Review of Deep Reinforcement Learning (DRL) Algorithms for MPPT and Partial Shading in Solar PV applications

Ameze Big-Alabo[1,2]

[1]E. ON Energy Research Center, RWTH Aachen University, Germany
[2]Department of Electrical/Electronic Engineering, Faculty of Engineering, University of Port Harcourt, PMB 5323, Port-Harcourt, Rivers Nigeria

Corresponding Author: Ameze Big-Alabo (Email: ameze.big-alabo@eonerc.rwth-aachen.de)
(Received: 16-October-2025; accepted: 24-February-2026; published: 30-June-2026)
http://dx.doi.org/10.55579/jaec.2026102.521

Abstract. *The present study reviews Deep Reinforcement Learning (DRL) algorithms as applied to Photovoltaic (PV) systems. A literature survey was conducted on various DRL techniques for Maximum Power Point Tracking (MPPT) and Partial Shading Conditions. The survey shows Deep Deterministic Policy Gradient (DDPG) to be the most implemented technique because of its fast convergence speed. Deep Q-Network (DQN) was considered to achieve faster response than DDPG. Twin Delayed Deep Deterministic Policy Gradient (TD3) was considered preferable, while Soft Actor-Critic (SAC), approach better eliminates power oscillations, under partially shaded conditions.*

The implementation of DRL-based MPPT for critical and effective learning requires defining the state variable, action variable and reward function of the PV module. It is therefore important to observe the voltage, current, irradiance, and temperature data that can allow for easy adaptation to changing environmental conditions. DRL requires higher computational effort compared to conventional methods due to its training phase. However, the trained models can operate with relatively low computational effort, thus making it a promising approach for real-time applications.

The literature survey also showed that the exploration–exploitation trade-off is a fundamental challenge in DRL-based MPPT control. Therefore, effective management of this trade-off, as well as bridging the gap between simulation and real-world hardware implementation, will enable DRL to become a practical solution for MPPT in PV systems.

Keywords: *DRL Algorithms, Exploration-Exploitation, MPPT techniques, Solar - PV.*

1. Introduction

Maximum Power Point Tracking (MPPT) algorithms are essential for maximizing the energy output of PV arrays that are continuously exposed to varying environmental conditions. Nonuniform irradiance and temperature variations affect PV modules by altering the PV system dynamics. This has a direct impact on the available power generated, thereby increasing the complexity of the MPPT control task. PV systems comprise a combination of solar cells connected in series and parallel to generate power output. The solar cell is basically represented by a single diode model [1], as shown in Figure 1. The governing equation of the PV cell

is stated in Equation (1), which describes the current (I)–voltage (V) characteristics of a PV cell [2].

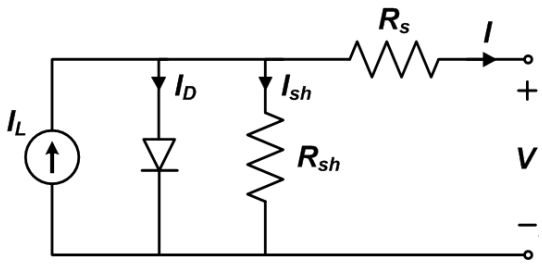

Figure 1: Circuit Model of a Single-diode solar cell [2].

$$I = I_l - I_s \left(\exp\left(\frac{V + IR_s}{mn_{se}V_T} \right) - 1 \right) - \frac{V + IR_s}{R_{sh}} \tag{1}$$

Where I_l, I_s, m, R_s, R_{sh}, n_{se}, and V_T are the light-induced current, the diode dark saturation current, the diode quality factor, the series resistance, the shunt resistance, the number of series-connected cells in the PV module, and the thermal voltage, respectively. The light-induced current, I_l, is affected by changes in the irradiance and temperature [3]. Furthermore, the dynamics of solar panels that capture the interdependencies between current, voltage, power, solar irradiation, and temperature have been studied in the literature [4]. The efficiency of the solar PV system depends upon its power output, which is a product of the panel's output voltage and current. The operating point at which the PV current and voltage are maximum is termed MPP, and this can be maintained via MPPT strategies [5, 6].

Moreover, the performance of the PV systems is directly affected by either partial or complete shading from the profile of the solar irradiation. The shaded portion of the PV modules acts as a resistance of loads consuming electrical power. Thus, the unshaded PV module parts cause the shaded part to operate in other reverse-bias conditions. For the reverse-bias conditions, the shaded portion dissipates the obtained power, causing an increase in the surface temperature, leading to hotspots. Experimental approaches to single-cell shading in high-efficiency monocrystalline silicon PV Passivated Emitter and Rear Cell (PERC) modules have been investigated [7].

Power electronic converters are becoming more relevant in recent times because of the easy integration of renewable energy sources into the electrical grid [8–11]. There are many useful applications of boost converters such as 'battery chargers for traction, energy storage systems [12, 13], photovoltaic voltage regulators [14,15], power factor correction [16–18], distributed generation systems [19], and microgrid applications [20,21]'. The boost converter is required for stepping up the DC voltage output obtained from solar panels from a relatively low and unstable voltage to a higher and more stable voltage. This helps to overcome the fluctuating output of PV panels due to variations in irradiance and temperature and, more importantly, is crucial for electronic inverters or DC loads that require a stable voltage to operate efficiently. Boost converters are commonly integrated with MPPT algorithms to allow dynamic adjustment of the MPP of the PV array under varying environmental conditions. Thus, the converter can maintain the PV panel at its optimal voltage and current so that energy harvesting can be maximized [22, 23].

Structure of the Study

This review study is structured into five sections. Section 1 introduces the concept of PV system dynamics and the concept of MPPT. Section 2 reviews conventional and advanced MPPT techniques, with particular emphasis on the challenges introduced by partial shading conditions. Section 3 surveys DRL algorithms applied to MPPT and partial shading mitigation, highlighting their operating principles and comparative performance. Section 4 provides a critical discussion of reward function design, state representation, and control architectures in DRL-based PV systems. It also outlines open challenges, practical considerations, and emerging research directions. Finally, a summary of key findings and insights is presented in Section 5.

2. MPPT Strategies and Partial Shading Effects: A Systematic Technical Review

This section reviews the evolution of MPPT strategies from conventional algorithms to advanced and intelligent methods, highlighting their suitability under Partial Shading Conditions (PSC). Perturb and Observe (P&O), Incremental Conductance (INC), and Constant Voltage (CV) [24–26] are conventional MPPT techniques that are simple and computationally efficient. However, under PSC, they are unable to differentiate between local maxima and the Global Maximum Power Point (GMPP), which results in incorrect tracking and significant energy losses. These limitations have motivated the development of more advanced MPPT methods such as metaheuristic optimization, artificial intelligence (AI) approaches, and a hybrid combination of these techniques.

Metaheuristic algorithms such as 'Particle Swarm Optimization (PSO) [27,28], Genetic Algorithms (GA) [29, 30], and Differential Evolution (DE) [31]' offer global search capability and are therefore well suited to PSC. Nevertheless, they struggle with rapidly changing environmental conditions and must be reinitialized to detect new GMPPs. This incurs high computational cost and slow convergence, making them less practical for real-time deployment.

AI techniques including fuzzy logic control [32, 33] and artificial neural networks (ANNs) [34] provide adaptability, robustness, and strong nonlinear modeling capability. These strengths make them effective for MPPT under PSC; however, their performance is dependent on the quality of training data and careful parameter tuning, which can limit generalization.

Hybrid MPPT strategies attempt to combine the strengths of conventional, metaheuristic, and AI-based approaches to balance tracking accuracy and convergence speed under PSC. Several such techniques have been widely reported in the literature [35–38].

Moreover, physical reconfiguration methods have been employed to restructure the interconnections of PV modules either manually or through automated schemes to reduce mismatch losses under PSC. This may involve rearranging modules to cluster shaded cells or using scanning algorithms to identify optimal configurations [39, 40].

When physical rewiring becomes computationally infeasible, metaheuristic-based electrical reconfiguration offers an alternative. Techniques such as neuro-fuzzy control with switch matrices [41], PSO [42, 43], War Strategy Optimization [44], and Grasshopper Optimization Algorithm (GOA) [45] can search for optimal array configurations, though they often introduce additional implementation complexity.

To address this complexity, a Rule-Based Adaptive Approach [46] has been proposed, where the array topology is adjusted after detecting irradiance thresholds. Although computationally light, its fixed rules limit adaptability to shading scenarios not explicitly considered during the design stage.

Reinforcement Learning (RL) provides a more flexible solution by learning optimal reconfiguration or control policies through interaction with the environment. Unlike rule-based or metaheuristic approaches, an RL agent can generalize to previously unseen shading patterns, explore the state space more comprehensively, and learn strategies that locate the global optimum without explicit programming for each scenario. This makes RL particularly suitable for highly dynamic and unpredictable shading conditions. However, these benefits come at the cost of increased computational complexity and the need for extensive training data.

A comprehensive review of the various MPPT algorithms and their effectiveness in mitigating PSC is presented in Table 1.

