IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Received 7 February 2024, accepted 14 March 2024, date of publication 18 March 2024, date of current version 26 March 2024.

*Digital Object Identifier 10.1109/ACCESS.2024.3378749*

RESEARCH ARTICLE

# Dandelion Optimizer-Based Reinforcement Learning Techniques for MPPT of Grid-Connected Photovoltaic Systems

**GHAZI A. GHAZI[1,2], ESSAM A. AL-AMMAR[1], (Senior Member, IEEE), HANY M. HASANIEN[3,4], (Senior Member, IEEE), WONSUK KO[1], JAESUNG PARK[5], DONGSU KIM[6], AND ZIA ULLAH[7], (Member, IEEE)**

[1]Department of Electrical Engineering, College of Engineering, King Saud University, Riyadh 11421, Saudi Arabia
[2]K. A. CARE Energy Research and Innovation Center, King Saud University, Riyadh 11421, Saudi Arabia
[3]Electrical Power and Machines Department, Faculty of Engineering, Ain Shams University, Cairo 11517, Egypt
[4]Faculty of Engineering and Technology, Future University in Egypt, Cairo 11835, Egypt
[5]Energy Efficiency Center, Korea Conformity Laboratories, Cheongwon-gu, Cheongju-si, Chungbuk 28115, South Korea
[6]Department of Architecture Engineering, Hanbat National University, Daejeon 34158, South Korea
[7]School of Electrical and Electronic Engineering, Huazhong University of Science and Technology, Wuhan 430074, China

Corresponding author: Ghazi A. Ghazi (439106681@student.ksu.edu.sa)

This work was supported in part by the Korea Institute of Energy Technology Evaluation and Planning (KETEP); and in part by the Ministry of Trade, Industry and Energy (MOTIE), Republic of Korea, under Grant 20228500000020.

**ABSTRACT** The integration of photovoltaic (PV) into electric power systems has been widely explored and adopted to address the problems associated with the depletion of fossil fuels and the release of greenhouse gases. PV panels convert sunlight into electricity, minimizing the reliance on fossil fuels and mitigating environmental pollution. It is crucial to optimally utilize the PV power in the system; hence maximum power point tracking (MPPT) algorithms have been developed to ensure optimal performance of grid-connected PV systems at the maximum power point (MPP) despite changes in weather conditions. Moreover, deep reinforcement learning (DRL) developments provide a promising approach for optimizing grid-connected PV systems, replacing the conventional proportional-integral-derivative (PID) controllers. However, there is limited research evaluating the efficiency of these systems using DRL techniques. This paper proposes a new dandelion optimizer (DO)-based DRL for MPPT of grid-connected photovoltaic systems and evaluates the proposed method for a 100-MW PV plant connected to a 33-kV distribution system. The proposed DRL technique uses proximal policy optimization (PPO) and deep deterministic policy gradient (DDPG) algorithms for continuous states and discrete or continuous action spaces to adjust the PV-measured voltage based on a reference one produced via DO-PPO and DO-DDPG methods. To test the effectiveness and practicality of the introduced methods, simulations were conducted using actual input data of a 100 MW PV plant connected to a 33-kV distribution system for typical days in summer and winter seasons using MATLAB/Simulink software. The proposed implemented methods were evaluated by comparing their simulation results with other techniques: DO-PID, particle swarm optimization (PSO), and incremental conductance (InC-PI). The findings revealed that the efficiencies of the DC-DC boost and the voltage source converters using the introduced methods were 84.25%- 85.90%, and 78.33%- 81.10% on a summer day while they were 92.77%- 95% and 86.70%- 89.50% on a winter day, respectively, which proves that these methods were efficient and effective, indicating their promising potential for future applications.

**INDEX TERMS** Dandelion optimizer, deep deterministic policy gradient, deep reinforcement learning, maximum power point tracking, PV systems, proximal policy optimization.

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

## I. INTRODUCTION

Renewable energy resources are becoming more important in electricity generation due to concerns about the unsustainable

nature of fossil fuels and greenhouse gas emissions [1]. PV energy resource, in particular, has seen noteworthy growth in clean energy generation in recent years; however, PV efficiency is still relatively low [2]. Therefore, to improve its efficiency, it is vital to continuously accurately pursue the PV resource's maximum power point (MPP). Hence, tracking the MPP for maximizing the energy capture from such resources under varying solar irradiance conditions is becoming essential.

Various approaches have been developed in the literature to enhance the efficiency of PV systems by monitoring the maximum power point. These approaches can be grouped according to multiple criteria, such as the type of tracking techniques used, the implementation of sensors, and the level of modernity [3]. The conventional MPPT techniques that are frequently utilized include perturb and observe (P&O) [4], hill-climbing [5], fractional short-circuit current and open-circuit voltage techniques in [6] and [7], respectively, and incremental conductance (InC) [8]. Nevertheless, when the PV modules are subjected to partial shading conditions (PSCs) in which the solar irradiance curve, as an input to these techniques, exhibits multiple peaks (global and local), these techniques may struggle to track or reach the MPP.

Furthermore, various novel MPPT methods have been suggested, including fuzzy and neuro-fuzzy inference systems in [8] and [9], respectively, artificial neural network [10], artificial bee colony [11], genetic algorithm [12], particle swarm optimization [13], grey wolf technique [14], salp-swarm optimization [15], cuckoo search algorithm [16], ant colony optimization [17], and firefly algorithm [18], These techniques can quickly achieve the MPP with less oscillation around it. However, their implementation cost is high compared to conventional techniques due to their complex algorithms. Hybrid methods have been introduced to enhance the efficiency and overcome limitations of the novel techniques. These techniques integrate multiple searching mechanisms to create a combined approach that improves performance and eliminates drawbacks. By using the hybrid methods, the limitations of one method can be enhanced through the implementation of another [19], [20], [21], [22]. Ultimately, all MPPT techniques aim to guarantee that PV modules operate at their MPP under any weather condition.

The DRL algorithm is a machine learning method where the agent enhances its effectiveness via a trial-and-error policy to optimize the overall cumulative reward [23]. The DRL's agent engages with the corresponding environment and acquires knowledge of the dynamics by experimenting with numerous actions and perceiving the outcomes via rewards. Thus, DRL focuses on decision-making to optimize the overall discounted reward while interacting with the environment. By utilizing powerful function approximators like neural networks, DRL techniques have been effectively employed in complex domains such as video games [24], robotic control [25] and electric vehicles [26]. Consequently, due to the achievements of DRL in various domains, numerous researchers have suggested utilizing DRL-based approaches to tackle the MPP issue and overcome the shortcomings of existing MPPT techniques. These limitations include disturbances caused by MPP's oscillations in the P&O method, the complexity involved in the controlling procedure of the InC method, extended training time for neural network methods, and the influence of designer preferences in fuzzy logic methods. In line with this, Kofinas et al. [27] proposed Q-learning, and SARSA approaches to pursue the MPP under diverse environmental circumstances and exhibited no oscillations around it. This study utilized a Q table of 4000 states and five possible actions to enhance computational efficiency.

Hsu et al. [28] developed an RL-based MPPT regulator using four states and seven actions to represent the operating point's movements of the MPP. Similarly, Youssef et al. [29] employed four actions and the same states to trail the MPP. Nonetheless, some oscillations were found around the MPP. And, Kofinas et al. [27] introduced a universal RL-MPPT approach to track the MPP of a PV source without requiring prior information. The effectiveness of this method was assessed across various environmental and operational scenarios, demonstrating its superior performance and faster response compared to the P&O method. A hybrid approach of a Q-learning algorithm and the P&O method was applied. The Q-learning algorithm learned the optimum duty cycles considering temperature and solar radiation levels. This information was then transferred into the P&O method, reducing its step size [30]. Chou et al. [31] developed two RL-based MPPT algorithms using a Q table and a deep Q network (DQN) but did not address the issue of PSCs. On the other hand, approaches [32], [33] dealt with MPPT control under PSCs by using multiple agents. Furthermore, novel DRL and transfer RL-based MPPT methods for tackling the MPP issue of the PV systems under PSCs were introduced in [32] and [33], respectively. Similarly, Pan et al. [34] presented an RL-based MPPT algorithm using a DQN with continuous state and discrete action spaces. The proposed method showed a significant improvement in tracking accuracy when compared with the P&O and InC methods and it converges to the MPP faster and has better steady-state performance under PSCs when compared to PSO and GWO methods. Moreover, besides the MPPT, the DQN has proven to be quite effective in other applications such as the speed control of a DC motor [35].

To enhance a 100 MW PV plant's performance connected to a 33-kV distribution system, an MPPT method is required to regulate the PV operating voltage to the MPP under different atmospheric conditions. The DO algorithm, which is a metaheuristic approach known for its simplicity and minimal design requirements, is proposed in this study. It has been effectively utilized for solving numerous engineering problems, including parameter estimation of fuel cell models [36], reactive power dispatch optimization [37], and steel frame design [38]. In this study, the DO algorithm is responsible for

tracking the MPP and generating a reference voltage signal. At the same time, the DRL PPO or DDPG controller regulates the PV-measured voltage signal to match the DO reference voltage signal. Furthermore, the DRL controller's output is then utilized in pulse-width modulation (PWM) to produce the corresponding duty cycles for the DC boost converter. The major contributions of this work indulging the proposal of a robust and novel MPPT control based on DRL techniques, named DO-PPO and DO-DDPG, where the proposed novel techniques were successfully implemented and validated in the MATLAB/Simulink environment. A detailed comparison between the outcomes of the proposed approaches and those of DO-PID, PSO-PPO, PSO-DDPG, and InC-PI methods, the findings illustration is presented to show the relevance and effectiveness of the proposed DRL technique.

The rest of the paper is arranged in the following manner: Section II briefly describes the introduced grid-connected PV system model. Section III describes the methodology that was employed, including the use of DRL and DO methods. The simulation results from the study and subsequent discussions are offered in Section IV. The paper's conclusion is provided in Section V.

## II. SYSTEM MODELING

The proposed grid-connected PV system's configuration is depicted in Fig. 1. In this configuration, a capacity of 100-MW PV plant is connected to a DC bus bar via a DC-DC boost converter whose duty cycles are produced using the proposed DO based DRL method. The DO approach works as MPPT generating a reference PV voltage while DRL method works as controller that regulates the measured PV voltage with the reference one which allowing more efficient power transfer between the PV plant and the DC bus bar. Then, the DC bus is connected to the AC system via a voltage source inverter (VSI) which converts the DC power generated by the PV plant into AC power that can be fed into the grid. An LCL filter ensures that the power electronics within the system do not introduce harmonics into the grid. This filter helps reduce the harmonics injected by the power electronics, ensuring that the PV system's power output is clean and the harmonics are kept within allowable limits. Furthermore, the LCL filter is connected to a utility grid via a 33-kV step-up distribution transformer.

The PV array is formed by connecting the PV panels in both series ($N_s$) and parallel ($N_p$). This configuration affects the PV array's overall current and voltage output, as shown in Fig. 2. Therefore, the PV array's output current ($I_{pv}$) can be determined as follows [39]:

$$I_{pv} = N_s I_s - N_p I_o \left( \exp\left( \frac{q\left(V_{pv} + R_s I_{pv}\right)}{AkTN_s} \right) - 1 \right) - N_p \frac{V_{pv} + R_s I_{pv}}{N_s R_{sh}} \quad (1)$$

Here, $I_s$ and $I_o$ are the photocurrent and saturation currents, $R_s$ and $R_{sh}$ are the series and shunt resistors, respectively, $k$ is the Boltzmann constant, $q$ is the electronic charge, and $A$ is the diode ideality factor.

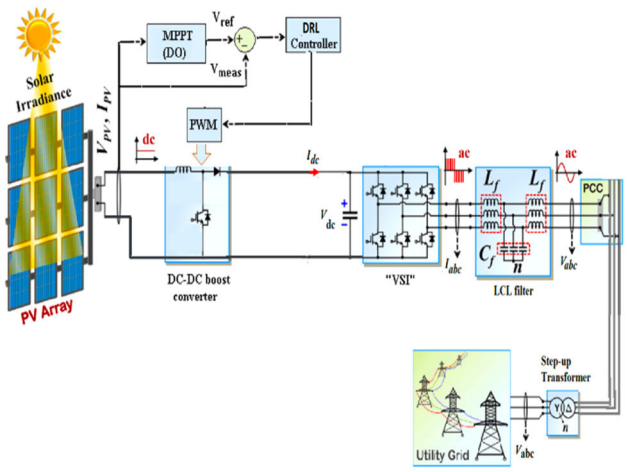



**FIGURE 1.** The grid-connected PV system's configuration.

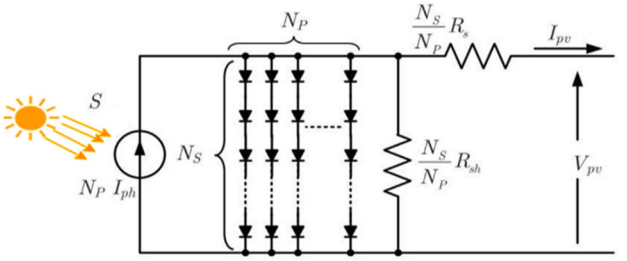



**FIGURE 2.** Circuit schematic of a PV array.

The photocurrent current $I_s$ can be specified as follows:

$$I_s = (I_{sc} - k_i(T - T_{ref}))\frac{G}{G_{ref}} \quad (2)$$

The equation involves several variables, including, $I_{sc}$, and $k_i$ which are the short-circuit current and its coefficient of the PV cell, $G_{ref}$ and G, which are the reference and operating solar radiations, $T_{ref}$ and $T$, which are the reference and ambient temperatures, respectively.

