Deep Reinforcement Learning Based Transferable EMS for Hybrid Electric Trains

Yogesh Wankhede[1], Sheetal Rana[2], Faruk Kazi[3]
1- Department of Electrical Engineering, Veermata Jijabai Technological Institute, Mumbai, India.
Email: yewankhede_p19@el.vjti.ac.in (Corresponding author)
2- Department of Electrical Engineering, Veermata Jijabai Technological Institute, Mumbai, India.
Email: sprana_m17@et.vjti.ac.in
3- Department of Electrical Engineering, Veermata Jijabai Technological Institute, Mumbai, India.
Email: fskazi@vjti.org.in

Received: 11 June 2023 Revised: 27 July 2023 Accepted: 15 August 2023

ABSTRACT:
The hybrid electric train which operates without overhead wires or traditional power sources relies on hydrogen fuel cells and batteries for power. These fuel cell-based hybrid electric trains (FCHETs) are more efficient than those powered by diesel or electricity because they do not produce any tailpipe emissions making them an eco-friendly mode of transport. The target of this paper is to propose low-budget FCHETs that prioritize energy efficiency to reduce operating costs and minimize their impact on the environment. To this end, an energy management strategy [EMS] has been developed that optimizes the distribution of energy to reduce the amount of hydrogen required to power the train. The EMS achieves this by balancing battery charging and discharging. To enhance the performance of the EMS, proposes to use of a deep reinforcement learning (DRL) algorithm specifically the deep deterministic policy gradient (DDPG) combined with transfer learning (TL) which can improve the system's efficiency when driving cycles are changed. DRL-based strategies are commonly used in energy management and they suffer from unstable convergence, slow learning speed, and insufficient constraint capability. To address these limitations, an action masking 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 is proposed. The DDPG+TL agent consumes up to 3.9% less energy than conventional rule-based EMS while maintaining the battery's charge level within a predetermined range. The results show that DDPG+TL can sustain battery charge at minimal hydrogen consumption with minimal training time for the agent.

KEYWORDS: Fuel Cell, State of Charge, Energy Management Strategy, Deep Reinforcement Learning, Deep Deterministic Policy Gradient, Transfer Learning.

1. INTRODUCTION

Exhaust emissions and their consequences on climate change and health need rethinking transport modalities. The increasing demand for energy-efficient and environmentally friendly transportation has led to the development of innovative solutions such as fuel-cell hybrid electric trains. Several countries including Japan, Germany, and the United States have already started using fuel cell technology to power trains. In India, the Ministry of Railways has also shown interest in this technology and several pilot projects are currently underway as the Indian Railways is one of the largest railway systems in the world and is responsible for a significant portion of the country's transportation emissions. Fuel cell technology has been around for several decades and is a well-established technology for generating electricity. The technology works by converting hydrogen and oxygen into electricity and water with only water and heat being emitted as byproducts. FCHETs are hybrid electric trains that use both a fuel cell and a battery to power the train. The battery is used to store excess energy and regenerative energy and provide additional power during acceleration and other high-power demand scenarios.

The EMS is responsible for the equitable distribution of power from more than one energy source. In this regard, upcoming technologies like fuel cells and auxiliary energy source batteries give the ability to create and achieve railway transportation scenarios that are less carbon-intensive and friendlier to the environment. Therefore, the EMS should refine, design, and control to optimize performance across a variety of use cases. For hybrid electric vehicles and trains, 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,

Paper type: Research paper
DOI: 10.30486/mjee.2023.1988714.1158
How to cite this paper: Y. Wankhede, Sh. Rana, and F. Kazi, **"Deep Reinforcement Learning Based Transferable EMS for Hybrid Electric Trains"**, *Majlesi Journal of Electrical Engineering*, Vol. 17, No. 3, pp. 165-179, 2023.

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. Given that there are often trade-offs to be made between competing optimization goals, it can be difficult to choose which path to take.

1.1. Rules-based EMS and Optimization-based EMS

Various EMS has been developed in the field of automotive like fuzzy logic control gives improved performance as compared to logic threshold strategy but it is observed that the quality of fuzzy control is affected by various control rules and also developer requires skillful knowledge regarding problem, issue of rules comprehensiveness and their persistency [1] . The energy management [EM] algorithm for an FC-battery system uses linear control techniques but the proposed scheme does not handle the load transients effectively, during load transients it disturbs the DC link. The performance of linear controllers strongly depends on system parameters and remains optimal around a specific operation point [2]. Genetic Algorithm (GA) has been applied to FC hybrid systems for electric ships. Depending on collected data from the electric ship driveline model. GA has some drawbacks such as the time for convergence is high there is no guarantee of finding global maxima[3]. The Power management strategy for FC-battery-SC has been investigated through a dynamic programming technique that slows down the system performance and needs more memory [4]. The adaptive control strategy is demonstrated for an FC-battery hybrid system consisting of a single unidirectional boost converter therefore battery operation is not controlled [5]. According to research, the effectiveness of a linear controller is influenced by the system parameter. To address this limitation many researchers have explored the application of nonlinear theory in hybrid systems. One proposed solution is an energy management system that uses an extreme seeking process, which has been tested on a test bench consisting of a fuel cell and a battery. The system's performance can be compared with semi-empirical models using an adaptive recursive least square algorithm with a three-step process [6]. The hybrid model lacks control over battery charging and discharging. To address this issue PBC-IDA approach has been proposed for the interconnection of the ultracapacitor system and the fuel cell in [7,8]. In this study, the outer side loop has a closed-loop port Hamiltonian structure. Sliding mode control and passivity-based control are used to address the DC link voltage control problem. The authors proposed a control strategy that uses sliding mode control principles and linear controllers based on the passivity approach [9]. Additionally, the authors proposed a differential flatness-based controller, which was experimentally validated. The proposed controller works without algorithm computations by projecting the desired trajectory of electrostatic energy stored in the capacitors and considering it as an output component [10]. EMS based on a frequency separation was proposed and validated through a predefined driving cycle and in this research proposes the fixed separation of frequency for different driving conditions is presented, Where a sliding mode controller used to regulate the state of charge of ultracapacitor with the help of fuel cell [11]. Backstepping sliding mode control was implemented for power sharing of fuel cells and ultracapacitor hybrid power systems for the EV's applications. They simulated and verified first the backstepping algorithm and then implemented the same using National Instrument hardware[12]. More aspects are yet to be studied for hybrid energy structure in the areas of absolute stability, robustness, and efficiency.

1.2. Learning-based EMS

Several approaches have been proposed to improve the EM of hybrid electric vehicles [HEV], including the use of past data for online learning [13]. To enhance HEV EM, researchers suggest leveraging instantaneous data obtained from intelligent infrastructures, which can be combined with cloud computing. However, these energy management systems require complex control models and expert knowledge in addition to learning from previous or expected data [14]. Reinforcement learning-based techniques have also been applied to HEV EM, but they require addressing challenges related to sparse, noisy and delayed scalar reward signals and highly correlated states [15]. Learning-based energy management systems offer promising results as they can adapt to diverse driving conditions but with certain drawbacks. Recent studies have proposed online learning control techniques such as neural dynamic programming [16] and fuzzy Q-learning [17] which do not depend on past driving conditions and can self-tune algorithm settings. Deep reinforcement learning has been used to solve complicated control issues and handle vast state spaces as demonstrated by its successful application in games such as Atari and Go [18]. Several studies have proposed DRL-based EM strategies for train traction systems [18], energy-efficient train control systems [19], and high-speed train systems [20]. In these studies, a deep Q-network is used to learn the policy that minimizes fuel use while ensuring a safe and comfortable ride for the passengers. Another study proposed a DRL-based EM system for FC hybrid railway vehicles considering fuel cell aging [21]. In this study, a deep Q-network is used to learn the optimal control policy that minimizes energy consumption while ensuring that the fuel cell operates within safe limits and mitigating the effects of aging.

