International Journal on Recent and Innovation Trends in Computing and Communication
ISSN: 2321-8169 Volume: 11 Issue: 9s
DOI: https://doi.org/10.17762/ijritcc.v11i9s.7396
***Article Received**: 02 May 2023 **Revised:** 28 June 2023 **Accepted:** 20 July 2023*

Deep Reinforcement Learning DDPG Algorithm with AM based Transferable EMS for FCHEVs

Yogesh Wankhede[1], Sheetal Rana[2], Dr. Faruk Kazi[3]
[1]Department of Electronics Engineering
Veermata Jijabai Technological Institute
Mumbai, India
e-mail: yewankhede_p19@el.vjti.ac.in
[2]Department of Electronics Engineering
Veermata Jijabai Technological Institute
Mumbai, India
e-mail: sprana_m17@et.vjti.ac.in
[3]Department of Electronics Engineering
Veermata Jijabai Technological Institute
Mumbai, India
e-mail: fskazi@el.vjti.ac.in

Abstract— Hydrogen fuel cell is used to run fuel cell hybrid electrical vehicles (FCHEVs). These FCHEVs are more efficient than vehicles based on conventional internal combustion engines due to no tailpipe emissions. FCHEVs emit water vapor and warm air. FCHEVs are demanding fast dynamic responses during acceleration and braking. To balance dynamic responsiveness, develop hybrid electric cars with fuel cell (FC) and auxiliary energy storage source batteries. This research paper discusses the development of an energy management strategy (EMS) for power-split FC-based hybrid electric cars using an algorithm called deep deterministic policy gradient (DDPG) which is based on deep reinforcement learning (DRL). DRL-based energy management techniques lack constraint capacity, learning speed, and convergence stability. To address these limitations proposes an action masking (AM) technique to stop the DDPG-based approach from producing incorrect actions that go against the system's physical limits and prevent them from being generated. In addition, the transfer learning (TL) approach of the DDPG-based strategy was investigated in order to circumvent the need for repetitive neural network training throughout the various driving cycles. The findings demonstrated that the suggested DDPG-based approach in conjunction with the AM method and TL method overcomes the limitations of current DRL-based approaches, providing an effective energy management system for power-split FCHEVs with reduced agent training time.

Keywords- Deep reinforcement learning; Energy management strategy; Fuel cell ; State of Charge; Transfer learning.

I. INTRODUCTION

The energy management (EM) system plays a crucial role in distributing power from various energy sources. Its primary objective is to ensure that the energy demand for everyday use is met while ensuring fair distribution of electricity. With the emergence of advanced technologies such as fuel cells and auxiliary batteries EMS has the potential to revolutionize the road transportation sector by promoting low carbon emission and eco-friendly practices. The EMS system enables efficient allocation and utilization of energy resources which leads to reduced environmental impact and enhanced sustainability in the long run.Therefore, the EMS should refine design and control to optimize performance across a variety of use cases. For FCHEVs, several researchers have examined various optimization targets for EMS. Some research takes a more narrow approach to find a solution by focusing just on reducing fuel use. However, fuel cells now have a high production cost and a limited lifespan. Consequently, numerous researchers have considered power system longevity to be an additional optimization target. When it comes to selecting the most appropriate approach for optimizing a system, there are typically trade-offs to consider among competing optimization objectives. This can make it challenging to determine whether to use a rule-based approach, an optimization-based method, or a learning-based technique. However, this study has specifically chosen to concentrate on the learning-based EMS as a viable solution.

A literature review on learning-based EM for FCHEVs reveals several approaches that have been proposed in recent years. One of the primary challenges in developing effective EMS for FCHEVs is the highly nonlinear and dynamic nature of the system, which makes it difficult to optimize performance in real-time. [1-3] In the recent years, learning-based EM have gained increasing attention for hybrid vehicles. These EMSs

International Journal on Recent and Innovation Trends in Computing and Communication
ISSN: 2321-8169 Volume: 11 Issue: 9s
DOI: https://doi.org/10.17762/ijritcc.v11i9s.7396
***Article Received**: 02 May 2023 **Revised:** 28 June 2023 **Accepted:** 20 July 2023*

utilize learning algorithms such as reinforcement learning (RL) and DRL. RL-DRL based EM rely solely on data, enabling them to achieve optimal control outcomes through trial and error learning and also interactions between agent, environment and exhibiting excellent adaptability and real-time performance [4-6]. The RL algorithm has found extensive use in different types of hybrid vehicles including engine-motor hybrids, FCHEVs, engine-ultracapacitor hybrids vehicle. Exploitation of opportunities for learning and the choice of such opportunities were studied in a parametric analysis of RL-based EM for hybrid vehicles in a research article [7-8]. Hybrid vehicle RL-based EMSs often use an offline training and online application mode to improve control optimality, convergence rate, and flexibility. To enhance the performance of RL-based EM, various techniques have been developed [9]. One approach to address this challenge is through the use of machine learning (ML) techniques. In [10] DRL based EM was proposed for FCHEVs. In which the system used a DRL agent to learn the optimal power management strategy. Results showed that the DRL-based approach was effective in improving the fuel economy and reducing emissions of the FCHEV. DRL has been widely applied in various domains, including robotics [11], building heating ventilation air conditioning control [12], ramp metering [13], and more. DRL is also extensively used in automotive systems, such as lane keeping assist [14], automated braking, and driverless cars [15]. These applications take advantage of the ability of DRL to enable an agent to learn from its environment and optimize its behavior based on the feedback received through trial-and-error interactions. The flexibility and adaptability of DRL make it a promising approach for developing intelligent control systems that can improve safety, efficiency, and performance in a wide range of applications. A deep Q-network (DQN) algorithm has been utilized for fitting the Q-table in order to consider more state variables and achieve accurate identification of these variables, which can reflect any continuous changes through decision-making system [16]. Research in [17] a Deep Q network based EM was suggested for hybrid electric bus. Based on the data, it seems that proposed EMS achieved better fuel economy than the RL-based EM. An EM framework that incorporates expert knowledge into the DDPG algorithm was proposed by [18-20]. This framework was designed to address the problem of increased control variables. It resulted in accelerated learning and improved fuel economy. However, Reinforcement Learning often encounters sparse rewards, which necessitates complex reward engineering.

This paper proposes an innovative approach to improve the EM of hybrid electric vehicles by implementing an action restriction technique that restricts the set of actions an agent can take in certain states to prevent invalid or prohibited actions. The proposed method also ensures the protection of the battery from overcharge and under discharge by power constraint of vehicle and reduces the training time required to avoid prolonged periods. In addition, the system monitors and logs the fuel cell and battery parameters to maintain fuel efficiency. To further enhance the effectiveness of the EM system, combining DDPG with TL which transfers knowledge from one domain to another resulting in a more efficient EMS selection process. This approach avoids the need for retraining the network after changing driving cycles which can be time-consuming.

