Deep Reinforcement Learning Based Optimal Perturbation for MPPT in Photovoltaics

Salman Sadiq Shuvo
Electrical Engineering
University of South Florida
Tampa, FL, USA
salmansadiq@usf.edu

Huruy Gebremariam
Electrical and Electronics Engineering
Sakarya University
Sakarya, Turkey
huruy.gebremariam1@ogr.sakarya.edu.tr

Yasin Yilmaz
Electrical Engineering
University of South Florida
Tampa, FL, USA
yasiny@usf.edu

***Abstract*—Methods to draw maximum power from Photovoltaic (PV) modules are an ongoing research topic. The so-called Maximum Power Point Tracking (MPPT) method aims to operate the PV module at its maximum power point (MPP) by matching the load resistance to its characteristic resistance, which changes with temperature and solar irradiance. Perturbation and Observation (P&O) is a popular method that lays the foundation for many advanced techniques. We propose a deep reinforcement learning (RL) based algorithm to determine the optimal perturbation size to reach the MPP. Our method utilizes an artificial neural network-based predictor to determine the MPP from temperature and solar irradiance measurements. The proposed technique provides an effective learning-based solution to the classical MPPT problem. The effectiveness of our model is demonstrated through comparative analysis with respect to the popular methods from the literature.**

***Index Terms*—Photovoltaics, MPPT, deep reinforcement learning, Markov decision process.**

I. Introduction

Solar Photovoltaic (PV) system is one of the fastest-growing renewable energy sources as it is inexhaustible and eco-friendly. A PV module can generate electrical energy directly from the sun without producing air pollution or greenhouse gases. However, PV energy generation systems have two problems, which are low efficiency of electrical power generation (52.94%), especially in times of low solar irradiation, and continuous fluctuations of produced electrical energy with changes in weather [1].

Maximum power point tracking (MPPT) methods are mainly used to extract maximum power from the PV module. Various MPPT algorithms have been discussed in the literature with their pros and cons [2]. These optimizing methodologies can be categorized into two major categories as indirect and direct techniques. Indirect techniques are mainly based on precomputed informational data about output power and voltage from a specific PV module for different environmental conditions. These data from experimentation are stored in terms of mathematical functions. Open Circuit Voltage and Short Circuit Current are the most commonly used indirect methods [3]. Another example of indirect methods is the lookup table method, which can extract the maximum power and voltage data from an experimental setup. A lookup table does not require complicated computations but rather provides a reference voltage for varying insolation [4]. The main advantage of these techniques is their simplicity. However, they cannot easily adapt to the external irradiance, and temperature changes as their core structures are based on estimated data.

On the other hand, direct techniques are based on instantaneous current and voltage measurements and thus are more accurate and have a faster response than the indirect methods. Also, depending on their structure and complexity, direct methods can further be categorized into different types. In this article, the conventional and heuristic types are discussed. Perturbation and observation (P&O), Incremental conductance (INC), and hill-climbing (HC) are common examples of conventional MPPT techniques, and these algorithms are based on fixed step size [5], [6]. The structure of these algorithms is simple with fewer implementation expenses, but they face problems of reaching the optimal value whenever solar irradiation becomes nonuniform during partially shaded conditions [6]. The major concerns of these techniques are high steady-state oscillation, slow convergence, and slow tracking speeds during fluctuations in temperature and solar irradiation.

In order to improve the transient and steady-state performance, meta-heuristic artificial intelligence (AI) based MPPT techniques have been proposed in the past, e.g., fuzzy logic (FL) and artificial neural network (ANN) controllers. According to [3] the ANN methodology has demonstrated better performance in comparison to FL-based MPPT controller for a wide range of solar irradiation, 1000 W/m^2-200 W/m^2, under rapidly varying irradiance. Especially in times of low irradiance, the efficiency difference is more significant. Also, [7] proposed AI-based technique known as Adaptive Neural-Fuzzy Interface System (ANFIS), which is an integration of ANN and FL. This MPPT technique has been designed in Matlab/Simulink and proved better convergence under varying solar irradiance. However, its main weakness is high cost and computational complexity due to integrating two smart methods. Moreover, in [8], hybrid techniques consisting of two MPPT techniques were shown to provide promising performance. For instance, the combination of ANN and P&O outperforms the hybrid method of particle swarm optimization (PSO) and P&O.

