Hierarchical Deep Reinforcement Learning Framework for Short-Term Solar Forecasting

1st Meng Liu
Department of Electrical and Electronic Engineering
The Hong Kong Polytechnic University
Hong Kong Special Administrative Region, China
meng7659.liu@connect.polyu.hk

2nd Siqi Bu*
Department of Electrical and Electronic Engineering
The Hong Kong Polytechnic University
Hong Kong Special Administrative Region, China
siqi.bu@polyu.edu.hk

***Abstract*—Accurate short-term renewable energy forecasting is crucial for grid operation under variable meteorological conditions. This paper proposes a hierarchical deep reinforcement learning (HDRL) framework that integrates local Q-learning agents and a global Proximal Policy Optimization (PPO) coordinator for adaptive photovoltaic (PV) forecasting. The internal Q-agents refine basic forecasters through error-driven interaction, while the external PPO layer dynamically optimizes four ensemble weights to enhance global coordination. The proposed framework was validated using the GEFCom2014 solar dataset, demonstrating superior accuracy and stability over conventional models. Simulation results indicate that Q-learning adjustment achieves up to 30.5% and 35.7% reductions in RMSE and MAE, respectively. Consequently, the ensemble model achieves the lowest RMSE and MAE and a higher R^2, outperforming conventional and single-layer DRL baselines. Comparative residual analyses confirm narrower error dispersion and improved robustness under non-stationary conditions. These results highlight the capability for the proposed HDRL framework in achieving accurate and reliable solar energy forecasting.**

***Index Terms*—PV forecasting, deep reinforcement learning, hierarchical model, and ensemble learning.**

I. INTRODUCTION

The growing penetration of PV generation brings both opportunities and challenges to modern power systems. While PV energy contributes to carbon neutrality and decentralized power production, its stochastic and intermittent nature causes severe fluctuations in output power. Short-term PV forecasting, typically covering a few minutes to several hours ahead, enables grid operators to anticipate rapid variations in solar output and take timely control actions. Accurate short-term PV forecasting is essential for ensuring reliable grid operation and maximizing renewable energy utilization [1].

In general, PV forecasting methods can be categorized into four main groups, including physical models [2], statistical models [3], artificial intelligence (AI) models and hybrid models. Physical and statistical approaches rely on meteorological variables and regression-based relationships but often fail to capture nonlinear temporal dependencies. This limitation becomes increasingly evident with the expansion of distributed PV systems, where solar irradiance exhibits strong non-stationarity under rapidly changing weather conditions. In this context, data-driven deep learning methods such as long short-term memory (LSTM) [4], gated recurrent unit (GRU) [5], and related bidirectional variants have demonstrated strong capability in modeling temporal dependencies for PV forecasting tasks [6,7]. For instance, a BiLSTM-based photovoltaic power forecasting model using correlation heatmap feature selection and two-stage decomposition is developed in [8]. Recently, hybrid frameworks combining multiple deep models have gained attention as ensemble forecasting can reduce individual model bias and improve overall stability [9,10]. Hui Li et al. proposed a hybrid PSO-BiLSTM forecasting framework combined with multi-stage feature extraction and symplectic geometry model decomposition to improve wind and PV power prediction accuracy, and a CNN-GRU-LSTM hybrid architecture was employed to forecast solar radiation and temperature across multiple time horizons [11,12]. Nevertheless, traditional ensemble or hybrid models rely on static fusion mechanisms and exhibit limited adaptability, which results in ineffective actions under varying meteorological conditions. Their fixed weights design leads to insensitivity to distributional shifts and reduced reliability.

To overcome this limitation, deep reinforcement learning (DRL) has been introduced as a promising approach for dynamic model combination. RL enables an agent to learn adaptive strategies by interacting with the environment and receiving feedback in the form of rewards [13]. Prior studies have applied Q-learning or policy-gradient methods for wind and solar forecasting [14–16]. However, most of these RL-based forecasting frameworks adopt a single-layer design, where the agent merely learns to optimize high-level decision policies without directly regulating the dynamics of predictors. Such a design constrains the adaptability because it fails to establish a hierarchical coordination between local model optimization and global ensemble control.

