Optimization of the Layout of Distributed Photovoltaic Access Points Using Sparse Autoencoders and Graph Embedding

Song Wan[1], Jing Tan[1], Juncheng Zhang[1], Xiaohong Tan[1], Rongyao Li[1]
[1]Power Grid Planning and Research Center of Guangxi Power Grid Co., LTD, Nanning 530000, Guangxi, China
Corresponding author: Song Wan (e-mail: 13457277979@163.com)

ABSTRACT Efficient planning of distributed photovoltaic (PV) access is increasingly important for intelligent power systems and electromagnetic energy infrastructure, yet existing approaches often suffer from insufficient extraction of high-dimensional electrical features and inadequate representation of network topology. To address these challenges, this paper proposes a hybrid optimization framework that combines Sparse Autoencoders (SAEs) with graph embedding for coordinated PV layout design. A three-layer SAE regularized by Kullback–Leibler divergence learns sparse feature representations emphasizing voltage deviation, line loading, and other electrically sensitive variables associated with PV integration. Simultaneously, the distribution network is modeled as a weighted graph in which Node2Vec generates topology-aware node embeddings that preserve electrical proximity. A bi-level optimization architecture is adopted, where the upper layer employs NSGA-II to maximize photovoltaic penetration and the lower layer performs OpenDSS-based power flow analysis with an electrical similarity mechanism for adaptive capacity adjustment under voltage constraints. Experimental results demonstrate that the proposed SAE effectively extracts key electrical characteristics while accurately preserving both global and local topological structures. The resulting layout strategy significantly improves technical performance and economic efficiency by increasing PV utilization and reducing operational costs. Beyond photovoltaic planning, the proposed feature–topology collaborative optimization paradigm provides a practical computational framework for electromagnetic energy distribution networks and intelligent grid infrastructure requiring adaptive spatial configuration and reliable power delivery.

INDEX TERMS distributed photovoltaic, sparse autoencoder, graph embedding, layout optimization, electromagnetic energy infrastructures

I. INTRODUCTION

BACKGROUND AND SIGNIFICANCE

HIGH penetration of distributed photovoltaic (PV) systems significantly affects both the steady-state operation and dynamic behavior of distribution networks. In industrial zones—particularly those with energy-intensive manufacturing—voltage fluctuations induced by PV intermittency can compromise the reliable operation of sensitive equipment. The coupling relationship between spatial distribution and grid topology and electrical parameters directly affects the safety and economy of the system [1,2]. Modern distribution networks generate high-dimensional, highly nonlinear data from heterogeneous sources. Traditional analysis methods cannot precisely characterize the electrical correlation between nodes [3]. At the same time, the spatiotemporal correlation between PV output and load requires that layout optimization consider both local electrical characteristics and global topological constraints [4-6]. However, existing data-driven methods still have theoretical gaps in the coordinated optimization of feature selection and graph structure modeling. A new optimization framework that integrates deep learning and complex network theory is urgently needed to effectively manage the grid complexities and stability demands inherent in supporting large-scale industrial consumers, such as those within the manufacturing sector.

EXISTING PROBLEMS AND PURPOSE OF THIS PAPER

The core challenge in optimizing distributed PV integration lies in the coordinated modeling of high-dimensional grid characteristics and complex topological relationships. In the feature extraction process, existing data-driven methods have difficulty distinguishing key electrical parameters from redundant noise, and some low-variance but high-sensitivity features are easily ignored, affecting the accuracy of subsequent optimization decisions [7]. The electrical distance of the power grid topology is not simply determined by the physical connection. Dynamic factors such as line impedance and node power injection make the correlation between nodes highly nonlinear [8-10]. Conventional graph embedding methods cannot precisely characterize this time-varying coupling relationship. PV layout optimization must simultaneously maximize penetration while satisfying operational constraints. However, the search for feasible solutions in high-dimensional decision spaces suffers from low efficiency [11,12], and most randomly generated candidate solutions violate voltage or power flow constraints, resulting in a serious waste of computing resources. It should be emphasized that although the basic electrical quantity form is simple, the joint feature space composed of multi-node, multi-time, and multi-type indicators (such as node voltage phase angle, branch power flow direction, impedance phase, etc.) has strong nonlinearity and high redundancy, which makes it difficult for linear or statistical correlation methods to effectively identify the hidden feature dimensions sensitive to photovoltaic disturbance.

This paper constructs a coordinated optimization framework that integrates sparse feature learning and graph structure embedding, overcoming the dual bottlenecks of high-dimensional feature screening and topological relationship maintenance in distributed photovoltaic layout optimization. The core contribution is to build a new decision-making paradigm driven by a data-physics hybrid approach, realizing the joint optimization of electrical feature sensitivity analysis and grid topological representation. This provides a planning method with high precision and strong interpretability for high-proportion photovoltaic access. In terms of innovation, the feature-topology joint modeling mechanism is adopted to design the coupling architecture of the KL divergence-constrained sparse autoencoder and impedance-weighted Node2Vec, which simultaneously retains electrical sensitivity and topological proximity in the feature dimension reduction stage. Through the probabilistic constraints of hidden layer activation, key parameters sensitive to photovoltaic access are automatically identified, and random walk weights are dynamically adjusted based on line impedance to ensure that the embedding space precisely reflects the electrical distance of nodes. A two-layer optimization solution paradigm is designed, and a decision framework for coordinated optimization of NSGA-II and OpenDSS is established to generate candidate solutions in the embedding space and verify their physical feasibility. An innovative solution update strategy guided by electrical similarity is applied: when voltage exceeds the limit, the photovoltaic capacity is adjusted preferentially in the nearest neighborhood of topological embedding, concentrating computing resources in the high-potential solution space and improving search efficiency. The feature importance-topology weight linkage update algorithm is used to enhance dynamic adaptive capability. In the scenario of PV output fluctuation, the feature selection threshold and edge weight coefficient are recalibrated according to real-time power flow data to ensure that the model remains sensitive to changes in operating conditions. The method in this paper addresses the problem of the separation of feature selection and topological modeling from a theoretical perspective and reduces the need for post-processing adjustments of planning schemes at the engineering level, providing a scalable technical path for the planning of new power systems.

II. RELATED WORK

Existing studies have applied various feature selection and graph modeling methods in PV access optimization to improve model performance. Regarding feature selection, some studies use principal component analysis (PCA) to compress electrical feature dimensions [13,14]. Chen Q. proposed a lightweight PCA data compression method based on physical knowledge enhancement, which effectively maintained the key physical characteristics (vibration shapes) in structural vibration data, making it suitable for resource-constrained edge devices. This method significantly outperformed traditional compression techniques in both numerical simulation and actual structural testing, with the highest precision improvement reaching 56% [15]. However, the linear transformation of this method easily leads to the loss of nonlinear features. Some studies have proposed feature screening based on Lasso regression, which can obtain sparse solutions but cannot capture high-order interactions between features [16,17]. The common defects of these methods are: feature selection and topological modeling are separated, which destroys the electrical-topological coupling relationship; physical constraints are not sufficiently embedded, and most of the generated solutions require post-processing adjustment.