The research gap is the lack of research on the impact of uncertainties on the performance of DRL-based EMS for FC-based hybrid vehicles. For instance, the battery

capacity, fuel cell efficiency, and driving conditions may vary which can affect the performance of DRL-based EM strategy. Therefore, there is a need for further research to investigate the robustness of DRL-based EMS to uncertainties in the fuel cell-based hybrid train system. Also, there is a need for research on the comparison of the performance of DRL-based EMS with other control strategies. Another research gap is retraining the network after having the driving cycles changed is a time-consuming and tedious process.

This research introduces a DRL-based EMS to address the above-mentioned challenges. The key contributions of this study include:

- Conducting a comparative analysis of DQN, DDPG, and DDPG with transfer learning.
- Developing an intelligent EM strategy for an FC and battery train using modern DRL technology such as DDPG+TL approach. A new reward function is designed to stabilize the training process, which involves battery charge sustaining.
- Proposes action masking technique used to restrict the set of actions that an agent can take in a particular state to prevent the agent from taking actions that are not valid or allowed in a given state
- Creating a stochastic training environment for the railway that simulates real driving scenarios using driving data from Jind to Sonipat (Indian Railway).
- Transferring training to a new domain.

This paper is structured as follows: Section 2 provides a comprehensive modeling of the vehicle, along with the hybrid power system that includes both fuel cell and batteries. Section 3 outlines the framework of the DRL-based EM challenge, while Section 4 analyzes and discusses the results obtained from training and simulation. Finally, Section 5 presents concluding remarks.

2. MODELLING OF ENERGY SOURCES

2.1. FC model

The proposed Dick-larminie electric circuit model is used to model concentration, activation, Ohmic polarization, and Nernst voltage of FC. An electrical equivalent model [22] of FC is introduced in Fig. 1.

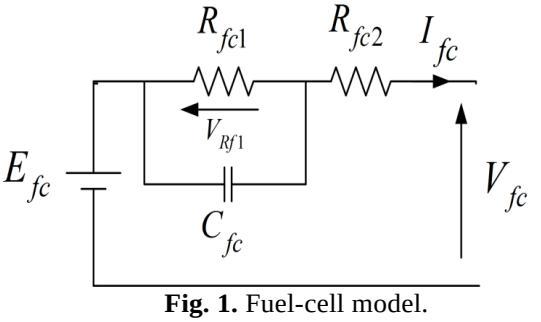


Fig. 1. Fuel-cell model.

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

The equation describes the various components and parameters involved in the output voltage of a fuel cell. The activation and concentration losses associated with the double-layer capacitance are represented by R_{fc1}, while R_{fc2} is associated with the movement of hydrogen and electrons. C_{fc} is the capacitor linked to the dissipation of electronic charges. E_{fc} represents OC voltage, while V_{fc} represents the voltage supplied by the fuel cell to the motor and auxiliaries. The output voltage of the fuel cell can be expressed using the given equation.

$$V_{fc} = n_{cell}[E_{nst} - V_{ac} - V_{con} - V_{ohm}] \quad (2)$$

Where, n_{cell} no of single FC, E_{nst} nernst electromotive force, $V_{ac}+V_{con}=V_{rf1}$ referred to as a vtg reduction due to phenomenon of activation and concentration polarization, $V_{ohm}=R_{fc2}I_{fc}$ ohmic vtg loss. The following is a detailed description of the design for each FC component.

$$E_{nst} = E_{fc} + \frac{\Delta TS}{nC_{fc}} - \frac{R_{fc1}T}{nC_{fc}}\ln\left\{\frac{P_{H_2O}}{P_{H_2}\sqrt{P_{O_2}}}\right\}$$
$$V_{ac} = \frac{R_{fc1}T}{\alpha C_{fc}}\ln\left\{\frac{i_{fc}+i_{loss}}{i_o}\right\} \quad (3)$$
$$V_{con} = \frac{R_{fc1}T}{nC_{fc}}\ln\left\{\frac{i_{\lim}}{i_{\lim}-i_{fc}}\right\}$$

Where, E_{fc} is 0.9 V OC vtg per cell of FC reaction at normal atmospheric force, R_{fc1} is 8.2316 gas constant, T is 333.15K is the FC temp., C_{fc} 87576 is faraday constant, α = 1 x'fer coefficient at atmospheric pressure, P is the pressure exerted by the reactants, i_{fc} is the c/n density. i_{loss} = 2.6 $mA/cm2$ is the c\n loss, i_0 is 0.0043$mA/cm2$ is the exchange c\n density. i_{lim} is 1.7 $A/cm2$ is the limiting c\n density. R_{fc2} is the FC resistance.

$$\dot{m}_{H_2} = M_{H_2}\frac{I_{fc}}{nF} \Rightarrow P_{fc} = \frac{2V_{fc}F}{M_{H_2}}.\dot{m}_{H_2} \quad (4)$$

Where, $\dot{m}_{H_2}$ - consumption rate of hydrogen, M_{H_2} - Hydrogen's molar weight, P_{fc}- o/p power of the FC. As shown in Fig. 4 unidirectional DC-DC boost converter is responsible for energy transfer from the fuel cell to the inverter also it's responsible for to adjust

boost converter voltage level as per the common dc link voltage. The DC-DC converter model for FC is

$$P_{fc} = P_{fc}^{'} / \eta_{dc}(P_{fc}^{'}) + P_{aux} \quad (5)$$

Where, $P_{fc}^{'}$ is o/p power of the FC. Assume this power is requested by control strategy, n_{dc} DC-DC converter efficiency for FC, P_{aux} auxiliary system assumes as a constant c\n load I_{aux}= 2 amp. The FC parameter is n_{cell} = 200, FC effective electrode area is A_{fc}=324cm^2, the force of anode hydrogen is 4 kPa, and oxygen used at the anode comes from the atmosphere. Fuel cells like many other forms of energy sources, have varying degrees of efficiency depending on the load. 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 efficiency begins to drop. A diagram depicting fuel cell performance in relation to power demand is presented in Fig. 2.

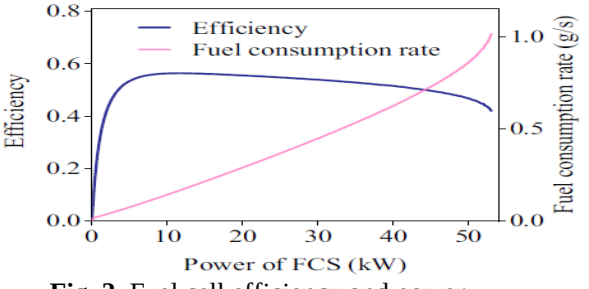


Fig. 2. Fuel cell efficiency and power characteristics.

2.2. Battery Model

To obtain accurate state-of-charge estimates, it is crucial to have precise battery models. There are three types of battery models - mathematical, electrochemical, and electrical equivalent circuit models. While mathematical models are relatively easy to calculate, modelling the external properties of the battery mathematically can be challenging. On the other hand, electrochemical models are highly accurate but their complex structure makes them unsuitable for modelling real-world operating conditions. In this paper 2-stage RC circuit which is equivalent to the battery model is considered [25].

