Enhanced Spectral Ensemble Clustering for Fault Diagnosis: Application to Photovoltaic Systems

Mohsen Zargarani, Claude Delpha, Demba Diallo, Anne Migan-Dubois, Chabakata Mahamat, Laurent Linguet

IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Received 4 October 2024, accepted 25 October 2024, date of publication 13 November 2024, date of current version 26 November 2024.

Digital Object Identifier 10.1109/ACCESS.2024.3497977

RESEARCH ARTICLE

Enhanced Spectral Ensemble Clustering for Fault Diagnosis: Application to Photovoltaic Systems

MOHSEN ZARGARANI[1], CLAUDE DELPHA[2], (Senior Member, IEEE), DEMBA DIALLO[3], (Senior Member, IEEE), ANNE MIGAN-DUBOIS[3], CHABAKATA MAHAMAT[1], AND LAURENT LINGUET[1]

[1]UMR-Espace Dev, Université de Guyane, 97300 Cayenne, France
[2]University Paris-Saclay, CNRS, CentraleSupelec, L2S, Gif-Sur-Yvette 91192, France
[3]University Paris-Saclay, CentraleSupelec, CNRS, GeePs, Gif-Sur-Yvette 91192, France

Corresponding author: Mohsen Zargarani (zargarani.mn@gmail.com)

ABSTRACT The role of clustering in unsupervised fault diagnosis is significant, but different clustering techniques can yield varied results and cause inevitable uncertainty. Ensemble clustering methods have been introduced to tackle this challenge. This study presents a novel integrated technique in the field of fault diagnosis using spectral ensemble clustering. A new dimensionality reduction technique is proposed to intelligently identify faults, even in ambiguous scenarios, by exploiting the informative segment of the underlying bipartite graph. This is achieved by identifying and extracting the most informative sections of the bipartite graph based on the eigenvector centrality measure of nodes within the graph. The proposed method is applied to experimental current-voltage (I-V) curve data collected from a real photovoltaic (PV) platform. The obtained results remarkably improved the accuracy of aging fault detection to more than 83.50%, outperforming the existing state-of-the-art approaches. We also decided to separately analyze the ensemble clustering part of our FDD method, which indicated surpassing performance compared to similar methods by evaluating commonly used datasets like handwritten datasets. This proves that the proposed approach inherently holds promise for application in various real-world scenarios that are indicated by ambiguity and complexity.

INDEX TERMS Enhanced spectral ensemble clustering (ESEC), bipartite graph partitioning, eigenvector centrality, neural networks, fault detection and diagnosis (FDD), photovoltaic (PV) system.

I. INTRODUCTION

In PV systems, fault detection and diagnosis (FDD) ensures superior reliability, safety, and energy efficiency. AI-based mechanisms are an effective solution to attain acceptable FDD efficiencies. Robust methods for employing machine learning in FDD exist, many based on supervised learning, and require well-labeled datasets. However, accumulating and accessing the history of events and faults in systems such as PV plants can be quite costly due to inherent obstacles [61]. Unsupervised learning is valued mainly because it relies less on well-labeled datasets. Unsupervised learning methods often use clustering, a key data analysis technique. To apply clustering for FDD purposes, clusters must be interpreted to determine the type of faults [7], [20], [22]. It must be noted that since clustering techniques generally depend on dissimilarity and similarity means, they can generate different results. The outcomes may change even when using a similar clustering method with only different parameters or initializations [1]. In addition, finding a clustering approach suitable for diverse cases is still an unsolved challenge [2]. Specifically in PV systems, a single clustering approach may result in lower performance when dealing with unseen I-V curve data from PV plants with larger sizes or different string architectures. Ensemble clustering is a promising solution to address these limitations. Firstly, multiple base clustering methods are used to obtain original base clusterings. Then, the outputs from these clustering methods are integrated within a consensus framework. Among many ensemble clustering methods, spectral ensemble clustering is particularly

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

TABLE 1. Nomenclature.

Symbol	Description	Symbol	Description
ψ^i	i^{th} Partition of base clustering	G	Bipartite graph
Ψ	The set of all partitions	EC	Eigenvector centrality
$\boldsymbol{Z}$	Dataset (I-V characteristics)	$\lambda_{\max(G)}$	The highest eigenvalue of G
θ_{ij}	j^{th} cluster inside the i^{th} partition	q	Number of neighboring nodes for vertex v_i
$\boldsymbol{A}$	Total affinity matrix	X	Dataset before K-means discretization
$\hat{\boldsymbol{A}}$	Adjacency matrix	k	Number of clusters for K-means partitioning
Θ	The set of all ensemble of clusters	c	The set of cluster centroids
θ_f	f^{th} cluster inside Θ	$\hat{\theta}$	The set of clusters generated by K-means
P_i	Number of clusters inside the i^{th} partition.	d	Distance function
N	Number of base clusterings	$\mathcal{F}(\theta)$	K-means objective function
M	Number of ensemble clusters	$\Gamma(\theta_j)$	K-means dispersion function for cluster j
r_{lf}	Member of cross-affinity matrix	G'	Bipartite graph after outlier removal
m	Number of data rows	G_d	Graph of the reduced eigenproblem
$\boldsymbol{R}$	Cross affinity matrix	ω	Eigenvector for reduced eigen problem
$\hat{\boldsymbol{P}}$	Matrix of transition probability	λ	Eigenvalue for reduced eigenproblem
h	Conversion function for reduced eigenproblem	$\boldsymbol{D_z}$	Diagonal matrix in reduced eigenproblem
γ	Eigenvalue for the bipartite graph G	ρ	Eigenvector for the bipartite graph G
σ	Standard deviation	$e(G)$	Number of edges in graph G
L	Laplacian matrix	D	Degree matrix
L_d	Laplacian matrix of reduced eigenproblem	n_v	Number of all vertices in the graph

effective for training nonlinear systems and managing high-dimensional data [63], [64]. In this article, spectral ensemble clustering is selected for FDD due to its promising function for both base clustering and consensus clustering steps. The main concern in spectral ensemble clustering is the complexity cost [62]. The main challenge is developing methods to minimize dimensionality while keeping essential information. Techniques such as PCA [3], outlier detection [4], and backward feature elimination [5] are insufficient for maintaining valuable underlying information in clustering [6]. This issue remains open in most FDD approaches, even when clustering techniques are used. Determining which data partitions carry the most valuable information is the primary task to detect faults in large volumes of data. In this article, a new method for FDD using spectral ensemble clustering augmented by smart outlier detection is designed. We believe this approach can potentially make significant advancements in PV systems diagnosis. We list our contributions as follows.

1. Integration of Spectral Ensemble Clustering and Fault Diagnosis: Historically, there has been limited integration between spectral ensemble clustering and fault diagnosis in PV systems. To bridge this gap, a novel integrated framework that combines spectral ensemble clustering with PV FDD is proposed.

2. Revealing Hidden Patterns: To reach an accurate diagnosis, a profound comprehension of the underlying meanings and patterns within the data is vitally important. The methodology presented defines a new bipartite graph partitioning stage in the segmentation part. As an overpowering mechanism for uncovering ambiguous or concealed defects, the new intelligent outlier detection is really advantageous. Also, It can predict weaknesses or early-stage faults profoundly.

The notations used in this article are overally displayed in Table 1. The contents of the rest of this paper are arranged as follows. The related works are discussed broadly in section II, where relevant context and background are offered. Section III delves into the enhanced spectral ensemble clustering methodology, while Section IV presents the application and in-depth analysis of the results.

II. RELATED WORKS

A. PV FDD USING CLUSTERING-BASED METHODS

For our specific case study, we focus on PV systems. The field of fault diagnosis for PV systems has witnessed numerous studies that employ clustering methods directly or combined with other techniques [19]. Among clustering-based FDD methods, K-means is a commonly used clustering algorithm known for its robust performance on diverse feature-rich data. For instance, it has been applied in fault detection analysis with thermal images and also for real-time PV diagnosis [23], [24], [25], [26]. However, it must be noted that K-means can be affected by initial cluster centers and may even converge to an unwanted local optimum [27], [56]. Another common clustering method used in PV FDD is "Fuzzy C-Means" (FCM). An FDD method using an integrated clustering technique is developed by Lei et al. They designed a two-stage feature selection technique, followed by an embedded weighting inside the FCM algorithm [21]. Honglu Zhu et al. introduced an "FDD model based on FCM.

TABLE 2. Different base clustering methods.

Clustering method	Advantage	Disadvantage	Clustering method	Advantage	Disadvantage
K-means	Simple, fast, and flexible [36]	Sensitive to non-convex shape clusters, outliers, initialization [56],[58]	OPTICS	Handles datasets with varying densities [41]	Dependency to distance measurement [41]
Agglomerative	understandable Tree-like output, visualization [55]	weak for categorical data [37], need manual interpretation, time cost,	BIRCH	Proper noise handling. It is suitable for spherical clusters [38]	Sensitive to the cluster shape [38]
Spectral	Handles non-globular clusters [39]	High computation and time cost. It needs postprocessing [40]	Affinity Propagation	No need to specify the number of clusters [43]	Slow algorithm [43]
DBSCAN	Handles arbitrary shape, Roust to outlier [57]	Sensitive to density hyperparameters [41]	Gaussian Mixture	handles missing data, Robust to outliers [44]	Sensitive to initialization, computational cost [44]
Mean shift	robust to cluster shape Handles arbitrary feature spaces [42]	Sensitive to window size selection [42]	FCM	Handles overlapped datasets [45]	Inclination to local optima[45]

They subsequently fed the generated labeled features to a ''probabilistic neural network'' (PNN) for classification. This hybrid method combines clustering and classification and is referred to as the ''cluster-then-label'' (CTL) algorithm [19]. Although the FCM method is effective, it has difficulty identifying non-spherical clusters. Additionally, FCM requires pre-defined cluster centers before the clustering process [28]. Shengyang Liu et al. developed a clustering-based algorithm for FDD using dilation and erosion theory. Notably, their method can detect unknown faults, as the number of clusters is not predetermined [29]. Yongjie Liu et al. proposed a fault diagnosis approach based on ''Clustering by Fast Search and Find of Density Peaks'' (CFSFDP) clustering and stacked auto-encoders. They achieved dimensionality reduction using t-distributed stochastic neighbor embedding [28]. Yuqiao Cai et al. presented an online fault diagnosis method based on data stream clustering. Specifically, they utilized the ''Density-Based Spatial Clustering of Applications with Noise'' (DBSCAN) technique with input from PV array data streams. This approach is explicitly for grid-connected PV plants [30]. The critical aspects of DBSCAN involve selecting core points based on predefined density thresholds. However, it is worth noting that these types of density thresholds, which are usually artificially set, can crucially change the clustering performance [28]. de Guia et al. used the mean-shift method for both outlier detection and then classification in an ensemble routine [31]. Hierarchical clustering has also been studied using similar FDD methods [32]. Shushan Wu et al. applied ''Multivariate Functional Principal Component Analysis'' (MFPCA) clustering. Their algorithm extracts current and voltage data from coupling nodes [33]. Pham et al. introduced an FDD approach based on dynamic clustering, primarily focused on detecting open circuit faults in inverters [34]. Most of the aforementioned fault diagnosis methods rely on a single clustering technique. However, there is a scarcity of research on PV FDD methods based on ensemble clustering [35].

B. SPECTRAL ENSEMBLE CLUSTERING

There are different ways of ensemble clustering [8], [9]. In this article, spectral ensemble clustering is selected, which has emerged as a key technique explored by various research approaches. For instance, Liang et al. formulated a spectral ensemble clustering technique that constructs a representative co-association matrix within a unified constrained framework for optimization [10]. Li et al. defined spectral ensemble clustering for large-scale data by generating basic clusterings and combining them using a bipartite graph [11]. Dong Huang et al. proposed a two-stage spectral ensemble clustering approach. A bipartite graph is designed to relate objects and base clusters in their generation stage, and the consensus clustering results from partitioning this bipartite graph [12]. Wenguang Wang et al. addressed ensemble clustering by utilizing correlation links to select proper hyperspectral bands. Their method involves an agglomerative clustering approach to shape the consensus function [13]. Fei et al. introduced a selective spectral ensemble clustering method [14]. The authors used stochastic initialization and ''normalized mutual information'' (NMI) to shape basic clusterings in their method. The consensus function is based on the density peaks phenomenon. In this paper, spectral ensemble clustering plays a key role, but it incurs significant computational costs. To tackle this issue, diverse approaches have been examined. Above all, techniques such as weighted cluster ensemble [17], sparsification [16], and landmarking [15] have been verified. For instance, a divide-and-conquer method is introduced by Hongmin Li et al. They accelerated the base clustering generation process by applying K-nearest neighbors [18]. Comparably, Dong Huang et al. used an approximation technique for K-nearest selection to build a sub-matrix of sparse affinity. This affinity sub-matrix efficaciously relates objects and cluster representatives with lower computational overhead [12]. In spectral ensemble clustering, it is vitally important to consider how the proposed method can reduce the drawbacks of dimensionality.