Combining graph embedding with microgrid planning can improve the diversity of solution sets. Spectral clustering is a common method based on graph structure learning, which can effectively deal with complex nonlinear data. Graph structure learning clusters data through the spectral information of graphs and has been widely used in power networks and other fields [18,19]. However, existing hybrid methods still have shortcomings. The sparsity control lacks electrical sensitivity guidance. The graph embedding does not consider weight adjustment under dynamic operating conditions. The interaction between the optimization layer and the simulation layer is inefficient. To address these challenges, this paper uses SAE to automatically screen key electrical parameters that are sensitive to photovoltaic access and reduce the optimization dimension. Embedding vectors of grid nodes are generated through the Node2Vec algorithm, retaining the electrical distance relationship. Combining NSGA-II multi-objective optimization and OpenDSS power flow verification, a Pareto optimal solution set that meets the operational constraints is generated, and an electrical-topological dual-driven optimization framework is constructed.

III. METHODS

Figure 1 shows a feature-driven optimization architecture for distribution networks with PV access, including data acquisition, feature screening, graph embedding learning, two-layer optimization modeling, and dynamic neighborhood search. First, the input electrical parameters and topology data are used to construct a sparse autoencoder and a weighted graph model, respectively, to extract key features sensitive to PV access and generate node embedding vectors. The embedding space supports a two-layer optimization model for solution generation and power flow constraint screening to ensure stable system operation. On this basis, a dynamic neighborhood search strategy is adopted to locally adjust the photovoltaic capacity according to electrical similarity, further improving computational efficiency and convergence speed. The output is a set of layout solutions that meet physical constraints and have a high photovoltaic penetration rate. This architecture realizes a closed-loop optimization process of feature-structure-physical coupling.

ADAPTIVE SCREENING MECHANISM OF POWER GRID FEATURES BASED ON SAE

As an unsupervised learning method, the autoencoder has been widely used in dimensionality reduction and feature extraction. Autoencoders can effectively capture the high-dimensional features of data and have been successfully applied in many fields, such as image processing, speech recognition, and power system optimization [20,21].

A three-layer sparse autoencoder network is constructed. The input layer contains multi-dimensional electrical parameters such as voltage deviation and line load rate. The hidden layer activation is constrained by the KL divergence penalty term, and the feature vector with a reconstruction error of less than 5% is output to screen out the feature subset with the highest sensitivity to photovoltaic access.

THREE-LAYER NETWORK STRUCTURE AND SPARSE CONSTRAINT DESIGN

When constructing the network, the number of nodes in the input layer is consistent with the dimensions of the original electrical parameters, such as voltage deviation and line load rate. The number of nodes in the first hidden layer is set to 0.8 times the original dimension for the preliminary extraction of electrical features. The number of nodes in the second hidden layer is further compressed to 0.6 times the first hidden layer to highlight key patterns. The number of nodes in the output layer matches the input layer and is used to reconstruct the original information. The three-layer structure and the reduction ratio ($0.8 \rightarrow 0.6$) are determined by ablation experiments: the two-layer network cannot effectively separate the sensitive features while keeping the reconstruction error below 5%, and networks with four or more layers cause training instability and failure of sparsity control. The ratios of 0.8 and 0.6 achieve the best balance between retaining key electrical features and suppressing redundancy. The weight matrix is initialized using a small-scale random Gaussian distribution, and the bias vector is initialized to zero. The activation function is the sigmoid function in the hidden layer, so the activation range is limited to (0, 1), which is convenient for subsequent sparsity measurement. The loss function consists of two parts: one is the reconstruction error E, defined as follows:

$$E = \frac{1}{N}\sum_{i=1}^{N}\left\|x^{(i)} - \tilde{x}^{(i)}\right\|^2 \tag{1}$$

Here, N represents the number of samples; $x^{(i)}$ is the i-th input vector; $\tilde{x}^{(i)}$ is the corresponding reconstructed vector. The other term is the sparse penalty term P, which constrains the average activation value $\hat{\rho}_j$ of the j-th hidden layer neuron to be close to the preset sparsity parameter ρ in the form of KL divergence:

$$P = \beta\sum_{j=1}^{H}\left[\rho\log\left(\frac{\rho}{\hat{\rho}_j}\right) + (1-\rho)\log\left(\frac{1-\rho}{1-\hat{\rho}_j}\right)\right] \tag{2}$$

Here, H is the total number of hidden layer neurons, and β is the penalty weight coefficient. The overall training objective is to minimize $(E + P)$, and the weight is updated using the Adam optimizer. The initial values of the learning rate and the batch size are set and dynamically adjusted within a fixed range. Since the KL divergence constraint forces the hidden layer neurons to maintain a low average activation rate and only respond significantly to the high-sensitivity electrical features in the input, the importance of the original features can be quantitatively traced by analyzing the coupling relationship between the input-hidden layer weights and activation sparsity. This structure and constraint work together to significantly reduce the overactivation of hidden layer neurons while ensuring reconstruction precision, so that the network automatically focuses on the dimension of electrical features that are most sensitive to photovoltaic access. To ensure the selectivity and stability of the sparse autoencoder, the sparse target is systematically verified. By setting the sparse target for training at different values, redundant features are avoided, and the effect of feature selection is optimized.

Table 1 shows the core sparsity control parameter configuration for the three-layer sparse autoencoder network training phase. Sparsity target represents the target sparsity of the activation value of each hidden layer neuron. When the sparsity target is set to 0.1, the model can effectively avoid the retention of redundant features while ensuring a low reconstruction error, thus optimizing the feature selection effect. The propagation of redundant information is suppressed. The penalty weight is used to adjust the impact of the KL divergence penalty term and achieve a balance between sparsity control and reconstruction error in training loss. The learning rate is set to a smaller range to control the gradient descent step size, reducing the risk of training oscillation and improving convergence stability. The batch size is fixed at 64 to achieve the optimal scheduling between

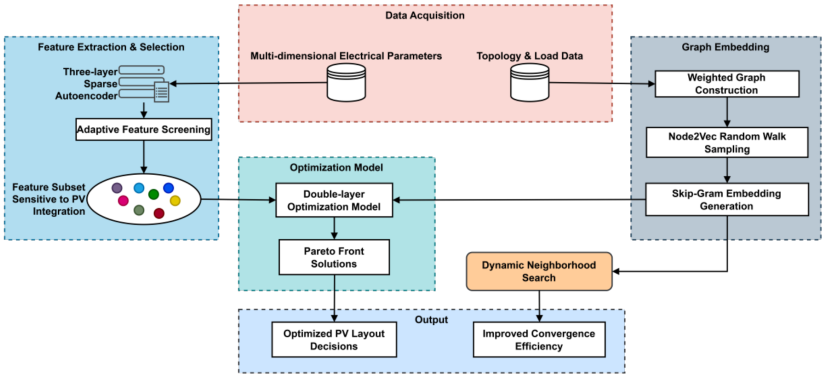

FIGURE 1. Distribution network feature-driven optimization for PV access

training efficiency and memory resources. The parameter configuration ensures that the network is selective and stable in extracting electrical features, effectively highlighting the potential factors most sensitive to the dynamic response of photovoltaic access.