The saturation current $I_o$ can be computed as follows:

$$I_o = I_{Rs}\left(\frac{T}{T_{ref}}\right)^3 exp\left(\frac{qE_g}{Ak}\left(\frac{1}{T_{ref}} - \frac{1}{T}\right)\right) \quad (3)$$

To clarify, $I_{Rs}$ refers to the reverse-saturation current while $E_g$ represents the energy gap of the semiconductor.

Moreover, the proposed grid-connected PV system utilizes solar irradiance and ambient temperature data obtained from a weather-monitoring station at King Saud University in Riyadh, Saudi Arabia for a typical day in each of summer and winter seasons in 2021, as shown in Figs. 2 and 3.

## III. METHODOLOGY

The DO technique is employed to optimize the grid-connected PV plant's power output. Additionally, the DRL controller

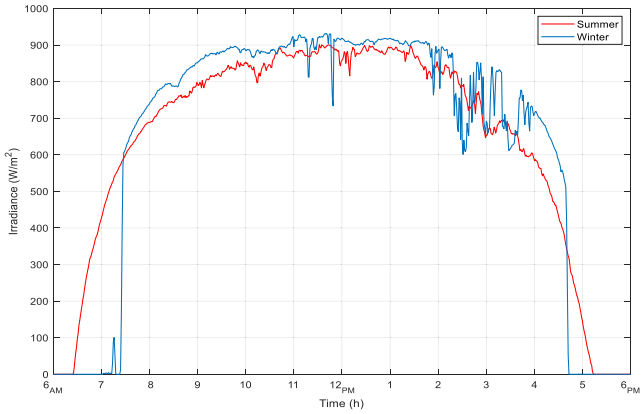



**FIGURE 3.** Solar irradiance (W/m$^2$).

matches the PV-measured voltage with the reference voltage generated by the DO technique. As a result, the objective function for MPPT can be represented using Eq.(4) as follows:

$$MaxP_{pv} = V_{pv} \times I_{pv} \tag{4}$$

$I_{pv}$ is the PV cell's output current, and $V_{pv}$ is the corresponding output voltage.

### A. DEEP REINFORCEMENT LEARNING

#### 1) BASIC CONCEPT OF DRL

Reinforcement learning (RL) is a type of machine learning where an agent learns a behavior policy through interacting with an environment and receiving cumulative rewards for its actions [40]. RL techniques have gained popularity due to recent advancements in computer science and their ability to resolve decision-making problems [41]. The components of the RL model are an agent, states, environment, actions, and rewards. The entity being acted upon is the environment, and the agent is the RL algorithm. The agent works in response to the state it receives from the environment, and as a result, it gets the subsequent form and reward. As a result, the agent refreshes its knowledge and assesses its prior activity, which continues until a set of conditions are satisfied [42]. Deep reinforcement learning (DRL) is a powerful technique combining RL principles and deep learning techniques. It has been widely applied in diverse domains such as finance management, language processing, games, and robotics [43].

Also, DRL algorithms can be classified into subcategories, including on-policy and off-policy, online and offline learning, and model-based and model-free algorithms [44]. Off-policy algorithms use a greedy learning approach where the agent selects actions that have performed well in the past from memory. On the other hand, on-policy algorithms involve the agent tracking the rewards according to the current action it takes. Online learning algorithms involve training the agent using real-world environments, unlike the offline ones that train the RL agent using virtual training environments or historical data. Model-based algorithms, unlike model-free, learn the system dynamics before planning, which are usually more expensive as they require learning a precise environment model and then finding an optimal policy. However, model-free algorithms like PPO and DDPG are more popular due to their lower cost. This work uses Actor-Critic DRL methods, such as PPO and DDPG, both online, model-free algorithms, while PPO on-policy is the DDPG off-policy algorithm.

Furthermore, the PPO and DDPG algorithms utilize different functions to train their agents. PPO uses the value function $V^{\pi}(s)$ to determine the optimal policy measures the agent's achievement for reaching a given state and yields the anticipated cumulative reward when succeeding a prearranged policy $\pi$ of current state $s$. On the other hand, DDPG uses the action-Q function $Q^{\pi}(s, a)$ which estimates the anticipated cumulative reward of a specific action in the current state $s$ of policy $\pi$. These functions are calculated as follows [42], [45]:

$$V^{\pi}(s_t) = E\left\{\sum_{k=0}^{\infty} r_{t+k+1} \cdot \gamma^t | s_t = s\right\}; \tag{5}$$

$$Q^{\pi}(s_t, a_t) = E\left\{\sum_{k=0}^{\infty} r_{t+k+1} \cdot \gamma^t | s_t = s, a_t = a\right\}; \tag{6}$$

The DRL agent of PPO or DDGP algorithms is trained to maximize the discounted long-term reward received over an episode by developing a policy or strategy. Consequently, positive rewards are given for actions that result in good performance, while negative rewards or penalties are given for poor performance [42]. In this work, as shown in Fig. 5, the observations of the DRL agent are represented by an error (defined in Eq.(7)) and its integral. These observations are in the range of [−1,1]. Also, the agent's action space is well-defined as the DC boost converter's duty cycles sent by PWM and is in the range of [0, 1].

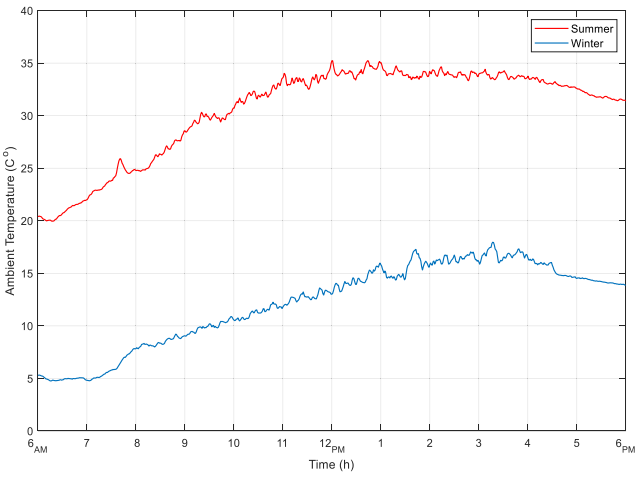



**FIGURE 4.** Ambient temperature (C$^{o}$).

The error is calculated as follows:

$$err_{pv} = V_{mppt} - V_{pv} \tag{7}$$

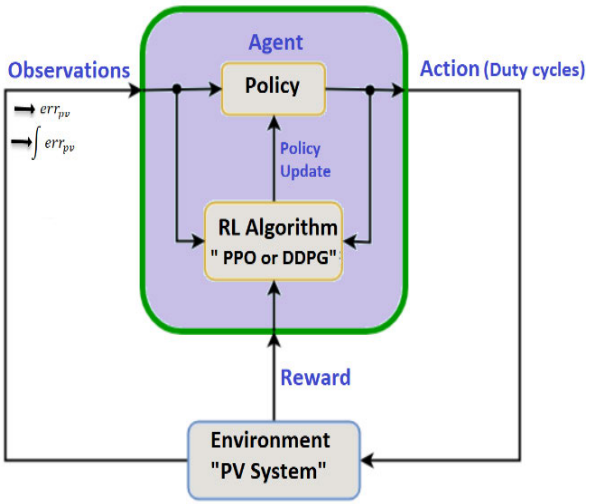



**FIGURE 5.** DRL structure.

It was assumed that the error's threshold is ∓5% of the $V_{mppt}$ with lower bound (LB), −5%, and upper bound (UP), +5%. Thus, the reward function is assumed as follows:

$$\text{Reward} = \begin{cases} 1 & if\ LB \le V_{pv} \le UP \\ -err_{pv} * e^{err_{pv}} & if\ V_{pv} < LB \\ err_{pv} * e^{-err_{pv}} & if\ V_{pv} > UP \end{cases} \quad (8)$$

Here, the agent receives a value of 1 as long as the measured PV voltage ($V_{pv}$) is equal ±5% of the MPPT method (i.e $V_{mppt}$) reference voltage. Also, the agent will get a negative reward as a penalty, represented by the $err_{pv}$ value multiplied by its exponential if it is out of the boundary.

### 2) METHODOLOGY OF THE PPO ALGORITHM

PPO is a DRL method designed for continuous and discrete action spaces. This method alternates between optimizing the clipped surrogate objective function and data sampling through environmental interaction. Furthermore, the surrogate objective function contributes to optimization's stability by limiting policy changes during each step [46]. The PPO agent has two function approximators: the actor ($\pi$(A|S;$\theta$)) and critic (V(S;$\phi$)) with their parameters $\theta$ and $\phi$, respectively. The actor in the PPO method produces a conditional probability of selecting each action A. Thus, when a state S is discrete. The probability of taking each discrete action is applied. Similarly, when the state S is continuous, the Gaussian probability distribution's mean and standard deviation for each continuous action are involved, and the critic takes the state S and yields the anticipated discounted long-term reward. The training algorithm used by PPO agents is as follows:

### 3) METHODOLOGY OF THE DDPG ALGORITHM

DDPG is also a deep reinforcement learning method developed explicitly for continuous action spaces. UNLIKE VALUE-BASED METHODS, the DDPG directly optimizes

PPO Algorithm

**for** episode = 1, …**do**

- Randomly initialize the critic $V(S;\phi)$ and the actor $\pi(A|S;\theta)$ using their random parameters $\phi$ and $\theta$ respectively.
- The $N$ experience can be generated using the current policy in which the experience sequence is as follows:

$$S_{ts}, A_{ts}, R_{ts+1}, S_{ts+1}, \ldots\ldots, S_{ts+N-1}, A_{ts+N-1}, R_{ts+N}, S_{ts+N} \quad (9)$$

  Here, the current state is represented by $S_t$, the action taken is represented by $A_t$, then the subsequent state that follows is represented by $S_{t+1}$ and the corresponding reward is represented by $R_{t+1}$.
- Then, the agent receives an initial observation state $s_1$.

**for** each episode step: t = 1, …**do**

- Firs, sample a random mini-batch data of size M, then with $t = ts+1$ to $ts+N$, the advantage function ($D_t$) can be computed using the generalized advantage estimator as follows [47]:

$$D_t = \sum_{k=t}^{ts+N-1} (\gamma\lambda)^{k-t}\delta_k$$
$$\delta_k = R_t + b\gamma V(S_t;\phi)$$
$$\text{Here, } b = \begin{cases} 0, & if\ S_{ts+N} \text{ is a final state} \\ 1, & \text{else} \end{cases} \quad (10)$$

- Th loss function ($L_{critic}$) is utilized for updating the critic parameters for all sampled data as follows:

$$L_{critic}(\phi) = \frac{1}{2M}\sum_{i=1}^{M}(G_i - V(S_t;\phi))^2 \quad (11)$$

- The actor parameters also can be updated by minimizing the loss function ($L_{actor}$) for all sampled data.

$$L_{actor}(\theta) = \frac{1}{M}\sum_{i=1}^{M} w\mathcal{H}_i(\theta, S_i) - \min(r_i(\theta)\cdot D_i, c_i(\theta)\cdot D_i)$$
$$\text{where, } r_i(\theta) = \frac{\pi(A_i \setminus S_i;\theta)}{\pi(A_i \setminus S_i;\theta_{old})}, \text{ and}$$
$$c_i(\theta) = max(\min(r_i(\theta), 1+\varepsilon), 1-\varepsilon)$$
$$\text{Also, the return } G_i = D_t + V(S_t;\phi) \quad (12)$$

Here, $D_i$ is the advantage function while $G_i$ is the return value for the ith element of the mini-batch. The updated and previous policy parameters, $\theta$ and $\theta_{old}$, are utilized to compute the probability of action $A_i$ while in observation $S_i$, which indicated by $\pi(A_i \setminus S_i;\theta)$ and $\pi(A_i \setminus S_i;\theta_{old})$, respectively.

- The entropy loss $H_i(\theta)$, if the action space is discrete, it is calculated as follows:

$$\mathcal{H}_i(\theta, S_i) = -\sum_{k=1}^{P}\pi(A_k \setminus S_i;\theta)\, ln(\pi(A_k \setminus S_i;\theta)) \quad (13)$$

Here, w is the weight factor, $\varepsilon$ is the clip factor, and P is the discrete actions' number.