To address these challenges, this study proposes an HDRL framework for short-term PV forecasting. The primary contributions of this paper can be summarized as follows:

- A Q-learning–based adaptive regulation mechanism is developed to dynamically adjust the forecasting behavior of individual models according to real-time performance feedback for enhancing local prediction accuracy.
- A PPO-driven hierarchical ensemble framework is proposed to integrate the adjusted models through global weights optimization, which improves forecasting stability and generalization.

979-8-3315-8138-1

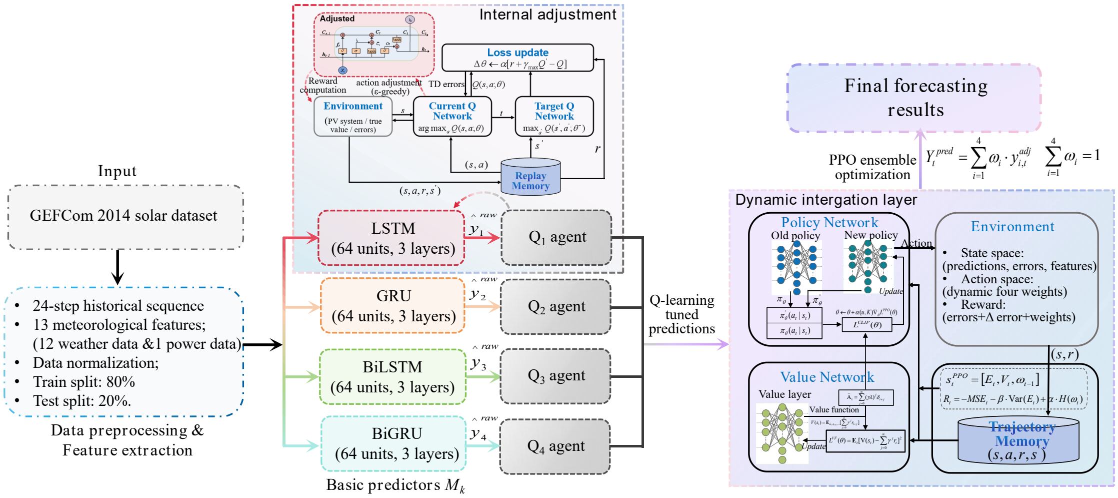

Fig. 1. Overall architecture of the proposed HDRL forecasting framework.

The remainder of the paper is organized as follows: The proposed methodologies are elaborated in Section II. Section III details the simulation setup and results analysis, and Section IV draws the conclusion.

II. Methodology

This section presents the proposed HDRL framework for short-term PV forecasting. The architecture comprises four recurrent neural forecasters refined by Q-learning agents at the internal layer, and utilizing a PPO policy to dynamically optimize ensemble weights at the external layer. The complete forecasting workflow is illustrated in Fig. 1.

A. Four Basic Forecasting Models

1) LSTM: The LSTM network maintains an updated cell state c_t and hidden state h_t at each time step to effectively capture long-term temporal dependencies through gated memory mechanisms, such as illustrated in (1).

$$\begin{cases} f_t = \sigma(W_f[h_{t-1}, x_t] + b_f), \\ i_t = \sigma(W_i[h_{t-1}, x_t] + b_i), \\ c_t = f_t \odot c_{t-1} + i_t \odot \tilde{c}_t, \\ o_t = \sigma(W_o[h_{t-1}, x_t] + b_o), \\ h_t = o_t \odot \tanh(c_t) \end{cases} \tag{1}$$

where x_t and h_{t-1} denote the current input and the previous hidden state, respectively; f_t, i_t, and o_t represent the forget, input, and output gates activated by the sigmoid function σ.

