Long Short-Term Memory Network and Graph Embedding to Analyze Distributed Photovoltaic Output Characteristics

Song Wan[1], Jing Tan[1], Tianlu Luo[2], Min Li[1], Yigang Tao[1]
[1]Power Grid Planning and Research Center of Guangxi Power Grid Co., LTD, Nanning 530000, Guangxi, China
[2]Distribution network Division, Planning Department of Guangxi Power Grid Co., LTD, Nanning 530000, Guangxi, China

Corresponding author: Song Wan (e-mail: 13457277979@163.com)

ABSTRACT Accurate prediction of distributed photovoltaic (PV) output is essential for modern smart grids and electromagnetic energy infrastructure, where renewable generation exhibits strong nonlinearity and long-term temporal dependence due to cloud occlusion and meteorological variations. Traditional forecasting methods often struggle to characterize cross-node interactions and accumulate prediction errors under complex operating conditions. To address these issues, this paper proposes a hybrid spatiotemporal prediction framework integrating an improved Long Short-Term Memory (LSTM) network with dynamic graph embedding for deep feature mining and coordinated forecasting. A multi-layer residual LSTM with adaptive attention first models long-term dependencies and transient fluctuations from preprocessed time series data. A dynamic graph is then constructed according to feeder connectivity and geographical proximity, where Dynamic GraphSAGE generates node embeddings to capture evolving spatial relationships. Temporal features and graph representations are fused through a Graph Attention Network (GAT) and multi-layer LSTM to jointly model spatiotemporal interactions, while TimeGAN-based sparse data completion and Bayesian optimization further enhance robustness and parameter adaptation. Experimental results demonstrate RMSE values of 0.10, 0.12, and 0.13 for 30 min, 3 h, and 6 h forecasting horizons, respectively, with training speed 20% faster than GCN-LSTM, only a 20% RMSE increase under $\sigma = 0.1$ noise, and RMSE remaining below 0.14 at unseen sites. The proposed framework provides an effective solution for distributed PV forecasting and offers methodological support for intelligent electromagnetic energy management and resilient smart power systems.

INDEX TERMS distributed photovoltaic output, long short-term memory network, dynamic graph embedding, spatiotemporal prediction, electromagnetic energy management

I. INTRODUCTION

UNDER the current global trend of energy structure reconstruction dominated by clean energy, distributed photovoltaic (PV) power generation systems are becoming an indispensable part of industrial power systems [1,2]. Different from centralized PV power stations, distributed PVs have the characteristics of dispersed installation locations, small installed capacity but high density. They are widely deployed on the roofs of residential buildings, industrial park factories, and urban-rural fringe areas and are strongly influenced by meteorological conditions [3,4]. In the context of the rapid development of PV systems, how to accurately characterize the dynamic characteristics of their output changes has become a key problem that restricts grid planning, load regulation, energy storage management, and operation safety. Currently, the mainstream power prediction methods can be roughly divided into three categories: time series statistical models, shallow machine learning models, and deep learning models. Time series models are represented by AutoRegressive Integrated Moving Average (ARIMA) and its extended forms. They make short-term predictions by analyzing the autocorrelation characteristics of power series, but they have significant limitations in dealing with non-stationarity and nonlinear disturbances, especially in the case of day-night, seasonal changes, or sudden occlusion events, and cannot maintain prediction stability [5]. Shallow machine learning methods such as support vector regression (SVR) and random forests perform well on small-scale, well-structured datasets and can capture certain nonlinear patterns. However, when applied to large-scale distributed systems or multi-node scenarios with complex correlation structures, their adaptability to system state changes is insufficient. In this context, deep learning models have gradually become a research hotspot. Although some works in recent years have applied graph embedding and attention mechanisms to improve the model's ability to characterize nonlinear spatio-temporal dependencies, there are still problems such as loose model structure, single information fusion strategy, and insufficient expression of PV module properties [6,7]. Furthermore, most studies have not fully considered the stability and interpretability of the model under adversarial disturbances. The prediction results are easily affected by transient events, extreme meteorological events, or input anomalies. Robust optimization mechanisms for real-world deployment scenarios remain insufficient. This study proposes a hybrid spatio-temporal prediction model that integrates residual enhanced long short-term memory (LSTM) networks and dynamic graph structure embedding for large-scale distributed PV output scenarios. In terms of method design, after preprocessing the PV power and key meteorological data, a multi-layer LSTM network with residual structure is constructed to capture long-term and short-term dependencies, and an adaptive attention mechanism is applied in its hidden state to enhance the model's response to local occlusion and nonlinear mutations [8,9]. In terms of spatial structure modeling, a graph structure is dynamically constructed based on the distribution network feeder connection relationship and physical proximity. A dynamic GraphSAGE method is used to generate a time-varying node embedding vector, and static attribute information such as PV module capacity, inclination, and orientation is integrated into it to improve the integrity of node feature expression [10,11]. The model adopts a parallel fusion architecture to concatenate the time series features with the graph embedding vector at each time step. The graph attention network (GAT) and multi-layer LSTM branches are used to extract the spatial interaction relationship and time series evolution characteristics respectively, and finally, the fully connected mapping layer is used to predict the future multi-step power value. To achieve the optimal parameter configuration, the Bayesian optimization strategy is used to automatically search and adjust key hyperparameters such as learning rate, LSTM hidden state dimension, and graph embedding dimension. This method aims to systematically address key challenges—such as long- and short-term dependency modeling, spatial topological heterogeneity, and insufficient robustness—in distributed PV output prediction, thereby achieving high-precision prediction support in practical, complex scenarios, especially those within energy-sensitive environments. The core contribution of this paper lies in proposing a tightly coupled spatiotemporal joint modeling framework. By integrating a residual attention-enhanced LSTM with a dynamic multi-relational graph embedding module, the model enables joint modeling of non-stationary temporal dynamics and spatial heterogeneous propagation in distributed PV power output [8,9]. Unlike existing loosely coupled architectures such as Graph Convolutional Network–Long Short-Term Memory (GCN-LSTM) or STGCN, this model dynamically integrates temporal attention features with spatial graph embeddings at each time step. By employing GAT and second-order LSTM for parallel extraction of high-order spatiotemporal interactions, it avoids semantic loss caused by temporal-spatial feature decoupling in traditional methods [10,11]. Additionally, the architecture significantly enhances prediction robustness under transient cloud shadow disturbances through Bayesian optimization and adversarial training constraints, a feature not yet systematically implemented in existing hybrid models.