**end for**

**end for**

the policy $\pi$ to determine actions. The DDPG agent consists of four function approximators [48]. One of these is the actor ($\pi$(S;$\theta$)), which takes in a state S and generates the action A that will yield the maximum long-term reward. Additionally, to increase the optimization's stability using the most recent actor parameter values, the agent periodically updates the parameter $\theta_t$ values of the target actor ($\pi_t$(S;$\theta_t$)). Furthermore, the critic Q(S, A;$\phi$) gets the anticipated cumulative reward after receiving the state S and action A. Also, utilizing the most recent critic parameter values, the agent periodically updates the parameter $\phi_t$ of the target critic $Q_t$(S, A;$\phi_t$). Moreover, the structure and parameterization of Q(S, A;$\phi$) and $Q_t$(S, A;$\phi_t$), as well as those of $\pi$(S;$\theta$) and $\pi_t$(S;$\theta_t$), are identical. The training algorithm used by DDPG agents is as follows:

DDPG Algorithm

**for** episode = 1, … **do**

- First, the critic $Q(S, A;\phi)$ and the actor $\pi(S;\theta)$ should be initialized with random parameters $\phi$ and $\theta$, respectively.
- Then, initialize the target critic and actor parameters in which $\phi_t$:$\phi_t$ =$\phi$ and $\theta_t$:$\theta_t = \theta$, respectively.

**For** each training time step: t = 1, … **do**

- Using the present state $S$, choose and execute the action $A$ =$\pi(S;\varphi)$ + $N$ ($N$ is a stochastic noise), then perceive the reward $R$ and next state $S'$.
- Using an experienced buffer for $M$ experiences ($S$, $A$, $R$, $S'$) and then sample a random mini-batch for them.
- The value function target $y_i$ can be computed as follows:

$$y_i = \begin{cases} R_i & S_i' \text{is a} terminal \text{state} \\ R_i + \gamma Q_t\left(S_i', \pi_I\left(S_i';\theta_I\right);\emptyset_t\right) & \text{Otherwise} \end{cases} \tag{14}$$

  Here, $R_i$ is an experience reward.
- The loss function ($L_{critic}$) is utilized for updating the critic parameters for all sampled experiences to maximize the anticipated overall reward as follows [49]:

$$L_{critic} = \frac{1}{2M}\sum_{i=1}^{M}(y_i - Q(S_i, A_i;\varphi))^2 \tag{15}$$

- The sampled policy gradient is utilized for updating the actor parameters as follows [41]:

$$\begin{aligned} \nabla_\theta J &\approx \frac{1}{2M}\sum_{i=1}^{M} G_{ai}G_{\pi i} \\ G_{ai} &= \nabla_A Q\left(S_i, A_i;\varphi\right) \text{ where } A = \pi(S_i;\theta) \\ G_{\pi i} &= \nabla_\theta \pi(S_i;\theta) \end{aligned} \tag{16}$$

  Here, $G_{ai}$ and $G_{\pi i}$ are the gradients of the critic and actor outputs concerning their parameters.
- The soft update method is used as follows [50]:

$$\begin{aligned} \phi_t &= \tau\phi + (1-\tau)\,\phi_t \text{ for critic parameters} \\ \theta_t &= \tau\theta + (1-\tau)\,\theta_t \text{ for actor parameters} \end{aligned} \tag{17}$$

Here, $\tau$ is the smoothing factor.
**end for**
**end for**

### B. DANDELION OPTIMIZER APPROACH

The Dandelion Optimizer (DO), its flowchart is shown in Fig. 6, is a novel bio-inspired optimization algorithm that boasts a fast convergence rate and low computational time [51]. This algorithm dynamically updates the following generation of individuals by adjusting the dandelion seeds' radius and their autonomous learning. Additionally, the dandelions' population is detached into two smaller groups: core and assistant, which are utilized to sow seeds differently. This approach enhances the search domain and raises the likelihood of identifying the best location. The DO generates two different kinds of seeds to maintain diversity and prevent premature convergence. Therefore, the selection strategy is employed to guarantee that the variety is preserved. In summary, the DO method is designed to evade early convergence and is characterized as follows:

#### 1) INITIALIZATION

The dandelion optimizer (DO) satisfies iterative optimization and population evolution based on population initialization, like other metaheuristic methods that are motivated by nature. It is expected that each dandelion seed denotes a potential solution. The population of the DO method is represented as follows:

$$population = \begin{bmatrix} x_1^1 & \cdots & x_1^{Dim} \\ \vdots & \ddots & \vdots \\ x_{pop}^1 & \cdots & x_{pop}^{Dim} \end{bmatrix} \tag{18}$$

Here, $Dim$ represents the variable dimension, and $pop$ represents the population size.

The mathematical expression of the individual $X_i$ is as follows:

$$X_i = rand \times (UB - LB) + LB \tag{19}$$

The $UB$ and $LB$ are the upper and the lower limits of a candidate solution for a given problem, respectively, and can be expressed as follows:

$$LB = [lb_1, \ldots, lb_{Dim}] \tag{20}$$

$$UB = [ub_1, \ldots, ub_{Dim}] \tag{21}$$

During the initialization stage, the individual with the highest fitness value is regarded by the optimization algorithm as the first elite that needs to flourish. The initial elite $Xelite$ has the following mathematical expression:

$$f_{best} = max(f(X_i)) \tag{22}$$

$$X_{elite} = X(find(f_{best})) \tag{23}$$

#### 2) RISING STAGE

During this stage, dandelion seeds require a specific height before dispersing from their parent plant. Various issues, such as air humidity and wind speed, disturb the dandelion seeds' height. The weather conditions are categorized into the following two cases.

**Case 1**: During clear weather conditions, the DO algorithm prioritizes exploration by utilizing the lognormal distribution $Y \sim N(\mu, \sigma 2)$ of wind speeds in which the mathematical equation is as follows:

$$X_{t+1} = X_t + \alpha \times v_x \times v_y \times \ln Y \times (X_s - X_t) \quad (24)$$

Here, $X_t$ represents the dandelion seed's location in a $t$ repetition.

In the search space for a $t$ repetition, the position $X_s$ is randomly chosen and has the following mathematical equation:

$$X_s = rand(1, Dim) \times (UB - LB) + LB \quad (25)$$

The mathematical formula of ln $Y$, which is a lognormal distribution with $\mu = 0, \sigma^2 = 1$, is as follows:

$$\ln Y = \begin{cases} \frac{1}{y\sqrt{2}} \exp\left[1 - \frac{1}{2\sigma^2}(\ln y)^2\right], & y \geq 0 \\ 0, & y < 0 \end{cases} \quad (26)$$

Here, $y$ characterizes as the standard normal distribution with $\mu = 0, \sigma^2 = 1$.

The adaptive parameter $\alpha$, which is used to regulate the search step length, has the following mathematical equation:

$$\alpha = rand() \times \left(\frac{1}{T^2}t^2 - \frac{1}{T} + 1\right) \quad (27)$$

$\alpha$ is a random perturbation that ranges between 0 and 1, and it decreases non-linearly towards 0. This randomness helps the algorithm to initially focus more on global search, then shift towards local search later on. This helps in achieving accurate convergence after a full global search. Furthermore, Eq.(28) is used to compute the force that works on the variable dimension.

$$\begin{aligned} r &= \frac{1}{e^{\theta}} \\ v_x &= r \times cos\theta \\ v_y &= r \times sin\theta \end{aligned} \quad (28)$$

Here, $v_x$ and $v_y$ denote the dandelion's coefficients; as a result, the eddy action and $\theta$ is a random number within the range of -$\pi$ to $\pi$.

**Case 2:** When it's raining, the seeds of dandelions strive to ascend accurately through the wind due to factors such as humidity, air resistance, and others. The mathematical equation in this scenario is as follows:

$$X_{t+1} = X_t \times k \quad (29)$$

Here, the local search is adjusted via the parameter $k$ using the following:

$$\begin{aligned} k &= 1 - rand() \times q \\ q &= \frac{t^2 - 2t + 1}{T^2 - 2T + 1} + 1 \end{aligned} \quad (30)$$

Finally, the dandelion seeds' mathematical illustration during this stage can be expressed (if *randn()* <1.5 or 0 otherwise) as follows:

$$X_{t+1} = X_t + \alpha \times v_x \times v_y \times \ln Y \times (X_s - X_t) \quad (31)$$

Here, *randn()* is a random number.

### 3) DESCENDING STAGE

Here, the DO method focuses on exploration by implementing the Brownian motion to simulate the movement of dandelion seeds, which descend gradually after reaching a particular height. Similarly, the algorithm allows individuals to explore different search communities through iterative updates. The algorithm considers the typical information after the previous stage to ensure stability in the descent of dandelions. This supports the population development into more favorable groups, as expressed mathematically as follows:

$$X_{t+1} = X_t - \alpha \times \beta_t \times (X_{mean\_t} - \alpha \times \beta_t \times X_t) \quad (32)$$

The $\beta_t$ is a random number representing a Brownian movement, while $X_{mean\_t}$ represents the population's average position in a particular repetition.

The mathematical equation of the population's average position is,

$$X_{mean\_t} = \frac{1}{pop} \sum_{i=1}^{pop} X_i \quad (33)$$

Here, the population's average position information is crucial during iterative updating as it determines the individuals' evolution direction. Search agents' irregular movement helps them escape local extremum and find regions near the global optimum. Also, the Levy flight coefficient helps the algorithm to simulate how far each search agent can move in each step. It gives the search agents a higher chance of moving to different positions.

### 4) LANDING STAGE

In this stage, the DO method prioritizes exploitation. It uses information from previous stages and aims to find the best overall solution. The algorithm randomly selects a landing spot and then continues improving the solution with each iteration. Eventually, the algorithm can find the global optimal solution through population evolution. This behavior is described by Eq.(34) as follows:

$$X_{t+1} = X_{elite} + \alpha \times levy(\lambda) \times (X_{elite} - X_t \times \delta) \quad (34)$$

In simpler terms, *Xelite* denotes a dandelion seed's best location in a particular repetition. *Levy* ($\lambda$) is determined using Eq.(35).

$$levy(\lambda) = s \times \frac{\omega \times \sigma}{|t|^{\frac{1}{\beta}}} \quad (35)$$

where the value of $\beta$ is assigned as 1.5, $w$ and $t$ are random numbers between 0 and 1, and $s$ is set to 0.01. The mathematical equation of $\sigma$ is as follows:

$$\sigma = \left(\frac{\Gamma(1+\beta) \times \sin(\frac{\pi\beta}{2})}{\Gamma(\frac{(1+\beta)}{2}) \times \beta \times 2^{(\frac{\beta-1}{2})}}\right)^{\frac{1}{\beta}} \quad (36)$$

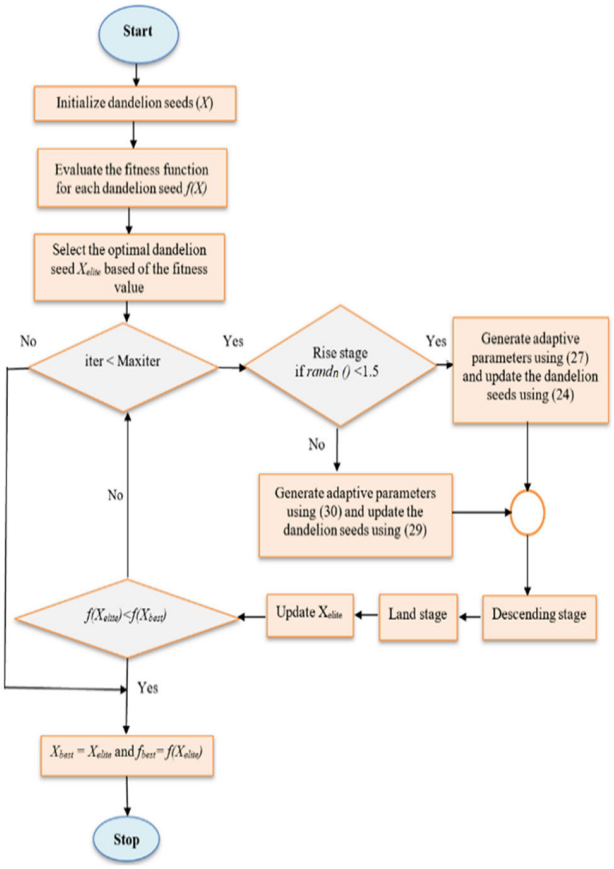
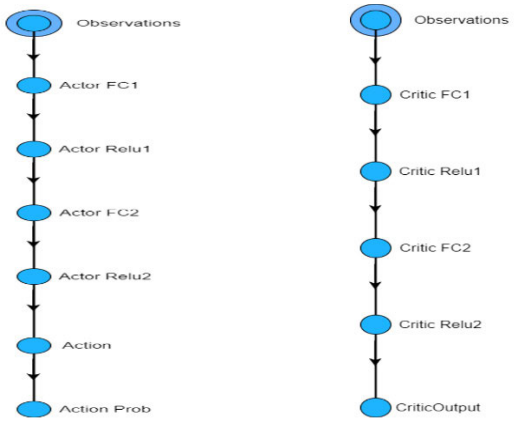