2) GRU: The GRU combines the forget and input gates into a single update gate and introduces a reset gate to control the past information, which provides efficient temporal modeling with fewer parameters than the LSTM, as given in (2)

$$\begin{cases} z_t = \sigma(W_z[h_{t-1}, x_t]), \\ r_t = \sigma(W_r[h_{t-1}, x_t]), \\ \tilde{h}_t = \tanh(W_h[x_t, r_t \odot h_{t-1}]), \\ h_t = (1 - z_t) \odot h_{t-1} + z_t \odot \tilde{h}_t) \end{cases} \tag{2}$$

where z_t and r_t denote the update and reset gate, respectively; h_t is the updated hidden state, $\tilde{h}_t$ is the candidate hidden state.

3) BiLSTM: The BiLSTM extends the standard LSTM by processing the input sequence in both forward and backward directions to capture both past and future contextual information. At each time step t, the forward and backward hidden states are concatenated as

$$h_t = [\overrightarrow{h_t}; \overleftarrow{h_t}] \tag{3}$$

where $\overrightarrow{h_t}$ and $\overleftarrow{h_t}$ denote the hidden states from the forward and backward passes, respectively.

4) BiGRU: Similarly, BiGRU adopts a bidirectional GRU structure to capture information from both temporal directions, enhancing temporal feature representation while retaining the parameter efficiency of GRU.

B. Internal Layer: Deep Q-Learning Based Model Tuning

Model-free Q-learning enables agents to learn adjustment policies through trial-and-error interaction with the environment without requiring explicit knowledge of the underlying dynamics. In this study, a lightweight Q-agent is employed to refine the raw output of each basic forecaster M_k ($k \in LSTM, GRU, BiLSTM, BiGRU$) based on the current prediction error and meteorological context.

State. At time step t, the agent observes

$$s_t = [\epsilon_t, x_{t,1}, \ldots, x_{t,j}] \in \mathbb{R}^{j+1}, \quad \epsilon_t = y_t - \hat{y}_t^{\text{raw}} \tag{4}$$

where ϵ_t is the instantaneous prediction error, and $x_{t,j}$ are meteorological features derived from the dataset.

Action. A discrete adjustment is applied to the raw prediction with a small step size δ.

$$a_t \in \{-\delta, 0, +\delta\}, \quad \hat{y}_t^{\text{adj}} = \hat{y}_t^{\text{raw}} + a_t \tag{5}$$

The adjusted value is projected to the physical range as

$$\hat{y}_t \leftarrow \max(0, \hat{y}_t^{\text{adj}}) \tag{6}$$

Reward. The reward function is designed to jointly balance accuracy improvement, action smoothness, and physical feasibility during model adjustment.

$$r_t = -\epsilon_{t+1}^2 + \lambda(\epsilon_t^2 - \epsilon_{t-1}^2) - \eta|a_t - a_{t-1}| - \zeta\mathbb{I}\{\hat{y}_t^{\text{adj}} < 0\} \tag{7}$$

979-8-3315-8138-1

where $-\epsilon_{t+1}^2$ penalizes large errors, $\lambda(\epsilon_t^2 - \epsilon_{t-1}^2)$ rewards error reduction, $-\eta|a_t - a_{t-1}|$ discourages abrupt actions, and $-\zeta \mathbb{I}\{\hat{y}_t^{\text{adj}} < 0\}$ maintains physical feasibility. λ, η, and ζ are balancing coefficients.

Policy and learning. With ϵ-greedy exploration, the action is chosen as (8), and the Q-network is updated through the temporal-difference rule in (9).

$$a_t = \begin{cases} \text{rand}(A), & \text{with prob. } \epsilon \\ \arg\max_a Q(s_t, a), & \text{with prob. } 1-\epsilon \end{cases} \tag{8}$$

$$Q_{t+1}(s_t, a_t) \leftarrow Q_t(s_t, a_t) + \alpha\big[r_t + \gamma \max_{a'} Q_t(s_{t+1}, a') - Q_t(s_t, a_t)\big] \tag{9}$$

Each basic model M_k is trained jointly with a dedicated Q-agent, yielding refined predictions $\hat{y}_k^{\text{adj}}$ that achieve improved accuracy and adaptability under non-stationary conditions. All models in this layer are trained using the same dataset.