This paper is structured as follows: Part II represents the modelling of FC, battery and FCHEVs dynamics, Implementation of a DRL-based energy management strategy is discussed in Part III, The results of the study and a discussion of the findings are presented in Part IV, Part V concludes the paper.

II. MODELLING OF ENERGY SOURCES

A. *FC Model*

Dick - larminie proposes the physical and electrochemical phenomena of fuel cell to its equivalent electric equivalent circuit. The proposed Dick-larminie electric circuit model use to model concentration, activation and Ohmic polarization and Nernst voltage. An electrical equivalent model of fuel cell is introduced in Figure 1.

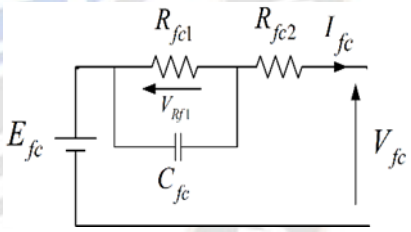


Figure. 1 Fuel cell model

$$I_{fc} = \frac{V_{\mathrm{Rf1}}}{R_{\mathrm{fc1}}} + C_{\mathrm{fc}} \frac{dV_{\mathrm{Rf1}}}{dt} \qquad (1)$$
$$V_{\mathrm{fc}} = E_{\mathrm{fc}} - V_{\mathrm{Rf1}} - R_{\mathrm{fc2}} I_{\mathrm{fc}}$$

Here, *Rfc1* is activation and concentration losses of double layer capacitance. *Rfc2* represents the flow of hydrogen and electrons. Capacitor *Cfc* is double layer charge, *Efc* is the voltage source as open circuit voltage and *Vfc* represent the fuel cell voltage supplying to the load.

$$P_{fc} = P_{fc}^{'} / \eta_{dc}(P_{fc}^{'}) + P_{aux} \qquad (2)$$

Where, $P_{fc}^{'}$ - o/p power FC. Assume $P_{fc}^{'}$ power requested by control strategy, η_{dc} - DC-DC converter efficiency for FC, P_{aux} - auxiliary system assume as a constant c\n load I_{aux} = 2 amp

Since PEM fuel cells operate most inefficiently in low power conditions, they are rarely used in such situations. After a certain point, where power consumption has reached its peak,

International Journal on Recent and Innovation Trends in Computing and Communication
ISSN: 2321-8169 Volume: 11 Issue: 9s
DOI: https://doi.org/10.17762/ijritcc.v11i9s.7396
***Article Received**: 02 May 2023 **Revised:** 28 June 2023 **Accepted:** 20 July 2023*

efficiency begins to drop. A diagram depicting fuel cell performance in relation to power demand is presented in Figure 2.

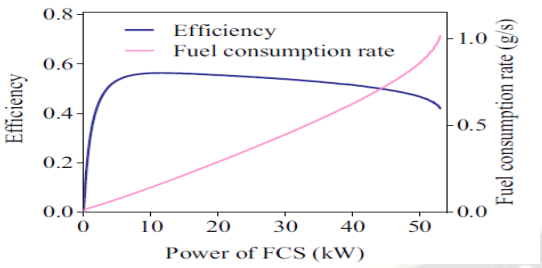


Figure. 2 Fuel cell efficiency and power characteristics

B. *Battery Model*

Two level register capacitor based electrical equivalent battery model is considered as basic necessity for battery state of charge estimation [21].

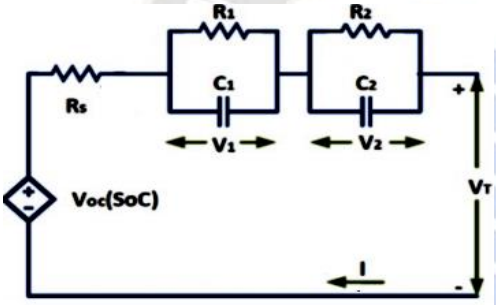


Figure. 3 Two stage RC equivalent circuit

$$\frac{d}{dt}V_1(t) = \frac{-V_1(t)}{R_1C_1} + \frac{I(t)}{C_1} \quad (3)$$

$$\frac{d}{dt}V_2(t) = \frac{-V_2(t)}{R_2C_2} + \frac{I(t)}{C_2} \quad (4)$$

$$\frac{d}{dt}SoC(t) = -\frac{\eta I}{Q} \quad (5)$$

(Q is the rated capacity of the battery, η - coulombic efficiency)

By applying Kirchoff's voltage law to the Figure 3.

$$V_T(t) = Voc(SoC(t)) - R_sI - V1 - V2 \quad (6)$$

Equations (3), (4), and (5) are linear state equations and because of Voc(SoC(t)) term in eq. (6), eq. (6) is nonlinear o/p equation. The Taylor's series expansion around SoC operating point SoC_0 is used for linearizing nonlinear system at every time step.

Battery o/p power

$$P_{bat} = V_{oc}(SOC_{bat})I_{bat} - I_{bat}^2R_{bat}(SOC_{bat}) \quad (7)$$

Where, V_{oc} battery open circuit voltage, I_{bat} - battery o/p current, R_{bat} - battery internal resistance.

$$P_{bat} = \left\{P_{bat}^{'} / \eta_{bdc}\right\}.....(P_{bat}^{'} > 0)$$
$$P_{bat} = \left\{P_{bat}^{'}.\eta_{bdc}\right\}.....(P_{bat}^{'} < 0) \quad (8)$$

Where, $P_{bat}^{'}$ is the o/p power of the power converter whose efficiency is η_{abc}

C. *FCHEV's Vehicle Dynamics*

Consider a FCHEVs driving at v on a road with gradient θ

$$F_m = F_{air} + F_f + F_s + F_a \quad (9)$$

$$= \frac{1}{2}C_DA\rho v^2 + Gf\cos\theta + G\sin\theta + m\frac{dv}{dt} \quad (10)$$

Where, F_m - diving force delivered by motor, F_{air}- air resistance (1.2), F_f- rolling resistance(0.0013), F_s - slop resistance, F_a - acceleration resistance, ρ - air density coefficient(1.50 kg/m2) C_D-air resistance coefficient(0.26), A - windward surface volume of the vehicle(1.8m2), *v*- vehicle velocity, m - vehicle mass (1449 kg), G=mg gravity of the vehicle (9.8m/s2), *f*- vehicle slideing resistance coefficient (0.4) The required power for the FCHEV :

$$P_{veh} = F_m.v / \eta_m \quad (11)$$

Where, P_{veh} - required power of the FCHEV motor, η_m - x'mission efficiency of electric machine (90%) FC and battery provide the motor's power, according to the power balance.

$$P_{veh} = P_{fc} + P_{bat} \quad (12)$$

Figure 4 shows the configuration of FCHEVs and in this configuration main power source is fuel cell and assisting power source battery is used. FC is connected to common DC link using unidirectional boost converter and battery is connected to common DC link using bidirectional buck boost converter.