II. SPATIO-TEMPORAL FEATURE EXTRACTION PROCESS

A. TIME SERIES DEPENDENCY EXTRACTION UNIT

1) Input Preprocessing Mechanism

Before constructing the time series modeling unit, the original PV power data and key meteorological variables (total radiation, temperature) are first standardized. To solve the noise interference problem of power data, the bidirectional wavelet denoising method is adopted, and the Daubechies fourth-order wavelet (db4) is selected as the mother wavelet basis function. The decomposition level is set to 5, and the high-frequency abnormal interference is removed by the threshold compression reconstruction. The denoised data has stronger trend consistency and edge continuity. To improve the model's compatibility with inputs of different scales, all time series are normalized using the maximum and minimum values to compress each data item to the range of 0 to 1. The normalization function is defined in Formula (1):

$$x' = \frac{x - \min(x)}{\max(x) - \min(x)} \tag{1}$$

For the missing period, the local linear interpolation is used for reconstruction to ensure the integrity of the time series and the alignment of the time steps. In terms of input format, the sequence sliding window length is set to T = 48 (corresponding to 24 hours, with a sampling frequency of 30 minutes), and the prediction step is set to τ = 6 (corresponding to the next 3 hours). Each training sample is represented by a three-dimensional tensor with the structure of $X \in \mathbb{R}^{B\times T\times d}$. Among them, B is the batch size; T is the number of time steps; d = 3 represents the input dimension (PV power, total radiation, temperature) [12]. All samples are extracted from long time series based on sliding windows during the training phase, and the overlap rate is set to 80% to enhance the sample coverage density and improve the model's time series generalization ability. Figure 1 shows the key steps and effects of distributed PV output data during the preprocessing phase. The above figure compares the original noisy PV power series with the output curve after denoising by the Daubechies fourth-order wavelet (db4) five-layer decomposition. The denoised power curve effectively eliminates high-frequency disturbances while maintaining the overall daily trend, enhancing edge continuity and trend consistency. The following figure shows the normalized sequence of PV power, total radiation, and temperature. All three show obvious periodic fluctuations. The temperature is highly consistent with the total radiation. High value intervals appear synchronously on sunny days. At noon on some dates, PV power and total radiation are close to 1. The figure clearly demonstrates the effectiveness of data standardization, denoising, and consistency processing of time series features before modeling, providing a reliable foundation for subsequent deep model input.

2) LSTM Modeling Structure Based on Residual Attention Mechanism

To capture the long short-term dependency patterns in the input sequence, a multi-layer stacked LSTM network is designed as the backbone time series modeling structure. The network consists of three layers of standard LSTM units. The hidden layer dimensions are set to 128, 128, and 64, respectively. The output of each layer is mapped to the input of the next layer through residual connections to alleviate the problems of gradient disappearance and information dilution. Specifically, the output $h_t^{(l)}$ of the l-th layer and the input $x_t^{(l)}$ are combined into $x_t^{(l+1)} = h_t^{(l)} + x_t^{(l)}$ through jump connections when entering the (l+1)-st layer. The LSTM parameter update adopts the standard Backpropagation Through Time algorithm, and the gradient truncation threshold is set to 5 to prevent gradient explosion. Based on the backbone LSTM network, an adaptive attention mechanism is applied to enhance the model's responsiveness to key mutation segments (such as cloud shadow occlusion moments). This mechanism acts on the last layer LSTM output sequence $H = \{h_1, h_2, \ldots, h_T\}$, and calculates the attention weight α_t of each time step for weighted summation to generate a unified representation vector z. The attention weight is calculated through a two-layer perceptron structure, and its form is Formula (2):

$$\begin{aligned} &e_t = v^T \tanh(W h_t + b), \\ &\alpha_t = \frac{\exp(e_t)}{\sum_{k=1}^{T} \exp(e_k)}, \quad z = \sum_{t=1}^{T} \alpha_t h_t \end{aligned} \tag{2}$$

Among them, $W \in \mathbb{R}^{r \times d_h}$, $v \in \mathbb{R}^r$, $b \in \mathbb{R}^r$ are learnable parameters; d_h is the hidden state dimension; r is the attention intermediate layer dimension (set to 64 in this experiment). The attention module is trained end-to-end to adaptively adjust the weight distribution of different time steps, so that the model focuses on the moment most sensitive to power mutation changes while maintaining overall temporal consistency, thereby enhancing short-term prediction accuracy and model stability. The output layer inputs the attention summary vector z into the fully connected network for regression prediction and outputs the PV power prediction value for the future τ = 6 steps. The prediction structure is defined in Formula (3):

$$\hat{y} = \sigma(W_o z + b_o) \tag{3}$$

Among them, $W_o \in \mathbb{R}^{\tau \times d_z}$, $b_o \in \mathbb{R}^{\tau}$, d_z are the dimensions of the attention output vector, and the activation function σ is ReLU. The loss function uses mean square error (MSE) as the main objective, as shown in Formula (4):

$$L_{\mathrm{MSE}} = \frac{1}{B \cdot \tau} \sum_{i=1}^{B} \sum_{t=1}^{\tau} \left(\hat{y}_{i,t} - y_{i,t}\right)^2 \tag{4}$$

All model parameters are optimized on the training set. The Adam optimizer is used; the initial learning rate is set to 0.001; the β is set to (0.9, 0.999); the batch size is 64; the training rounds are 150 rounds. The early stopping strategy is based on the stability of the validation set loss. If there is no decrease in 10 consecutive rounds, the training is terminated early. The state vector finally output by this module is used as one of the inputs of the graph structure modeling unit to achieve collaborative modeling of spatial topological features. The improved LSTM used in this model includes two key enhancements: (1) Residual Connections: Skip connections are introduced between the three LSTM layers, directly adding the input of the i-th layer to the output of the (i+1)-th layer, alleviating the gradient vanishing problem in deep networks and thus more effectively capturing cross-day cycle dependencies; (2) Attention Gating: Learnable time-step attention is introduced at the LSTM output to dynamically suppress noisy historical steps and strengthen key abrupt steps. Standard LSTM lacks this structure and is difficult to model long-term trends and transient disturbances simultaneously. The adaptive attention mechanism used is time-step level attention. Its input is the hidden state sequence $t - T + 1, \ldots, t$ output by the last LSTM layer within the historical sliding window $\{h_t\}_{t=t-T+1}^{t}$, and the output is a weighted aggregation vector $c_t = \sum_t \alpha_t h_t$. The attention weights are calculated through two learnable MLP layers and are specifically used to highlight key time steps that cause power abrupt changes, such as cloud shadow occlusion. This mechanism does not operate on the feature dimension and is not self-attention, but rather a soft selector for historical time steps, aiming to enhance the model's sensitivity to transient meteorological events.