TABLE 1. Sparsity Constraint Settings

Parameter Name	Description	Value Range	Functional Role
Sparsity Target	Desired average activation of hidden neurons	0.05–0.15	Controlling the sparsity level of hidden layer activations
Penalty Weight	KL divergence penalty coefficient	1.0–5.0	Balancing reconstruction accuracy and activation sparsity
Learning Rate	Step size for weight updates	0.0001–0.001	Ensuring stable convergence and preventing gradient explosion/vanishing
Batch Size	Number of samples per training batch	64	Enhancing training efficiency and memory balance

FEATURE SENSITIVITY CALCULATION AND SUBSET EXTRACTION PROCESS

After network training is completed, a complete forward propagation is performed on all training samples; the average activation value of each hidden layer unit is calculated; the absolute difference between the activation value and the preset sparsity is $|\hat{\rho}_j - \rho|$ as the sparse sensitivity index of the unit. Then, the weight matrix W between the input layer and the first hidden layer is used to weight the contribution of each original feature dimension to all hidden layer units to calculate the feature sensitivity score S_i:

$$S_i = \sum_{j=1}^{H} |W_{ij}| \cdot |\hat{\rho}_j - \rho| \tag{3}$$

Here, W_{ij} represents the connection weight from input node i to hidden node j. The higher the score, the more significant the influence of the corresponding electrical parameter in the reconstruction process and the stronger the correlation with photovoltaic access. All feature scores are arranged in descending order, and the top dimensions with cumulative scores accounting for 90% are selected as feature subsets. The subset is input into the encoder again to verify that the reconstruction error E does not exceed the 5% threshold, ensuring that dimensionality reduction does not lose key information. If the error exceeds the limit, the dimension is gradually expanded in the order of scores while retaining the original subset until the reconstruction error meets the requirements. The feature subset finally output not only has a significantly reduced dimension but also contains the most sensitive parameters in the photovoltaic access decision, providing guarantees for both computational efficiency and decision accuracy.

ELECTRICAL TOPOLOGICAL EMBEDDING REPRESENTATION LEARNING

The distribution network is abstracted into a weighted graph structure, where the node weight corresponds to the load power, and the edge weight is the inverse of the line impedance. The Node2Vec algorithm is used for random walk sampling, and the node embedding vector is generated through the Skip-Gram model so that the Pearson correlation coefficient between the node distance and the electrical distance in the embedding space is ≥ 0.85.

WEIGHTED GRAPH CONSTRUCTION AND WEIGHTED RANDOM WALK

The nodes of the distribution network are mapped as vertices in the graph, and the load power is assigned to the corresponding node weight. The line impedance is taken as its inverse, and the weight of the connected edge is assigned to form a weighted adjacency matrix W_{uv}. Its elements are defined as follows:

$$W_{uv} = \begin{cases} Z_{uv}^{-1}, & \text{if nodes } u \text{ and } v \\ & \text{are physically connected,} \\ 0, & \text{otherwise.} \end{cases} \tag{4}$$

Here, Z_{uv}^{-1} represents the line impedance between nodes u and v. The node weights are stored in a diagonal matrix D, and $D_{uu} = P_u$, where P_u is the load power of node u.

Biased random walk sampling is performed on this graph structure to capture both electrical proximity and long-range correlations. During the walk process, if the current walk path is from the predecessor node t to the current position v, the transition probability of selecting the next target x is calculated as follows:

$$\pi_{v \to x} = \frac{\alpha_{t,v,x} W_{vx}}{\sum_{y \in N(v)} \alpha_{t,v,y} W_{vy}} \tag{5}$$

Here, $N(v)$ represents the neighbor set of node v. The bias factor $\alpha_{t,v,y}$ is determined by the following rules: If $x = t$, then $\alpha = 1/p$, where p is the probability adjustment parameter for returning to the original node; if x is directly connected to t, then $\alpha = 1$; otherwise, $\alpha = 1/q$, where q is used to control the tendency to explore remote nodes. The parameters p and q are adjusted in multiple groups of walk experiments so that the sampling sequence can retain local electrical tightness and cover the network's important long-range connections. This sampling strategy effectively solves the problem that it is difficult to reveal the nonlinear coupling of electrical distance by relying solely on physical adjacency.

SKIP-GRAM TRAINING AND EMBEDDING SPACE CALIBRATION

Multiple fixed-length walk sequences are input into the Skip-Gram model, and the context windows in which the nodes appear in the sequences are used as positive sample pairs to minimize the objective function:

$$J = \sum_{u \in V} \sum_{v \in C(u)} - \ln \frac{\exp(z_u^\top z_v)}{\sum_{w \in V} \exp(z_u^\top z_w)} \tag{6}$$

Here, V is the set of all nodes; $C(u)$ is the set of context neighbors of node u; $z_u \in \mathbb{R}^d$ and $z_v \in \mathbb{R}^d$ are the embedding vectors of nodes u and v, respectively; $\mathbb{R}^d$ is the embedding dimension. To verify the influence of embedding dimension on the retention of topological structure, experiments are conducted with different embedding dimensions. The experimental results show that when the embedding dimension is 128, the retention of electrical distance between nodes (Spearman rank correlation coefficient) is the highest, ensuring high-precision representation of topological relationships in the embedding space. Negative sampling is used in the optimization process to reduce the computational complexity. The Adam algorithm is used for embedding vector updates. The learning rate is linearly decayed from the initial 0.01 to 0.001. After training is complete, the Pearson correlation coefficient between the embedding spatial distance $d_{ij}^{\text{embed}} = \|z_i - z_j\|$ and the corresponding electrical distance d_{ij}^{elec} of all node pairs (i, j) is calculated:

$$r = \frac{\sum_{i<j} \left(d_{ij}^{\text{elec}} - \mu_e\right)\left(d_{ij}^{\text{embed}} - \mu_m\right)}{\sqrt{\sum_{i<j} \left(d_{ij}^{\text{elec}} - \mu_e\right)^2} \sqrt{\sum_{i<j} \left(d_{ij}^{\text{embed}} - \mu_m\right)^2}} \tag{7}$$

Here, μ_e and μ_m represent the means of all electrical distances and embedding distances, respectively. The model is iterated until $r \geq 0.85$ to ensure that the distance between nodes in the embedding space is highly consistent with the physical electrical distance. This process solves the problem that graph embedding often ignores the electrical characteristics of the line and provides a reliable topological representation for subsequent optimization steps.