2.1. Concept of Reinforcement Learning

Reinforcement Learning is based on Markov Decision Process (MDP) [62], whereby the agent learns to make sequential decisions by interacting with an environment. RL framework for PV systems can be represented by the tuple

Table 1: Review of MPPT Control Algorithms in PV Applications.

Algorithms	Technique	Advantages	Research Gap
Conventional: Hill-Climbing (HC) [24], Incremental Conductance (INC) [25], Perturb & Observe (P&O) [26]	Model-based	Easy implementation; effective PV power tracking under static irradiation	Exhibit slower tracking, steady-state oscillations, and poor convergence. Not suitable under partial shading conditions.
Metaheuristic Algorithms: PSO [27,28], GA [29]	Model-based	Suitable for non-linearity analyses of the PV system, optimization of PV power under uniform and partial shading conditions	It cannot store optimal data and thus does not have learning ability where optimal data can be reused.
Q-learning [47, 48] and SARSA [49]	Model-free RL algorithm implemented in the MDP framework	Assumes discrete state and action spaces	In practice, Q-learning requires a quantized model of state and action space for non-linearities and noise in PV systems. Unsuitable for continuous action and state spaces.
DQN [50–53]	Deep neural networks approximate Q-value function	Applicable for continuous state spaces	Not suitable for continuous action spaces. Discretization of the action space may reduce control resolution and precision, leading to performance degradation. Moreover, the method tends to overestimate Q-values, particularly in noisy environments, which can result in suboptimal or unstable policies.
DDPG and TD3 [54, 55]	Model-free RL	Suitable for continuous state and action spaces	Hyperparameter sensitivity, brittle training, and slow convergence rate.
SAC [56–58]	Model-free reinforcement learning based on the maximum entropy framework	Suitable for continuous action spaces; ideal for PV systems requiring fine control, such as continuous duty-cycle adjustment of boost converters	Requires two Q-networks, an actor network, a target Q-network, and temperature tuning, making it computationally heavier and more complex. Also it is Hyperparameter Sensitive such that poor tuning of the learning rates, entropy temperature or Q-function update rates can lead to instability or slow learning.
PPO [59–61]	Model-free, on-policy reinforcement learning belonging to the policy-gradient family	Supports both discrete and continuous action spaces. Simpler than SAC, requiring only an actor network and a value function	Not suitable for applications requiring high control precision and resolution.

Table 2: Performance Metrics of DRL Algorithms.

Algorithm	Tracking Efficiency / Power Deviation	Convergence / Training Time	Stability & Oscillations
DQN	Faster response than DDPG [50]; Outperforms PSO & GWO [69]	Trains faster than DDPG [50]; faster than P&O and PSO [69]	Some oscillations; less stable at low irradiance
DDPG	<1% deviation from theoretical GMPP [54]; improved efficiency [55]	∼0.8 s convergence [55]	Good performance but hyperparameter sensitive [54]
TD3	Higher power output (up to +8.37%) vs. DDPG at high irradiance [70]	Slower than DQN; comparable to DDPG [70]	Most stable among DQN, DDPG, and TD3 [70]
SAC	Reduces oscillations compared to DDPG [56]	Slower than DQN [56]	More stable than DDPG [56]
PPO	Demonstrated in PV systems [59]; 96–98% efficiency reported [60, 61]	Moderate; slower than DQN but stable [60]	Stable but lower precision for fine MPPT control

(S, A, P, R, γ) where S, A, P, R, γ represent the state, action spaces, transition probability function, non-negative reward function and discount factor respectively. When the RL agent takes an action, it receives feedback in the form of rewards or penalties, which makes it learn the optimal policy that guides it for maximum cumulative expected rewards [63, 64]. The RL technique is considered a useful tool for MPPT in PV systems particularly for PSC. An example is the Q-learning-based Global Flexible Power Point Tracking (GFPPT) algorithm [65] which dynamically adapts to changes due to PSC, thereby improving the performance of the system. The RL agent explores the PV characteristics over time to improve convergence and robustness compared to conventional and metaheuristic approaches.

3. Survey of DRL Control Techniques for Partial Shading Conditions

Deep Reinforcement Learning (DRL) is an advancement of Reinforcement Learning (RL). It offers adaptive, model-free control under dynamic and complex environmental conditions. The agent senses its environmental state '$s \in S$' and takes action '$a \in A$' according to a policy π that is optimized through learning to maximize future rewards '$r \in R$'. In this case, a deep neural network is used to represent the policy π, which is referred to as a deep policy network. Thus, DRL is based on training deep neural networks to approximate the optimal policy π [66]. Various DRL algorithms include 'Q-learning, Deep Q-Network (DQN), Proximal Policy Optimization (PPO), Soft Actor-Critic (SAC), and Deep Deterministic Policy Gradient (DDPG)', as well as hybrid combinations of these methods. The implementation of DRL for MPPT control under partial shading conditions has been shown to achieve faster response times and better adaptability to dynamic and complex shading environments [50, 54, 67, 68]. Table 2 presents the performance metrics of various DRL techniques reported in the literature, in terms of tracking efficiency, output power improvement, convergence or training time, and stability.

The DDPG showed deviations of less than 1% from the theoretical MPPT [54]. A fuzzy-DDPG hybrid attained 95% tracking efficiency [71], while DQN achieved faster response than DDPG [50]. Other advantages include reduced oscillatory output power, which can be achieved using the SAC approach [56]. SAC exhibits

Table 3: Computational Complexity, Strengths and Limitations of DRL Algorithms.

Algorithm	Computational Complexity	Strengths	Limitations
DQN	Low–Moderate	Fast convergence; simple implementation	Requires action space discretization (reduced precision)
DDPG	Moderate–High	Works well in continuous state–action spaces	Training brittleness; slower than DQN
TD3	High	Robust to noise; avoids overestimation bias	Lower output than DQN at low irradiance
SAC	High (two critics + temperature tuning)	Smooth continuous duty-cycle control	Computationally heavier; complex tuning
PPO	Moderate	Easy to implement; supports discrete and continuous actions	Less precise than SAC/DDPG for high-resolution control

reduced steady-state power oscillations primarily due to its entropy-regularized objective, which promotes smooth stochastic policies and suppresses aggressive action updates near the MPP. It employs dual critic networks, which further mitigate overestimation bias and stabilize control actions. TD3 improves stability and can achieve higher output power of up to 8.37% compared to DDPG at high irradiance, but it produces lower output power than DQN at low irradiance [70]. Table 3 summarizes the computational complexity, strengths, and limitations of each DRL algorithm.

3.1. Comparisons: Experimental vs. Simulation

The review of DRL MPPT literatures applicable to PV systems reveals limited but informative evidence regarding experimental versus simulation performance differences. Literatures [72, 73] were able to give a direct experimental vs simulation comparison, with divergent findings. For example, the literature [72] found that DQN's superiority over P&O observed in simulation did not entirely replicate in real-world testing, while [73] reported that DDPG experimental improvements of 51.45% substantially exceeded simulation predictions of 10.45%. This evidence indicates that experimental-simulation gaps are algorithm-dependent and condition-specific rather than systematic. Thus, the relative advantage of DDPG over P&O, as presented in Table 4, was found to be greater in experimental conditions than that of simulations.

4. Formulation of DRL for PV Systems

This section describes the application of DRL to PV systems. The MPPT problem is formulated based on a Markov Decision Process (MDP) that enables the agent to learn optimal control policies through interaction with a dynamic PV environment, especially under PSC. The following terminologies are described as follows:

4.1. Markov Decision Process (MDP) Framework for PV Systems

1) Agent

This is the DRL algorithm that learns optimal decisions that maximizes reward, by determining the best action to take in each state to maximize long term rewards [74].

Table 4: Comparison of Experimental and Simulation Studies.

Study	Simulation Performance	Experimental Performance	Performance Gap
DQN [72]	Consistently superior to P&O	Consistently superior to P&O	Not entirely replicable in real-world testing
DDPG [73]	10.45% power improvement; settling time 24.54 times faster than P&O	51.45% power improvement; settling time 0.25 s vs. 4.26 s for P&O	Experimental results outperformed simulation predictions

2) Environment and States

Environment models the PV panels, DC-DC boost converter and load dynamics based on a Markov decision process (MDP) [62]. The MDP consists of a set of states S, a set of action A and a set of rewards R. In relation to PV systems, the set of states S observed by the agent for MPPT under steady state conditions are the instantaneous measurements of PV panel parameters of voltage, current and PV power. When considering partial shading conditions, the observable states are a combination of PV panel parameters with change in power, ΔP_{pv} and previous duty cycle of the boost converter. Optional observable states that can be considered include the irradiance profiles or temperature. These observations determine the state space that set up the learning environment.

3) Action and Reward

In DRL for PV converters, the reward is a scalar function of the state–action transition, typically designed to encourage maximum power extraction while respecting voltage, current, and stability limits. The RL agent interacts with the environment by adjusting the duty cycle or voltage reference of the MPPT boost converter in the PV system. In return, the environment provides a reward to the agent. This reward may be defined based on maximum power output, efficiency improvement, or constant voltage regulation. Common reward formulations used in previous works are discussed in detail in Section 3.1.