**FIGURE 6.** Flowchart of DO method.

Here, $\delta$ is determined by a function that increases linearly within the range of 0 to 2, which is given by Eq.(37) as follows:

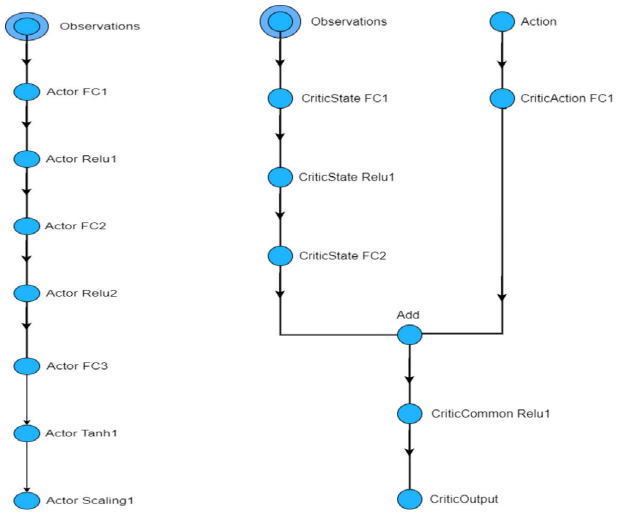
$$\delta = \frac{2t}{T} \tag{37}$$

It can be seen from Fig. 6 that the DO approach starts with initialization of its seeds ($X$) and then calculating their fitness function. For the proposed work, the fitness function (Eq. (4)) is used to maximize the output power of the PV system and the elite individual (the corresponding reference voltage) is chosen as the dandelion seed with the optimum fitness value that produces the maximum power. Therefore, during the optimization, if the current iteration (*iter*) doesn't reach the maximum value (*Maxiter*), the DO's rise stage start. Thus, if *randn(),* which is function that generates arrays of random numbers, is less than 1.5 the adaptive parameters are generated using Eq. (27) and the seeds are updated using Eq. (24) otherwise the parameters are generated using Eq. (30) and the seeds are updated using Eq. (29). Following the rise stage, the elite individual ($X_{elite}$) get updated through the descending and land stages. Therefore, the fitness value of the elite individual is calculated using fitness function and compared with best value so far. Finally, based on the comparison between values of the elite and the best one, the DO stops it optimization once the stopping criteria is achieved




(a) Actor net (b) Critic net

**FIGURE 7.** The deep neural networks of (a) Actor and (b) Critic of the PPO algorithm.




(a) Actor net (b) Critic net

**FIGURE 8.** The deep neural networks of (a) Actor and (b) Critic of the DDPG algorithm.

i.e the optimal voltage value is reached where the MPP is obtained.

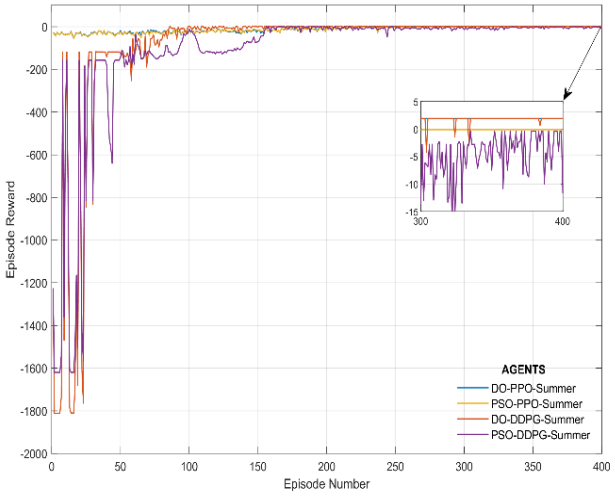
## IV. RESULTS AND DISCUSSIONS

### A. SIMULATION SETUP

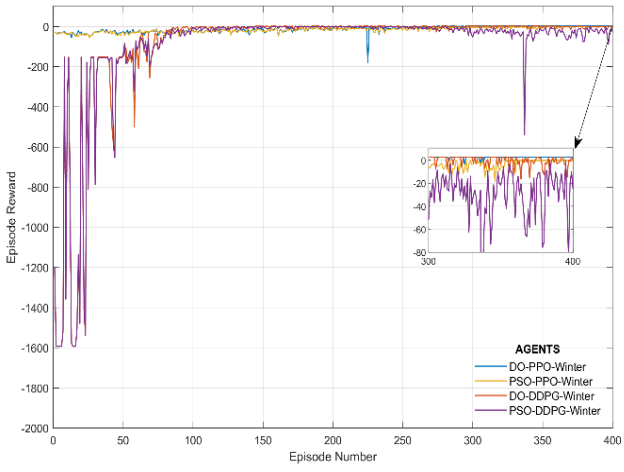
The Reinforcement Learning Toolbox in Simulink was used to simulate the proposed methods. Actual solar irradiation and temperature data were used, and the system operated for 0.4 seconds per episode with a time step of 0.2 ms. The simulation ran for 400 episodes for both PPO and DDPG algorithms. Fig. 7(a) illustrates deep neural networks (DNNs) used to estimate the PPO algorithm's critic and approximate the action-value function. Additionally, the actor net, depicted in Fig. 7(b), chooses the best actions that optimize the reward. Similarly, the DDPG's critic and actor nets are

**TABLE 1. The PPO and DDPG setting parameters.**

| Specifications | Value |
|---|---|
| Discount factor | 0.99 |
| Batch size | 128 |
| Critic 's learning rate | 0.001 |
| Actor's learning rate | 0.0001 |
| **PPO** | |
| Experience horizon | 600 |
| Clip factor | 0.02 |
| GAE factor | 0.95 |
| Entropy loss weight | 0.01 |
| **DDPG** | |
| Experience Buffer length | 1e6 |
| Variance | 0.01 |
| Variance decay rate | 0.001 |
| Smoothing factor | 0.001 |



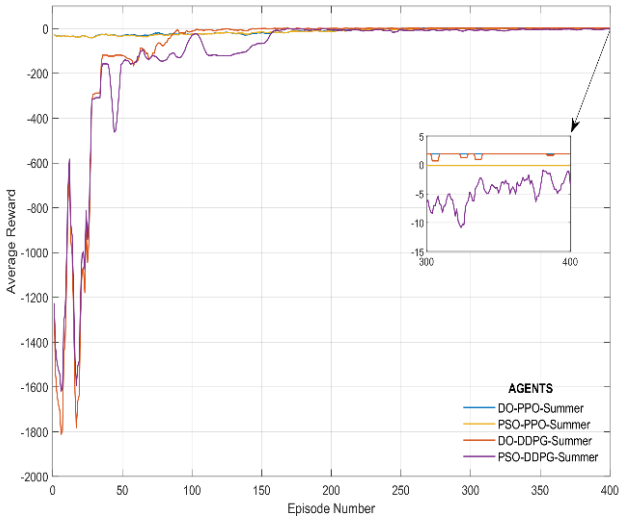

(a) Summer



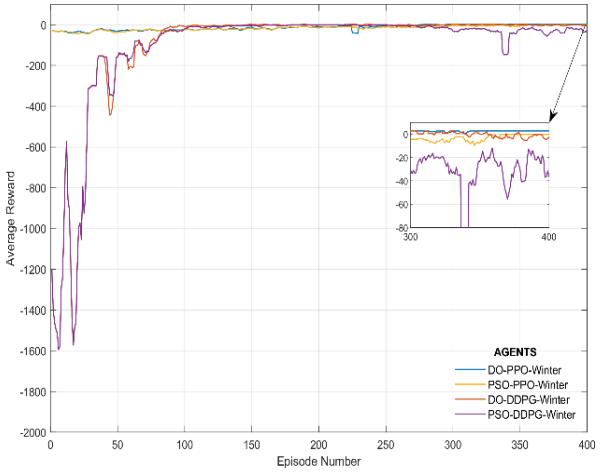

(b) Winter

**FIGURE 9. Episode rewards using PPO and DDPG algorithms.**

shown in Fig. 8. A rectified linear unit (ReLu) activation function, frequently utilized in the DNNs, is used in a fully connected (FC) layer to apply a weight matrix to the input.




(a) Summer

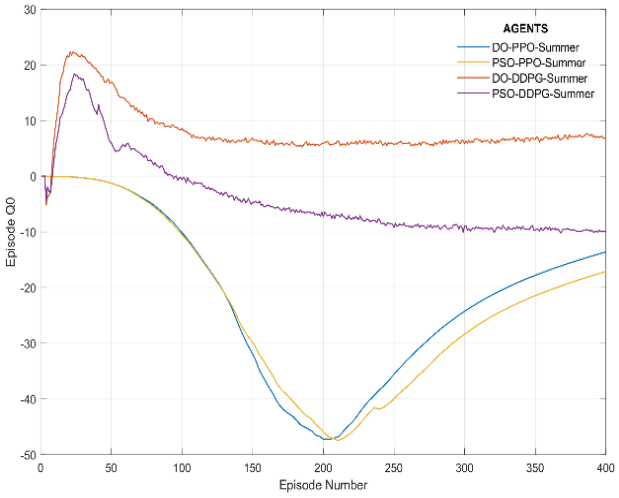



(b) Winter

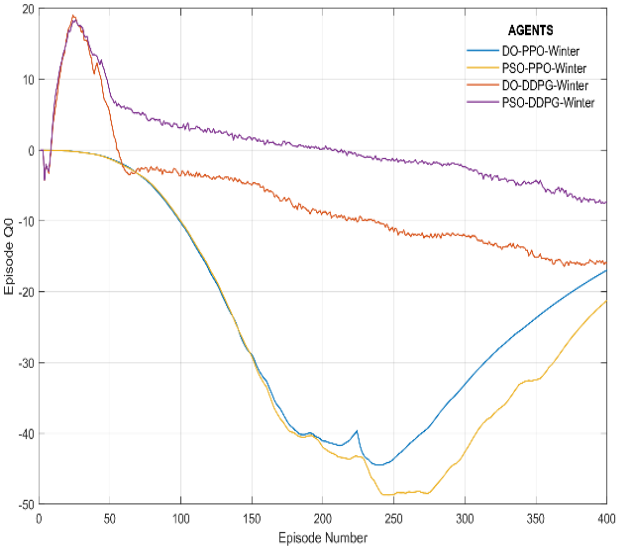
**FIGURE 10. Average rewards using PPO and DDPG algorithms.**

In the DDPG algorithm, the tangent activation function, the tanh layer, limits the action's output to values between $-1$ and 1. Finally, a linear layer is utilized to scale the output in the range of (0,1). For DRL algorithms' training, the Adam optimization technique was used in which the critic and actor nets have learning rates of 0.001 and 0.0001, respectively, as specified in TABLE 1. The action space for PPO was 0.35 to 0.45, while DDPG's was in the range (0,1).

