Despite its efficiency for small and medium-sized datasets, PAM is computationally expensive with $O(k(m-k)^2)$, particularly for larger datasets [53], [54].

D. POSTPROCESSING

The clustering algorithm performs better in a reduced dimensional subspace $\mathbb{R}^r$ rather than the original subspace $\mathbb{R}^n$. Although obtaining labels or clusters in the original subspace is straightforward, the medoids in original sub-space are not directly provided. To retrieve medoids in the original dimension, the decoder $g(\cdot)$ part of DAE is used to return the medoids to the original subspace. However, this method may result in some information loss from the original data during the compression process. This is because autoencoders are designed to learn an approximation of the input data and may introduce some amount of distortion or error in the reconstruction process [50]. Additionally, the encoding process may introduce some negligible noises on data points, particularly on reconstructing data that was not part of the training phase, which could affect the accuracy of the reconstructed data. Similarly, inverse functions in PCA or approximation functions in KPCA are associated with some degree of error due to information loss [19], [24], [50], [55]. Therefore, this study introduces a new approach, algorithm 1, to retrieving medoids from the original subspace after clustering.

The proposed algorithm retrieves the medoids in the original subspace after clustering the data in a lower dimension. Once the clusters are formed, the algorithm proceeds with the following steps:

1) Set the number of clusters (k) and initialize variables.
2) Iterate over each cluster in the original subspace.
3) Calculate pairwise distances, considering the chosen distance metric in clustering step $dist(\cdot)$, between samples within the cluster and store them in a matrix, $\boldsymbol{M}$.
4) Calculate the sum of distances for each sample in the cluster from $\boldsymbol{M}$ and store it in $\boldsymbol{S}$.
5) Find the index (idx) of the data associated with the minimum sum of distances in $\boldsymbol{S}$ using $argmin$.
6) Assign the data at the corresponding index (idx) as the pattern (the medoid) for the current cluster.
7) Repeat for all clusters.
8) Return all the retrieved patterns (medoids) for each cluster ($\boldsymbol{P}$).

The medoid is the data point in the cluster with the smallest distance to all other samples within the same cluster. The proposed algorithm guarantees that the medoids are not subject to any error as they are derived directly from the original subspace.

Algorithm 1 Identifying Medoids in Original Sub-Space

Input: Original dataset and their labels
Output: The medoids (or data points) for each cluster in original sub-space

Let k = number of cluster
Let l_k number of samples in k$^{\text{th}}$ cluster
Let P patterns (medoids) in original sub-space
  For *cluster* = 1 to k **do**
    Let a $l_k \times l_k$ matrix M, where M[i,j] is the distance between sample i and sample j in k$^{\text{th}}$ cluster
    Let *data* = all samples in the *cluster*(k$^{\text{th}}$)
    For $i = 1$ to l_k **do**
      For $j = i$ to l_k **do**
        $distance = dist(data[i, :], data[j, :])$
        M[i,j] = *distance*
        M[j,i] = *distance*
      end for
    $S_i = \sum_i \sum_j$ M[i, j]
    $idx \leftarrow$ index of data associated to $argmin(S_i)$
    $P_k \leftarrow data[idx, :]$
  end for
Return P

The overall time complexity of the algorithm is $O(kl^2n)$, where l is a number of samples in the biggest cluster. However, more efficient algorithms presented in NumPy or SciPy [56], such as the k-d tree algorithm, can reduce the time complexity of calculating the pairwise distance matrix to $O\,(kl(\log l)n)$, making it more suitable for large datasets.

E. KNOWLEDGE EXTRACTION

The final stage is to extract new knowledge and analyze the data mining results. In KDD process, knowledge extraction refers to the process of identifying valuable and actionable information or patterns from large datasets. Knowledge extraction aims to transform raw data or patterns into

interpretable knowledge for decision-making, prediction, and understanding of the underlying data.

In the process of extracting typical patterns of PV power production, different criteria can be used, such as taking the average of all samples or the center of centroids in each cluster. In this study, the sample closest to the cluster center is chosen as a better representative of the cluster than the mean [51], [52], [57]. This is because the mean can be affected by outliers, which can skew the results. On the other hand, the sample closest to the center of centroids is not influenced by outliers and is more stable. Furthermore, the closest sample provides a more accurate representation of the typical pattern for that cluster.

Moreover, in order to gain deeper insights into the extracted patterns and to obtain more knowledge from them, various statistical techniques are utilized. Apart from clustering, the Pearson correlation coefficient and cumulative distribution function are employed to analyze the trends and correlation between the patterns in two different case studies. The Pearson correlation coefficient is a measure of the linear relationship between two variables, and it ranges from -1 (perfect negative correlation) to 1 (perfect positive correlation). By calculating the Pearson correlation coefficient between the patterns, we can identify which ones are highly correlated and which ones are not. The cumulative distribution function is another statistical technique used to analyze the distribution of data. By applying this technique to the patterns, we can observe the behavior of the data over time and gain insights into its characteristics, such as its variability and trends.

TABLE 1. Dataset description.

Database	Rated power [kW]	Average power [kW]	Standard Deviation	Time Span
$\mathcal{D}_1$	243	35.68	57.15	2015÷2018
$\mathcal{D}_2$	75	10.05	16.44	2015÷2017

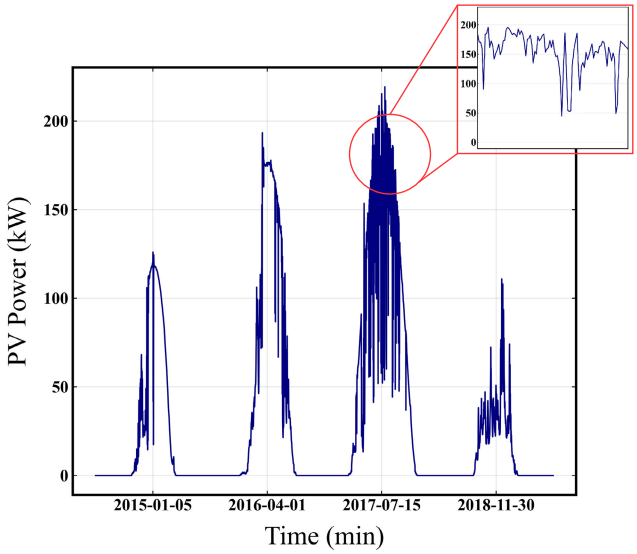


FIGURE 3. PV power outputs for four random days of $\mathcal{D}_1$.

IV. CASE STUDY

As a case study, this paper employs real PV power generation using a public database from the National Institute of Standards and Technology (NIST) campus in Gaithersburg, Maryland [58], [59]. This database contains historical PV power generations of two different PV power farms located near each other (within a proximity of 1.25 km), Table 1. The database has been divided into two datasets: $\mathcal{D}_1$ consists of PV power generation data for a larger farm with a rated power of 243 kW, while $\mathcal{D}_2$ contains PV power generation data for a smaller farm with a rated power of 75 kW, with one sample recorded per minute [60]. $\mathcal{D}_1$ comprises a dataset spanning four years or 1,461 days, consisting of samples of daily PV power generation. Each sample represents data collected at one-minute intervals, resulting in a total of 1,440 samples per day. $\mathcal{D}_2$ comprises 1,095 days of daily PV production data, covering the period from 2015 to 2017.