C. External Layer: PPO-Based Ensemble Optimization

While the internal Q-learning agents locally refine individual forecasters, the external layer employs a PPO agent to achieve global coordination by dynamically adjusting ensemble weights $\boldsymbol{\omega} = [\omega_1, \omega_2, \omega_3, \omega_4]$ across the four DRL-tuned predictors. The ensemble prediction is expressed as

$$\hat{Y}_t^{pred} = \sum_{i=1}^{4} \omega_i \hat{y}_{i,t}^{adj}, \quad \sum_{i=1}^{4} \omega_i = 1 \tag{10}$$

where the weights w_i are constrained within $[w_{\min}, w_{\max}]$ and normalized through a softmax function to ensure feasibility.

State and Action. At each time step t, the PPO agent observes a global state $s_t^{\text{PPO}} = [E_t, V_t, \boldsymbol{\omega}_{t-1}]$, where E_t denotes recent forecasting errors and V_t represents weather variability indicators. The action a_t^{PPO} adjusts ensemble weights as $\boldsymbol{\omega}_t = \text{Softmax}(\boldsymbol{\omega}_{t-1} + a_t^{\text{PPO}})$.

Reward and Policy Optimization. The reward balances three objectives: forecasting accuracy, stability, and diversity of weight allocation, defined as

$$R_t = -\text{MSE}_t - \beta\, \text{Var}(E_t) + \alpha\, H(\boldsymbol{\omega}_t) \tag{11}$$

where $H(\cdot)$ is the entropy term encouraging weight diversity, and β and α are balancing coefficients. The PPO objective follows the clipped surrogate loss:

$$L^{\text{CLIP}}(\theta) = \mathbb{E}_t\left[\min\left(r_t(\theta)\hat{A}_t,\ \text{clip}(r_t(\theta), 1-\epsilon, 1+\epsilon)\hat{A}_t\right)\right] \tag{12}$$

where $r_t(\theta) = \pi_\theta(a_t|s_t)/\pi_{\theta_{\text{old}}}(a_t|s_t)$ and $\hat{A}_t$ is the estimated advantage. The clipping term limits policy deviation for stable learning in the bounded simplex of ensemble weights.

Through iterative policy interaction, the PPO agent adaptively allocates ensemble weights based on temporal dynamics, meteorological variability and forecasting performance to achieve global optimization of accuracy and robustness while maintaining smooth temporal transitions.

D. Overall Framework and Algorithm

The dual-layer framework integrates internal Q-learning agents and an external PPO coordinator to achieve adaptive and stable PV forecasting. The internal layer tunes individual forecasters based on local error feedback, while the external layer ensures global coordination through dynamic ensemble weights. The overall procedure is summarized in **Algorithm 1** and the schematic structure is illustrated in Fig. 1.

Algorithm 1 The Proposed HDRL for PV Forecasting

```
Initialize: Q-agent for basic model M_i and PPO-agent for ensemble coordination.
for each training episode do
    for each time step t do
        for each forecasting model i = 1, ..., 4 do
            Observe local state s_t^i and select action a_t^i via ε-greedy policy.
            Update Q_i(s_t^i, a_t^i) using reward r_t^i from Eq. (7).
            Obtain tuned output ŷ_{t,i}^(adj).
        end for
        Collect refined predictions {ŷ_{t,i}^(adj)}_{i=1}^4.
        PPO agent computes ensemble weights ω_t = Softmax(ω_{t-1} + a_t^PPO).
        Final ensemble output: Ŷ_t^pred = Σ_{i=1}^4 ω_{t,i} ŷ_{t,i}^(adj).
        Store (s_t^PPO, ω_t, R_t) for PPO update.
    end for
    Update PPO parameters using clipped objective L^CLIP from Eq. (12).
end for
```