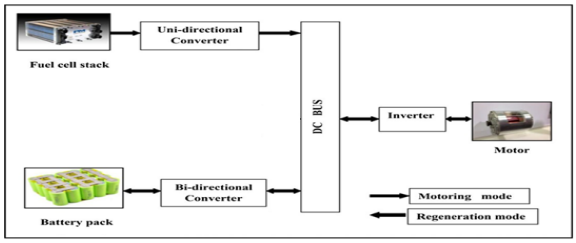


Figure. 4 FC-battery configuration

International Journal on Recent and Innovation Trends in Computing and Communication
ISSN: 2321-8169 Volume: 11 Issue: 9s
DOI: https://doi.org/10.17762/ijritcc.v11i9s.7396
***Article Received**: 02 May 2023 **Revised:** 28 June 2023 **Accepted:** 20 July 2023*

III. DRL ALGORITHMS BASED TRANSFERABLE EMS FOR FCHEVS

A. Implementation of DRL Based Energy Management Strategy

Reinforcement learning is a technique used to find the optimal actions to take in a given state in order to maximize future rewards as shown in equation 13. In the case of FCHEVs, the hybrid drivetrain is the controllable object and various factors such as powertrain condition, driving conditions, and driver needs can represent the state at a given time (t). The energy distribution system is the action (a) taken at time t. Real-time metrics such as instantaneous fuel consumption, SoC variations, speed, and acceleration are monitored and used to calculate the reward (r) for the energy distribution system. This method of regulation is commonly referred to as the energy management system[22].

$$\pi^* = \arg\max E_{a(t)\in A}\left[\sum_{t=0}^{N-1} r(s(t),a(t))T_s\right]_{s(o)=s_0} \quad (13)$$

Where, pi - ideal EMS, A - action space, T_s - sampling time, S_0 -starting state and N - time sequence length of the finite step Markov decision process problem. Through reinforcement learning, the EMS [36] learns to efficiently distribute energy by exploring the vehicle and its driving environment while receiving feedback. In the Markov decision process (MDP), Equation 14 defines the action-value function as the expected cumulative reward obtained by taking action a and following the optimal policy.

$$Q_\pi(s,a) = E_\pi\left[\sum_{t=0}^{N-T} r(s(t),a(t))\gamma^t \mid s(0)=s_T, a(0)=a_T\right] \quad (14)$$

Where, S_T and a_T- state, action at time T, γ discount rate determining the present value of future reward (γ =0.99).

Reinforcement learning approaches are designed with the Bellman Equation at the core. One popular approach is the value-based reinforcement learning method, where the action-value function is updated iteratively through interactions between agent-environment [23] based on the likelihood of an actual action. Updating the action-value function immediately using Equation 14 can be inefficient as it requires retracing the entire control sequence. To address this, the Temporal Difference (TD) update concept is commonly utilized in reinforcement learning to expedite the learning process of estimating Q(s,a), as demonstrated in Equation 15.

$$Q_{new}(s,a) \leftarrow Q_{old}(s,a) + \alpha\left[(r + \gamma \max_{a'} Q_{old}(s',a')) - Q_{old}(s,a)\right] \quad (15)$$

Where, $(r + \gamma \max_{a'} Q_{old}(s',a')) - Q_{old}(s,a)$ is TD error (Temporal Difference δ), Q_{old} (s,a) represents the Estimate_{old}, $(r + \gamma \max_{a'} Q_{old}(s',a'))$ represents target, (s, a) represents s(t),a(t), (s',a') represents s(t+1),a(t+1), r represents r(s(t),a(t)), s = {v, SOC, acceleration, P_{dem}, θ_{slope}, SOC_{ref}, v-1,v-2,v-N), $a(t) \in$ A,t=1,2,3……N

B. Deep Q Learning (DQN) Algorithm Based EMS

In a Q-table the dimensions of state and action are used to hold the Q-value. In reinforcement learning [24], the Bellman equation-based temporal difference technique is often used. The learning algorithm heavily relies on this equation, which can be adapted to different strategies. The calculation of the Q-value involves the consideration of time difference, and the following equation, Eq. 16, illustrates the use of the temporal difference update to estimate Q(s,a) and speed up the learning process.

$$Q^{new}(s_t,a_t) \leftarrow Q(s_t,a_t) + \alpha(r_t + \gamma.\max_a Q(s_{t+1},a) - Q(s_t,a_t)) \quad (16)$$

Algorithm: Deep Q Learning

1: perform an initialization of replay memory D with capacity N.

2: perform some random weighting at the beginning of the Q that is at the beginning of the action-value function.

3: for epi = 1, M do

perform an initialization and preprocessed sequenced as s1 = {x1}, ϕ1 = ϕ (s1) resp.

4: for t = 1, T do

along with the probability ϵ perform the selection of any action a_t else select $a_t = \max_a Q^*(\phi(s_t), a; \theta)$

5: perform a_t and notice reward r_t and x_{t+1}

6: assign $s_{t+1} = s_t, a_t, x_{t+1}$ and executes $\phi_{t+1} = \phi(s_{t+1})$

save transition $(\phi_t, a_t, r_t, \phi_{t+1})$ in memory D

7: Take sample random minibatch (transitions) $(\phi_j, a_j, r_j, \phi_{j+1})$ in memory D

8: set $y_{j=} r_j$ for terminal ϕ_{j+1} and set

$y_{j=} r_j + \gamma \max a' Q(\phi_{j+1}, a'; \theta)$ for non-terminal ϕ_{j+1}

9: Execute gradient descent step on $(y_j\ Q(\phi_j, a_j; \theta))$

International Journal on Recent and Innovation Trends in Computing and Communication
ISSN: 2321-8169 Volume: 11 Issue: 9s
DOI: https://doi.org/10.17762/ijritcc.v11i9s.7396
***Article Received**: 02 May 2023 **Revised:** 28 June 2023 **Accepted:** 20 July 2023*

10: end for

11: end for

In the problem at hand, the states include the requested power P_{dem}, the power derived from the battery P_b, the power from the fuel cell P_{fc}, state of charge of battery is SoC and the deviation of the charge available in the battery SoC_{des}. While these are the most commonly used states for these variables there can be other ways of specifying them. During training various combinations of states are tested until the reward maximization is achieved. In the case of DQN the following states can be considered:

$$S = \{SoC - SoC_{des}, P_{dem}\} \tag{17}$$

The system has no bounds because the state P_{dem} is the input. However, there are constraints on the initial state if SoC is less than SoC_{des}, as shown by the following:

$$-SoC_{diff,\lim it} < SoC - SoC_{des} < SoC_{diff,\lim it} \tag{18}$$

One version of the objective function represented by the reward function below is arrived at by learning.