To further improve the efficiency, recently reinforcement learning (RL) has been used for MPPT control. For instance, [9] proposes a RL technique that takes temperature and

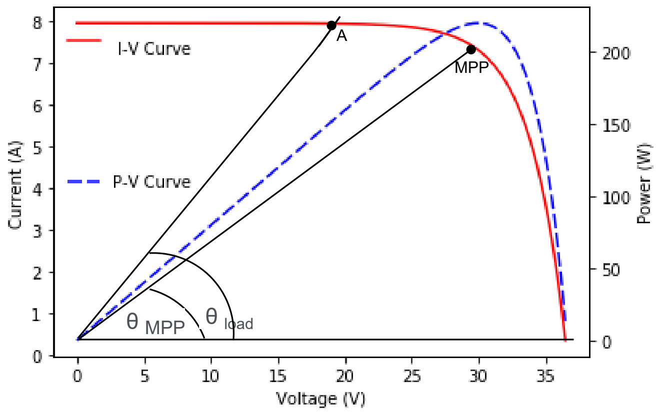

Fig. 1. I-V and P-V curve for the PV Module 1STH-220-P.

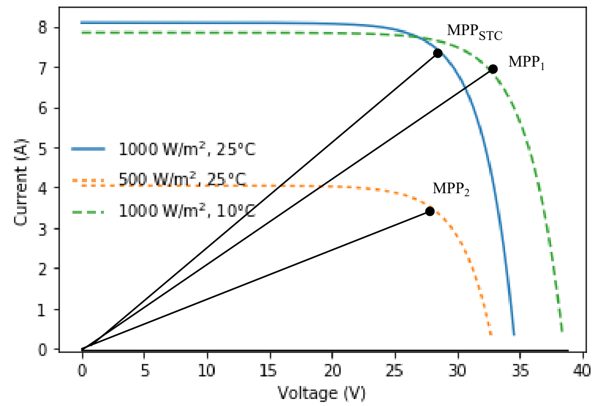

Fig. 2. MPP shifts with change in temperature and irradiance.

irradiance sensor data as state variables, apart from the voltage and current values. However, this method assumes stationary load resistance and thus becomes infeasible for practical application. This work proposes a novel deep RL technique to obtain fine-grained duty cycle adjustments for optimal MPPT considering variable load resistance. We support the deep RL method with maximum power point (MPP) and MPP resistance predictions using temperature and irradiance observations within an ANN regression model. Furthermore, our reward definition with predicted maximum power provides more stable feedback for the RL agent than the change in power used in the literature. The proposed deep RL method outperforms the state-of-the-art method in [9] in terms of both the transient and steady-state performance (Section IV). The main contributions of this work are:

- A novel deep RL technique which provides fine-grained duty cycle control under varying load resistance;
- An ANN based MPP and MPP resistance predictor for any temperature and solar irradiance;
- An efficient way based on the Newton-Raphson method to train the ANN regressor.

The remainder of the paper is organized as follows. Section II gives the necessary background. The proposed technique is explained in Section III, and the experimental results are presented in Section IV. Finally, concluding discussions and remarks are given in Section V and Section VI, respectively.

II. Background

In a PV system, the operation point determines the produced power, which can be defined as the product of the generated current I_{pv} and voltage V_{pv} at any moment in the current versus voltage (I-V) curve, for fixed environmental conditions. The solid red curve in Fig. 1 shows the I-V curve for a solar cell under the standard test conditions (STC) of $T_r = 25°\text{C}$ panel temperature and $G_r = 1000\text{W/m}^2$ irradiance. The same figure also shows the P-V curve with the dashed blue line where P=(V×I) represents power. The point on the operating curve where the generated power is maximum is called the Maximum Power Point (MPP), labeled in the figure. The voltage, current, and power for this point are 29.3 V, 7.47 A, and 218.87 W, respectively. The operating point of the PV system is the point where the load line intersects the I-V curve. And, the slope of the load line is equal to inverse of load resistance. The MPP creates the angle θ_{MPP} with the origin, whose slope can be written as