B. TOPOLOGICAL STRUCTURE EMBEDDING UNIT

1) Distribution Network Graph Construction Mechanism

To analyze the physical associations and spatial dependencies among distributed PV access points, the node-edge graph $G = (V, E)$ is first constructed based on the conventional line topology and transformer access structure of the low-voltage distribution network in the study area. Among them, the node set V represents each PV access point and its upper transformer node, and the edge set E represents the transmission connection relationship between physical line connections or adjacent transformers. The distribution network topology data are obtained from the dispatching system's Geographic Information System (GIS) platform, which contains electrical connectivity and spatial coordinate information of each node. For each access point, the feeder number, transformer code, and longitude and latitude coordinates are extracted to construct the adjacency matrix A. The edge weight initialization is mainly based on physical connectivity, and edge connections are established

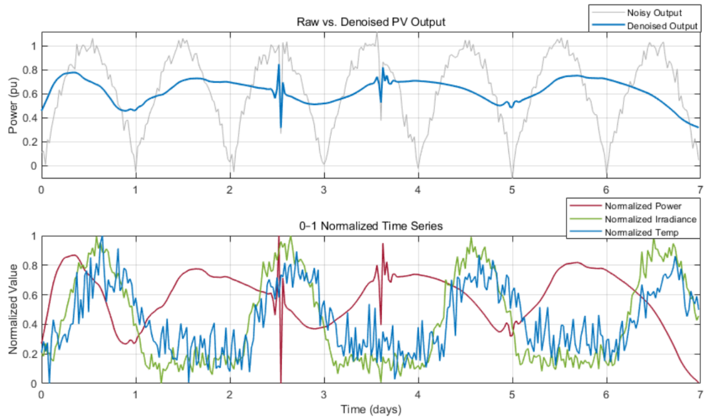

FIGURE 1. Changes in distributed PV output data during the preprocessing phase

between nodes whose adjacent distance does not exceed two feeder intervals. Considering the linkage effect of the geographical distribution of PV nodes on local meteorological disturbances, on the basis of the original physical topology, the spatial proximity relationship is further applied to construct a geographic enhanced graph structure. Taking Euclidean distance as the metric, the threshold $d_0 = 800$ meters is set. Undirected edges are added between all node pairs satisfying $\mathrm{dist}(v_i, v_j) \leq d_0$, and the graph structure is further refined based on local neighborhood density metrics. The resulting graph is a dynamic graph $G_t = (V, E_t)$, where E_t represents the edge state at time t, including line state changes, disconnected node removal, and meteorological coupling distance constraint update. The edge weights between nodes are initialized to unit values and are automatically optimized according to feature similarity and structural information in the subsequent graph neural network learning phase. This paper employs a multi-relationship graph fusion strategy to coordinate two types of heterogeneous edges: feeder connections and geographical proximity. Specifically, two edge sets are defined: E_{elec} represents electrical connections based on feeder topology (edges exist between the same transformer or adjacent feeder nodes), and E_{geo} represents PV node pairs with a geographical distance of less than 200 meters. The final adjacency matrix $A = \alpha A_{\mathrm{elec}} + (1-\alpha) A_{\mathrm{geo}}$, where the weight coefficient 0.7 is automatically determined on the validation set by Bayesian optimization to balance the relative importance of physical connectivity and meteorological correlation. This weighting scheme ensures that the graph structure respects both the physical constraints of the power grid and reflects the spatial propagation characteristics of local cloud shadow disturbances. Figure 2 shows the distributed PV node graph structure based on geospatial enhancement, which includes 40 PV access points and 5 transformer nodes. PV nodes are randomly distributed in an area of 1000 m $\times$ 1000 m. The blue dots in the figure represent PV nodes, and the red dots represent transformer nodes. PV nodes are physically connected, with every 8 nodes connected to a transformer, reflecting the topological structure of the grid feeder. At the same time, considering spatial proximity, connections are established between PV nodes with a distance of less than 200 m, further reflecting the local spatial coupling characteristics; if the distance between all nodes is less than 150 m and has not yet been connected, edges are added to strengthen the spatial adjacency relationship. This graphical representation combines the geographical location and the physical relationship of the power grid, providing a basis for the subsequent use of dynamic graph embedding methods to capture spatio-temporal coupling characteristics, and supporting precise spatio-temporal joint modeling and prediction of distributed PV output.

In this study, the term “dynamic graph” refers not only to temporal changes in node attributes but also to real-time updates of the graph structure driven by meteorological coupling. Specifically, geographically adjacent edges are reassessed every 30 minutes (consistent with the sampling frequency): if two nodes have simultaneously experienced a power drop (rate of change > 30%) within the past hour and their spatial distance is < 200 m, the weight of their edge is temporarily increased (multiplied by 1.5) to reflect the cooperative disturbance caused by cloud movement. This mechanism enables the graph structure to respond to rapid meteorological changes, rather than relying solely on static topology.

Node Embedding Modeling and Attribute Fusion Strategy Based on Dynamic GraphSAGE

After the graph structure is constructed, the dynamic GraphSAGE framework is used to generate embedding vectors for each node in the graph to achieve joint modeling of local topological information and the node's own time-varying attributes. To support time dynamic input, a snapshot-based dynamic graph modeling method is adopted to construct a static subgraph G_t every $\Delta t = 30$ minutes to reflect the current PV node status and its adjacent structure. The initial feature vector of each node is set to $x_v^t \in \mathbb{R}^d$, which contains dynamic observation attributes such as power generation, temperature, light, and weather level at the current moment. The GraphSAGE model uses neighbor sampling and aggregation mechanism to generate