III. Simulations and Results

A. Dataset and Simulation Setup

The proposed model was evaluated using the GEFCom2014 solar power dataset [17]. This dataset contains hourly PV generation records and 12 corresponding weather variables collected from April 2012 to April 2013. In this study, 5000 hourly samples were selected for model training. The data were preprocessed through Z-score normalization and a 24-hour sliding window approach to predict the next-hour output, and the dataset was chronologically split into 80% for training and 20% for testing. All models were trained and evaluated under multiple different random seeds.

The simulations were conducted on a computing platform configured with Python 3.8 and PyTorch 1.12. The detailed model architecture and hyperparameter configurations are provided in the tables referenced in Appendix: Parameter settings.

B. Evaluation Metrics

To quantitatively evaluate the forecasting performance, three commonly utilized statistical metrics were adopted, including the Root Mean Square Error (RMSE), Mean Absolute Error (MAE), and the Coefficient of Determination (R^2). The mathematical formulations are expressed as follows:

$$\text{RMSE} = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2} \tag{13}$$

$$\text{MAE} = \frac{1}{n}\sum_{i=1}^{n}|y_i - \hat{y}_i| \tag{14}$$

$$R^2 = 1 - \frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2} \tag{15}$$

where y_i and $\hat{y}_i$ denote the actual and predicted value, respectively. $\bar{y}$ is the mean of the actual values, and n is the number

979-8-3315-8138-1

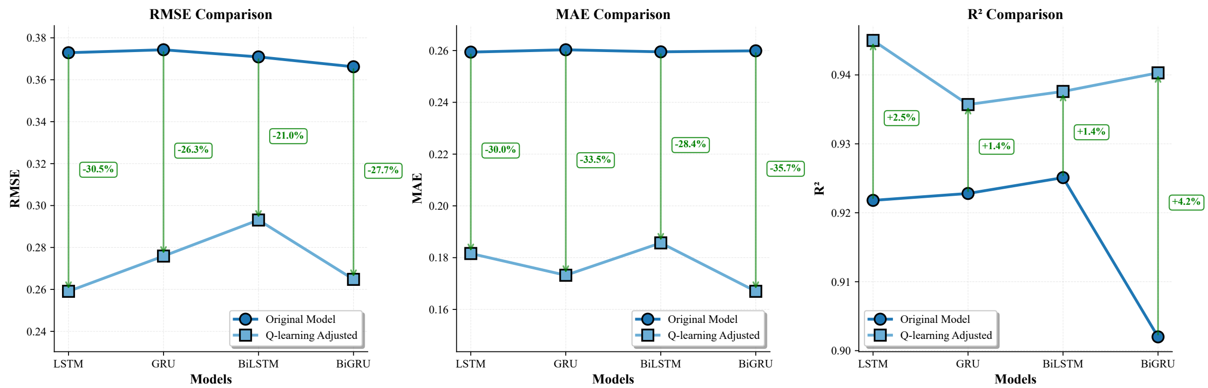

Fig. 2. Comparison of baseline and Q-learning adjusted models.