$$S_{MPP} = \frac{I_{MPP}}{V_{MPP}} = \frac{1}{R_{MPP}},$$

where R_{MPP} is the load resistance at MPP, also known as MPP characteristic resistance. The electrical load of resistance R_A connected to the PV source defines the operating point A, and the corresponding produced power. When the load resistance equals the MPP resistance, i.e., $R_A = R_{MPP}$, the operation point overlaps with the MPP, requiring no further tracking. However, the PV system does not generate maximum power when a load resistance with a different value is connected. Thus, for MPPT, it is essential to ensure that the load line passes through the MPP to deliver the maximum power to the output. A DC/DC converter between the PV source and load must be connected for this to happen. The DC/DC buck converter can move the operation point of the PV source from A to MPP by changing the slope S_A of the resistive load line as

$$D^2 \times S_A = \frac{D^2}{R_A} = \frac{1}{R_{MPP}} = S_{MPP},$$

where $D \in [0, 1]$ is the duty cycle of the converter. Notably, the DC/DC buck converter can move the operating point only to the right side. Hence, this converter can only reach MPP when $R_A < R_{MPP}$, limiting its operability for heavier loads. As most PV systems are connected to a variable resistive load, the MPPT task is essential to match the value of source resistance to the value corresponding to the MPP. Electrical resistive loads' variable or dynamic behavior always tries to shift the operating point from the MPP to the right or left depending on the power demand from the load side. Therefore, considering constant environmental conditions, the applied technique must continuously locate and track the MPP for efficient output power in the system.

Temperature and solar irradiance are the other two factors that affect the MPP. Fig. 2 shows that the I-V curve changes when the environmental condition changes from STC (solid

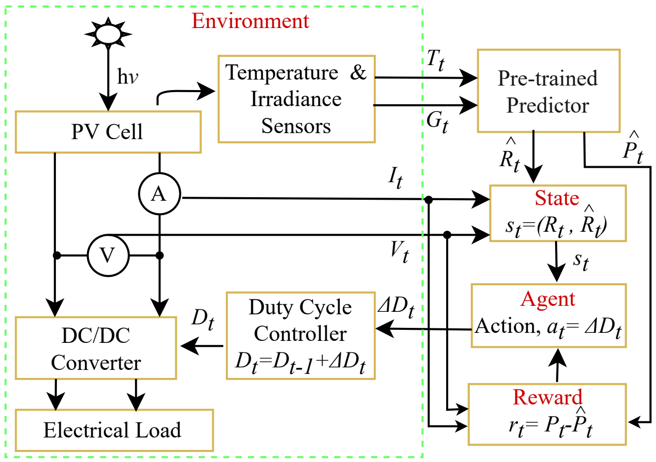

Fig. 3. Proposed System model. RL components are marked in red.

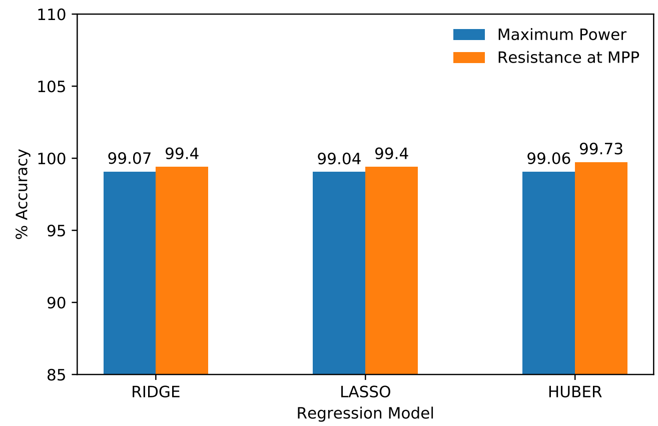

Fig. 4. Average accuracy for different regressor models for predicting P_{max} and R_{MPP} from temperature and irradiance values.

curve). In fact, for a particular PV cell, there is a distinct I-V curve for each temperature and irradiance pair. Along with the I-V curve, the MPP also changes. Fig. 2 shows that when the temperature or irradiance changes to 10 °C or 500 W/m^2, MPP changes to MPP$_1$ and MPP$_2$, respectively. To summarize, MPP for a particular PV is a function of temperature and irradiance. The environmental conditions differ continuously, and the I-V curve of the PV source changes consequently. So, the change in environment makes the MPPT task more essential and challenging. For brevity, we collectively call the load resistance, temperature, and irradiance as the *PV dynamics* hereafter.

