Results in Engineering 28 (2025) 107369

Contents lists available at ScienceDirect

Results in Engineering

journal homepage: www.sciencedirect.com/journal/results-in-engineering

Research paper

Interpretable prediction of multi-photovoltaic power stations via spatial-temporal multi-task learning with Transformer-XLSTM

Rongquan Zhang [a,b], Xiupeng Wan [a,c,d], Siqi Bu [b,*], Min Zhou [c], Qiangqiang Zeng [c], Zhe Zhang [c,*]

[a] College of Transportation, Nanchang JiaoTong Institute, Nanchang, 330044, China
[b] Department of Electrical and Electronic Engineering, Hong Kong Polytechnic University, 999077, Hong Kong
[c] College of Engineering Physics, Shenzhen Technology University, Shenzhen, 518118, China
[d] School of Information and Control Engineering, Liaoning Petrochemical University, Fushun, 113000, China

ARTICLE INFO

Keywords:
Photovoltaic power
Multi-task explainable forecasting
Multi-task anomaly detection
Spatial-temporal network
Extended long short-term memory
Transformer

ABSTRACT

With numerous distributed photovoltaic (PV) power stations integrated into the electrical energy system, accurate power forecasting for multiple PV stations is essential to ensure grid stability. However, due to the intricate correlations and spatiotemporal dynamics among PV stations, prevailing methods struggle to capture and interpret cross-station interactions induced by meteorological variations and distribution. For this purpose, a novel interpretable prediction approach using spatial-temporal multi-task learning with Transformer-extended long short-term memory (XLSTM) for multi-PV power stations is proposed in this paper. First, a spatial feature extractor combining rotary position embedding, Transformer, dilated causal convolutional, and residual connection is presented to model PV power global-local features (meteorological distribution), and a temporal feature extractor based on the XLSTM with global attention mechanisms is designed to effectively capture long-term dependencies and critical temporal features (meteorological variations). Then, a multi-task prediction model based on spatial-temporal feature extractors is proposed for learning the coupling relationships for PV stations. In addition, to ensure training data quality, an isolation forest-based multi-task outlier detection model is incorporated into the forecasting approach. Finally, the Shapley additive explanations model is utilized to elucidate the relationships between the coupled features and multi-task outputs. Simulation experiments are conducted using operational data from PV stations in western China. Compared to 17 advanced benchmarks, the proposed approach achieves a root mean square error reduction in PV power prediction, with an average value of 91.63%. The experimental results also demonstrate that the proposed approach can interpret the importance of relevant features in multi-task outputs.

1. Introduction

1.1. Background

Since the dawn of the 21st century, rapid global economic development has driven sustained growth in electricity demand worldwide [1]. However, this growth has led to excessive extraction, exploitation, and utilization of fossil fuels, resulting in the energy crisis and escalating environmental pollution [2]. To address this global challenge of environmental pollution, the international community has reached a consensus on the concentrated implementation of the Paris Agreement. For instance, China announced at the 75th UN General Assembly its strategic objective to achieve carbon neutrality by 2060 [3]. To achieve the Paris Agreement goals and mitigate the energy crisis, numerous countries and organizations worldwide are actively engaged in developing sustainable, secure, and reliable new energy sources. Photovoltaic (PV) power has become the environmentally optimal renewable energy source today due to its renewable nature, zero-emission characteristics, and noiseless operation. These advantages have accelerated its integration into modern electrical energy systems [4]. However, due to its susceptibility to geographical location and meteorological conditions, PV power exhibits significant characteristics of randomness, in-

* Corresponding author.
E-mail addresses: zhangrq19931102@163.com (R. Zhang), 15717084766@163.com (X. Wan), siqi.bu@polyu.edu.hk (S. Bu), minzhou2020@163.com (M. Zhou), qiangqiang9672@163.com (Q. Zeng), zhangzhe@sztu.edu.cn (Z. Zhang).

https://doi.org/10.1016/j.rineng.2025.107369
Received 12 July 2025; Received in revised form 10 September 2025; Accepted 18 September 2025

Available online 1 October 2025

Nomenclature

PV	Photovoltaic
LSTM	Long short-term memory
XLSTM	Extended LSTM
SVR	Support vector regression
XGBoost	Extreme gradient boosting
CNN	Convolutional neural networks
RNN	Recurrent neural networks
SLSTM	Scalar LSTM
MLSTM	Matrix LSTM
AM	Attention mechanism
RoPE	Rotary position embedding
DCC	Dilated causal convolutional
DBSCAN	Density-based spatial clustering with noise
IF	Isolated forest
IFMD	IF-based multi-task outlier detection
MAE	Mean absolute error
RMSE	Root mean square error
OCSVM	One-class support vector machine
SHAP	Shapley additive explanations
ML	Multi-task learning model
SL	Single-task learning model
MA	Multi-head attention mechanism
GA	Global attention mechanism
Proposed_SL	The proposed model with SL
DT+RG	DCC+Transformer+RNN+GA
DT+LG	DCC+Transformer+LSTM+GA
CNNXLSTM	CNN model with XLSTM

termittency, and uncertainty in power output [5,6]. These features pose significant technical challenges to various aspects of electrical energy systems after grid integration, including electricity market operations, voltage stability maintenance, frequency regulation, grid security operations, and power system planning [7,8]. To reduce the negative effects of PV power on electrical energy systems, there is an urgent need for advanced, highly accurate forecasting techniques.

1.2. Literature review

PV power forecasting approaches have been widely studied in academia and industry, primarily categorized into three types: physical, statistical, and machine learning models [9]. The physical model uses the PV effect principle to forecast future PV power output through mathematical models. Input parameters include meteorological data like solar irradiance, temperature, humidity, and cloud thickness [10]. While physical models generally don't need historical operational data from PV plants, their predictive accuracy and robustness are limited by modeling complexities and parameter uncertainties, including inaccuracies in meteorological data [11].

Statistical models primarily utilize historical operational data from PV power stations to establish mapping relationships between input and output variables through statistical methods such as linear regression and correlation analysis, thereby enabling the prediction of future PV power generation [11]. Commonly used statistical models—including regression analysis [12], autoregressive moving average [13], and Markov chains [14]—demonstrate limited prediction accuracy for stochastic and fluctuating PV power output due to their inherent dependence on strongly linearly correlated data patterns [15]. To tackle the low accuracy of physical and statistical models, recent studies have increasingly and successfully employed machine learning models—including support vector regression (SVR) [16], extreme learning machines [17], extreme gradient boosting (XGBoost) [18], convolutional neural networks (CNN) [11], recurrent neural networks (RNN) [19], attention mechanism (AM) [10], long short-term memory (LSTM) [20,21], and Transformer [22]—to enhance PV power forecasting performance through more sophisticated network architectures.

However, the previously mentioned PV power forecasting methods depend on single-task learning models and neglect feature-sharing mechanisms between tasks, which limits their capacity to capture inter-task correlations crucial for multi-task collaboration, potentially reducing generalization performance [23]. Multi-task learning, a learning approach that focuses on sharing knowledge across different tasks, can overcome the limitations of single-task learning, which often struggles to extract relevant features from related tasks during prediction [24]. Due to these benefits, multi-task learning has gained increasing research interest in the field of PV power forecasting in recent years. In [25], a multi-task learning-based nonlinear weather correction model is developed to enhance PV power forecasting accuracy. In [26], an integrated forecasting framework combining attention mechanisms with multi-task learning is proposed for simultaneous accurate prediction of wind power generation, PV output, and system load demand. In [27], a PV power prediction model based on a self-attention mechanism and multi-task learning is proposed. In [28], a RNN-based multi-task forecasting model for multi-PV stations is proposed. Current research on multi-task learning for power forecasting across PV stations remains insufficient, with existing approaches generally lacking systematic consideration for enhancing the extraction capabilities of temporal and spatial features.

Systematic consideration of spatial-temporal feature extraction constitutes a prerequisite for multi-task PV power forecasting, primarily driven by the following three rationales: (1) PV power curves and meteorological variations such as solar irradiance exhibit dynamic periodic patterns, necessitating temporal feature extraction to capture long-term cyclical trends [29]; (2) The power output of PV stations is profoundly affected by meteorological distributions, including the spatial correlations from adjacent stations [30], which can be exploited by the spatial feature extractor of multi-task learning models to enhance forecasting accuracy; (3) The introduction of a spatial-temporal feature extraction as a multi-task sharing layer facilitates transfer learning across various PV station forecasting tasks, enhancing model generalization in diverse operational conditions [31].

LSTM, as a specialized type of RNN, can capture long-term dependencies (temporal features) in meteorological variation data by introducing three critical gating mechanisms—the forget gate, input gate, and output gate—to control the flow and forgetting of information. Given these advantages, multi-task learning models based on LSTM have been successfully applied in the PV power forecasting [32,33]. However, LSTM-based PV power prediction exhibits two fundamental limitations: (1) The model cannot revise storage decisions, restricting dynamic information updating in nearest-neighbor search problems; (2) The limited storage capacity forces information compression into scalar cell states, degrading prediction accuracy for rare patterns. To overcome these limitations, the extended LSTM (XLSTM) architecture proposed in 2024 incorporates two key innovations: scalar LSTM (SLSTM) with exponential gating for dynamic memory updating, and matrix LSTM (MLSTM) with covariance update rules for expanded memory capacity [34]. This enhanced framework has demonstrated the most advanced performance in energy prediction applications and has become a research hotspot over the past year [35]. For instance, in [36], an XLSTM-based model is investigated for wind power forecasting, where the superior performance is demonstrated in handling non-linear wind power features compared to the LSTM. However, XLSTM-based single-task and multi-task learning models have been rarely reported in PV power forecasting research. Moreover, although XLSTM enhances the flexibility of informa-

tion flow through covariance update rules, its relatively fixed structure limits its ability to capture these dynamic key temporal features. To address this issue, the global attention mechanism—an advanced machine learning technique—dynamically assigns weights to all positions in an input sequence, enabling focused extraction of critical temporal features [37,38]. Nevertheless, the use of a temporal feature extractor that integrates XLSTM with a global attention mechanism has not been extensively studied in energy system forecasting.

Meanwhile, considering that multi-task PV power forecasting involves not only the extraction of temporal features but also spatial features (meteorological distribution), this implies that the model should also incorporate a well-designed spatial feature extractor. Spatial features can be further classified into local and global spatial features. Local features encompass station-specific factors such as solar irradiance, ambient temperature, and humidity. In contrast, global features primarily involve meteorological phenomena for all PV stations, including cloud movement patterns and their meteorological spatial differences [39]. The dilated causal convolution controls the receptive field through the dilation rate, which enhances local feature extraction while maintaining the ability to capture long-range spatial dependencies [40]. However, dilated causal convolution suffers from a lack of global feature extraction capability due to its sparse sampling mechanism (where the convolutional kernel operates with a localized receptive field) and unidirectional causal constraints that restrict information flow. The Transformer can dynamically capture critical global patterns by computing attention weights in parallel across multiple subspaces through the multi-head attention mechanism, compensating for the limitations of dilated causal convolutions in global feature extraction [41,42]. With these advantages, dilated causal convolutions and Transformer have been successfully applied in research fields such as wind power [43]. To improve the spatial feature extraction ability of multi-task learning, inspired by the complementary advantages of Transformer and dilated causal convolutional neural networks, this approach holds significant potential for multi-task PV power prediction research. However, studies combining dilated causal convolution and Transformer in multi-task PV power prediction remain scarce.

In addition, the above multi-task PV power forecasting approach also encounters two primary challenges. The first challenge issue is their failure to account for the impact of outliers, which consequently leads to overfitting and compromised robustness in multi-task forecasting. Commonly used outlier detection methods for PV power forecasting primarily include the box plot [44], outlier clustering [45], and z-score analysis [46]. However, these approaches often struggle with large-scale datasets and are prone to overfitting. To address this challenge, the isolation forest demonstrates great efficiency to detect anomalous points in large-scale datasets through random partitioning and path length-based anomaly scoring [47]. Given these advantages, the isolation forest algorithm has been explored for PV power forecasting over the past two years [11]. However, the aforementioned traditional detection methods and the isolation forest primarily focus on single-task anomaly detection, which restricts their ability to leverage cross-task shared features and results in lower detection accuracy and suboptimal generalization performance.

The second major challenge arises from the intrinsic complexity of multi-task learning models, which complicates understanding the connections between input features and multiple output tasks, as well as determining the importance of each feature. In contrast to interpretable approaches like partial dependence plots [48], attention weight [30], transfer learning-based interpretation [49], and local surrogate models [50], the Shapley Additive exPlanations (SHAP) method quantifies feature importance by computing Shapley values—a game-theoretic measure of marginal contribution each feature—thereby enabling both local and global model interpretability, and has been successfully applied in the energy field of PV power prediction [51]. However, there are few multi-task interpretable PV power prediction reports based on SHAP. In fact, achieving interpretability in multi-task PV power forecasting is more difficult than in single-task forecasting. The primary reasons can be summarized as follows: (1) Multi-task learning necessitates that the model captures the correlations among multiple tasks within a shared representation layer. This results in nonlinear interactions between the input features and output variables, complicating the accurate differentiation of the dependencies of individual tasks. (2) While multi-task learning enhances generalization ability by sharing underlying feature representations, the weight optimization in the shared layer is the negotiation outcome among multiple tasks. Consequently, the decision-making logic of this shared layer may not directly correspond to an intuitive explanation for any single task. Given these difficulties and the limited research on multi-task interpretability through SHAP, investigating the interpretability of multi-task models is a significant endeavor.

1.3. Research gaps and contributions

Based on the above literature review, the following unresolved issues are summarized:

- Multi-task learning can enhance feature extraction ability and robustness by jointly modeling the correlations among multiple PV stations. Temporal and spatial feature extractors can capture long-term dependencies and global-local dynamics of PV stations through spatiotemporal modeling. However, the multi-task learning model based on temporal-spatial feature extractors remains underreported in PV power forecasting.
- XLSTM can better dynamic state updating and parallel processing by enhancing memory capacity, improving flexibility, and optimizing computational efficiency compared with LSTM. However, there are relatively few reports on PV power prediction models based on XLSTM for both single-task and multi-task learning.
- The global attention mechanism can better capture key temporal features of the XLSTM, but there is a notable paucity of research on PV power prediction that leverages the combination of XLSTM and the global attention mechanism.
- Dilated causal convolution and Transformer are capable of effectively capturing both local and global features of PV power. However, research on spatial feature extractors based on dilated causal convolution and Transformer in multi-task PV power prediction remains insufficient.
- Isolation forests efficiently detect anomalies in large-scale datasets via random partitioning and path-length scoring. While multi-task detection improves efficiency and accuracy by processing related tasks jointly, its integration with isolation forests remains unexplored in PV power forecasting.
- Although SHAP can effectively interpret the mapping relationships between features and predictive outcomes, research on applying SHAP to explain multi-task output results in PV power forecasting remains scarce.

To fill the identified research gaps, this paper introduces a novel multi-task spatial-temporal interpretable forecasting method specifically designed for multiple PV power stations. Table 1 provides a comparative analysis highlighting the research gaps between the proposed multi-task spatial-temporal forecasting model and related works. In comparison to existing studies, the key contributions of this paper are outlined as follows:

- For the first time, a multi-task spatial-temporal interpretable prediction model is proposed, which includes a multi-task shared layer that employs a spatiotemporal feature extractor, along with task-specific layers built on feedforward neural networks. The shared layer leverages the spatiotemporal feature extractor to identify complex, high-dimensional features that are correlated across different tasks. At the same time, the task-specific layers employ fully

Table 1
Comparative analysis of research gaps between the proposed multi-task spatial-temporal forecasting approach and similar literature.