The PV power generation exhibits diverse output shapes each day, as depicted in Fig. 3. This figure displays four random days from $\mathcal{D}_1$, and reveals that the output patterns vary across different months. Moreover, by zooming in on the third sample, it can be observed that the PV output fluctuates rapidly within each minute. For instance, it can drop from 200 kW to less than 50 kW within a couple of minutes. The identification of dominant patterns from a vast amount of PV datasets is not feasible with the naked eye. Therefore, to overcome this challenge, a data-driven framework presented in Fig. 1 is employed to detect similarities in high-resolution daily PV datasets. Through clustering, the framework enables the extraction of patterns linked with each cluster, making it possible to identify and analyze the underlying patterns.

One possible method for identifying PV power generation patterns is by grouping them according to the seasons. However, as illustrated in Fig. 4, this approach is insufficient in determining the typical pattern that occurs over the course of years. For instance, the patterns observed during two different seasons might have nearly identical shapes. Additionally, this method fails to acknowledge the possibility of encountering the same pattern in various seasons or days of the year. Essentially, this approach does not take into account the likelihood of finding the same pattern across different months of a year. Therefore, there is a need for a more effective method that can capture and analyze the patterns of data more comprehensively.

The dataset was split into 80% for training the autoencoder and 20% for testing its generalization capability. Once the optimal autoencoder is obtained, the entire dataset is utilized to feed into the network for further analysis.

This study employed Python as the programming language and utilized various libraries and packages such as NumPy and Pandas for data pre-processing, Keras and

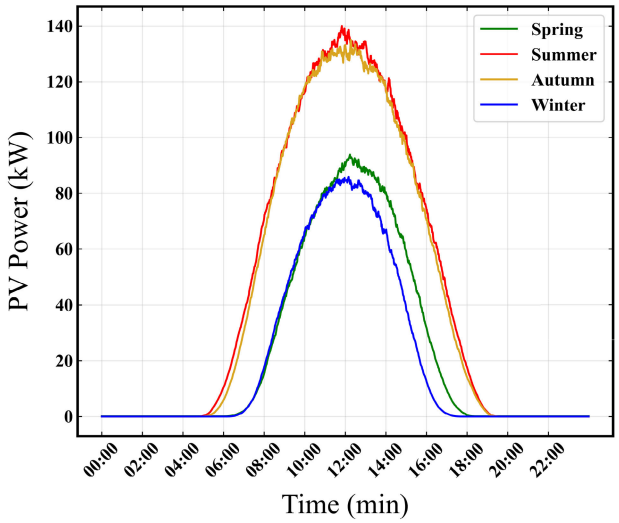


FIGURE 4. The average PV power production in each season for $\mathcal{D}_1$.

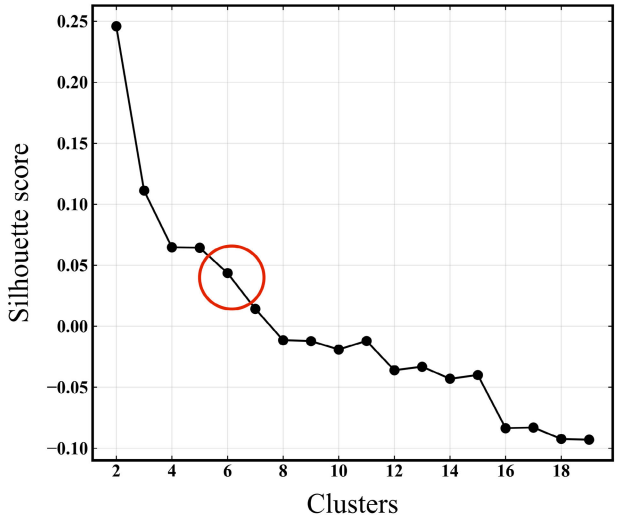


FIGURE 5. The silhouette score for PAM with Manhattan distance metric for a different number of clusters.

TensorFlow for neural network modeling and optimization, and Matplotlib and Seaborn for creating visualizations. The computational resources utilized for this research included an i7-8700K CPU, 16 GB RAM, and Nvidia RTX GeForce 2080 with 16 GB RAM. To expedite the training process, the models were trained on a GPU experimental environment, leveraging parallel computing capabilities and reducing training time.

V. RESULTS

The model presented in this study is applied to both datasets separately. Initially, the data is preprocessed, and then the DAE model is applied to capture the complex, non-linear relationships within the dataset and reduce the dimensionality to enhance the performance of the clustering algorithm. Furthermore, using DAE allows for extracting relevant features from the data in the lower dimension, which aids in pattern recognition and identifying dominant patterns. The outcome of the proposed model can be used in a variety of applications, including load forecasting and energy management systems.

TABLE 2. Deep autoencoder structure.

Encoder		Decoder	
Layers	Parameters	Layers	Parameters
FC 1	Neurons: 1440 Activation: SELU	FC 7	Neurons: 124 Activation: SELU
FC 2	Neurons: 672 Activation: SELU	FC 8	Neurons: 240 Activation: SELU
FC 3	Neurons: 480 Activation: SELU	FC 9	Neurons: 480 Activation: SELU
FC 4	Neurons: 240 Activation: SELU	FC 10	Neurons: 672 Activation: SELU
FC 5	Neurons: 124 Activation: SELU	FC 11	Neurons: 1440 Activation: SELU
FC 6	Neurons: 60 Activation: Sigmoid Regularization: KL divergence		

This study extensively analyzed various neural network architectures to identify the optimal approach for capturing high spikes in the dataset. Experimental investigations revealed that the feed-forward neural network without batch normalization exhibited superior performance in effectively capturing the high oscillations present in the data. Table 2 illustrates the optimized structure of the DAE through Bayesian optimization [48] in this study. The DAE architecture comprises eleven fully connected neural network layers, utilizing the Scaled Exponential Linear Unit (SELU) activation function to ensure efficient learning. Additionally, the sixth layer of the autoencoder serves as a bottleneck, which maps the data to 60 dimensions with the Kullback-Leibler (KL) divergence Regularization function. The deep neural network was trained with the Mean Absolute Error (MSE) loss function and Adam optimizer. The training process was carried out with 20000 epochs with an exponential decay learning rate resulting in an MSE of 6.922×10^{-4}. Although MSE reaches a value of 7.518×10^{-4} after 4000 epochs, the model continues to run longer to achieve even lower MSE values. The total training time was approximately 5 hours and 35 minutes. When dealing with large datasets, incorporating batch and layer normalization techniques can greatly enhance the training process by reducing the number of epochs needed and improving convergence speed.

The Kullback-Leibler (KL) divergence, also known as relative entropy, is a statistical measure that quantifies the difference between two probability distributions. Specifically, it calculates the expected logarithmic difference between the probabilities of two distributions (5), where one represents the true distribution (P) and the other represents

the approximated distribution (Q) [61]. In the context of autoencoders, KL divergence is commonly utilized as a regularization term in the loss function to encourage the encoded representation of data to follow a specific probability distribution, such as a Gaussian distribution, and enforce sparsity [49].

$$D_{KL}(P|Q) = \sum_{i=1}^{n} P(i) \log \frac{P(i)}{Q(i)} \tag{5}$$

By using KL divergence, the model can avoid overfitting and improve its generalization performance. Moreover, by regularizing the encoded representation of data to follow a specific distribution, the model is forced to learn a more robust and meaningful representation of the input data, which can lead to better performance on downstream tasks, such as clustering or classification. Additionally, using KL divergence can make the autoencoder more interpretable, allowing the encoded representation of data to be more easily visualized and understood [49], [61], [62].