LAYOUT DECISION GENERATION COUPLED WITH PHYSICAL CONSTRAINTS

A two-layer optimization model is established. The upper layer aims to maximize the photovoltaic penetration rate and uses the NSGA-II algorithm to generate candidate solutions in the embedding space. The lower layer calls OpenDSS to perform power flow calculations and screen the Pareto frontier solution set that simultaneously satisfies voltage deviation $\leq 5\%$ and line load rate $\leq 90\%$. It should be noted that the low-dimensional electrical features selected by SAE are used to construct OpenDSS simulation inputs, while the embedded vectors generated by Node2Vec are exclusively employed to define the decision space structure of NSGA-II. These two components operate in tandem through a dual-channel mechanism combining embedded-guided search and feature-driven validation.

CANDIDATE SOLUTION GENERATION AND MULTI-OBJECTIVE EVOLUTION IN EMBEDDING SPACE

The node embedding vector is mapped to the decision variable space, and the decision vector $c = [c_1, c_2, \ldots, c_i]$ is defined, where c_i represents the planned installed capacity of the distributed photovoltaic at node i, and M is the total number of candidate nodes. The objective function is set to maximize the photovoltaic penetration rate:

$$f_1 = \frac{\sum_{i=1}^{M} c_i}{\sum_{k=1}^{N} P_{\text{load},k}} \tag{8}$$

Here, $P_{\text{load},k}$ is the total load power of the k-th branch of the distribution network, and N is the number of branches. When the genetic algorithm is initialized, the population size is set, and individuals are randomly generated in the embedding space according to a uniform distribution corresponding to different c values. Two sets of parent vectors are selected through the crossover operation. Single-point exchanges are performed at the vector level, and the offspring are generated. Through the mutation operation, a small perturbation related to the Euclidean distance of the node embedding vector is added or subtracted to several randomly selected components in vector c. Non-dominated sorting and crowding distance are used for population selection to ensure penetration rate diversity and computing power balance. During the iteration process, individuals in each generation of the population are continuously updated so that high-potential areas in the embedding space are more sampled and deep mining of possible high-penetration solutions is achieved.

PHYSICAL SIMULATION VERIFICATION AND SOLUTION SET SCREENING

Several candidate decision vectors output by the genetic algorithm are exported as OpenDSS simulation scripts. The distribution network modeling file is called, and the photovoltaic capacity of each node is dynamically loaded. After the power flow calculation is completed, the voltage V_i is extracted for each node, and the node voltage deviation rate is calculated according to the formula:

$$\Delta V_i = \frac{|V_i - V_{\mathrm{nom}}|}{V_{\mathrm{nom}}} \tag{9}$$

Here, V_{nom} is the rated voltage of the system. The actual load rate $L_j = S_j/S_{j,\max}$ is read for all lines, where S_j represents the apparent power flowing through the line, and $S_{j,\max}$ is its rated capacity. The solution set screening rule is limited to $\max_i \Delta V_i \leq 0.05$ and $\max_j L_j \leq 0.90$. During the NSGA-II search process, voltage offset and load rate constraints are directly embedded into the optimization objective function as hard constraints. These constraints are managed through penalty functions to ensure that each candidate solution meets the voltage and load requirements from the outset. Specifically, if the voltage offset or load rate exceeds the limit, the relevant solutions are adjusted according to the penalty factor to prevent solutions that violate the physical constraints from advancing to the next iteration. All individuals that meet the above dual physical constraints constitute the final Pareto frontier solution set, which not only retains the optimality of penetration rate but also ensures voltage stability and branch safety load, solving the problem of relying only on theoretical space solution generation and ignoring actual operational restrictions. It should be emphasized that when a candidate solution violates voltage constraints, this method first activates a local capacity redistribution mechanism based on electrical similarity. Within the Node2Vec embedding space, the system identifies the nearest neighbors of the violating node using Euclidean distance and transfers a portion of photovoltaic capacity according to their load margin ratios. The newly generated candidate solution undergoes re-validation. This strategy better preserves high-penetration potential solutions compared to standard penalty functions, effectively preventing premature convergence to low-penetration regions. To distinguish hard constraints from soft constraints, voltage offset and line load rate are regarded as hard constraints, which must be satisfied in the optimization process. Other objectives, such as maximum photovoltaic penetration, are soft constraints, which can be optimized on the basis of meeting hard constraints.

IV. LAYOUT OPTIMIZATION PERFORMANCE VERIFICATION

IMPORTANCE OF GRID FEATURES AND RECONSTRUCTION ERROR

The data used for the importance of power grid features and reconstruction error analysis are from the measured electrical parameters of a regional distribution network, covering ten types of indicators. After the data are processed, outliers are removed and normalized, and then input into a three-layer sparse autoencoder model. The model uses reconstruction error as the objective function and uses a KL divergence penalty term to constrain the hidden layer activation to ensure the sparsity of the encoding process. After training, the reconstruction error of each feature is calculated. The lower the error, the better the feature information is retained. The sensitivity score is obtained by statistically analyzing the correlation between the activation of hidden layer neurons and the feature input, reflecting the influence of the feature on the model output. This method ensures that the selected feature subset has a high identification ability for photovoltaic access perturbations, providing a solid foundation for subsequent analysis.

Figure 2 shows the importance of power grid features and reconstruction error under the sparse autoencoder. Figure 2(a) shows the ability of the sparse autoencoder to reconstruct the input features. The horizontal axis is the 10 main electrical features, and the vertical axis is the mean-square error (MSE) of these features in the autoencoder output. The reconstruction errors of Voltage Deviation and Active Power Flow are relatively high, indicating that these features change dramatically during the compression process and may better reflect the state changes of the system after photovoltaic access. The errors of other features, such as Impedance Magnitude or Current Magnitude, are small, indicating that they can still be well reconstructed under low-dimensional coding and have high redundancy.

Figure 2(b) integrates the dual-axis information of sparse sensitivity score and cumulative contribution. The horizontal axis is the feature name, and the left vertical axis corresponds to the bars, which represent the sensitivity scores obtained after KL divergence regularization, indicating the contribution of each feature to the sparsity of activation units. The right vertical axis corresponds to the lines, which represent the cumulative percentage contribution of each feature, showing that the top four features can explain nearly 60% of the sensitivity variation. This shows that Voltage Deviation, Line Load Rate, Power Factor, and Frequency are parameters that are more sensitive to system state changes and can be used preferentially for the subsequent construction of photovoltaic access judgment models. Features with low scores have limited contributions to the model, which may be due to their hysteresis in changes or limited spatial distribution range, resulting in their weakened role in photovoltaic perturbation perception. This visualization method makes feature screening more interpretable and helps to make feature selections with clear physical meanings in subsequent deployment steps.