4) Episodes

Episodes consist of sequences of states, actions, and rewards that terminate at a terminal state, thereby reflecting the time required for the agent to converge to an optimal and stable policy. A lower number of episodes to convergence indicates faster learning as compared to a higher number of episodes. Thus, for practical deployment, having a low number of episodes to convergence is preferable. Moreover, this helps to show the computational effort required by the agent in reaching its target and hence the viability of the DRL technique. Longer episodes require more simulation time which suggests more computational effort and larger memory requirement.

5) Policy

This is the mapping of states to action; thereby defining the best action to take for a given state as represented in Equation (2):

$$\pi(s, a) = P_r(A_t = a \mid S_t = s) \tag{2}$$

where S_t and A_t denote the state and action random variables at time step t, respectively. Thus, the optimal policy represents the probability of taking action 'a' given the state 's', in order to maximize the cumulative future rewards. Policies used in DRL algorithms include discrete action policies such as the ε-Greedy policy [75–77], which is widely used in DQN, and continuous action policies such as Gaussian policies [78], which can be applied in DDPG, PPO, and SAC. Entropy-based exploration, which incorporates entropy into the reward objective to encourage broader exploration, can be applied in SAC [56]. Deterministic policies, applicable to

DDPG [79], are also commonly used. A combination of these policy techniques within DRL algorithms provides a balanced trade-off between exploration and exploitation strategy.

4.2. Exploration-Exploitation trade-off

The exploration–exploitation trade-off is a fundamental challenge in DRL-based MPPT. Several studies highlight the importance of balancing these two aspects to optimize system performance. The study in [56] on SAC-based MPPT control discusses the challenge of balancing exploration and exploitation to achieve efficient and accurate tracking under varying environmental conditions. An Exploration Decay Policy (EDP) has been designed to enhance this trade-off within DDPG [80]. This approach adaptively modulates exploration over time to improve continuous control performance and prevent suboptimal convergence. A chaotic K-best gravitational search algorithm that nonlinearly balances exploration and exploitation through chaotic dynamics has been shown to improve convergence and avoid local optima [81]. Furthermore, a framework enabling adaptive exploration strategies via Universal Value Function Approximators (UVFA) has been proposed [82]. This concept is further advanced by introducing Meta-SAC, which automatically tunes the entropy temperature parameter in the SAC algorithm [83]. Thus, the choice of exploration strategy significantly influences learning efficiency and control stability.

4.3. Types of reward formulations used for MPPT

1) Maximum Power Reward

Many MPPT studies define the reward as the instantaneous power or its normalized form, as described in Equation (3) [54].

Instantaneous Power

The reward r_t at each time step t, is defined by Equation (3):

$$r_t = \begin{cases} \dfrac{P_t}{c}, & \text{if } P_t > 0 \\ -1, & \text{if } P_t \leq 0 \end{cases} \tag{3}$$

where, $P_t(t) = V_t \times I_t$ instantaneous PV power, V_t and I_t are the PV array voltage and current at time step t respectively, and c is a positive scalar used to scale down the reward magnitude. This prevents excessively large rewards, improving numerical stability during training and critic learning. If the power becomes negative, it indicates unstable or non-physical operating points; thus, the agent receives a fixed negative reward to discourages actions that drive the PV system into infeasible regions.

2) Threshold Reward

This reward constitutes a threshold-based power reward whereby a predefined voltage window acts as a gating condition that helps reinforce operation near the GMPP and suppress local optima. The reward at each time step t, denoted as r_t is defined by Equation (4) as [84]:

$$r_t = \begin{cases} \left(\dfrac{P_t}{P_{\max}}\right)^4 & \text{if } |V_t - V_{\max}| \leq 2.5 \\ 0 & \text{otherwise} \end{cases} \tag{4}$$

Where P_{max} is the maximum power of the PV array at GMPP, V_{max} is voltage corresponding to GMPP and the $2.5V$ is the voltage tolerance window around V_{max}.

3) Hybrid Shaped Reward

A typical hybrid approach example is one that embeds Incremental Conductance (IC) physics inside a PPO framework. Although the reward is not explicitly formulated in closed form, the PPO agent in [85] is guided by Incremental Conductance (IC) principles. The underlying decision logic can be equivalently expressed as a binary reward that enforces the IC equilibrium condition as defined by Equations (5) – (7). From the IC theory in Equation (5):

$$\frac{dP}{dV} = I + V\frac{dI}{dV} \tag{5}$$

Such that at MPP,

$$\frac{dP}{dV} = 0 \;\Rightarrow\; \frac{dI}{dV} = -\frac{I}{V} \tag{6}$$

An IC-consistent reward can therefore be expressed as defined in Equation (7):

$$r_t = \begin{cases} +1, & \left|\frac{\Delta I_t}{\Delta V_t} + \frac{I_t}{V_t}\right| \le \epsilon \\ -1, & \text{otherwise} \end{cases} \tag{7}$$

Where ΔV_t and ΔI_t are the time step change in voltage and current respectively. This reward implicitly informs the agent whether it operates to the left or right of the MPP, emulating classical IC decision rules.

4) Fuzzified Reward Function

A fuzzified reward mechanism is introduced in [71] to enhance DRL-based MPPT by replacing a crisp scalar reward with a fuzzy inference-driven feedback signal. The motivation is to overcome the limitations of sparse or discontinuous rewards when dealing with the nonlinear and continuously varying power–voltage characteristics of PV systems.

The reward formulation is based on the slope of the PV power–voltage curve, defined by Equation (8) as:

$$e(t) = \frac{\Delta P_{pv}(t)}{\Delta V_{pv}(t)} \tag{8}$$

Where $\Delta P_{pv}(t)$ and $\Delta V_{pv}(t)$ are the instantaneous change in PV power and voltage respectively.

Equation (8) provides information about the relative position of the operating point with respect to the MPP. A positive slope indicates operation on the left of the MPP, while a negative slope indicates operation on the right. To capture the dynamic behavior of convergence, the temporal variation of this slope is also considered in Equation (9):

$$\Delta e(t) = e(t) - e(t-1) \tag{9}$$

These two signals, $e(t)$ and $\Delta e(t)$, are used as inputs to a Mamdani-type fuzzy inference system. Each input is fuzzified into linguistic variables such as "*Negative Big*, *Negative Small*, *Zero Error*, *Positive Small*, and *Positive Big*", thereby allowing the controller to interpret MPPT behavior at different abstraction levels.

The fuzzy rule base maps these linguistic inputs to a reward level categorized as *Low*, *Medium*, or *High*, such that large deviations from the MPP or incorrect movement directions result in low rewards. In contrast, stable operation near the MPP yields high rewards. The fuzzy output is then defuzzified to generate a continuous and bounded reward signal, which is supplied to the DRL agent.

By embedding expert knowledge into the reward design, the fuzzified reward provides smoother gradients for learning, accelerates convergence toward the MPP, and reduces steady-state oscillations.

5) Weighted Sum of Multiple Sub-Rewards

The total reward at each time step is defined as the weighted sum of multiple sub-rewards, each targeting a specific control objective of the MPPT problem, as shown in Equation (10) [71]:

$$R_t = r_{\text{mpp}} + r_1 + r_2 + r_3 + r_4 + r_5 \tag{10}$$

Each of these sub-rewards is described in Equations (11)–(17) as follows:

r_{mpp}: Deviation from GMPP

This term penalizes deviation of the measured power P from the maximum observed power value, $P_{\text{MPP,max}}$, during training as shown in Equation (11):

$$r_{\text{mpp}} = \begin{cases} -2\left|\dfrac{P - P_{\text{MPP,max}}}{P_{\text{MPP,max}}}\right|, & P_{\text{MPP,max}} > 0 \\ 0, & \text{otherwise} \end{cases} \tag{11}$$

This component drives the agent toward the GMPP rather than local maxima.

r_1: Power-Proportional Reward

A normalized reward proportional to the instantaneous extracted power P_t is defined as in Equation (12):

$$r_1 = \begin{cases} 0.1\dfrac{P_t}{P_{\text{MPP,max}}}, & P_{\text{MPP,max}} > 0 \\ 0, & \text{otherwise} \end{cases} \tag{12}$$

This ensures continuous positive reinforcement as power extraction improves.

r_2: Sign-Based Power Trend Reward

To encourage monotonic improvement in power, a sign-based term is included as shown in Equation (13):

$$r_2 = \text{sign}(\Delta P_t) \tag{13}$$

where

$$\Delta P_t = P_t - P_{t-1} \tag{14}$$

This component rewards actions that increase power and penalizes those that reduce it.

r_3: Voltage Constraint Penalty

To ensure safe operation, voltage excursions beyond the open-circuit voltage (V_{oc}) are penalized as defined in Equation (15):

$$r_3 = \begin{cases} 0, & 0 \le V_t \le V_{oc} \\ -2, & \text{otherwise} \end{cases} \tag{15}$$

This introduces a hard safety constraint into the learning process.

r_4: Voltage Change Penalty (Linear)

Sudden voltage changes are discouraged using a linear penalty as defined by Equation (16):

$$r_4 = -0.1\,|\Delta V_t| \tag{16}$$

This promotes smoother control actions and reduces oscillations around the MPP.

r_5: Voltage Change Penalty (Quadratic)

A stronger penalty is applied for large voltage variations as defined in Equation (17):

$$r_5 = -0.5\,|\Delta V_t|^2 \tag{17}$$

The quadratic term significantly penalizes aggressive control actions that may destabilize the system. The total reward function adopts a hybrid shaping strategy that integrates global optimality enforcement r_{mpp}, continuous power maximization r_1, sign-based power trend guidance r_2, threshold-based operational safety constraints r_3, and smoothness penalties r_4 and r_5.