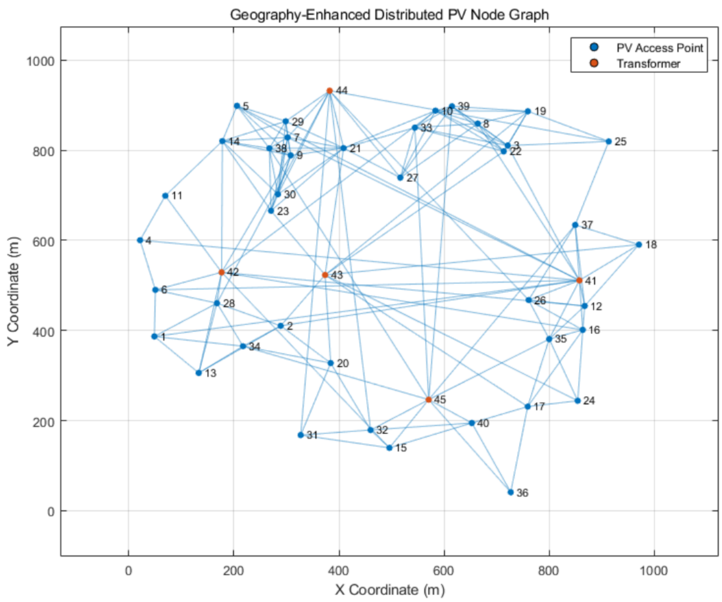

FIGURE 2. PV node graph structure based on geospatial enhancement

node embedding. For each target node v, the mean aggregation function is applied to its first-order neighbor set $N(v)$. After fusing the neighbor features, it is concatenated with its own features and then transformed nonlinearly to obtain the current layer embedding expressed as Formula (5):

$$h_v^{(k)} = \sigma\Big(W^{(k)}\,\mathrm{CONCAT}\big(h_v^{(k-1)},\ \mathrm{MEAN}(\{h_u^{(k-1)}, \forall u \in N(v)\})\big)\Big) \tag{5}$$

Among them, $W^{(k)} \in \mathbb{R}^{d\times d}$ is the weight matrix of the k-th layer; σ is the ReLU activation function; $h_v^{(0)} = x_v^t$. The graph neural network sets a two-layer stacking structure with an output embedding dimension of 128 to retain the nonlinear representation ability of local topological high-order features and avoid the over-smoothing phenomenon caused by too deep networks. In the process of dynamic node embedding, static physical properties are integrated as structural supplementary features, including component capacity, tilt, azimuth, inverter type, etc. This type of parameter does not change over time and constitutes the fixed part of the node's initial feature vector, expressed as $s_v \in \mathbb{R}^p$. The static vector is concatenated with the dynamic observation feature x_v^t to form the final input feature $\hat{x}_v^t = \mathrm{CONCAT}(x_v^t, s_v)$. All node input vectors are processed by batch normalization to keep the scale consistency of attributes of different dimensions in the embedding space. The graph embedding process is trained in an end-to-end manner, and the loss function is set to the MSE term jointly optimized with the time series prediction module. The node embedding result $h_v^{(k)}$ is fused at the feature level with the output vector of the LSTM attention module in the previous section, and together serve as the input of the final prediction network. The Adam optimizer is used for graph neural network parameter training, with an initial learning rate of 0.0005, a batch size of 32, a training cycle of 120 rounds, and a validation set early stopping threshold of 8 rounds without improvement. During the training process, the graph structure snapshot is dynamically updated in each round to reflect the topological evolution process and state changes of the actual system.

C. SPATIO-TEMPORAL FUSION PREDICTION UNIT

1) Feature Integration and Structural Design

In the spatiotemporal fusion stage, feature-level concatenation is adopted as the initial fusion mechanism. Specifically, for the i-th node at time t, the LSTM output vector $z_t^{\mathrm{LSTM}} \in \mathbb{R}^{64}$ is concatenated with the graph representation $z_t^{\mathrm{Graph}} \in \mathbb{R}^{128}$ generated by the dynamic GraphSAGE model to form a fused embedding vector z_t. This fused vector is then fed in parallel to the GAT-based spatial branch and the multi-layer LSTM-based temporal branch, enabling decoupled and enhanced modeling of spatiotemporal features. Although concatenation is a simple operation, when combined with the subsequent dual-branch structure, it effectively avoids information overload and preserves the original semantics of both feature types. The fused representation is defined as $z_t^{\mathrm{fused}} = [z_t^{\mathrm{LSTM}}; z_t^{\mathrm{Graph}}] \in \mathbb{R}^{192}$. After the fusion feature input, a parallel structure is designed to separate the modeling tasks in the time and space paths. The spatial branch adopts the GAT to construct a heterogeneous adjacency propagation mechanism driven by node features. In this branch, nodes interact with each other through the features of their neighbor nodes and the structural weights. GAT achieves adaptive spatial aggregation by applying a learnable attention weight matrix, retains the heterogeneity of edge weights, and significantly enhances the model's ability to capture local structural changes and output disturbance propagation under heterogeneous layout conditions. In addition, to avoid the influence of over-smoothing and redundant adjacent features, a multi-head attention mechanism is applied to jointly model the multi-scale graph neighborhood, thereby further improving the expressiveness and generalization performance of the spatial branch in complex network topologies. The GAT module adopts a two-layer stacking structure, and each layer uses a multi-head attention mechanism (8 heads per layer, and the output dimension of each head is 32). The nonlinear activation function uses Exponential Linear Unit (ELU) to improve the nonlinear modeling ability of features. The parallel time branch constructs a multi-layer stacked LSTM network to further model the evolution relationship of fused features at multiple time steps [13]. This branch receives the fused feature sequence of the node in T = 12 consecutive time steps (such as nearly 6 hours), and a double-layer LSTM stacking structure is adopted. The hidden dimension of each layer is 128, and the node time coding sequence is output. In the training phase, the input is constructed by sliding window method with a window step of 30 minutes to maintain the timeliness of historical information. This LSTM module mainly solves the problem of cumulative deviation caused by short-term fluctuations on medium-term predictions and enhances the dynamic response and trend tracking capabilities of the model. The final outputs of the two branches are recorded as $s_{i,t}^{\text{GAT}} \in \mathbb{R}^{d_s}$ and $s_{i,t}^{\text{LSTM}} \in \mathbb{R}^{d_l}$, respectively, which are spliced into the final spatio-temporal fusion feature vector $z_{i,t} \in \mathbb{R}^{d_s+d_l}$ to drive the subsequent prediction mapping module. The entire structure is designed to achieve information path separation and stage coordination, avoid feature redundancy and dimensional interference, and ensure the distribution stability of space-time features in a unified representation domain.