One of the advantages of the PAM clustering algorithm is that it is possible to use various distance metrics to compute the dissimilarities between samples. This study evaluated different distance metrics, and the best results were obtained when the Manhattan distance was used, as presented by (6). The Manhattan distance has several advantages, including its simplicity and computational efficiency [52]. It is also robust to noise and outliers because it does not square the differences between coordinates, which can lead to the amplification of noise.

$$dist_{Manhattan}\left(\boldsymbol{x}^{(j)}, \boldsymbol{m}^{(k)}\right) = \sum_{i=1}^{n} \left| x_i^{(j)} - m_i^{(k)} \right| \tag{6}$$

where $\boldsymbol{x}^{(j)}$ is the j^{th} sample point, and $\boldsymbol{m}^{(k)}$ is the k^{th} data point representing k^{th} medoid.

After training the optimal autoencoder, its encoder function $f(\cdot)$ maps the input data to a lower dimension to be fed into the clustering algorithm for determining the number of clusters, K. PAM clustering algorithm with Manhattan distance is run for various K values to identify the optimal number of clusters. The number of clusters is determined by computing the silhouette score for different K values [57], as shown in Fig. 5. However, it is worth to mention that the optimal number of clusters may not be unique, and other factors, such as the interpretability of the resulting clusters or domain-specific knowledge, should also be considered. Therefore, a cluster number of six (with silhouette score of 0.044) was chosen in this study to further explore the patterns.

Fig. 6 illustrates the dataset prior to the transformation and data mining steps. The t-SNE visualization technique is employed to project the high-dimensional input data (1440 dimensions) onto a two-dimensional space, providing a condensed representation [63].

Fig. 7 displays the results obtained when K is set to six. This figure showcases the encoder output that is mapped to a 2D dimension using the t-SNE visualization technique. Each cluster is represented by a distinct color, and the center of each cluster or medoids $\boldsymbol{m}^{(j)}$ is denoted by a cross in the 2D plane. It should be noted that the size of each cluster is not uniform, with some containing more samples than others. This variation in cluster size may indicate that the data distribution is not evenly spread and may require further investigation to uncover any underlying patterns or trends.

The projection of the dataset into a two-dimensional (2D) space using the learned latent features by the autoencoder will change. It is important to note that the data in the latent space, which has 60 dimensions, exhibits greater separability. However, due to the limitations of visualizing high-dimensional data in a 2D representation, this enhanced separability may not be readily apparent in the resulting figures. Nonetheless, the underlying improvements in separability contribute to the overall effectiveness of the autoencoder in capturing meaningful features and patterns within the dataset.

The clustering algorithm has revealed six distinct clusters, each of which corresponds to a unique daily PV power generation pattern. To represent each cluster, the PAM algorithm has computed medoids that have 60 dimensions. Unlike the centroids in K-means, which are calculated as the mean of all points in a cluster, medoids are actual data points from the dataset and have the smallest average dissimilarity to all other points in the same cluster, with non-sphere shapes. By mapping these medoids back to the original vector space with 1440 dimensions using the algorithm 1, the PV output power patterns for each cluster can be obtained, as shown in Fig. 8. Consequently, this presentation enables a clearer understanding of the different PV power generation patterns represented by each cluster.

The extracted PV output power patterns are associated with different types of days and seasons, providing valuable insights into solar energy generation. For instance, the pattern or cluster 3, which has the highest peak value, represents sunny days with a maximum capacity of PV power generation, while pattern or cluster 2 demonstrates mostly sunny days. Cluster 5, on the other hand, represents partially sunny days, which are characterized by a mix of sunshine and wind. Patterns or clusters 1 and 6 can be categorized as mostly and completely rainy or cloudy days, respectively. Lastly, pattern 4 is associated with zero or negligible PV generation and contrasts most with patterns of clusters 2 and 3. Since these patterns serve as the centers of their respective clusters, they exhibit smoother behavior instead of having spiky shapes with large spikes.

VI. DISCUSSION

This paper conducts further analysis to gain a deeper insight into the identified PV power patterns. This analysis provides a more comprehensive understanding of the characteristics and behaviors of each pattern.

The distribution of the PV power patterns identified by the proposed data-driven framework is not uniform. As illustrated in Fig. 9, some clusters contain a larger number of samples, indicating that those patterns occur more frequently

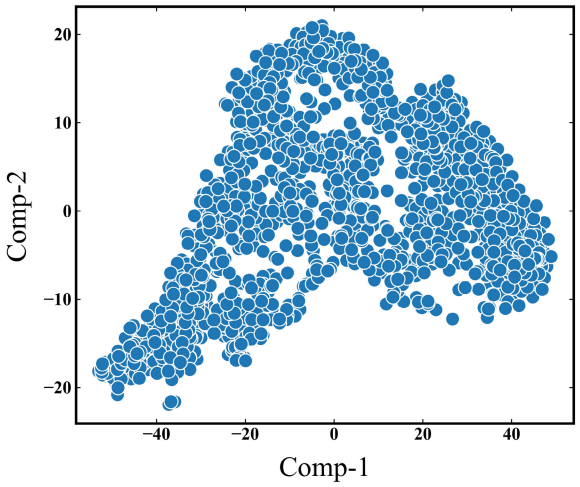


FIGURE 6. The original $\mathcal{D}_1$ dataset before the transformations and data mining procedures.

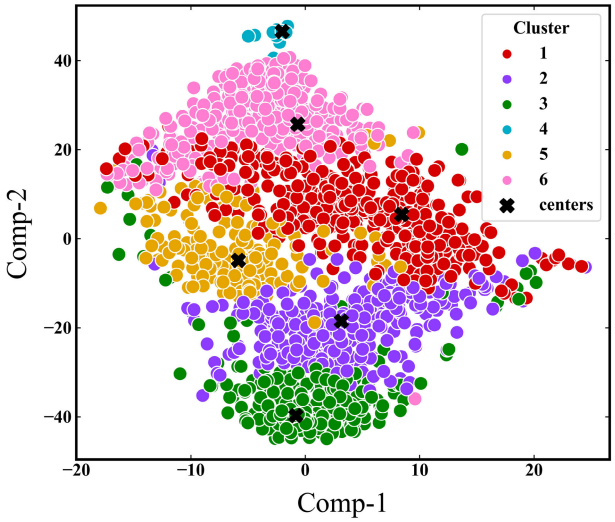


FIGURE 7. The clustering results using DAE and PAM algorithms in two dimensions for $\mathcal{D}_1$. The samples belonging to each cluster are depicted with the same color for better distinction.

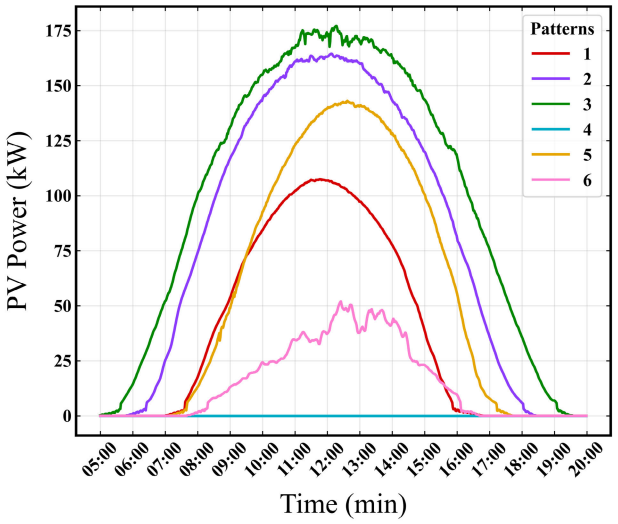


FIGURE 8. The PV power generation patterns recognized by the proposed data-driven framework for $\mathcal{D}_1$.