Ref.	Year	Prediction objection	Outlier detection	Multi-task	Spatial-temporal	Prediction model	Explainability
[10]	2025	PV power	×	×	×	Attention mechanism	SHAP(SL)
[11]	2022	PV power	IF	×	×	CNN	×
[19]	2025	PV power	×	×	×	LSTM-RNN	×
[21]	2025	PV power	×	×	×	Monte Carlo+LSTM	×
[22]	2024	PV power	×	×	×	Transformer-RNN	×
[24]	2024	Multi-energy load	×	✓	✓	Gated convolutional	Attention weight(ML)
[23]	2025	Wind power	×	✓	×	Transformer+LSTM+ML	×
[25]	2025	Weather&PV power	×	✓	×	LSTM+ML	Interaction weight(ML)
[26]	2024	Wind-solar-load	×	✓	×	AM+ML	×
[28]	2024	Multi-PV stations	×	✓	×	RNN+ML	×
[29]	2024	PV power	×	×	✓	Attention mechanism	×
[30]	2022	Multi-PV stations	×	×	✓	Graph attention network	Attention weight(SL)
[31]	2024	Multi-meteorology	×	✓	✓	Transformer+ML	×
[32]	2025	PV power	×	✓	×	LSTM+ML	×
[33]	2023	Wind-PV power	×	✓	×	LSTM+ML	×
[36]	2025	Wind power	×	×	×	XLSTM	×
[38]	2024	PV power	×	×	×	CNN+BiLSTM+Attention	×
[43]	2023	Wind power	×	×	×	DCC+Transformer	×
[51]	2024	PV power	×	×	×	Deep reinforcement learning	SHAP(SL)
[52]	2024	Traffic flow	×	×	✓	DCC+MA	×
Our	2025	Multi-PV stations	IFMD	✓	✓	DT+XLSTM+GA+ML	SHAP(ML)

connected neural networks to capture features unique to each task, ensuring precise predictions.

- To comprehensively capture the interactions between stations influenced by meteorological changes and distributions, a new spatiotemporal feature extractor combining Transformer and XLSTM is introduced to extract nonlinear characteristics of PV power. The spatial feature extractor, which integrates rotary position embedding, Transformer, dilated causal convolution, and residual connections, is initially developed to jointly model both local and global features for multi-task PV power forecasting. Subsequently, the temporal feature extractor, based on XLSTM and global attention mechanisms, is employed to identify long-term dependencies and essential temporal features.
- To guarantee high-quality training data and enhance prediction accuracy, a multi-task outlier detection model based on isolation forests is incorporated into the forecasting method to detect anomalies in historical PV power data. This approach effectively minimizes overfitting and improves generalization by utilizing multi-task anomaly detection alongside randomized subspace partitioning techniques.
- An interpretable model based on SHAP is initially introduced into a multi-task spatial-temporal approach, enabling the elucidation of coupled feature interactions and their causal mappings to multi-task output variables in PV power forecasting.

1.4. Paper organization

The structure of the remainder of this paper is presented as follows. The proposed multi-task spatial-temporal forecasting model for PV power is introduced in Section 2. The proposed approach, incorporating multi-task outlier detection and interpretability, along with its detailed prediction procedures, is presented in Section 3. Section 4 presents the experimental configuration and simulation analysis of the proposed approach. The concluding remarks are summarized in Section 5.

2. The proposed multi-task spatial-temporal forecasting model for multi-PV power stations

In this section, we first propose a multi-task sharing layer based on spatial and temporal feature extractors for learning correlated feature of PV power prediction tasks. Then, a novel multi-task spatial-temporal forecasting model, consisting of a multi-task sharing layer and a task-specifying layer using a fully connected neural network, is proposed for prediction of multi-PV power stations, as shown in Fig. 1.

2.1. The proposed spatial-temporal feature extractor

Typically, the output of PV power time series is influenced by two major categories of features: temporal and spatial features. To comprehensively account for these influences and improve the prediction accuracy of PV stations, this paper proposes a novel multi-task sharing layer based on spatial and temporal feature extractors. In these two feature extractors, the spatial feature extractor is first used to capture global and local features, followed by the temporal feature extractor, which learns long-term dependency and critical temporal features. The placement of the spatial feature extractor before the temporal feature extractor offers two key advantages: (1) PV power is directly influenced by spatially distributed factors such as irradiance, temperature, and cloud movement. Spatial feature extractor can effectively capture spatial topology relationships, providing temporally explicit physical characteristics as inputs for subsequent temporal feature extraction; (2) The sequential architecture prioritizing spatial feature extraction over temporal processing prevents gradient entanglement caused by hybrid spatial-temporal parameter updates, thereby enhancing training stability and convergence efficiency.

2.1.1. Spatial feature extractor based on convolutional Transformer

The traditional Transformer is composed of multi-head attention, residual connections, layer normalization, and a feed-forward neural network [53]. Although the multi-head attention mechanism, as the core component of the Transformer, can capture the dependency relationships between any positions in the input sequence through dynamic weight allocation (global features), it lacks explicit local feature extraction capability [23]. Considering that dilated causal convolution possesses the ability for local feature extraction while still maintaining the capacity to capture long-range spatial dependencies, a spatial feature extractor based on the convolutional Transformer in this work is designed to learn global-local features from PV power prediction tasks. The structure of the spatial feature extractor based on a convolutional Transformer is shown in Fig. 1. The multi-head attention mechanism of the Transformer projects input data into query-key-value subspaces via independent linear transformations of each head, facilitating parallelized multi-dimensional feature extraction to capture global features in PV power. However, as the standard Transformer is inherently position-agnostic, this study introduces rotary position embedding to encode sequential ordering into the query-key attention computation [54]. Rotary position embedding (RoPE) employs unitary transformations in complex space, where positional offsets are represented as phase rotations in the complex plane. This design inherently preserves relative

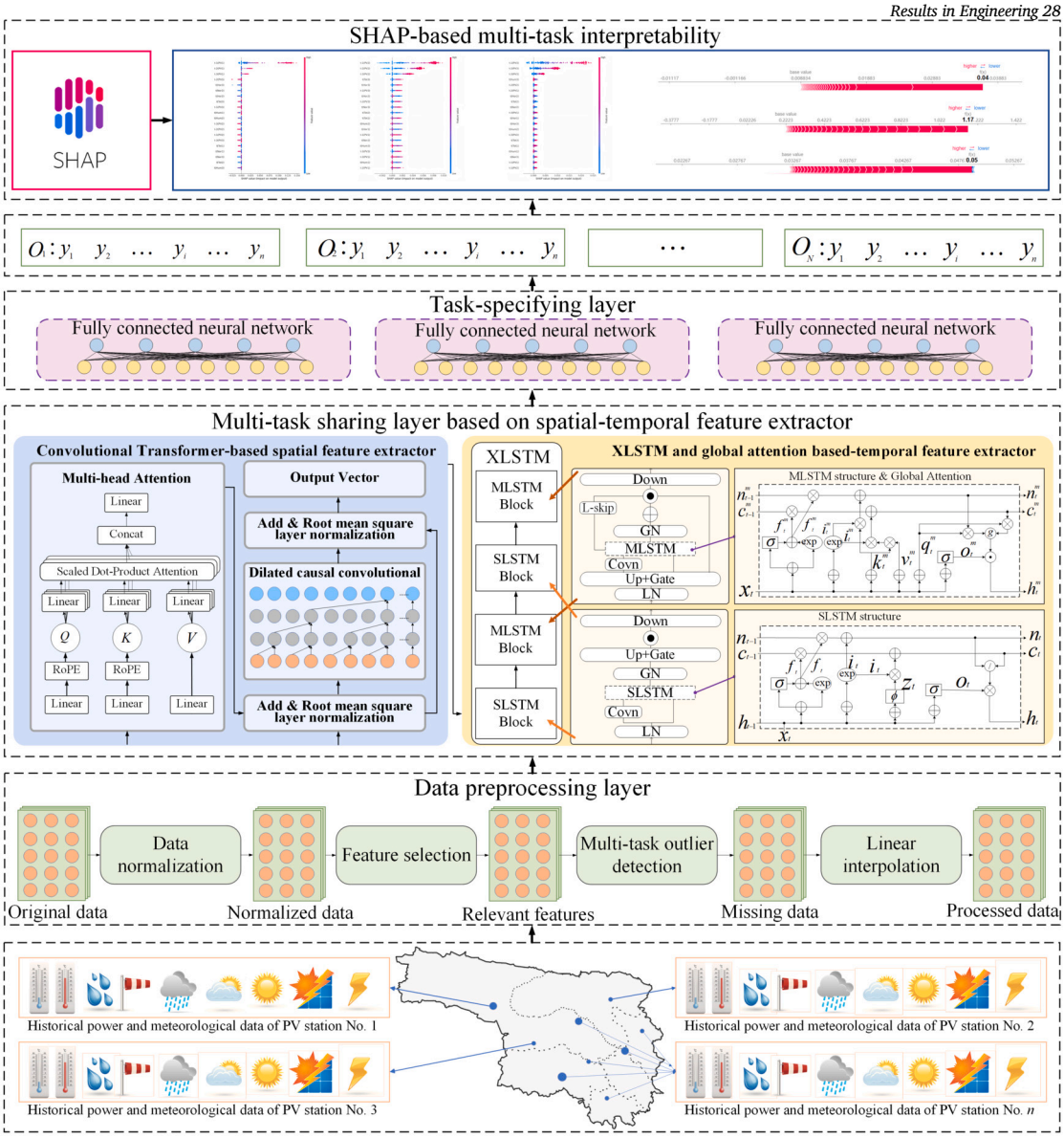

Fig. 1. Network structure of the proposed multi-task spatial-temporal interpretable prediction approach for multi-PV stations.

positional relationships through the group-theoretic properties of rotation operators. Compared to absolute position encoding, rotary position embedding demonstrates superior generalization to variable sequence lengths and improved efficiency [54]. The rotary position embedding is mathematically described as follows:

$$f_{\{q,k\}}(x_j, j) = G^d_{\phi,j} W_{\{q,k\}} x_j \tag{1}$$

$$G^d_{\phi,j} = \begin{bmatrix} H^0_{\phi,j} & 0 & .... & 0 \\ 0 & H^1_{\phi,j} & .... & 0 \\ ... & ... & ... & ... \\ 0 & 0 & ... & H^{d/2-1}_{\phi,j} \end{bmatrix} \tag{2}$$

$$H^{d/2-1}_{\phi,j} = \begin{bmatrix} \cos j\theta_{d/2-1} & -\sin j\theta_{d/2-1} \\ \sin j\theta_{d/2-1} & \cos j\theta_{d/2-1} \end{bmatrix} \tag{3}$$

$$\phi = \{\phi_j = 10000^{-2(j-1)/d}, j \in [1, 2, ..., d/2]\} \tag{4}$$

where x_j denotes the j-th feature. The transformation matrix $W_{\{q,k\}}$ is used to calculate the weights for the query and key vectors, respectively. The vector $f_{\{q,k\}}$ represents the query/key embedding of the feature after incorporating rotational positional encoding. The angle θ is a predefined hyperparameter controlling the magnitude of rotational adjustments applied to the query and key representations. The operator $G^d_{\phi,j}$ corresponds to an orthogonal matrix that preserves the vector norm during transformations. $H^{d/2-1}_{\phi,j}$ is a rotation matrix parameterized by θ within an d-dimensional space, enabling dimension-specific angular transformations.

Following position encoding, the input sequence undergoes attention calculation through a multi-head attention structure. This mechanism captures richer feature representations and multiple semantic associations by executing multiple parallel attention heads. Within each head, an independent query, key, and value weight project the input into distinct representation subspaces. The attention computation begins by applying rotary positional encoding to both query and key vectors (g_q and g_k), which preserves relative positional relationships through rotational transformations. These enhanced vectors then participate in scaled dot-product operations to determine attention scores, effectively modeling token dependencies. Then, to formalize this computation, the attention scores are normalized through a differentiable Softmax operator ϖ, which induces a probabilistic distribution over the input tokens. This distribution subsequently orchestrates a context-aware aggregation of the value vectors via a linear combination operation. Specifically, the

weights $a_{w,n}$ are employed as coefficients to compute a convex combination of the value embeddings, thereby synthesizing latent semantic representations for each attention head. Finally, the multi-head outputs are aggregated via concatenation and projected into a unified semantic space through a learned linear transformation ω, yielding the final multi-head attention output MA_t, as described below:

$$m_w = \varpi(g_q * g_k) = x_j^T W_q (G_{\phi,j}^d)^T G_{\phi,k}^d W_k x_k \tag{5}$$

$$MA_t = \omega(m_{w,1} v_1, ..., m_{w,n} v_n, ..., m_{w,N} v_N) \tag{6}$$

where v_n represents the value embedding associated with the $n-th$ attention head. $m_{w,n}$ is the attention weight for head n.

Following the multi-head attention mechanism, dilated causal convolution is utilized as an enhanced 1D CNN structure to effectively capture localized features in PV power forecasting tasks. Dilated convolution enhances the network receptive field through strategically spaced kernel elements (controlled by a dilation rate), effectively improving feature spatial modeling while preserving computational efficiency and parameter count. Causal convolution ensures that every output element relies only on the current and past input values, maintaining the temporal sequence by not incorporating any future data points. Fig. 1 illustrates a dilated causal convolution architecture comprising an input layer, two convolutional feature mapping layers, and an output layer. Unlike standard convolution, dilated causal convolution strategically skips adjacent neurons through a dilation rate mechanism, enabling exponential receptive field expansion at the output layer. For a multi-head attention sequence $O_{m,s-p,k}$ and filter f_p with dilation rate p, the dilated causal convolution operation DC is defined as:

$$DC_t = (MA_t *_c f_p) = \sum_{p=1}^{C_p - 1} f_p O_{m,s-p,k}, t > 0 \tag{7}$$

where p is the dilation rate. The operator $*_c$ represents convolution parameterized by c, where the effective receptive field of the kernel expands to C_p. $s-p,k$ denotes the convolution operation direction. The DC_t enforces causality by exclusively aggregating input features from the current timestep t and all preceding timesteps, ensuring that no future information ($t > 0$) influences the output.

To alleviate the issue of network degradation, residual connections are introduced after the multi-head attention and the dilated causal convolution. Meanwhile, the root mean square layer normalization $LN(i_x)$ is applied to impose constrained weight scaling within a specific range, which enhances the model training efficacy,

$$LA_t = LN(x_t + MA_t) \; and \; LD_t = LN(DC_t + LA_t) \tag{8}$$

$$LN(i_x) = g_w \frac{i_x - \mu}{\sqrt{(\sigma^2 + \epsilon)}} + \beta_w \tag{9}$$

where LA_t and LD_t denote the outputs obtained by applying layer normalization to the residual connections following the multi-head attention mechanism and the dilated causal convolution, respectively. The parameters g_w and β_w are learnable scaling and shifting factors. The symbol ϵ represents a small constant used for numerical stability. The variables μ and σ correspond to the mean and variance of the feature i_x, which specifically refer to the residual connections $x_t + MA_t$ and $DC_t + LA_t$, respectively.