Thus, the reward enables the agent to distinguish between local and global maxima and reduces steady-state oscillations, achieving fast and stable convergence to the GMPP under partial shading conditions.

5. Comparative Analysis of DDPG, SAC, and PPO Training Algorithms

5.1. SAC training algorithm

SAC is an off-policy DRL algorithm as described in Figure 2. The SAC is designed to optimize a stochastic policy by jointly maximizing the expected cumulative reward and the policy entropy, thereby promoting efficient exploration. The SAC framework adopts an actor–critic architecture consisting of a stochastic actor and twin critic networks. The actor outputs the parameters of a Gaussian action distribution, and actions are sampled using the reparameterization trick and passed through a hyperbolic tangent function to enforce bounded control actions. The overall SAC training flowchart is developed based on the literature [57]. The learning process begins with the observed state s, which is provided as input to the actor network. The actor produces the mean and standard deviation of the stochastic policy, from which a bounded action A is sampled and applied to the environment. The environment returns a reward R and the subsequent state s'. The resulting

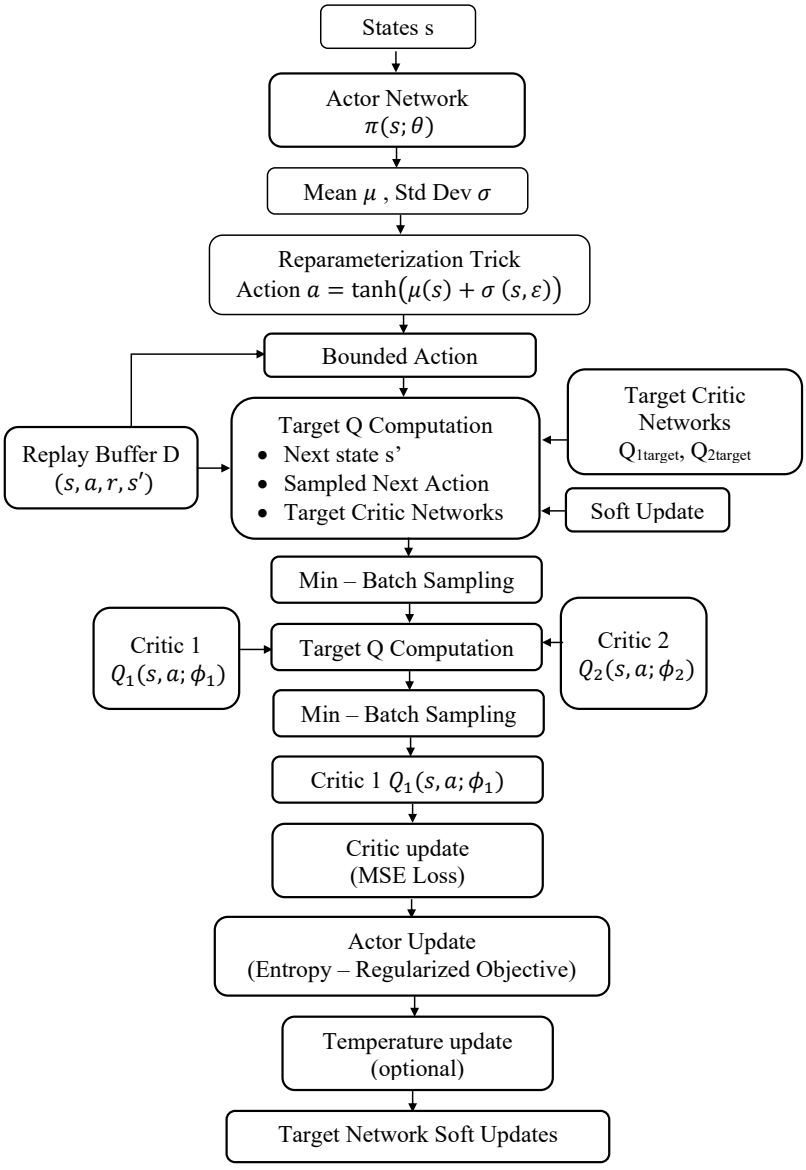

Figure 2: Training flow chart for SAC.

transition (s, a, r, s') is stored in the replay buffer D.

During training, mini-batches sampled from the replay buffer are used to update two critic networks, $Q_1(s, a; \phi_1)$ and $Q_2(s, a; \phi_2)$, which estimate state–action values to mitigate overestimation bias. Target Q-values are computed using the target critic networks, $Q_{1,\text{target}}$ and $Q_{2,\text{target}}$, along with the entropy-regularized expected future return. The critic parameters ϕ_1 and ϕ_2 are updated by minimizing the mean-squared Bellman error, while the actor parameters θ are updated to maximize the entropy-regularized objective. An optional temperature update adjusts the entropy coefficient to regulate policy stochasticity. The target critic networks are softly updated to ensure stable and consistent learning dynamics.

5.2. DDPG training algorithm

The DDPG is an off-policy actor–critic reinforcement learning algorithm as shown in Figure 3, that is designed for continuous control problems. It employs a deterministic actor network that directly maps the observed state to a continuous action and a critic network that evaluates the corresponding state–action value.

The training flow chart begins with the State 's' block, which represents the observation obtained from the environment at each interval of time. This state is provided to the Actor Network, which outputs a deterministic control action. To enable the exploration during the training process, the action is perturbed by the 'N' external Noise Model, resulting in the final action 'a' that is applied to the environment. The environment replies by yielding the 'r' Reward, and the 's'' Next State together with the current state and action, form an experience tuple that is stored in the Replay Buffer.

During the learning process, mini-batch sampling is used to randomly select the Replay Buffer experiences. The Target Actor Network processes the sampled next state, while the Target Critic Network evaluates the resulting action within the Target Q computation block to produce a stable target value for learning. The Critic Network is updated by minimising the Bellman errors within the computed target value and its current Q value estimate. Simultaneously, the Actor Update block revises the parameters of the actor by using the critic's gradients to maximise the evaluated state–action values. Finally, the Target Network Soft Updates block slowly corrects the parameters of the target critic networks and the target actor to follow the learned networks, guaranteeing steady and constant training dynamics. This iterative procedure persists until the procedure converges. The flow chart is based on existing literature [86].

5.3. PPO Training Algorithm

The PPO training flowchart in Figure 4 illustrates the interaction between the actor policy and critic value networks within an on-policy DRL framework. The actor network $\pi(s;\theta)$ maps the state 's' to a stochastic action distribution from which an action 'a' is sampled, while the critic network $V(s;\phi)$ estimates the state value. s, a, r and s' represents the current state, selected action, received reward, and next state s' respectively. These are experience tuples that are grouped into trajectories and stored in a trajectory buffer for learning. Based on these trajectories, returns and advantage estimates are computed, optionally using generalized advantage estimation (GAE), and advantage normalization is applied to improve training stability. The collected trajectory data are then reused for multiple optimization epochs, during which the actor network is updated using a clipped policy objective to constrain policy updates, and the critic network is trained by minimizing the value estimation loss. The iterative optimization process continues with newly collected trajectories, forming a stable on-policy learning loop that balances policy improvement and value function approximation. The flowchart is developed based on the literature [87].

5.4. Similarities in Training Characteristics

DRL algorithms such as DDPG, SAC and PPO share a common actor–critic training framework in which an actor network generates control actions, and a critic network evaluates their quality through value function approximation. In all three methods, learning is achieved through gradient-based optimization, where neural network parameters are updated by minimizing loss functions derived from Bellman equations or policy gradient objectives using mini-batches of experience. The agent interacts iteratively with the environment in an episodic manner by observing the current state, selecting an action, receiving a reward, and transitioning to the next state. Future rewards are weighted using a discount factor γ allowing the algorithms to optimize long-term performance rather than immediate gains. By leveraging deep neural networks as function approximators, these methods can effectively address high-dimensional, nonlinear control problems encountered in complex dynamical systems.