RELATIONSHIP BETWEEN EMBEDDING SPATIAL DISTANCE AND ELECTRICAL DISTANCE, POWER-AWARE ELECTRICAL TOPOLOGY, AND SAMPLING WALK PATH

The actual operating parameters of the regional distribution network contain the load power and line impedance information of multiple nodes. The distribution network is modeled as a weighted graph structure, in which the weight of the node is the load power and the weight of the edge is the inverse of the line impedance, reflecting the strength of the electrical connection. Subsequently, the Node2Vec algorithm is used to perform multiple random walk samplings, and the sampling path captures the complex topological relationships and electrical features between nodes. The sampled data are input into the Skip-Gram model. A low-dimensional embedding vector is generated during the training process to ensure that the distance between nodes in the space reflects the actual electrical distance to the greatest extent. The Euclidean distance of the node pair in the embedding space is calculated and paired with the electrical distance of the original node pair. Scatter points are drawn to verify the ability of the node distance in the graph embedding space to maintain the original electrical structure.

Figure 3 shows the correspondence between the embedding spatial distance and the electrical distance between nodes in the distribution network. The horizontal axis represents the electrical distance between node pairs, which is defined as the impedance value of the connecting line. The vertical axis is the Euclidean distance of the nodes obtained after graph embedding training. In the scatter plot, each point corresponds to a pair of nodes, and its coordinate position reflects its relative distribution in the original electrical space and the embedding space. From the overall trend, the data points show an obvious positive correlation, indicating that the physical connectivity between nodes is consistent with the distance between their embedding vectors. This correspondence shows that the embedding model maintains the electrical characteristics of the original graph structure well during the training process. The embedding distance between nodes is not completely linear with the electrical distance, and some data points have slight deviations, especially in the higher electric distance range, where the dispersion is slightly higher.

Figure 4 shows the expanded electrical network graph structure, which includes 12 nodes and 16 weighted connecting edges. The connection relationships between the nodes are more hierarchical, reflecting the branch distribution characteristics of the actual distribution network. The node size in the figure is scaled according to the load power value. The maximum load corresponds to the largest dot. The edge weight label shows the inverse of the line impedance. The larger the value, the stronger the transmission capacity. The red edge is the path trajectory of a round of random walks, starting from node 1, passing through nodes 2, 5, 10, and finally reaching node 12. The black-filled nodes identify the key nodes on this path. The walk path clearly shows that the algorithm tends to choose paths with high edge weights (i.e., low impedance) to achieve more effective electrical connectivity feature learning. The node loads show a certain non-uniform distribution. For example, the loads of nodes 4 and 8 are high (large nodes), while the loads of nodes 11 and 7 are small. This difference helps the model strengthen the weight perception of important nodes during the embedding process. The weights of the edges in the path are $10 \rightarrow 4 \rightarrow 1 \rightarrow 1$, and the overall path weight is high. The model can preferentially capture low-impedance and high-connectivity paths during the sampling process, thereby enhancing the ability of the embedding vector to express the strong and weak relationships of electrical topology in space.

BALANCES BETWEEN PENETRATION RATE, LINE PRESSURE, AND VOLTAGE STABILITY

The data used come from the layout candidate solution set in the embedding space, and the specific generation process includes two stages. In the first stage, the optimization framework is used to construct the photovoltaic layout candidate solution, and the node configurations with different penetration rate levels are randomly generated in the multi-objective space through the evolutionary search mechanism. At the same time, the node embedding representation is combined to control the rationality of the layout distribution. In the second stage, each set of solutions is input into the power flow analysis module, and the OpenDSS simulation engine is called to calculate the corresponding voltage deviation and line load rate indicators. The maximum voltage deviation value and the maximum line load rate are used as the basis for evaluating the system operation constraints so as to select the solutions that meet voltage $\leq 5\%$ and load $\leq 90\%$ to form the Pareto frontier. The penetration rate, voltage, and load data of all solutions are mapped to two-dimensional space for chart drawing. Figure 5 shows the balances between penetration rate, line pressure, and voltage stability.

In Figure 5(a), the horizontal axis represents the proportion of photovoltaic access in the candidate solution in the entire load, i.e., the penetration rate. The vertical axis is the maximum voltage deviation generated by the electrical node under the corresponding layout scheme. The blue dots show that as the penetration rate increases, the voltage deviation shows an overall downward trend but is accompanied by discrete fluctuations. The Pareto points marked by red crosses are concentrated in the medium and high penetration rate areas, representing Pareto solutions that simultaneously meet the physical constraint that the voltage deviation is not higher than 0.05. The schemes within this range can keep the voltage deviation below the threshold at a higher photovoltaic share. This phenomenon stems from the accurate characterization of the electrical distance of the grid nodes by the decision vector in the

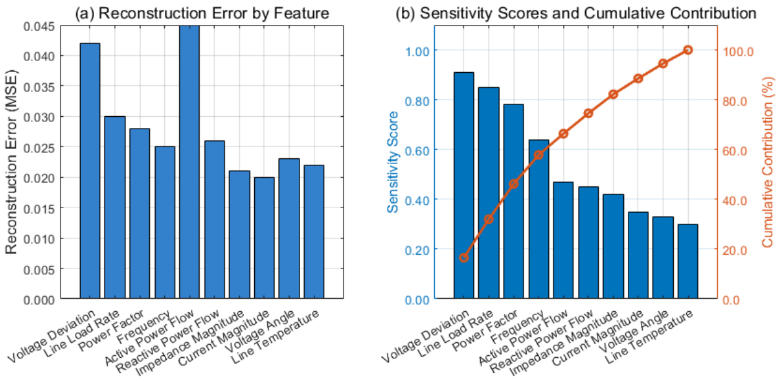

FIGURE 2. Importance of power grid features and reconstruction error under sparse autoencoder. Figure 2(a). Reconstruction Error Feature. Figure 2(b). Sensitivity Score and Cumulative Contribution