2.1.2. Temporal feature extractor based on XLSTM and global attention

In this subsection, an XLSTM and global attention mechanism-based temporal feature extractor is proposed to capture long-term dependencies and critical features in PV power prediction tasks. XLSTM is a next-generation recurrent neural network architecture proposed in 2024 by Sepp Hochreiter, the original creators of LSTM. It introduces two key innovations—exponential gating and matrix memory—to systematically address the critical limitations of traditional LSTM in long-term memory storage, parallel computation efficiency, and model capacity. The architecture of XLSTM comprises two core modules: scalar LSTM (SLSTM) and matrix LSTM (MLSTM). SLSTMis designed to enhance the stability of conventional LSTM architectures, prevent numerical overflow, and mitigate gradient-related issues through two core innovations: exponential gating and new memory mixing, as follows: (1) Traditional LSTMs use sigmoid functions to control their gating mechanisms. In contrast, SLSTM replaces these with exponential functions to better regulate gate values, allowing for more flexible modulation of gate dynamics, reducing nonlinear saturation effects, and suppressing numerical overflow. (2) In traditional LSTMs, the cell state controls information retention and updates through forget and input gates. The SLSTM introduces a mechanism called new memory mixing, which combines exponential gating with normalized states to optimize memory management. This method not only improves numerical stability by constraining activation magnitudes but also enhances the model's ability to capture long-term dependencies through dynamic memory consolidation. The mathematical definition of SLSTM is as follows:

$$c_t = f_t c_{t-1} + i_t z_t \tag{10}$$

$$n_t = f_t n_{t-1} + i_t \tag{11}$$

$$h_t = o_t c_t / n_t \tag{12}$$

$$z_t = \phi(w_z^T x_t + r_Z h_{t-1} + b_z) \tag{13}$$

$$i_t = \exp(w_i^T x_t + r_i h_{t-1} + b_i) \tag{14}$$

$$f_t = \sigma(w_f^T x_t + r_f h_{t-1} + b_f) \; or \exp(w_f^T x_t + r_f h_{t-1} + b_f) \tag{15}$$

$$o_t = \sigma(w_o^T x_t + r_o h_{t-1} + b_o) \tag{16}$$

where c_t, n_t, h_t, z_t, i_t, f_t, and o_t denote the cell state, normalizer state, hidden state, cell input, input gate, forget gate, and output gate. w_z, w_i, w_f, and w_o denote the weights of the cell input, input gate, forget gate, and output gate. b_z, b_i, b_f, and b_o denote the bias of the cell input, input gate, forget gate, and output gate.

The derivative value of an exponential function increases exponentially as the numerical value grows, which is prone to cause overflow. Therefore, XLSTM adopts stabilizing of the output to stabilize the input gate and the forget gate to achieve numerical stability, as described below:

$$m_t = \max(\log(f_t) + m_{t-1}, \log(i_t)) \tag{17}$$

$$i_t' = \exp(\log(i_t) - m_t) \tag{18}$$

$$f_t' = \exp(\log(f_t) + m_{t-1} - m_t) \tag{19}$$

where m_t, i_t', and f_t' denote the stable values of the state, input, and forget gates, respectively.

MLSTM of the XLSTM aims to enhance the memory and expression capabilities of the model by using matrix-form memory. In addition to introducing the exponential gating mechanism with SLSTM, MLSTM also has two significant improvements, such as matrix memory and covariance update rule, as follows: In traditional LSTM, the cell state is a scalar value or vector, while in MLSTM, the cell state is extended to a matrix, enabling it to store more context information. In MLSTM, memory is expanded into a matrix form, allowing the model to store multiple key-value pairs. To effectively manage and update these key-value pairs, an update mechanism similar to a covariance matrix is introduced, which is mathematically described as follows:

$$c_t^m = f_t^m c_{t-1}^m + i_t^m v_t^m k_t^{T,m} \tag{20}$$

$$n_t^m = f_t^m n_{t-1}^m + i_t^m k_t^m \tag{21}$$

$$h_t^m = o_t^m \odot \{c_t^m q_t^m / \max[\left| n_t^{T,m}, q_t^m \right|, 1]\} \tag{22}$$

$$q_t^m = W_q x_t + b_q \tag{23}$$

$$k_t^m = \frac{1}{\sqrt{d}} W_k x_t + b_k \tag{24}$$

$$v_t^m = W_v x_t + b_v \tag{25}$$

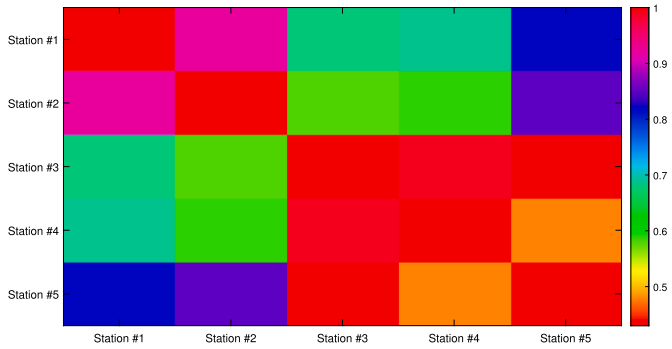

Fig. 2. The correlation coefficient of multi-PV power stations.

$$i_t^m = \exp(W_i^T x_t + b_i) \tag{26}$$

$$f_t^m = \sigma(W_f^T x_t + b_f) \; or \exp(W_f^T x_t + b_f) \tag{27}$$

$$o_t^m = \sigma(W_o x_t + b_o) \tag{28}$$

where c_t^m, n_t^m, h_t^m, q_t^m, k_t^m, v_t^m, i_t^m, f_t^m, and o_t^m are respectively the outputs of the cell state, normalizer state, hidden state, query state, key input, value input, input gate, forget gate, and output gate for MSLTM.

Considering that MLSTM may fail to capture long-range dependencies in sequences and has a limited ability to capture key temporal features, a global attention mechanism is introduced into the MLSTM of XLSTM to address these limitations. Unlike multi-head attention, global attention performs pairwise interaction computations across all positions of the input sequence, establishing direct relationships between each element and the entire sequence. This characteristic makes it particularly suitable for integration with MLSTM to capture global consistency features and enhance long-range dependency modeling. In this study, the global attention dynamically weights the hidden state sequence of MLSTM to identify importance variations across different timesteps in its output, as described below:

$$\chi_k = \sum_{t=1}^{T} h_t^m \frac{\exp(S_k^T W^I h_t^m)}{\sum_{t=1}^{T} \exp(S_k^T W^I h_t^m)} \tag{29}$$

where S_k and W^I are the hidden state and the learnable weight matrix of the MLSTM, respectively. Given the MLSTM hidden states and environmental variables, the information from both vectors can be combined via an activation-based concatenation method, generating an attentive hidden state $\tilde{S}_k$. Finally, a linear projection is applied to produce the MLSTM's output y_k, as described below:

$$y_k^m = \tilde{S}_k W^S, \tilde{S}_k = tanh([c_k^m : S_k] * W^C) \tag{30}$$

where W^C and W^S denote respectively the weight matrices of the attention hidden state and the linear output.

2.2. Spatial-temporal feature extractor for building multi-task learning model in PV power forecasting

In this study, a novel multi-task sharing layer based on spatial and temporal feature extractors is proposed, which has at least four distinct advantages. (1) The spatial feature extractor, empowered by multi-head attention of the Transformer with rotary position embeddings, achieves multidimensional feature awareness, enabling effective global feature extraction for PV power forecasting tasks. (2) The spatial feature extractor effectively enhances the receptive field while preserving local detail information through dilated causal convolutional, enabling robust extraction of localized features for PV power forecasting. (3) The temporal feature extractor utilizes SLSTM and MLSTM architectures of the XLSTM to enhance memory storage and parallel computation efficiency, effectively capturing long-term dependencies in PV power prediction tasks. (4) The global attention mechanism is introduced into the XLSTM to improve the dynamic weight allocation and the capability to focus on key features of XLSTM. The superiority of the proposed spatial-temporal feature extractor over similar advanced PV power prediction models has been demonstrated in Section 4.2.

Furthermore, considering the interactions among different PV power stations, it is necessary to investigate the correlation of PV power prediction. Taking the PV power data of the PV stations in Northwest China in December 2020 as an example, Fig. 2 shows the correlation coefficients of five PV power stations. The color bar in Fig. 2 represents the correlation of PV power among multi-PV stations. It is clearly shown in Fig. 2 that numerous rose-red, black, and green squares are distributed across the figure, indicating that the correlation coefficients of PV power prediction exceed 0.7. To comprehensively account for the inter-dependencies among PV power stations, multi-task learning offers potential advantages over single-task learning by explicitly capturing inter-dependencies through shared representations. The main difference between multi-task learning and single-task learning lies in the output structure (i.e., single-task learning predicts only a single target, whereas multi-task learning jointly predicts multiple targets), while their input features remain consistent. While single-task learning can achieve high accuracy for individual stations at the cost of increased model complexity, multi-task learning improves efficiency through parameter sharing and may enhance generalization by leveraging spatial-temporal correlations across tasks.

Based on the above four major advantages of the proposed spatial-temporal feature extractor, and considering the strong correlation in PV power stations, this paper proposes a novel multi-task spatial-temporal forecasting model for power predictions across multiple PV power stations. The task-sharing layer captures important global and local features as well as long-term dependencies from various PV power plants. The task-specific layer focuses on extracting distinct features for each individual task to ensure precise task results. The proposed multi-task spatial-temporal prediction model, in terms of its network architecture, is categorized as a bottom-layer shared structure of multi-task learning models. In comparison to the other two representative multi-task learning structures (multi-gate mixture-of-experts and the subnet routing), the structure employs a shared layer that serves as the feature extractor for all tasks [55]. This design offers advantages such as ease of model implementation, a reduction in the number of model parameters, and a mitigation of the overfitting risk. Assume there are Z tasks in multi-task PV power forecasting. The model parameters are optimized by minimizing the loss function with L2 regularization, as shown below:

$$\ell = \min_{\{\lambda_s, \lambda_m\}} \sum_{t=1}^{T} \sum_{z=1}^{Z} \alpha_z L_z\{f[g(x_t, \lambda_s), \lambda], y_t^z\} + \pi\Psi(\lambda_s, \lambda_m) \tag{31}$$

where λ_s and λ_m are the parameters of the multi-task shared layer and the task-specific layer of task z. x_t and y_t^z represent the relevant features and observed values of task z at time t. f and g represent the functions of the multi-task sharing layer and task-specific layers. α_z is the weight value used to balance the PV power prediction task, and λ is the regularization parameter for reducing overfitting. L_z and Ψ are respectively the loss function and L2 regularization term of the PV power prediction tasks.

3. The proposed multi-task spatial-temporal interpretable prediction approach considering multi-task outlier detection

In this section, an isolation forest-based multi-task outlier detection model is first introduced into the proposed multi-task spatial-temporal forecasting approach. Subsequently, the SHAP model is employed to elucidate the relationships between the coupled features and multi-task outputs in the proposed approach. Finally, the detailed prediction procedures of the proposed multi-task spatial-temporal interpretable forecasting approach of PV power are presented.

3.1. Isolation forest-based multi-task outlier detection

Outliers in multi-task PV power forecasting datasets—caused by factors such as sensor malfunctions, partial shading, or extreme weather conditions—can significantly distort model parameter optimization and compromise prediction accuracy. To enhance PV power dataset quality and mitigate the adverse effects of outliers on multi-task forecasting performance, the development of robust anomaly detection algorithms is essential for improving PV power prediction precision. Compared to traditional anomaly detection methods such as box plot, outlier clustering, and z-score analysis, the isolation forest method demonstrates superior performance by randomly partitioning feature space to efficiently isolate anomalies without relying on data distribution assumptions, achieving higher classification prediction accuracy [11]. In addition, current methods primarily focus on single-task anomaly detection, which limits the learning of inter-variable correlations and results in poor generalization performance.

To address this limitation, this study proposes an isolation forest-based multi-task approach for anomaly detection of PV power. The isolation forest-based multi-task outlier detection model leverages a shared bottom-layer architecture for feature extraction and outputs task-specific anomaly scores through dedicated branches tailored to different tasks. The detailed implementation steps are as follows: (1) Randomly selecting multivariate samples from multiple PV station datasets (including PV power and meteorological parameters) to construct an ensemble of isolation trees; (2) Each isolation tree recursively splits samples by simultaneously selecting features and split thresholds from the shared multi-task feature space until either isolation occurs or reaching the predefined depth limit; (3) The average path length (multi-task outlier score) of the multivariate sample is calculated based on the mean path length $E(h(x,z))$ and the normalization factor $c(m)$ of the sample number m in isolation tree to determine the degree of anomaly. The outlier scores $S(x,m,z)$ are described as follows:

$$S(x,m,z) = 2^{-\frac{E(h(x,z))}{c(m)}} \tag{32}$$

$$c(m) = 2H(m-1) - \frac{2(m-1)}{m} \tag{33}$$

where $E(h(x,z))$ is the mean path length of sample x task z in all isolation tree. $H(i) = ln(i) + 0.5772156649$, which is the harmonic function.

3.2. SHAP-based spatial-temporal interpretability for multi-task forecasting of PV power

In this subsection, the SHAP model—grounded in cooperative game theory—is introduced to interpret multi-task spatial-temporal prediction results of PV power through quantitative feature contribution analysis. The SHAP model operates in three sequential stages: (1) Based on the cooperative game theory, SHAP values are computed to quantify the marginal contributions of individual features to model outputs, enabling global interpretability(the model holistic behavior) and local interpretability(single-sample predictions) [56]. This quantification is achieved by evaluating the predictive impact of varying feature values across multi-task prediction. (2) Marginal contributions are aggregated via SHPA value-weighted averaging, ensuring axiomatic fairness while enabling cross-task feature importance estimation. (3) Interpretability visualizations are produced to elucidate how feature interactions collectively drive multi-task spatial-temporal predictions. In the cooperative game theory, if any subset S of the feature set (relevant feature combination) is associated with a contribution function $v(S)$, and any two disjoint subsets S_1 and S_2 satisfy the super-additivity condition $v(S_1) \cup v(S_2) > v(S_1) + v(S_2)$, then cooperative behavior emerges only when the allocated value (prediction accuracy) for each subset exceeds its non-cooperative baseline. In the context of multi-task spatial-temporal prediction, the SHAP model quantifies the intrinsic decision-making mechanisms by computing SHAP values for each multi-output prediction under relevant feature combinations. These SHAP values can analyze the fundamental principles behind the prediction results and the degree of influence of related features. Specifically, the SHAP value v_k^z for the kth feature in the zth task can be formally defined as:

$$v_k^z = \sum_{k \in N(k)} \frac{(|k|-1)!\,(W-|k|)!}{|W|!} m_k^z \tag{34}$$

where W denotes the cardinality (total count) of the feature set utilized for predicting PV power outputs. Define $N(k)$ as the subset of features within the associated feature space that includes feature k. m_k^z represents the marginal contribution of the feature subset containing k for output task z to the predictive model performance [57]. In the multi-task spatial-temporal forecasting of PV power, the SHAP value attributed to each feature contribution for a given sample adheres to the following formulation:

$$p_t^z = y_b^z + f\left(x_{t1}^z\right) + f\left(x_{t2}^z\right) + f\left(x_{t3}^z\right) +, \ldots + f\left(x_{tk}^z\right) \tag{35}$$

where y_b^z denotes the model baseline prediction, typically defined as the mean of the predicted values across all samples in the dataset. x_{tk} is the k-th feature for the task z within the t-th sample. SHAP value $f\left(x_{tk}^z\right)$ quantifies the marginal contribution of feature x_{tk}^z to the final predicted value p_t^z for sample t. If the SHAP value $f\left(x_{tk}^z\right) > 0$, the feature x_{tk}^z positively contributes to the predicted value p_t^z, thereby increasing it relative to the baseline. Conversely, if $f\left(x_{tk}^z\right) < 0$, the feature x_{tk}^z negatively contributes to p_t^z, resulting in a reduction of the predicted value.