III. NOVELTY IN SPECTRAL ENSEMBLE CLUSTERING

A. BASE CLUSTERING GENERATION

Different clustering techniques are gathered to evaluate the true efficiency of the proposed ensemble clustering method, each with its distinct philosophy. A non-exhaustive list of these clustering methods is displayed in Table 2. Since this paper focuses mainly on the design of the consensus core [11], [12], the base clustering methods are not deeply studied. The base clustering data in our proposed method is built by results from ten base techniques shaping the required "cross-affinity matrix."

B. NOVELTY IN CONSENSUS CLUSTERING

Although the reasons mentioned above underscore the capability of spectral ensemble clustering, one question still needs to be addressed, particularly for PV FDD using I-V curve data: How can we improve the consensus function for cluster aggregation? The spectral ensemble clustering inputs are the matrix of labels generated by all base clustering routines. The main steps in our proposed spectral ensemble clustering involve computing the cross-affinity matrix, shaping the bipartite graph, performing dimensionality reduction, and solving the new eigenproblem. In this paper, we offer new outlier detection and outlier removal after constructing the bipartite graph based on the centralities of vertices in the graph. Adding this outlier removal has several advantages, including higher robustness and suitability for systems with ambiguities in their inherent distribution or fault behavior. Consequently, the proposal is named "Enhanced Spectral Ensemble Clustering (ESEC)," which has no constraints for cluster shapes like other graph-based clustering methods [27]. We explain the novel spectral ensemble clustering step by step, As shown in the flowchart in Figure 2.

1) CONSTRUCTING BIPARTITE GRAPH

In the following, we assume that there are N base clusterings. Each generates a partition denoted ψ^i, where i = 1, …, N. The union of all N partitions forms, Ψ defined as

$$\Psi = \left\{\psi^1, \psi^2, \ldots, \psi^N\right\} \tag{1}$$

The collection of all cluster subsets generated by different base clusterings is denoted as Θ.

$$\Theta = \bigcup_{i=1}^{N} \left(\bigcup_{j=1}^{P_i} \theta_{ij} \right)$$

where θ_{ij} represents the jth cluster within partition $\psi^i$and P_i (for $i = 1, \ldots, N$) denotes the number of clusters within each base clustering. Note that the value P_i, may differ for each clustering. The total number of clusters in the ensemble is denoted as M.

$$M = \sum_{i=1}^{N} P_i \tag{2}$$

After creating the ensemble set of clusters, the membership of each cluster to its base clustering is not considered. Instead, the membership of a cluster to the ensemble set is studied. Let's rewrite "$\theta = \{\theta_f\}_{1\times M}$" as the set of collected clusters from different base clustering methods.

$$\theta = \{\theta_1, \theta_2, \ldots, \theta_M\}$$

The bipartite graph G is defined as:

$$G = \{\boldsymbol{Z}, \theta, \boldsymbol{R}\} \tag{3}$$

Here, Zrepresents the featured dataset of m objects denoted as z_l (for $l = 1,\ldots, m$). The bipartite graph vertices consist of the m objects z_l and related clusters θ_f, and the total number of clusters is M. Total vertices of both partition sides of the bipartite graph form $|Z \cup O|_{(m+M)}$is $(m + M)$. The variable R represents the cross-affinity matrix in the bipartite graph. Each element r_{lf} of the matrix R determines the membership of object node z_l,in the cluster θ_f. If an edge connects two vertices from opposite sides of the bipartite graph, the membership equals one.

$$\boldsymbol{R} = \{r_{lf}\}_{m\times M} \tag{4}$$

where r_{lf} equals 1 if related data belongs to cluster θ_f and r_{lf}changes to zero if the above-mentioned condition is invalid. "A" represents the total affinity matrix, defined as:

$$\boldsymbol{A} = \begin{pmatrix} 0 & \boldsymbol{R} \\ \boldsymbol{R^T} & 0 \end{pmatrix} \tag{5}$$

2) EIGENVECTOR CENTRALITY

Eigenvector centrality is a key parameter for assessing the importance and impact of a vertex within a graph (or a node in a network). Noticeably, Google's PageRank algorithm shares similarities with eigenvector centrality [46]. Denoted as EC_i for node i, this centrality measure initially assumes a value of one and is calculated based on the following formula [47].

$$\mathrm{EC_i} = \frac{1}{\lambda_{max}\,(G)} \sum_{j=1}^{q} \hat{\mathrm{a}}_{ij}.\mathrm{EC_j} \tag{6}$$

where $\lambda_{max}\,(G)$ represents the maximum eigenvalue of the graph G. $\hat{\mathbf{A}}$ is the adjacency matrix and $\hat{\mathrm{a}}_{ij}$ are adjacency matrix elements connected to the ith node. It is determined by summing EC_jdenoted as the eigenvector centrality of all nodes that are directly connected to node "i"via edges within its neighborhood q. Eigenvector centrality considers two informative indicators: the number of nodes connected to node i and the importance of the given node in terms of the information stream by looking at other nodes [48]. In this section, we will examine eigenvector centrality in the context of spectral ensemble clustering. In a bipartite graph, high eigenvector centrality has different implications. High eigenvector centrality in object nodes suggests membership in multiple clusters, showing overlap between clusters. On the other hand, in cluster nodes, high eigenvector centrality reveals large clusters covering many object nodes.

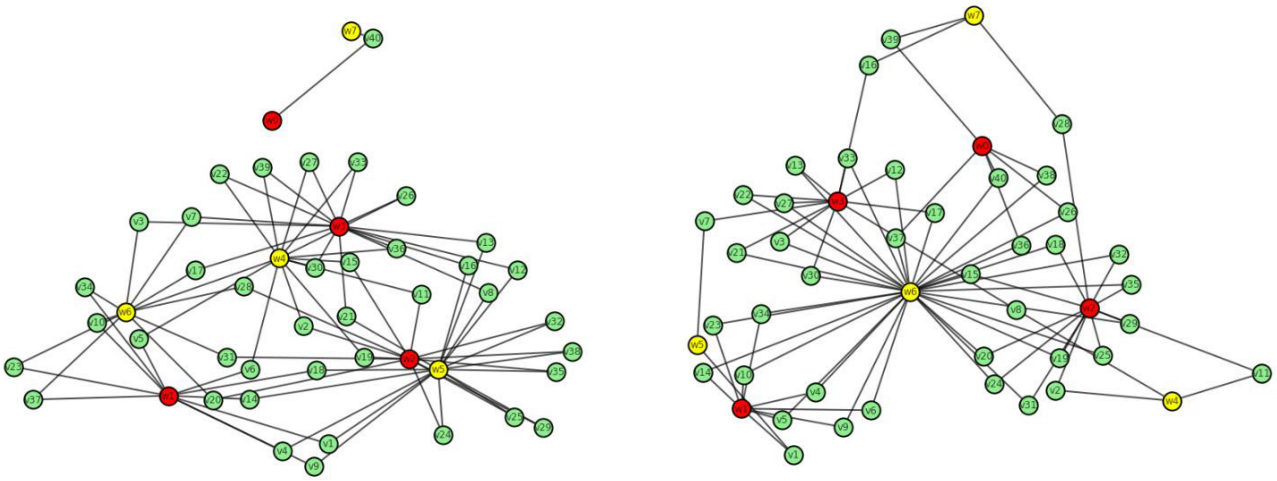

FIGURE 1. Illustrative examples of outliers. The green vertices represent data nodes, while the red and yellow vertices represent base clustering nodes. The left graph shows three outlier nodes with low eigenvector centrality. The right graph demonstrates incorrect high eigenvector centrality on a single cluster (w6).

Notably, major differences in eigenvector centrality among cluster nodes reflect unbalanced clustering. Very high and very low eigenvector centralities in object and cluster nodes, as shown in Figure 1, identify outliers. In summary, the use of eigenvector centrality provides three key indicators.

- Clustering overlap indication
- Unbalanced clustering detection
- Outlier detection

Low eigenvector centrality (EC) values often indicate outliers, while high EC values, as formulated in equation (7), may suggest overlapping or unbalanced clusters and, in some cases, outliers. This work primarily focuses on outlier detection, but it also briefly addresses the detection of overlapping and unbalanced clusters.

$$\max\left(\mathrm{EC_i}\right) \gg \frac{1}{n_v}\sum_{i=1}^{n_v}\mathrm{EC_i} \tag{7}$$

In equation (7), the variable "n_v" represents the number of vertices in the bipartite graph. It is important to note that base clustering methods have no overlapping occurrences inside, as each data point is assigned to a unique cluster within a specific base clustering. Obviously, a data object cannot belong to two clusters within the same base clustering method at the same time. In spectral ensemble clustering, overlapping occurs only between clusters from different base clusterings. It is possible to have scenarios where all basic clustering methods produce the same clustering, resulting in completely overlapped clusters. Conversely, there are scenarios where all of the clusters, from different base clustering methods, have only partial overlapping. However, these scenarios are empirically infrequent. Clusters with high eigenvector centrality in their vertices often indicate the presence of clusters with many members. Although it is common for certain clusters to encompass larger partitions, the crucial consideration is determining whether these larger clusters are disproportionately sized or not. The unbalanced cluster characteristics depend on the underlying data distribution. In datasets marked by high standard deviation, distinguishing between normal clusters and oversized unbalanced ones causes a challenge, while for datasets with lower standard deviation, the identification of misclassified enormous clusters is more straightforward. The main task is to interpret the unbalanced clustering and the reasons behind it. Some systems have an asymmetric nature and naturally cause unbalanced clusters, while in others, unbalanced clustering may imply misguided clustering. Such erroneous imbalance is studied more in the outlier section. The pure unbalanced clustering analysis is out of the scope of this research, and this paper touches on some parts relevant to the outlier detection.

3) OUTLIERS

This section analyzes the bipartite graph generated by spectral ensemble clustering. The main focus is on consensus clustering, which comes after the generation of clusters from the "N" base clustering partitions. This article tries to introduce a novel approach for dimensionality reduction based on the eigenvector centrality of the bipartite graph. In current state-of-the-art methods, reducing data to a smaller size by removing members associated with small clusters may lead to misinterpretation. For instance, a member of a small cluster in one base clustering technique could be part of larger clusters in others. Due to this reason and many other benefits, which we will explain in upcoming sections, this article uses eigenvector centrality (EC) to identify and remove outliers

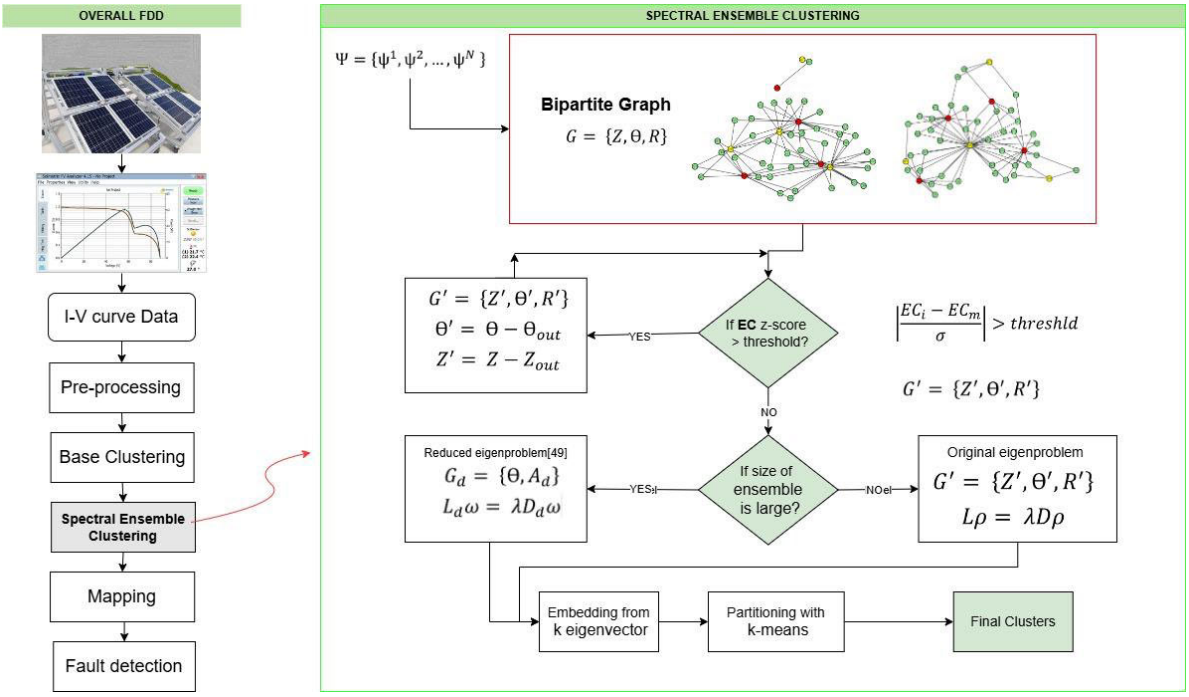

FIGURE 2. Flowchart of the proposed ESEC method.

and reduce dimensions in the bipartite graph. This helps extract a more informative subset of data, thereby improving subsequent consensus clustering. This enhancement is based on the recognition that very low EC values not only indicate a lack of connectivity to large clusters but also to nodes with high centrality. Conversely, very high EC values may result from erroneous clusters that indiscriminately cover most of the data.

a: THRESHOLD

We use the z-score method on EC values to detect outliers. First, we use logarithmic transformation to reduce the skewness and compress the wide range of values into a more manageable scale. After applying the logarithmic transformation, we use the z-score method to detect outliers in the transformed data. The z-score identifies outliers based on the number of standard deviations from the mean. This paper sets the z-score threshold at ''three'' based on empirical reasons. Once we have the indices of the outliers in the transformed data, we can map them back to the original data to identify the actual outliers. In general, both very low and very high eigenvector centrality values that their z-score are substantially low or high (here if the absolute value of the z-score is greater than three) are listed as outliers. For each vertex, v_i in graph G where $v_i \in$V(G), the mean value of eigenvector centralities EC_m, which is equivalent to the right side of (7), and the criteria for identifying outliers are determined as follows:

$$\mathrm{EC}_m = \frac{1}{(m+M)\,.\lambda_{max}\,(G)} \sum_{i=1}^{m+M} \sum_{j=1}^{q} \hat{\mathrm{a}}_{ij}.\mathrm{EC}_j \qquad (8)$$

$$\left|\frac{\mathrm{EC_i} - \mathrm{EC_m}}{\sigma}\right| > 3.0 \qquad (9)$$

The standard deviation (σ) of EC values is needed to calculate the z-score as formulated on the left side of the inequality (9). In Figure 1, there are two typical small examples of bipartite graphs from different base clusterings. Green vertices represent data nodes, while red and yellow represent two base clustering nodes. The edges indicate which data nodes belong to which base clustering nodes. In the left graph, three nodes are separated from the rest, which can be categorized as outliers due to their low eigenvector centrality. On the right graph, there is a scenario where a base clustering wrongly assigns almost all data nodes to a single cluster (w6). This results in a lack of distinction between the data nodes. The yellow cluster node (w6) shows high eigenvector centrality. However, this high centrality is misleading as it fails to differentiate between the data nodes.