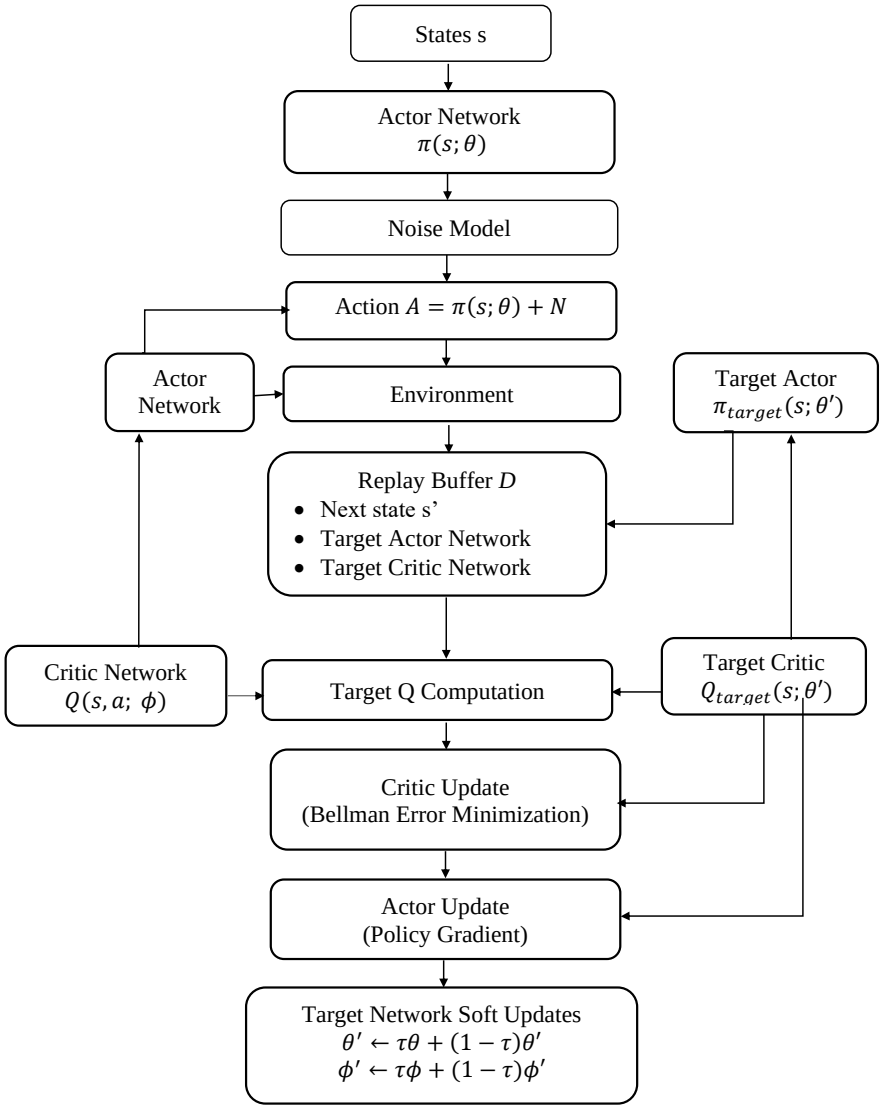

Figure 3: Training flow chart for DDPG.

5.5. Differences in Training Characteristics

In terms of differences, the DDPG, SAC, and PPO differ significantly in their training philosophies and stability mechanisms. DDPG employs a deterministic policy, where exploration is introduced through externally injected noise. It relies on a single critic network, which can result in overestimation bias and sensitivity to hyperparameter tuning, particularly in noisy environments. SAC adopts a stochastic policy formulation that explicitly maximizes both expected return and policy entropy, promoting robust exploration. The use of twin critics and a minimum Q-value selection further mitigate overestimation bias, while automatic tuning of the entropy temperature enables a balanced trade-off between exploration and exploitation, resulting in improved stability and sample efficiency. PPO follows a fundamentally different on-policy training approach, where data are collected in trajectories and discarded after each update. Training stability is enforced through a clipped

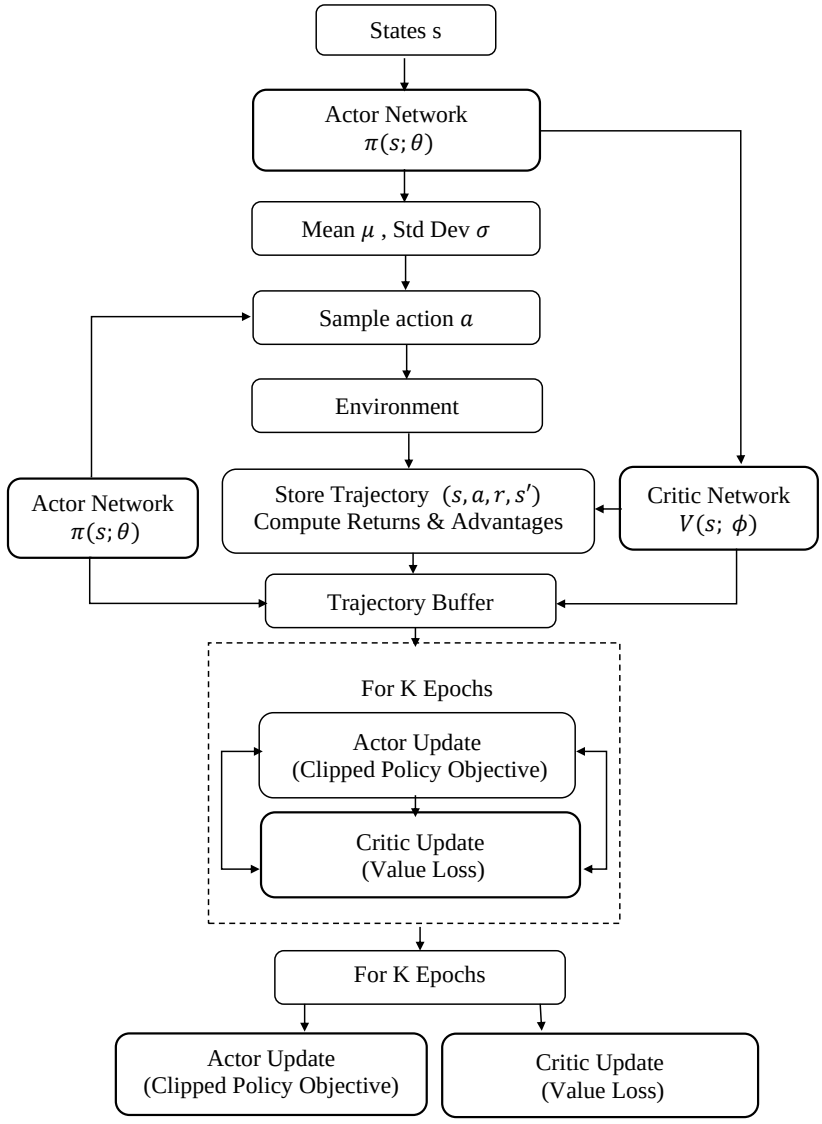

Figure 4: Training flow chart for PPO.

surrogate objective that constrains policy updates, and variance is reduced using advantage estimation techniques such as generalized advantage estimation (GAE). While PPO is generally easier to tune and exhibits strong training stability, it is typically less sample-efficient than off-policy methods such as DDPG and SAC.

5.6. Practical Implications for Control Applications

From a control perspective, DDPG is well suited for fast, deterministic continuous-control tasks but may suffer from stability issues due to its sensitivity to exploration noise and hyperparameter selection. SAC is generally preferred for complex nonlinear systems, such as power electronic converters and microgrids, as its entropy-regularized stochastic policy and

twin-critic structure provide improved robustness and exploration efficiency.

In contrast, PPO prioritizes training stability and implementation simplicity through constrained policy updates, making it attractive for applications where reliable convergence is more important than sample efficiency.

5.7. Challenges and Future works

The DRL approach is promising for MPPT and partial shading mitigation; however, several challenges remain before these methods can achieve widespread practical adoption. These challenges include scalability, real-time stability and safety, generalization and adaptability, computational and hardware constraints, as well as benchmarking and standardized evaluation.

The scalability of DRL to larger systems remains an open research direction, as most DRL techniques are implemented on a single PV array–converter system. Future PV farms will require multi-agent DRL frameworks capable of coordinating PV arrays and grid interfaces. Real-time stability under strict grid codes must also be ensured, since exploratory actions of DRL may lead to suboptimal or unstable control.

With respect to generalization and adaptability, DRL agents should be trained under varying irradiance and temperature profiles to enable adaptation to new climates and load conditions without retraining. Although the execution phase of DRL is lightweight, the training process remains computationally expensive. Therefore, future research should focus on resource-efficient DRL solutions deployable on low-cost embedded platforms and validated through Hardware-in-the-Loop (HIL) testing prior to large-scale deployment. Finally, standardized benchmarks and test protocols will enable more rigorous comparisons of DRL algorithms with respect to MPPT efficiency and response time.

6. Conclusion

This study provides a comprehensive review of DRL techniques for MPPT and partial shading mitigation in PV systems. The progression of MPPT strategies—from conventional model-based methods, through metaheuristic and artificial intelligence approaches, to modern DRL controllers—has been systematically reviewed. The analysis demonstrates that DRL algorithms exhibit superior adaptability and robustness in handling the nonlinear, stochastic, and time-varying characteristics of PV systems, particularly under partial shading conditions, where conventional and heuristic techniques often struggle to maintain global optimality.

A comparative evaluation of prominent DRL algorithms reveals distinct performance trade-offs. Value-based approaches such as DQN offer rapid convergence but are inherently limited by discrete action spaces, which restrict control resolution. Actor–critic methods such as DDPG enable continuous control but are sensitive to hyperparameter tuning and exploration noise, which may affect training stability. TD3 improves upon DDPG by mitigating overestimation bias through twin critics and delayed policy updates, while SAC further enhances robustness by incorporating entropy regularization and stochastic policies. The resulting effect is reduced steady-state power oscillations and improved exploration–exploitation balance.