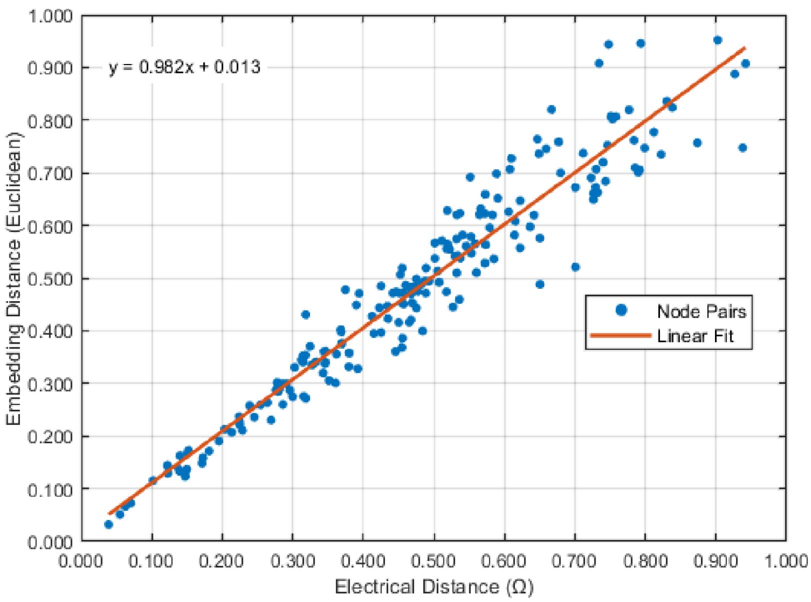

FIGURE 3. Scattering of embedding and electrical distance

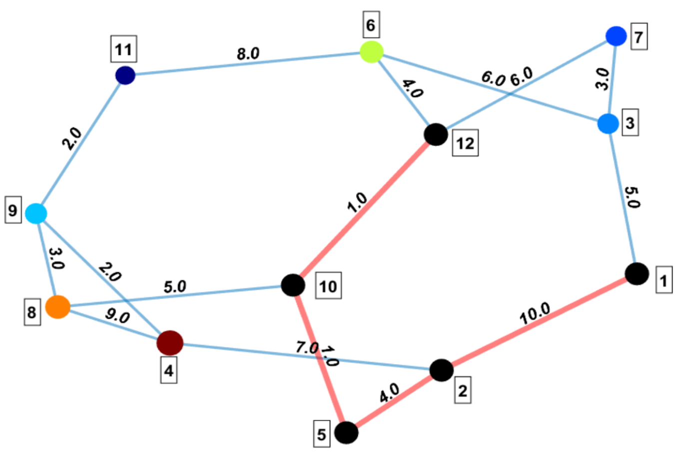

FIGURE 4. Power-aware electrical topology and sampling path

embedding space, which enables more sampling in high-potential areas, thereby generating a better layout. The fluctuation of voltage deviation reflects the unevenness of electrical coupling between nodes.

In Figure 5(b), the horizontal axis also represents the penetration rate. The vertical axis is the maximum line load rate in each scheme. The green dots correspond to invalid schemes that do not meet the load constraints, which are mainly concentrated in the low-penetration rate and high-branch load range, highlighting the risk of overload at network bottlenecks. The dark green cross is the Pareto solution that satisfies the load rate not exceeding 0.9, showing a trend of steadily decreasing load rate as the penetration rate increases. This trend is due to the dynamic adjustment strategy of node electrical similarity in the decision-making process, which can prioritize the expansion of low-resistance paths in the electrical neighborhood and disperse the pressure of traffic concentration points, thereby reducing the probability of overload. Through linkage observation, the Pareto scheme considers both high penetration rate and physical stability, indicating that the coupled simulation verification and multi-objective screening process effectively avoids the safety hazards brought about by single-objective optimization.

FEATURE SELECTION EFFECTIVENESS INDEX AND COMPARISON OF DIFFERENT OPTIMIZATION METHODS

Three feature dimensionality reduction methods—SAE, PCA, and Lasso—are processed and compared in terms of their information retention and redundancy suppression performance. The specific approach is to set different information loss rate conditions and use the inverse reconstruction error to adjust the dimensionality reduction strength of each method so that the information entropy retained after dimensionality reduction is consistent. The mutual information retention rate is calculated by measuring the information interconnection between the nested coding variables and the target variable. The redundant feature elimination rate is measured by the suppression ratio of linearly correlated dimensions in the covariance matrix of the final feature set. All data are averaged after five-fold cross-validation to avoid accidental errors affecting the stability of the conclusions.

The two sub-figures in Figure 6 show the performance differences among the three feature selection methods in terms of mutual information retention rate and redundant feature elimination rate at different information loss levels.

Figure 6(a) shows that the SAE method can better maintain the high-quality original information mapping relationship during the information loss process and can maintain a mutual information retention rate of 0.85 at an information loss of 25%. Its curve is smooth and maintains a high value throughout the process, indicating that the expressive ability of the feature reconstruction is not significantly weakened. This result is attributed to the KL divergence sparsity constraint mechanism used in SAE, which makes the hidden layer more inclined to activate dimensions with a strong correlation with the input, thereby effectively avoiding the risk of the main component of information being mistakenly deleted during the dimensionality reduction process. In contrast, PCA is limited by the linear projection structure and has insufficient ability to retain high-order statistical features, resulting in more significant information loss. Although Lasso has a certain dimensional compression capability, it is prone to redundant expression under input conditions with strong feature correlation, and its retention rate drops rapidly. Figure 6(b) further verifies the above observations from the perspective of redundancy control. SAE can more precisely eliminate low-contribution feature dimensions while increasing the information loss rate. The redundant feature elimination rate is 0.85 at 25% information loss, showing a gradually increasing trend. This is because, with higher information loss, the algorithm discards more features, naturally including more redundant parts, and the elimination rate increases. However, this may also lead to the loss of effective features, and the mutual information retention rate may decrease. The data show that the adaptive structure has a stronger ability to identify invalid information than traditional sparse constraints. PCA has a delayed response to redundant dimensions. Although Lasso has some improvement in some stages, its robustness is poor. SAE achieves a better balance between information fidelity and dimensionality reduction redundancy control, with a significant effect on the representation ability of the feature space.

Table 2 compares the performance of four different optimization methods: SAE + Node2Vec + NSGA-II, PCA + Node2Vec + NSGA-II, Lasso + Node2Vec + NSGA-II, and the traditional non-embedded NSGA-II algorithm. By comparing key indicators such as photovoltaic penetration, maximum voltage deviation, maximum line load, optimization cost, and convergence speed, the optimization effect after introducing SAE and Node2Vec can be observed. The experimental results show that the integrated method based on SAE and Node2Vec is superior to the traditional NSGA-II algorithm in terms of photovoltaic penetration, maximum voltage deviation, and line load, while the optimization cost and convergence speed are also significantly improved. Especially in terms of convergence speed, the traditional NSGA-II algorithm requires more iterative generations to achieve the ideal optimization effect, while the model using the embedded method can achieve convergence in fewer generations. This comparison verifies the advantages of the embedded method in improving optimization performance and computational efficiency.