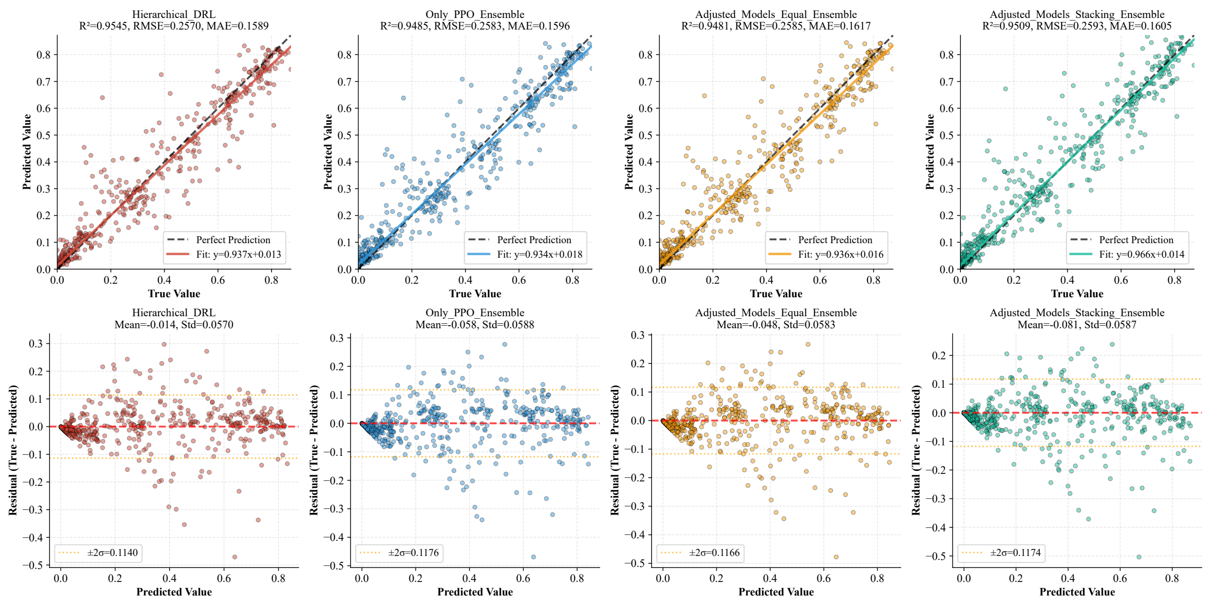

Fig. 3. Forecasting performance comparison of the proposed method and other ensemble models.

of samples. RMSE and MAE measure the average deviation between predictions and observations, while R^2 indicates the degree of correlation between the predicted and actual data.

C. Comparison of Forecasting Results

1) Performance Comparison between Q-learning Adjusted and Baseline Models: To validate the effectiveness of the proposed Q-learning adjustment strategy, four baseline forecasting models were trained and compared with Q-learning adjusted counterparts. The forecasting errors and accuracy metrics are exhibited in Fig. 2, where the Q-learning adjustment consistently enhances prediction accuracy across all models.

Specifically, the adjusted models exhibit an overall reduction in RMSE and MAE, accompanied by an increase in R^2 values. Among the four basic models, LSTM achieves the largest RMSE reduction (30.5%), while BiGRU achieves the largest MAE improvements (35.7%). This enhancementment can be attributed to the adaptive adjustment mechanism introduced by Q-learning. Through a reward driven optimization process, Q-learning dynamically updates the action policy of the model based on historical forecasting errors, enabling adaptive correction under varying meteorological conditions. The ε-greedy strategy enables more effective parameter adaptation that allows Q-learning to enhance model adaptability under dynamic weather conditions and achieve more stable forecasting. The improvement observed in R^2 further confirms that the adjusted models capture the temporal dependency of PV generation more accurately than the original versions.

2) Comprehensive Analysis of the proposed HDRL Models: Based on the results presented in Section III-C1, the Q-learning adjustment effectively improved the forecasting performance of individual basic models. To further enhance global coordination, the adjusted models were dynamically integrated using the PPO-based ensemble strategy. Fig. 3 illustrates the comprehensive comparison among the HDRL framework, the single PPO ensemble method, and two conventional ensemble baselines (equal-weight and stacking).

The HDRL achieves the best overall performance, with an RMSE of 0.2570, MAE of 0.1589, and an R^2 value of 0.9545. Compared with the single PPO ensemble and traditional static

979-8-3315-8138-1

integration, it shows higher prediction accuracy and lower residual variance. In the residual plots, the HDRL exhibits the narrowest error band and the most concentrated distribution around zero, which suggests effective bias correction and noise suppression. This improvement stems from the two-level mechanism. The internal layer adaptively adjusts model parameters based on recent performance to enhance precision, while the PPO layer learns four dynamic ensemble coefficients. This hierarchical coordination balances local accuracy and global consistency, resulting in more stable and accurate predictions than single-level DRL and traditional ensembles.