III. Proposed Technique

The proposed technique consists of an ANN regressor and a RL agent which controls the duty cycle of the DC/DC converter. In this work, we use the advantage actor-critic (A2C) deep RL structure, which can accommodate arbitrarily fine-grained action space [10]. As shown in Fig. 3, data from temperature T_t and irradiance G_t sensors, connected to the PV module, are fed into an ANN which predicts the MPP (P_{max}) and the MPP resistance (R_{MPP}) as a state variable $\hat{R}_t$ for the RL agent. The RL agent also has access to an online voltmeter and ammeter that provides voltage V_t and current I_t measurements. The agent decides the required change in duty cycle, ΔD_t, based on the current load resistance $R_t = V_t/I_t$ and target resistance $\hat{R}_t$. The agent's selected duty cycle change, ΔD_t, is implemented by the Duty Cycle Controller. The controller updates the duty cycle $D_t = D_{t-1} + \Delta D_t$ and achieves the target duty cycle through an Integrator and Pulse Width Modulator (PWM) to set the DC/DC buck converter. Finally, the action taken is evaluated to inform and direct the next action by subtracting the predicted MPP $\hat{P}_t$ from the actual power P_t.

A. ANN Predictor

If $R_t = R_{MPP}$, then V_t, I_t and power $P_t = P_{max}$, all becomes deterministic through a complex relationship. In order to attain unique R_{MPP} and P_{max} for a particular environmental condition, we train fully connected ANNs (i.e., multilayer perceptrons) using the temperature T_t and irradiance G_t input to predict the R_{MPP} and P_{max}. We consider several regression models (Ridge, Lasso, and Huber) with two fully connected layers and select Huber regressor, which gives the maximum accuracy, for our experiments. Notably, all of the regressors provide around 99% accuracy, as shown in Fig. 4. The real-time predictor continuously provides prediction $\hat{R}_t$ for R_{MPP}, which is in turn used as input for the deep RL state. Similarly, the predictor provides prediction $\hat{P}_t$ for maximum power P_{max} that we use to calculate the reward of the RL agent. Besides, we use $\hat{P}_t$ to estimate the performance of the algorithms.

To train the predictor, the voltage and current values are generated from solar irradiance G_t at temperature T_t using the equations [11]

$$\begin{aligned} I_t &= I_{sc}\frac{G_t}{G_r}(1 + n_{iscT}(T_t - T_r)) - a_1 e^{b_1 V_t}, \\ V_t &= I_t \times R_t, \\ b_1 &= \frac{b_{STC}}{1 + n_{vocT}(T_t - T_s)}, \end{aligned}$$

where a_1, b_{STC}, I_{sc}, n_{vocT}, and n_{iscT} are constant values for a particular PV module. These parameters for our experimental PV module 1STH-220-P are given in Table I. $T_r = 25$°C and $G_r = 1000$W/m^2 are the reference temperature and irradiance values under STC. P_{max} and R_{MPP} are found by iteratively solving the above interlocked equations using the Newton-Raphson method. Using those equations and iterative optimization we avoid the time and computations needed for the Simulink simulations commonly used in the literature.

B. MDP Model

To formulate the problem for the RL agent we develop a Markov Decision Process (MDP) model, which is based on the Markov Property: the future state is dependent only on the current state and action taken by the agent. The physical part within the green dashed box in Fig. 3 constitues the Environment of the MDP model. The MPPT controller is the MDP agent that takes action a_t about change in duty cycle ΔD of the DC/DC Buck converter.