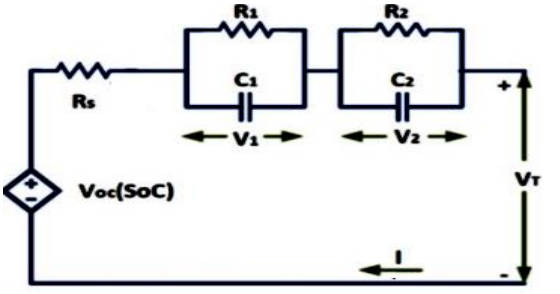


Fig. 3. Two-stage RC equivalent circuit.

$$\frac{dV_1(t)}{dt} = \frac{-V_1(t)}{R_1C_1} + \frac{I(t)}{C_1} \quad (6)$$

$$\frac{dV_2(t)}{dt} = \frac{-V_2(t)}{R_2C_2} + \frac{I(t)}{C_2} \quad (7)$$

$$\frac{dSoC(t)}{dt} = -\frac{\eta I}{Q} \quad (8)$$

(Q is the rated-capacity of the battery, η - coulombic efficiency) By applying Kirchoff's voltage law to the Fig. 3.

$$V_T(t) = Voc(Soc(t)) - R_sI - V1 - V2 \quad (9)$$

Eq. (6), (7), and (8) represent linear state equations and because of $Voc(Soc(t))$ term in eq. (9), eq. (9) is nonlinear o/p eq. The nonlinear system is linearized at each time step by utilizing Taylor's series expansion around the SOC operating point SOC0.

Battery o/p power

$$P_{bat} = V_{oc}(SOC_{bat})I_{bat} - I_{bat}^2R_{bat}(SOC_{bat}) \quad (10)$$

Where, V_{oc} - battery open circuit vtg, I_{bat} - battery o/p current, R_{bat} - battery internal resistance and battery efficiency is

$$\eta_{bat} = \left\{\frac{V_{oc}(SOC_{bat}) - I_{bat}R_{bat}(SOC_{bat})}{V_{oc}(SOC_{bat})}\right\}_{(I_{bat}>0)}$$

$$\eta_{bat} = \left\{\frac{V_{oc}(SOC_{bat})}{V_{oc}(SOC_{bat}) - I_{bat}R_{bat}(SOC_{bat})}\right\}_{(I_{bat}<0)} \quad (11)$$

V_{oc} and R_{bat} are two empirical functions of SOC and battery output power is expressed as

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

$$P_{bat} = \left\{P_{bat}^{'}.\eta_{bdc}\right\}.....(P_{bat}^{'} < 0) \quad (12)$$

Where, $P_{bat}^{'}$ is the o/p converter power with efficiency is $n_{bdc,}$

2.3. The Train Performance Model

Consider a FCHETs driving at v on a track with gradient θ and consider basic sources of resistant force, aerodynamic drag, rolling resistance, gravitational force, transient force.

$$F_m = F_{air} + F_f + F_s + F_a \quad (13)$$

$$= \frac{1}{2} C_D A \rho v^2 + Gf \cos\theta + G \sin\theta + m\frac{dv}{dt} \quad (14)$$

Where, F_m- driving force delivered by motor, F_{air} - air resistance(1.2), F_f rolling resistance (R1= 0.0018, R2=0.000016), F_s- slop resistance, F_a - acceleration resistance, ρ - air density coefficient(1.184 kg/m3), C_D air resistance coefficient(0.26), A - windward surface volume of the vehicle(15.33 m2), v - vehicle velocity, m - vehicle mass(170551 kg) 3300 passenger have an average weight of 50 kg including luggage weight, G-gravity of the vehicle(9.8m/s2), f - vehicle slideing resistance coefficient (0.4), track grade $\theta = \tan^{-1}(\Delta k / \Delta n)$ where, k represents altitude and n represents distance for the drive cycle route. The requested power for the FCHETs

$$P_{train} = F_m . v / \eta_m \quad (15)$$

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

$$P_{train} = P_{fc} + P_{bat} \quad (16)$$

Table 1 shows the specification of the modified DEMU train (modified in the term of existing DEMU train without conventional fuel weight but assuming hydrogen tank, fuel cell, and battery pack weight) and The requirements for the design of a hybrid power train system that operates between and Jind and Sonipat (India) using hydrogen fuel cells. Drive Details: 1) The overall distance traveled by the Sonepat-Jind section of Northern Railways is 89 kilometers. This stretch features 12 stations. Approx. 300 m above sea level 2) Indicative power train component rating based on genuine drive cycle between Sonepat and Jind, including operational margins for driving cycle changes that are part of normal train operation:

a. 800 KW fuel cell-based power stacks, especially in 50-kW increments (KW).

b. Secondary energy source, such as a 400 KW, 330–380 KWH battery bank. The drive cycle of the Modified DEMU (Diesel Electric Multiple Unit) that operates between Jind to Sonipat based on some assumption is shown in the result section.

Table 1. Modified DEMU train specification.

Parameter	Range
Train Length	195 m (12 coach)
Track gauge	1676 mm
Train Height	1434.7 mm
Train Weight	170551 kg
Seating capacity	790 seating, 2510 standing
Maximum Acceleration	0-60 km/h -0.9 m/s^2 60-120 km/h – 0.08 m/s^2
Maximum Deceleration	120 – 60 km/h – 0.9 m/s^2 60 – 0 km/h – 1.0 m/s^s
Maximum Speed	120 km/h
Power output	1200 KW

Fig. 4 shows the configuration of FCHETs and in this configuration main power source is the fuel cell and an assisting power source battery is used. By implementing a unidirectional power converter, the fuel-cell was linked to the common DC link whereas a bidirectional power converter was utilized to connect the battery to the same common DC link.

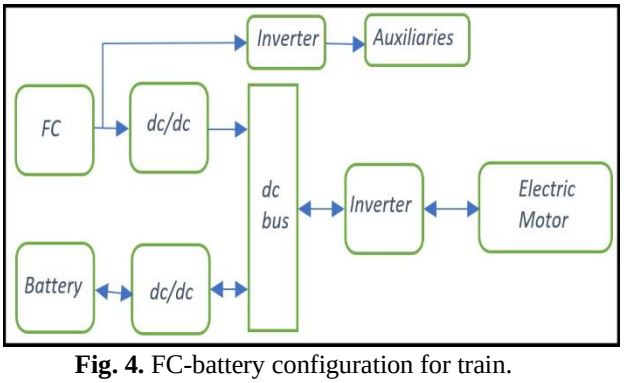


Fig. 4. FC-battery configuration for train.

3. DRL ALGORITHMS BASED TRANSFERABLE EMS FOR FCHET'S

3.1. Implementation of DRL-based Energy Management Strategy

Reinforcement learning [RL] maps states to actions to maximize future cumulative rewards. as shown in Equation 17. The hybrid drivetrain is the controllable object in FC-HETs. Powertrain condition, driving conditions, driver needs, etc., might all stand in for the state (s) at time t. At time t action (a) is the energy distribution system. Metrics like instantaneous fuel consumption, SOC variations, etc. may be monitored in real time and used to calculate the reward r for an energy

distribution system. The EMS is the name of this method of regulation:

$$\pi^* = \arg\max E_{a(t)\in A}\left[\sum_{t=0}^{N-1} r(s(t),a(t))T_s\right]_{s(o)=s_0} \quad (17)$$

Where, pi represents optimal EM strategy, A represents action space, T_s represents sampling time, s_0 represents the starting state and N represents the time sequence length of Markov decision process problem. The proposed EMS which is based on reinforcement learning explores the train and its driving environment and receives feedback to learn the most efficient way to distribute energy. The essential components of this interactive learning process are discussed below.

i) The Agent and EMS Pi - Within reinforcement learning, the agent makes all of the calls. Its function in energy management is analogous to that of a train controller, which regulates power flow in real-time. The EMS is the controller's energy management control software utilized to translate between status s(t) and action a(t).

ii) The environment - In the context of FCHETs EM, the regulated system is analogous to the hybrid powertrain and the environment is the representation of this environment. As soon as the controller (agent) issues an action signal, the powertrain will react accordingly. This reaction may manifest itself in the form of immediate fuel usage, a shift in SOC, etc.