4) OUTLIER REMOVAL

Let G be a bipartite graph with its vertex set V(G) and its edge set E(G). Assume H is an outlier subgraph of $G\,(V\,(G)\,, E\,(G))$, then its vertex set is $V\,(H)$, and the edge set is $E\,(H)$:

$$\text{Subject to} \begin{cases} V\,(H) \subseteq V\,(G) \\ E\,(H) \subseteq E\,(G) \end{cases}$$

The new bipartite graph after outlier elimination, which is explained in algorithm 2, is

$$G' = \left\{Z', O', R'\right\} \qquad (10)$$

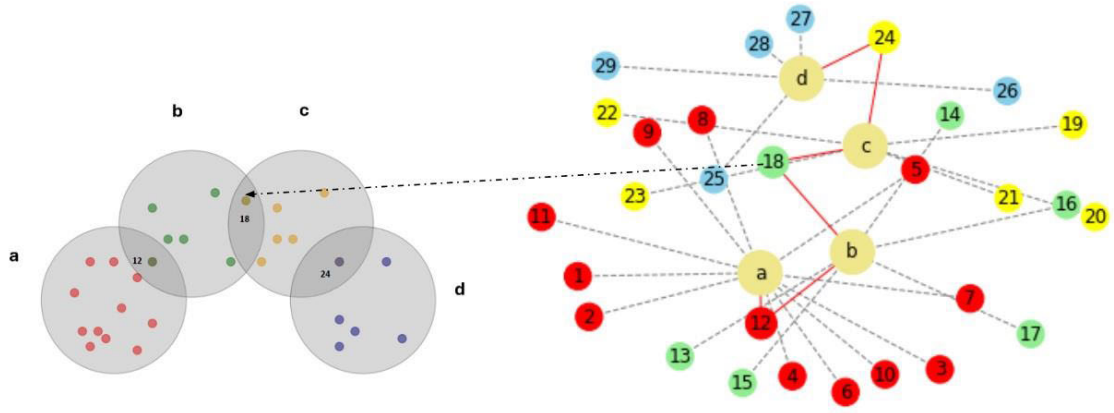

FIGURE 3. An illustrative example of a connected bipartite graph is discussed in theorem 3. Overlapping clusters (left) indicate common nodes. The common nodes in the overlapping regions make the bipartite graph connected (right).

where $G' = (V', E')$ is the bipartite graph after outlier removal and vertex set V' is representad as follows.

$$V' = V(G)\, V(H) \tag{11}$$

5) IMPACT OF OUTLIER REMOVAL ON COMPUTATIONAL COMPLEXITY

It is evident that the removal of outliers leads to a reduction in both time and memory costs. In the context of the eigenproblem, the cross-affinity matrix (denoted as R) transitions from an $m \times M$ matrix to a lower-dimensional cross-affinity matrix of dimensions $(m - m_{out}) \times (M - c_{out})$, where m_{out} and c_{out} represent the number of outlier elements and outlier clusters, respectively. Given that the number of non-zero elements in each row of the cross-affinity matrix equals N (where N represents the number of base clustering methods), the computational complexity for shaping such a matrix shifts from $O(m.N)$ to $O((m - m_{out}).N)$ [49]. We examine the time cost before and after outlier removal in result section.

6) OUTLIER REMOVAL ON SPECTRAL ENSEMBLE CLUSTERING

In general, spectral clustering is sensitive to outliers, thereby emphasizing the need for precise measures. Outliers tend to yield eigenvalues that are proximal to "one," along with eigenvectors that substantially differ from others involved in the eigen-decomposition and clustering process. Hence, it is optimal to remove outliers before the eigen-decomposition stage [50].

7) IMPACT OF OUTLIER REMOVAL ON DISCRETIZATION STEP (K-MEANS ALGORITHM)

Applying outlier removal to the bipartite graph G induces changes in the cross-affinity matrix R and, consequently, the affinity matrix A by reducing the number of rows and columns. This poses a new challenge in assessing the influence of outlier removal on the performance of the discretizing step (K-means), which is used as a discretization tool within the last step of spectral ensemble clustering. To investigate the impact of outlier removal, we endeavor to establish the following theorem, which explains the efficiencies before and after outlier removal:

Theorem 1: Let the objective function in the k-means algorithm be:

$$\mathcal{F}\left(\hat{\theta}\right) = \sum_{j=1}^{k} \sum_{x \in \hat{\theta}_j} d\left(c_j, x\right) \tag{12}$$

where $d\left(c_j, x\right)$is the distance between datapoints and cluster centers. Applying outlier removal after base clustering generation improves the K-means discretization efficiency (it decreases the K-means objective function, $\mathcal{F}\left(\hat{\theta}\right)$, inside bipartite graph partitioning).

Proof: See the Appendix A.

8) UPPER AND LOWER BOUNDS OF EIGENVECTOR CENTRALITY

Theorem 2: Let G be a bipartite graph between data points and cluster points from different base clusterings in the spectral ensemble clustering algorithm. Upper and lower bounds for the eigenvector centrality (EC_i) of any node in G is represented as.

$$\frac{1}{\sqrt{m.N}} \sum_{j=1}^{q} \hat{a}_{ij}.\mathrm{EC}_j < \mathrm{EC_i} < \frac{1}{\lambda_{max}(G)} \sum_{j=1}^{m} \hat{a}_{ij}.\mathrm{EC_j} \tag{13}$$

where m is the number of data rows, N is the number of base clusterings, and q is the number of surrounding nodes neighboring each vertex of G.

Proof: See Appendix B.

9) EIGENVECTOR CENTRALITY IN BIG DATA

In this paper, the eigenvector centrality in big data is studied only from a purely theoretical perspective. The clusters generated by base clustering overlap more in ultra-large datasets, particularly with an increasing number of ensembles. So, it is very important to study the large clusters with overlapping members. Such overlapping members may make the bipartite graph a connected graph. Graph connectivity is a key element in the analysis of bipartite graphs within spectral ensemble clustering. Figure 3 illustrates a typical brief example of how the overlapping ensemble clusters impact the connectivity of a bipartite graph. Nodes in the shared area create connecting edges in the bipartite graph. In our proposed method for outlier removal based on eigenvector centrality, this article studies the influence of graph connectivity on eigenvector centrality and its lower bound in Theorem 4. Before that, we must describe bipartite graph connectivity in spectral ensemble clustering.

Theorem 3: Let $G = \{Z, \Theta, A\}$ be a bipartite graph constructed from N base clustering with total M clusters. $\Theta = \{\Theta_1, \Theta_2, \ldots, \Theta_M\}$,and let the partition of whole Θ into two nonempty subsets be Θ_v and Θ_u, that are constructed from any combination of clusters inside Θset while $\Theta = \Theta_v \cup \Theta_u$. If for all possible partitions of G into Θ_v and Θ_u, it holds:

$$|\Theta_{\mathrm{v}} \cap \Theta_{\mathrm{u}}| \geq 1 \tag{14}$$

Then, G is connected.

Proof: See Appendix C.

If we take into account the overlapping behavior in ensemble clustering in big data, as well as the graph connectivity explained in theorem 3, we can now establish a new lower bound for eigenvector centrality.

Theorem 4: Let G be a bipartite graph with conditions in theorem 3. The lower bound of its eigenvector centrality is:

$$\frac{1}{\sqrt{m\,(2N-1)-M+1}} \sum_{j=1}^{q} \hat{\mathrm{a}}_{ij}.\mathrm{EC_j}{<}\mathrm{EC_i} \tag{15}$$

M is the number of all clusters from different base clusterings.

Proof: See Appendix D.

10) CUTTING THE EIGENVECTORS

The generalized eigenproblem for the aforementioned bipartite graph $G = \{\boldsymbol{Z}, \theta, \mathbf{R}\}$ is represented as:

$$\boldsymbol{L\rho} = \boldsymbol{\gamma D\rho} \tag{16}$$

where L denotes the Laplacian matrix ($\boldsymbol{L} = \boldsymbol{D} - \boldsymbol{A}$), " $\boldsymbol{A}= \boldsymbol{R^T D^{-1} R}$" is the full affinity matrix, and $\boldsymbol{D}$ is the degree matrix. Let the corresponding first k eigenpairs of G be denoted as "$\{(\gamma_i, \rho_i)\}_{i=1}^{k}$". Li et al. [49] demonstrated the equivalence between the eigenproblem of the bipartite graph G and the subsequent eigenproblem of the reduced-size graph "$G_d = \{O, \boldsymbol{A_d}\}$." The affinity matrix of the reduced eigenproblem is $\boldsymbol{A_d} = \boldsymbol{R^T D_z^{-1} R}$. Where D_z is a diagonal matrix ($\boldsymbol{D_z} \in \mathbb{R}^{m\times m}$) whose entries along the main diagonal (i, i) are equal to the sum of the i$^{\mathrm{th}}$ rows of $\boldsymbol{R}$. Here, G_d comprises graph nodes solely from cluster vertices. Its eigenproblem is expressed as:

$$\boldsymbol{L_d\omega} = \boldsymbol{\lambda D_d\omega} \tag{17}$$

where $\boldsymbol{D_d} \in \mathbb{R}^{M\times M}$ is the degree matrix of G_d and $\boldsymbol{L_d} = \boldsymbol{D_d} - \boldsymbol{A_d}$ denotes the Laplacian matrix of the reduced eigenproblem. The first k eigenpairs "$\{(\lambda_i, \omega_i)\}_{i=1}^{k}$"are also determined for the eigenproblem of G_d. Thus, instead of computing k eigenvectors of G (denoted as ρ_1, ρ_2,…, ρ_k), k eigenvectors of G_d (denoted as ω_1, ω_2,…, ω_k) are computed. Finally, the eigenvectors of G, ρ_i, are obtained by the following equations:

$$\gamma_i = (2 - \gamma_i)\,\lambda_i \tag{18}$$

$$g_i = \frac{1}{1-\gamma_i}\hat{\boldsymbol{P}}\omega_i \tag{19}$$

$$\rho_i = \begin{bmatrix} g_i \\ \omega_i \end{bmatrix} \tag{20}$$

where $\hat{\boldsymbol{P}}$ is the matrix of transition probability [49].

$$\hat{\boldsymbol{P}} = \boldsymbol{D_z^{-1} R} \tag{21}$$

Algorithm 1 : Enhanced Spectral Ensemble Clustering (ESEC)

Input: *Dataset of features (V_{ocn}, I_{scn}, V_{mpn}, I_{mpn})*
Output: *Ensemble clustering*

1. *Generate various base clustering.*
2. *Compute the cross-affinity matrix via (4)*
3. *Shape the bipartite graph.*
4. *Outlier detection using Algorithm 2*
5. *Solve the new eigenproblem via (17)*
6. *Stack the eigenvectors in a new matrix.*
7. *Apply K-means for final clustering.*