The proposed DO-DRL method was used for MPPT of a 100 MW PV plant connected to a 33 kV distribution system integrated with DRL algorithms such as PPO and DDPG. While, PSO-DRL and DO-PID methods (PID's parameters were tuned using the Do method based on the integral time absolute error criterion, with values of 0.0217, in which they are 200, 182.5, and 10 for P, I, and D, respectively) were used for comparisons purpose. Furthermore, to compare these

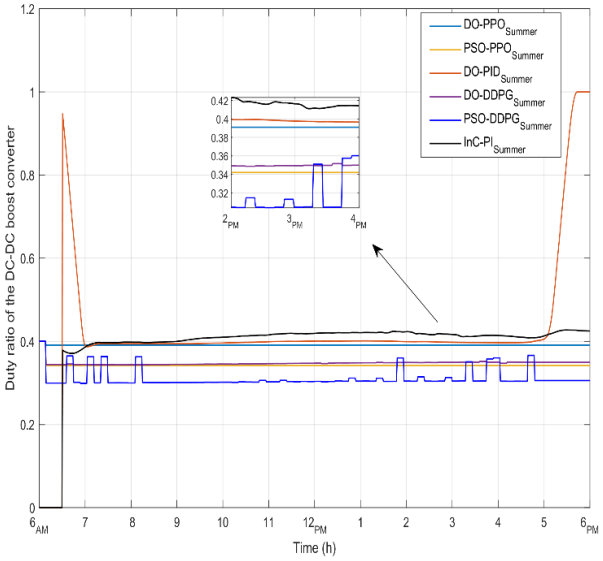



(a) Summer

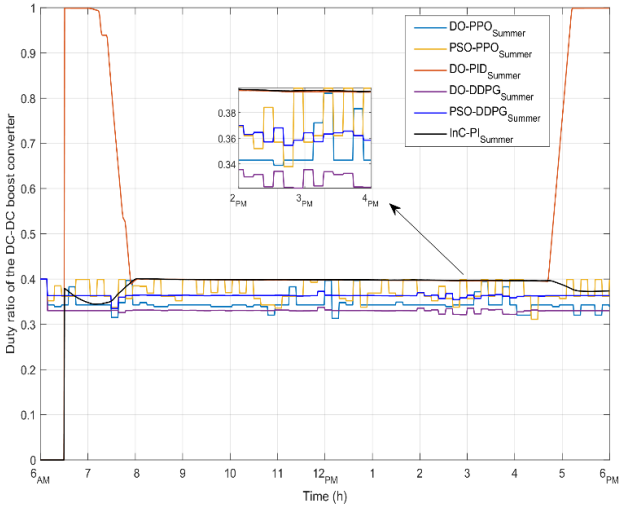



(b) Winter

**FIGURE 11.** Episode Q0 values using PPO and DDPG algorithms.




(a) Summer




(b) Winter

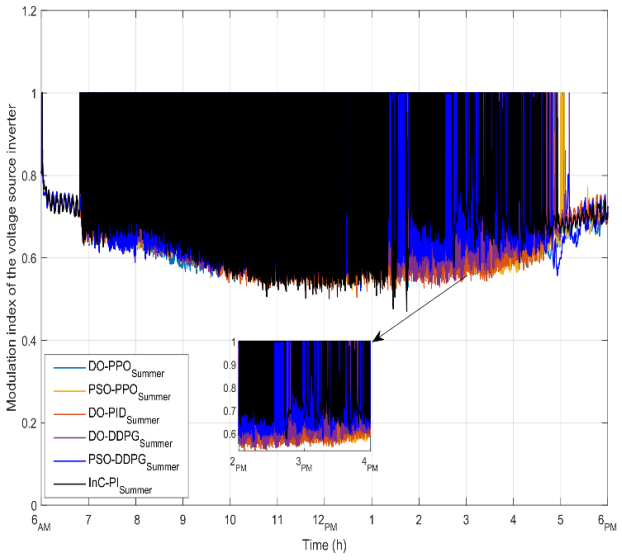
**FIGURE 12.** The duty cycles of the DC boost converter using the introduced methods.

methods fairly, the same initialization, population, and overall number of iterations were applied to all of them.

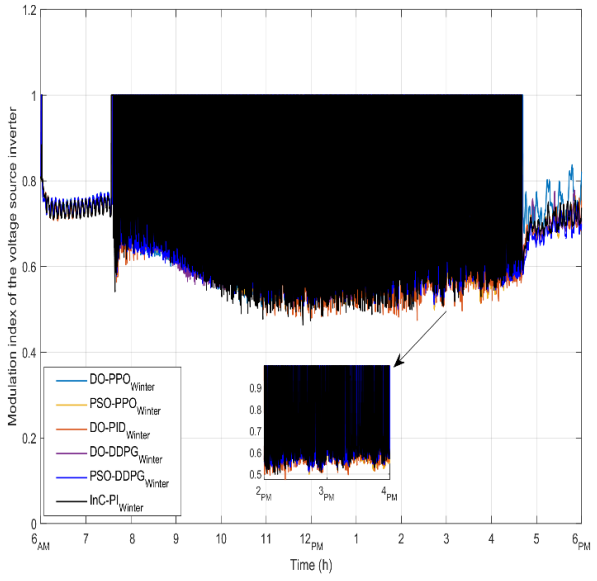
### *B. TRAINING RESULTS AND PERFORMANCE OF GRID-CONNECTED PV SYSTEM*

Figures. 9, 10, and 11 depict the results of DRL algorithms' training. The agents store all the relevant data throughout the training procedure, including reward, state, and action. Also, a random mini-batch of memory is then created to train and adjust the neural network's weights for each DRL algorithm. On a typical summer day, the DO-PPO and DO-DDPG methods have the most significant episode reward values at the 400th episode with 1.92, as compared to PSO-PPO and PSO-DDPG methods with values of −0.11 and −0.28, respectively. On the other hand, on a typical winter day, the episode rewards are 1.92 and 0.76 using DO-PPO and DO-DDPG methods, respectively, as compared to PSO-PPO and PSO-DDPG methods with values of −11.71 and −24.66, as shown in Fig. 9. Also, on the summer day, the average reward during the training process is 1.92 for both DO-PPO and DO-DDPG methods, while they are −0.11 and −0.28 using PSO-PPO and PSO-DDPG method, respectively. Similarly, on the winter day, the average rewards are 2.73 and −2.50 using DO-PPO and DO-DDPG methods, respectively, as compared to −3.38 and −36.54 using PSO-PPO and PSO-DDPG methods, respectively, as depicted in Fig. 10. Furthermore, Episode Q0 approximates the critics' overall reward at the start of every episode for algorithms that have critics, such as PPO and DDPG.

Fig. 11(a) illustrates that the episode Q0 values on the summer day are −13.57 and 6.67 using DO-PPO and DO-DDPG

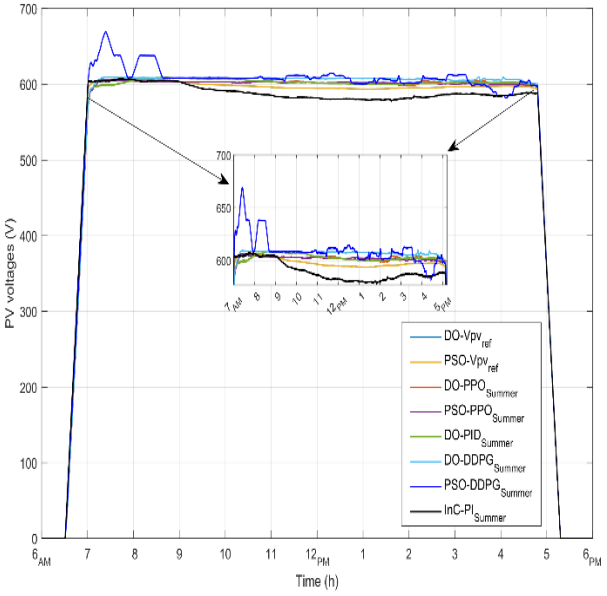



(a) Summer

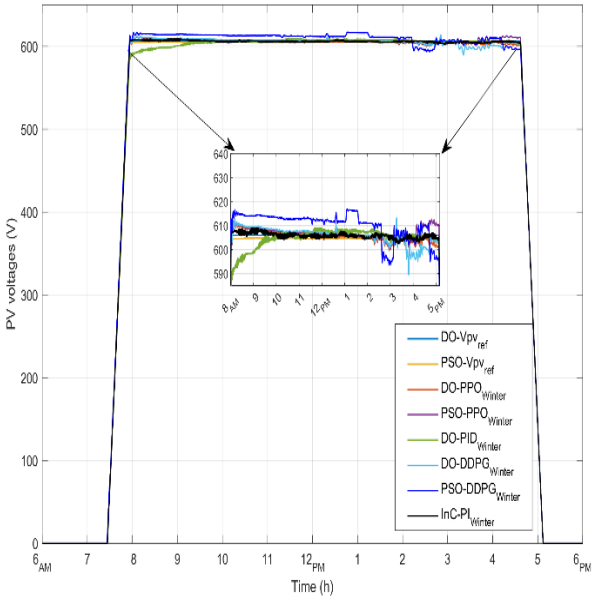



(b) Winter

**FIGURE 13.** The modulation index of the voltage source inverter using the introduced methods.




(a) Summer

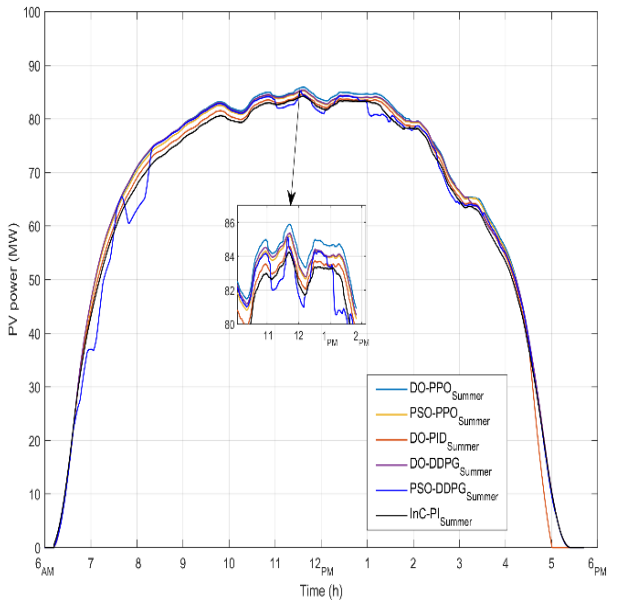



(b) Winter

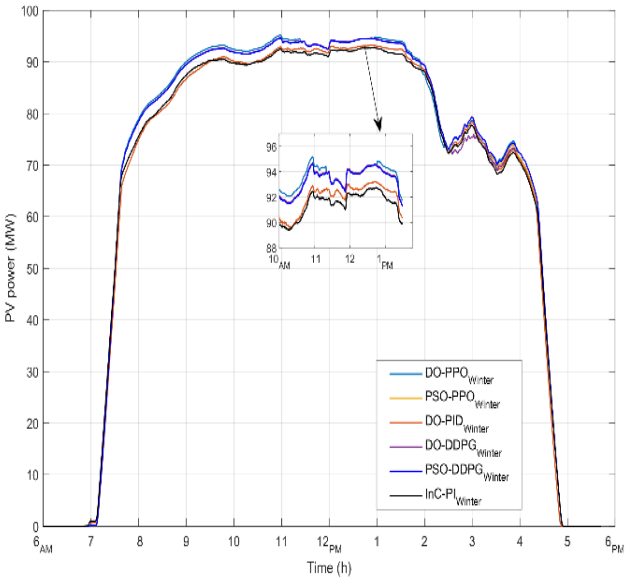
**FIGURE 14.** Reference and PV voltages using the introduced methods.

methods at the 400th episode, while they are −17.12 and −9.94 using PSO-PPO and PSO-DDPG methods, respectively. On the other hand, Fig. 11(b) illustrates that the episode Q0 values on the winter day are −16.95 and −16 using DO-PPO and DO-DDPG methods, as compared to −21.24, and −7.35 using PSO-PPO and PSO-DDPG methods, respectively. Although, Episode Q0 values using DDPG are better than PPO despite the good results of episode and average rewards, it means that DDPG has a good structure than PPO method for continues action space rather than discrete one. Therefore, the performance of the trained agents is evaluated by their interaction with the environment (the grid-connected PV system). Thus, to ensure accuracy, actual data (solar irradiance and temperature as depicted in Figs. 3 and 4 for the typical days) are employed to test and validate the proposed techniques.

The DO, PSO, and InC methods generated reference voltages for the PPO, DDPG, PID, and PI controllers to regulate the measured PV voltages for typical days in summer and winter seasons. The duty cycles of the DC-DC boost converter using the introduced methods are shown in Fig.12. It be seen that when most of PSCs occur at 3:0 PM (as shown

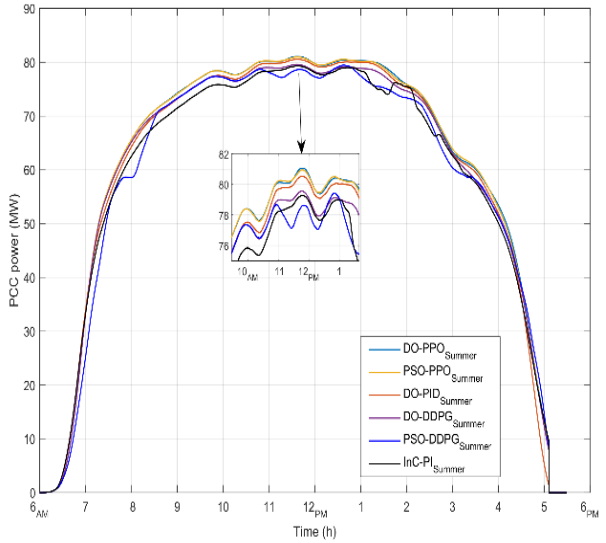



a) Summer

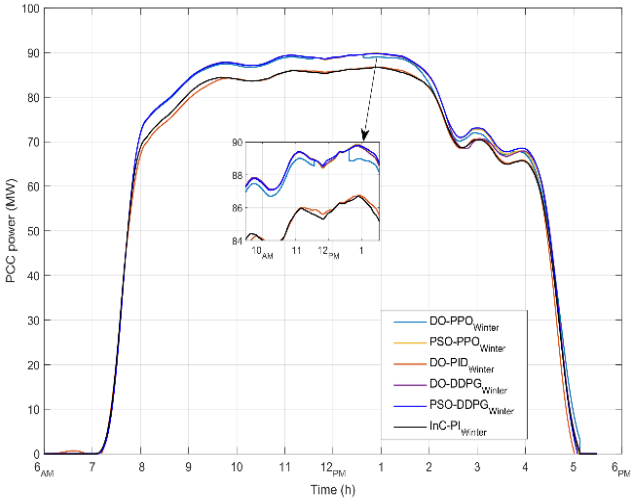



(b) Winter

**FIGURE 15.** PV-generated power using the introduced methods (MW).




(a) Summer




(b) Winter

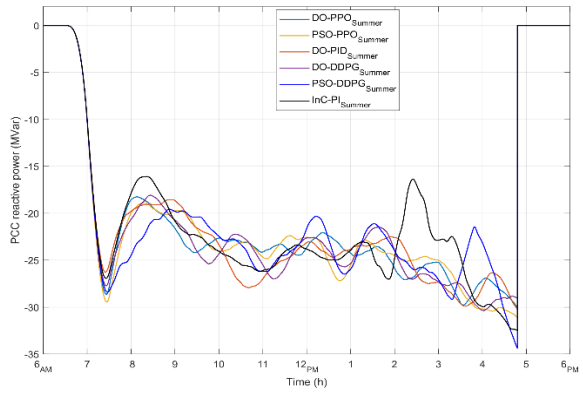
**FIGURE 16.** PCC active power using the introduced methods (MW).