PPO provides a stable and computationally efficient alternative through constrained on-policy updates and advantage estimation, but at the cost of lower sample efficiency and control precision compared to off-policy methods. These findings confirm that no single DRL algorithm is universally optimal, and algorithm selection should be guided by application-specific requirements such as control accuracy, convergence speed, and computational constraints.

Despite significant progress, several challenges remain, including effective management of the exploration–exploitation trade-off and limited experimental validation. The absence of standardized benchmarks and evaluation protocols further complicates fair performance comparison among DRL algorithms. Ultimately, effective

handling of the exploration–exploitation trade-off and bridging the gap between simulation and real-world hardware implementation will enable DRL to become a practical solution for MPPT in PV systems.

Acknowledgement

The author acknowledges the Alexander von Humboldt Foundation for sponsoring the research fellowship, of which this work forms a part.

Declaration of Competing Interest

The author declares that there is no competing interests whether financial or personal that could have appeared to influence the work reported in this paper.

References

[1] D. S. Chan, and J. C. Phang. Analytical methods for the extraction of solar-cell single- and double-diode model parameters from I–V characteristics. *IEEE Transactions on Electron Devices*, 34(2):286–293, 1987. DOI: 10.1109/T-ED.1987.22920.

[2] M. G. Villalva, J. R. Gazoli, and E. R. Filho. Comprehensive approach to modeling and simulation of photovoltaic arrays. *IEEE Transactions on Power Electronics*, 24(5):1198–1208, 2009. DOI: 10.1109/T-PEL.2009.2013862.

[3] P. Bharadwaj, K. N. Chaudhury, and V. John. Sequential optimization for PV panel parameter estimation. *IEEE Journal of Photovoltaics*, 6(5):1261–1268, 2016. DOI: 10.1109/JPHOTOV.2016.2574128.

[4] L. Guanghua, F. A. Siddiqui, M. M. Aman, S. H. H. Shah, A. Ali, A. M. Soomar, and S. Shaikh. Improved maximum power point tracking algorithms by using numerical analysis techniques for photovoltaic systems. *Results in Engineering*, 21:101740, 2024. DOI: 10.1016/j.rineng.2023.101740.

[5] P. K. Pathak, A. K. Yadav, and P. A. Alvi. A state-of-the-art review on shading mitigation techniques in solar photovoltaics via meta-heuristic approach. *Neural Computing and Applications*, 34(1):171–209, 2022. DOI: 10.1007/s00521-021-06586-3.

[6] M. H. Ali, M. Zakaria, and S. El-Tawab. A comprehensive study of recent maximum power point tracking techniques for photovoltaic systems,*Scientific Reports*, 15(1):14269, 2025. DOI: 10.1038/s41598-025-96247-5.

[7] N. Kumari, S. K. Singh, S. Kumar, and V. K. Jadoun. Performance analysis of partially shaded high-efficiency mono PERC/mono crystalline PV module under indoor and environmental conditions. *Scientific Reports*, 14(1):21587, 2024. DOI: 10.1038/s41598-024-72502-z.

[8] A. S. Valarmathy and M. Prabhakar. High gain interleaved boost-derived DC–DC converters—A review on structural variations, gain extension mechanisms and applications. *e-Prime*, 8:100618, 2024. DOI: 10.1016/j.prime.2024.100618.

[9] B. N. Reddy, B. S. Goud, C. N. Sai Kalyan, P. K. Balachandran, B. Aljafari, and K. Sangeetha. The design of 2S2L-based buck-boost converter with a wide conversion range. *International Transactions on Electrical Energy Systems*, 2023(1):4057091, 2023. DOI: 10.1155/2023/4057091.

[10] E. Bushra, K. Zeb, I. Ahmad, and M. Khalid. A comprehensive review on recent trends and future prospects of PWM techniques for harmonic suppression in renewable energy-based power converters. *Results in Engineering*, 22:102213, 2024. DOI: 10.1016/j.rineng.2024.102213.

[11] R. M. A. Velásquez. Thermal influence in the design of DC to AC converters due to climatic change for photovoltaic solar

plants. *Results in Engineering*, 23:102480, 2024. DOI: 10.1016/j.rineng.2024.102480.

[12] J. Wang, B. Wang, L. Zhang, J. Wang, N. I. Shchurov, and B. V. Malozyomov. Review of bidirectional DC–DC converter topologies for hybrid energy storage system of new energy vehicles. *Green Energy and Intelligent Transportation*, 1(2):100010, 2022. DOI: 10.1016/j.geits.2022.100010.

[13] L. Pirashanthiyah, H. N. Edirisinghe, W. M. P. De Silva, S. R. A. Bolonne, V. Logeeshan, and C. Wanigasekara. Design and analysis of a three-phase interleaved DC-DC boost converter with an energy storage system for a PV system. *Energies*, 17(1):250, 2024.

[14] S. Ramasamy, V. Sivasubramaniyam, G. Gatto, and A. Kumar. DC link voltage control based energy management strategy for standalone solar PV fed hybrid system. *AEIT Automotive*, 2023. DOI: 10.23919/AEITAUTOMOTIVE58986.2023.10217257.

[15] J. S. Oliver, P. W. David, P. K. Balachandran, and L. Mihet-Popa. Analysis of grid-interactive PV-fed BLDC pump using optimized MPPT in DC–DC converters. *Sustainability*, 14(12):7205, 2022. DOI: 10.3390/su14127205.

[16] N. Bhati and U. K. Kalla. Power factor corrected Y-cell modified boost converter fed battery charger for EV applications. *e-Prime*, 8:100550, 2024. DOI: 10.1016/j.prime.2024.100550.

[17] G. Sarowar, I. Ahmed, S. Rahman, A. Al Mamun, and K. M. Salim. Investigation of a power factor correction converter utilizing SEPIC topology with input current switching. *Results in Engineering*, 22:102271, 2024. DOI: 10.1016/j.rineng.2024.102271.

[18] A. K. Pallekonda and R. K. Ch. High gain interleaved PFC converter for torque ripple minimization in industrial PMBLDC motor based drives. *Results in Engineering*, 23:102413, 2024. DOI: 10.1016/j.rineng.2024.102413.

[19] A. S. Al-Khayyat, M. J. Hameed, and A. A. Ridha. Optimized power flow control for PV with hybrid energy storage system HESS in low voltage DC microgrid. *e-Prime*, 6:100388, 2023. DOI: 10.1016/j.prime.2023.100388.

[20] D. A. Palash, Z. Alam, T. K. Roy, and A. M. T. Oo. Designing robust hybrid controllers for enhancing output voltage regulation in CPL feed boost converters. *e-Prime*, 8:100532, 2024. DOI: 10.1016/j.prime.2024.100532.

[21] M. E. T. Souza Junior and L. C. G. Freitas. Power electronics for modern sustainable power systems: Distributed generation, microgrids and smart grids—A review. *Sustainability*, 14(6):3597, 2022. DOI: 10.3390/su14063597.

[22] K. Akter, S. M. A. Motakabber, A. Z. Alam, and S. H. B. Yusoff. Design and investigation of high power quality PV fed DC-DC boost converter. *e-Prime*, 9:100649, 2024. DOI: 10.1016/j.prime.2024.100649.

[23] S. E. Babaa, G. El Murr, F. Mohamed, and S. Pamuri. Overview of boost converters for photovoltaic systems. *Journal of Power and Energy Engineering*, 6(4):16–31, 2018. DOI: 10.4236/jpee.2018.64002.

[24] V. Jately, B. Azzopardi, J. Joshi, A. Sharma, and S. Arora. Experimental Analysis of hill-climbing MPPT algorithms under low irradiance levels. *Renewable and Sustainable Energy Reviews*, 150:111467, 2021. DOI: 10.1016/j.rser.2021.111467.

[25] G. Dhaouadi, O. Djamel, S. Youcef and C. Salah. Implementation of incremental conductance based MPPT algorithm for photovoltaic system. *ICPEA*, 2019. DOI: 10.1109/ICPEA1.2019.8911186.

[26] S. Thakran, J. Singh, R. Garg, and P. Mahajan. Implementation of P&O algorithm for MPPT in SPV system. *PEEIC*, 2018. DOI: 10.1109/PEEIC.2018.8665588.

[27] H. Li, D. Yang, W. Su, J. Lü, and X. Yu. An overall distribution particle swarm optimization MPPT algo-

rithm for photovoltaic system under partial shading. *IEEE Transactions on Industrial Electronics*, 66(1):265–275, 2018. DOI: 10.1109/TIE.2018.2829668.

[28] V. Loganathan and J. Swaroopan N. M. MPPT of solar PV systems using PSO memetic algorithm considering the effect of change in tilt angle. *Scientific Reports*, 15(1):1–17, 2025. DOI: 10.1038/s41598-025-92598-1.

[29] P. Kumar, G. Jain, and D. K. Palwalia. Genetic algorithm based maximum power tracking in solar power generation. *ICPACE*, 2015. DOI: 10.1109/ICPACE.2015.7274907.