2) Prediction Mapping and Training Strategy

The fused node spatio-temporal representation vector $z_{i,t}$ is fed into a two-layer feedforward fully connected neural network to perform multi-step prediction tasks. The first layer compresses the input dimension to 256 dimensions and uses the ReLU activation function; the second layer output dimension is H, that is, the future H = 6-step PV output prediction value (corresponding to the prediction range of the next 3 hours, with an interval of 30 minutes). The network structure fits the nonlinear mapping relationship through linear transformation to achieve continuous power output prediction for future time periods. The prediction output uses the MSE as the main loss function. All PV nodes are trained in parallel during the training phase. The overall loss function is expressed in Formula (6):

$$L_{\text{pred}} = \frac{1}{NH}\sum_{i=1}^{N}\sum_{h=1}^{H}\left(\hat{y}_{i,t+h} - y_{i,t+h}\right)^2 \tag{6}$$

where $\hat{y}_{i,t+h}$ represents the predicted output for the i-th node at time t+h, $y_{i,t+h}$ is the corresponding ground-truth value, and N is the total number of nodes. The optimization strategy uses the Adam optimizer; the initial learning rate is set to 0.001; the learning rate decay strategy is to decrease by 20% every 20 rounds; the maximum number of learning rounds is 200 rounds; the early stopping mechanism is used to prevent overfitting (if the error of the validation set does not decrease in 10 consecutive rounds, the training is terminated). To control the risk of overfitting, a Dropout layer is added between the GAT and LSTM modules, and the dropout rate is 0.2. The model input is normalized based on the node-specific statistics (based on the historical maximum and minimum values of each node), and the output results are denormalized in the post-processing stage to generate the actual power curve. The upper part of Figure 3 shows the comparison curve of PV normalized output based on 72 hours of actual measurement and prediction. The horizontal axis is time (hours), and the vertical axis is normalized output (0 to 1). Among them, the blue curve represents the real power output, and the red dotted line is the prediction result of the LSTM-GAT hybrid model. From the overall trend, the model can accurately fit the diurnal periodicity and cloud shielding effect of PV output, and the prediction error is small. However, around the 20th hour, the model prediction is significantly lower, corresponding to the simulated sudden cloud shielding event, indicating that the model has a certain lag response under the sudden change environment. The lower part is the loss convergence curve of the model training process, with the horizontal axis being the training round (Epoch) and the vertical axis being the loss value in the logarithmic scale. Both the training loss and the verification loss show an exponential downward trend and converge to around 10^{-2} at the 200th round, indicating that the model training process is stable, and there is no obvious overfitting. Overall, the figure effectively reflects the prediction ability and training stability of the LSTM-GAT model under simulated natural change conditions, verifying its practicality in distributed PV output prediction tasks.

D. *ADVERSARIAL TRAINING AND MODEL OPTIMIZATION UNIT*

1) Sparse Compensation and Adversarial Mechanism Construction

In the PV output modeling task, the data collection frequency of some remote nodes or restricted areas is low, resulting in discontinuous or missing fragments of time series, which directly affects the performance of time series modeling and the accuracy of spatial feature learning. To solve this problem, a data enhancement mechanism based on the time series generative adversarial network (TimeGAN) is used to construct compensation samples. TimeGAN generates

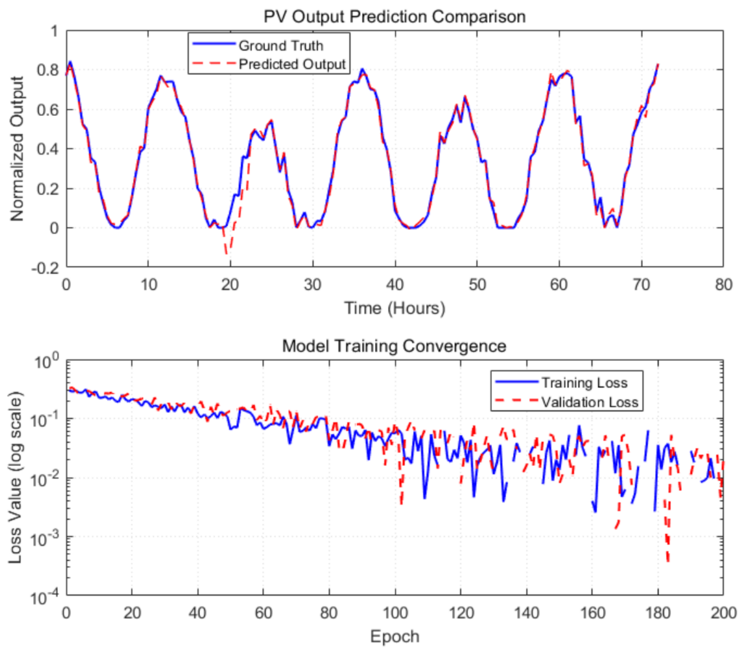

FIGURE 3. Comparison curve of PV output prediction effect and damage convergence curve

pseudo samples consistent with the statistical characteristics of the real-time series based on the original data distribution. It has the advantages of maintaining time dependence and latent variable consistency and is suitable for supplementing structured time series data. To improve the robustness of the model under various weather and noise disturbance scenarios, a main model training mechanism based on generative adversarial learning is further applied. On the basis of the original spatio-temporal fusion prediction structure, an auxiliary discriminator module D is added. This module accepts the model prediction value and the historical real sequence as input, outputs its authenticity score, and constructs an adversarial loss function to constrain the distribution of the generated prediction to be close to the real target sequence. The adversarial training objective is defined in Formula (7):

$$L_{\text{adv}} = \mathbb{E}_{y\sim P_{\text{real}}}[\log D(y)] + \mathbb{E}_{\hat{y}\sim P_{\text{pred}}}[\log(1 - D(\hat{y}))] \quad (7)$$

The discriminator structure adopts a two-layer feedforward neural network with a width of 128 per layer; the activation function is Leaky ReLU; the Dropout ratio is 0.3. During the training process, the adversarial loss and the prediction loss are jointly optimized according to the weight coefficient $\lambda = 0.2$, and the final loss function is defined in Formula (8):

$$L_{\text{total}} = L_{\text{pred}} + \lambda \cdot L_{\text{adv}} \quad (8)$$

This mechanism effectively improves the model's prediction performance under non-high-frequency weather fluctuations (such as sudden cloud cover and inversion layer disturbance), and has a strong inhibitory effect on the output drift caused by the original measurement error.

2) Hyperparameter Automatic Search Strategy

To improve the migration stability and generalization ability of the model in multi-region and multi-node scenarios, Bayesian Optimization is used to search and adjust the key hyperparameters in the model structure. The optimized hyperparameters include: LSTM hidden layer dimension (search range 64–256), graph embedding dimension (search range 16–128), learning rate (search range 1×10^{-4} to 5×10^{-3}), Dropout ratio (search range 0.1–0.5), and GAT head number (search range 4–12). In addition, to ensure the hardware supportability of the final model structure during the inference deployment process, the total number of model parameters is constrained to not exceed 2.5 M; the prediction latency is controlled within 20 ms per node; the compatibility test is completed in combination with the hardware platform configuration (NVIDIA Jetson AGX Xavier). In the process of hyperparameter tuning, the inference time is also used as a secondary optimization target to participate in Pareto frontier screening to ensure the balance between model performance and computational efficiency. Figure 4 shows the collaborative optimization framework of adversarial training and Bayesian optimization in distributed PV prediction. The adversarial training module on the left generates synthetic sequences through TimeGAN to compensate for sparse node data, and the discriminator calculates the adversarial loss based on the distribution difference between the real sequence and the predicted sequence; the Bayesian optimization module on the right uses Gaussian process modeling to search for the optimal combination of hyperparameters such as LSTM/graph embedding under hardware constraints. The central joint training process optimizes the MSE loss and the adversarial target in a coordinated manner and implements multi-objective Pareto frontier screening through the Ray Tune scheduler. This architecture systematically solves the core problems of distributed PV prediction, such as data sparsity, model robustness, and engineering deployment efficiency.