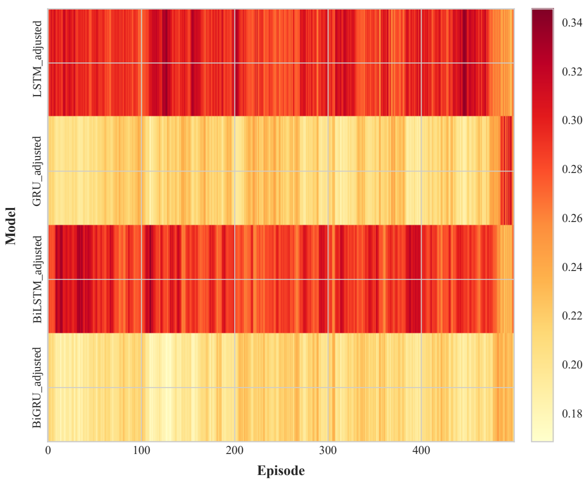

Fig. 4. Dynamic weights evolution under PPO coordination.

To evaluate the real-time capability of the proposed framework, Table I summarizes the offline training cost and the online inference latency. After offline training, all baseline forecasters can be deployed with millisecond-level inference latency. With the Q-learning adjustment layer, the inference latency is reduced by up to 21.6% across the baseline models. Moreover, the integrated PPO-based predictor and the proposed HDRL scheme exhibit inference latencies of 8.33 ms and 7.50 ms, respectively, which are negligible compared with the hourly forecasting horizon and support real-time PV forecasting deployment.

TABLE I
TIME COST OF BASELINE AND THE PROPOSED MODELS.

Model	Offline training time (min)			Online inference time (ms)		
	Model	Q-agent	Total	Model	Model+Q	Δ(%)
LSTM	5.00	4.10	9.10	1.29	1.22	-5.2
GRU	4.67	3.94	8.61	3.52	2.76	-21.6
BiLSTM	5.83	3.92	9.75	2.09	2.13	+2.0
BiGRU	5.50	3.91	9.41	5.82	5.11	-12.2
PPO Ens.	6.19	–	6.19	8.33	–	$>$+43.1
Proposed HDRL	6.43	4.31	10.74	7.50	–	$>$+28.9

3) Adaptive Weights Coordination under PPO Policy: Fig. 4 illustrates the dynamic weight evolution of four adjusted forecasters under PPO coordination. As training converges, the PPO agent adaptively assigns higher weights to the LSTM and BiLSTM models. The final normalized weight combination converges to [0.34, 0.29, 0.26, 0.21].

IV. CONCLUSION

This study proposed an HDRL framework for short-term PV power forecasting, which integrates Q-learning–based model adjustment and PPO-based global ensemble optimization. Through a two-stage training process, Q-learning effectively refined the forecasting behavior of individual basic models, while the PPO layer dynamically coordinated their interactions through adaptive weights. Simulation results demonstrated that the HDRL framework consistently outperformed both conventional ensemble methods and single-level DRL approaches. The internal layer regulation reduced local prediction errors and improved adaptability under non-stationary weather conditions, and the external ensemble layer further enhanced global stability and generalization. Overall, the proposed method achieved more accurate and robust forecasting performance, providing a promising direction for intelligent solar prediction.

REFERENCES

[1] Y. Ma, Y. Huang, and Y. Yuan, "The short-term forecasting of distributed photovoltaic power considering the sensitivity of meteorological data," *Journal of Cleaner Production*, vol. 486, p. 144599, 2025.