1) State, s_t: The agent collects the MPP resistance prediction $\hat{R}_t$ for the current environment condition from the predictor and V_t, I_t from the voltmeter and ammeter. We calculate the load resistance $R_t = V_t/I_t$ and define the two-input MDP state

$$s_t = (R_t, \hat{R}_t).$$

2) Action, a_t: The MDP agent's action a_t is selecting the duty cycle change ΔD_t of the DC/DC Buck converter. We consider a fine-grained action space for A2C and choose the duty cycle change $\Delta D_t \in \{-0.05, -0.04, ..., 0.04, 0.05\}$, i.e., a_t takes a value from these 11 actions . Ideally, the optimal duty cycle change ΔD_t satisfies

$$\frac{\Delta D_t^2}{R_t} = \frac{1}{\hat{R}_t}.$$

Since, in practice, R_t changes continuously, a sequential data-driven RL controller is an ideal fit for determining ΔD_t. The agent's selected duty cycle change, ΔD_t, is implemented by the Duty Cycle Controller. This controller includes an Integrator Pulse Width Modulator (PWM) to achieve the target duty cycle $D_t = D_{t-1} + \Delta D_t$.

3) Reward, r_t: We define reward as the difference between output power and the predicted maximum power of the ANN regressor

$$r_t = P_t - \hat{P}_t \tag{1}$$

The agent tries to maximize the reward, i.e., maximize the power output. Our reward selection provides the RL agent a stable target to reach, instead of the floating incremental power $\Delta P_t = P_t - P_{t-1}$ used in [9]. Since the maximum reward an agent can achieve is zero, once the MPP is reached, changing duty cycle will incur negative reward and the agent is expected to select making no change in the duty cycle. i.e. $\Delta D = 0$.

4) State Transition: If the PV dynamics do not change, the agent's action sets the state parameters R_t and $\hat{R}_t$ deterministically. After taking the action, the agent takes the voltmeter and ammeter reading V_{t+1} and I_{t+1} and calculates R_{t+1} to determine the next state for the RL agent. This completes one complete step for the RL decision-making. If the PV dynamics change, the state inputs will be different from the estimated next state. However, we assume the next state for MDP estimation to be free of such change due to instant impact of duty cycle change.

C. Solution Approach

Our RL agent aims to maximize the discounted total reward in T time steps,

$$r_T = \sum_{t=0}^{T} \gamma^t r_t, \tag{2}$$

Algorithm 1 A2C algorithm for MPPT.

Input: discount factor γ, learning rate, and number of episodes E
Input: Irradiance $\{G_T\}$, temperature $\{T_t\}$, and resistance $\{R_t\}$
Initialize: Actor network with random weights and critic network with random weights
for episode $= 1, 2, ..., E$ **do**
  for t$= 1, 2, ..., T$ **do**
    ANN predictor predicts $\hat{R}_t$, and $\hat{P}_t$.
    Select action a_t for state $s_t = (R_t, \hat{R}_t)$ using actor network.
    Execute action a_t and observe reward r_t from Eq. (1).
    Store transitions (s_t, a_t, r_t, s_{t+1}).
    Update actor network via advantage function.
    Update critic network through back propagation.
  end for
end for

Parameter	Definition	Value
V_{ocr}	Open circuit voltage	36.6 V
I_{scr}	Short circuit current	7.97 A
V_{mppr}	MPP voltage	29.3 V
I_{mppr}	MPP current	7.47 A
n_{iscT}	Short Circuit Current Temp Coefficient	0.102 (% /°C)
n_{vocT}	Open Circuit voltage Temp Coefficient	-0.361 (% /°C)
b_{STC}	$\frac{\log_e(1-\frac{I_{mppr}}{I_{scr}})}{V_{mppr}-V_{ocr}}$	0.37929
a_1	$I_{scr}e^{-b_{STC}V_{oc}}$	7.46×10^{-6}

Table I: PV Module 1STH-220-P details for Standard Test Conditions (STC).

where $\gamma \in (0, 1)$ is the discount factor for future cost. There are two popular approaches to find the optimal policy $\{a_t\}$, value-based methods (e.g., deep Q-learning) and policy-based methods (e.g., policy gradient). The Advantage Actor-Critic (A2C) is it's popular deep RL algorithm for continuous state environments [10], hence we consider A2C in this work. A2C is a hybrid deep RL method which consists of a policy-based actor network and value-based critic network. A pseudo code for the A2C algorithm is given in Algorithm 1.