11) MAPPING AND FAULT LABELING

Merely obtaining cluster labels, as explained in Algorithm 1, does not suffice for FDD; it is imperative to interpret the cluster centroids. Following the final clustering phase, all cluster representatives should undergo interpretation and categorization into distinct fault groups [51]. Clusters are categorized into main PV fault categories, such as partial shading, open circuit fault, and aging, based on the alignment of each cluster center with fault electrical characteristics, as illustrated in Figure 4. Various measurements from I-V curve characteristics, such as Vocn, Iscn, $\mathrm{V_{mpn}}$, and $\mathrm{I_{mpn}}$, are relevant fault signatures. Potential faults can be deduced from the abnormal variations in electrical features. Significantly, a decline in currents ($\mathrm{I_{scn}}$ and $\mathrm{I_{mpn}}$) is a probable sign of open circuit faults. Partial shading indicates a decrease in $\mathrm{V_{mpn}}$ and $\mathrm{I_{mpn}}$. In more serious shading cases, it manifests a more rapid reduction in $\mathrm{I_{mpn}}$.Aging faults (R_s degradation) display drops in $\mathrm{V_{mpn}}$ and $\mathrm{I_{mpn}}$, with a more notable decline in $\mathrm{V_{mpn}}$ [53]. Subsequently, a fault label is assigned to each row of the

data according to the inferred fault or health status of cluster representatives,

Algorithm 2: Dimensionality Reduction

Input: $\boldsymbol{G} = \{\boldsymbol{Z}, \Theta, \boldsymbol{A}\}, \boldsymbol{\Theta} = \{\theta_1, \theta_2, \ldots, \theta_M\}$,
$Z = \{z_1, z_2, \ldots, z_m\}$, threshold,
Output: $\boldsymbol{G'}$ (Reduced bipartite graph)
Compute $\boldsymbol{EC_i}$ ***for*** $i = 1$ *to* $(m + M)$ *by equation* (6)
Compute the mean value of $\boldsymbol{EC_m}$ *by equation* (8)
for $i = 1$*to* $(m + M)$ ***do***
if $\left|\frac{EC_i - EC_m}{\sigma}\right| >$ *threshold* ***then***
Update Z_{out} ***if*** $1 \leq i \leq m$
Update Θ_{out} ***if*** $(m + 1) \leq i \leq M$
end if
end for
Obtain final (Θ_{out}, Z_{out})
Compute Θ' *by* $\Theta' = \Theta - \Theta_{out}$
Compute Z' *by* $Z' = Z - Z_{out}$
Update $\boldsymbol{R'}$ *via* (4) for its new dimension
Obtain $\boldsymbol{G'}$, *via* (10),
Return $\boldsymbol{G'}$

IV. EXPERIMENT AND RESULTS

The I-V curve data measured from PV systems are the core data for the diagnostic model presented herein. The PV systems can be effectively and comprehensively described and modeled by the "Single Diode Model" [50].

A. EXPERIMENTAL SETUP AND DATASETS

Among the various electrical features of PV systems, four primary attributes are selected: short-circuit current (I_{sc}), open-circuit voltage (V_{oc}), maximum power point current (I_{mpp}), and maximum power point voltage (V_{mpp}). The range of environmental parameters is confined to typical minimum and maximum summer day irradiance and temperature within the PV test platform located south of Paris. The irradiance spans from 200 W/m^2 and 16°C in the morning to 1180 W/m^2 and 46°C in the early afternoon. Various fault scenarios are considered for fault generation, including open circuit faults, partial shading, and aging, as these represent the most common faults.

1) FEATURES FROM I-V CURVES

To remove the environmental effects from I-V measurements, the data of the I-V curve are converted to the standard irradiance and temperature at STC (1000 W/m^2 and 25°C). In addition, to facilitate the comparison of clustering results, a normalization of the I-V curve is needed. The results will shape the per unit values in the [0, 1] range. Then, the features change to I_{scn}, V_{ocn}, I_{mpn}, and V_{mpn}. The preprocessing concepts in this stage are formulated based on the correction formula (22) and (23) proposed by Li Baojie et al. [53].

$$I_2 = I_1 \left(1 + \alpha_{rel} \left(T_{m2} - T_{m1}\right) \frac{G_2}{G_1}\right) \tag{22}$$

$$\begin{aligned} V_2 = V_1 + V_{oc1} \left(\beta_{rel} \left(T_{m2} - T_{m1}\right) + a.ln\left(\frac{G_2}{G_1}\right)\right) \\ - R_s \left(I_2 - I_1\right) \\ - k.I_2 \left(T_{m2} - T_{m1}\right) \end{aligned} \tag{23}$$

where α_{rel} and β_{rel} are the relative coefficients of temperature correction for I_{sc} and,V_{oc}. The internal series resistance is R_s. k and a are the curve and irradiance correction factors, respectively. Indices 1 and 2 represent the before and after-correction for current $I_1$and I_2, voltage V_1 and V_2, irradiance G_1, $G_2$and temperature T_{m1} and T_{m2}. The PV module is the BlueSolar Monocrystalline Panels series SPM040551200 (55W-12V Mono 545 × 668 × 25mmseries 4a), whose characteristics at STC are listed in Table 3 [54].

TABLE 3. Electrical data of PV module under STC [54].

Parameters	Values
Short circuit current	3.22 A
Open circuit voltage	22.9 V
Nominal Power	55 W
Maximum Power Voltage	18.8 V
Maximum Power Current	2.94 A

2) DATA GATHERING

Various fault scenarios are implemented in a PV system consisting of 8 PV panels. Four PV panels are lined in series, and then this line is connected in parallel with another similar four PV panels. Shading faults are induced at two distinct partial shading levels by partially covering the PV panels. For open circuit faults, one of the parallel lines is deliberately disconnected. Aging faults are intentionally introduced by incorporating resistors with varying severity levels.

Ambiguous and moderate aging fault scenarios are purposely crafted to assess the diagnostic model's efficacy.

3) DATASETS

Diverse datasets encompassing different fault scenarios are integrated into the final dataset, comprising 860 rows of electrical features such as I_{sc}, V_{oc}, I_{mpp}, and $V_{mp.p}$. In fact, each row of these features is extracted from a unique I-V curve, while each I-V characteristic curve is a set of 100 data points for voltage, current, and power. Base clustering techniques encompass K-means, Agglomerative clustering, Spectral clustering, FCM, Mean-shift, OPTICS, DBSCAN, Affinity propagation, Gaussian mixture model, and Birch. Each method, as listed in Table 4, is applied, resulting in 60 base clustering outcomes in separate datasets. Subsequently, the results from different clustering approaches are concatenated after the initial clustering step for further processing through spectral ensemble methods. Figure 2 conveys the flowchart of the introduced methodology.

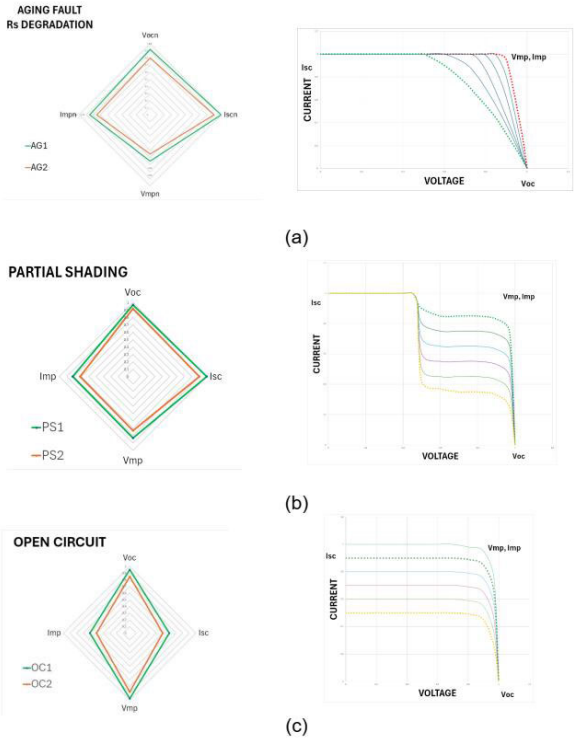

FIGURE 4. Following the clustering process, the faults are, mapped and labeled according to their normalized value of electrical features (Iscn, Vocn, Impn, Vmpn). The faults are (a) aging (Rs degradation), (b) partial shading, and (c) open circuit faults.

TABLE 4. Different base clustering datasets.

Clustering method	Max. number of datasets in ensemble clustering	Data rows in each dataset
K-means Agglomerative Spectral DBSCAN Mean shift OPTICS Birch Affinity propagation Gaussian Mixture FCM	60 (10 methods, and each method has six datasets)	871

4) EVALUATION METRICS

The accuracy of prediction is computed as:

$$\mathrm{ACC} = \frac{\mathrm{TP} + \mathrm{TN}}{\mathrm{TP} + \mathrm{TN} + \mathrm{FP} + \mathrm{FN}} \tag{24}$$

TN (true negative) and TP (true positive) are regarded as the correct predictions, while FP (false positive) and FN (false negative) stand for wrong predictions [51]. In addition, NMI is used for the evaluation of clustering performance [52]. Two partitions ψ, ψ' have NMI as follows:

$$\mathrm{NMI}\left(\psi, \psi'\right) = \frac{\sum_{i=1}^{\pi}\sum_{j=1}^{\pi'} b_{ij} \log\left(b_{ij}.b/b_i.b_j\right)}{\sqrt{\left(\sum_{i=1}^{\pi} b_i \log\left(b_i/b\right)\right)\left(\sum_{j=1}^{\pi'} b_j \log\left(b_j/b\right)\right)}} \tag{25}$$

where $\boldsymbol{b_j}$ refers to the number of elements in the j$^{\text{th}}$ cluster of available ground truth dataset, the $\boldsymbol{b_i}$ represents the number of elements in the i$^{\text{th}}$ cluster of the proposed clustering method and, $\boldsymbol{b_{ij}}$ means the common objects in $\boldsymbol{b_i}$ and $\boldsymbol{b_j}$ ($\forall \boldsymbol{i} = \{\boldsymbol{1}, \ldots, \boldsymbol{\pi}\}$, and $\forall \boldsymbol{j} = \{1, \ldots, \boldsymbol{\pi}'\}$). In this study, the results of spectral ensemble clustering are also compared with the cases when only one clustering method is used, like K-means, agglomerative clustering, and FCM.

B. RESULTS AND DISCUSSION

1) CLUSTERING RESULTS

The K-means clustering analysis was chosen for comparison due to its superior performance compared to other base clustering methods. The outcome of K-means is illustrated in Figure 5. Despite setting the number of clusters to 8, the algorithm only identifies three central regions. It does not effectively differentiate between healthy areas and aging faults, nor does it accurately discern various levels of partial shading. In Figure 6, the ESEC clustering outcome is depicted. It effectively differentiated between healthy areas and areas with ambiguous aging faults, and it also exposed regions with partial shading faults at both high and low levels. It's worth noting that there are several new preprocessing methods for I-V curves to improve partial shading diagnosis using step detection [72]. However, these methods were not used or discussed in this study, as the primary focus was applying spectral ensemble clustering with outlier removal to PV FDD. The visualization of clusterings in Figures 5 and 6 offers only an initial conceptual representation of ESEC performance.

However, the system's faultiness or health can be discerned by the subsequent pivotal mapping step.

2) PERFORMANCE

The performance of the enhanced spectral ensemble clustering method, ESEC, introduced in this study was evaluated against two established spectral ensemble clustering methods, named U-SENC [12] and LSEC [11]. Tables 5, 6, and Figure 10 visually compare the performance of three ensemble spectral clustering methods, U-SENC, LSEC, and our proposed method, ESEC, using a PV dataset from a real test platform. They are all spectral ensemble clustering techniques that differ primarily in their underlying algorithms. The methods differ in their algorithms, with ESEC utilizing eigenvector centrality to identify outliers. U-SENC and LSEC have distinct ensemble generation algorithms but similar consensus clustering processes. Each method was tested with ensemble sizes (n) of 20, 40, and 60. Table 5 shows accuracy rates for detecting partial shading, open circuit, healthy

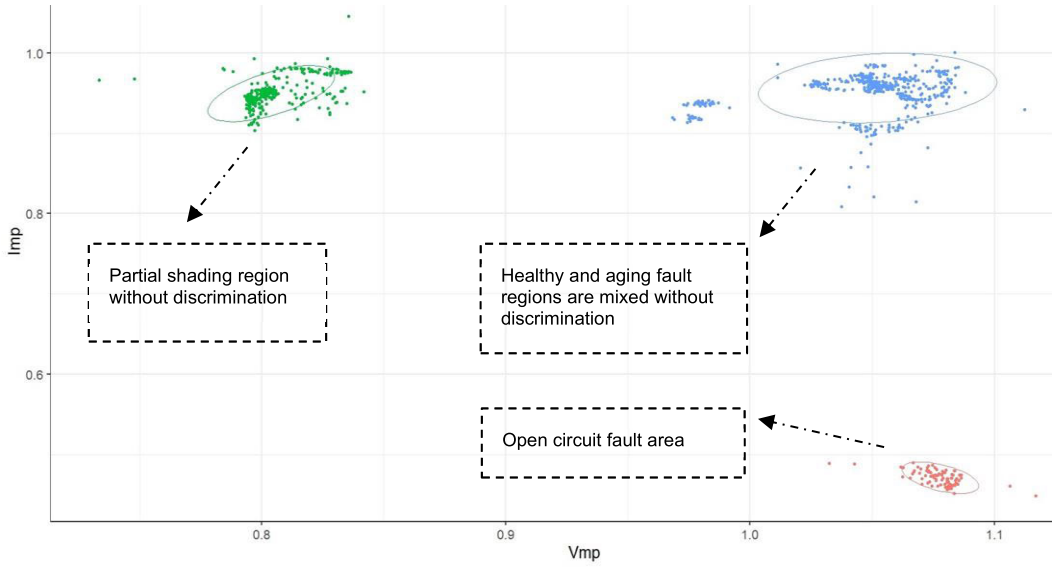

FIGURE 5. K-means clustering results. Despite the setting of 8 clusters, it can detect only three major regions without proper discrimination.