[2] M. J. Mayer and D. Yang, "Pairing ensemble numerical weather prediction with ensemble physical model chain for probabilistic photovoltaic power forecasting," *Renewable and Sustainable Energy Reviews*, vol. 175, p. 113171, 2023.

[3] M. AlKandari and I. Ahmad, "Solar power generation forecasting using ensemble approach based on deep learning and statistical methods," *Applied Computing and Informatics*, vol. 20, no. 3/4, pp. 231–250, 2024.

[4] S. Han, Y.-h. Qiao, J. Yan, Y.-q. Liu, L. Li, and Z. Wang, "Mid-to-long term wind and photovoltaic power generation prediction based on copula function and long short term memory network," *Applied energy*, vol. 239, pp. 181–191, 2019.

[5] Y. Dai, Y. Wang, M. Leng, X. Yang, and Q. Zhou, "Lowess smoothing and random forest based gru model: A short-term photovoltaic power generation forecasting method," *Energy*, vol. 256, p. 124661, 2022.

[6] J. Gong, Z. Qu, Z. Zhu, and H. Xu, "Parallel timesnet-bilstm model for ultra-short-term photovoltaic power forecasting using stl decomposition and auto-tuning," *Energy*, vol. 320, p. 135286, 2025.

[7] A. Mellit, A. M. Pavan, and V. Lughi, "Deep learning neural networks for short-term photovoltaic power forecasting," *Renewable Energy*, vol. 172, pp. 276–288, 2021.

[8] W. Lin, B. Zhang, H. Li, and R. Lu, "Multi-step prediction of photovoltaic power based on two-stage decomposition and bilstm," *Neurocomputing*, vol. 504, pp. 56–67, 2022.

[9] Y. Dai, W. Yu, and M. Leng, "A hybrid ensemble optimized bigru method for short-term photovoltaic generation forecasting," *Energy*, vol. 299, p. 131458, 2024.

[10] H. Li, Z. Ren, Y. Xu, W. Li, and B. Hu, "A multi-data driven hybrid learning method for weekly photovoltaic power scenario forecast," *IEEE Transactions on Sustainable Energy*, vol. 13, no. 1, pp. 91–100, 2021.

[11] S. Peng, J. Zhu, T. Wu, C. Yuan, J. Cang, K. Zhang, and M. Pecht, "Prediction of wind and pv power by fusing the multi-stage feature extraction and a pso-bilstm model," *Energy*, vol. 298, p. 131345, 2024.

[12] M. Fathi, S. Allahyari, A. Ahmadi, A. Zahedi, and S. Asiaei, "Enhanced global solar irradiance and temperature forecasting for rooftop pv power generation prediction using multi-step lstm and gru based hybrid models," *Energy Conversion and Management: X*, p. 101351, 2025.

[13] B. Singh, R. Kumar, and V. P. Singh, "Reinforcement learning in robotic applications: a comprehensive survey," *Artificial Intelligence Review*, vol. 55, no. 2, pp. 945–990, 2022.

[14] R. Zhang, S. Bu, M. Zhou, G. Li, B. Zhan, and Z. Zhang, "Deep reinforcement learning based interpretable photovoltaic power prediction framework," *Sustainable Energy Technologies and Assessments*, vol. 67, p. 103830, 2024.

[15] J. Liu, H. Zang, L. Cheng, T. Ding, and Z. Wei, "Robust photovoltaic power forecasting against multi-modal adversarial attack via deep reinforcement learning," *IEEE Transactions on Sustainable Energy*, 2025.

[16] J. Yan, C. Feng, M. Sun, and J. Zhang, "Short-term probabilistic solar forecasting via reinforcement learning over ecmwf," in *2024 IEEE Power & Energy Society General Meeting (PESGM)*. IEEE, 2024, pp. 1–5.

[17] T. Hong, P. Pinson, S. Fan, H. Zareipour, A. Troccoli, and R. J. Hyndman, "Probabilistic energy forecasting: Global energy forecasting competition 2014 and beyond," pp. 896–913, 2016.

979-8-3315-8138-1