IV. Results

A. Experimental Setup

In our experiments, we use the PV Module 1STH-220-P, whose operation details are provided in Table I. All the experiments are performed in Python 3.6.8 version. Fig. 5 shows the convergence of our A2C deep RL algorithm for varying irradiance. The y-axis represents the episodic output energy difference with respect to the ideal case (if the PV always operates at MPP). The smoothed reward is the running mean of the last 10 episodes of raw (actual) rewards. The algorithm converges within 4000 episodes and minimizes this energy difference to 0.36 kJ, where the total ideal output is 32.7 kJ for the episode duration (200 s). This minimal 1.1% loss of energy happens during the irradiance change time step, which is impossible to nullify.

B. Benchmark Policies

We compare our method with the following policies.

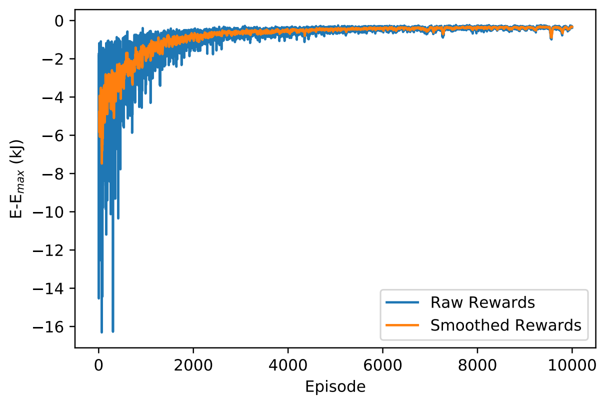

Fig. 5. Convergence of our A2C deep RL method under varying irradiance.

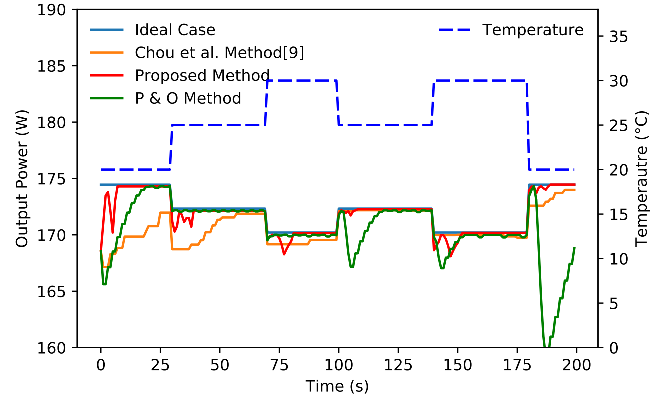

Fig. 7. Power output for different methods for varying temperature.

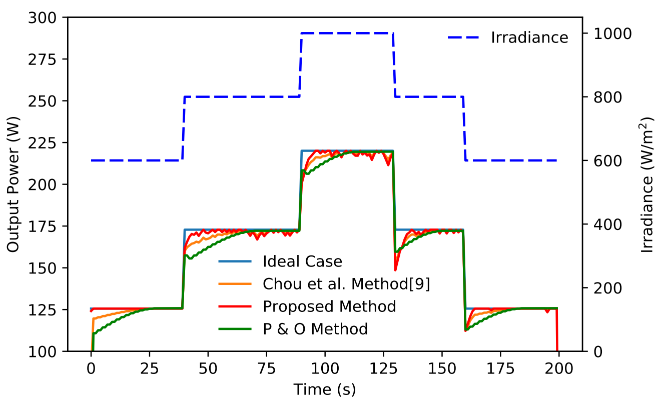

Fig. 6. Power output for different methods for varying irradiance.