$$r = -\tanh(\alpha \dot{m}_{H_2} + \beta \mid \Delta SoC_{ref} \mid^2) \tag{19}$$

Equation 16 shows that the key to finding the best action is to learn to select an action with a high expected reward return. If the state exceeds the limit, the simulation terminates and the agent is penalized to discourage such behavior in the future. The loss function of the network is given by

$$losses = (r + \gamma \max_{a'} Q(s', a'; \theta') - Q(s, a; \theta))^2 \tag{20}$$

C. Deep Deterministic Policy Gradient Based EMS

DDPG employs separate actor and critic networks that independently evaluate and critique each other using the state and behavior inputs. The actor network generates actions based on the input states, utilizing a deterministic policy gradient instead of a probability distribution to produce a deterministic action. As the problem involves continuous actions and states, this approach is well-suited for solving it. Meanwhile, the critic network takes in states and their associated behaviors as inputs and outputs a Q-value. To facilitate exploration, a noise strategy is utilized. DDPG distinguishes itself from DQN by utilizing a soft update mechanism for its parameters, loss function, and policy gradient [25].

$$\begin{aligned} loss &= \frac{1}{n}\sum [Z_t - Q(s_t, a_t \mid \theta_c \mid_{a_t = \mu(s_t \mid \theta_a}]^2 t \\ a_t &= \mu(s_t \mid \theta_a) \\ z_t &= r + \gamma Q'(s_{t+1}, \mu'(s_{t+1} \mid \theta_a' \mid \theta_c') \\ \nabla_{\theta_a} loss &= \frac{1}{n}\sum [\nabla_{a_t} Q(s_t, a_t \mid \theta_c) \nabla_{\theta_a} \mu(s_t \mid \theta_a)] \end{aligned} \tag{21}$$

Where, n – no. of minibatch, θ_a and θ_c - parameters for actor and critic, θ_a' and θ_c' - parameters of the target actor and target critic, μ - function to map action, r – reward, γ - discount rate.

In contrast to DQN, the action space and state space in DDPG are not distinctly separated. However, there are slight differences in the state space between the two algorithms. The reward function is nearly identical between the two algorithms, with a small modification that is defined as follows:

$$r = -\tanh(\alpha \mid \Delta \dot{m}_{H_2} \mid^2 + \beta \mid \Delta SOC_{ref} \mid^2) \tag{22}$$

The simulation will end with the agent receiving a penalty if state boundaries are violated.

Algorithm: DDPG Algorithm with action mask

1: Initialize critic network $Q(s, a / \theta^Q)$ and actor $u(s / \theta^u)$ with weights θ^Q and θ^u

2: Initialize target network θ' and u' with weights $\theta^{Q'} \leftarrow \theta^Q$, $\theta^{u'} \leftarrow \theta^u$

3: Initialize buffer R

4: for loop for epi = 1, M do

set N for action exploration as a random process

5: Initial observation is received as a state s_1

6: for t = 1, T do

Select the action $a_t = u(s_t / \theta^u) + N_t$ for exploration of the noise with action mask clip function $P_{FC}(t)$ = clip [P_{min} FC (t), P_{max} FC (t)]

7: Execute a_t and observe r_t and observe s_{t+1}

8: Store changes (s_t, a_t, r_t, s_{t+1}) in buffer R

9: Sample N transitions (s_i, a_i, r_i, s_{i+1}) from buffer R where as N is random minibatch

10: Set $z_i = r_i + \gamma Q'(s_{i+1}, u'(s_{i+1} / \theta^{u'}) / \theta^{Q'})$

11:Update: $L = \frac{1}{N}\sum i(y_i - Q(s_i, a_i / \theta^Q))^2$ at critic

International Journal on Recent and Innovation Trends in Computing and Communication
ISSN: 2321-8169 Volume: 11 Issue: 9s
DOI: https://doi.org/10.17762/ijritcc.v11i9s.7396
***Article Received**: 02 May 2023 **Revised:** 28 June 2023 **Accepted:** 20 July 2023*

12:Update the sampled policy gradient:

$$\nabla_{\theta^u} J \approx \frac{1}{N}\sum i \nabla_a Q(s,a|\theta^Q)\Big|_{s=si,a=u(si)} \nabla_{\theta^u} u(s|\theta^u)\Big|_{si}$$

13: Update the target networks:

$$\theta^{Q'} \leftarrow \tau\theta^Q + (1-\tau)\theta^{Q'}$$
$$\theta^{u'} \leftarrow \tau\theta^u + (1-\tau)\theta^{u'}$$

14: end for

15: end for

Environments having Markov property are met by DRL agent. Each time the agent makes a decision on what to do next, the environment either rewards or punishes. Combining the experience of FCHEV experts with the DDPG algorithm, this study can determine the best course of action for EMS. DDPG-based EMS rewards based on immediate use of fuel cell and battery charge sustaining costs based on this two points multi-objective reward function is fc

$$state = \{SoC, velocity, acceleration\}$$
$$action = \{continuous\_power\}$$
$$reward = -\tanh(\alpha |\Delta \dot{m}_{H_2}|^2 + \beta |\Delta SOC_{ref}|^2)$$

(23)

The fuel utilization rate factor and SoC deviation are denoted by α and β respectively, where a larger β value of indicates a greater reward for following the SoC recommendations carefully. Similarly, a larger value of α indicates significant rewards for reducing fuel consumption. The primary objective of fine-tuning and is to maximize fuel efficiency while ensuring that the trained strategy satisfies the SoC requirements. After several rounds of tuning, the value of α is set to 50, while the values of β and SoC_{ref} are set to 150 and 0.65 respectively as the initial SoC value. To achieve good battery efficiency the SoC must meet the top and lower constraints.

D. Action Masking for DDPG

The action mask is a widely used method for preventing invalid actions in reinforcement learning, where the agent explores the action space through trial-and-error. However, unrestricted exploration can pose safety risks, particularly in scenarios such as vehicle control, where constraints limit the number of available actions. In such cases, soft constraints are typically employed to assign a large negative reward to invalid actions, incentivizing the agent to avoid them. Despite being effective, this approach has limitations, such as a longer training time and increased model instability. To address these concerns, this study employs a hard-constrained approach using an action mask, which filters out invalid actions and restricts the agent to selecting only valid ones. In order to prevent DDPG from engaging in pointless learning exploration and selecting invalid actions, action masking is necessary. DDPG's main objective is to create effective long-term planning strategies by learning the distribution of states, actions, and state transitions in the environment. To meet this goal, AM must satisfy two criteria: first, it should not alter the original action space distribution, which could otherwise damage the potential state transfer probability function. Second, incorrect samples must be excluded from training data as they would not be stored in the experience replay buffer. To apply AM at each time step t, DDPG follows a specific set of procedures.