iii) State space - Agent's state is a description of its current surroundings. Train features, track and traffic signal circumstances, driving demand, etc. are all represented by a set of state vectors s(t), s(t) belongs to s at t equals to 1,2....upto N. It's important that the s(t) accurately reflects the condition of the train and the driving conditions and that it's straightforward to inspect. In most cases, we may use the train current velocity v, the percentage of battery life, notch state (6 Notch used in DEMU), etc. to represent the motor force, which is primarily influenced by the longitudinal dynamics of the train. The motor power demand may be represented by several signals such as the notch signal representing the (a) acceleration, the estimated driving torque or power requirement etc. It is possible to learn about the driving cycle from prior velocities, such as the speed in the preceding in second: v-01, v-02, upto v-k. States that included in the state space as

$$S = \{v, SOC, a, P_{demand}, \theta_{slope}, SOC_{ref} v-1, v-2, v-N)$$

iv) The action space A - A control action by an agent is meant here. All conceivable a(t) (action vector) that define the energy distribution on the drive cycle are included in action space A for all actions. $a(t) \in A, t = 1, 2, upto....N$

v) State transition $s \to s'$ - The current reaction of the controlled object following the execution of action triggers a transition in the environment from its present state to the subsequent states. Markovian features may be seen in this process as the state transfer.

vi) The reward R - When a state changes a reward signal is sent to indicate how successful the new approach reward will be

$$r = -\tanh(\alpha \dot{m}_{H_2} + \beta \mid \Delta SOC_{ref} \mid^2) \quad (18)$$

Here, α and β represent the fuel utilization rate and SOC deviation. A high beta means that a higher indicates a greater reward for carefully following the SOC's recommendations. Larger alpha shows decreasing fuel consumption can yield significant rewards. The primary goal of fine-tuning alpha and beta is to maximize fuel efficiency while guaranteeing that the trained strategy can always satisfy SOC prerequisites. After repeated tuning the value of beta was set to 50 and values of alpha was set to 150 and SOC_{ref} was 0.65 as affixed initial SOC. Equation 18 demonstrates that the secret to identifying the best course of action is learning to choose a course of action that has a high anticipated reward return.

vii) Q value - As shown by Equation 19, the action-value function in markov decision process issues is designed to express the anticipated reward return over time after acting on action a:

$$Q_\pi(s,a) = E_\pi\left[\sum_{t=0}^{N-T} r(s(t),a(t))\gamma^t \mid_{s(0)=s_T, a(0)=a_T}\right] \quad (19)$$

Here, s_T represents state, and a_T represents action at time T, γ discount rate represents calculating the current value of future reward (γ =0.92 to estimation efficiency for balance) The best strategy π^*has the largest action value Q^*=(s,a) is

$$Best\_action\_value \to Q^*(s,a) = \max \pi Q_\pi(s,a) \quad (20)$$

If Q^*=(s,a) has been gained, the best strategy as shown in eq. 17 reformulated as

$$\pi^* = \arg\max_{a(t)\in A}\left[Q^*(s(t),a(t)) \mid s(0) = s_0\right] \quad (21)$$

As per bellman equation, To resolve the optimal decision process described in equation 22, equations 20 and 21 need to be broken down into a series of single-step decisions.

$$Q^*(s(t),a(t)) = E\left[r(s(t),a(t)) + \max_{a(t+1)\in A} Q^*(s(t+1),a(t+1))\right] \quad (22)$$

Reinforcement learning strategies are developed keeping the Bellman Equation in mind. Consider the value-based RL method in which the action-value function is modified repeatedly in combination with agent-environment interaction. Based on the probability of an actual action value as indicated in Equation 23, in order to estimate error target estimate to become 0 the future action value Estimate is updated repeatedly by a set step Step Size.

$$Estimate_{new} \leftarrow\mid Estimate_{old} + StepSize(T\arg et - Estimate_{old}) \quad (23)$$

If the action value is updated immediately by Equation 19, however, the solution procedure becomes very inefficient since it must retrace the full control sequence. Therefore, the concept of Temporal Difference (TD) update is widely used in reinforcement learning to speed up the learning procedure of Q(s,a) estimate, as illustrated in Equation 24.

$$Q_{new}(s,a) \leftarrow Q_{old}(s,a) + \alpha\left[(r + \gamma \max_{a'} Q_{old}(s',a')) - Q_{old}(s,a)\right] \quad (24)$$

Where, $(r + \gamma \max_{a'} Q_{old}(s',a')$ is TD error (Temporal Difference δ), $Q_{old}(s,a)$ represents the $Estimated_{old}$, $(r + \gamma \max_{a'} Q_{old}(s',a')$ 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)). Iteratively solve the Bellman eq. to get the optimum EMS eq. 21.

Viii) FC-HETs Energy Management - FC-HETs energy management strategy is developed to obtain an optimal distribution between sources and lood. FCHETs EM represents the control signal level given to DC/DC converter which is the turn control motor.

- It simply imposes a maximum and minimum limit to the storage elements' energy or SOC
- As shown in Fig. 5. SOC_MIN and SOC_MAX are two limits for battery and FC_eff_MIN and FC_eff_MAX are two limits for FC
- The general idea behind choosing limits for storage element energy is to make its SOC in an acceptable functioning area.
- For FC system its power limits offer a supplementary degree of freedom serving to optimal the FC functioning that is fuel consumption minimization.

In these three train operational modes are possible

1) Stop mode and start mode – The train is powered by a battery

2) Traction motor mode – Balancing FC and battery as per SOC

3) Braking mode – Generate regenerative energy.

The goal of EMS is to maintain the SOC at the optimal interval while minimizing fuel use. All that has to be constructed are the cost function, state variable, and suitable action variable. In this work, we choose to have the constant power from the FC be the only action variable, while the state variables include the speed, acceleration, and state of charge of the battery. In this paper we assume the regenerative energy store in the battery. The following inequalities must be met to ensure the components work safely and dependably.

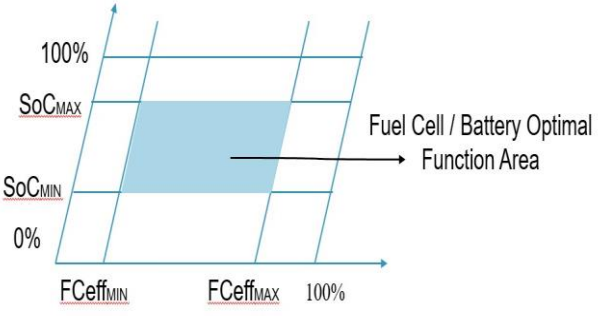


Fig. 5. FC-battery optimal function area.

$$\begin{aligned} &P_{Demanded,\min} \le P_{Demanded}(t) \le P_{Demanded,\max} \\ &SOC_{\min} \le SOC(t) \le SOC_{\max} \\ &FC\_eff\_MIN \le FC\_eff(t)\_MAX \end{aligned} \quad (25)$$

3.1.1. Deep Q Learning (DQN) Algorithm Based EMS

In Q learning Q-value is remembered by a Q-table, where state and action are the dimensions. The Bellman equation includes the temporal difference approach. The eq. is the fundamental building block of the learning process and its parameters were modified for this research. Here is how to determine the Q value by taking into account the time difference.

$$Q^{new}(s_t, a_t) \leftarrow Q(s_t, a_t) + \alpha(r_t + \gamma.\max_a Q(s_{t+1}, a) - Q(s_t, a_t)) \quad (26)$$

When the Q-value can no longer increase because the optimal set of actions for each state has been discovered and the table is stopped being updated with the learning rate multiplier marked by α. It's possible that greedy goals are permanently stuck in some actions. The most typical issue is the trade-off between discovery and development that is outlined by ϵ. The exploration rate increases as ϵ defined between 0 and 1, leading to more random action selection and a greater emphasis on future rewards. Algorithms tend to be more acquisitive and prioritize reward when the value is close to zero. The major drawback of the method is that it necessitates a Q-table with a size equal to the product of the no. of states and the no. of actions. Instead of using a Q table, DQN employs a deep neural network (DNN) to handle issues with a large state-action space quickly or at all. DNN can

benefit from a number of hidden layers to allow for the discovery of more complicated associations. All of the nodes in each of those levels are interconnected. Fig. 6 depicts the network in its simplest form, however, the connections might vary depending on the application. The activation function's parameters are adjusted at each node to provide an accurate mapping between input and output. The no. of nodes and the no. of layers both influence the precision with which mapping or fitting may be accomplished. Even though there is no foolproof method for determining which values will get the greatest outcomes, the computing time grows dramatically with the number of layers and nodes which may be deceptive in terms of convergence. In such a scenario it may take a long time. Although defining complexity is not easy. Algorithm 1 shows the DQN algorithm for EM in FCHETs.