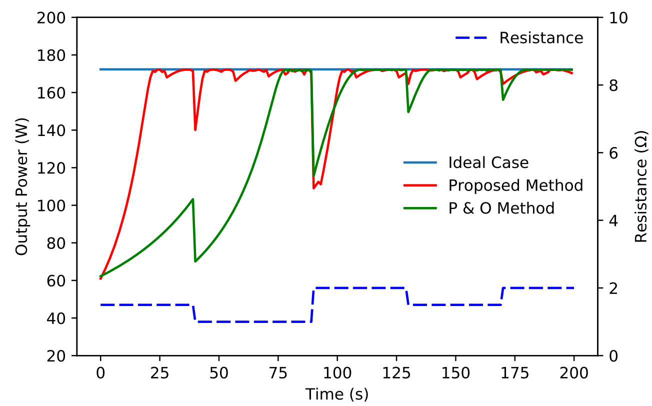

Fig. 8. Power output for different methods for varying resistance.

1) Perturb and Observe (P&O): We use the popular P&O method [6] as our baseline policy. We determine 0.01 to be a suitable step size for perturbation through a grid search.

2) RL-based approach: Chou et al. [9] propose a deep Reinforcement Learning (RL) based MPPT. Their method uses temperature, irradiance, and duty cycle of the DC/DC converter as the RL state, so they require pretty much the same setup as ours except the ANN predictor. Also, the reward used in [9] is the change in power, $\Delta P_t = P_t - P_{t-1}$, which provides a less stable (i.e., more fluctuating) feedback than our prediction-based reward $R_t = P_t - \hat{P}_t$. As the hardware requirement is quite similar, this method provides a fair comparison for our method.

C. Performance Analysis

We aim to test our method for different environments. Hence, we provide three sets of case simulations where we examine our method by changing either irradiance, temperature, or load resistance. The experiment duration is 200 time steps (seconds) for each analysis.

1) Varying Irradiance (G): We keep the temperature (25 °C) and the load resistance (5 Ω) stationary for this setup. The right y-axis in Fig. 6 represents the irradiance value that changes between 600, 800, and 1000 W/m^2, shown by the dashed line. The left y-axis shows the output power for different methods. The solid blue line represents the ideal output power that all the methods try to reach. P&O method is the slowest to reach, and our proposed method is the fastest. Chou et al.'s method [9] lie in between.

2) Varying Temperature (T): We set the irradiance at 800 W/m^2 and the load resistance (5 Ω) stationary for this setup. The right y-axis in Fig. 7 represents the temperature value that changes between 20, 25, and 30 °C, shown by the dashed line. The left y-axis represents the output power, and the solid blue line shows the ideal output power. Our method performs significantly better than the other methods.

3) Varying Load Resistance (R): The temperature and irradiance are fixed at 25 °C and 800 W/m^2 respectively for this case. But the load resitance changes among 1, 1.5 and 2 Ω as shown in Fig. 8. We don't include Chou et al. method [9] here as their model does not consider variable load. The maximum power remains stable at 172.8 W as it is free of load variability. The P&O method cannot reach the MPP fast enough due to small step size. We also experimented with bigger step sizes, which provided worse results and unstable output power. Our method uses its variable step size to provide the optimal solution. Clearly, MPPT for variable load is a more challenging task as it shifts the operation point

	Time to reach MPP for each change in operating condition (s)			Energy Output (kJ)			
Case	P&O	Chou et al.	Proposed	Ideal	P&O	Chou et al.	Proposed
Variable G	25, 22, 19, 18, 19	18, 14, 13, 13, 15	2, 6, 6, 6, 4	32.68	31.78	32.2	32.34
Variable T	19, 2, 3, 13, f/r*	f/r*, 22, 13, 2, 13	8, 6, 3, 5, 4	34.52	34.08	34.11	34.27
Variable R	25, 22, 19, 18, 19	n/a**	2, 6, 6, 6, 4	34.56	27.76	n/a**	31.84

* Fails to reach the MPP, ** Not applicable

Table II: Summary of performances under different cases considered in Figs. 6–8. The five numbers in each cell represent the performance under five time intervals in each case.

further from the MPP.

Table II shows the summary of the performance for the methods for different cases. Our proposed deep RL method is the fastest to track the MPP and maximizes the power output for each case. All the methods does well to maximize the output for variable irradiance and temperature; however, our method outperforms the others to be the closest to ideal case. The benefit of our method is more evident in the variable resistance case, where it outputs 13 % more energy than the P&O method. The time to reach the MPP after every change in PV dynamics is also provided in Table II, which is consistent with the output energy results.