To determine the appropriate working range at each time step t, the following three stages are utilized: Firstly, the Action set is defined as $\{P_{FC}|P_{FC} \in$ [0 KW power, Max power in KW]}, which is then discretized to generate a(t). Secondly, the fuel cell's maximum and minimum power at time t is computed by traversing a(t) in accordance with the dynamics of the driving cycle. Finally, the new fuel cell max and min power at time t are obtained. Subsequently, the FC energy $P_{FC}(t)$ output from the actor network in the DDPG based algorithm is subjected to a clip operation $P_{FC}(t)$ = clip [P_{FC}, P_{minFC} (t), $P_{max\ FC}$ (t)]. Since the clip function does not modify the initial action "A" there is no impact it can have on the DDPG. It should be emphasized that the first stage, which involves traversing the A-action space to eliminate erroneous actions, is a crucial step employed by a wide range of mathematical model-based techniques. This step is extensively used for good reason. Furthermore, it is essential to note that the action masking strategy must be implemented for both the actor and target actor networks otherwise, the algorithm's ability to learn will be severely hindered. Moreover, it is important to mention that the method of concealing faulty actions using the clip function is only applicable to algorithms based on actor-critic and deterministic policy, such as DDPG. This is because the clip function ensures that the output of the actor network is within the acceptable range of values, which is a critical requirement for these algorithms to functioning effectively.

E. Transferable DRL

A bi-level control structure used as a solution to the EM problem in FCHEVs. To learn the EMS [25], the driving cycles are divided into intervals with three different speeds. The DRL algorithms are then used to search for the best control strategy for the examined powertrain. TL is also implemented to accelerate the training process. In summary, the bi-level control structure is proposed as a solution to the EM problem in FCHEVs. The training process involves dividing driving cycles into intervals with different speeds, using DQN-DDPG to identify the best control strategy and applying transfer learning to speed up the convergence of the training procedure.

International Journal on Recent and Innovation Trends in Computing and Communication
ISSN: 2321-8169 Volume: 11 Issue: 9s
DOI: https://doi.org/10.17762/ijritcc.v11i9s.7396
***Article Received**: 02 May 2023 **Revised:** 28 June 2023 **Accepted:** 20 July 2023*

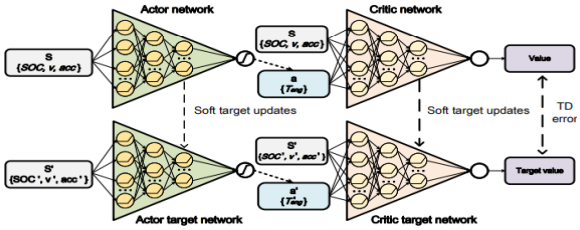


Figure. 5 DRL and TL based EMS structure

In this paper, the DDPG algorithm is used to investigate the EMS issue in FCHEVs. Developing a exercise model for a new driving cycle can be a time-consuming process. The driving cycle's vehicle speeds are first divided into three categories based on TL theory. To transfer the DDPG parameters from one speed interval to another, TL is helpful due to the similarities in the learning problem and feature space across different speed ranges. Only the parameters of the network's outer most layer need to be retrained and the rest of the network's parameters can remain unchanged. The EMS is trained for various speed ranges using the DDPG method and the corresponding parameters are stored in memory. Since FCHEVs driving cycles share the same feature space and are correlated they must all be driven in the same way. The learned parameters are then applied to the new driving cycle at various speeds to increase computation efficiency. This allows for the effective creation of the EMS and ensures its optimality.

Figure 5 illustrates a DDPG-based control structure used for developing an optimal control strategy. Standard deep learning techniques handle training and testing data from the same domain. However, when the criteria are not met rebuilding the model and retraining the data can be expensive and time consuming. Transfer learning can address this issue by reusing most of the neural network settings when two study issues are similar. In this TL method, there is a source domain (Sm) and a learning task (Tm), as well as a target domain (Sn) and a corresponding task (Tn). Applying what is learned in one setting (Sm) to another (Tm), where Sm Sn or Tm Tn is an example of transfer learning. This enables the goal prediction function (f) in Sn to be learned more effectively. The results section examines the simulation findings and draws conclusions about the efficacy of the DDPG and TL-enabled EMS.

Table 1. H parameters for DDPG training algorithm

Name of the parameters	Parameters value
Experience Buffer Length	1e6
Critic Learning Rate	1e-4
Simulation Time	15
Agent Noise Variance Decay Rate	1e-3
Mini Batch Size	128
Actor Learning Rate	1e-4
Agent Noise Variance	0.1
Sample Time	0.1
Discount Factor	0.99

IV. RESULTS AND DISCUSSION

In this section, discuss the implementation of reinforcement learning-based EMS and evaluating the optimum performance, flexibility, and maximum efficiency in all circumstances of the proposed EMS for FCHEVs. In addition, we look at how to implement RL-based EMS. The DQN and DDPG control techniques are evaluated as a benchmark to determine whether or not the DDPG + TL approaches are effective. In the next step, the Q-value table's convergence rate is evaluated by doing an in-depth comparison between the DDPG + TL and the DDPG with AM. DDPG with transfer learning techniques are given to the new but same driving cycles in order to validate adaptation. The Q-Learning approach in this study includes a gradual decrease in exploration rate from 1.0 to 0.001 to achieve optimal outcomes. A learning rate of 0.01 and a decay rate of 0.9 are assigned to this setup.The state and control variables are discretized in this study with P_{veh}, SoC_{bat}, and P_{FC} having a step size of 1 kW, while P_{FC} has a step size of 10%. The exploration rate decreases from 1.0 to 0.001 in the Q-Learning setup to achieve the desired outcomes. A total of one thousand episodes are performed under these conditions.

To determine the effectiveness of the suggested EMS based on transfer learning (TL) and deep deterministic policy gradient (DDPG), baseline measures including DDPG with action masking and DQN were used. The globally optimum control actions were created by DQN, making it the benchmark to compare the effectiveness of the proposed strategy. The default settings for both DDPG+TL and DDPG were identical, allowing for a fair comparison. To facilitate this comparison, the standardized driving cycle New European Driving Cycle (NEDC) was used. The differences between the proposed strategy and the baseline measures were used to evaluate the degree of optimality of the suggested EMS. In the context of an EMS, action masking is a useful approach that involves limiting the set of actions that an agent can take. This technique not only helps to streamline the learning process, but it also contributes to improving the efficiency and effectiveness of the system by enabling the agent to focus on a narrower set of actions that are more likely to yield positive outcomes. Additionally, by reducing the complexity of the decision-making process, action masking can enhance the speed and accuracy of the agent's actions which can be particularly important in time-sensitive scenario.