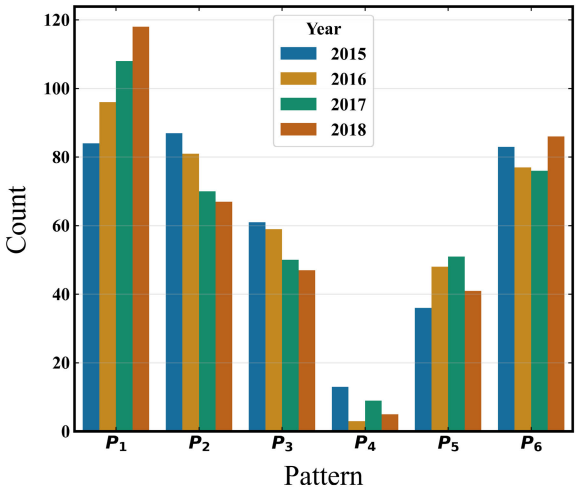


FIGURE 9. The occurrence of PV power patterns in $\mathcal{D}_1$ for each year.

throughout the year such as P_1. Conversely, clusters with fewer samples, such as P_4, represent PV power generation patterns that occur less frequently. This uneven distribution of patterns provides valuable insight into the varying nature of PV power generation across different periods of the year. Fig. 9 depicts that the patterns P_1 and P_6, with a maximum power output of 110 kW and 50 kW, respectively, have a higher frequency of occurrence compared to other patterns. Furthermore, patterns P_2 and P_3, with a maximum power output of 175 kW and 160 kW, respectively, also have a relatively high occurrence rate. These patterns correspond to the days with the highest PV power generation, as they have larger peak values compared to other patterns.

A dedicated figure, depicted in Fig. 10, has been designed to visualize the distribution of the identified PV power output patterns across different days.

Fig. 10 allows for a more comprehensive understanding of the occurrence of the patterns throughout the year. Each color in this figure corresponds to a distinct PV pattern, such as the green color representing P_3, which is associated with the highest PV power output, as shown in Fig. 8. The Figure reveals that some patterns, such as P_1 and P_6 with a maximum of 110 and 50 kW, are more frequent than others and seen across a year. In the contrary, the occurrence of P_2 and P_3 patterns, which are related to days with high PV power generation of up to 175 and 160 kW, respectively, is more frequent during specific weeks of the year. P_3 appears only from week 21 to week 33, which corresponds to summers, while P_2 occurs during weeks 12 to 21 and weeks 33 to 41, corresponding to spring and autumn. It is noteworthy that the model did not receive any information regarding the time of day for PV power generation but rather determined these

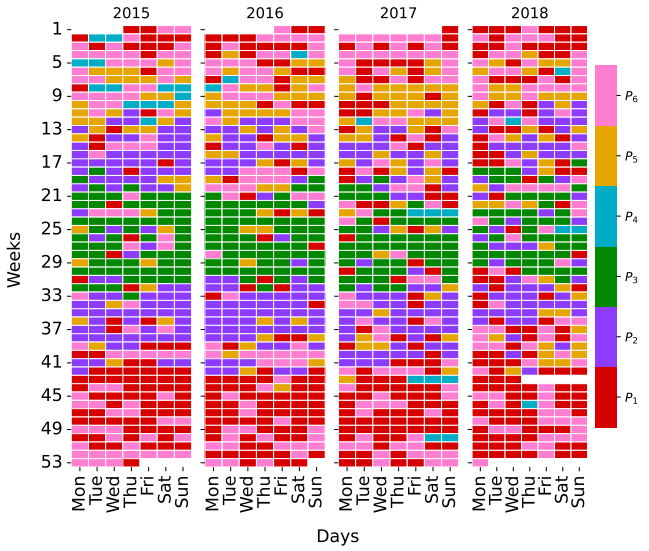


FIGURE 10. The occurrence of identified PV power generation patterns across different days in $\mathcal{D}_1$.

patterns by itself, which occur during specific times of the years.

Fig. 10 provides a more detailed insights and knowledge into PV power generation beyond the traditional approach of considering only the average seasonal PV production. By analyzing these patterns, new information can be extracted about PV power generation. For instance, pattern P_5, with a peak value of 140 kW, is predominantly observed during weeks 6 to 13, which usually corresponds to the months of February, March, and April. After week 41, only patterns P_1 and P_6 are visible, with peak values of 110 kW and 50 kW, respectively. Furthermore, these patterns offer additional information about the sequence of PV power generation, which can be quantified by considering the characteristics of each pattern.

Fig. 11 shows the PV power generation for 12 consecutive days in 2018 with corresponding patterns for production of each day. This figure displays the daily PV power generation, color-coded to correspond with the identified daily patterns. As an example, the PV power pattern associated with the highest peak values, P_3, is highlighted in green in all the figures. The PV power time-series in Fig. 11 has a high resolution of one sample per minute. This level of detail makes it possible to improve the approximation of daily PV production using the six identified patterns, as shown in the figure. This information can also be beneficial for energy management systems that incorporate storage units, electric vehicle charging stations, or a day-ahead PV power prediction.

To further analyze the extracted PV power output patterns, the proposed method is also applied to $\mathcal{D}_2$, resulting in the identification of six distinct patterns, as shown in Fig. 12.

TABLE 3. Sorted identified patterns based on peak values.

Dataset	Highest	High	Medium	Low	Lowest	Zero
$\mathcal{D}_1$	P_3	P_2	P_5	P_1	P_6	P_4
$\mathcal{D}_2$	P_8	P_{10}	P_{12}	P_7	P_{11}	P_9

To investigate the correlation between the patterns extracted from these two datasets, a correlation matrix is calculated and presented in Fig. 13. The patterns denoted from 1 to 6 correspond to dataset $\mathcal{D}_1$, whereas patterns labeled from 7 to 12 correspond to dataset $\mathcal{D}_2$. The top-left square in this figure represents the inner Pearson correlation of the identified patterns in $\mathcal{D}_1$, while the bottom-right square represents the inner Pearson correlation of the identified patterns in $\mathcal{D}_2$. The correlation coefficients between patterns in $\mathcal{D}_1$ reveal that P_1 and P_6 have a higher correlation coefficient, as well as P_2 with P_5. Conversely, P_3 displays the lowest correlation coefficient with other patterns. High correlation coefficients are observed between patterns P_7 and P_{11}, as well as between P_{10} and P_{12} in $\mathcal{D}_2$.

The correlation between observed patterns in $\mathcal{D}_1$ and $\mathcal{D}_2$ is displayed in the bottom-left part of the Fig. 13. The red circles highlight the highly correlated patterns between the two datasets. Specifically, P_1 with P_7, P_2 with P_{10}, P_3 with P_8, P_5 with P_{12}, and P_6 with P_{11} show high correlation coefficients. On the other hand, P_4 and P_9 have zero correlation with other patterns since they always have a constant value of zero, indicating no PV power generation. Additionally, sorting the patterns in each dataset based on their peak values, as shown in Table 3, also supports the high correlation between P_3 and P_8, as they both have the highest peak values.