3.3. Prediction steps of the proposed interpretable forecasting approach for multi-PV stations

In this paper, a novel interpretable forecasting approach using spatial-temporal multi-task learning with Transformer-XLSTM is proposed for predicting multi-PV power stations. The prediction procedure of the proposed multi-task spatial-temporal forecasting approach is illustrated in Fig. 3, with detailed descriptions provided as follows. (1) Collect and input external variables and historical data from multiple PV power stations. (2) The collected PV power and external variable data are normalized to 0 and 1. (3) Feature selection is performed on the normalized dataset through the random forest algorithm to determine the feature variables. Compared to conventional feature selection techniques such as Pearson correlation, random forest offers distinct advantages. Random Forest intrinsically evaluates feature importance by measuring the mean decrease in Gini impurity or information gain during node splitting. Additionally, its embedded feature subspace sampling mechanism ensures robust performance in high-dimensional data scenarios. The feature selection based on random forests for PV power prediction can be specifically observed in our previous research [51]. (4) An isolation forest-based multi-task anomaly detection approach is employed to identify and filter out outliers in PV power datasets using Eqs. (32)-(33). (5) The dataset after outlier filtering is divided into training and testing datasets according to different seasons. Note that the testing dataset does not require screening via the multi-task outlier detection method. (6) A new spatial feature extractor based on rotary position embedding, Transformer, dilated causal convolutional, and residual connection is constructed using Eqs. (1)-(8) to extract the global-local features of PV power. (7) A temporal feature extractor is constructed using Eqs. (10)-(30) by integrating XLSTM with a global attention mechanism, enabling the model to effectively capture both long-term dependencies and discriminative features in PV power data. (8) A novel multi-task sharing layer based on spatial and temporal feature extractors is designed for learning the nonlinear features of PV power correlated tasks, in which the loss function with L2 regularization is constructed using Eq. (31). (9) The weights, biases, and hyperparameters of the PV power multi-task prediction model are initialized, and the weights and biases are iteratively optimized using the Adam algorithm under the training dataset. (10) Upon completion of each training episode, both the model parameters (weights and biases) are persisted. (11) Perform

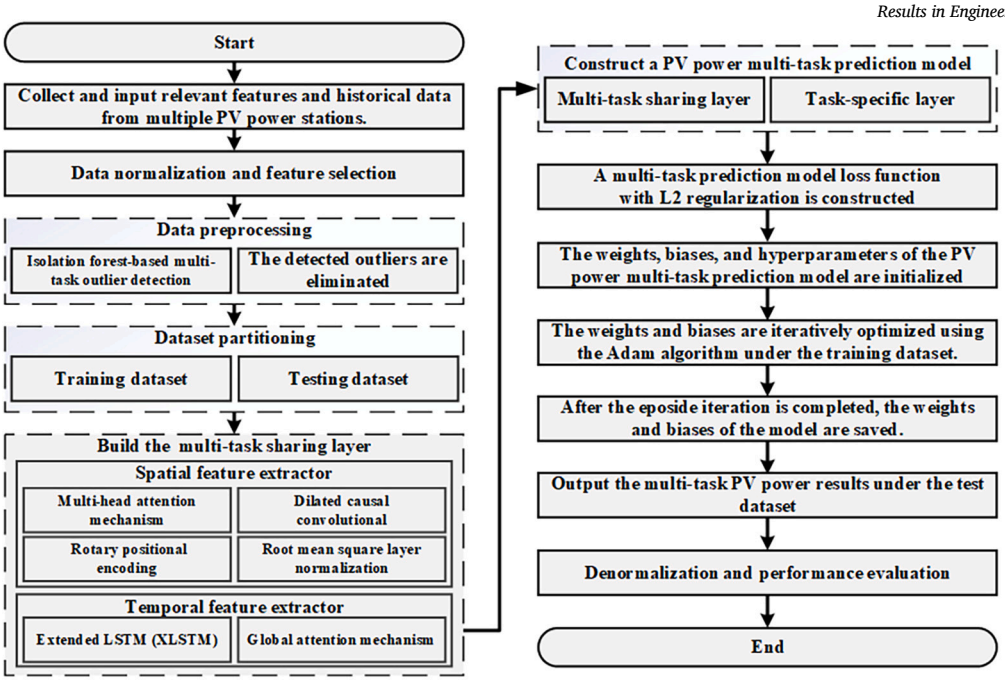

Fig. 3. Flowchart of the proposed multi-task spatial-temporal interpretable forecasting approach of PV power.

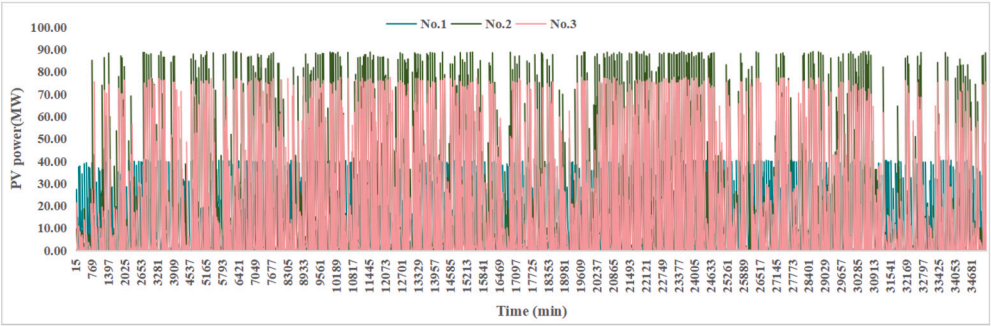

Fig. 4. The original data of the three PV power stations.

Table 2
PV power statistics of the three power stations in 2020.

PV station	Min(MW)	Max(MW)	Mean(MW)	SD(MW)
No. 1	0.00	40.38	9.51	13.39
No. 2	0.00	88.97	15.75	26.04
No. 3	0.00	77.44	13.75	22.50

denormalization and output the multi-task spatial-temporal prediction results of PV power under the test set. (12) The SHAP model is employed to interpret multi-task spatial-temporal prediction results and elucidate the importance of relevant features for multiple-task outputs. (13) The multi-task prediction results are evaluated using two error performance metrics: mean absolute error (MAE) and root mean squared error (RMSE) [47]. MAE represents the average absolute differences between predicted and actual values, while RMSE denotes the square root of the mean squared differences. Lower values of both metrics indicate superior predictive performance of the proposed multi-task spatial-temporal interpretable forecasting approach.

4. Case studies

4.1. Experimental settings

In this paper, we propose a novel multi-task interpretable forecasting approach, which consists of a multi-task shared layer (spatial-temporal feature extractor) and task-specific layers for forecasting multiple PV power stations. The spatial feature extractor combines rotary position embedding, Transformer, dilated causal convolutional, and residual connections to achieve joint modeling of local and global features. The temporal feature extractor utilizes an XLSTM and a global attention mechanism, enabling the model to effectively capture long-term dependencies and key features in the temporal dimension. Subsequently, to ensure the quality of the training data, a multi-task outlier detection method based on an isolation forest is introduced into the forecasting model to identify anomalies. Finally, the SHAP model is utilized to elucidate the relationships between the coupled features and multi-task outputs in the proposed approach. The hyperparameters of the proposed multi-task spatial-temporal interpretable prediction model are mainly obtained through the trial-and-error method, which is described as follows. In the spatial feature extractor, the convolution kernel size is 3, and the dilation rate for the dilated convolution is 2. The number of attention heads, the model dimension, and the dimension per attention head are 8, 512, and 64, respectively. The tiny threshold for root mean square layer normalization is set to 1E-6. For the temporal feature extractor, the global attention head dimension, the number of network layers, the projection factor for SLSTM (which controls the compression ratio of the hidden state dimension), and the projection factor for MLSTM (which adjusts the hidden state dimension) are 32, 4, 2, and 4/3, respectively. In multi-task anomaly detection using an isolation forest, the parameters are set as follows: the number of tasks is 3, the random

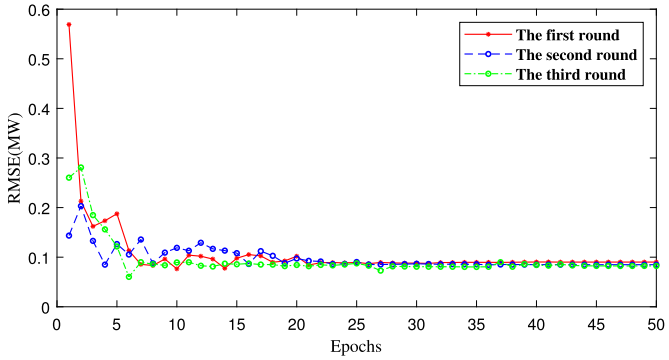

Fig. 5. The RMSE value of the model training in 50 epochs.

seed is 42, the number of decision trees is 200, the anomaly ratio is 0.05, and the interpolation method used is linear interpolation.

The proposed multi-task spatial-temporal interpretable forecasting approach is validated using data from three actual PV power stations in northwestern China, collected in 2019 and 2020. The PV power data employed are obtained from the monitoring and data acquisition systems of PV operators and include total irradiance (Tot), normal direct irradiance (Nor), diffuse horizontal irradiance (Hor), air temperature (Tem), air pressure (press), and relative humidity (Hum), with a resolution of 15 minutes. The above collected data has been transformed into a two-dimensional matrix, where the rows represent potential relevant features of all tasks, and the columns represent the time dimension, as illustrated in Fig. 1. Considering visual effects, Fig. 4 and Table 2 only display the raw data and relevant statistical information of three PV plants for the year 2020. The statistical information of PV power includes the mean, maximum, minimum, and standard deviation (SD). The PV power dataset is partitioned into the training and testing datasets. The training dataset is employed to optimize the weights and biases of the proposed multi-task spatial-temporal forecasting model, as well as to fine-tune the hyperparameters using a trial-and-error approach. The testing dataset is used to evaluate the model's performance in predicting PV power. Considering the seasonal and spatial-temporal variations in PV power output, the test datasets from the three PV power stations are further divided into four parts: spring (the last week of January 2020), summer (the last week of May 2020), autumn (the last week of August 2020), and winter (the last week of November 2020). For each PV test dataset, data from the preceding two months are utilized as the training dataset. Additionally, for improved prediction of multiple power station tasks in PV multi-task spatiotemporal power forecasting, it is necessary to ensure that the time dimensions of the test datasets for the three PV power stations are consistent. The multi-task spatiotemporal interpretable forecasting method and benchmarks presented in this paper are developed using the Python 3.8 environment.

Choosing the maximum number of epochs is crucial for optimizing neural network parameters in multi-task spatial-temporal forecasting of PV power, with typical values ranging from 5 to 100 iterations. The number of epochs selected greatly affects the performance of the PV power prediction model. Insufficient epochs (e.g., < 10) may lead to incomplete feature extraction and underfitting, while excessive epochs (e.g., > 100) may cause overfitting and unnecessary computational overhead. To determine the optimal maximum epoch value and evaluate its impact on model performance, an experiment is conducted using the spring test dataset at a specific timestamp. Multiple simulation trials are performed to analyze the RMSE of the trained model while keeping all other hyperparameters constant. Fig. 5 presents the RMSE values of the trained model over 50 epochs. As shown in the experimental data (Fig. 5), the RMSE values of the three training curves stabilize after the 30th epoch and reach their minimum at 50 epochs. Therefore, the maximum epoch number is set to 50.

4.2. Simulation results and analysis

To validate the effectiveness and superiority of the proposed multi-task spatiotemporal interpretable forecasting approach for PV power, Case 1 first analyzes the anomaly detection performance of the multi-task method based on the isolation forest. Subsequently, Case 2 presents a comparative evaluation of the forecasting performance of the proposed model against eight single-task learning models. In addition, Cases 3 and 4 further verify the proposed model error metric comparisons with five multi-task learning benchmarks under different seasons and varying forecasting horizons. Finally, Case 5 is utilized to elucidate the relationship between the prediction results of the proposed multi-task spatial-temporal forecasting model and the relevant features. Here, except for Case 4, which involves multi-step forecasting, the forecasting horizons for all other cases are set at 15 minutes ahead.

4.2.1. Performance comparison of multi-task outlier detection based on isolated forests

This case analyzes the effectiveness and superiority of the proposed multi-task outlier detection model based on isolated forests (IFMD), taking spring as an example. The three outlier detection methods, namely One-class support vector machine (OCSVM) [58], density-based spatial clustering of applications with noise (DBSCAN) [59], and isolation forest (IF) [60], serve as comparative benchmarks for the proposed multi-task outlier detection approach. Note that, except for the different outlier detection methods, the prediction model and outlier filling are consistent, namely the proposed multi-task spatial-temporal prediction model and linear interpolation. Fig. 6 presents the performance comparison of the proposed multi-task outlier detection method and three benchmarks at PV station No. 3 in spring. The linearized equation in Fig. 6 represents the linear relationship between the predicted PV power values and their corresponding actual values. The coefficient of determination (R^2) is a statistical measure that evaluates the goodness of fit of a model. An R^2 value closer to 1 indicates a better fit of the model to the data. The comparison of error metrics between the proposed multi-task outlier detection method and three benchmark methods in spring is presented in Fig. 7.

As illustrated in Fig. 1, the actual PV power values and predicted values from all four anomaly detection methods demonstrate excellent linear fitting, confirming the feasibility of the proposed multi-task spatiotemporal forecasting model for PV power prediction. Notably, compared with the three benchmark methods, the proposed multi-task anomaly detection approach exhibits the most perfect linear fitting (with R^2 closest to 1) and fewer deviating points. This superior performance can be attributed to its enhanced capability in effectively handling anomalies, which potentially reduces overfitting and improves generalization ability. Fig. 1 reveals that the average MAEs for OCSVM, DBSCAN, IF, and the proposed IFMD across three PV plants are 1.1060, 0.9443, 0.8621, and 0.5637, respectively, with corresponding RMSEs of 2.2380, 1.9739, 1.8321, and 1.2978. Compared to OCSVM, DBSCAN, and IF, the proposed IFMD demonstrates significant improvements, achieving MAE reductions of 96.21%, 67.52%, and 52.94%, respectively, along with RMSE reductions of 72.44%, 52.10%, and 41.17%. Compared with OCSVM and DBSCAN, the proposed IFMD demonstrates superior outlier detection capability. This is primarily because the isolation forest calculates the path length of samples in isolation trees (outliers have shorter paths), which quantifies the degree of outliers and improves detection accuracy. Moreover, it maintains stable performance with high-dimensional data and shows robustness against the curse of dimensionality. The proposed IFMD achieves more satisfactory outlier detection performance than the original IF, mainly due to its simultaneous handling of multiple related tasks, thereby enhancing collaborative detection efficiency and generalization ability.

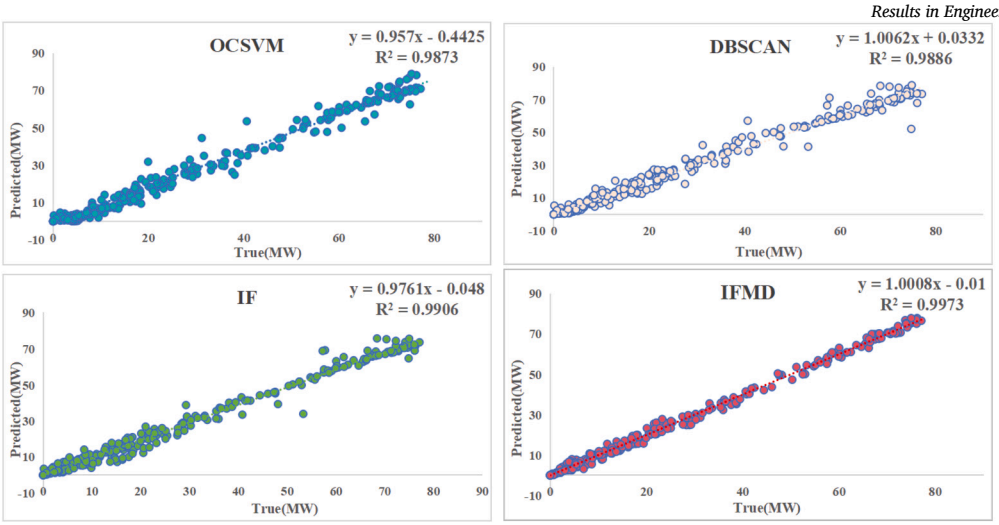

Fig. 6. Performance comparison of the proposed multi-task outlier detection method and three benchmarks at PV station No. 3 in spring.