V. Discussions

The MPPT task aims to reach MPP by shifting the load resistance towards the MPP resistance through duty cycle change. We define our MDP state as estimated MPP resistance and current load resistance, which has enough information to change the duty cycle. This effective breakdown of the problem helps us keep the RL state small and to the point, which is the underlying reason for the success of this model. This model is suitable for large-scale PV units where temperature and irradiance from multiple sensors may keep it apart from unnecessary noise from those sensors. Furthermore, periodical (yearly) calibration of the ANN predictor may compensate for degradation and corresponding changes in the I-V curve of the PV module over long-term usage. Deep RL algorithms with continuous action may further benefit this approach; however, the action range is a matter of deliberation as a significant change in the duty cycle may complicate the action of the DC/DC converter. The real-life implementation of this simulation-based method may provide further insights into this technique. Our discrete action setup has small granularity and a suitable operating range for the DC/DC converter. Our method addresses the major problem of optimal perturbation size of the P&O method by providing flexible duty cycle change based on the state of the MDP model.

VI. Conclusion

This work aims to provide a state-of-the-art solution to the MPPT task for photovoltaics by modeling a deep RL-based technique. We integrated an ANN-based pre-trained predictor into the deep RL model that predicts power and resistance at MPP for a given irradiance and temperature. These two parameters help to shape the state and reward of the RL model. This process breaks down the task for the deep RL-based algorithm, resulting in superior performance than the existing P&O and a recent deep RL-based method [9]. Our method is robust and can be used for any PV module by training the predictor with the module's I-V data.

References

[1] R. Faranda and S. Leva, "Energy comparison of mppt techniques for pv systems," *WSEAS transactions on power systems*, vol. 3, no. 6, pp. 446–455, 2008.

[2] X. Li, H. Wen, Y. Hu, Y. Du, and Y. Yang, "A comparative study on photovoltaic mppt algorithms under en50530 dynamic test procedure," *IEEE Transactions on Power Electronics*, vol. 36, no. 4, pp. 4153–4168, 2020.

[3] T. Anuradha, V. Senthil Kumar, and P. Senthil Kumar, "A comparison of existing mppt techniques for a pv system with interleaved converter," *Australian Journal of Basic and Applied Sciences*, vol. 10, no. 5, pp. 69–75, 2016.

[4] S. Malathy and R. Ramaprabha, "Maximum power point tracking based on look up table approach," in *Advanced Materials Research*, vol. 768. Trans Tech Publ, 2013, pp. 124–130.

[5] H. Abouadane, A. Fakkar, D. Sera, A. Lashab, S. Spataru, and T. Kerekes, "Multiple-power-sample based p&o mppt for fast-changing irradiance conditions for a simple implementation," *IEEE Journal of Photovoltaics*, vol. 10, no. 5, pp. 1481–1488, 2020.

[6] N. Femia, G. Petrone, G. Spagnuolo, and M. Vitelli, "Optimization of perturb and observe maximum power point tracking method," *IEEE transactions on power electronics*, vol. 20, no. 4, pp. 963–973, 2005.

[7] M. R. Javed, A. Waleed, U. S. Virk, and S. Z. ul Hassan, "Comparison of the adaptive neural-fuzzy interface system (anfis) based solar maximum power point tracking (mppt) with other solar mppt methods," in *2020 IEEE 23rd international multitopic conference (INMIC)*. IEEE, 2020, pp. 1–5.

[8] M. T. Mitsuya and A. A. de Moura Meneses, "Efficiency of hybrid mppt techniques based on ann and pso for photovoltaic systems under partially shading conditions," *Am. J. Eng. Appl. Sci*, vol. 12, pp. 460–471, 2019.

[9] 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, 2019.

[10] P.-H. Su, P. Budzianowski, S. Ultes, M. Gasic, and S. Young, "Sample-efficient actor-critic reinforcement learning with supervised data for dialogue management," *arXiv preprint arXiv:1707.00130*, 2017.

[11] P. Kofinas, S. Doltsinis, A. Dounis, and G. Vouros, "A reinforcement learning approach for mppt control method of photovoltaic sources," *Renewable Energy*, vol. 108, pp. 461–473, 2017.