The Pearson correlation matrix does not consider the correlation of each pattern with respect to time. The temporal correlation between patterns is taken into account by calculating the cumulative summation of each pattern with respect to time. This requires normalizing the patterns in each dataset based on their maximum values, followed by computing their cumulative summations, as shown in Fig. 14. This approach provides additional information on the temporal correlations between the patterns, which cannot be inferred from the Pearson correlation matrix alone. It can be observed from this figure that the correlated patterns exhibit similar trends and follow similar progressions over time, even if these patterns are obtained from two distinct PV plants with varying rated power.

It can be inferred from the figure that there exist some patterns, such as P_2 (P_{10}), P_3 (P_8), P_4 (P_9), and P_6 (P_{11}), which have high correlations and similar temporal trends in both datasets after normalization. These patterns can be categorized as typical patterns. Moreover, they can be interpreted as representing different weather conditions, including mostly sunny, totally sunny, zero or negligible generations, and cloudy days, respectively. Considering dataset $\mathcal{D}_1$, these

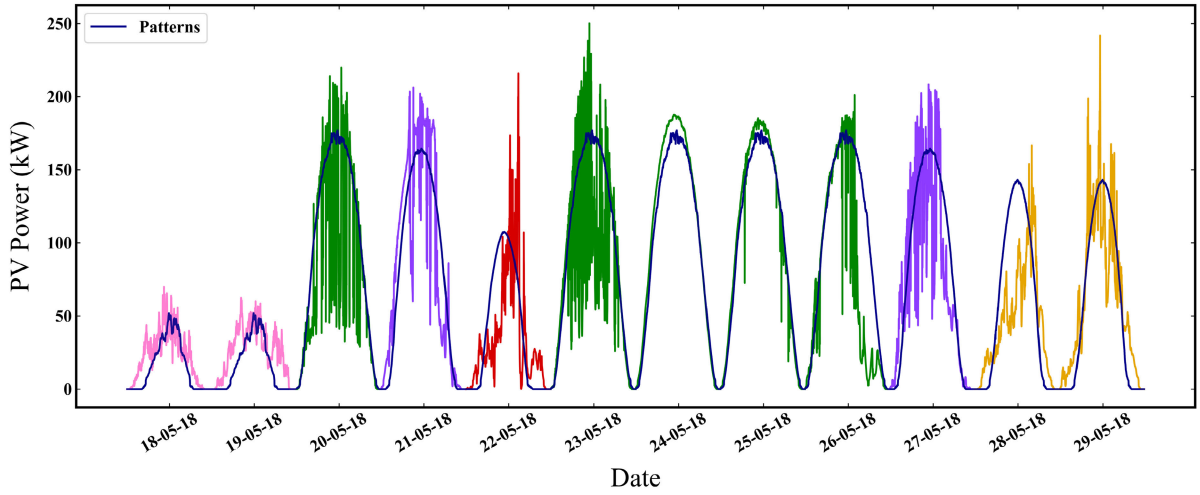


FIGURE 11. The PV power generation and the corresponding patterns for each day's production in 2018. the daily PV power generation, color-coded to correspond with the identified patterns.

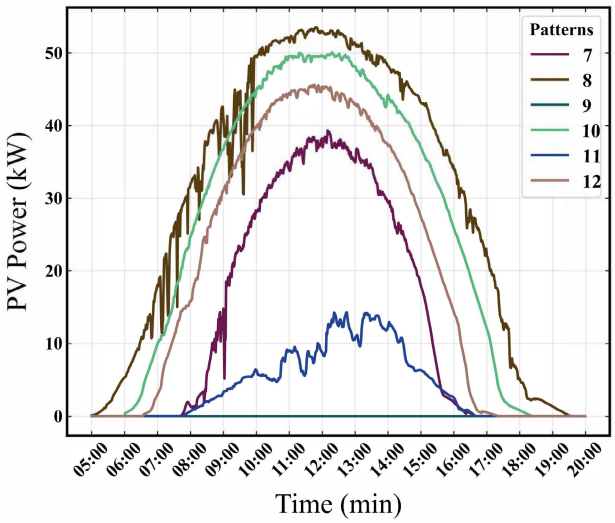


FIGURE 12. The PV power generation patterns recognized by the proposed data-driven framework for $\mathcal{D}_2$ with rated power of 75 kW.

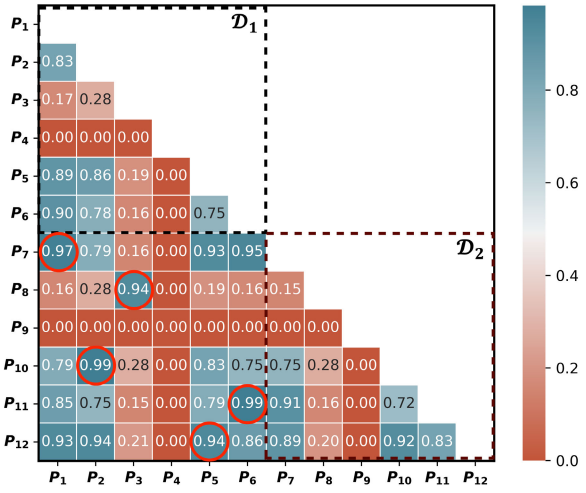


FIGURE 13. The correlation matrix between the identified patterns from two datasets, and $\mathcal{D}_2$. The patterns labeled from 1 to 6 are associated with $\mathcal{D}_1$, while patterns labeled from 7 to 12 are associated with $\mathcal{D}_2$.

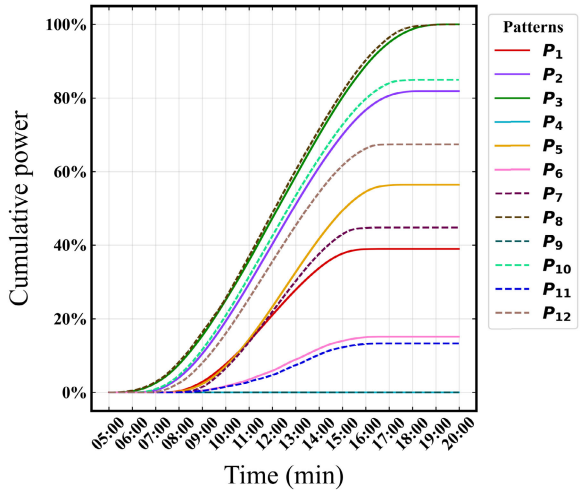


FIGURE 14. Cumulative distribution of patterns identified over time.

patterns have occurred in approximately 61% of the total time period analyzed. This information provides valuable insights into the varying levels of PV power production based on the prevailing weather conditions. On the other hand, the remaining patterns can be classified as local patterns. These typical patterns are anticipated to be observed in other locations as well since they represent common trends in PV power generation.

The potential applications of the identified PV power generation patterns are significant. They can be utilized to enhance the design and operation of PV systems by providing a deeper understanding of their power generation patterns. In addition, they can help to identify opportunities for implementing energy storage or demand response programs and scheduling, which can lead to more efficient energy management strategies [3], [6], [7], [8]. Moreover, these patterns can contribute to the development of more precise forecasting models, which can ultimately result in more accurate predictions of PV power output.