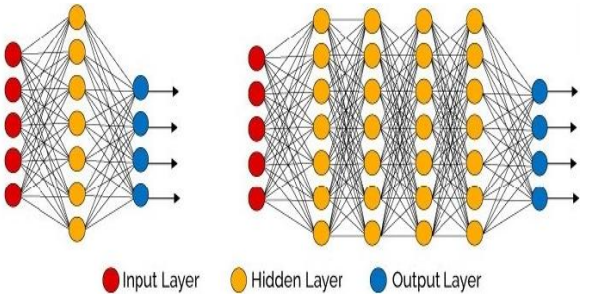


Fig. 6. Neural network and Deep learning neural network.

Algorithm 1 Deep Q Learning

perform an initialization of replay memory D with capacity N.
perform some random weighting at the beginning Q (action-value function).
for (loop) epi = 1, M do
perform an initialization and preprocessed sequenced as $s1 = \{x1\}, \varphi 1 = \varphi(s1)$ resp.
for (loop) t = 1, T- do
along with probability ϵ perform the selection of a random-action and else select $a_t = \max_a Q^*(\phi(s_t), a; \theta)$
perform action a_t and notice reward r_t and image x_{t+1}
set $s_{t+1} = s_t, a_t, x_{t+1}$ and execute $\phi_{t+1} = \phi(s_{t+1})$
Store the transition $(\phi_t, a_t, r_t, \phi_{t+1})$ in memory D
Take sample random minibatch (transitions) $(\phi_t, a_t, r_t, \phi_{t+1})$ from memory D
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}
Execute a gradient descent step on $(z_j\ Q(\phi_j, a_j; \theta))$
end for (loop)
end for (loop)

DQN utilized the targeted value function, in that the total of just the current and future rewards is taken into consideration as opposed to the temporal difference technique in the value function and the subsequent storage of this Q value in a table. After that, the network is refreshed using gradient descent on the loss function. DQN yields a form of action that specifies what percentage of requested energy will be met by the battery. The converter's switches manage the flow of electricity. The gain of the converter model chosen for action F_{bat}, F_{FC}. In this research, one of F_{FC} and F_{bat} shall be used as an action variable. One is dependent and the other an independent variable, and their total is constant. Gains and battery current [23] are related as shown in the following equation.

$$I_{bat} = \frac{F_{bat}}{F_{bat} + F_{FC}} * I_{in} \tag{27}$$

The system is resilient if and only if the total of F_{bat} and F_{FC} is bigger than one. As per Table 3 Computational investigations show that adjusting this sum enhances system responsiveness, hence the value of 5 is chosen.

$$a = \{F_{bat}\} where\_F_{FC} = 5 - F_{bat} \tag{28}$$

After some trial and error, the following was determined to be the optimal range for the action. The optimal range is divided into steps (step size = 0.5).

$$-3 < F_{bat} < 5 \tag{29}$$

Table 2. Modes for power utilization.

Mode	Operation	F_{FC}	F_{bat}
Mode 1	FC charges the battery and also provides power to the motor and auxiliaries	7	-3
Mode 2	Only FC provide power to the motor and auxiliaries.	5	3
Mode 3	FC and battery supplies equal current to the motor	3	3
Mode 4	Only Battery supplies power to the motor	0	5

Requested energy P_{demand}, battery derived energy P_{bat} and FC energy P_{fc} , FC efficiency FC_{eff}, SOC of the battery, and deviation of remaining charge available in the battery is $SOC_{desired}$ are all possible states in the problem. They are the most likely possibilities for the variables, however, they can be specified in other ways. Many combinations are tested out during training before the reward maximization is achieved. For DQN, the following are the states that can be in :

$$S = \{SOC - SOC_{desired}, P_{demand,}\} \tag{30}$$

The system has no bounds because the state P_{demand} is the input. However, there are constraints on the initial state if SOC is less than $SOC_{desired}$, as shown by the following equation.

$$-SOC_{diff,\lim it} < SOC - SOC_{desired} < SOC_{diff,\lim it} \quad (31)$$

One version of the objective function is represented by the reward function.

$$r = -\tanh(\alpha \dot{m}_{H_2} + \beta \mid \Delta SOC_{ref} \mid^2) \quad (32)$$

If the state goes over the limit depicted by equation 22, the simulation is terminated and the agent is punished in the form of a penalty so that it will learn to avoid this behavior in the future. The network loss function is described as:

$$loss\_function = (r + \gamma \max_{a'} Q(s', a'; \theta') - Q(s, a; \theta))^2 \quad (33)$$

3.1.2. Deep Deterministic Policy Gradient-based EMS

Since DDPG consists of an actor-critic network both the actor and the critic have their own independent networks which are used to evaluate and critique the other. The assessment network takes states as input and produces an action as output. By using a deterministic policy gradient, the approach produces a deterministic action from the actor network rather than a probability of actions. Since this is a problem with continuous actions and states, the DDPG algorithm can solve it. The critic network takes in states and their associated behaviors as inputs and produces a Q-value as an output. The noise strategy is adopted to boost exploration and the replay experience pool method is employed to minimize data correlation during training while the utilization of priority experience replay [26] further aids in reducing the duration of training unlike DQN, DDPG uses a soft update for its parameters. There is no clear separation between the DDPG algorithm's action space and the action space described for DQN. On the other hand state space is a little different. Once again, the state variables are selected as follows after multiple training episodes.

$$s = \{SOC, H_{2,eff}\} \quad (34)$$

Similarly to DQN, DDPG has a finite number of state variables. These constraints apply to the SOC (Battery) and $H_{2,eff}$ (FC efficiency) variables as follow,

$$SOC_{\min} < SOC < SOC_{\max}$$
$$H_{2,eff\,\min} < H_{2,eff} < H_{2,eff\,\max} \quad (35)$$

The amount of electricity that a fuel cell can produce is proportional to its efficiency. The agent receives a reward while in the condition when it is observing the efficiency value but is unaware of the amount of electricity delivered by the fuel cell. The agent's knowledge of fuel cell power is superfluous, all they require is the efficiency curve. Therefore, it makes little difference whether the power is shifting to the left or the right as long as it is being used effectively. The other variable of state SOC is used to determine the direction of the power differential. The definition of the reward function is almost similar with one small change defined as follows:

$$r = -\tanh(\alpha \mid \Delta \dot{m}_{H_2} \mid^2 + \beta \mid \Delta SOC_{ref} \mid^2) \quad (36)$$

The simulation will end with the agent receiving a penalty if state boundaries are violated. DDPG algorithm with action mask as algorithm 2.

Algorithm 2 DDPG

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

Assign target network Q' and u' with weights $\theta^{Q'} <-- \theta^Q, \theta^{u'} <-- \theta^u$

Assign buffer R

for loop for epi = 1, M do

set N for action exploration as a random process

for t = 1, T do

Select the action for the acc. of the policy and for exploration of the noise $a_t = u(s_t/\theta^u) + N_t$ with clip function $P_{FC}(t)$ = clip [$P_{FC,}$ Pmin FC (t), Pmax FC (t)]

Execute a_t and observe r_t and observe $s_{t+1,}$ i.e Action reward for new state resp.