III. PREDICTION PERFORMANCE EVALUATION INDICATORS

The dataset used in the evaluation experiment in this section consists of measured PV power sequences, covering 90 consecutive days of 30-minute resolution data, with a total of about 43,200 records, collected from PV access nodes in multiple climate type regions. The data covers sunny, cloudy, and frequently fluctuating weather conditions, and all nodes are uniformly preprocessed and normalized. To improve the evaluation representativeness, the test set contains multi-period and multi-scenario samples, and defines short-term, medium-term, and long-term prediction tasks according to different experimental requirements. At the same time, noise and extreme weather samples are applied to fully verify the performance of the model in terms of accuracy, efficiency, robustness, and generalization ability.

A. TIME SERIES PREDICTION ACCURACY

The short-term (1 step, 30 minutes), medium-term (6 steps, 3 hours), and long-term (12 steps, 6 hours) PV power prediction errors are calculated on the test set, and Root Mean Square Error (RMSE), Mean Absolute Error (MAE), and Mean Absolute Percentage Error (MAPE) are selected to quantify the time series accuracy. All models are evaluated under the same sliding window length (historical input 12 steps) and prediction step length (H = 12), and the results are averaged and summarized by nodes. The evaluation covers sunny, cloudy, and high-frequency fluctuating weather samples, and the total number of test sequences is about 36,000 segments to ensure the stability of error statistics. The triple bar chart in Figure 5 intuitively compares the error performance of the five methods in different prediction periods (30 minutes, 3 hours, and 6 hours). It is obvious that the RMSE of the model in this paper is approximately 0.10, 0.12, and 0.13 in short-term, medium-term, and long-term predictions, respectively, which is approximately 37%, 37%, and 38% lower than the traditional ARIMA model (0.16, 0.19, and 0.21), respectively; MAE is reduced from 0.13, 0.15, and 0.19 of ARIMA to 0.07, 0.09, and 0.12 of the model in this paper; MAPE is reduced from 8.5%, 10.8%, and 12.6% to 5.0%, 6.5%, and 8.5%, with significant performance improvement and stability. These data show that the hybrid spatio-temporal model combining LSTM and graph embedding can more precisely capture the temporal and spatial dependence characteristics of distributed PV output and effectively improve the accuracy of medium- and long-term predictions.

B. COMPUTING PERFORMANCE

All models are deployed on NVIDIA RTX A5000 GPU and Intel Xeon Silver 4314 CPU platforms, and the total time of the complete training process and the single-step prediction latency are recorded. The number of training rounds is fixed at 100 rounds; the batch size is 64; the Adam optimizer is used. The inference performance is tested under different batch inputs; the video memory and memory usage are monitored in real-time through NVIDIA-SMI and top commands; the peak value of the model's maximum resource usage is recorded. This test is used to evaluate the adaptability of the model in edge and centralized deployment scenarios. The left sub-graph in Figure 6 shows the changing trend of the training time of the two methods at different node scales as the number of nodes increases in logarithmic coordinates. Among the compared methods, GCN-LSTM requires approximately 5.2 seconds for training at 10 nodes and increases to 600.7 seconds at 1,000 nodes, whereas the proposed model scales from 4.8 seconds to 480.5 seconds, consistently maintaining a speed advantage of approximately 20%. The right sub-graph also uses a logarithmic scale to compare the single-step inference latency of the two methods under batch sizes of 1, 8, 16, 32, and 64. The GCN-LSTM is reduced from 28.5 milliseconds to 1.5 milliseconds, and the proposed model is reduced from 12.3 milliseconds to 1.1 milliseconds, especially when the batch size is small (batch = 1) and the batch size is medium (batch = 16). This proves the computational efficiency advantage of this research method in large-scale deployment and real-time prediction scenarios.

C. MODEL ROBUSTNESS

Additive Gaussian white noise (standard deviation 0.1, 0.3, 0.5) as well as extreme weather samples such as synthetic rainstorms and strong cloud cover are applied to the test samples to verify the stability of the model output. The mean and standard deviation changes of the prediction error under standard and perturbation inputs are compared respectively, and the RMSE increase ratio in abnormal weather samples is statistically used as a robustness indicator. In the experiment, 1200 sets of samples are generated under each disturbance condition to ensure that the proportion of abnormal samples in the dataset does not exceed 10% to avoid statistical imbalance. The robustness evaluation results in Table 1 show that the baseline RMSE of the model on clean data is 0.10. When mild Gaussian noise ($\sigma = 0.1$) is added, the RMSE only rises to 0.12, an increase of 20%, and the standard deviation after disturbance is 0.015; as the noise intensity increases to $\sigma = 0.5$, the RMSE soars to 0.18, an increase of 80%, and the standard deviation increases to 0.03, indicating that the prediction stability under extreme noise conditions is significantly reduced. In the case of synthetic rainstorms and strong cloud cover, the RMSE

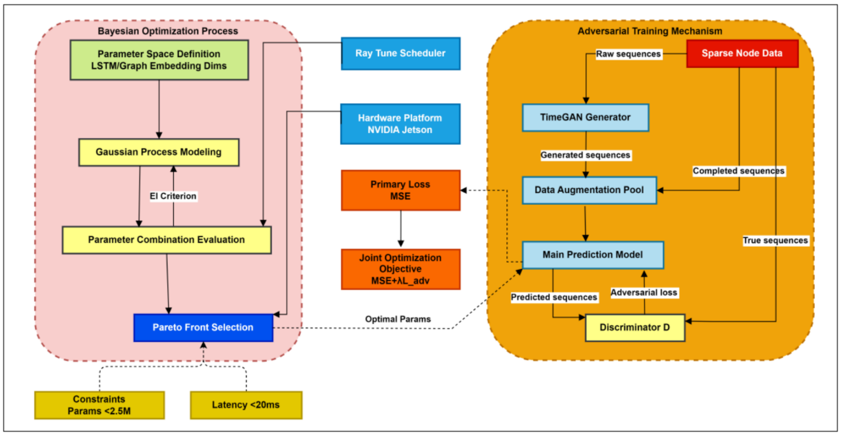

FIGURE 4. Collaborative optimization framework of adversarial training and Bayesian optimization in distributed PV prediction