VII. CONCLUSION

This paper presented PV power generation patterns extracted by the proposed data-driven framework from oscillated daily PV power outputs per minute. The proposed data-driven model employs a deep autoencoder to capture complex and non-linear relationship in input dataset, reducing 1440 dimensions of the original dataset to a lower 60-dimensional latent feature space. The partitioning around medoids algorithm then clusters the data into six distinct groups. Using the proposed new algorithm, the center, or medoid closest to the center of each cluster, is determined in the original subspace to represent each cluster as its pattern. This model was applied to two real PV plant datasets with different rated power separately, identifying six distinct daily PV power generation patterns for each PV plant. Compared to the seasonal average behavior, these patterns provide new and more profound knowledge and insights into the daily PV power output repeated over the years. Statistical analysis showed that four patterns from the first dataset exhibited high correlation and distribution trends with the identified patterns from the second dataset, making them typical patterns that are expected to be observed in other locations as well.

The proposed framework is computationally efficient, scalable, and robust, making it ideal for identifying typical patterns in big data. It can provide critical insights and new knowledge into underlying patterns and trends in PV power generation, enabling informed decision-making in areas like energy management, PV hosting capacity determination, and policy-making. Furthermore, the flexibility of framework allows it to adapt to various datasets and applications, increasing its potential utility since it uses only one variable as an input. However, the challenge of applying the proposed method is designing and hyperparameter tuning a suitable deep learning autoencoder requires significant computational resources and is a time-consuming process. This is due to the complex architecture of deep learning models and the need to optimize numerous hyperparameters to achieve satisfactory performance. The availability of data poses another limitation when applying the proposed method. These techniques rely heavily on the availability of large and diverse datasets to effectively capture patterns, learn representations, and make accurate discoveries. Ensuring access to such datasets is essential for successfully applying these techniques.

ACKNOWLEDGMENT

This manuscript reflects only the authors' views and opinions, neither the European Union nor the European Commission can be considered responsible for them.

REFERENCES

[1] R. Ahmed, V. Sreeram, Y. Mishra, and M. D. Arif, "A review and evaluation of the state-of-the-art in PV solar power forecasting: Techniques and optimization," *Renew. Sustain. Energy Rev.*, vol. 124, May 2020, Art. no. 109792.

[2] G. G. Kim, J. H. Choi, S. Y. Park, B. G. Bhang, W. J. Nam, H. L. Cha, N. Park, and H. Ahn, "Prediction model for PV performance with correlation analysis of environmental variables," *IEEE J. Photovolt.*, vol. 9, no. 3, pp. 832–841, May 2019, doi: 10.1109/JPHOTOV.2019.2898521.

[3] D. Azuatalam, K. Paridari, Y. Ma, M. Förstl, A. C. Chapman, and G. Verbič, "Energy management of small-scale PV-battery systems: A systematic review considering practical implementation, computational requirements, quality of input data and battery degradation," *Renew. Sustain. Energy Rev.*, vol. 112, pp. 555–570, Sep. 2019, doi: 10.1016/j.rser.2019.06.007.

[4] N. T. Mbungu, R. C. Bansal, R. M. Naidoo, M. Bettayeb, M. W. Siti, and M. Bipath, "A dynamic energy management system using smart metering," *Appl. Energy*, vol. 280, Dec. 2020, Art. no. 115990, doi: 10.1016/j.apenergy.2020.115990.

[5] S. M. Miraftabzadeh and M. Longo, "High-resolution PV power prediction model based on the deep learning and attention mechanism," *Sustain. Energy, Grids Netw.*, vol. 34, Jun. 2023, Art. no. 101025, doi: 10.1016/j.segan.2023.101025.

[6] E. Bullich-Massagué, F.-J. Cifuentes-García, I. Glenny-Crende, M. Cheah-Mañé, M. Aragüés-Peñalba, F. Díaz-González, and O. Gomis-Bellmunt, "A review of energy storage technologies for large scale photovoltaic power plants," *Appl. Energy*, vol. 274, Sep. 2020, Art. no. 115213, doi: 10.1016/j.apenergy.2020.115213.

[7] A. Cabrera-Tobar, E. Bullich-Massagué, M. Aragüés-Peñalba, and O. Gomis-Bellmunt, "Review of advanced grid requirements for the integration of large scale photovoltaic power plants in the transmission system," *Renew. Sustain. Energy Rev.*, vol. 62, pp. 971–987, Sep. 2016, doi: 10.1016/j.rser.2016.05.044.

[8] A. Koirala, T. Van Acker, R. D'Hulst, and D. Van Hertem, "Hosting capacity of photovoltaic systems in low voltage distribution systems: A benchmark of deterministic and stochastic approaches," *Renew. Sustain. Energy Rev.*, vol. 155, Mar. 2022, Art. no. 111899, doi: 10.1016/j.rser.2021.111899.

[9] T. W. Liao, "Clustering of time series data—A survey," *Pattern Recognit.*, vol. 38, no. 11, pp. 1857–1874, Nov. 2005, doi: 10.1016/j.patcog.2005.01.025.

[10] T. Schütz, M. H. Schraven, M. Fuchs, P. Remmen, and D. Müller, "Comparison of clustering algorithms for the selection of typical demand days for energy system synthesis," *Renew. Energy*, vol. 129, pp. 570–582, Dec. 2018, doi: 10.1016/j.renene.2018.06.028.

[11] J. Li, H. Izakian, W. Pedrycz, and I. Jamal, "Clustering-based anomaly detection in multivariate time series data," *Appl. Soft Comput.*, vol. 100, Mar. 2021, Art. no. 106919, doi: 10.1016/j.asoc.2020.106919.

[12] S. M. Miraftabzadeh, M. Longo, F. Foiadelli, M. Pasetti, and R. Igual, "Advances in the application of machine learning techniques for power system analytics: A survey," *Energies*, vol. 14, no. 16, p. 4776, Aug. 2021, doi: 10.3390/en14164776.

[13] G. Yao, Y. Wu, X. Huang, Q. Ma, and J. Du, "Clustering of typical wind power scenarios based on K-means clustering algorithm and improved artificial bee colony algorithm," *IEEE Access*, vol. 10, pp. 98752–98760, 2022, doi: 10.1109/ACCESS.2022.3203695.

[14] P. Gianniou, X. Liu, A. Heller, P. S. Nielsen, and C. Rode, "Clustering-based analysis for residential district heating data," *Energy Convers. Manag.*, vol. 165, pp. 840–850, Jun. 2018, doi: 10.1016/j.enconman.2018.03.015.

[15] S. Yilmaz, J. Chambers, and M. K. Patel, "Comparison of clustering approaches for domestic electricity load profile characterisation—Implications for demand side management," *Energy*, vol. 180, pp. 665–677, Aug. 2019, doi: 10.1016/j.energy.2019.05.124.

[16] G. Chicco, O. Ionel, and R. Porumb, "Electrical load pattern grouping based on centroid model with ant colony clustering," *IEEE Trans. Power Syst.*, vol. 28, no. 2, pp. 1706–1715, May 2013, doi: 10.1109/TPWRS.2012.2220159.

[17] K. Li, Z. Ma, D. Robinson, and J. Ma, "Identification of typical building daily electricity usage profiles using Gaussian mixture model-based clustering and hierarchical clustering," *Appl. Energy*, vol. 231, pp. 331–342, Dec. 2018, doi: 10.1016/j.apenergy.2018.09.050.

[18] K. Ø. Mikalsen, F. M. Bianchi, C. Soguero-Ruiz, and R. Jenssen, "Time series cluster kernel for learning similarities between multivariate time series with missing data," *Pattern Recognit.*, vol. 76, pp. 569–581, Apr. 2018, doi: 10.1016/j.patcog.2017.11.030.

[19] G. Chicco, R. Napoli, and F. Piglione, "Comparisons among clustering techniques for electricity customer classification," *IEEE Trans. Power Syst.*, vol. 21, no. 2, pp. 933–940, May 2006, doi: 10.1109/TPWRS.2006.873122.