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

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

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

$L = \frac{1}{N}\sum i(z_i - Q(s_i, a_i/\theta^Q))^2$

Update: at critic using minimizing the loss

Update the sampled policy gradient: as a updating of Actor_policy

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

Update the target networks:

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

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

end for

end for

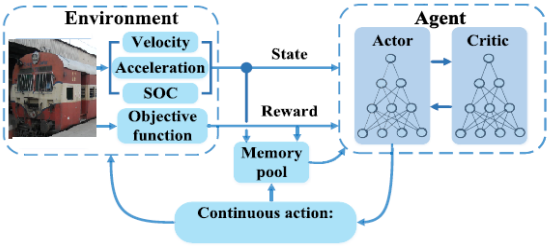


Fig. 7. FCHET's Agent environment interaction.

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 FCHETs experts with the DDPG algorithm, this study can determine the best course of action for EMS. Agent environment interaction of FCHETs EM as shown in Fig. 7 also is known as the interface between the EM system and the train. DDPG-based EMS rewards are based on the immediate use of fuel cell and battery charge sustaining costs are based on this two-point multi-objective reward function.

$$\begin{aligned} state\_vector &= \{SOC, velocity, acceleration\} \\ action &= \{continuous\_power\} \\ reward &= -\{\alpha[\, fuel(t) + \beta[SOC_{ref} - SOC(t)]^2\} \end{aligned} \tag{37}$$

Here, alpha represents the weight of hydrogen consumption, beta represents battery charge weight, SOC_{ref} reference value for maintaining battery charge. Parameter tweaking between alpha and beta is a significant obstacle to the multi-objective reward function. Parameter tuning's primary goal is to maximize fuel efficiency within the constraints of the battery's needs. In consideration of the features of the battery the reference state of charge (SOC_{ref}) has been set at 0.6 based on the minimum charge-discharge internal resistance and attaining high battery efficiency requires the SOC to operate within the predefined upper and lower limits.

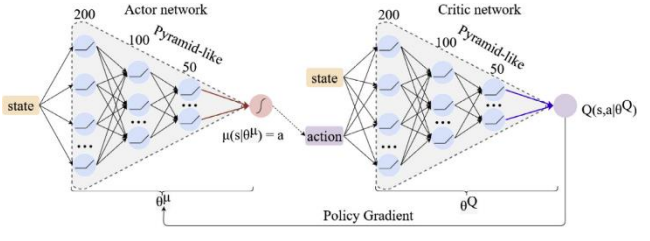


Fig. 8. Architecture of actor-critic network.

Fig. 8 further demonstrates how the actor-critic network architecture was developed specifically for the EMS. Pyramidal in shape and the size of a neural network reduces as it progresses through its layers. As part of the DDPG framework, prioritized experience replay (PER) is implemented to make the most efficient use of the memory pool for learning and to speed up the convergence process. As opposed to randomly replaying experiences, PER prioritises re-visiting key observation data based on the severity of errors in order to boost RL's learning efficiency. Similar to DDPG in that it uses the same set of state variables and reward functions for its analyses. The ideal efficiency curve of the fuel cell is used to determine the increment or decrement in value specified as action variables.

3.1.3. Action masking for DDPG

As compared to other deep reinforcement learning-based EMS the DDPG randomly discovers the whole action space to learn how to limit the chance of restriction into a local optimum, which would result in actions that are outside of the fuel cell's practical operating range. Hence, action masking must be created in order to filter out invalid actions and stop DDPG from engaging in pointless learning exploration. Given that the main goal of DDPG is to create long-term planning strategies through the distribution of states, distribution of actions, and transition of states in the learning environment. For Action Masking to avoid violating the DDPG algorithm's guiding principles, it must adhere to two criteria. The first is that AM would not alter the initial action space distribution in order to prevent the destruction of the environment's potential state transfer probability function. The second is that incorrect samples will not be used in the training since they will not be gathered into the experience replay buffer. The following procedures are specifically repeated by DDPG to apply AM at each time step t.

The following three stages are used to determine the appropriate working range at each time step t.: (1) Action = $\{P_{FC} | P_{FC} \in [0$ KW power, Max power in KW$]\}$ this discretized to form a(t); (2) Compute the fuel cell maximum and minimum power at time t by traversing a(t) as per the dynamics of the driving cycle and (3) Obtain new fuel cell max and min power at time t. afterword's, the FC energy $P_{FC}(t)$ output form actor network in the DDPG based algorithm by the clip operation $P_{FC}(t) = \text{clip}\,[P_{FC,}\, P_{min}\ \text{FC}\ (t), P_{max}\ \text{FC}\ (t)]$, due to the fact that the clip function does not alter the initial action A, there is no impact that it can have on the DDPG.

It is crucial to point out that the initial step is employed by a broad variety of mathematical model-based techniques, all of which eliminate erroneous actions by traversing the A-action space and it is also necessary to note that this step is extensively used. In addition to this, the action masking strategy has to be implemented for both the actor, target actor networks. If this does not occur, the algorithm's capacity for learning will be severely compromised. In addition to this, the method of concealing faulty actions via the use of the clip

function is only relevant to algorithms such as DDPG that are founded on the actor critic and deterministic policy.

3.1.4. Transfer Deep Reinforcement Learning

In this section, a solution to the issue of EM for a FCHET's is developed in the form of a bi-level control structure and first and foremost, the process of training is described. In order to learn the EMS [24] the driving cycles are broken up into intervals of three different speeds. After that, the deep Q network and DDPG are implemented in order to look for the best possible control strategy for driving cycle that was examined. In conclusion, the use of transfer learning is implemented in order to improve the speediness of the convergence of the training process.

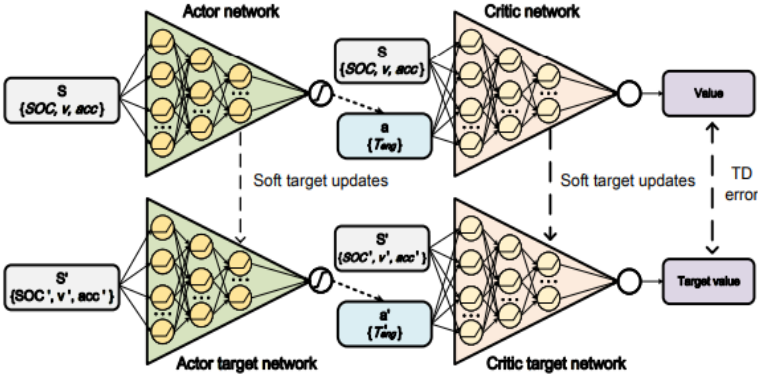


Fig. 9. DRL and TL-based EMS structure.

Fig. 9 shows a DDPG-based control structure used to build an optimum control methodology. The train speed dataset comes from a real-world train driving cycle. To shorten training time and increase control precision, train speed is separated into three speed intervals: [Speed Interval 0-39], [Speed Interval 40-79], [Speed Interval 80-120] km/hr. The categorized speed slots are used to train the DDPG algorithm individually until convergence, as shown in Fig. 10. Lower level control process applies obtained EMSs and TL method applies the already trained network to the new driving cycle domain. A unique drive cycle's best control method is derived in seconds.