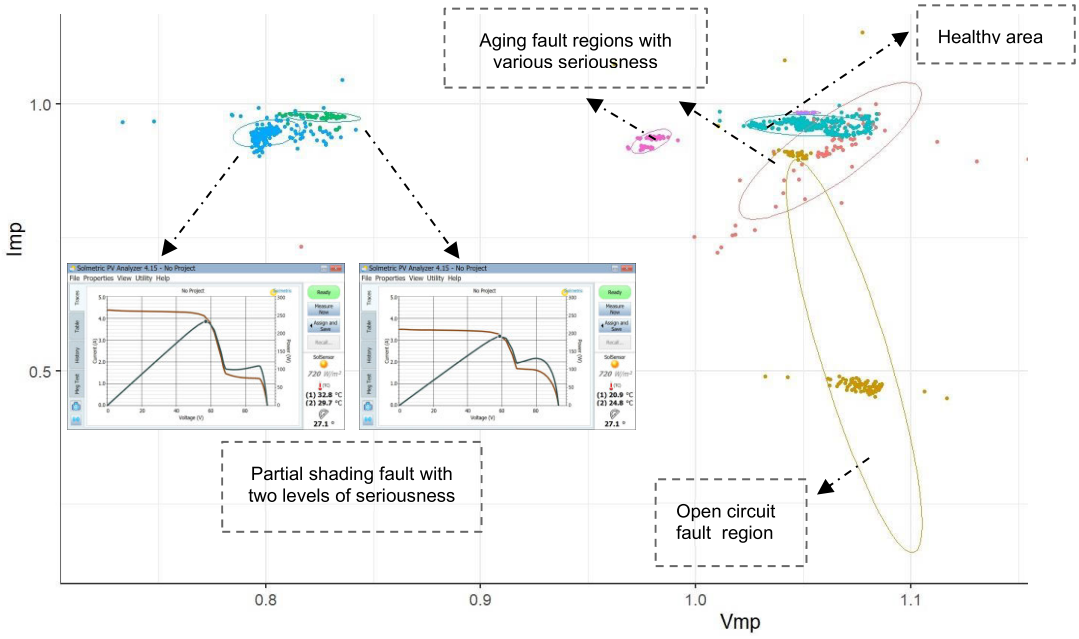

FIGURE 6. ESEC clustering results. It illustrates distinct fault areas and better separation ability, specifically for the aging fault, which is intentionally designed with ambiguity.

status, and aging faults. Partial shading and open circuit faults are easier to detect. The distinct electrical signatures of these faults facilitated their detection, resulting in comparable performance across all spectral ensemble methods. But, aging fault detection due to its ambiguous characteristics is more challenging. ESEC showed its effectiveness in the detection of elusive anomalies characteristic of aging, signifying its higher performance for ensemble sizes of 40 and 60. For n=20, the ESEC remained competitive even though the accuracy was a little lower. Again, the importance of robust methods like ESEC can be underscored by the challenging nature of aging fault detection. Table 6 presents RMSE values, reinforcing the findings from Table 5. ESEC demonstrates superior performance in aging fault detection at larger ensemble sizes, while all of the listed methods perform well in detecting partial shading and open circuit faults. Nonetheless, distinguishing between a healthy status and an aging fault poses a challenge due to intentionally complex and

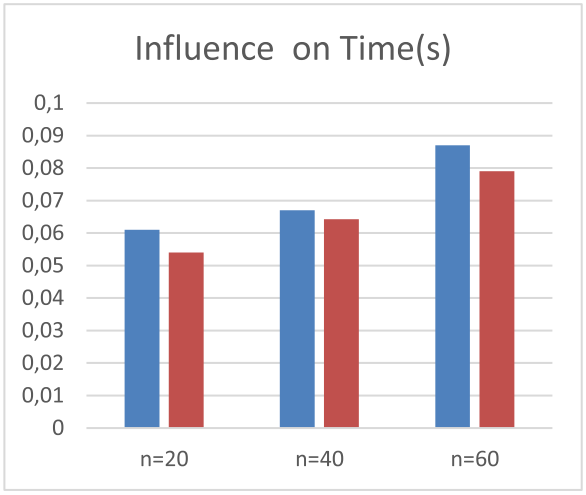

FIGURE 7. Computation time for bipartite graph partitioning with (red) and without (blue) outlier removal.

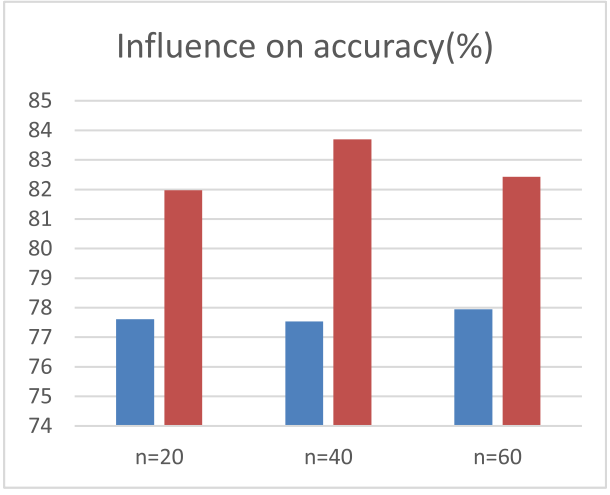

FIGURE 8. Accuracy of bipartite graph partitioning with (red) and without (blue) outlier removal.

ambiguous data distribution in that area. Most state-of-the-art methods find it difficult to differentiate between the two because of the weak impact of aging faults. ESEC methods consistently achieve higher accuracy for healthy status across all ensemble sizes, including 20, 40, and 60. Table 9 demonstrates that ESEC exhibits superior accuracy to all four base clustering methods in distinguishing aging faults and healthy statuses.

3) INFLUENCE OF OUTLIER REMOVAL ON COMPUTATION TIME

Figure 7 demonstrates the impact of outlier removal on the computation time required for bipartite graph partitioning. While the inclusion of an outlier removal step significantly reduces the computation time for the graph partitioning process itself, it is important to note that calculating eigenvector centrality to identify outliers also requires additional time. Consequently, the overall computation time for the entire process is not the fastest but remains competitive compared to other mainstream methods. This balance highlights the efficiency of the proposed method in maintaining a competitive edge while improving specific computational steps.

4) INFLUENCE OF OUTLIER REMOVAL ON ACCURACY

Figure 8 compares cases with and without outlier removal. It illustrates the effect of outlier removal on the accuracy of fault detection in PV systems. The x-axis represents the different sizes of the ensemble (n=20, n=40, n=60), and the y-axis measures accuracy in percentage. The bars show that incorporating an outlier removal step improves accuracy by up to 5% on average compared to cases without outlier removal. This enhancement in accuracy is expected, as removing outliers helps in achieving a more precise clustering, thereby improving the overall fault detection performance.

5) INFLUENCE OF ENSEMBLE SIZE ON PERFORMANCE

Figure 9 shows the Normalized Mutual Information (NMI) scores for U-SENC, LSEC, and ESEC across different numbers of base clusters (n=10 to 60). Our proposed method (ESEC) consistently achieves higher NMI scores compared to U-SENC and LSEC, indicating superior performance in detecting PV faults. In addition, ESEC's NMI scores remain relatively stable and high across the range of base clusters, while U-SENC and LSEC show more fluctuation. This comparison highlights the robustness and superior performance of ESEC, while U-SENC and LSEC exhibit lower and more fluctuating NMI scores, suggesting they are more sensitive to the number of base clusters.

6) COMPARATIVE ANALYSIS FOR NEW CASE STUDIES

The primary dataset utilized in this study is derived from a PV test platform, resulting in a somewhat limited diversity in data distribution due to the inherent limited ranges in I-V curve values. In addition, spectral ensemble clustering has a key role in the success of our proposed FDD approach. It is deliberately chosen to analyze the clustering aspect of our overall FDD method separately. To effectively demonstrate the efficiency of the proposed method, three well-known datasets were employed, "PenDigit" [68], [71],"Letters" [68], [70], and "USPS" [69]. The number of used base clusterings is fixed at n=20 for better comparison.

As mentioned previously, this study introduced the ESEC method and compared it with two well-established spectral ensemble clustering methods: U-SENC [12] and LSEC [11]. Specifically, as presented in Table 7, the proposed method in this article (ESEC) attained the highest NMI scores of 85.81% and 48.83% on the PenDigits and Letters datasets, respectively. For the USPS dataset, the ESEC method achieved an NMI score of 73.88%, which is remarkably close to the highest NMI score of 73.92% obtained by the U-SENC method. It should be noted that base clustering methods such

TABLE 5. Accuracy (%) comparison of spectral ensemble clustering methods (PV datasets).

PV FAULT	U-SENC			LSEC			ESEC		
	n= 20	n=40	n =60	n=20	n=40	n=60	n=20	n=40	n=60
Average Aging	76.57	75.31	77.26	79.33	79.32	79.84	**81.97**	**83.69**	**82.43**
Partial Shading Fault	99.08	99.05	99.11	95.86	97.12	95.40	**99.51**	**99.31**	**99.37**
Open Circuit Fault	**99.77**	**99.70**	**99.68**	98.62	97.35	97.24	99.28	99.43	99.54
Healthy Status	76.40	75.84	71.39	79.40	79..61	79.93	**81.63**	**82.40**	**82.31**

TABLE 6. RMSE comparison of spectral ensemble clustering methods (PV datasets).

PV FAULT	U-SENC			LSEC			ESEC		
	n= 20	n=40	n =60	n=20	n=40	n=60	n=20	n=40	n=60
Average Aging	0.4839	0.4968	0.4767	0.4545	0.455	0.437	0.4558	0.430	**0.410**
Partial Shading Fault	0.0958	0.0957	0.096	0.2033	0.1694	0.2143	0.1071	**0.0830**	0.1123
Open Circuit Fault	0.0479	**0.0478**	0.048	0.1173	0.1625	0.1660	0.1016	0.0829	0.0677
Healthy Status	0.485	0.464	0.494	0.450	0.438	0.428	0.458	0.431	**0.419**

TABLE 7. NMI comparison (handwritten datasets).

DATASET	BASE CLUSTERING METHODS			SPECTRAL ENSEMBLE CLUSTERING METHODS		
	K-means	Agglomerative	FCM	U-SENC	LSEC	ESEC
PenDigits	69.13	72.81	54.25	84.47	81.69	**85.81**
Letters	36.24	39.25	21.74	45.78	46.47	**48.83**
USPS	41.51	57.53	24.23	**73.92**	67.62	73.88

TABLE 8. TIME(s) comparison (handwritten datasets).

DATASET	U-SENC	LSEC	ESEC	ESEC (base clustering generation included)
PenDigits	19.2557	**5.5602**	9.9020	15.66
Letters	21.3069	**7.7363**	26.0879	33.50
USPS	27.7585	11.04	**9.3488**	20.35

as k-means, agglomerative clustering, and FCM are included solely as benchmarks for NMI score comparison. It is evident that the clustering part of our overall method also outperformed similar spectral ensemble clustering approaches when assessed on commonly used datasets. This strongly indicates that our approach inherently holds the potential for diverse real-world scenarios characterized by complexity and ambiguity. Table 8 provides an overview of the computation time

TABLE 9. Accuracy (%) comparison with base clustering methods (PV datasets).

PV FAULT	K-means	Agglomerative	Spectral	FCM	ESEC (n=20)
Aging	68.30	63.03	63.09	61.02	**81.97**
Healthy status	67.24	61.89	61.02	60.97	**81.63**
Partial Shading Fault	**99.65**	99.54	99.19	99.51	99.51
Open Circuit Fault	99.31	**99.50**	99.01	99.46	99.28

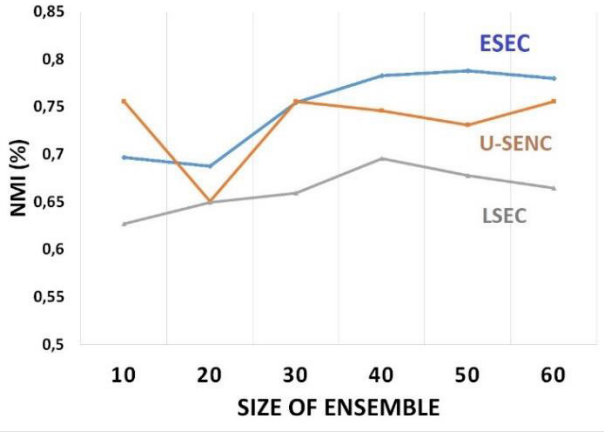

FIGURE 9. Influence of ensemble size on clustering performance.