Table 3
Comparison of the prediction errors of the proposed model and 11 single-task learning models for three power stations in autumn.

Indexes	No.1		No.2		No.3	
	RMSE(MW)	MAE(MW)	RMSE(MW)	MAE(MW)	RMSE(MW)	MAE(MW)
XGBoost	3.0207	1.7363	4.7198	1.9688	3.5859	1.8147
SVR	3.1016	1.7689	4.5367	2.3394	3.4187	1.9184
RNN	2.8637	1.4826	3.9033	1.7731	2.9779	1.2945
LSTM	2.4993	1.3169	3.8495	1.6835	2.9470	1.2524
DCC+MA	2.4720	1.3042	3.5642	1.7441	2.7551	1.2887
XLSTM	2.3517	1.1904	3.5181	1.7615	2.4338	1.2758
XLSTM+GA	2.2919	1.1642	3.3364	1.5941	2.1897	1.2195
Transformer	2.3388	1.3398	3.3459	1.6845	2.4482	1.0627
Transformer+XLSTM	2.2065	1.0982	3.2050	1.6241	2.4396	1.0477
Transformer+DCC	2.2159	1.0709	3.3240	1.5371	2.4419	0.9787
Proposed_SL	2.0138	0.9223	3.0928	1.3161	1.8267	1.0371
Proposed	1.3282	0.9138	1.9783	0.9802	1.2692	1.0293

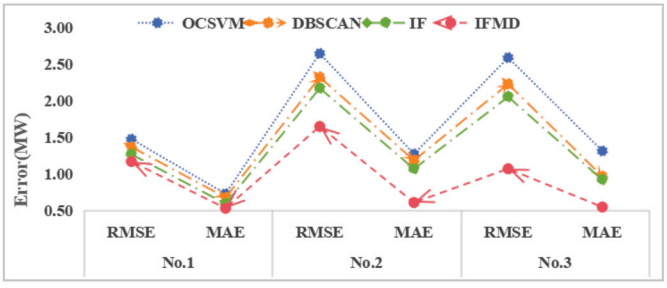

Fig. 7. Comparison of error metrics between the proposed multi-task outlier detection method and three benchmark methods in spring.

4.2.2. Comparison of wind power forecasting performance with single-task learning models

Building upon the multi-task anomaly detection framework in Case 1, this case conducts a comprehensive performance evaluation of the proposed multi-task learning model for PV power predictions through comparative analysis with 11 distinct single-task learning benchmarks, taking autumn as an example. The 11 single-task learning models include XGBoost [18], SVR [16], RNN [19], LSTM, dilated causal convolutional with the multi-head attention (DCC+MA) [61], XLSTM [34], XLSTM with the global attention (XLSTM+GA), Transformer, Transformer with the XLSTM (Transformer+XLSTM), Transformer with the DCC (Transformer+DCC), and the proposed model with the single-learning learning (Proposed_SL). To ensure fair performance comparison, the hyperparameters of XGBoost, SVR, RNN, LSTM, DCC+MA, XLSTM, XLSTM+GA, Transformer, Transformer+XLSTM, Transformer+DCC, and Proposed_SL are optimized using a trial-and-error approach based on the last 7-day PV power data from the training dataset, given their substantial number of tunable parameters. For single-task learning models, the PV power multi-station prediction must follow a sequential approach where each model performs predictions one after another. The proposed multi-task PV power forecasting model and single-task learning models are illustrated in Figs. 8-10. Table 3 presents a comprehensive performance comparison between the proposed model and 11 single-task prediction models for PV power forecasting.

As can be seen from Figs. 8-10, the proposed multi-task spatial-temporal prediction model (red curve) closely matches the actual values (black curve), demonstrating the effectiveness of the proposed multi-task spatial-temporal prediction model for multi-plant PV power forecasting. Furthermore, compared with single-task learning models, the proposed model shows overall better agreement with actual values at peak points, highlighting its performance advantages in multi-task PV power prediction. As shown in Table 3, the proposed multi-task spatial-temporal prediction model achieves MAEs of 0.9138, 0.9802, and 1.0293, and RMSEs of 1.3282, 1.9783, and 1.2692 for the three PV power stations, respectively. The proposed multi-task spatial-temporal prediction model demonstrates superior performance compared to XGBoost, SVR, RNN, and LSTM, achieving average reductions in MAE of 2.2502, 2.1604, 1.7231, and 1.5734, respectively, across three power stations. Correspondingly, the RMSE values decreased by an average of 0.8655, 1.0345, 0.5423, and 0.4432. This significant improvement

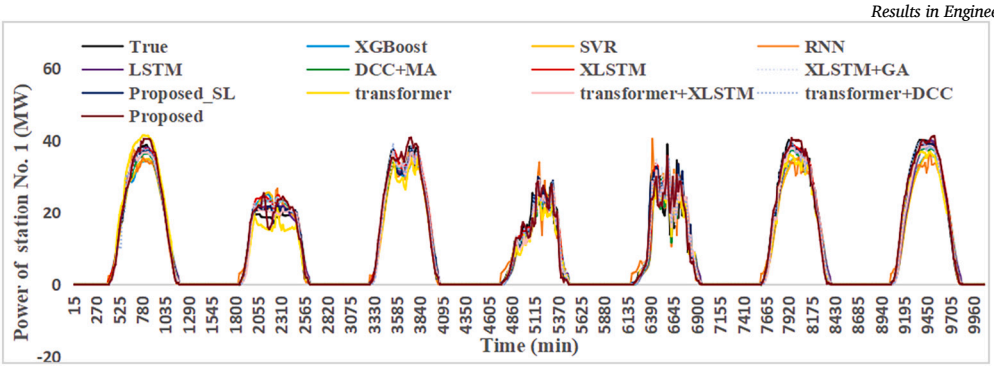

Fig. 8. The prediction curves of the proposed model and 11 single-task learning models for power station No. 1 in autumn.

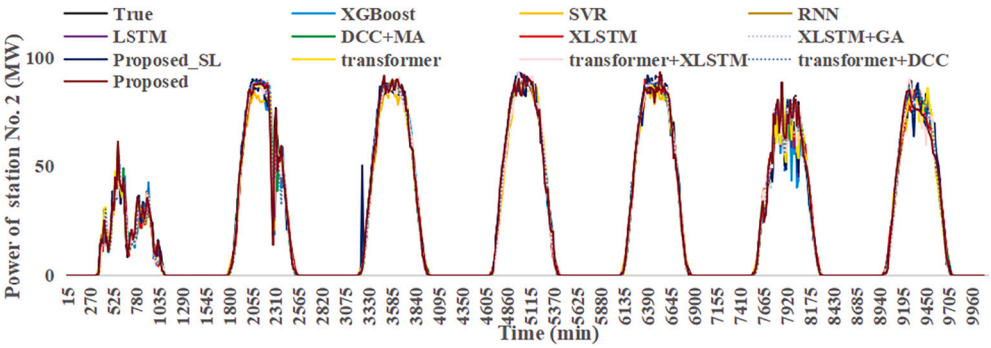

Fig. 9. The prediction curves of the proposed model and 11 single-task learning models for power station No. 2 in autumn.

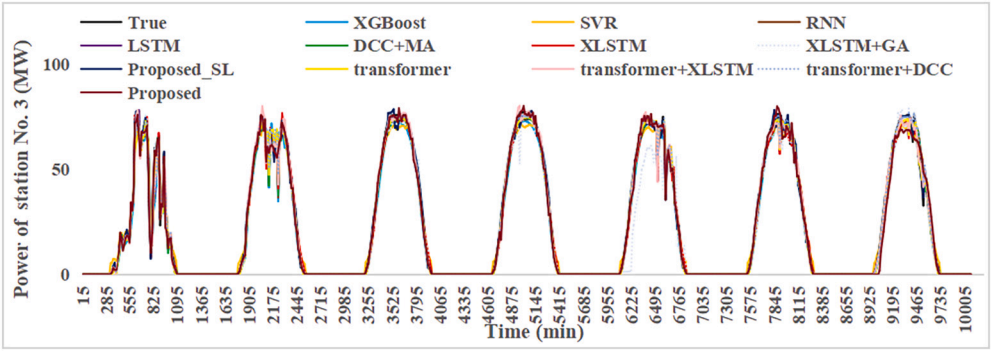

Fig. 10. The prediction curves of the proposed model and 11 single-task learning models for power station No. 3 in autumn.

in both MAE and RMSE metrics highlights the model's enhanced capability in handling the highly nonlinear characteristics of PV power. The limitations of the four baseline models primarily stem from their relatively simplistic network architectures, which struggle to capture the complex spatial-temporal dependencies inherent in PV power data. The proposed multi-task spatial-temporal prediction model outperforms DCC+MA with average reductions of 0.4712 in MAE and 1.4052 in RMSE, primarily attributed to its XLSTM architecture and global attention mechanism that effectively capture long-term dependencies and critical features in PV power time-series data. The proposed multi-task spatial-temporal prediction model achieves superior performance over XLSTM and XLSTM+GA, with average MAE reductions of 0.4348 and 0.3515, and RMSE reductions of 1.2426 and 1.0808, respectively. This improvement stems from its integration of rotary position embedding, multi-head attention mechanisms, and dilated causal convolution to simultaneously capture global spatial patterns and local spatial features in PV power time-series data. Additionally, it can be observed that XLSTM+GA exhibits smaller prediction errors compared to XLSTM. This is primarily because the introduction of global attention into the MLSTM of XLSTM enables more effective capture of long-range dependencies and dynamic key features in PV power time series. Compared with Transformer, Transformer XLSTM, and Transformer DCC, the proposed multi-task learning prediction model achieves reductions in MAE of 0.3879, 0.2822, and 0.2211, respectively, and reductions in RMSE of 1.1857, 1.0918, and 1.1354, respectively. The primary reason for the decrease in these error metrics is the full utilization of the advantages of XLSTM and DCC. The proposed multi-task spatial-temporal prediction model demonstrates superior performance over the Proposed-SL model, achieving average reductions of 0.1174 in MAE and 0.7859 in RMSE across three PV power stations. This improvement primarily stems from its multi-task learning architecture, which effectively captures feature correlations among multiple PV stations while enhancing the model generalization capability.

4.2.3. Comparison of PV power forecasting performance with the multi-task learning benchmark models

To comprehensively validate the superiority of the proposed model in PV power forecasting, this case is conducted by comparing it with the multi-task learning benchmark model across four seasons. The graph convolutional networks-based multi-task learning model (GCN+ML)

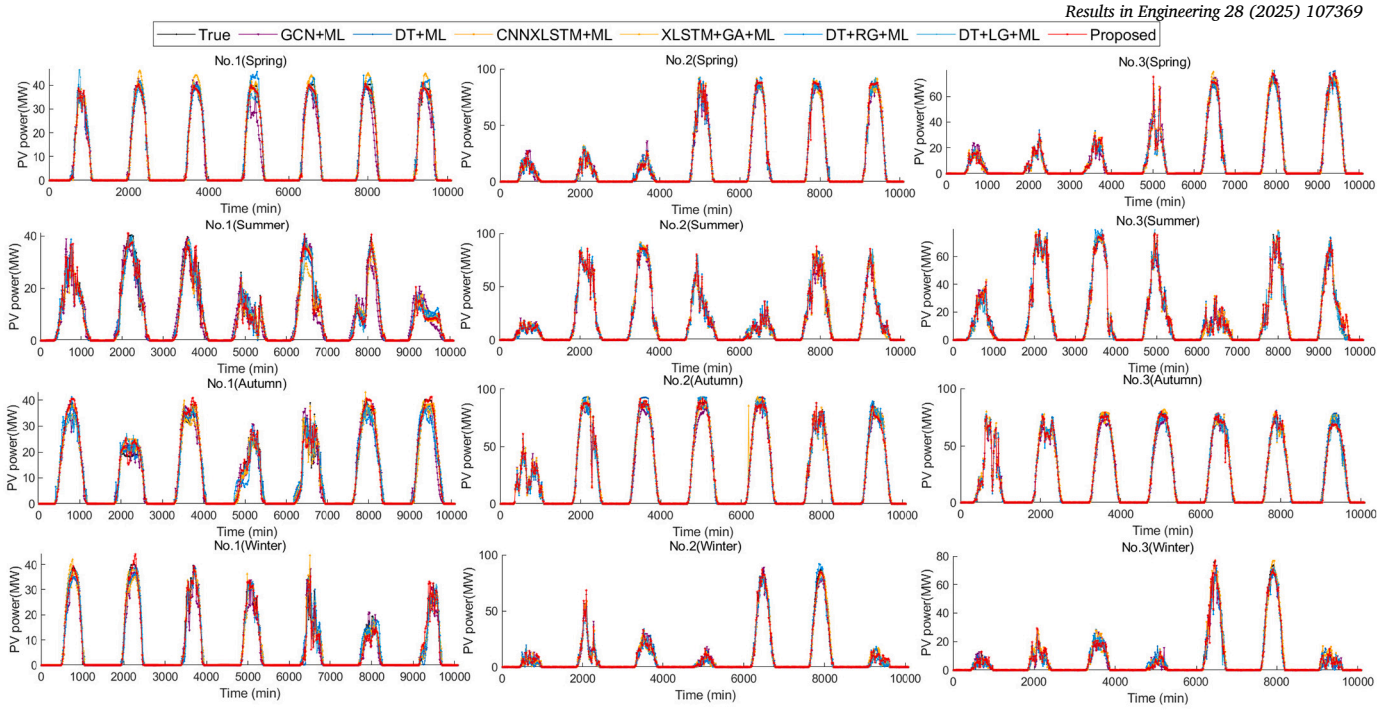

Fig. 11. The PV power predicted results of the proposed model and multi-task learning benchmarks of three PV stations in four seasons.

[62,63], DCC+Transformer-based multi-task learning model (DT+ML), CNNXLSTM-based multi-task learning model (CNNXLSTM+ML), XLSTM+GA-based multi-task learning model (XLSTM+GA+ML), DCC+Transformer+RNN+GA-based multi-task learning model (DT+RG+ML), and DCC+Transformer+LSTM+GA-based multi-task learning model (DT+LG+ML) are used as prediction performance comparison approaches. The benchmark model for multi-task learning incorporates various architectures in its task-sharing layer, specifically utilizing GCN, DT, CNNXLSTM, XLSTM+GA, DT+RG, and DT+LG. For the task-specifying layer, a fully connected neural network is employed. The method for determining the hyperparameters of these benchmark models remains consistent with the approach outlined in Section 4.2.2. Fig. 11 presents the PV power forecasting results of the proposed model and the multi-task learning benchmark models across three PV power stations in the four seasons. Fig. 12 shows a comparison of the MAE and RMSE statistical results between the proposed model and the multi-task learning benchmark models for the three PV power stations in the four seasons. Fig. 13 shows the comparison of the convergence speeds of the proposed multi-task spatial-temporal prediction model and five benchmarks at 10:00 AM on the first day of the spring test set.