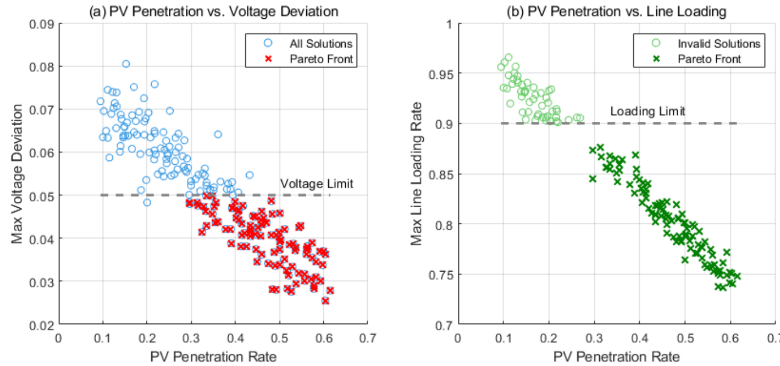

FIGURE 5. Balances between penetration rate, line pressure, and voltage stability. Figure 5(a). PV Penetration vs. Voltage Deviation. Figure 5(b). PV Penetration vs. Line Loading

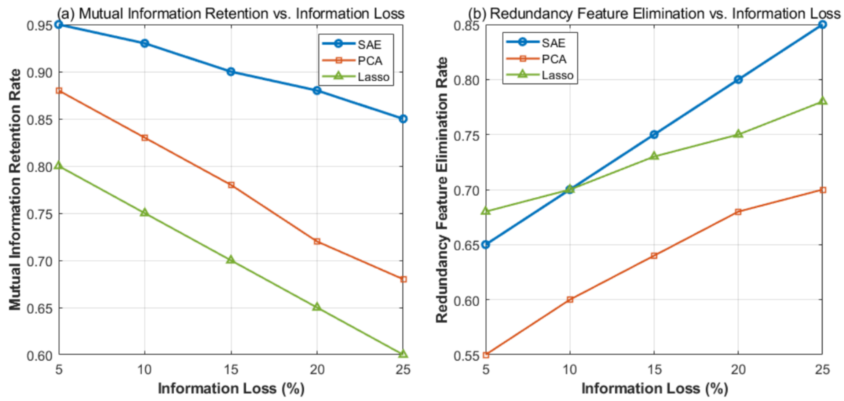

FIGURE 6. Comparison of mutual information retention rate and redundant feature elimination rate. Figure 6(a). Mutual Information Retention vs. Information Loss. Figure 6(b). Redundancy Feature Elimination vs. Information Loss

TABLE 2. Performance of Different Optimization Methods

Method	PV Penetration Rate	Max Voltage Deviation (%)	Max Line Load (%)	Optimization Cost ($\times 10^6$ USD)	Convergence Speed (Generations)
SAE + Node2Vec + NSGA-II	20.40%	4.2	85.3	7.83	50
PCA + Node2Vec + NSGA-II	19.10%	4.5	86.1	8.00	55
Lasso + Node2Vec + NSGA-II	19.60%	4.3	85.7	7.95	52
NSGA-II (No Embedding)	18.30%	5.0	88.2	8.50	70

TOPOLOGY RETENTION EVALUATION

A weighted undirected graph containing node load power and line impedance information is constructed, and the actual electrical distance between node pairs is determined by the branch impedance-weighted path length. Subsequently, the node vectors are generated using Node2Vec and the currently popular LINE, DeepWalk, and GraphSAGE embedding methods, respectively, and the Spearman rank correlation coefficient between the Euclidean distance in the embedding space and the electrical distance is calculated for all node pairs. By setting the reconstruction edge threshold, the overlap between the top-ranked edges in the embedding space and the actual high electrical connection edges is compared to obtain the accuracy of key edge reconstruction.

The above process is repeated five times, and the Wilcoxon test is used to evaluate the significance of the differences between the methods.

The evaluation results in Table 3 quantify the ability of different embedding algorithms to retain electrical topological structure features. The Spearman rank correlation coefficient measures the monotonic relationship between the embedding node distance and the actual electrical distance, reflecting the degree of maintenance of relative proximity. The key edge reconstruction accuracy evaluates the accuracy of the retention of key electrical connections in the embedding space, which is crucial for reliable topological representation.

TABLE 3. Comparison of Topology Retention Evaluation

Method	Spearman Rank Correlation ↑	Key Edge Reconstruction Accuracy (%) ↑	p-value (vs. Node2Vec)
Node2Vec	0.872	91.4	—
LINE	0.749	78.2	0.0027
DeepWalk	0.803	83.7	0.0142
GraphSAGE	0.765	79.5	0.0096

V. CONCLUSION

This paper focuses on the optimization of distribution network layout under distributed PV access and constructs a systematic methodological framework that integrates feature screening, graph embedding representation, and multi-objective decision-making. This framework offers a crucial planning tool for sectors like the industry to efficiently integrate clean energy while maintaining the stable power supply required for continuous manufacturing processes. A sparse autoencoder network is constructed to extract the core indicators that are most sensitive to photovoltaic access from multi-dimensional electrical quantities, achieving the unity of feature dimension compression and selection efficiency. On this basis, the distribution network is mapped into a weighted graph structure, and structured random walks are combined with embedding learning methods to obtain a node vector expression that considers both electrical relationships and topological semantics, thereby improving the physical consistency of subsequent searches. Furthermore, a set of feasible layout schemes is generated by coordinating voltage and load constraints through a two-layer optimization model, and a neighborhood search strategy driven by physical similarity is adopted to effectively accelerate the convergence process and reduce the algorithm's computational complexity. Multi-angle evaluation is carried out from the perspectives of feature effectiveness, topology retention ability, and economy, verifying the comprehensive advantages of the method in terms of precision, stability, and practicality. The overall research not only enhances the level of intelligent decision-making in photovoltaic distributed deployment but also provides theoretical support and a practical path for building an adaptable and scalable new energy access control system, which is vital for guaranteeing the necessary power flexibility and resilience for continuous, high-demand industrial operations like manufacturing.

AUTHOR CONTRIBUTIONS

Conceptualization –Song Wan, Jing Tan, Juncheng Zhang, Xiaohong Tan and Rongyao Li; methodology – Song Wan, Jing Tan, Juncheng Zhang, Xiaohong Tan and Rongyao Li; investigation – Song Wan, Jing Tan, Juncheng Zhang and Xiaohong Tan; writing-original draft preparation – Song Wan, Jing Tan, Juncheng Zhang, Xiaohong Tan and Rongyao Li. All authors have read and agreed to the published version of the manuscript.

CONFLICTS OF INTEREST

The authors declare no conflict of interest.

FUNDING

This research received no external funding.