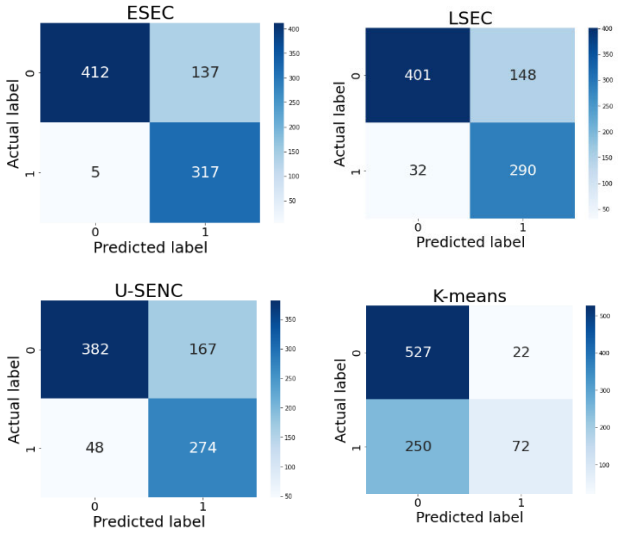

FIGURE 10. Confusion matrices of ESEC, LSEC, U-SENC (number of ensemble is 40), and K-means for Aging fault.

costs associated with the three spectral ensemble clustering methods. Given that base clustering methods generally incur lower computation costs than ensemble clustering methods, they are deliberately excluded from the computation time comparison. For the proposed ESEC method, the table includes a column detailing the time required for the consensus clustering, as well as a column for the total time encompassing both the ESEC and base clustering generation steps. Although the primary focus of this paper is on the consensus clustering step, the base clustering generation step is included for reference purposes only. In terms of computation time, while the ESEC method does not exhibit the lowest costs, its time efficiency is competitive and comparable to other methods within the field. Except for the Letters dataset, the EESC method demonstrates competitive performance on both the PenDigits and USPS datasets. The ESEC algorithm has proven its effectiveness over other spectral ensemble clustering methods, even when accounting for the computational time required to calculate eigenvector centrality.

7) PERSPECTIVE

In scenarios with reliable ground truth data, fault labels can be directly used for prediction. However, many systems may lack proper archives of their fault events. In such cases, the Cluster-Then-Label (CTL) technique generates labeled datasets used as inputs to various classifiers. Thereby, integrating ESEC with neural networks is helpful. The ESEC method can create new labeled datasets split for training and testing as inputs to the ANN for classification.

V. CONCLUSION

This study introduces a new method called ESEC (enhanced spectral ensemble clustering), which is used for fault detection and diagnostics (FDD) in photovoltaic (PV) systems. The results show that this method effectively identifies faults and performs competitively. This study is the first to use the graph's centralities to analyze outliers in spectral ensemble clustering. Eigenvector centrality can improve the information flow between nodes and also give deeper insights into the key role of nodes and the characteristics of each node's neighbors. By efficiently integrating these techniques, the underlying structure of ambiguities and pattern complexities within the data can be captured, which leads to more robust and insightful clustering outcomes. Additionally, by incorporating centrality-based graph techniques into spectral ensemble clustering, new creative avenues for upcoming research are opened. For photovoltaic (PV) systems, it would be beneficial perspective to focus on exploring

the interpretation of different fault conditions within I-V curves in order to validate the effectiveness of our proposed FDD approach.

APPENDIX A
PROOF OF THEOREM 1

Let X be a set with at least two members; the K-means clustering partitions X into k subsets that are disjoint and nonempty. $\hat{\theta}$ is the set of all clusters after K-means clustering.

$$\hat{\theta} = \left\{\hat{\theta}_1, \ldots, \hat{\theta}_k\right\} \bigcup_{j=1}^{k} \hat{\theta}_j = X$$

$$\left|\hat{\theta}_j\right| \geq 1$$

Let $c = \{c_1, \ldots, c_k\}$ represents the set of cluster centers for the partition $\hat{\theta}$. In the K-means clustering algorithm, F is the objective function [60] as:

$$\mathcal{F}\left(\hat{\theta}\right) = \sum_{j=1}^{k} \sum_{x \in \hat{\theta}_j} d\left(c_j, x\right) \tag{26}$$

where d is the distance between cluster centers and data points, the K-means algorithm tries to find the optimal partitioning by solving related optimization function

$$\underset{\hat{\theta}}{\operatorname{argmin}}\, \mathcal{F}\left(\hat{\theta}\right) \tag{27}$$

For simplicity, let $\Gamma\left(\hat{\theta}_j\right)$ be the dispersion inside each cluster as:

$$\Gamma\left(\hat{\theta}_j\right) = \sum_{x \in \hat{\theta}_j} d\left(c_j, x\right) \tag{28}$$

After applying outlier removal, X_{out} is removed as outliers from the initial set X. A new partition $\hat{\theta}' = \left\{\hat{\theta}'_1, \ldots, \hat{\theta}'_k\right\}$ is obtained. It holds

$X' = X - X_{out}, |X_{out}| \geq 1$

$$X' = \bigcup_{j=1}^{k} \hat{\theta}'_j = \bigcup_{j=1}^{k} \hat{\theta}_j - X_{out} \tag{29}$$

The new partition $\hat{\theta}'$may find a new optimal k partition with a lower $\mathcal{F}$ value or at least repeat the previous optimal partition without outlier elements. In the latter case, there is at least one cluster denoted as $\hat{\theta}'_s$ whose dispersion function changes to

$$\Gamma\left(\hat{\theta}'_s\right) = \sum_{x_i \in \hat{\theta}'_s} d\left(c_s, x_i\right) \tag{30}$$

While this cluster before outlier removal is $\hat{\theta}_s$, and its dispersion function is

$$\Gamma\left(\hat{\theta}_s\right) = \sum_{x_i \in \hat{\theta}_s} d\left(c_s, x_i\right)$$

$$\Gamma\left(\hat{\theta}_s\right) = \sum_{x_i \in \hat{\theta}'_s} d\left(c_s, x_i\right) + \sum_{x_{oi} \in X_{out}} d\left(c_s, x_{oi}\right) \tag{31}$$

Hence, there must be at least one outlier element in X_{out} denoted $x_o \in X_{out}$ as the difference between $\hat{\theta}_s$ and $\hat{\theta}'_s$ and

$$\left|\hat{\theta}_s\right| - \left|\hat{\theta}'_s\right| \geq 1$$

For such a specific cluster, we have

$$\Gamma\left(\hat{\theta}_s\right) \geq \sum_{x_i \in \hat{\theta}'_s} d\left(c_s, x_i\right) + d\left(c_s, x_o\right) \tag{32}$$

One can formulate the objective function as:

$$\mathcal{F}\left(\hat{\theta}\right) = \sum_{j=1}^{k} \sum_{x \in \hat{\theta}_j} d\left(c_j, x\right)$$

$$\mathcal{F}\left(\hat{\theta}\right) = \sum_{j=1}^{k-1} \Gamma\left(\hat{\theta}_j\right) + \Gamma\left(\hat{\theta}_s\right)$$

Applying (32), following inequality is obtained as:

$$\mathcal{F}\left(\hat{\theta}\right) \geq \sum_{j=1}^{k-1} \Gamma\left(\hat{\theta}_j\right) + \sum_{x_i \in \hat{\theta}'_s} d\left(c_s, x_i\right) + d\left(c_s, x_o\right) \tag{33}$$

The right side of (33) includes $\mathcal{F}\left(\hat{\theta}'\right)$, which is

$$\mathcal{F}\left(\hat{\theta}'\right) = \sum_{j=1}^{k-1} \Gamma\left(\hat{\theta}_j\right) + \sum_{x_i \in \theta'_s} d\left(c_s, x_i\right)$$

The right side of (33) can be rewritten as:

$$\mathcal{F}\left(\hat{\theta}\right) \geq \mathcal{F}\left(\hat{\theta}'\right) + d\left(c_s, x_o\right) \tag{34}$$

Considering that$d\left(c_s, x_o\right) > 0$, therefore

$$\mathcal{F}\left(\hat{\theta}\right) > \mathcal{F}\left(\hat{\theta}'\right) \tag{35}$$

□

APPENDIX B
PROOF OF THEOREM 2

There are two partitions in bipartite graph, here the vertices of data side and of cluster side. On the data side, there are equal edges connecting each vertex and this makes the variation in data side partition limited. By emperical reasons, we expect to find the largest EC in cluster side, specifically for the large size clusters.

UPPER BOUND:

The highest assumed degree of cluster nodes is for the case when all data are members of one specific cluster in a base clustering method. In this case, we have

$$\mathrm{EC_H} = \frac{1}{\lambda} \sum_{j=1}^{m} \hat{\mathrm{a}}_{ij}.\mathrm{EC_j} \tag{36}$$

But since there is no isolated vertex in a bipartite graph, each cluster should have at least one member, so the possible largest eigenvector centrality is as follows:

$$\mathrm{EC}_{H'} = \frac{1}{\lambda} \sum_{j=1}^{q'} \hat{\mathrm{a}}_{ij}.\mathrm{EC}_j \tag{37}$$

where $q' \leq (m - P_n + 1)$ and P_n is the number of clusters in related base clustering.

then, it holds,

$$\begin{aligned}(m - q') &\geq (P_n - 1)\\ \mathrm{EC_H} &= \frac{1}{\lambda}\sum_{j=1}^{m} \hat{a}_{ij}.\mathrm{EC_j}\\ &= \frac{1}{\lambda}\sum_{j=1}^{q'} \hat{a}_{ij}.\mathrm{EC_j}\\ &\quad + \frac{1}{\lambda}\sum_{j=1}^{(m-q')} \hat{a}_{ij}.\mathrm{EC}_j \end{aligned} \tag{38}$$

So, there are some vertices out of q', and then, for the eigenvector centrality of this neighborhood out of q', we have

$$\frac{1}{\lambda}\sum_{j=1}^{(m-q')} \hat{a}_{ij}.\mathrm{EC_j} > 0 \tag{39}$$

Considering (36), (37), and (39), we have $EC_{H'} < EC_H$. On the other side, note that

$$\mathrm{EC_{H'}} = \max\left(\{\mathrm{EC_i} : i = 1, \ldots, m\}\right)$$

Thus, $EC_i \leq EC_{H'}$. So, It holds

$$\mathrm{EC_i} \leq \mathrm{EC_{H'}} < \mathrm{EC_H} \tag{40}$$

It proves the upper bound of this theorem.

LOWER BOUND:

For the lower bound, Let $e\,(G)$ denote the edge number in our bipartite graph G, and V_z, V_θ be two sets of vertices in two partition sides.

Note that $e\,(G)$ is equal to the sum of all vertex's degrees on the one side of bipartite graph G. $V_z = \{v_{z1}, v_{z2}, \cdots, v_{zm}\}$ is the set of vertices in data node side, and deg v_{zi} is the degree of vertices v_{zi}.Note that

$$e\,(G) = \sum_{i=1}^{m} \deg v_{zi} \tag{41}$$

It has been proved before [67] that

$$\lambda_{max}\,(G) \leq \sqrt{e\,(G)} \tag{42}$$

The equality holds if and only if the G is a complete bipartite graph, which in the spectral ensemble clustering is not true. Because there are two sets of vertices V_z and V_θ. Each v_{zi} can connect to only one cluster vertex in each base clustering. So the inequality 42 changes to

$$\lambda_{max}\,(G) < \sqrt{e\,(G)} \tag{43}$$

Since each z_i is a member of one cluster in each base clustering, and there is no isolated vertex in such bipartite graph G, the number of edges connected to each vertex v_{zi} is equal to the number of base clusterings, here as N. Then,

$$\sum_{i=1}^{m} \deg v_i = m \cdot N \tag{44}$$

applying (41) and (44) in (43), it becomes

$$\lambda_{max} < \sqrt{m.N} \tag{45}$$

Let $EC_l be$ the eigenvector centrality based on this new λ_{max}

$$\mathrm{EC_l} = \frac{1}{\sqrt{m.N}}\sum_{j=1}^{q} \hat{a}_{ij}.\mathrm{EC}_j \tag{46}$$

Then it holds the lower bound, and we have both bounds as claimed in this theorem,

$$\frac{1}{\sqrt{m.N}}\sum_{j=1}^{q} \hat{a}_{ij}.\mathrm{EC_j} < \mathrm{EC_i} < \frac{1}{\lambda}\sum_{j=1}^{m} \hat{a}_{ij}.\mathrm{EC}_j$$

□

APPENDIX C
PROOF OF THEOREM 3

The partition of ensemble cluster set $\Theta = \{\Theta_1, \Theta_2, \ldots, \Theta_M\}$ to separated sub-clusters Θ_v and Θ_u, is equivalent to the partition of bipartite graph $G = \{Z, \Theta, R\}$ into subgraphs G_v and G_u where their vertex sets are V_v and V_u. In addition, the non-zero intersection of overlapping clusters is equivalent to the validity of connecting edges between two vertices across these two subgraphs. Then, we need to prove the bipartite graph $G = \{Z, \Theta, R\}$ is connected considering the condition in the theorem. We assume that the partition of Θ fulfil the condition $|\Theta_v \cap \Theta_u| \geq 1$ for all possible partition of G into Θ_v and Θ_u, and $G$is still disconnected. Then, let there be G_u and G_v as components, which are subgraphs equivalent to subsets Θ_u and Θ_v. Here, the vertices of G_u is $V_u = V\,(\mathrm{G}_u) = V\,(G) - V\,(\mathrm{G}_v)$. If $\mathrm{G}_u$is a subgraph with maximal connectivity; there is no edge with one endpoint in G_u and the other endpoint outside of G_u.So, there is no connecting edge between both partitioned components G_u and G_v, which contradicts the condition $|\Theta_v \cap \Theta_u| \geq 1$. It proves that G is a connected bipartite graph. □