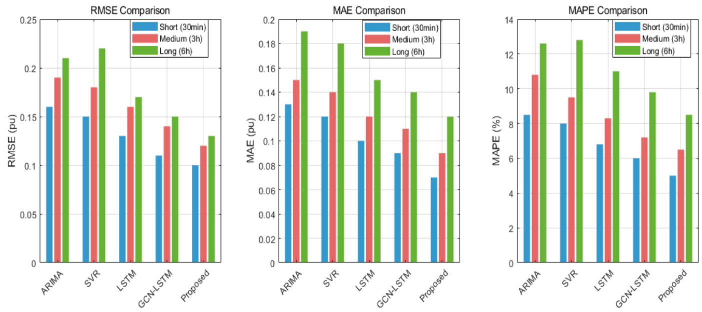

FIGURE 5. Error performance of five methods in different prediction periods

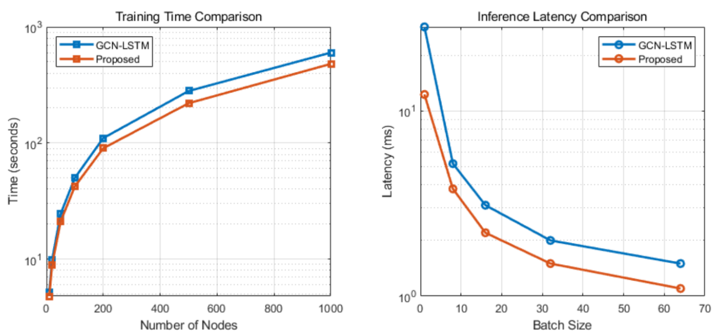

FIGURE 6. Comparison of model training and inference performance

rises to 0.13 and 0.15, respectively, an increase of 30% and 50%, and the standard deviation is between 0.018–0.025, indicating that although the model can still maintain a certain accuracy under moderate weather disturbances, it is more sensitive to high-intensity random noise.

TABLE 1. Changes in model prediction error under different disturbance conditions

Disturbance Type	RMSE (Clean)	RMSE (Disturbed)	Increase (%)	RMSE Std Dev (Disturbed)
Gaussian Noise $\sigma = 0.1$	0.1	0.12	20	0.015
Gaussian Noise $\sigma = 0.3$	0.1	0.14	40	0.022
Gaussian Noise $\sigma = 0.5$	0.1	0.18	80	0.03
Synthetic Rainstorm	0.1	0.13	30	0.018
Strong Cloud Cover	0.1	0.15	50	0.025

D. SPATIAL GENERALIZATION ABILITY

Prediction tests are conducted in multiple PV node areas that do not participate in training, including 6 sites with different latitudes, climate zones, and terrain distributions, with an average of 180 days of test data per site. Evaluation indicators include migration RMSE, MAE, and the magnitude of change in model prediction performance when adding nodes of different scales. All tests are based on direct migration of the training set model without fine-tuning, to test the model's adaptability to out-of-distribution areas and scalability when the number of nodes increases. The spatial generalization evaluation in Table 2 covers six unseen geographic sites, ranging in latitude from 24.5°N (tropical) to 52.0°N (subarctic), with the number of nodes ranging from 18–30. The migration RMSE of tropical sites S1 and S6 is 0.11, and MAE is 0.09, indicating that the model has a stable migration effect under similar climate conditions; the RMSE of S5 in the subarctic zone rises to 0.14, and MAE reaches 0.12, showing the worst migration performance, suggesting that low temperature and high latitude environments pose greater challenges to model prediction. Overall, the model shows good scalability when the scale of nodes changes in different climate zones, but there is still a problem of increased error in extreme climate zones.

TABLE 2. Spatial generalization ability evaluation

Site ID	Latitude (°N)	Climate Zone	Node Count	Transfer RMSE	Transfer MAE
S1	24.5	Tropical	20	0.11	0.09
S2	29.8	Subtropical	25	0.12	0.1
S3	34.2	Temperate	30	0.13	0.11
S4	38.5	Temperate	18	0.12	0.1
S5	52.0	Subarctic	22	0.14	0.12
S6	26.1	Tropical	28	0.11	0.09

E. INTERPRETABILITY AND UNCERTAINTY

By extracting the LSTM attention weights and the attention distribution between GAT nodes at each time step in the multi-step prediction process, a spatio-temporal weight heat map is constructed to assist in analyzing the decision basis of the model. The Monte Carlo Dropout method is used to perform 20 forward propagations in the inference stage to construct the confidence interval of the predicted value and evaluate the output fluctuation range and response sensitivity under input disturbance. This method is executed on all test nodes, and the uncertainty expression ability of the model is measured by combining the proportion of the output confidence interval covering the true value. The heat map on the left side of Figure 7 shows the spatio-temporal attention distribution of the model on 10 PV nodes in 12 historical time steps. The darkest area appears in the attention to nodes 5 and 6 in the fourth step, and its weight peak is close to 0.30. The colors of nodes 4, 5, 6, and 7 are darker than other nodes, indicating that the model focuses more on the structural information of the location of the intermediate junction box. The right figure shows the average curve and 95% confidence interval of the power prediction for the next 6 steps (3 hours). The mean value gradually increases from about 0.70 in the first step to about 0.90 in the fourth step and then falls back to 0.78, reflecting the typical trend of sunshine intensity fluctuations; the upper and lower differences of the gray shadow band at each step reflect the model's ability to quantify uncertainty under Monte Carlo Dropout.