Figure 7 displays the o/p power of the FC - battery with respect to the power requirements of the FCHEV. The numerical values presented in the figure demonstrate the advantages of the proposed technique in terms of power

International Journal on Recent and Innovation Trends in Computing and Communication
ISSN: 2321-8169 Volume: 11 Issue: 9s
DOI: https://doi.org/10.17762/ijritcc.v11i9s.7396
***Article Received**: 02 May 2023 **Revised:** 28 June 2023 **Accepted:** 20 July 2023*

distribution at the control level. Figure 8 shows the cumulative state-of-charge trajectory of batteries using DDPG + TL, DQN, and DDPG algorithms under the NEDC driving cycle as depicted in Figure 6. While the terminal SoC values are almost identical, it can be observed that the battery's remaining charge state in DDPG is quite similar to that in DQN implying that their control sequences are comparable. Moreover the changes in SoC reveal that DDPG plus transfer learning performs better than traditional DDPG in terms of the battery's power output. The patterns of variation in DDPG and DQN are similar which means that their fuel efficiency performance may vary. Therefore, by utilizing transfer learning from previous knowledge about EM, the DDPG plus transfer learning approach has the potential to reduce fuel consumption and ensure optimality.

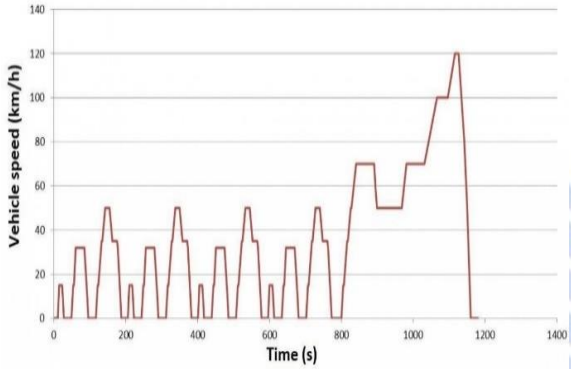
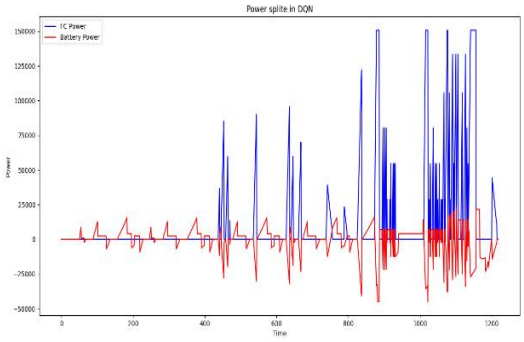
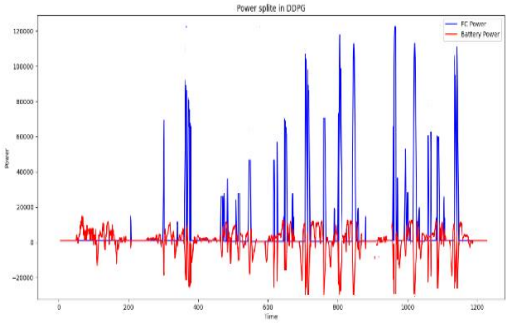


Figure. 6 New european driving cycle (NEDC)





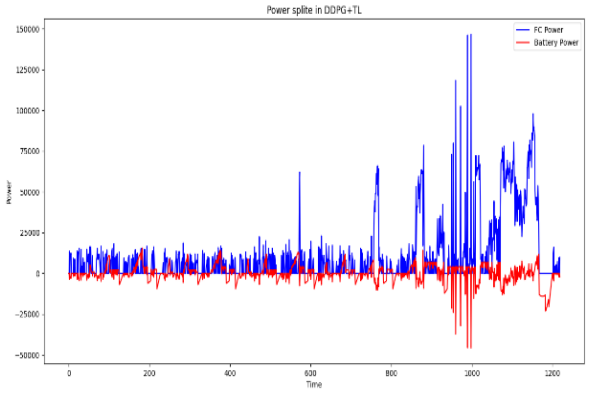
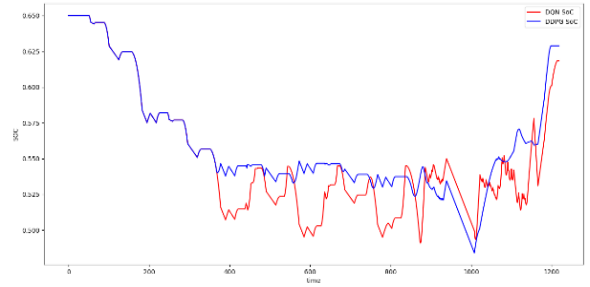


Figure. 7 Power distribution between FC and battery in three EMS models.



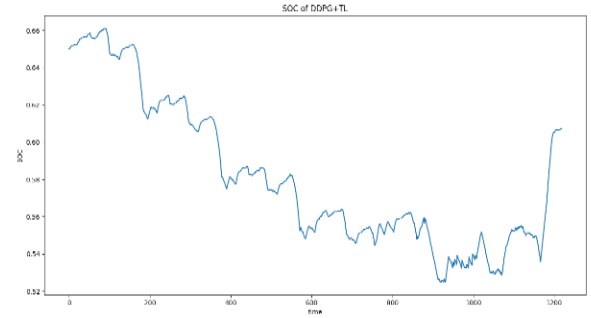


Figure. 8 SoC trajectories of three EMS models

The DDPG + TL algorithm is an effective method used during the training phase to accelerate the convergence process of the agent. This is achieved by playing crucial samples from the experience pool at a higher frequency, which helps to reinforce the learning of the most important actions and strategies. By repeating these essential samples, the agent can quickly learn to prioritize the most rewarding actions and achieve higher levels of performance in a shorter amount of time.

The DDPG plus TL algorithm is a powerful technique designed to expedite the process of identifying optimal controls by leveraging the parameters learned by a neural network. In comparison to other EMSs, the primary distinction lies in the length of the driving cycle. The TL theory posits that only the outer most layer of the neural network needs to be retrained to

International Journal on Recent and Innovation Trends in Computing and Communication
ISSN: 2321-8169 Volume: 11 Issue: 9s
DOI: https://doi.org/10.17762/ijritcc.v11i9s.7396
***Article Received**: 02 May 2023 **Revised:** 28 June 2023 **Accepted:** 20 July 2023*

adapt to a new driving cycle which can significantly reduce the time required to master a new control regime. To assess the efficacy of DDPG plus TL we examined its convergence rate and training time in comparison to the standard DDPG algorithm. Both DRL approaches approximate the Q-value table using a neural network. The results, as shown in Figure 9 reveal a decreasing trend in the average inaccuracy of the Q table and the training sessions, indicating that the quality of the control sequence obtained improves with each subsequent episode. Moreover, compared to the standard DDPG the mean error value in DDPG plus TL is more reasonable for each episode suggesting that this approach may provide a better understanding of the agent's environment leading to more effective regulation. Therefore it is apparent that DDPG+TL is more effective at learning than the DDPG algorithm.