APPENDIX D
PROOF OF THEOREM 4

To prove Theorem 4, we consider this assumption that if base clusterings of big data make many overlapping clusters, the same as conditions in Theorem 3, then these overlappings make the resulting bipartite graph connected, which is a key difference from Theorem 2. Considering the bipartite graph is connected, it has been proved by [65] and [66] that the upper bound of the largest eigenvalue for the connected graph changes to

$$\lambda_{max}\,(G) < \sqrt{2e - n_v + 1} \tag{47}$$

where e is the number of edges, n_v denotes the number of all vertices in G, and m, M, and N are the number of data points, number of clusters, and number of base clusterings, respectively. Here, we have

$$e = m.N$$
$$n_v = m + M$$

Substituting these variables for (47), inequality changes to

$$\lambda_{max}(G) < \sqrt{m(2N-1) - M + 1} \tag{48}$$

So, it holds the following lower bound of eigenvector centrality

$$\frac{1}{\sqrt{m(2N-1) - M + 1}} \sum_{j=1}^{q} \hat{a}_{ij}.\mathrm{EC_j} < \mathrm{EC_i}$$

□

REFERENCES

[1] H. Li, X. Ye, A. Imakura, and T. Sakurai, "Ensemble learning for spectral clustering," in *Proc. IEEE Int. Conf. Data Mining (ICDM)*, Sorrento, Italy, Nov. 2020, pp. 1094–1099, doi: 10.1109/ICDM50108.2020.00131.

[2] T. Alqurashi and W. Wang, "Object-neighbourhood clustering ensemble method," in *Proc. Int. Conf. Intell. Data Eng. Automated Learn.*, in Lecture Notes in Computer Science, vol. 8669, 2014, pp. 142–149, doi: 10.1007/978-3-319-10840-7_18.

[3] I. T. Jolliffe and J. Cadima, "Principal component analysis: A review and recent developments," *Phil. Trans. Roy. Soc. A, Math., Phys. Eng. Sci.*, vol. 374, no. 2065, Apr. 2016, Art. no. 20150202, doi: 10.1098/rsta.2015.0202.

[4] A. Zimek and P. Filzmoser, "There and back again: Outlier detection between statistical reasoning and data mining algorithms," *WIREs Data Mining Knowl. Discovery*, vol. 8, no. 6, p. e1280, Nov. 2018, doi: 10.1002/widm.1280.

[5] G.-H. Fu, J.-B. Wang, and W. Lin, "An adaptive loss backward feature elimination method for class-imbalanced and mixed-type data in medical diagnosis," *Chemometric Intell. Lab. Syst.*, vol. 236, May 2023, Art. no. 104809, doi: 10.1016/j.chemolab.2023.104809.

[6] N. Amruthnath and T. Gupta, "Fault diagnosis using clustering. What statistical test to use for hypothesis testing?" *Mach. Learn. Appl., Int. J.*, vol. 6, no. 1, pp. 17–33, Mar. 2019, doi: 10.5121/mlaij.2019.6102.

[7] J.-H. Park, H. Cho, S.-M. Gil, K.-B. Choo, M. Kim, J. Huang, D. Jung, C. Yun, and H.-S. Choi, "Research on clustering-based fault diagnosis during ROV hovering control," *Appl. Sci.*, vol. 14, no. 12, p. 5235, Jun. 2024, doi: 10.3390/app14125235.

[8] T. Wu, J. Fan, and P. Wang, "An improved three-way clustering based on ensemble strategy," *Mathematics*, vol. 10, no. 9, p. 1457, Apr. 2022, doi: 10.3390/math10091457.

[9] H.-Y. Du and W.-J. Wang, "A clustering ensemble framework with integration of data characteristics and structure information: A graph neural networks approach," *Mathematics*, vol. 10, no. 11, p. 1834, May 2022, doi: 10.3390/math10111834.

[10] Y. Liang, Z. Ren, Z. Wu, D. Zeng, and J. Li, "Scalable spectral ensemble clustering via building representative co-association matrix," *Neurocomputing*, vol. 390, pp. 158–167, May 2020, doi: 10.1016/j.neucom.2020.01.055.

[11] H. Li, X. Ye, A. Imakura, and T. Sakurai, "LSEC: Large-scale spectral ensemble clustering," *Intell. Data Anal.*, vol. 27, no. 1, pp. 59–77, Jan. 2023, doi: 10.3233/IDA-216240.

[12] D. Huang, C.-D. Wang, J.-S. Wu, J.-H. Lai, and C.-K. Kwoh, "Ultra-scalable spectral clustering and ensemble clustering," *IEEE Trans. Knowl. Data Eng.*, vol. 32, no. 6, pp. 1212–1226, Jun. 2020, doi: 10.1109/TKDE.2019.2903410.

[13] W. Wang, W. Wang, and H. Liu, "Correlation-guided ensemble clustering for hyperspectral band selection," *Remote Sens.*, vol. 14, no. 5, p. 1156, Feb. 2022, doi: 10.3390/rs14051156.

[14] B. Fei, D. Liu, S. Bi, G. Wu, X. Zhang, J. Ouyang, Y. Zhou, and X. Zhu, "Selective ensemble method based on spectral clustering," in *Proc. 5th Int. Conf. Big Data Inf. Analytics (BigDIA)*, Kunming, China, 2019, pp. 144–149, doi: 10.1109/BigDIA.2019.8802842.

[15] D. Rafailidis, E. Constantinou, and Y. Manolopoulos, "Landmark selection for spectral clustering based on weighted PageRank," *Future Gener. Comput. Syst.*, vol. 68, pp. 465–472, Mar. 2017.

[16] D. Cai and X. Chen, "Large scale spectral clustering via landmark-based sparse representation," *IEEE Trans. Cybern.*, vol. 45, no. 8, pp. 1669–1680, Aug. 2015, doi: 10.1109/TCYB.2014.2358564.

[17] D. Huang, C.-D. Wang, and J.-H. Lai, "Locally weighted ensemble clustering," *IEEE Trans. Cybern.*, vol. 48, no. 5, pp. 1460–1473, May 2018, doi: 10.1109/TCYB.2017.2702343.

[18] H. Li, X. Ye, A. Imakura, and T. Sakurai, "Divide-and-conquer based large-scale spectral clustering," *Neurocomputing*, vol. 501, pp. 664–678, Aug. 2022, doi: 10.1016/j.neucom.2022.06.006.

[19] H. Zhu, L. Lu, J. Yao, S. Dai, and Y. Hu, "Fault diagnosis approach for photovoltaic arrays based on unsupervised sample clustering and probabilistic neural network model," *Sol. Energy*, vol. 176, pp. 395–405, Dec. 2018, doi: 10.1016/j.solener.2018.10.054. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0038092X18310338

[20] T. Ma, F. Wang, J. Cheng, Y. Yu, and X. Chen, "A hybrid spectral clustering and deep neural network ensemble algorithm for intrusion detection in sensor networks," *Sensors*, vol. 16, no. 10, p. 1701, Oct. 2016, doi: 10.3390/s16101701.

[21] Y. Lei, Z. He, Y. Zi, and X. Chen, "New clustering algorithm-based fault diagnosis using compensation distance evaluation technique," *Mech. Syst. Signal Process.*, vol. 22, no. 2, pp. 419–435, Feb. 2008, doi: 10.1016/j.ymssp.2007.07.013.

[22] M. Golyadkin, V. Pozdnyakov, L. Zhukov, and I. Makarov, "SensorSCAN: Self-supervised learning and deep clustering for fault diagnosis in chemical processes," *Artif. Intell.*, vol. 324, Nov. 2023, Art. no. 104012, doi: 10.1016/j.artint.2023.104012.

[23] A. Et-taleby, M. Boussetta, and M. Benslimane, "Faults detection for photovoltaic field based on K-means, elbow, and average silhouette techniques through the segmentation of a thermal image," *Int. J. Photoenergy*, vol. 2020, pp. 1–7, Dec. 2020, doi: 10.1155/2020/6617597.

[24] S. Wang, F. Gao, J. Wu, C. Zheng, X. Fu, and F. Duan, "Online clustering based fault data detection method for distributed PV sites," in *Proc. 39th Chinese Control Conf. (CCC)*, Shenyang, China, 2020, pp. 4341–4346, doi: 10.23919/CCC50068.2020.9188826.

[25] S. Park, S. Park, M. Kim, and E. Hwang, "Clustering-based self-imputation of unlabeled fault data in a fleet of photovoltaic generation systems," *Energies*, vol. 13, no. 3, p. 737, Feb. 2020, doi: 10.3390/en13030737.

[26] K. Dhibi, R. Fezai, M. Mansouri, M. Trabelsi, A. Kouadri, K. Bouzara, H. Nounou, and M. Nounou, "Reduced kernel random forest technique for fault detection and classification in grid-tied PV systems," *IEEE J. Photovolt.*, vol. 10, no. 6, pp. 1864–1871, Nov. 2020, doi: 10.1109/JPHOTOV.2020.3011068.

[27] U. von Luxburg, "A tutorial on spectral clustering," *Statist. Comput.*, vol. 17, no. 4, pp. 395–416, Dec. 2007, doi: 10.1007/s11222-007-9033-z.

[28] Y. Liu, K. Ding, J. Zhang, Y. Li, Z. Yang, W. Zheng, and X. Chen, "Fault diagnosis approach for photovoltaic array based on the stacked auto-encoder and clustering with I-V curves," *Energy Convers. Manage.*, vol. 245, Oct. 2021, Art. no. 114603, doi: 10.1016/j.enconman.2021.114603.

[29] S. Liu, L. Dong, X. Liao, Y. Hao, X. Cao, and X. Wang, "A dilation and erosion-based clustering approach for fault diagnosis of photovoltaic arrays," *IEEE Sensors J.*, vol. 19, no. 11, pp. 4123–4137, Jun. 2019, doi: 10.1109/JSEN.2019.2896236.

[30] Y. Cai, P. Lin, Y. Lin, Q. Zheng, S. Cheng, Z. Chen, and L. Wu, "Online photovoltaic fault detection method based on data stream clustering," *IOP Conf. Ser., Earth Environ. Sci.*, vol. 431, no. 1, 2020, Art. no. 012060, doi: 10.1088/1755-1315/431/1/012060.

[31] J. D. de Guia, R. S. Concepcion, H. A. Calinao, S. C. Lauguico, E. P. Dadios, and R. R. P. Vicerra, "Application of ensemble learning with mean shift clustering for output profile classification and anomaly detection in energy production of grid-tied photovoltaic system," in *Proc. 12th Int. Conf. Inf. Technol. Elect. Eng. (ICITEE)*, Yogyakarta, Indonesia, 2020, pp. 286–291, doi: 10.1109/ICITEE49829.2020.9271699.

[32] E. H. S. Oviedo, L. Travé-Massuyés, A. Subias, C. Alonso, and M. Pavlov, "Hierarchical clustering and dynamic time warping for fault detection in photovoltaic systems," in *Proc. 10th Congreso Internacional CIMM Ingeniería Mecánica, Mecatrónica y Automatización*, May 2021, pp. 31–32. [Online]. Available: https://hal.science/hal-03355362v1

[33] S. Wu, L. Fang, J. Zhang, T. N. Sriram, S. J. Coshatt, F. Zahiri, A. Mantooth, J. Ye, W. Zhong, P. Ma, and W. Song, "Unsupervised anomaly detection and diagnosis in power electronic networks: Informative leverage and multivariate functional clustering approaches," *IEEE Trans. Smart Grid*, vol. 15, no. 2, pp. 2214–2225, Mar. 2024, doi: 10.1109/TSG.2023.3325276.

[34] T. H. Pham, S. Lefteriu, E. Duviella, and S. Lecoeuche, "Auto-adaptive and dynamical clustering for double open-circuit fault diagnosis of power inverters," in *Proc. 4th Conf. Control Fault Tolerant Syst. (SysTol)*, Casablanca, Morocco, Sep. 2019, pp. 306–311, doi: 10.1109/SYSTOL.2019.8864777.

[35] M. Zargarani, A. Migan-Dubois, D. Diallo, C. Delpha, S. Zermani, and C. Mahamat, "Fault diagnosis of grid-connected photovoltaic systems based on unsupervised ensemble clustering and multi-layer perceptron model," in *Proc. 40th Eur. Photovoltaic Solar Energy Conf. Exhib.*, Lisbon, Portugal, Sep. 2023, doi: 10.4229/EUPVSEC2023/4CV.1.27. [Online]. Available: https://userarea.eupvsec.org/proceedings/EU-PVSEC-2023/4CV.1.27/

[36] M. E. Celebi, H. A. Kingravi, and P. A. Vela, "A comparative study of efficient initialization methods for the k-means clustering algorithm," *Exp. Syst. Appl.*, vol. 40, no. 1, pp. 200–210, Jan. 2013, doi: 10.1016/j.eswa.2012.07.021.