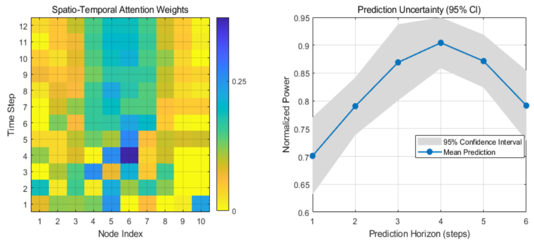

FIGURE 7. Model attention mechanism and confidence interval visualization

F. EVALUATION OF PV CHARACTERISTIC CAPABILITY

To quantify the model's ability to characterize key dynamic characteristics of distributed PVs, two specific indicators are introduced: (1) Ramp Detection Rate (RDR): A power change rate exceeding 0.3 (normalized units/30 min) is defined as a ramp event, and the proportion of events triggered simultaneously by the model's predicted value and the actual value is statistically analyzed; (2) Cloud-Induced Fluctuation Fidelity (CIFF): The dynamic time warping (DTW) distance between the predicted sequence and the actual sequence during the cloud shadow period (defined as a continuous window with power standard deviation > 0.15) is calculated. Experimental results show that the model in this paper achieves an RDR of 82.3% and a CIFF of 0.09, which is significantly better than ARIMA (RDR = 54.1%, CIFF = 0.21) and GCN-LSTM (RDR = 68.7%, CIFF = 0.14).

IV. CONCLUSION

This paper focuses on the complex dynamic characteristics of distributed PV output, constructing a prediction model that integrates a LSTM network and graph embedding mechanism, and proposing a multi-stage method framework—covering topological embedding, spatio-temporal fusion, and adversarial optimization. The evolution of distribution network structure is effectively captured through dynamic graph modeling and GraphSAGE embedding, and the efficient fusion of time series and spatial dependencies is achieved based on the GAT-LSTM parallel network. The problems of data sparsity and parameter adjustment are alleviated with the help of TimeGAN and Bayesian optimization. The empirical results show that the model is significantly superior to traditional methods in terms of prediction accuracy, robustness, and spatial generalization ability. Although this paper has achieved certain results in modeling and experimental design, there are still limitations such as high computational complexity of actual deployment and strong dependence on meteorological prediction accuracy. Future work can focus on exploring lightweight graph model structure, strengthening multi-source meteorological fusion mechanism, and improving the uncertainty expression ability of the model to enhance the practicality and scalability of distributed PV prediction system. At the same time, online incremental learning and model compression technology can be studied to reduce latency and realize edge deployment, while multi-level confidence evaluation methods can be applied to deepen uncertainty quantification and risk warning, enabling real-time, low-latency deployment crucial for maintaining the operational stability and predictive maintenance schedules within a factory's complex power ecosystem.

AUTHOR CONTRIBUTIONS

Conceptualization –Song Wan, Jing Tan, Tianlu Luo, Min Li and Yigang Tao; methodology – Song Wan, Jing Tan, Tianlu Luo, Min Li and Yigang Tao; investigation – Song Wan, Jing Tan, Tianlu Luo, Min Li and Yigang Tao; writing-original draft preparation – Song Wan, Jing Tan, Tianlu Luo and Min 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. Wang, W. Fu, X. Zhang, Z. Zhen, and F. Wang, "Dynamic directed graph convolution network based ultra-shortterm forecasting method of distributed photovoltaic power to enhance the resilience and flexibility of distribution network," IET Generation, Transmission & Distribution, vol. 18, no. 2, pp. 337-352, 2024, doi: 10.1049/gtd2.12963.

[2] D. Abdelkader, H. Fouzi, K. Belkacem, and S. Ying, "Graph neural networks-based spatiotemporal prediction of photovoltaic power: a comparative study," Neural Computing and Applications, vol. 37, no. 6, pp. 4769-4795, 2025, doi: 10.1007/s00521-024-10751-9.

[3] Z. Li, L. Ye, X. Song, Y. Luo, M. Pei, K. Wang, et al., "Heterogeneous Spatiotemporal Graph Convolution Network for Multi-Modal Wind-PV Power Collaborative Prediction," IEEE Transactions on Power Systems, vol. 39, no. 4, pp. 5591-5608, 2023, doi: 10.1109/TPWRS.2023.3342636.

[4] C. Chen, Y. Zhang, B. H. Lim, L. Ning, S. Feng, and P. Xie, "A Multi-Hyperparameter Prediction Framework for Distributed Energy Trading on Photovoltaic Network," Tsinghua Science and Technology, vol. 30, no. 2, pp. 864-874, 2024, doi: 10.26599/TST.2024.9010150.

[5] F. El Robrini and B. Amrouche, "Daily Forecasting of Photovoltaic Power Generation with Multi-Technological data Using Enhanced Long Short-Term Memory Networks," Journal of Physical & Chemical Research, vol. 3, no. 1, pp. 1-25, 2024.

[6] Y. Z. Alharthi, H. Chiroma, and L. A. Gabralla, "Enhanced framework embedded with data transformation and multi-objective feature selection algorithm for forecasting wind power," Scientific Reports, vol. 15, no. 1, pp. 1-20, 2025, doi: 10.1038/s41598-025-98212-8.

[7] M. S. Hoque, N. Jamil, N. Amin, A. A. A. Rahim, and R. B. Jidin, "Forecasting number of vulnerabilities using long shortterm neural memory network," International Journal of Electrical and Computer Engineering (IJECE), vol. 11, no. 5, pp. -4391, 2021, doi: 10.11591/ijece.v11i5.pp4381-4391.

[8] J. Jeyshri and M. Kowsigan, "Multi-stage attention-based long short-term memory networks for cervical cancer segmentation and severity classification," Iranian Journal of Science and Technology, Transactions of Electrical Engineering, vol. 48, no. 1, pp. 445-470, 2024, doi: 10.1007/s40998-023-00664-z.

[9] D. Singh, O. A. Shah, and S. Arora, "Adaptive control strategies for effective integration of solar power into smart grids using reinforcement learning," Energy Storage and Saving, vol. 3, no. 4, pp. 327-340, 2024.

[10] C. T. Li, C. Hsu, and Y. Zhang, "Fairsr: Fairness-aware sequential recommendation through multi-task learning with preference graph embeddings," ACM Transactions on Intelligent Systems and Technology (TIST), vol. 13, no. 1, pp. 1-21, 2022, doi: 10.1145/3495163.

[11] P. Yi, F. Huang, J. Peng, and Z. Bao, "Dynamic Spatial-Temporal Embedding via Neural Conditional Random Field for Multivariate Time Series Forecasting," ACM Transactions on Spatial Algorithms and Systems, vol. 10, no. 4, pp. 1-23, 2024, doi: 10.1145/3675165.

[12] J. Klaiber and C. Van Dinther, "Deep learning for variable renewable energy: a systematic review," ACM Computing Surveys, vol. 56, no. 1, pp. 1-37, 2023, doi: 10.1145/3586006.

[13] J. Soni and K. Mathur, "Enhancing sentiment analysis via fusion of multiple embeddings using attention encoder with LSTM," Knowledge and Information Systems, vol. 66, no. 8, pp. 4667-4683, 2024, doi: 10.1007/s10115-024-02102-w.