As can be seen from Fig. 11, the dark red curves (i.e., the proposed model) almost coincide with the black curves (i.e., the actual values) in the power prediction of three PV power stations across four seasons. However, the purple curves (GCN+ML), blue curves (DT+ML), orange curves (CNNXLSTM+ML), yellow curves (XLSTM+GA+ML), green curves (DT+RG+ML), and cyan curves (DT+LG+ML) show a certain deviation from the black curves. This coincidence with the black curves indicates that the proposed multi-task PV power prediction model achieves the most satisfactory prediction results compared to the baseline models. Similarly, it can be observed that the dark red curve at the extreme values in Fig. 11 better follows the black curve, indicating that the proposed multi-task learning model can more effectively handle complex feature extraction applications. Additionally, when comparing the various subplots in Fig. 11, there are significant differences in the degree of overlap between the dark red and black curves. The primary reason for this lies in the variations in the generalization ability of the proposed multi-task prediction model across different geographical locations and meteorological conditions of the PV stations.

From Fig. 12, it can be observed that the MAEs of the proposed multi-task PV power prediction model at the three power stations in spring respectively are 0.5317, 0.6101, and 0.5493, with an average value of 0.5637. Additionally, the RMSEs of the proposed model at these three power stations in spring respectively are 1.1718, 1.6498, and 1.0717, with an average value of 1.2978. In spring, the average MAEs of the multi-task learning comparison models, such as GCN+ML, DT+ML, CNNXLSTM+ML, XLSTM+GA+ML, DT+RG+ML, and DT+LG+ML, at the three power stations are 1.6721, 1.2054, 1.0601, 0.9577, 1.1411, and 0.9570, respectively. Correspondingly, their average RMSEs are 3.4809, 2.3187, 2.2496, 1.9447, 2.2689, and 1.8045, respectively. Compared to the models GCN+ML, DT+ML, CNNXLSTM+ML, XLSTM+GA+ML, DT+RG+ML, and DT+LG+ML, the proposed multi-task prediction model exhibits reductions in the average MAEs at the three PV stations in spring by 1.1084, 0.6417, 0.4964, 0.3940, 0.5774, and 0.3933, respectively. Additionally, the corresponding average RMSEs are reduced by 2.1831, 1.0209, 0.9518, 0.6469, 0.9711, and 0.5067, respectively. The inferior performance of GCN+ML in PV power forecasting primarily stems from its static graph structure and limited temporal feature extraction capabilities. Unlike dynamic models, GCN relies on a predefined adjacency matrix, which restricts its ability to adaptively capture spatially evolving correlations caused by transient meteorological factors (e.g., moving cloud cover). Additionally, its fixed spatial dependency modeling fails to account for the asymmetric and long-range temporal interactions inherent in PV systems, where historical weather patterns exhibit non-stationary influences on current power generation. In contrast, sequential models like Transformers or hybrid XLSTM-based approaches demonstrate superior adaptability by dynamically adjusting both spatial and temporal feature representations. The proposed multi-task PV power prediction model achieves relatively lower MAE and RMSE values primarily due to two key innovations. (1) Spatial Feature Extractor: The novel spatial feature extractor integrates rotary position embedding, Transformer, DCC, and residual connection. This architecture enables joint modeling of local and global features for multi-task PV power prediction. (2) Temporal Feature Extractor: The proposed temporal feature extractor employs XLSTM with a global attention mechanism, which effectively captures long-term dependencies and identifies critical features in PV power data.

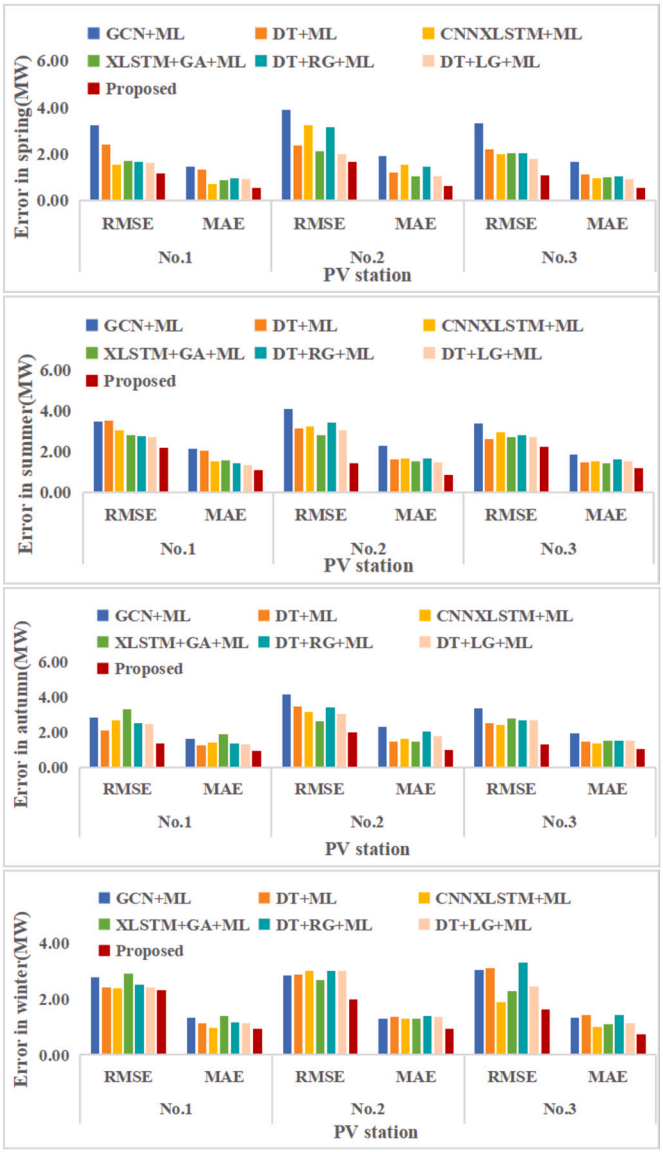

Fig. 12. Comparison of error performance between the proposed model and the multi-task learning benchmark models across four different seasons.

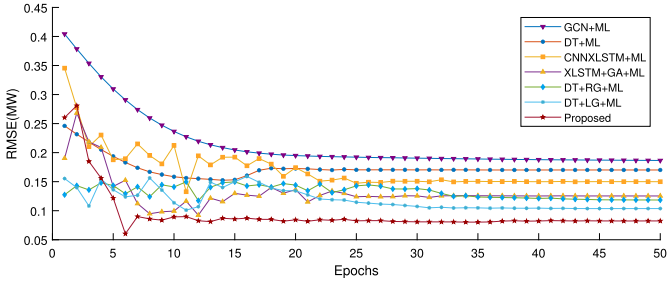

Fig. 13. Comparison of the convergence speeds of the proposed model and five benchmarks.

As can be seen from Fig. 13, the proposed multi-task spatial-temporal forecasting model exhibits the fastest convergence speed, which can be primarily attributed to the following two aspects: (1) Spatial feature extractor: Dilated convolutions expand the receptive field to capture local spatial features, while the Transformer with rotary position embedding focuses on global spatial features. This design enables the model to rapidly capture both global and local features of PV power station outputs during the early stages of training, thereby accelerating convergence. (2) Temporal feature extractor: The XLSTM is capable of processing long-term sequential data, and the global attention mechanism dynamically adjusts the weights of different time steps to highlight critical temporal features. This design facilitates the capture of the impacts of meteorological dynamics on PV power station outputs, thus expediting the model convergence. The RMSE of the model decreases rapidly at the beginning of training, then experiences a period of increase, and finally stabilizes. This may be due to adjustments in the learning rate during the training process. Initially, a larger learning rate allows the model to converge quickly, but subsequently, an excessively considerable learning rate may cause the model to oscillate near a local optimum. By employing optimization strategies to reduce the learning rate, the model can fine-tune its parameters and ultimately stabilize.

4.2.4. Comparison with the multi-step prediction performance of the multi-task learning benchmark models

To further verify the superiority of the proposed multi-task spatial-temporal prediction model, a series of multi-step prediction experiments are carried out in this case, taking summer as an example. The adopted comparison models are the DT+ML, CNNXLSTM+ML, XLSTM+GA+ML, DT+RG+ML, and DT+LG+ML, and their methods for determining hyperparameters are the same as those in Section 4.2.2. The multi-step prediction range of the proposed multi-task spatial-temporal prediction model is 15 minutes to 90 minutes in advance. The PV power multi-step forecasting methods can be primarily categorized into three types: the multi-output strategy (a single model simultaneously predicts multiple steps), the recursive strategy (iteratively using previous predictions as inputs), and the direct strategy (training independent models for each forecasting step) [23]. Among these, the direct strategy constructs multiple dedicated PV power forecasting models, effectively avoiding the multi-output strategy neglect of adjacent features and the recursive strategy error accumulation issue. Therefore, in this paper, the direct strategy is adopted to model the multi-task spatial-temporal prediction of PV power. Figs. 14-16 present the comparison of the multi-step prediction performance of PV power between the proposed model and the benchmark models in PV power stations. The multi-step prediction error statistics of the proposed PV power multi-task spatial-temporal forecasting model and five benchmarks are shown in Fig. 17.

It can be seen from Figs. 14-16 that the proposed multi-task spatial-temporal model has the lowest RMSE and MAE values in different multi-step predictions of the three PV power stations, indicating that it is superior to other comparison models and has higher prediction accuracy. As evidenced in Figs. 14-16, both the proposed multi-task spatial-temporal forecasting model and the five benchmark models demonstrate superior performance with lower prediction errors during the 15-45 minute horizon, followed by progressively increasing deviations. This temporal pattern primarily arises from cloud cover dynamics being the dominant factor governing PV power output fluctuations. The enhanced accuracy within the 45-minute window benefits from the quasi-linear continuity characteristics of cloud advection movements, whereas longer-term predictions must account for nonlinear cloud behaviors including abrupt dissipation/formation and splitting-merging events. Such complex spatial-temporal interdependencies pose significant challenges for modeling frameworks to capture.

From Fig. 17, it can be observed that the minimum RMSE values of the proposed multi-task spatiotemporal prediction model for the three PV stations are 2.1621, 1.4019, and 2.2244, respectively; the maximum RMSE values are 4.6061, 4.6598, and 4.6527, respectively; and the median RMSE values are 2.9263, 3.0714, and 2.7225, respectively. The median RMSE values of the five benchmark models for PV station No. 1 are 5.0559, 4.7317, 4.2395, 3.9962, and 2.9263, respectively. For PV station No. 2, the median RMSE values are 5.5208, 5.22835, 5.179, 5.10015, and 4.8834, respectively. For PV station No. 3, the median RMSE values are 5.5334, 5.1650, 5.4397, 4.4880, and 4.3075, respectively. It is evident that the proposed multi-task spatial-temporal

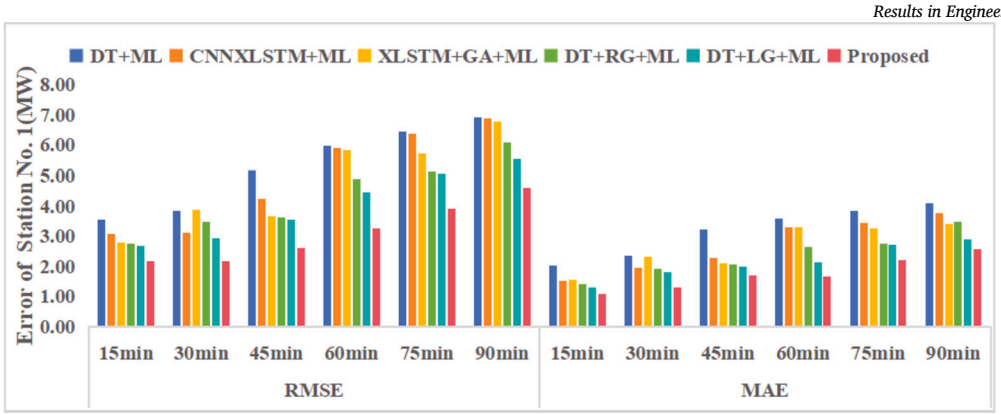

Fig. 14. Comparison of the multi-step prediction performance of PV power between the proposed model and the benchmark models in PV Power Station No. 1.

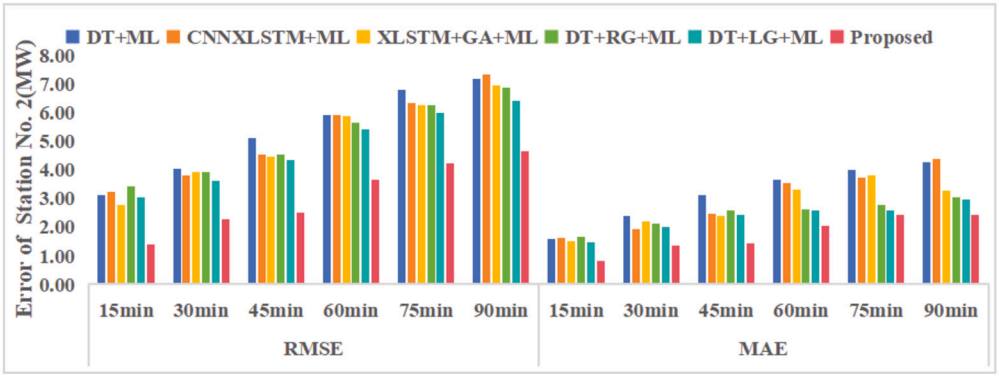

Fig. 15. Comparison of the multi-step prediction performance of PV power between the proposed model and the benchmark models in PV power station No. 2.

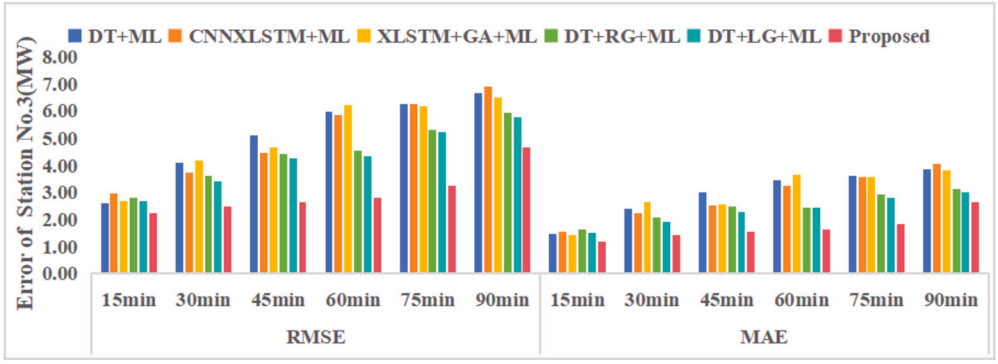

Fig. 16. Comparison of the multi-step prediction performance of PV power between the proposed model and the benchmark models in PV power station No. 3.

prediction model exhibits the best predictive performance, followed by DT+LG+ML, DT+RG+ML, XLSTM+GA+ML, CNNXLSTM+ML, and DT+ML. The proposed multi-task spatial-temporal prediction model demonstrates superior performance in PV power forecasting compared to DT+LG+ML and DT+RG+ML. This enhancement is attributed to its advanced dynamic state updating mechanism and efficient parallel processing capabilities, achieved through optimized memory architecture, increased model flexibility, and improved computational efficiency. Compared with the DT+ML, the superior predictive performance of the proposed model is primarily attributed to the introduction of not only XLSTM but also a global attention mechanism, which enables better capture of critical temporal features. Compared with XLSTM+GA+ML, the proposed model exhibits superior predictive performance, primarily due to the introduction of DCC and rotary position embedding, which can better capture local spatial features. From the above analysis, we can further demonstrate the effectiveness of the proposed multi-task spatial-temporal model in PV power prediction based on the spatial-temporal feature extractors.