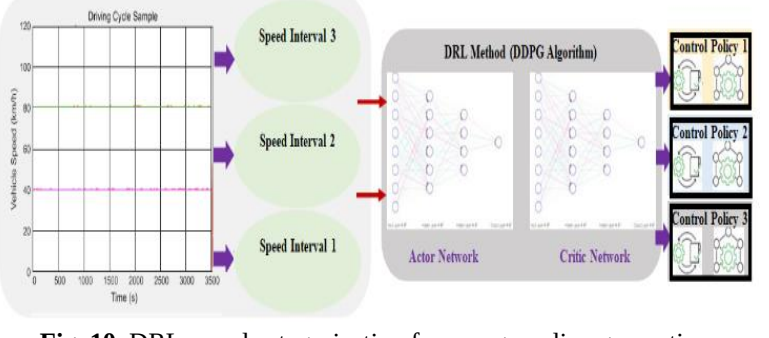


Fig. 10. DRL speed categorization for energy policy generation.

The standard deep learning technique handles testing and training data from the same drive cycle. When the criterion isn't met, rebuilding the model and retraining the data is expensive and time-consuming. TL can help solve this issue. As the two study problems are similar the TL may reuse most neural network settings. For this, TL method has a source domain (Sx) and a learning task (Tx), as well as a target domain (Sy) and a corresponding task (Ty). Applying what was learned in one setting (Sx) to another (Tx), where Sx Sy or Tx Ty is an example of transfer learning. This is done so that the goal prediction function (f) in Sy may be learned more effectively. The EMS issue in FCHETs is investigated in this study by employing the DDPG algorithm. Developing a training model for a brand-new driving cycle might be a time-consuming process. The driving cycle's train speeds are first divided into three categories based on TL theory. The EMS is trained for various speed ranges using the DDPG method and the corresponding parameters are then stored in memory as shown in Fig. 10. Since the

FCHETs driving cycles share the same feature space and are correlated, it follows that they must all be driven in the same way. The learned parameters are then applied to the new driving cycle at various speeds. This would allow the EMS to be created effectively and its optimality could be ensured. The result section examines the simulation findings and draws conclusions about the efficiency of the DDPG and TL-enabled EMS.

Table 3. H parameters for DDPG training algorithm.

Name of the parameters	Parameters value
Experience Buffer Length	1e6
Critic Learning Rate	1e-4
Simulation Time	15
Mini Batch Size	256
Sample Time	0.1
Actor Learning Rate	1e-4
Agent Noise Variance	0.1
Agent Noise Variance Decay Rate	1e-3
Discount Factor	0.99

4. RESULTS AND DISCUSSION

This work discusses RL-based EMS implementation and evaluates its optimality, flexibility, and collative efficiency for FCHETs. DQN and DDPG based action mask optimum control policies are used to evaluate DDPG + TL methods. Comparing the DDPG + TL and DDPG methods determines Q value table convergence rate. DDPG and TL-enabled control methods validate adaptability to the new driving cycle domain. In the Q framework, the exploration rate is gradually reduced from 0.9 to 0.002 to achieve the desired outcomes. A learning rate of 0.01 and a decay rate of 0.9 are used in this scenario. The state variables and the control variables both have a discrete form. P_{train}, SOC_{bat}, and P_{fc} each have a step size of 10 kW, but P_{fc} only has a step size of 10%. Under the conditions, a total of one thousand episodes are performed. The DQN and the DDPG are used as baseline measures in order to determine whether or not the suggested EMS that is based on TL and DDPG is the most effective one. The DQN was able to create the globally optimum control actions, thus the differences that exist between the technique that was described and the DQN are utilized to measure the degree to which the offered strategy is optimal. The default settings for both DDPG+TL and regular DDPG are entirely identical to one another. In order to facilitate this comparison a standardized train driving cycle known as the Jind to Sonipat is being used. Also By restricting the set of actions that the agent can take, action masking can help focus the learning process and improve the efficiency and effectiveness of the EMS

Fig. 12 depicts the output power values of FC and batteries, along with the power requirements of the FCHETs. Additionally, Fig. 13 displays the cumulative state-of-charge trajectory of batteries equipped with DDPG + TL, DQN, and DDPG under the Jind to Sonipat driving cycle, as driving cycle shown in Fig. 11. The terminal SOC values, on the other hand, are almost identical. The changes in the state of charge further indicate that the DDPG+TL method outperforms the conventional DDPG with regard to the battery's power delivery capacity. Fig. 12 analyzes the power distribution between the FC and the battery, demonstrating the control-level advantages of the proposed technique. The DDPG and DDPG+TL display similar patterns of variation. However, the DQN approach differs from the other two approaches, and thus fuel efficiency performance may vary. With transformed learning derived from previously acquired knowledge of energy management, the DDPG+TL technique has the potential to minimize fuel consumption and ensure optimal performance.

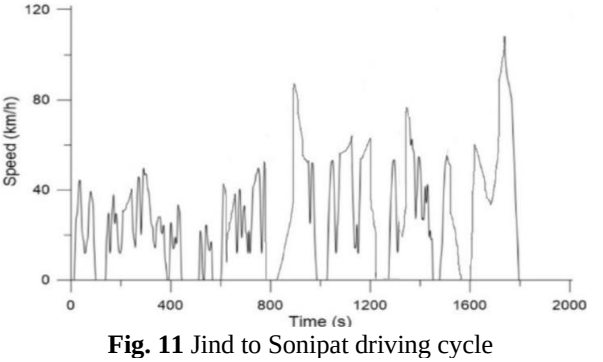
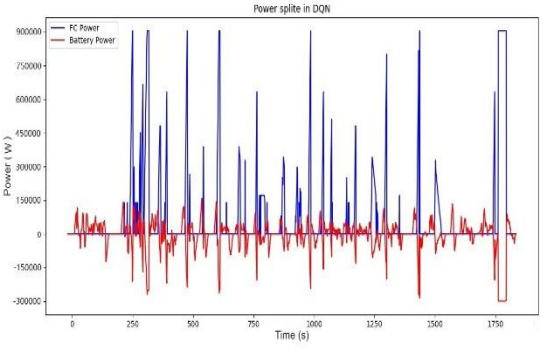


Fig. 11 Jind to Sonipat driving cycle



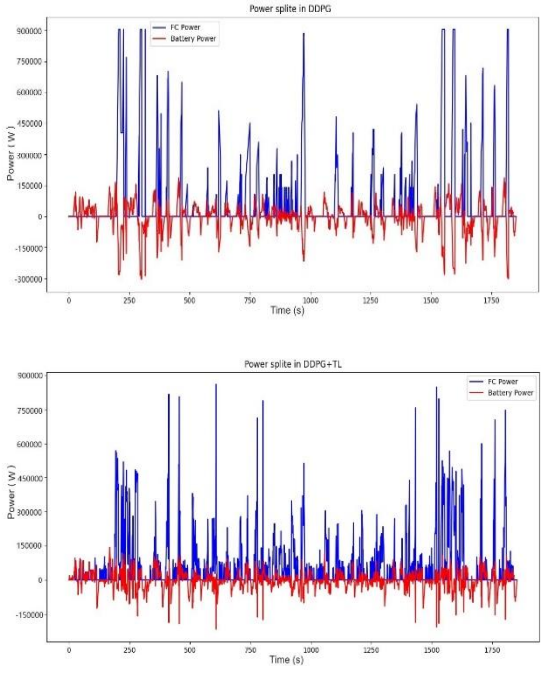


Fig. 12. Power distribution between FC and battery in three EMS models.