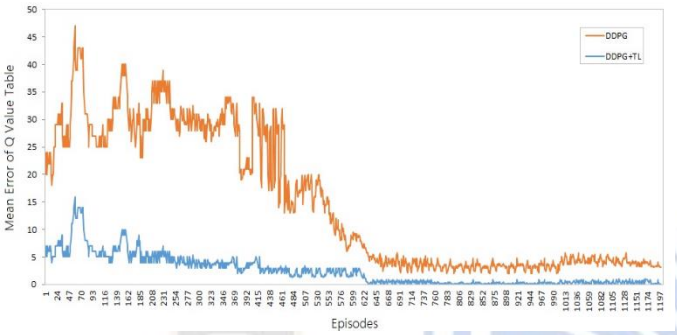


Figure. 9 Mean error under DDPG and DDPG plus TL

The effectiveness of Deep Reinforcement Learning training can be quantified by the cumulative rewards, which increase with the number of episodes as illustrated in Figure 10. Notably, the addition of transfer learning to DDPG results in significantly higher cumulative rewards for the same number of episodes compared to the DDPG and DQN. At the beginning episodes of DDPG and DDPG + TL the reward values for both methods are similar due to the use of the same random seeds. This findings suggest that the addition of AM does not impact the learning performance of DDPG, nor does it affect its stability or slow down the learning process. This observation can be attributed to the fact that AM does not interfere with the distribution of the environment, thus complying with the mathematical concepts underlying DDPG. To provide further technical insights, Table 4 presents the total training time for both DDPG strategies on the same driving cycle. The results indicate that DDPG with TL is more efficient than the regular DDPG in terms of reducing the learning time. This feature makes it possible to apply the planned EMS in practical driving situations with greater feasibility.

Table 2 Learning duration for DDPG and DDPG plus TL

Algorithm	Training Time (hrs)
DDPG	04.07
DDPG plus TL	0.82
Intel(R) Core(TM) i3- 3110 M Processor @ 2.40 GHz and RAM Primary Memory 8.00 GB (7.83 GB usable)	

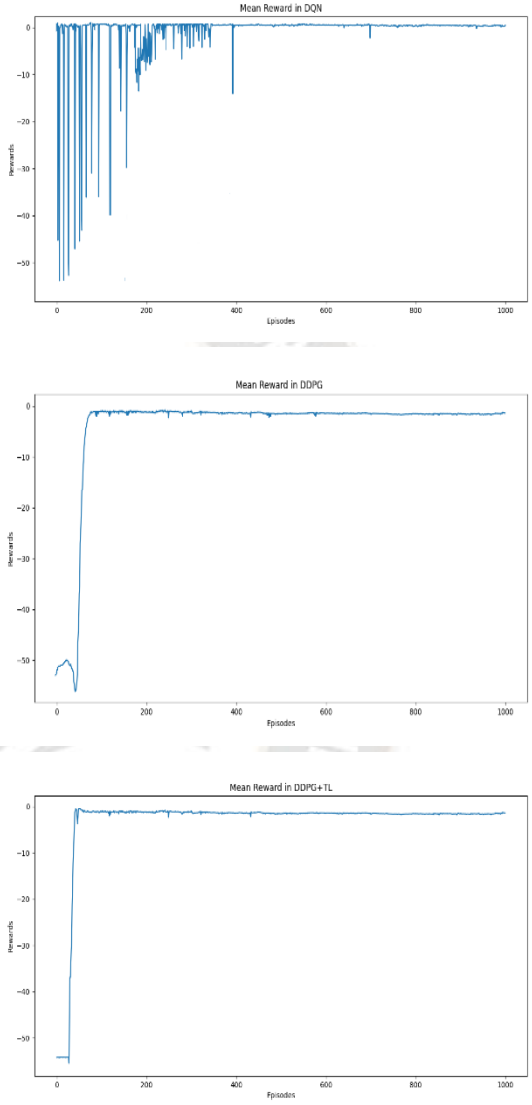


Figure.10 Rewards under in three EMS models

V. CONCLUSION

The focus of this paper is to introduce a learning-based energy management system that is model-free and uses the DDPG algorithm in deep reinforcement learning. The DDPG method employs trial and error to identify the most optimal EMS solution by collecting a large number of actual samples from the real environment, resulting in improved performance. The study utilizes a two-level register-capacitor based equivalent circuit model to nonlinearly model the battery and

International Journal on Recent and Innovation Trends in Computing and Communication
ISSN: 2321-8169 Volume: 11 Issue: 9s
DOI: https://doi.org/10.17762/ijritcc.v11i9s.7396
***Article Received**: 02 May 2023 **Revised:** 28 June 2023 **Accepted:** 20 July 2023*

fuel cell in simulation. The primary target of this research was to investigate the feasibility of incorporating reinforcement learning algorithms into EMS. Despite Q-learning's widespread use in EMS this study employed a model-free DQN algorithm in conjunction with Q-learning for FC FCHEVs with the training of the agent alongwith a drive cycle the agent with the highest reward was chosen and its performance was evaluated. A similar process was followed for the DDPG algorithm which has a continuous action space and is ideal for FCHEVs. The use of action masking and reward shaping techniques improves the DDPG agent's performance in complex environments by restricting the set of actions that the agent can take and providing a measure of success or failure for each state-action pair. However, the training time for another drive cycle is significantly longer which is a drawback. To address this issue, this study combines DDPG with transfer learning to construct an adaptive EM controller for FCHEVs, reducing the laborious training time associated with the DRL technique. The control architecture can be easily adapted to other hybrid powertrains, making it a promising approach for future research.

ACKNOWLEDGEMENTS

Authors acknowledge Centre of Excellence in Complex and Nonlinear Dynamical Systems laboratory for providing support and platform for research.

DECLARATIONS

Conflict of interest The authors declare that they no conflict of interest.

REFERENCES

[1] Heeyun L, Suk C, “Energy Management Strategy of Fuel Cell Electric Vehicles Using Model-Based Reinforcement Learning With Data-Driven Model Update”, IEEE, (2021) pp. 59244-59254.