4.2.5. *SHAP-based multi-task interpretability of PV power forecasting*

The last case is employed to analyze the mapping relationships between the multi-task spatial-temporal output results of PV power and their associated features based on the SHAP interpretability model, using the prediction 15 minutes in advance during spring as an example. Figs. 18-20 illustrate the SHAP values of associated features in the multi-task spatial-temporal output results of PV power for three PV stations. The notation $t - N$ signifies the time point that is $15 * N$ minutes ahead of the prediction moment t. PV1, PV2, and PV3 correspond to the power outputs of three different power plants, with other related features labeled in a similar way. Figs. 18-20 visualize feature importance using color (black-to-red gradient for low-to-high values) and point size.

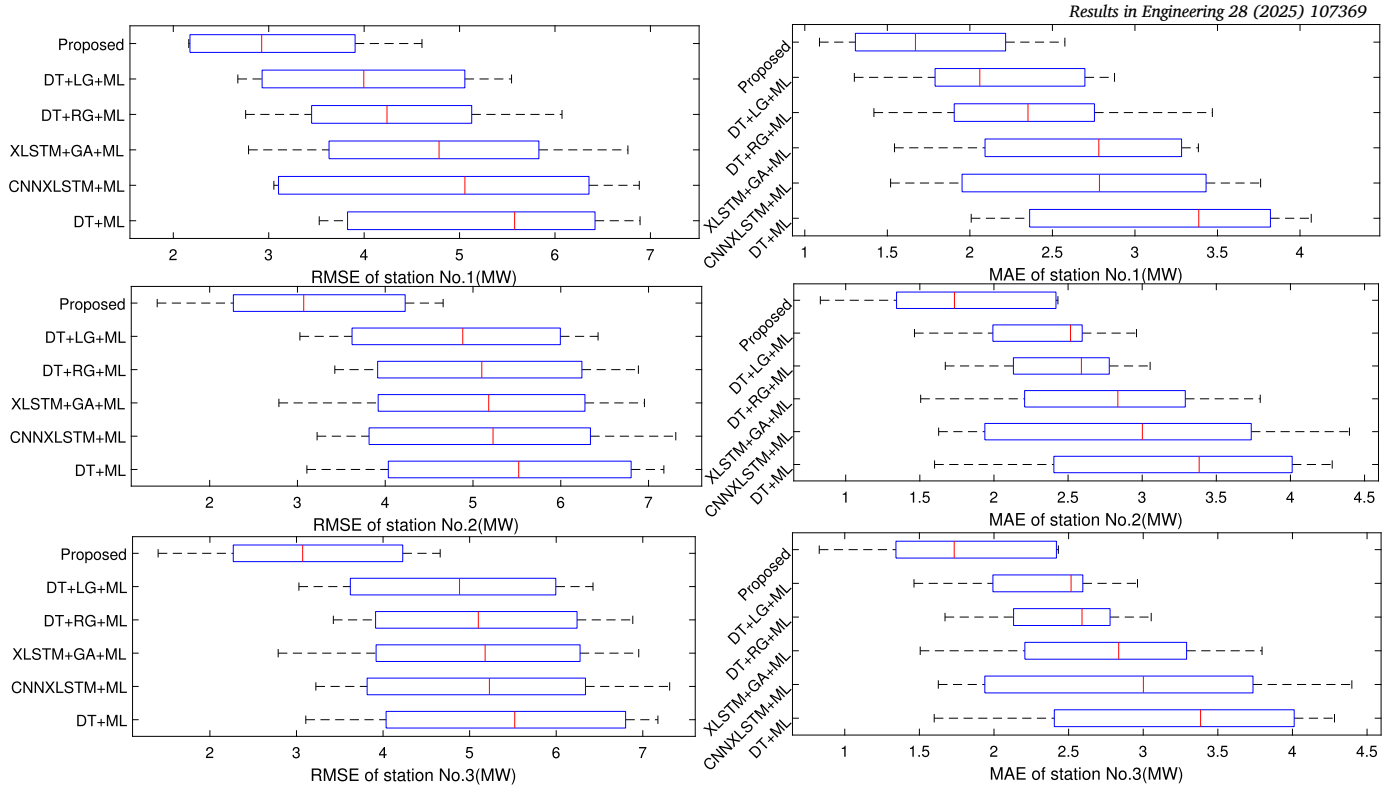

Fig. 17. Multi-step prediction error statistics of the proposed PV power multi-task spatial-temporal forecasting model and five benchmarks.

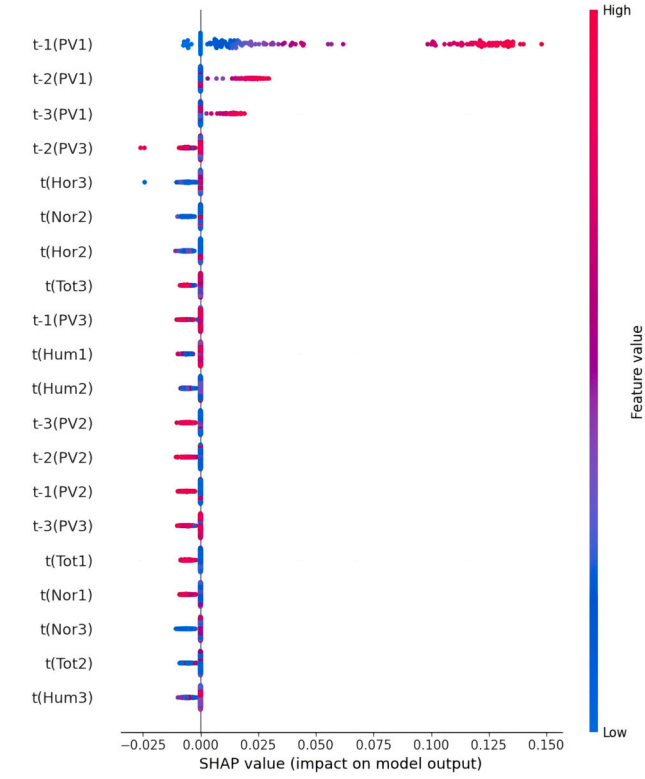

Fig. 18. The interpretability of the multi-task spatial-temporal prediction results and correlated features for PV Station No. 1.

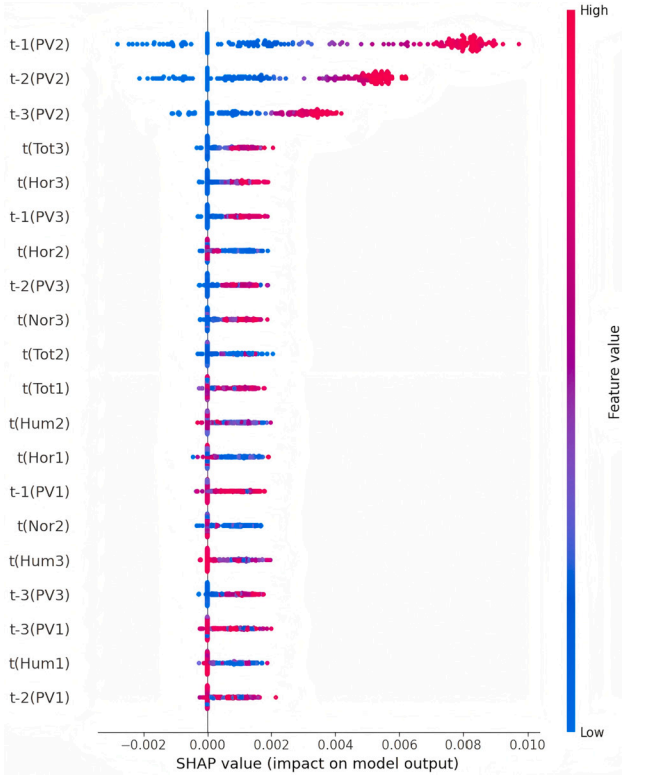

Fig. 19. The interpretability of the multi-task spatial-temporal prediction results and correlated features for PV Station No. 2.

Red-highlighted features exhibit a more substantial influence on model outputs at corresponding time points.

From Figs. 18-20 show that the relevant features of the three PV stations have a significant impact on the prediction for each station. This indicates that the temporal and spatial correlation of data between different PV stations significantly affects the prediction results, further demonstrating the effectiveness and superiority of the proposed multi-task spatial-temporal interpretable prediction approach. As illustrated

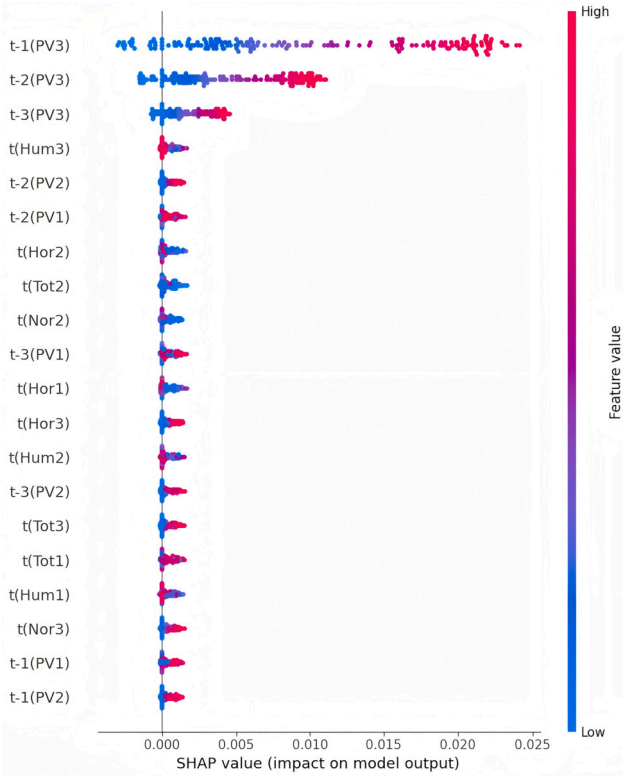

Fig. 20. The interpretability of the multi-task spatial-temporal prediction results and correlated features for PV Station No. 3.

in Fig. 18, the PV power data from 15 minutes ahead t-1(PV1) up to 45 minutes ahead t-3(PV1) of PV station No. 1 displays the greatest number of red dots (denoting larger SHAP values). This indicates that these particular features exert the most substantial influence on the PV power prediction outcomes for PV station No. 1, which is largely attributable to the time-lag effect presented in PV power output. Some SHAP values of the feature t-1 (PV1) are shown on the left side in Fig. 18. In specific prediction scenarios, such as those involving extreme weather conditions, this feature can negatively impact the model's prediction performance due to the combined effects of data distribution characteristics and changes in the model structure.

It can be observed from Fig. 19 total irradiance t(Tot3), diffuse horizontal irradiance t(Hor3) of PV station No. 3 significantly influence the PV power forecasting for PV station No. 2. This is primarily due to the powerful spatial correlation between these features, making them highly significant in explaining the power variations at PV station No. 2.

As shown in the SHAP interpretability Fig. 20 for multi-station multi-task spatial-temporal prediction of PV power at station No. 3, the SHAP value distribution for t-2(PV2) and t-2(PV1) reveals their significant impact on the prediction results, which can be attributed to several factors: (1) Temporal Correlation: The power data from stations No. 1 and No. 2 at t-2(PV2) and t-2(PV1), which represent the power levels 30 minutes and 15 minutes before the prediction time at station No. 3, respectively, provide historical context. If there are consistent time patterns in PV power across the different stations, the model can leverage this information to improve predictions for station No. 3. For instance, if stations No. 1 and No. 2 show a gradual increase or decrease in power generation leading up to the prediction time, the model might anticipate a similar trend at station No. 3. (2) Spatial Correlation: The influence of t-2(PV2) and t-2(PV1) may also stem from spatial correlations between the stations. Shared environmental factors, such as weather conditions, cloud movement, and geographical proximity, can lead to similar power generation patterns across stations.

5. Conclusion

In this paper, we propose a novel interpretable forecasting approach for multi-PV power stations based on spatial-temporal multi-task learning with Transformer-XLSTM. The proposed approach begins with a spatial feature extractor that combines rotary position embedding, Transformer, DCC, and residual connection to capture both global and local PV power patterns. For temporal modeling, an XLSTM-based feature extractor enhanced with global attention mechanisms is developed to learn long-term dependencies and key temporal features. Building on these components, a multi-task spatial-temporal prediction model is constructed to model the coupling relationships among PV stations. In addition, to address data reliability, the approach incorporates an isolation forest-based multi-task outlier detection module, which actively filters anomalies from historical PV power data. Finally, a SHAP-based model is integrated into the proposed methodology to elucidate the coupled feature interactions in PV power forecasting and their causal relationships with multi-task output variables. Empirical validation has been performed in case studies using operational datasets from PV power stations in Western China.

Based on experimental results and case analysis of the proposed multi-task spatial-temporal interpretable forecasting approach, the principal findings of this research are as follows: (1) The proposed IFMD outperforms OCSVM, DBSCAN, and IF by substantial margins, delivering relative reductions of 96.21%, 67.52%, and 52.94% in MAE, and 72.44%, 52.10%, and 41.17% in RMSE, respectively. The enhanced performance of the IFMD primarily stems from isolation forest-based path length measurement (shorter paths indicate outliers), which quantifies anomaly scores more precisely, while maintaining robust performance in high-dimensional spaces and effectively handling multiple correlated tasks through joint optimization for improved detection efficiency and generalization. (2) Compared with XGBoost, SVR, RNN, LSTM, DCC+MA, XLSTM, XLSTM+GA, Transformer, Transformer+XLSTM, Transformer+DCC, and proposd_SL, the proposed multi-task spatiotemporal forecasting model achieves MAE reductions of 0.8655, 1.0345, 0.5423, 0.4432, 0.4712, 0.4348, 0.3515, 0.3879, 0.2822, 0.2211, and 0.1174, respectively, along with corresponding RMSE decreases of 2.2502, 2.1604, 1.7231, 1.5734, 1.4052, 1.2426, 1.0808, 1.1857, 1.0918, 1.1354, and 0.7859. (3) The proposed model demonstrates significant performance improvements over the baseline methods (GCN+ML, DCC+MA+ML, CNNXLSTM+ML, XLSTM+GA+ML, DT+RG+ML, and DT+LG+ML) in terms of both MAE and RMSE metrics across three PV stations during spring. Specifically, it achieves average MAE reductions of 1.1084, 0.6417, 0.4964, 0.3940, 0.5774, and 0.3933, respectively, while the corresponding average RMSE values decrease by 2.1831, 1.0209, 0.9518, 0.6469, 0.9711, and 0.5067. (4) Compared with DT+ML, CNNXLSTM+ML, XLSTM+GA+ML, DT+RG+ML, and DT+LG+ML, the proposed multi-task spatial-temporal prediction model reduces RMSE by a minimum of 42.02%, a maximum of 70.72%, and an average of 57.55% in the multi-step prediction of PV power. Similarly, the MAE decreased by a minimum of 30.35%, a maximum of 78.65%, and an average of 54.32%. The enhanced predictive performance of the proposed multi-task spatial-temporal forecasting model primarily stems from two key architectural innovations: first, the temporal feature extractor based on the XLSTM with global attention mechanism, which optimizes memory architecture to improve model flexibility and capture critical temporal patterns; second, the spatial feature extractor incorporating Transformer and causal convolution operations, which synergistically enables comprehensive extraction of both global and localized spatial features. (5) The SHAP model can be utilized to interpret and analyze the relationships between the multi-task predictive variables of the proposed multi-task spatial-temporal prediction approach and their associated features. The above comprehensive cases confirm the accuracy and superiority of the proposed interpretable forecasting approach, demonstrating that the Transformer-XLSTM-based

spatial-temporal multi-task learning model holds significant potential for practical PV power forecasting in electric energy systems.