[30] S. Daraban, D. Petreus, and C. Morel. A novel MPPT (maximum power point tracking) algorithm based on a modified genetic algorithm specialized on tracking the global maximum power point in photovoltaic systems affected by partial shading. *Energy*, 74:374–388, 2014. DOI: 10.1016/j.energy.2014.07.001.

[31] K. S. Tey, S. Mekhilef, M. Seyedmahmoudian, B. Horan, A. T. Oo, and A. Stojcevski. Improved differential evolution-based MPPT algorithm using SEPIC for PV systems under partial shading conditions. *IEEE Transactions on Industrial Informatics*, 14(10):4322–4333, 2018. DOI: 10.1109/TII.2018.2793210.

[32] K. Ullah, M. Ishaq, F. Tchier, H. Ahmad, and Z. Ahmad. Fuzzy-based maximum power point tracking (MPPT) control system for photovoltaic power generation system. *Results in Engineering*, 20:101466, 2023. DOI: 10.1016/j.rineng.2023.101466.

[33] Y. Zou, F. Yan, X. Wang, and J. Zhang. An efficient fuzzy logic control algorithm for photovoltaic maximum power point tracking under partial shading condition. *Journal of the Franklin Institute*, 357(6):3135–3149, 2020. DOI: 10.1016/j.jfranklin.2019.07.015.

[34] A. G. Olabi, Mohammad Ali Abdelkareem, Concetta Semeraro, Muaz Al Radi, Hegazy Rezk, Omar Muhaisen, Omar Adil Al-Isawi, and Enas Taha Sayed. Artificial neural networks applications in partially shaded PV systems. *Thermal Science and Engineering Progress*, 37:101612, 2023. DOI: 10.1016/j.tsep.2022.101612.

[35] S. Chtita, S. Motahhir, A. El Hammoumi, A. Chouder, A. S. Benyoucef, A. El Ghzizal, A. Derouich, M. Abouhawwash, and S. S. Askar. A novel hybrid GWO–PSO-based maximum power point tracking for photovoltaic systems operating under partial shading conditions. *Scientific Reports*, 12(1):10637, 2022. DOI: 10.1038/s41598-022-14733-6.

[36] B. Aljafari, P. K. Balachandran, D. Samithas, and S. B. Thanikanti. Solar photovoltaic converter controller using opposition-based reinforcement learning with butterfly optimization algorithm under partial shading conditions. *Environmental Science and Pollution Research*, 30(28):72617–72640, 2023. DOI: 10.1007/s11356-023-27261-1.

[37] M. Melhaoui, M. Rhiat, M. Oukili, I. Atmane, K. Hirech, B. Bossoufi, M. M. Almalki, T. A. H. Alghamdi, and M. Alenezi. Hybrid fuzzy logic approach for enhanced MPPT control in PV systems. *Scientific Reports*, 15(1):19235, 2025. DOI: 10.1038/s41598-025-03154-w.

[38] C. H. Basha, M. Palati, C. Dhanamjayulu, S. M. Muyeen, and P. Venkatareddy. A novel on design and implementation of hybrid MPPT controllers for solar PV systems under various partial shading conditions. *Scientific Reports*, 14(1):1609, 2024. DOI: 10.1038/s41598-023-49278-9.

[39] P. R. Satpathy and R. Sharma. Power and mismatch losses mitigation by a fixed electrical reconfiguration technique for partially shaded photovoltaic arrays. *Energy Conversion and Management*, 192:52–70, 2019. DOI: 10.1016/j.enconman.2019.04.039.

[40] K. Ş. Parlak. PV array reconfiguration method under partial shading conditions.

International Journal of Electrical Power & Energy Systems, 63:713–721, 2014. DOI: 10.1016/j.ijepes.2014.06.042.

[41] H. I. Solis-Cisneros, P. Y. Sevilla-Camacho, J. B. Robles-Ocampo, M. A. Zuñiga-Reyes, J. Rodríguez-Reséndiz, J. Muñiz-Soria, and C. A. Hernández-Gutiérrez. A dynamic reconfiguration method based on neuro-fuzzy control algorithm for partially shaded PV arrays. *Sustainable Energy Technologies and Assessments*, 52:102147, 2022. DOI: 10.1016/j.seta.2022.102147.

[42] T. S. Babu, J. P. Ram, T. Dragičević, M. Miyatake, F. Blaabjerg, and N. Rajasekar. Particle swarm optimization based solar PV array reconfiguration of the maximum power extraction under partial shading conditions. *IEEE Transactions on Sustainable Energy*, 9(1):74–85, 2018. DOI: 10.1109/TSTE.2017.2714905.

[43] F. Iraji, E. Farjah, and T. Ghanbari. Optimisation method to find the best switch set topology for reconfiguration of photovoltaic panels. *IET Renewable Power Generation*, 12(3):374–379, 2018. DOI: 10.1049/iet-rpg.2017.0505.

[44] A. G. Alharbi, A. Fathy, H. Rezk, M. A. Abdelkareem, and A. G. Olabi. An efficient war strategy optimization reconfiguration method for improving the PV array generated power. *Energy*, 283:129129, 2023. DOI: 10.1016/j.energy.2023.129129.

[45] A. Fathy. Recent meta-heuristic grasshopper optimization algorithm for optimal reconfiguration of partially shaded PV array. *Solar Energy*, 171:638–651, 2018. DOI: 10.1016/j.solener.2018.07.014.

[46] P. Verma, G. Singh, and A. Singh. Partial shading effect minimization in a PV array using shading behaviour based adaptive reconfiguration algorithm. *IC-SIT*, 2024. DOI: 10.1109/IC-SIT63503.2024.10862256.

[47] P. Kofinas, S. Doltsinis, A. I. Dounis, and G. A. Vouros. A reinforcement learning approach for MPPT control method of photovoltaic sources. *Renewable Energy*, 108:461–473, 2017. DOI: 10.1016/j.renene.2017.03.008

[48] M. A. Zeddini, M. Turki, and M. F. Mimoun. Optimization of PV energy conversion system using reinforcement learning algorithm. *STA*, pp. 249–254, 2020. DOI: 10.1109/STA50679.2020.9329331

[49] K. Bavarinos, A. Dounis, and P. Kofinas. Maximum power point tracking based on reinforcement learning using evolutionary optimization algorithms. *Energies*, 14(2):335, 2021. DOI: 10.3390/en14020335

[50] B. C. Phan, Y.-C. Lai, and C. E. Lin. A deep reinforcement learning-based MPPT control for PV systems under partial shading condition. *Sensors*, 20(11):3039, 2020. DOI: 10.3390/s20113039

[51] Y. Singh and N. Pal. Reinforcement learning with fuzzified reward approach for MPPT control of PV systems. *Sustainable Energy Technologies and Assessments*, 48:101665, 2021. DOI: 10.1016/j.seta.2021.101665

[52] G. Muriithi and S. Chowdhury. Deep Q-network application for optimal energy management in a grid-tied solar PV-Battery microgrid. *The Journal of Engineering*, 2022(4):422–441, 2022. DOI: 10.1049/tje2.12128

[53] K. Vora, S. Liu, and H. Dhulipati. Deep reinforcement learning based MPPT control for grid connected PV system. In *2024 IEEE 7th International Conference on Industrial Cyber-Physical Systems (ICPS)*, pp. 1–5, 2024. DOI: 10.1109/ICPS59941.2024.10639977

[54] L. Avila, M. De Paula, M. Trimboli, and I. Carlucho. Deep reinforcement learning approach for MPPT control of partially shaded PV systems in Smart Grids. *Applied Soft Computing*, 97(Part B):106711, 2020. DOI: 10.1016/j.asoc.2020.106711

[55] R. Guntupalli and M. Sudhakaran. Modeling & implementation of DRLA based partially shaded solar system integration with

3-ϕ conventional grid using constant current controller. *Heliyon*, 8(6):e09669, 2022. DOI: 10.1016/j.heliyon.2022.e09669

[56] S. E. Nwachukwu, K. A. Folly, and K. O. Awodele. A comparative study between soft actor-critic (SAC) and deep deterministic policy gradient (DDPG) algorithms for solar PV MPPT control under partial shading conditions. *IEEE Access*, 13:71738–71754, 2025. DOI: 10.1109/ACCESS.2025.3561807

[57] T. Haarnoja, A. Zhou, K. Hartikainen, G. Tucker, S. Ha, J. Tan, V. Kumar, H. Zhu, A. Gupta, and P. Abbeel. Soft actor-critic algorithms and applications. *arXiv preprint arXiv:1812.05905*, 2018. DOI: 10.48550/arXiv.1812.05905

[58] D. Xu, Y. Cui, J. Ye, S. W. Cha, A. Li, and C. Zheng. A soft actor-critic-based energy management strategy for electric vehicles with hybrid energy storage systems. *Journal of Power Sources*, 524:231099, 2022. DOI: 10.1016/j.jpowsour.2022.231099

[59] D. Ning, X. Chen, J. Chen, T. Meng, B. Xu, and H. Zhang. PPO-MixClip: An energy scheduling algorithm for low-carbon parks. *Energy Reports*, 12:4195–4207, 2024. DOI: 10.1016/j.egyr.2024.09.042

[60] A. U. Rehman, Z. Ullah, H. S. Qazi, H. M. Hasanien, and H. M. Khalid. Reinforcement learning-driven proximal policy optimization-based voltage control for PV and WT integrated power system. *Renewable Energy*, 227:120590, 2024. DOI: 10.1016/j.renene.2024.120590

[61] G. A. Ghazi, E. A. Al-Ammar, and H. M. Hasanien. Maximum Power Point Tracking Based Proximal Policy Optimization Algorithm for Grid-Connected Photovoltaic Systems. In *2024 5th International Conference on Communications, Information, Electronic and Energy Systems (CIEES)*, pp. 1-7, 2024. DOI: 10.1109/CIEES62939.2024.10811382.