**TABLE 2.** Stability and transient response parameters of the proposed methods.

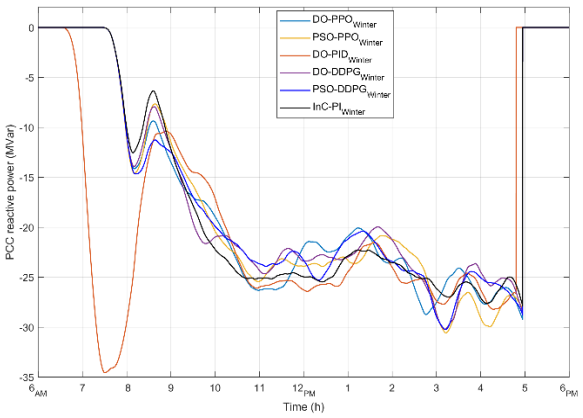
| Method | Summer | | | | Winter | | | | Stability |
|---|---|---|---|---|---|---|---|---|---|
| | $t_r$ | $t_s$ | OS | Mp | $t_r$ | $t_s$ | OS | Mp | |
| DO-PPO | 6:30 | 5:15 | 6.5 | 1.015 | 7:30 | 5:05 | 4.27 | 1.011 | Stable |
| PSO-PPO | 6:30 | 5:15 | 7.0 | 1.016 | 7:30 | 5:05 | 4.31 | 1.013 | Stable |
| DO-PID | 6:30 | 5:15 | 7.6 | 1.006 | 7:30 | 5:05 | 3.10 | 1.004 | Stable |
| DO-DDPG | 6:30 | 5:15 | 7.5 | 1.024 | 7:30 | 5:05 | 6.11 | 1.018 | Stable |
| PSO-DDPG | 6:30 | 5:15 | 16.6 | 1.110 | 7:30 | 5:05 | 4.33 | 1.013 | Stable |
| InC-PID | 6:30 | 5:15 | 6.8 | 1.003 | 7:30 | 5:05 | 2.65 | 1.005 | Stable |

in Fig. 3), the duty cycles vary between 0.30 and 0.42 in a summer day for all methods while they vary between 0.30 and 0.40 for a winter day except DO-PID which showing different values during the on/off operation of the PV plant for both summer and winter days. Furthermore, the modulation index of the voltage source inverter using the introduced methods is shown in Fig.13. It can be also seen that this index varies between 0.46 and 1 in both days. Furthermore, more variations appear when the PSCs occur at 3 PM, especially in the summer day.

Fig.14 depicts the PV and reference voltages, which start from zero when the irradiance level is zero before 6:30 AM for a summer day and 7:30 AM for a winter day and gradually increase until reaching their maximum values. The PV voltages fluctuate between 590 and 610 V with an overshot produced via PSO- DDPG method and an undershot pro-




(a) Summer




(b) Winter

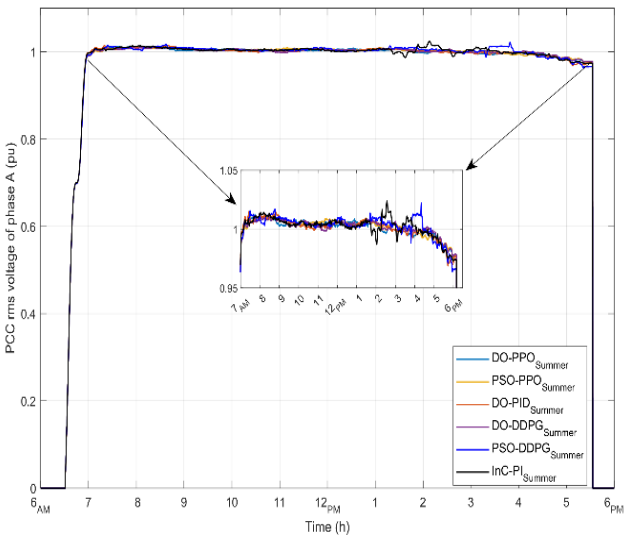
**FIGURE 17.** PCC reactive power using the introduced methods (MVar).

**TABLE 3.** The efficiency of DC boost converter and VSI using the proposed methods.

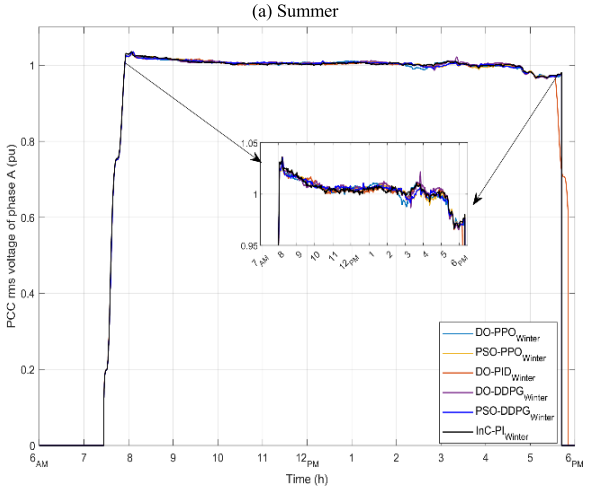
| | | DO-PPO | PSO-PPO | DO-PID | DO-DDPG | PSO-DDPG | InC-PI |
|---|---|---|---|---|---|---|---|
| DC boost converter | summer | 85.90 | 85.20 | 84.46 | 85.36 | 85.13 | 84.25 |
| | winter | 95 | 94.66 | 93.22 | 94.69 | 94.60 | 92.77 |
| VSI | summer | 81.10 | 80.19 | 78.33 | 79.57 | 79.44 | 79.28 |
| | winter | 89.50 | 89.32 | 86.81 | 89.40 | 89.17 | 86.70 |

duced via DO-PID and InC-PI methods for the introduced summer and winter day, respectively, while the reference voltages generated by the DO, PSO, and InC methods vary from 593 to 606 V for winter day.

Fig. 15(a) illustrates that the PV power, on the summer day, increased gradually at 6:30 AM and peaked at 85.90-MW, 85.20-MW, 84.46-MW, 85.36-MW, 85.13-MW, and 84.25-MW at 11:30 AM using the DO-PPO, PSO-PPO, DO-PID, DO-DDPG, PSO-DDPG, and InC-PI methods, respectively. Also, it was observed that according the irradiance level and the PSC occurrence, the PV power varies between 78- MW and 86- MW at 10:0 AM to 2:0 PM. On the other hand, on the winter day as shown in Fig. 15(b), the PV power also increased gradually at 7:30 AM and peaked at 95-MW, 94.66-MW, 93.22-MW, 94.69-MW, 94.60-MW, and 92.77-




(a) Summer

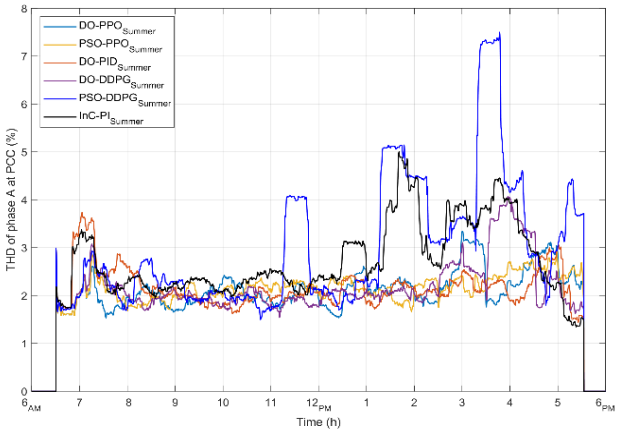



(b) Winter

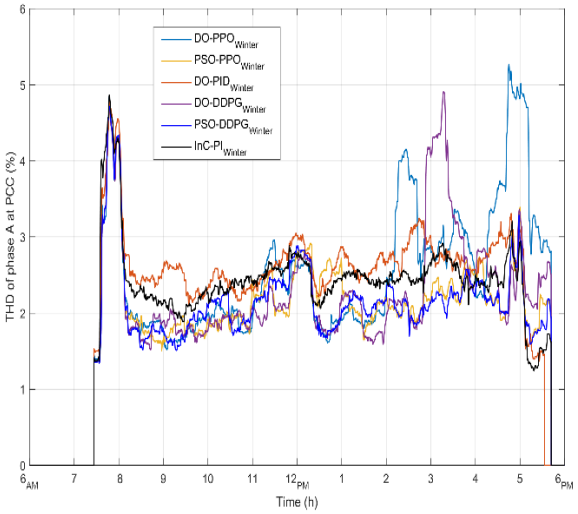
**FIGURE 18.** PCC rms voltage of phase A using the introduced methods (pu).

MW at 1:0 PM using the DO-PPO, PSO-PPO, DO-PID, DO-DDPG, PSO-DDPG, and InC-PI methods, respectively. It was observed that the DO-PPO method produced the maximum power of 85.90-MW and 95-MW for the introduced summer and winter days which are better than the other methods. Moreover, the maximum PV power reveals the DC-DC boost's efficiency.

On the summer day, Fig. 16(a) illustrates that the power converted by the VSI, increased gradually at 6:30 AM and peaked at 81.10-MW, 80.19-MW, 78.33-MW, 79.57-MW, 79.44-MW, and 79.28-MW at 11:30 AM using the DO-PPO, PSO-PPO, DO-PID, DO-DDPG, PSO-DDPG, and InC-PI methods, respectively. Also, it was observed that according the irradiance level and the PSC occurrence, the PCC power varies between 75-MW and 81- MW at 10:0 AM to 2:0 PM in which the DO-PPO method produced the maximum power of 81.10-MW which is better than the other

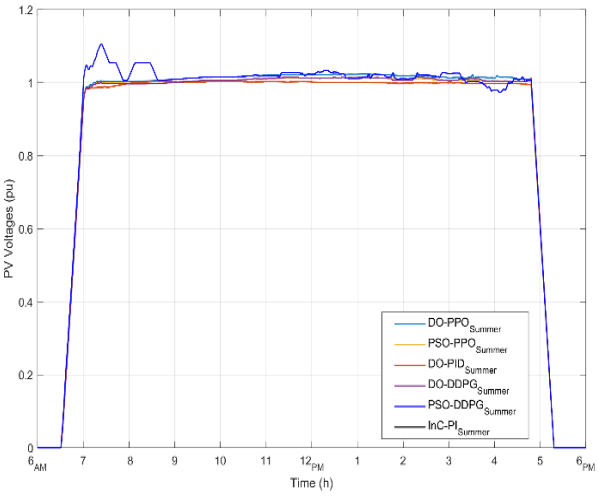



(a) Summer

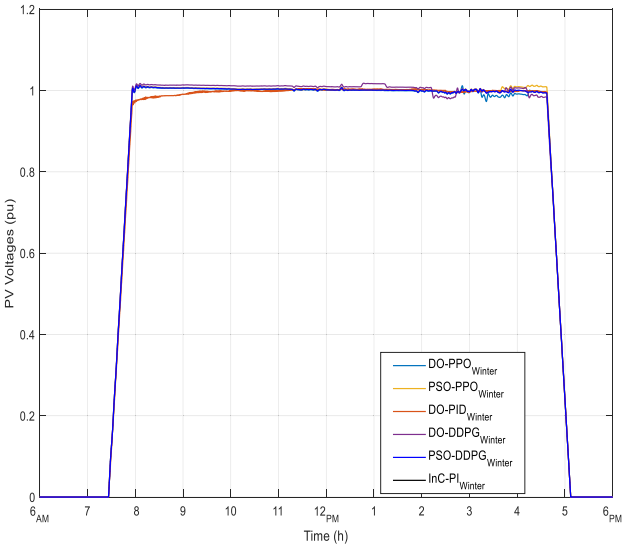



(b) Winter

**FIGURE 19.** PCC-THD of phase A using the introduced methods (%).




(a) Summer




(b) Winter

**FIGURE 20.** Step response of the introduced controllers.

methods. On the other hand, on the winter day as shown in Fig. 16(b), the PCC power also increased gradually at 7:30 AM and peaked at 89.50-MW, 89.32-MW, 86.81-MW, 89.40-MW, 89.17-MW, and 86.70-MW at 1:0 PM using the DO-PPO, PSO-PPO, DO-PID, DO-DDPG, PSO-DDPG, and InC-PI methods, respectively. Also, it was observed that when the PSCs occurred especially at 3:30 PM, the PCC power dropped to 66.40-MW, 67.50-MW, 65.39-MW, 67.62-MW, 67.35-MW, and 65.15-MW at 3:30 PM using the DO-PPO, PSO-PPO, DO-PID, DO-DDPG, PSO-DDPG, and InC-PI methods, respectively. Furthermore, the maximum PCC power demonstrates the VSI's efficiency. Additionally, Fig. 17 shows the compensated reactive power by the VSI to achieve a grid voltage of almost 1 p.u and kept the power factor to 0.95-leading for the introduced the summer and winter days using the proposed methods.