CRediT authorship contribution statement

Rongquan Zhang: Writing – original draft, Methodology, Investigation, Formal analysis, Conceptualization. **Xiupeng Wan:** Visualization, Software, Data curation. **Siqi Bu:** Writing – review & editing, Supervision, Funding acquisition, Formal analysis. **Min Zhou:** Writing – review & editing, Visualization, Supervision. **Qiangqiang Zeng:** Visualization, Software. **Zhe Zhang:** Writing – review & editing, Funding acquisition, Formal analysis, Conceptualization.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgements

This work was jointly supported by the Science and Technology Research Project of the Jiangxi Provincial Department of Education under Grant Nos. GJJ2403005 and GJJ2403008, in part by the Natural Science Foundation of Top Talent of SZTU under Grant No. GDRC202313, and in part by the National Natural Science Foundation of China under Grant No. 62305232.

Data availability

Data will be made available on request.

References

[1] M.A. Atiea, A.A. Abdelghaffar, H.B. Aribia, F. Noureddine, A.M. Shaheen, Photovoltaic power generation forecasting with Bayesian optimization and stacked ensemble learning, Results Eng. 26 (2025) 104950.

[2] M.O. Faruque, M.M. Islam, M.J. Talukder, A. Mia, S. Tasnim, M.A. Hossain, S. Muyeen, Enhancing microgrid forecasting accuracy with at cnn-tls framework: a novel approach to mitigating uncertainty in renewable energy and load predictions, Results Eng. 27 (2025) 105606.

[3] R. Chen, C. Ren, C. Liao, Y. Huang, Z. Liu, Spatial dynamics of per capita building carbon emissions in the greater bay area: pathways to net zero carbon by 2060, Build. Environ. 270 (2025) 112501.

[4] R.S. Kumar, P. Meera, V. Lavanya, S. Hemamalini, Brown bear optimized random forest model for short term solar power forecasting, Results Eng. 25 (2025) 104583.

[5] W. Liu, M. Gai, Pv-mlp: a lightweight patch-based multi-layer perceptron network with time–frequency domain fusion for accurate long-sequence photovoltaic power forecasting, Renew. Energy 251 (2025) 123277.

[6] S. Rahman, S. Saha, M. Haque, S. Islam, M. Arif, M. Mosadeghy, A. Oo, A framework to assess voltage stability of power grids with high penetration of solar pv systems, Int. J. Electr. Power Energy Syst. 139 (2022) 107815.

[7] R. Zhang, S. Bu, G. Li, Multi-market p2p trading of cooling–heating-power-hydrogen integrated energy systems: an equilibrium-heuristic online prediction optimization approach, Appl. Energy 367 (2024) 123352.

[8] N. Sushmi, D. Subbulekshmi, Real-time ultra short-term irradiance forecasting using a novel R-GRU model for optimizing pv controller dynamics, Results Eng. 26 (2025) 105046.

[9] C. Cai, L. Zhang, J. Zhou, L. Zhou, Sky images based photovoltaic power forecasting: a novel approach with optimized vmd and vision mamba, Results Eng. 24 (2024) 103022.

[10] H. Dai, Z. Zhen, F. Wang, Y. Lin, F. Xu, N. Duić, A short-term pv power forecasting method based on weather type credibility prediction and multi-model dynamic combination, Energy Convers. Manag. 326 (2025) 119501.

[11] R. Zhang, G. Li, S. Bu, G. Kuang, W. He, Y. Zhu, S. Aziz, A hybrid deep learning model with error correction for photovoltaic power forecasting, Front. Energy Res. 10 (2022) 948308.

[12] B.U.D. Abdullah, S.A. Khanday, N.U. Islam, S. Lata, H. Fatima, S.H. Nengroo, Comparative analysis using multiple regression models for forecasting photovoltaic power generation, Energies 17 (2024) 1564.

[13] S. A, M.S. Christo, J.V. Elizabeth, A hybrid approach to time series forecasting: integrating arima and prophet for improved accuracy, Results Eng. 27 (2025) 105703.

[14] J. Munkhammar, D. van der Meer, J. Widén, Very short term load forecasting of residential electricity consumption using the Markov-chain mixture distribution (mcm) model, Appl. Energy 282 (2021) 116180.

[15] X. Cheng, R. Zhang, S. Bu, A data-driven approach for collaborative optimization of large-scale electric vehicles considering energy consumption uncertainty, Electr. Power Syst. Res. 221 (2023) 109461.

[16] Y. ÖNAL, Gaussian kernel based svr model for short-term photovoltaic mpp power prediction, Comput. Syst. Sci. Eng. 41 (2022).

[17] D. El Bourakadi, H. Ramadan, A. Yahyaouy, J. Boumhidi, A novel solar power prediction model based on stacked bilstm deep learning and improved extreme learning machine, Int. J. Inf. Technol. 15 (2023) 587–594.

[18] J. Polo, N. Martín-Chivelet, M. Alonso-Abella, C. Sanz-Saiz, J. Cuenca, M. de la Cruz, Exploring the pv power forecasting at building façades using gradient boosting methods, Energies 16 (2023) 1495.

[19] T. Dewi, E.N. Mardiyati, P. Risma, Y. Oktarina, Hybrid machine learning models for pv output prediction: harnessing random forest and lstm-rnn for sustainable energy management in aquaponic system, Energy Convers. Manag. 330 (2025) 119663.

[20] A. Ait Mansour, A. Tilioua, M. Touzani, Bi-lstm, gru and 1D-cnn models for short-term photovoltaic panel efficiency forecasting case amorphous silicon grid-connected pv system, Results Eng. 21 (2024) 101886.

[21] S. Iqbal, M. Shafiullah, M. Aurangzeb, I. Jamil, A. Rehman, A. Islam, A. Ali, S.S. Alharbi, Forecasting large-scale solar power plant energy production based on Monte Carlo simulations and long-short-term memory, Results Eng. 27 (2025) 106269.

[22] J. Kim, J. Obregon, H. Park, J.-Y. Jung, Multi-step photovoltaic power forecasting using transformer and recurrent neural networks, Renew. Sustain. Energy Rev. 200 (2024) 114479.

[23] R. Zhang, S. Bu, Y. Zheng, G. Li, X. Wan, Q. Zeng, M. Zhou, A novel multi-task learning model based on transformer-lstm for wind power forecasting, Int. J. Electr. Power Energy Syst. 169 (2025) 110732.

[24] C. Song, H. Yang, J. Cai, P. Yang, H. Bao, K. Xu, X.-B. Meng, Multi-energy load forecasting via hierarchical multi-task learning and spatiotemporal attention, Appl. Energy 373 (2024) 123788.

[25] Z. Tian, Y. Chen, G. Wang, Enhancing pv power forecasting accuracy through nonlinear weather correction based on multi-task learning, Appl. Energy 386 (2025) 125525.

[26] H. Wang, J. Yan, J. Zhang, S. Liu, Y. Liu, S. Han, T. Qu, Short-term integrated forecasting method for wind power, solar power, and system load based on variable attention mechanism and multi-task learning, Energy 304 (2024) 132188.

[27] Y. Ju, J. Li, G. Sun, Ultra-short-term photovoltaic power prediction based on self-attention mechanism and multi-task learning, IEEE Access 8 (2020) 44821–44829, https://doi.org/10.1109/ACCESS.2020.2978635.

[28] H. Song, N.A. Khafaf, A. Kamoona, S.S. Sajjadi, A.M. Amani, M. Jalili, X. Yu, P. McTaggart, Multitasking recurrent neural network for photovoltaic power generation prediction, Energy Rep. 9 (2023) 369–376, 2022 the 3rd International Conference on Power, Energy and Electrical Engineering.

[29] H. Zang, D. Chen, J. Liu, L. Cheng, G. Sun, Z. Wei, Improving ultra-short-term photovoltaic power forecasting using a novel sky-image-based framework considering spatial-temporal feature interaction, Energy 293 (2024) 130538.

[30] J. Simeunović, B. Schubnel, P.-J. Alet, R.E. Carrillo, P. Frossard, Interpretable temporal-spatial graph attention network for multi-site pv power forecasting, Appl. Energy 327 (2022) 120127.

[31] T.-B. Li, A.-A. Liu, D. Song, W.-H. Li, J. Zhang, Z.-Q. Wei, Y.-T. Su, Multi-task spatial-temporal transformer for multi-variable meteorological forecasting, IEEE Trans. Knowl. Data Eng. 36 (2024) 8876–8888.

[32] C. Li, M. Wu, Y. Wu, Z. Zhang, Y. Qin, K. Sun, Photovoltaic power forecasting using multi-task learning considering spatio-temporal coupling relationship, in: Electrical Artificial Intelligence Conference, Springer, 2024, pp. 495–501.

[33] Y. Chen, J.-W. Xiao, Y.-W. Wang, Y. Li, Regional wind-photovoltaic combined power generation forecasting based on a novel multi-task learning framework and tpa-lstm, Energy Convers. Manag. 297 (2023) 117715.

[34] M. Beck, K. Pöppel, M. Spanring, A. Auer, O. Prudnikova, M. Kopp, G. Klambauer, J. Brandstetter, S. Hochreiter, XLSTM: extended long short-term memory, arXiv preprint, arXiv:2405.04517, 2024.

[35] M. Huang, L. Yang, G. Jiang, X. Hao, H. Lu, H. Luo, P. Wang, J. Li, Rescconv-XLSTM: an improved XLSTM model with spatiotemporal feature extraction capability for remaining useful life prediction of aero-engine, Results Eng. 26 (2025) 105513.

[36] Z. Barbre, G. Li, Enhanced wind energy forecasting using an extended long short-term memory model, Algorithms 18 (2025) 206.

[37] B. Ozdemir, I. Pacal, An innovative deep learning framework for skin cancer detection employing convnextv2 and focal self-attention mechanisms, Results Eng. 25 (2025) 103692.

[38] W. Liu, Z. Mao, Short-term photovoltaic power forecasting with feature extraction and attention mechanisms, Renew. Energy 226 (2024) 120437.

[39] T.-Y. Deng, W.-L. Li, F. Zhang, Q. Hua, C.-R. Dong, B.H. Lim, A multiscale global-local transformer for long-sequence pv power generation forecasting, in: International Conference on Parallel and Distributed Computing: Applications and Technologies, Springer, 2024, pp. 613–625.

[40] J. Tully, R. Haight, B. Hutchinson, S. Huang, J.-Y. Lee, S. Katipamula, Dilated causal convolutional neural networks for forecasting zone airflow to estimate short-term energy consumption, Energy Build. 286 (2023) 112890.

[41] M. Zhang, M.Z. Yousif, L. Yu, H.-C. Lim, Enhanced retrospective forecasting in dissipative dynamical systems using transformer and multi-scale esrgan models, Results Eng. 24 (2024) 103597.
[42] S. Al-Dahidi, H. Alahmer, B. Rinchi, A. Bani-Abdullah, M. Alrbai, O. Ayadi, L. Al-Ghussain, Multistep pv power forecasting using deep learning models and the reptile search algorithm, Results Eng. 27 (2025) 106265.
[43] N. Li, J. Dong, L. Liu, H. Li, J. Yan, A novel emd and causal convolutional network integrated with transformer for ultra short-term wind power forecasting, Int. J. Electr. Power Energy Syst. 154 (2023) 109470.
[44] G.G. Kim, J.H. Hyun, J.H. Choi, B.G. Bhang, H.-K. Ahn, et al., Quality analysis of photovoltaic system using descriptive statistics of power performance index, IEEE Access 11 (2023) 28427–28438.
[45] H. Liu, J. Li, Y. Wu, Y. Fu, Clustering with outlier removal, IEEE Trans. Knowl. Data Eng. 33 (2019) 2369–2379.
[46] N.B. Chikodili, M.D. Abdulmalik, O.A. Abisoye, S.A. Bashir, Outlier detection in multivariate time series data using a fusion of k-medoid, standardized Euclidean distance and z-score, in: International Conference on Information and Communication Technology and Applications, Springer, 2020, pp. 259–271.
[47] R. Zhang, G. Li, Z. Ma, A deep learning based hybrid framework for day-ahead electricity price forecasting, IEEE Access 8 (2020) 143423–143436.
[48] N.-H. Nguyen, J. Abellán-García, S. Lee, E. Garcia-Castano, T.P. Vo, Efficient estimating compressive strength of ultra-high performance concrete using xgboost model, J. Build. Eng. 52 (2022) 104302.
[49] J. Liu, L. Hou, R. Zhang, X. Sun, Q. Yu, K. Yang, X. Zhang, Explainable fault diagnosis of oil-gas treatment station based on transfer learning, Energy 262 (2023) 125258.
[50] M. Amer, U. Sajjad, K. Hamid, N. Rubab, Reliable prediction of solar photovoltaic power and module efficiency using Bayesian surrogate assisted explainable data-driven model, Results Eng. 24 (2024) 103226.
[51] R. Zhang, S. Bu, M. Zhou, G. Li, B. Zhan, Z. Zhang, Deep reinforcement learning based interpretable photovoltaic power prediction framework, Sust. Energy Technol. Assess. 67 (2024) 103830.
[52] X. Chen, H. Tang, Y. Wu, H. Shen, J. Li, et al., Adpstgcn: adaptive spatial–temporal graph convolutional network for traffic forecasting, Knowl.-Based Syst. 301 (2024) 112295.
[53] P. Nandal, N. Bohra, P. Mann, N.N. Das, Yolov11 with transformer attention for real-time monitoring of ships: a federated learning approach for maritime surveillance, Results Eng. 27 (2025) 106297.
[54] J. Su, M. Ahmed, Y. Lu, S. Pan, W. Bo, Y. Liu, Roformer: enhanced transformer with rotary position embedding, Neurocomputing 568 (2024) 127063.
[55] K. Fan, Y. Chen, C. Lai, Q. Cai, X. Wu, Energy-saving control of multi-zone purification ventilation system based on a novel multi-task learning framework, Energy (2025) 134744.
[56] M.P. Bakht, M.N.H. Mohd, B.S.K.K. Ibrahim, N. Khan, U.U. Sheikh, A.A.-H. Ab Rahman, Advanced automated machine learning framework for photovoltaic power output prediction using environmental parameters and shap interpretability, Results Eng. 25 (2025) 103838.
[57] W. Abdelfattah, M.K. Abosaoda, K. Vaghela, G. J, P.K. Sahu, K.U. Singh, R. Sivaranjani, R. Chauhan, S. Singla, S. Sherzod, Predicting biochar yield from biomass pyrolysis: a comprehensive data-driven approach using machine learning and shap analysis, Results Eng. 26 (2025) 105389.
[58] K. Suresh, K.J. Velmurugan, R. Vidhya, et al., Deep anomaly detection: a linear one-class svm approach for high-dimensional and large-scale data, Appl. Soft Comput. 167 (2024) 112369.
[59] M. Abboush, C. Knieke, A. Rausch, Intelligent back-to-back testing with denoising autoencoder-based fault detection and dbscan clustering, Results Eng. 27 (2025) 105900.
[60] F. Harrou, B. Bouyeddou, N. Zerrouki, A. Dairi, Y. Sun, Y. Zerrouki, Detecting the signs of desertification with landsat imagery: a semi-supervised anomaly detection approach, Results Eng. 22 (2024) 102037.
[61] R.A. Hamad, M. Kimura, L. Yang, W.L. Woo, B. Wei, Dilated causal convolution with multi-head self attention for sensor human activity recognition, Neural Comput. Appl. 33 (2021) 13705–13722.
[62] T. Kipf, Semi-supervised classification with graph convolutional networks, arXiv preprint, arXiv:1609.02907, 2016.
[63] P. Li, H. Yang, H. Wu, Y. Wang, H. Su, T. Zheng, F. Zhu, G. Zhang, Y. Han, Deep learning model for solar and wind energy forecasting considering northwest China as an example, Results Eng. 24 (2024) 102939.