[62] M. L. Puterman. Markov Decision Processes: Discrete Stochastic Dynamic Programming. *John Wiley & Sons*, edition 1st, chapter 2, pp. 17–32, 1994. DOI: 10.1002/9780470316887

[63] A. T. D. Perera and P. Kamalaruban. Applications of reinforcement learning in energy systems. *Renewable and Sustainable Energy Reviews*, 137:110618, 2021. DOI: 10.1016/j.rser.2020.110618

[64] D. Alfred, D. Czarkowski, and J. Teng. Reinforcement learning-based control of a power electronic converter. *Mathematics*, 12(5):671, 2024.

[65] E. Lioudakis and E. Koutroulis. Global flexible power point tracking based on reinforcement learning for partially shaded PV arrays. *IEEE Journal of Emerging and Selected Topics in Industrial Electronics*, 6(2):699–710, 2024. DOI: 10.1109/JESTIE.2024.3476695

[66] R. S. Sutton and A. G. Barto. Introduction to reinforcement learning. *Cambridge: MIT press*, 135:223-260, 1998.

[67] W. H. Yew, C. F. Chau, A. W. M. Zuhdi, W. S. W. Abdullah, W. K. Yew, and N. Amin. Investigating the performance of deep reinforcement learning-based MPPT algorithm under partial shading condition. In *2023 IEEE Regional Symposium on Micro and Nanoelectronics (RSM)*, pp. 9–12, 2023. DOI: 10.1109/RSM59033.2023.10326748

[68] T. Hoang and T. H. Le. Development of deep reinforcement learning for maximum power point tracking of photovoltaic systems. *Indonesian Journal of Electrical Engineering and Computer Science*, 33(2):707–714, 2024. DOI: 10.11591/ijeecs.v33.i2.pp707-714

[69] W. Pan, C. Cui, and H. Chen. Research on photovoltaic MPPT technique based on deep reinforcement learning under varying irradiance levels. In *2023 8th International Conference on Power and Renewable Energy (ICPRE)*, pp. 1794–1799, 2023. DOI: 10.1109/ICPRE59655.2023.10353820

[70] J. Panggabean, N. Sutisna, I. Syafalni, and T. Adiono. Comparison of MPPT based

on Deep Reinforcement Learning by DQN, DDPG and TD3. In *2023 Asia Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC)*, pp. 261-266, 2023 DOI: 10.1109/APSIPAASC58517.2023.10317341

[71] D. Ortiz-Munoz, D. Luviano-Cruz, L. A. Perez-Dominguez, A. G. Rodriguez-Ramirez, and F. Garcia-Luna. Hybrid fuzzy–DDPG approach for efficient MPPT in partially shaded photovoltaic panels. *Applied Sciences*, 15(9):4869, 2025. DOI: 10.3390/app15094869

[72] L. F. Giraldo, J. F. Gaviria, M. I. Torres, C. Alonso, and M. Bressan. Deep reinforcement learning using deep-Q-network for Global Maximum Power Point tracking: Design and experiments in real photovoltaic systems. *Heliyon*, 10(21):e37974, 2024. DOI: 10.1016/j.heliyon.2024.e37974

[73] E. Artetxe, J. Uralde, O. Barambones, I. Calvo, and I. Martin. Maximum power point tracker controller for solar photovoltaic based on reinforcement learning agent with a digital twin. *Mathematics*, 11(9):2166, 2023. DOI: 10.3390/math11092166

[74] Z. Ding, Y. Huang, H. Yuan, and H. Dong. Introduction to reinforcement learning. In Deep Reinforcement Learning: Fundamentals, Research and Applications, *Springer*, pp. 47–123, 2020. DOI: 10.1007/978-981-15-4095-0_2

[75] S. Zhang, H. Li, M. Wang, M. Liu, P.-Y. Chen, S. Lu, S. Liu, K. Murugesan, and S. Chaudhury. On the convergence and sample complexity analysis of deep Q-networks with ϵ-greedy exploration. *Advances in Neural Information Processing Systems*, 36:13064–13102, 2023. DOI: 10.48550/arXiv.2310.16173

[76] F. Ming, F. Gao, K. Liu, and C. Zhao. Cooperative modular reinforcement learning for large discrete action space problem. *Neural Networks*, 161:281–296, 2023. DOI: 10.1016/j.neunet.2023.01.046

[77] M. Ben-Akka, C. Tanougast, and C. Diou. Novel design of reward and epsilon-greedy decay strategy tailored for Q-learning in optimizing local mobile robot path planning. *Knowledge-Based Systems*, 324:113836, 2025. DOI: 10.1016/j.knosys.2025.113836

[78] P. Ladosz, L. Weng, M. Kim, and H. Oh. Exploration in deep reinforcement learning: A survey. *Information Fusion*, 85:1–22, 2022. DOI: 10.1016/j.inffus.2022.03.003

[79] E. H. Sumiea, S. J. Abdulkadir, H. S. Alhussian, S. M. Al-Selwi, A. Alqushaibi, M. G. Ragab, and S. M. Fati. Deep deterministic policy gradient algorithm: A systematic review. *Heliyon*, 10(9):e30697, 2024. DOI: 10.1016/j.heliyon.2024.e30697

[80] E. H. Sumiea, S. J. AbdulKadir, H. Alhussian, S. M. Al-Selwi, M. G. Ragab, and A. Alqushaibi. Exploration decay policy (edp) to enhanced exploration-exploitation trade-off in ddpg for continuous action control optimization. In *2023 IEEE 21st Student Conference on Research and Development (SCOReD)*, pp. 19–26, 2023. DOI: 10.1109/SCOReD60679.2023.10563810

[81] H. Mittal, R. Pal, A. Kulhari, and M. Saraswat, Mittal, H., Pal, R., Kulhari, A., and M., Saraswat. Chaotic kbest gravitational search algorithm (ckgsa). In *2016 ninth international conference on contemporary computing (IC3)*, pp. 1–6, 2016. DOI: 10.1109/IC3.2016.7880252

[82] A. P. Badia, P. Sprechmann, A. Vitvitskyi, D. Guo, B. Piot, S. Kapturowski, O. Tieleman, M. Arjovsky, A. Pritzel, and A. Bolt. Never give up: Learning directed exploration strategies. *arXiv preprint arXiv:2002.06038*, 2020. DOI: 10.48550/arXiv.2002.06038

[83] Y. Wang and T. Ni. Meta-SAC: Auto-tune the entropy temperature of soft actor-critic via metagradient. *arXiv preprint arXiv:2007.01932*, 2020. DOI: 10.48550/arXiv.2007.01932

[84] A. Wadehra, S. Bhalla, V. Jaiswal, K. P. S. Rana, and V. Kumar. A deep recurrent reinforcement learning approach for enhanced MPPT in PV systems. *Applied Soft Computing*, 162:111728, 2024. DOI: 10.1016/j.asoc.2024.111728

[85] S. F. Chevtchenko, E. J. Barbosa, M. C. Cavalcanti, G. M. S. Azevedo, and T. B. Ludermir. Combining PPO and incremental conductance for MPPT under dynamic shading and temperature. *Applied Soft Computing*, 131:109748, 2022. DOI: 10.1016/j.asoc.2022.109748

[86] T. P. Lillicrap, J. J. Hunt, A. Pritzel, N. Heess, T. Erez, Y. Tassa, D. Silver, and D. Wierstra. Continuous control with deep reinforcement learning. *US Patent 10,776,692*, 2020. DOI: 10.48550/arXiv.1509.02971

[87] J. Schulman, F. Wolski, P. Dhariwal, A. Radford, and O. Klimov. Proximal policy optimization algorithms. *arXiv preprint arXiv:1707.06347*, 2017. DOI: 10.48550/arXiv.1707.06347

About Authors

Big-Alabo Ameze is an Alexandra Humboldt research fellow at the E. ON Energy research Centre, Germany. She is also a Senior lecturer at the University of Port Harcourt, Nigeria. Her areas of specialization include Power Systems, Renewable Energy, and Thermoelectricity. She obtained her B.Eng and M.Eng (2008) at the University of Port Harcourt, Nigeria. Msc in Advanced Control and Systems Engineering, (2011) at the University of Manchester, UK and PhD in Electronic and Electrical Engr (2017) University of Glasgow, Scotland, UK. Her research interest include applications of deep reinforcement learning in renewable energy systems. She can be contacted at email: ameze.big-alabo@eonerc.rwth-aachen.de.