[20] S. M. Miraftabzadeh, M. Longo, M. Brenna, and M. Pasetti, "Data-driven model for PV power generation patterns extraction via unsupervised machine learning methods," in *Proc. North Amer. Power Symp. (NAPS)*, Oct. 2022, pp. 1–5.

[21] L. Wen, K. Zhou, and S. Yang, "A shape-based clustering method for pattern recognition of residential electricity consumption," *J. Cleaner Prod.*, vol. 212, pp. 475–488, Mar. 2019, doi: 10.1016/j.jclepro.2018.12.067.

[22] H. Li, "Multivariate time series clustering based on common principal component analysis," *Neurocomputing*, vol. 349, pp. 239–247, Jul. 2019, doi: 10.1016/j.neucom.2019.03.060.

[23] Y. Mohammadi, S. M. Miraftabzadeh, M. H. J. Bollen, and M. Longo, "An unsupervised learning schema for seeking patterns in rms voltage variations at the sub-10-minute time scale," *Sustain. Energy, Grids Netw.*, vol. 31, Sep. 2022, Art. no. 100773, doi: 10.1016/j.segan.2022.100773.

[24] Y. Mohammadi, S. M. Miraftabzadeh, M. H. J. Bollen, and M. Longo, "Seeking patterns in rms voltage variations at the sub-10-minute scale from multiple locations via unsupervised learning and patterns' post-processing," *Int. J. Electr. Power Energy Syst.*, vol. 143, Dec. 2022, Art. no. 108516, doi: 10.1016/j.ijepes.2022.108516.

[25] O. F. Özgül, B. Bardak, and M. Tan, "A convolutional deep clustering framework for gene expression time series," *IEEE/ACM Trans. Comput. Biol. Bioinf.*, vol. 18, no. 6, pp. 2198–2207, Nov. 2021, doi: 10.1109/TCBB.2020.2988985.

[26] S. M. Mousavi, W. Zhu, W. Ellsworth, and G. Beroza, "Unsupervised clustering of seismic signals using deep convolutional autoencoders," *IEEE Geosci. Remote Sens. Lett.*, vol. 16, no. 11, pp. 1693–1697, Nov. 2019, doi: 10.1109/LGRS.2019.2909218.

[27] C. Aytekin, X. Ni, F. Cricri, and E. Aksu, "Clustering and unsupervised anomaly detection with l2 normalized deep auto-encoder representations," in *Proc. Int. Joint Conf. Neural Netw. (IJCNN)*, Jul. 2018, pp. 1–6, doi: 10.1109/IJCNN.2018.8489068.

[28] C. Bian, X. Wang, Y. Su, Y. Wang, K.-C. Wong, and X. Li, "ScEFSC: Accurate single-cell RNA-seq data analysis via ensemble consensus clustering based on multiple feature selections," *Comput. Struct. Biotechnol. J.*, vol. 20, pp. 2181–2197, Jan. 2022, doi: 10.1016/j.csbj.2022.04.023.

[29] T. A. Geddes, T. Kim, L. Nan, J. G. Burchfield, J. Y. H. Yang, D. Tao, and P. Yang, "Autoencoder-based cluster ensembles for single-cell RNA-seq data analysis," *BMC Bioinf.*, vol. 20, no. S19, pp. 1–11, Dec. 2019, doi: 10.1186/s12859-019-3179-5.

[30] M. M. Fard, T. Thonet, and E. Gaussier. (2020). *Pattern Recognition Letters Deep k-Means: Jointly Clustering With k-Means and Learning Representations*. [Online]. Available: https://www.elsevier.com/open-access/userlicense/1.0/

[31] H. Teichgraeber and A. R. Brandt, "Clustering methods to find representative periods for the optimization of energy systems: An initial framework and comparison," *Appl. Energy*, vol. 239, pp. 1283–1293, Apr. 2019, doi: 10.1016/j.apenergy.2019.02.012.

[32] L. G. B. Ruiz, M. C. Pegalajar, R. Arcucci, and M. Molina-Solana, "A time-series clustering methodology for knowledge extraction in energy consumption data," *Expert Syst. Appl.*, vol. 160, Dec. 2020, Art. no. 113731, doi: 10.1016/j.eswa.2020.113731.

[33] F. Grasso, C. I. Garcia, G. M. Lozito, and G. Talluri, "Artificial load profiles and PV generation in renewable energy communities using generative adversarial networks," in *Proc. IEEE 21st Medit. Electrotech. Conf. (MELECON)*, Jun. 2022, pp. 709–714, doi: 10.1109/MELECON53508.2022.9843062.

[34] C. Dai, J. Wu, D. Pi, S. I. Becker, L. Cui, Q. Zhang, and B. Johnson, "Brain EEG time-series clustering using maximum-weight clique," *IEEE Trans. Cybern.*, vol. 52, no. 1, pp. 357–371, Jan. 2022, doi: 10.1109/TCYB.2020.2974776.

[35] Mustaqeem, M. Sajjad, and S. Kwon, "Clustering-based speech emotion recognition by incorporating learned features and deep BiLSTM," *IEEE Access*, vol. 8, pp. 79861–79875, 2020, doi: 10.1109/ACCESS.2020.2990405.

[36] O. Motlagh, A. Berry, and L. O'Neil, "Clustering of residential electricity customers using load time series," *Appl. Energy*, vol. 237, pp. 11–24, Mar. 2019, doi: 10.1016/j.apenergy.2018.12.063.

[37] W. W. Tso, C. D. Demirhan, C. F. Heuberger, J. B. Powell, and E. N. Pistikopoulos, "A hierarchical clustering decomposition algorithm for optimizing renewable power systems with storage," *Appl. Energy*, vol. 270, Jul. 2020, Art. no. 115190, doi: 10.1016/j.apenergy.2020.115190.

[38] G. Liu, L. Zhu, X. Wu, and J. Wang, "Time series clustering and physical implication for photovoltaic array systems with unknown working conditions," *Sol. Energy*, vol. 180, pp. 401–411, Mar. 2019, doi: 10.1016/j.solener.2019.01.041.

[39] L. Xu, Z. Pan, C. Liang, and M. Lu, "A fault diagnosis method for PV arrays based on new feature extraction and improved the fuzzy C-mean clustering," *IEEE J. Photovolt.*, vol. 12, no. 3, pp. 833–843, May 2022, doi: 10.1109/JPHOTOV.2022.3151330.

[40] T. B. P. Nguyen, Y. Wu, and M. Pham, "A novel data-driven method to estimate invisible solar power generation: A case study in Taiwan," *IEEE Trans. Ind. Appl.*, vol. 58, no. 6, pp. 7057–7067, Nov. 2022, doi: 10.1109/TIA.2022.3201810.

[41] F. Wang, Z. Zhen, Z. Mi, H. Sun, S. Su, and G. Yang, "Solar irradiance feature extraction and support vector machines based weather status pattern recognition model for short-term photovoltaic power forecasting," *Energy Buildings*, vol. 86, pp. 427–438, Jan. 2015, doi: 10.1016/j.enbuild.2014.10.002.

[42] J. Martinek and M. J. Wagner, "Efficient prediction of concentrating solar power plant productivity using data clustering," *Sol. Energy*, vol. 224, pp. 730–741, Aug. 2021, doi: 10.1016/j.solener.2021.06.002.