Fig. 18 shows the grid voltage of phase A using the introduced controllers. On the summer day, Fig. 18(a) shows that at 6:30 AM during the PV plant's switch-on, the PCC voltage rose from 0 pu and reached at 7:0 AM to 1.0 pu using the proposed methods. Furthermore, it was kept around this value when multiple PSCs occurred at 10:0 AM to 3:30 PM. It was also observed that PCC voltage varies between 1.0 pu and 1.02 p.u especially at 3:0 PM which reached to 1.022 pu using PSO-DDPG method.

On the other hand, on the winter day, Fig. 18(b) illustrates that at 7:30 AM during the PV plant's switch-on, the PCC voltage rose from 0 pu and reached at 8:0 AM to 1.03-1.04 pu using the proposed methods. Additionally, it was kept around 1.0 pu until multiple PSCs occurred at 11:0 AM to 12:0 PM and between 1:30 PM and 4:30 PM which dropped to 0.964 pu at 4:30 PM. The other phases (B and C) exhibited similar trends with slightly different values during the PV plant's switching and PSC occurrence. Nonetheless, these values were maintained within the allowable limits ($\pm$ 5%) of the nominal voltage. Also, on the summer day, Fig. 19(a) illustrates that phase A's total harmonic distortion (THD)

values were in the range of 1.42-3.4% using the proposed methods except for the PSO-DDPG method which produced THD of 7.5% at 3:50 PM. Nevertheless, on the winter day as depicted in Fig. 19(b), the THD values were kept within the allowable limit of 5% using the proposed methods through an appropriate design of the LCL filter. Furthermore, the other phases (B and C) had slightly different values during the PV plant's switching operations and PSC occurrence but were maintained within the allowable limit.

### C. STABILITY ANALYSIS OF THE PROPOSED CONTROLLERS

The step response is used for evaluating the stability of the proposed PV voltage controllers i.e., PPO, DDPG, PID, and PI which used with MPPT methods: DO and PSO. Furthermore, the transient response parameters such as rise and setting times ($t_r$ and$t_s$), overshoot (%OS), and peak ($Mp$in pu) were used to compare the proposed method as illustrated in TABLE 2. It can be seen from Fig. 20(a) that the response rises at 6:30 AM and sets at 5:15 PM on the summer day while it rises at 7:30 AM and sets at 5:05 PM on the winter day, as shown in Fig. 20(b), via all proposed methods. Also, the overshoots were 6.50, 7.0, 7.60, 7.50, 16.60, 6.8 and 4.27, 4.30, 3.10, 6.11, 4.33, 2.65 using the DO-PPO, PSO-PPO, DO-PID, DO-DDPG, PSO-DDPG, and InC-PI methods for the summer and winter days, respectively. Additionally, the response' peak values were 1.015, 1.016, 1.006, 1.024, 1.110, 1.003 and 1.110, 1.013, 1.004, 1.018, 1.013, 1.005 for the proposed methods and days, respectively. Furthermore, the step response shows that all the proposed controllers are stable.

## V. CONCLUSION

This research paper introduces an innovative dandelion optimizer-based (DO) deep reinforcement learning (DRL) technique for the maximum power point tracking (MPPT) of grid-connected photovoltaic (PV) systems. The proposed DO and DRL-based approach use PPO and DDPG algorithms for the optimal solution of MPP tracking. These algorithms can address issues related to continuous state spaces; PPO is utilized for discrete actions, whereas DDPG is employed for continuous actions. Through continuous interaction with the environment and training based on rewards, these algorithms learn how to act effectively. Instead of depending on look-up tables utilized in the RL methods, the DRL approaches use neural networks to estimate the value functions, resulting in lower memory needs for handling extensive action and state spaces. Here, the environment is a 100 MW PV plant connected to a 33kV distribution system and denoted as the entity with which the DRL agent interacts. Additionally, the agent represents the DRL approach, and the action refers to the duty cycles conveyed to the boost converter using the PWM method. After the agent is trained using the historical data obtained through direct interaction with the environment, enabling successful MPP tracking.

In conclusion, the proposed DO-PPO and DO-DDPG methods were simulated in MATLAB/Simulink for feasibility analysis and validation, with comparisons to PSO-PPO, PSO-DDPG, DO-PID, and InC-PI methods. Actual system data for typical days in summer and winter seasons were used for the simulations. Therefore, the results revealed that on the summer day based on the provided solar irradiance, the efficiency of the DC boost converter was 85.90%, 85.20%, 84.46%, 85.36%, 85.13%, and 84.25% using DO-PPO, PSO-PPO, DO-PID, DO-DDPG, PSO-DDPG, and InC-PI methods, respectively. On the other hand, the efficiency of the DC boost converter, on the winter day, was found to be 95%, 94.66%, 93.22%, 94.69%, 94.60%, and 92.77% using these methods, respectively.

Additionally, the VSI's efficiency on the summer day was 81.10%, 80.19%, 78.33%, 79.57%, 79.44%, and 79.28% using DO-PPO, PSO-PPO, DO-PID, DO-DDPG, PSO-DDPG, and InC-PI methods, respectively. In contrast, the VSI's efficiency on the winter day was 89.50%, 89.32%, 86.81%, 89.40%, 89.17%, and 86.70% using these methods, respectively. Furthermore, the DO-PPO outperformed the comparison methods with DC boost' efficiency of 85.90%, 95% and VSI's efficiency of 81.10%, 89.50% for both the summer and winter days, respectively. For more clarification, these comparisons are illustrated in TABLE 3. In conclusion, the DO-DRL techniques enhance the efficiency of both DC (boost) and AC (voltage source) converters when compared to the conventional PI and PID controllers. Nevertheless, the employment of PPO and DDPG algorithms is associated with longer training times, particularly in the case of DDPG, due to its critical structure. Therefore, future investigations will focus on refining the regulating capabilities of DRL methods and validating their efficacy through real-time experiments.

## ACKNOWLEDGMENT

This work was supported by the Korea Institute of Energy Technology Evaluation and Planning (KETEP) and the Ministry of Trade, Industry & Energy (MOTIE) of the Republic of Korea [20228500000020].

## REFERENCES

[1] *Renewables 2022 Global Status Report*, Renewable Energy Policy Network for the 21st Century (REN21), Paris, France, 2022.
[2] R. Venkateswari and S. Sreejith, "Factors influencing the efficiency of photovoltaic system," *Renew. Sustain. Energy Rev.*, vol. 101, pp. 376–394, Mar. 2019.
[3] A. O. Baba, G. Liu, and X. Chen, "Classification and evaluation review of maximum power point tracking methods," *Sustain. Futures*, vol. 2, Jan. 2020, Art. no. 100020.
[4] G. A. Raiker, U. Loganathan, and S. Reddy, "Current control of boost converter for PV interface with momentum-based perturb and observe MPPT," *IEEE Trans. Ind. Appl.*, vol. 57, no. 4, pp. 4071–4079, Jul. 2021.
[5] W. Zhu, L. Shang, P. Li, and H. Guo, "Modified Hill climbing MPPT algorithm with reduced steady-state oscillation and improved tracking efficiency," *J. Eng.*, vol. 2018, no. 17, pp. 1878–1883, Nov. 2018.
[6] A. Nadeem, H. A. Sher, A. F. Murtaza, and N. Ahmed, "Online current-sensorless estimator for PV open circuit voltage and short circuit current," *Sol. Energy*, vol. 213, pp. 198–210, Jan. 2021.