[2] N.Sulaiman, M.Hannan, A. Mohamed, E. Majlan, W.Wan, “A review on EMS for fuel cell hybrid electric vehicle: Issues and challenges”, Renew Sustain Energy Rev, vol. 52,(2015), pp. 802–814.

[3] C.Wang , M.Nehrir, “Power Management of a Stand-Alone Wind/Photovoltaic/Fuel Cell Energy System”, IEEE , vol. 23, no. 3, (2008), pp. 957-967.

[4] Haiying Z., Tenghai Q., Shuxiao L, Chengfei Z., Xiaosong L., Hongxing C., “Autonomous Navigation with Improved Hierarchical Neural Network Based on Deep Reinforcement Learning”, Chinese Control Conference, (2019).

[5] Runze L., Junghui C., Lei X., Hongye S., “Accelerating reinforcement learning with case-based model-assisted experience augmentation for process control”, Neural Networks vol 158, (2023), pp. 197-215.

[6] L. Guo, Z. Li , R. Outbib, “Reinforcement Learning based EMS for Fuel Cell Hybrid Electric Vehicles”, IECON 2021 – 47th Annual Conference of the IEEE Industrial Electronics Society, Toronto, ON, Canada, (2021), pp. 1-6.

[7] Chunhua Z, Wei L, Weimin L, Kun X, Lei P, Suk C, “A Deep RL-Based EMS for Fuel Cell Hybrid Buses”, International Journal of Precision Engineering and Manufacturing-Green Technology, (2021), pp. 885-897.

[8] H. sun, Z. Fu, F. Tao, L. Zhu, P. Si, “Data-driven reinforcement learning-based hierarchical EMS for fuel cell/battery/ultracapacitor hybrid electric vehicles”, J. Power Sources, vol. 455, Art. no. 227964, (2020).

[9] H. Lee, C. Kang, Y. Park, N. Kim, S.Cha, “Online data driven energy management of a hybrid electric vehicle using model-based Q-learning”, IEEE Access, vol. 8, (2020), pp. 84444–84454.

[10] Dhawal Khem, Shailesh Panchal, Chetan Bhatt. (2023). Text Simplification Improves Text Translation from Gujarati Regional Language to English: An Experimental Study. International Journal of Intelligent Systems and Applications in Engineering, 11(2s), 316–327. Retrieved from https://ijisae.org/index.php/IJISAE/article/view/2699

[11] Yue H, Weimin L, Kun X, Taimoor Z, Feiyan Q, Chenming L, “Energy Management Strategy for a Hybrid Electric Vehicle Based on DRL”, Applied Sciences, (2018), pp. 1-15.

[12] Tim B, Jens K, Karl T, Robert B, “Integrating state representation learning into deep reinforcement learning”, IEEE, vol. 3, (2018), pp. 1394-1401.

[13] Tianshu Wei, Yanzi Wang, Qi Zhu, “Deep reinforcement learning for building HVAC control”, 54th ACM/EDAC/IEEE Design Automation Conference (DAC), (2017), pp. 1-6.

[14] Bayen, Belletti F., Haziza D., Gomes G., “Expert level control of ramp metering based on multi-task DRL”, IEEE Transactions, no. 99, (2017), pp. 1–10.

[15] Qi X, Luo Y, Wu G, Boriboonsomsin K, Brath J, “Deep reinforcement learning-based vehicle energy efficiency autonomous learning system”, Proceedings IEEE Intelligent Vehicles Symposium , (2017), pp. 11–14.

[16] Hiroshi Yamamoto, An Ensemble Learning Approach for Credit Risk Assessment in Banking , Machine Learning Applications Conference Proceedings, Vol 1 2021.

[17] Tianho Z., Kanh G., Levine S., Abbeel P., “Learning deep control policies for autonomous aerial vehicles with MPC-guided policy”, Proceedings IEEE International Conference, (2016), pp. 528–535.

[18] Mr. Rahul Sharma. (2013). Modified Golomb-Rice Algorithm for Color Image Compression. International Journal of New Practices in Management and Engineering, 2(01), 17 - 21. Retrieved from http://ijnpme.org/index.php/IJNPME/article/view/13

[19] W. Jia, J. Li, Y. Zhao, “DQN Algorithm Based on Target Value Network Parameter Dynamic Update”, IEEE International Conference, (2021), pp. 285-28.

[20] F. Lewis, D. Vrabie, “Reinforcement learning and adaptive dynamic programming for feedback control”, IEEE, vol. 9, no. 3, (2009), pp. 32–50.

[21] Archie C., Zahra R., Gregor V., “Actor-critic learning for optimal building energy management with phase change materials”, Electric Power Systems Research, (2020).

International Journal on Recent and Innovation Trends in Computing and Communication
ISSN: 2321-8169 Volume: 11 Issue: 9s
DOI: https://doi.org/10.17762/ijritcc.v11i9s.7396
***Article Received**: 02 May 2023 **Revised:** 28 June 2023 **Accepted:** 20 July 2023*

[22] Jiageng R., Changcheng W., Zhaowen L., Kai L., Bin L., Weihan L., Tongyang L., "Exclude quotes On Exclude bibliography On Exclude matches Off The application of machine learning-based EMS in a multi-mode plug-in hybrid electric vehicle part II: Deep deterministic policy gradient algorithm design for electric mode", Energy vol. 269, (2023).

[23] Xiaolin T., Jiaxin C., Huayan P., Teng L., A. Khajepour, "Double Deep Reinforcement Learning-Based Energy Management for a Parallel Hybrid Electric Vehicle with Engine Start-Stop Strategy", IEEE Transactions, (2021), pp. 1376-1388.

[24] Yogesh W., Sheetal R., Faruk K., "SoC Estimation of Battery in FCHEVs Using Reformulated Constrained Unscented Kalman Filter", IEEE International Conference on Sustainable Technology for Power and Energy Systems (STPES),(2022), pp. 1-6.

[25] Yuecheng Li , Hongwen He, "DRL-based Energy Management for Hybrid Electric Vehicles. Synthesis Lectures on Advances in Automotive Technology " Springer Cham,(2022), pp. 1-123.

[26] Chunhua Z., Dongfang Z., Yao X., Wei L., "Reinforcement learning-based EMS of fuel cell hybrid vehicles with multi-objective control", Journal of Power Sources vol. 543, (2022).

[27] Zekeriya E., "Reinforcement learning based energy management strategy for fuel cell hybrid vehicles", Sabanci University, (2022), pp. 1- 56.

[28] Xiaowei G., Teng L., Bangbei T., Xiaolin T., Jinwei Z., Wenhao T., Shufeng J., "Transfer Deep Reinforcement Learning enabled Energy Management Strategy for Hybrid Tracked Vehicle", IEEE Access, (2020), pp. 1-11.