[37] D. Eppstein, "Fast hierarchical clustering and other applications of dynamic closest pairs," *ACM J. Experim. Algorithmics*, vol. 5, p. 1, Dec. 2000, doi: 10.1145/351827.351829.

[38] A. Lang and E. Schubert, "BETULA: Fast clustering of large data with improved BIRCH CF-trees," *Inf. Syst.*, vol. 108, Sep. 2022, Art. no. 101918, doi: 10.1016/j.is.2021.101918.

[39] R. Kannan, S. Vempala, and A. Vetta, "On clusterings," *J. ACM*, vol. 51, no. 3, pp. 497–515, May 2004, doi: 10.1145/990308.990313.

[40] H. Zare, P. Shooshtari, A. Gupta, and R. R. Brinkman, "Data reduction for spectral clustering to analyze high throughput flow cytometry data," *BMC Bioinf.*, vol. 11, no. 1, p. 403, Dec. 2010, doi: 10.1186/1471-2105-11-403.

[41] H. Kriegel, P. Kröger, J. Sander, and A. Zimek, "Density-based clustering," *WIREs Data Mining Knowl. Discovery*, vol. 1, no. 3, pp. 231–240, May 2011, doi: 10.1002/widm.30.

[42] D. Comaniciu and P. Meer, "Mean shift: A robust approach toward feature space analysis," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 24, no. 5, pp. 603–619, May 2002, doi: 10.1109/34.1000236.

[43] B. J. Frey and D. Dueck, "Clustering by passing messages between data points," *Science*, vol. 315, pp. 972–976, Feb. 2007, doi: 10.1126/science.1136800.

[44] N. Amruthnath and T. Gupta, "Fault class prediction in unsupervised learning using model-based clustering approach," in *Proc. Int. Conf. Inf. Comput. Technol. (ICICT)*, DeKalb, IL, USA, Mar. 2018, pp. 5–12, doi: 10.1109/INFOCT.2018.8356831.

[45] T. Banerjee, J. M. Keller, M. Skubic, and E. Stone, "Day or night activity recognition from video using fuzzy clustering techniques," *IEEE Trans. Fuzzy Syst.*, vol. 22, no. 3, pp. 483–493, Jun. 2014, doi: 10.1109/TFUZZ.2013.2260756.

[46] S. Brin and L. Page, "The anatomy of a large-scale hypertextual web search engine," *Comput. Netw. ISDN Syst.*, vol. 30, nos. 1–7, pp. 107–117, Apr. 1998, doi: 10.1016/s0169-7552(98)00110-x.

[47] E. J. Bienenstock and P. Bonacich, "Eigenvector centralization as a measure of structural bias in information aggregation," *J. Math. Sociol.*, vol. 46, no. 3, pp. 227–245, Jul. 2022, doi: 10.1080/0022250x.2021.1878357.

[48] C. F. A. Negre, U. N. Morzan, H. P. Hendrickson, R. Pal, G. P. Lisi, J. P. Loria, I. Rivalta, J. Ho, and V. S. Batista, "Eigenvector centrality for characterization of protein allosteric pathways," *Proc. Nat. Acad. Sci. USA*, vol. 115, no. 52, pp. 12201–12208, Dec. 2018, doi: 10.1073/pnas.1810452115.

[49] Z. Li, X.-M. Wu, and S.-F. Chang, "Segmentation using superpixels: A bipartite graph partitioning approach," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Providence, RI, USA, Jun. 2012, pp. 789–796, doi: 10.1109/CVPR.2012.6247750.

[50] V. Kongphet, A. Migan-Dubois, C. Delpha, J.-Y. Lechenadec, and D. Diallo, "Low-cost I–V tracer for PV fault diagnosis using single-diode model parameters and I–V curve characteristics," *Energies*, vol. 15, no. 15, p. 5350, Jul. 2022, doi: 10.3390/en15155350.

[51] D. Chicco and G. Jurman, "The advantages of the Matthews correlation coefficient (MCC) over F1 score and accuracy in binary classification evaluation," *BMC Genomics*, vol. 21, no. 1, pp. 1–6, Jan. 2020, doi: 10.1186/s12864-019-6413-7.

[52] T. Alqurashi and W. Wang, "Clustering ensemble method," *Int. J. Mach. Learn. Cybern.*, vol. 10, no. 6, pp. 1227–1246, Jun. 2019, doi: 10.1007/s13042-017-0756-7.

[53] B. Li, A. Migan-Dubois, C. Delpha, and D. Diallo, "Evaluation and improvement of IEC 60891 correction methods for I-V curves of defective photovoltaic panels," *Sol. Energy*, vol. 216, pp. 225–237, Mar. 2021, doi: 10.1016/j.solener.2021.01.010.

[54] *Blue Solar Monocrystalline Panels, By Victron Energy*. Accessed: Jun. 2023. [Online]. Available: https://www.victronenergy.com/upload/documents/Datasheet-BlueSolar-Monocrystalline-Panels-EN.pdf

[55] A. Fernández and S. Gómez, "Versatile linkage: A family of space-conserving strategies for agglomerative hierarchical clustering," *J. Classification*, vol. 37, no. 3, pp. 584–597, Oct. 2020, doi: 10.1007/s00357-019-09339-z.

[56] D. Arthur, B. Manthey, and H. Röglin, "K-means has polynomial smoothed complexity," in *Proc. 50th Annu. IEEE Symp. Found. Comput. Sci.*, Oct. 2009, pp. 405–414, doi: 10.1109/FOCS.2009.14.

[57] E. Schubert, J. Sander, M. Ester, H. P. Kriegel, and X. Xu, "DBSCAN revisited, revisited: Why and how you should (still) use DBSCAN," *ACM Trans. Database Syst.*, vol. 42, no. 3, pp. 1–21, Sep. 2017, doi: 10.1145/3068335.

[58] G. Hamerly and C. Elkan, "Learning the K in K-means," in *Proc. 16th Int. Conf. Neural Inf. Process. Syst.*, Whistler, BC, Canada. Cambridge, MA, USA: MIT Press, 2004, pp. 281–288. [Online]. Available: https://api.semanticscholar.org/CorpusID:3031883

[59] T. Zhang, R. Ramakrishnan, and M. Livny, "BIRCH: An efficient data clustering method for very large databases," *ACM SIGMOD Rec.*, vol. 25, no. 2, pp. 103–114, Jun. 1996, doi: 10.1145/235968.233324.

[60] R. C. de Amorim and C. Hennig, "Recovering the number of clusters in data sets with noise features using feature rescaling factors," *Inf. Sci.*, vol. 324, pp. 126–145, Dec. 2015, doi: 10.1016/j.ins.2015.06.039.

[61] K. Osmani, A. Haddad, T. Lemenand, B. Castanier, M. Alkhedher, and M. Ramadan, "A critical review of PV systems' faults with the relevant detection methods," *Energy Nexus*, vol. 12, Dec. 2023, Art. no. 100257, doi: 10.1016/j.nexus.2023.100257.

[62] C. P. S. Tautenhain and M. C. V. Nascimento, "An ensemble based on a bi-objective evolutionary spectral algorithm for graph clustering," *Exp. Syst. Appl.*, vol. 141, Mar. 2020, Art. no. 112911, doi: 10.1016/j.eswa.2019.112911.

[63] H. Liu, J. Wu, T. Liu, D. Tao, and Y. Fu, "Spectral ensemble clustering via weighted K-means: Theoretical and practical evidence," *IEEE Trans. Knowl. Data Eng.*, vol. 29, no. 5, pp. 1129–1143, May 2017, doi: 10.1109/TKDE.2017.2650229.

[64] Z. Tao, H. Liu, S. Li, Z. Ding, and Y. Fu, "Robust spectral ensemble clustering via rank minimization," *ACM Trans. Knowl. Discovery Data*, vol. 13, no. 1, pp. 1–25, Feb. 2019, doi: 10.1145/3278606.

[65] H. Yuan, "A bound on the spectral radius of graphs," *Linear Algebra Appl.*, vol. 108, pp. 135–139, Sep. 1988, doi: 10.1016/0024-3795(88)90183-8.

[66] J. Merikoski, R. Kumar, and R. Rajput, "Upper bounds for the largest eigenvalues of a bipartite graph," *Electron. J. Linear Algebra*, vol. 26, pp. 168–176, Jan. 2013, doi: 10.13001/1081-3810.1647.

[67] A. Bhattacharya, S. Friedland, and U. N. Peled, "On the first eigenvalue of bipartite graphs," *Electron. J. Combinatorics*, vol. 15, no. 1, Nov. 2008, Art. no. R144, doi: 10.37236/868.

[68] K. Bache and M. Lichman. (2017). *UCI Machine Learning Repository*. [Online]. Available: http://archive.ics.uci.edu/ml

[69] S. Roweis. *USPS Dataset*. Accessed: Jan. 2024. [Online]. Available: https://cs.nyu.edu/~roweis/data.html

[70] D. Slate, "Letter recognition," *UCI Mach. Learn. Repository*, 1991, doi: 10.24432/C5ZP40.

[71] E. Alpaydin and F. Alimoglu, "Pen-based recognition of handwritten digits," *UCI Mach. Learn. Repository*, 1996, doi: 10.24432/C5MG6K.

[72] X. Ma, W.-H. Huang, E. Schnabel, M. Köhl, J. Brynjarsdóttir, J. L. Braid, and R. H. French, "Data-driven *I–V* feature extraction for photovoltaic modules," *IEEE J. Photovolt.*, vol. 9, no. 5, pp. 1405–1412, Sep. 2019, doi: 10.1109/JPHOTOV.2019.2928477.

MOHSEN ZARGARANI received the M.S. degree in electrical engineering from the University of Grenoble Alpes, France. He is currently pursuing the Ph.D. degree with UMR-Espace Dev, Université de Guyane, Cayenne, France. In 2023, he was a Visiting Researcher with the Institute GeePs, CentraleSupélec, Université Paris-Saclay, France. His current research interests include both pure data analytics and machine learning models and their application to engineering problems in complex systems.

CLAUDE DELPHA (Senior Member, IEEE) received the degree in electrical and signal processing engineering, and the Ph.D. degree in signal processing with smart sensors-based systems applications from the University of Metz. Since 2001, he has been with the Laboratory of Signal and Systems, France. He is involved in signal processing for complex systems security and process monitoring (multimedia and smart systems). He is currently a Full Professor with Université Paris-Saclay, France. His research interests include multidimensional and statistical signal processing, data hiding (watermarking, steganography), pattern recognition, fault detection and diagnosis (incipient and intermittent), estimation, and detection.

DEMBA DIALLO (Senior Member, IEEE) received the M.Sc. and Ph.D. degrees in electrical and computer engineering from the National Polytechnic Institute of Grenoble, France, in 1990 and 1993, respectively. He is currently a Full Professor with the Group of Electrical Engineering Paris, Université Paris-Saclay, France. His current research interests include fault diagnosis, prognosis, fault tolerant control and energy management, and microgrids with renewable energies. He is also an Associate Editor of IEEE TRANSACTIONS ON CYBERNETICS.

ANNE MIGAN-DUBOIS received the M.Sc. and Ph.D. degrees in high frequency and optical telecommunications from the University of Limoges, France, in 1998 and 2001, respectively. She is currently a Full Professor with Paris-Saclay University and develops her research activities with the Group of Electrical Engineering Paris (GeePs). She is the Head of the group on "advanced photovoltaics characterizations in real outdoor conditions." Her research interests include soft integration of photovoltaics in smart grids and smart buildings and PV fault detection and diagnosis. She is an Editorial Board Member of the MDPI journals *Sustainability* and *Materials* and was the Guest Editor of the special issues on "Modeling and Forecasting for Energy Production of Photovoltaic (PV) Systems" in the *International Journal of Photoenergy* and "High Stability Perovskite Solar Cell: Progress and Prospects," respectively. She has chaired several sessions at EuPVSEC.

CHABAKATA MAHAMAT received the M.S. degree in electrical engineering from Polytech Nantes, University of Nantes, Nantes, in 2014, and the Ph.D. degree in electrical engineering from the University of Paris-Saclay (ENS Paris-Saclay, SATIE), in 2018. From 2018 to 2020, he was a Teacher-Researcher with the University of Angers, France. Since 2020, he has been an Associate Professor with the UMR Espace-Dev Laboratory, University of French Guiana. He is currently responsible for the Energy Master's degree at the University of French Guiana.

LAURENT LINGUET received the Ph.D. degree in electrical engineering from Thomson-CSF (now THALES). He then continued his research activities with the LESiR Laboratory, Ecole Normale Supérieure de Cachan. After several years spent with the Renewable Energy Research Group (GRER), Université des Antilles et de Guyane, in January 2011, he joined the Joint Space Research Unit for Development (IRD, UM2, UAG, UR). From 2015 to 2022, he was the Vice-President of Research at the University of French Guiana (UG), where he is currently the President and a Full Professor.