ACKNOWLEDGEMENTS

Not applicable.

REFERENCES

[1] Y. Wei, B. Zhao, J. Hu, X. Xiao, M. Shi, and Q. Zhou, "A two-layer planning method for distributed energy storage with multi-point layout in high photovoltaic penetration distribution network," Journal of Electrical Engineering & Technology, vol. 20, no. 1, pp. 1-12, 2025, doi: 10.1007/s42835-024-01945-1.

[2] N. Klasen, F. Lux, J. Weber, T. Roessler, A. Kraft, and A. comprehensive study of module layouts for silicon solar cells under partial shading, "IEEE Journal of Photovoltaics," 2022; 12(2):546-556, doi: 10.1109/JPHOTOV.2022.3144635.

[3] L. Wang, L. Jia, G. Xie, X. Chen, and Y. Liu, "Carbon emission reduction: Contribution of photovoltaic power and practice in China," China Geology, vol. 7, no. 2, pp. 371-376, 2024, doi: 10.31035/cg2024078.

[4] X. Li, L. Wang, N. Yan, and R. Ma, "Cooperative dispatch of distributed energy storage in distribution network with PV generation systems," IEEE Transactions on Applied Superconductivity, vol. 31, no. 8, pp. 1-4, 2021, doi: 10.1109/TASC.2021.3117750.

[5] X. Fu, "Statistical machine learning model for capacitor planning considering uncertainties in photovoltaic power," Protection and Control of Modern Power Systems, vol. 7, no. 1, pp. 1-13, 2022, doi: 10.1186/s41601-022-00228-z.

[6] X. Sun, J. Qiu, and J. Zhao, "Optimal local volt/var control for photovoltaic inverters in active distribution networks," IEEE Transactions on Power systems, vol. 36, no. 6, pp. 5756-5766, 2021, doi: 10.1109/TPWRS.2021.3080039.

[7] S. Zhang, Y. Fang, H. Zhang, H. Cheng, and X. Wang, "Maximum hosting capacity of photovoltaic generation in SOP-based power distribution network integrated with electric vehicles," IEEE Transactions on Industrial Informatics, vol. 18, no. 11, pp. 8213-8224, 2022, doi: 10.1109/TII.2022.3140870.

[8] A. M. Brockway, J. Conde, and D. Callaway, "Inequitable access to distributed energy resources due to grid infrastructure limits in California," Nature Energy, vol. 6, no. 9, pp. 892-903, 2021, doi: 10.1038/s41560-021-00887-6.

[9] S. Theocharides, C. Spanias, I. Papageorgiou, G. Makrides, S. Stavrinos, and Efthymiou V et al, "A hybrid methodology for distribution level photovoltaic power production forecasting verified at the distribution system of Cyprus," IET Renewable Power Generation, vol. 16, no. 1, pp. 19-32, 2022, doi: 10.1049/rpg2.12296.

[10] R. Xu, C. Zhang, Y. Xu, Z. Dong, and R. Zhang, "Multi-objective hierarchically-coordinated volt/var control for active distribution networks with droop-controlled PV inverters," IEEE Transactions on Smart Grid, vol. 13, no. 2, pp. 998-1011, 2021, doi: 10.1109/TSG.2021.3126761.

[11] J. Jiang, T. Zhang, and D. Chen, "Analysis, design, and implementation of a differential power processing DMPPT with multiple buck–boost choppers for photovoltaic module," IEEE Transactions on Power Electronics, vol. 36, no. 9, pp. 10214-10223, 2021, doi: 10.1109/TPEL.2021.3063230.

[12] M. Amir, Zaheeruddin, A. Haque, F. I. Bakhsh, V. S. B. Kurukuru, and M. Sedighizadeh, "Intelligent energy management scheme-based coordinated control for reducing peak load in grid-connected photovoltaicpowered electric vehicle charging stations," IET Generation, Transmission & Distribution, vol. 18, no. 6, pp. 1205-1222, 2024, doi: 10.1049/gtd2.12772.

[13] R. Pourramezan, R. Hassani, H. Karimi, M. Paolone, and J. Mahseredjian, "A real-time synchrophasor data compression method using singular value decomposition," IEEE Transactions on Smart Grid, vol. 13, no. 1, pp. 564-575, 2021, doi: 10.1109/TSG.2021.3114585.

[14] N. Hashemipour, J. Aghaei, A. Kavousi-Fard, T. Niknam, L. Salimi, and Granado PC et al, "Optimal singular value decomposition based big data compression approach in smart grids," IEEE Transactions on industry applications, vol. 57, no. 4, pp. 3296-3305, 2021, doi: 10.1109/TIA.2021.3073640.

[15] Q. Chen, J. Cao, and Y. Xia, "Physics-enhanced PCA for data compression in edge devices," IEEE Transactions on Green Communications and Networking, vol. 6, no. 3, pp. 1624-1634, 2022, doi: 10.1109/TGCN.2022.3171681.

[16] M. Nejati and N. Amjady, "A new solar power prediction method based on feature clustering and hybridclassification-regression forecasting," IEEE Transactions on Sustainable Energy, vol. 13, no. 2, pp. 1188-1198, 2021, doi: 10.1109/TSTE.2021.3138592.

[17] X. Ge, F. Xu, Y. Wang, H. Li, F. Wang, and Hu J et al, "Spatio-temporal two-dimensions data based customer baseline load estimation approach using LASSO regression," IEEE Transactions on Industry Applications, vol. 58, no. 3, pp. 3112-3122, 2022, doi: 10.1109/TIA.2022.3148351.

[18] J. Wu, X. Chen, S. Badakhshan, J. Zhang, and P. Wang, "Spectral graph clustering for intentional islanding operations in resilient hybrid energy systems," IEEE Transactions on Industrial Informatics, vol. 19, no. 4, pp. 5956-5964, 2022, doi: 10.1109/tii.2022.3199240.

[19] C. Tang, Z. Li, J. Wang, X. Liu, W. Zhang, and E. Zhu, “Unified one-step multi-view spectral clustering,” IEEE Transactions on Knowledge and Data Engineering, vol. 35, no. 6, pp. 6449-6460, 2022, doi: 10.1109/TKDE.2022.3172687.
[20] X. Chen, M. Ding, X. Wang, Y. Xin, S. Mo, and Wang Y et al, “Context autoencoder for self-supervised representation learning,” International Journal of Computer Vision, vol. 132, no. 1, pp. 208-223, 2024, doi: 10.1007/s11263-023-01852-4.
[21] C. Zhang, Y. Geng, and Z. Han, “Autoencoder in autoencoder networks,” IEEE transactions on neural networks and learning systems, vol. 35, no. 2, pp. 2263-2275, 2022, doi: 10.1109/TNNLS.2022.3189239.