[43] F. Wang, J. Li, Z. Zhen, C. Wang, H. Ren, H. Ma, W. Zhang, and L. Huang, "Cloud feature extraction and fluctuation pattern recognition based ultrashort-term regional PV power forecasting," *IEEE Trans. Ind. Appl.*, vol. 58, no. 5, pp. 6752–6767, Sep. 2022, doi: 10.1109/TIA.2022.3186662.

[44] Z. Zhen, S. Pang, F. Wang, K. Li, Z. Li, H. Ren, M. Shafie-Khah, and J. P. S. Catalão, "Pattern classification and PSO optimal weights based sky images cloud motion speed calculation method for solar PV power forecasting," *IEEE Trans. Ind. Appl.*, vol. 55, no. 4, pp. 3331–3342, Jul. 2019, doi: 10.1109/TIA.2019.2904927.

[45] F. Wang, K. Li, L. Zhou, H. Ren, J. Contreras, M. Shafie-Khah, and J. P. S. Catalão, "Daily pattern prediction based classification modeling approach for day-ahead electricity price forecasting," *Int. J. Electr. Power Energy Syst.*, vol. 105, pp. 529–540, Feb. 2019, doi: 10.1016/j.ijepes.2018.08.039.

[46] U. Fayyad, G. Piatetsky-Shapiro, and P. Smyth, "From data mining to knowledge discovery in databases," in *Proc. AAAI*, 1996, pp. 37–54. [Online]. Available: http://www.ffly.com/

[47] S. Aghabozorgi, A. S. Shirkhorshidi, and T. Y. Wah, "Time-series clustering—A decade review," *Inf. Syst.*, vol. 53, pp. 16–38, Oct. 2015, doi: 10.1016/j.is.2015.04.007.

[48] V. Perrone, M. Donini, M. B. Zafar, R. Schmucker, K. Kenthapadi, and C. Archambeau, "Fair Bayesian optimization," Jun. 2020, *arXiv:2006.05109*.

[49] D. P. Kingma and M. Welling, "An introduction to variational autoencoders," *Found. Trends Mach. Learn.*, vol. 12, no. 4, pp. 1–18, 2019, doi: 10.1561/2200000056.

[50] I. Goodfellow, Y. Bengio, and A. Courville, *Deep Learning*. Cambridge, MA, USA: MIT Press, 2016, pp. 504–526.

[51] P. D'Urso, L. D. Giovanni, R. Massari, R. L. D'Ecclesia, and E. A. Maharaj, "Cepstral-based clustering of financial time series," *Expert Syst. Appl.*, vol. 161, Dec. 2020, Art. no. 113705, doi: 10.1016/j.eswa.2020.113705.

[52] M. Mittal, L. M. Goyal, D. J. Hemanth, and J. K. Sethi, "Clustering approaches for high-dimensional databases: A review," *WIREs Data Mining Knowl. Discovery*, vol. 9, no. 3, p. e1300, May 2019, doi: 10.1002/widm.1300.

[53] A. E. Ezugwu, A. K. Shukla, M. B. Agbaje, O. N. Oyelade, A. José-García, and J. O. Agushaka, "Automatic clustering algorithms: A systematic review and bibliometric analysis of relevant literature," *Neural Comput. Appl.*, vol. 33, no. 11, pp. 6247–6306, Jun. 2021, doi: 10.1007/s00521-020-05395-4.

[54] E. Schubert and P. J. Rousseeuw, "Fast and eager k-medoids clustering: O(k) runtime improvement of the PAM, CLARA, and CLARANS algorithms," *Inf. Syst.*, vol. 101, Nov. 2021, Art. no. 101804, doi: 10.1016/j.is.2021.101804.

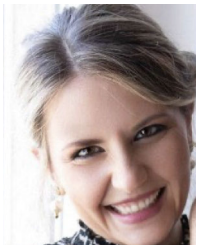
[55] F. Anowar, S. Sadaoui, and B. Selim, "Conceptual and empirical comparison of dimensionality reduction algorithms (PCA, KPCA, LDA, MDS, SVD, LLE, ISOMAP, LE, ICA, t-SNE)," *Comput. Sci. Rev.*, vol. 40, May 2021, Art. no. 100378, doi: 10.1016/j.cosrev.2021.100378.

[56] C. R. Harris et al., "Array programming with NumPy," *Nature*, vol. 585, no. 7825, pp. 357–362, Sep. 2020, doi: 10.1038/s41586-020-2649-2.

[57] E. Schubert and P. J. Rousseeuw, "Faster k-medoids clustering: Improving the PAM, CLARA, and CLARANS algorithms," in *Proc. 12th Int. Conf. Similarity Search Appl. (SISAP)*, Newark, NJ, USA, Oct. 2019, pp. 171–187.

[58] M. T. Boyd, "NIST weather station for photovoltaic and building system research," U.S. Dept. Commerce, Nat. Inst. Standards Technol., Gaithersburg, MD, USA, Tech. Rep., 2014, doi: 10.6028/NIST.TN.1913.
[59] *NIST Data Repository Page*. Accessed: Jun. 19, 2023. [Online]. Available: https://data.nist.gov/sdp/#/
[60] S. M. Miraftabzadeh, C. G. Colombo, M. Longo, and F. Foiadelli, "A day-ahead photovoltaic power prediction via transfer learning and deep neural networks," *Forecasting*, vol. 5, no. 1, pp. 213–228, Feb. 2023, doi: 10.3390/forecast5010012.
[61] R. Wei and A. Mahmood, "Recent advances in variational autoencoders with representation learning for biomedical informatics: A survey," *IEEE Access*, vol. 9, pp. 4939–4956, 2021, doi: 10.1109/ACCESS.2020.3048309.
[62] D. P Kingma and M. Welling, "Auto-encoding variational Bayes," Dec. 2013, *arXiv:1312.6114*.
[63] L. Van Der Maaten and G. Hinton, "Visualizing data using t-SNE," *J. Mach. Learn. Res.*, vol. 9, no. 11, pp. 1–27, 2008.

SEYED MAHDI MIRAFTABZADEH (Member, IEEE) received the M.S. degree in electrical engineering from the Politecnico di Milano, Italy, in 2017, and the Ph.D. degree, in 2021. He is currently an Assistant Professor with the Department of Energy, Politecnico di Milano. His current research interests include the application of artificial intelligence and deep learning in electrical systems and smart cities to improve power grid stability and promote e-mobilities.



MICHELA LONGO (Member, IEEE) received the M.Sc. degree in information engineering and the Ph.D. degree in mechatronics, information, innovative technologies, and mathematical methods from the University of Bergamo, Bergamo, Italy, in 2009 and 2013, respectively. She is currently an Associate Professor with the Department of Energy, Politecnico di Milano. Her research interests include electric power systems and electric traction. She is a member of the Italian Group of Engineering About Railways (CIFI) and the Italian Association of Electrical, Electronics, Automation, Information and Communication Technology (AEIT).

MORRIS BRENNA (Member, IEEE) received the M.S. degree in electrical engineering from the Politecnico di Milano, Italy, in 1999, and the Ph.D. degree, in 2003. He is currently a Full Professor with the Department of Energy, Politecnico di Milano. His current research interests include power electronics, distributed generation, electromagnetic compatibility, and electric traction system. He is a member of the Italian Association of Electrical, Electronics, Automation, Information and Communication Technology (AEIT) and the Italian Group of Engineering about Railways (CIFI).

Open Access funding provided by 'Politecnico di Milano' within the CRUI CARE Agreement