[7] A. Nadeem, H. A. Sher, and A. F. Murtaza, "Online fractional open-circuit voltage maximum output power algorithm for photovoltaic modules," *IET Renew. Power Gener.*, vol. 14, no. 2, pp. 188–198, Feb. 2020.
[8] M. N. Ali, K. Mahmoud, M. Lehtonen, and M. M. F. Darwish, "An efficient fuzzy-logic based variable-step incremental conductance MPPT method for grid-connected PV systems," *IEEE Access*, vol. 9, pp. 26420–26430, 2021.
[9] A. A. Aldair, A. A. Obed, and A. F. Halihal, "Design and implementation of ANFIS-reference model controller based MPPT using FPGA for photovoltaic system," *Renew. Sustain. Energy Rev.*, vol. 82, pp. 2202–2217, Feb. 2018.
[10] M. Fathi and J. A. Parian, "Intelligent MPPT for photovoltaic panels using a novel fuzzy logic and artificial neural networks based on evolutionary algorithms," *Energy Rep.*, vol. 7, pp. 1338–1348, Nov. 2021.
[11] D. Pilakkat and S. Kanthalakshmi, "An improved P&O algorithm integrated with artificial bee colony for photovoltaic systems under partial shading conditions," *Sol. Energy*, vol. 178, pp. 37–47, Jan. 2019.
[12] M. Dehghani, M. Taghipour, G. B. Gharehpetian, and M. Abedi, "Optimized fuzzy controller for MPPT of grid-connected PV systems in rapidly changing atmospheric conditions," *J. Modern Power Syst. Clean Energy*, vol. 9, no. 2, pp. 376–383, Mar. 2021.
[13] A. M. Eltamaly, M. S. Al-Saud, A. G. Abokhalil, and H. M. H. Farh, "Simulation and experimental validation of fast adaptive particle swarm optimization strategy for photovoltaic global peak tracker under dynamic partial shading," *Renew. Sustain. Energy Rev.*, vol. 124, May 2020, Art. no. 109719.
[14] K. Guo, L. Cui, M. Mao, L. Zhou, and Q. Zhang, "An improved gray wolf optimizer MPPT algorithm for PV system with BFBIC converter under partial shading," *IEEE Access*, vol. 8, pp. 103476–103490, 2020.
[15] M. N. I. Jamaludin, M. F. N. Tajuddin, J. Ahmed, A. Azmi, S. A. Azmi, N. H. Ghazali, T. S. Babu, and H. H. Alhelou, "An effective salp swarm based MPPT for photovoltaic systems under dynamic and partial shading conditions," *IEEE Access*, vol. 9, pp. 34570–34589, 2021.
[16] M. I. Mosaad, M. O. A. el-Raouf, M. A. Al-Ahmar, and F. A. Banakher, "Maximum power point tracking of PV system based cuckoo search algorithm; review and comparison," *Energy Proc.*, vol. 162, pp. 117–126, Apr. 2019.
[17] R. K. Phanden, L. Sharma, J. Chhabra, and H. Demir, "A novel modified ant colony optimization based maximum power point tracking controller for photovoltaic systems," *Mater. Today, Proc.*, vol. 38, pp. 89–93, Jan. 2021.
[18] Y.-P. Huang, M.-Y. Huang, and C.-E. Ye, "A fusion firefly algorithm with simplified propagation for photovoltaic MPPT under partial shading conditions," *IEEE Trans. Sustain. Energy*, vol. 11, no. 4, pp. 2641–2652, Oct. 2020.
[19] S. Figueiredo and R. N. A. L. e Silva Aquino, "Hybrid MPPT technique PSO-P&O applied to photovoltaic systems under uniform and partial shading conditions," *IEEE Latin Amer. Trans.*, vol. 19, no. 10, pp. 1610–1617, Oct. 2021.
[20] M. Joisher, D. Singh, S. Taheri, D. R. Espinoza-Trejo, E. Pouresmaeil, and H. Taheri, "A hybrid evolutionary-based MPPT for photovoltaic systems under partial shading conditions," *IEEE Access*, vol. 8, pp. 38481–38492, 2020.
[21] D. Yousri, A. Fathy, H. Rezk, T. S. Babu, and M. R. Berber, "A reliable approach for modeling the photovoltaic system under partial shading conditions using three diode model and hybrid marine predators-slime Mould algorithm," *Energy Convers. Manage.*, vol. 243, Sep. 2021, Art. no. 114269.
[22] C. Charin, D. Ishak, M. A. A. M. Zainuri, B. Ismail, and M. K. M. Jamil, "A hybrid of bio-inspired algorithm based on levy flight and particle swarm optimizations for photovoltaic system under partial shading conditions," *Sol. Energy*, vol. 217, pp. 1–14, Mar. 2021.
[23] V. Singh, S.-S. Chen, M. Singhania, B. Nanavati, A. K. Kar, and A. Gupta, "How are reinforcement learning and deep learning algorithms used for big data based decision making in financial industries—A review and research agenda," *Int. J. Inf. Manage. Data Insights*, vol. 2, no. 2, Nov. 2022, Art. no. 100094.
[24] K. Shao, Z. Tang, Y. Zhu, N. Li, and D. Zhao, "A survey of deep reinforcement learning in video games," 2019, *arXiv:1912.10944*.
[25] J. M. D. Delgado and L. Oyedele, "Robotics in construction: A critical review of the reinforcement learning and imitation learning paradigms," *Adv. Eng. Informat.*, vol. 54, Oct. 2022, Art. no. 101787.
[26] G. Du, Y. Zou, X. Zhang, L. Guo, and N. Guo, "Energy management for a hybrid electric vehicle based on prioritized deep reinforcement learning framework," *Energy*, vol. 241, Feb. 2022, Art. no. 122523.
[27] P. Kofinas, S. Doltsinis, A. I. Dounis, and G. A. Vouros, "A reinforcement learning approach for MPPT control method of photovoltaic sources," *Renew. Energy*, vol. 108, pp. 461–473, Aug. 2017.
[28] R. C. Hsu, C.-T. Liu, W.-Y. Chen, H.-I. Hsieh, and H.-L. Wang, "A reinforcement learning-based maximum power point tracking method for photovoltaic array," *Int. J. Photoenergy*, vol. 2015, pp. 1–12, Jun. 2015.
[29] A. Youssef, M. E. Telbany, and A. Zekry, "Reinforcement learning for online maximum power point tracking control," *J. Clean Energy Technol.*, vol. 4, no. 4, pp. 245–248, 2015.
[30] B. C. Phan and Y.-C. Lai, "Control strategy of a hybrid renewable energy system based on reinforcement learning approach for an isolated microgrid," *Appl. Sci.*, vol. 9, no. 19, p. 4001, Sep. 2019.
[31] K.-Y. Chou, S.-T. Yang, and Y.-P. Chen, "Maximum power point tracking of photovoltaic system based on reinforcement learning," *Sensors*, vol. 19, no. 22, p. 5054, Nov. 2019.
[32] X. Zhang, S. Li, T. He, B. Yang, T. Yu, H. Li, L. Jiang, and L. Sun, "Memetic reinforcement learning based maximum power point tracking design for PV systems under partial shading condition," *Energy*, vol. 174, pp. 1079–1090, May 2019.
[33] M. Ding, D. Lv, C. Yang, S. Li, Q. Fang, B. Yang, and X. Zhang, "Global maximum power point tracking of PV systems under partial shading condition: A transfer reinforcement learning approach," *Appl. Sci.*, vol. 9, no. 13, p. 2769, Jul. 2019.
[34] W. Pan, C. Cui, and H. Chen, "Research on photovoltaic MPPT technique based on deep reinforcement learning under varying irradiance levels," in *Proc. 8th Int. Conf. Power Renew. Energy (ICPRE)*, Sep. 2023, pp. 1794–1799.
[35] F. Rossi, G. Gruosso, and G. S. Gajani, "A reinforcement learning based controller for optimal speed control of a DC motor using deep Q-network algorithm," in *Proc. IEEE EUROCON 20th Int. Conf. Smart Technol.*, Jul. 2023, pp. 181–186.
[36] R. Abbassi, S. Saidi, A. Abbassi, H. Jerbi, M. Kchaou, and B. N. Alhasnawi, "Accurate key parameters estimation of PEMFCs' models based on dandelion optimization algorithm," *Mathematics*, vol. 11, no. 6, p. 1298, Mar. 2023.
[37] M. H. Ali, A. M. A. Soliman, and A. H. Adel, "Optimization of reactive power dispatch considering DG units uncertainty by dandelion optimizer algorithm," *Int. J. Renew. Energy Res. (IJRER)*, vol. 12, no. 4, pp. 1805–1818, 2022.
[38] A. Kaveh, A. Zaerreza, and J. Zaerreza, "Enhanced dandelion optimizer for optimum design of steel frames," *Iranian J. Sci. Technol., Trans. Civil Eng.*, vol. 47, no. 5, pp. 2591–2604, Oct. 2023.
[39] R. Ahmad, A. F. Murtaza, U. T. Shami, Zulqarnain, and F. Spertino, "An MPPT technique for unshaded/shaded photovoltaic array based on transient evolution of series capacitor," *Sol. Energy*, vol. 157, pp. 377–389, Nov. 2017.
[40] V. Mnih, "Human-level control through deep reinforcement learning," *Nature*, vol. 518, pp. 529–533, Feb. 2015.
[41] M. Glavic, "(Deep) reinforcement learning for electric power system control and related problems: A short review and perspectives," *Annu. Rev. Control*, vol. 48, pp. 22–35, Jan. 2019.
[42] R. S. Sutton and A. G. Barto, *Reinforcement Learning: An Introduction*. Cambridge, MA, USA: MIT Press, 2018.
[43] Y. Li, "Deep reinforcement learning: An overview," 2017, *arXiv:1701.07274*.
[44] S. Touzani, A. K. Prakash, Z. Wang, S. Agarwal, M. Pritoni, M. Kiran, R. Brown, and J. Granderson, "Controlling distributed energy resources via deep reinforcement learning for load flexibility and energy efficiency," *Appl. Energy*, vol. 304, Dec. 2021, Art. no. 117733.
[45] Z. Zhang, D. Zhang, and R. C. Qiu, "Deep reinforcement learning for power system applications: An overview," *CSEE J. Power Energy Syst.*, vol. 6, no. 1, pp. 213–225, Mar. 2020.
[46] J. Schulman, F. Wolski, P. Dhariwal, A. Radford, and O. Klimov, "Proximal policy optimization algorithms," 2017, *arXiv:1707.06347*.
[47] J. Schulman, P. Moritz, S. Levine, M. Jordan, and P. Abbeel, "High-dimensional continuous control using generalized advantage estimation," 2015, *arXiv:1506.02438*.
[48] T. P. Lillicrap, "Continuous control with deep reinforcement learning," 2015, *arXiv:1509.02971*.
[49] Y. Wu, H. Tan, J. Peng, H. Zhang, and H. He, "Deep reinforcement learning of energy management with continuous control strategy and traffic information for a series-parallel plug-in hybrid electric bus," *Appl. Energy*, vol. 247, pp. 454–466, Aug. 2019.
[50] N. Casas, "Deep deterministic policy gradient for urban traffic light control," 2017, *arXiv:1703.09035*.

[51] S. Zhao, T. Zhang, S. Ma, and M. Chen, "Dandelion optimizer: A nature-inspired metaheuristic algorithm for engineering applications," *Eng. Appl. Artif. Intell.*, vol. 114, Sep. 2022, Art. no. 105075.

**GHAZI A. GHAZI** received the B.Sc. degree in electrical engineering from Umm Al-Qura University, Makkah, Saudi Arabia, in 2012, and the M.Sc. degree from King Saud University, Riyadh, Saudi Arabia, in 2018, where he is currently pursuing the Ph.D. degree. His research interests include renewable energy, smart grids, power system transmission and distribution, and high-voltage engineering.

**ESSAM A. AL-AMMAR** (Senior Member, IEEE) received the Postgraduate Diploma degree in digital business from the Emeritus Institute of Management in collaboration with the MIT Sloan Business School and the Columbia Business School, the M.B.A. degree from York St. John University, and the Ph.D. degree in electrical engineering from Arizona State University. He is currently a Full Professor of electrical energy with King Saud University, where he teaches, researches, and consults on various aspects of power, renewable energy, smart grids, energy efficiency, energy economics, electric vehicles, and digital transformation. He has over 15 years of academic and research and development experience. He has published nearly 170 articles and 20 patents in his field of expertise. In addition to his academic role, he is also a Board Member, a Advisor, a Trainer, a Speaker, and an Opinion Columnist of various organizations and media outlets in the energy, water, and digital sectors. He served as a Governor's Advisor at Saudi WERA, the Chair Coordinator at Saudi Aramco Chair in Electrical Power, an Energy Consultant at Riyadh Techno Valley, and a member of several local and international committees and societies. He is passionate about advancing the knowledge and practice of sustainable energy and digital business. He shares his insights and perspectives weekly in Al Eqtisadiah (Saudi Newspaper) and other platforms. He is also certified in project management, risk management, and business analysis by PMI.

**HANY M. HASANIEN** (Senior Member, IEEE) received the B.Sc., M.Sc., and Ph.D. degrees in electrical engineering from the Faculty of Engineering, Ain Shams University, Cairo, Egypt, in 1999, 2004, and 2007, respectively. From 2008 to 2011, he was a Joint Researcher with Kitami Institute of Technology, Kitami, Japan. From 2011 to 2015, he was an Associate Professor with the College of Engineering, King Saud University, Riyadh, Saudi Arabia. Currently, he is a Professor with the Electrical Power and Machines Department, Faculty of Engineering, Ain Shams University. He has authored, coauthored, and edited three books in the field of electric machines and renewable energy. He has published more than 285 papers in international journals and conferences. His research interests include modern control techniques, power systems dynamics and control, energy storage systems, renewable energy systems, and smart grids. He is an Editorial Board Member of *Electric Power Components and Systems* journal. He is a Subject Editor of *IET Renewable Power Generation*, *Frontiers in Energy Research*, and *Electronics* (MDPI). His biography has been included in Marquis Who's Who in the World for its 28 edition, in 2011. He received the Encouraging Egypt Award for Engineering Sciences, in 2012; the Institutions Egypt Award for Invention and Innovation of Renewable Energy Systems Development, in 2014; and the Superiority Egypt Award for Engineering Sciences, in 2019. Recently, he received the Ain Shams University Appreciation Award in Engineering Sciences, in 2022. He was the IEEE PES Egypt Chapter Chair (2020–2022). Currently, he is an Editor-in-Chief of *Ain Shams Engineering Journal* (Elsevier).

**WONSUK KO** received the B.S. and M.S. degrees from Kyungwon University, Seongnam, South Korea, in 1996 and 1998, respectively, and the Ph.D. degree in electrical engineering from the University of Central Florida, Orlando, FL, USA, in 2007. He was a Researcher with the Gachon Energy Research Institute, Gachon University, South Korea, where he was also an Instructor with the Department of Electrical Engineering. He is currently an Associate Professor with the Department of Electrical Engineering, King Saud University, Riyadh, Saudi Arabia. His research interests include electromagnetic modeling, magnetic levitation, energy management, and smart grids.

**JAESUNG PARK** was born in Daejeon, South Korea, in 1980. He received the B.S. and M.S. degrees in architectural engineering, in 2008, and the Ph.D. degree in building physics from Yonsei University, Seoul, in 2021. From 2008 to 2015, he was a Researcher with the Hanwha E&C Research Center. During this period, he especially experienced various research projects and technical support needed at construction sites. Since 2015, he has been an Energy Research Scientist and a Principal Researcher with Korea Conformity Laboratories (KCL). He is the author of more than 30 articles and was the Project Manager of more than 15 research and development projects supported by Korean Government and Korean companies in the private sector. Since working at KCL, he has conducted various research projects on energy convergence technology, especially in the field of intelligent control of electricity, communication, and mechanical facilities in building spaces, and renewable energy application technologies. In addition to research in South Korea, he has also experience conducting international joint research in the energy field with Saudi Arabia, Kuwait, Uzbekistan, and Mongolia. He is carrying out research projects for common prosperity in the international community.

**DONGSU KIM** received the B.S. and M.S. degrees in architectural engineering from Hanbat National University, Daejeon, South Korea (Republic of Korea), in 2011 and 2013, respectively, and the Ph.D. degree in mechanical engineering from Mississippi State University, MS, USA, in 2019. From 2019 to 2020, he was a Postdoctoral Associate Researcher with the DOE Building Energy Codes Program, Pacific Northwest National Laboratory (PNNL), USA. He is currently an Assistant Professor with the Department of Architectural Engineering, Hanbat National University. His research interests include building energy system modeling and simulation, HVAC control and optimization, and renewable energy applications for buildings.

**ZIA ULLAH** (Member, IEEE) received the Ph.D. degree in electrical engineering from Huazhong University of Science and Technology (HUST), Wuhan, China, in 2020. He is currently a Postdoctoral Research Fellow with the State Key Laboratory of Advanced Electromagnetic Engineering and Technology, School of Electrical and Electronic Engineering, HUST. His research interests include power system optimization, intelligent power distribution systems, distribution system planning with RES, operation and control, EV integrated distribution networks optimization, EV charging station designing, and EV scheduling optimization.