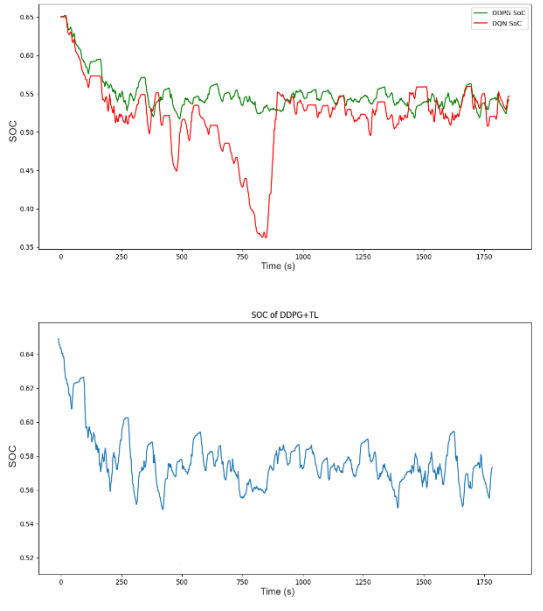


Fig. 13. SOC trajectories of three EMS models.

At the time of the training process, the DDPG + TL is used to speed up the convergence process by repeatedly playing significant samples from the experience pool at a higher frequency. As shown in Fig. 14, a comparison is made between the tendency of the Mean reward throughout training for the situations of DDPG + TL, DQN, and DDPG. This is done so that the efficacy of the DDPG may be verified. It has been discovered that the DQN method converges sometime in the range of 600, 800, and 1000 rounds, while it requires further episodes in order to achieve 0 mean. While the DDPG and DDPG + TL start to converge with around 50 and 70 rounds respectively, the proposed DDPG + TL EMS demonstrates superior convergence performance. Also, the DDPG plus TL is much better than the regular DDPG for the same episodes. Since both methods employ the same random seeds, the reward values for DDPG and DDPG -TL in the beginning episodes are quite similar to one another. This is the effect of the same random seeds being used. This finding suggests that AM has no effect on the learning performance of DDPG and as a result, DDPG does not slow down or affect the stability of DDPG learning. This may be explained by the fact that AM does not disrupt the distribution of the environment and does not go against the mathematical concepts outlined in DDPG.

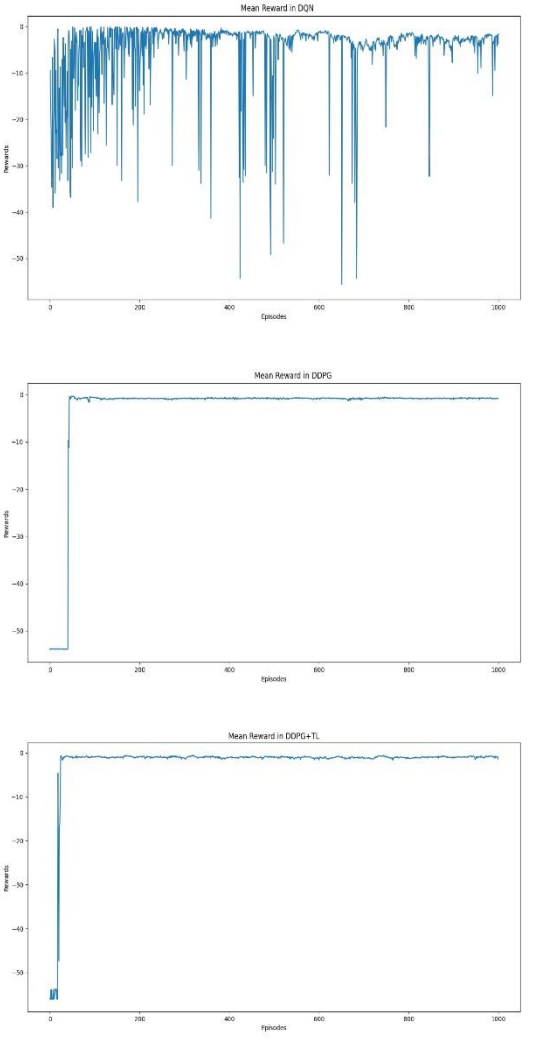


Fig. 14. Tendency of average reward during training in three EMS models.

The DDPG and TL were developed with the goal of speeding up the process of finding the best controls by making use of the parameters learned by a neural network. All other EMSs are the same, the only difference is the length of the driving cycle. According to TL theory, in the DDPG plus TL control instance, only the neural network's output layer has to be retrained for a new driving cycle. This might drastically shorten the amount of time required to master a new control regime. Here, we examine the convergence rate and training time of the DPPG plus TL versus the standard DDPG. To some extent, the Q table is approximated by the neural network in each of these DRL approaches. Fig.. 15 shows the average inaccuracy of Q table and the training sessions. The declining trends show that the quality of the control sequence obtained improves with each subsequent episode. Compared to regular DDPG the mean error value in DDPG plus TL is more reasonable for each episode. This finding suggests that the suggested method may be able to gain insight into its surroundings, leading to enhanced regulation. It follows that the DDPG plus TL is more efficient at learning than the standard DDPG.

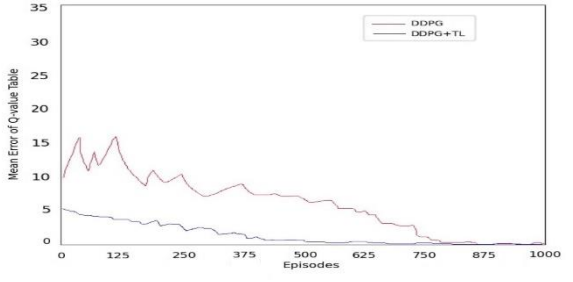


Fig. 15. Mean error under DDPG and DDPG plus TL.

To provide additional technical outcomes, Table 4 shows the total time spent training with both strategies on the same driving cycle. Because most of the parameters in the neural network may be recycled, the DPPG plus TL is clearly more efficient than regular DDPG. This feature makes it possible to implement the planned EMS in practical driving situations. Also, DDPG+TL algorithm-based EMS requires 1.5 % less hydrogen fuel as compared to DDPG based algorithm based EMS.

Table 4. Learning duration for DDPG and DDPG plus TL

Algorithm	Training Time (hrs)
DDPG	03.17
DDPG plus TL	0.59
System Configuration- Intel(R) i3-3110M CPU @ 2.40GHz and usable RAM 7.89 GB.	

5. CONCLUSION

This research provides a learning-based energy management method for FCHET's that is model-free and is constructed on the DDPG algorithm of DRL. The model-free DDPG method uses trial and error to discover the best possible EMS solution. In order to accomplish this goal, DDPG collects a sizable number of actual samples from the real environment and achieves improved performance.

In the Simulation implementation, two stages register capacitor equivalent circuit model (Lithium-ion battery) and Dick-laminae electric circuit model (PEMFC) were used to represent the nonlinear modeling of battery-FC for calculation of state space analysis. The prime objective of this research is to investigate the feasibility of incorporating algorithms for reinforcement learning into an energy management approach for hybrid trains. As a result of this, DDPG and DDPG with transfer learning outperformed in terms of fuel consumption, which indicates that the DDPG-based approaches can learn better control policies that lead to higher energy efficiency with reduced fuel required for the powertrain. Also, action masking restricts the set of actions that the agent can take, while the reward function shapes the behavior of the agent by providing a measure of success or failure for each state-action pair. Both AM and rewards techniques can be used together to improve the performance of a DDPG agent in complex environments. After putting the DDPG algorithm into practice using real train drive cycles (Jind -Sonipat) to train the agent, the agent with the highest reward is chosen and then that agent's performance is analyzed using again same drive cycle. An additional algorithm that is known as DDPG plus transfer learning is developed. The research integrates DDPG and transfer learning to achieve its objectives to construct an adaptive EM controller for FCHET's to decrease the laborious training time associated with the DRL technique. It is simple to generalize this control architecture such that it may be used in another hybrid power train.

ACKNOWLEDGEMENTS

The authors acknowledge Centre of Excellence in Complex and Nonlinear Dynamical Systems (CoE-CNDS) laboratory for providing support and a platform for research.

DECLARATIONS

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

REFERENCES

[1] Varan Navale , Timothy C. Havens, "**Fuzzy logic controller for energy management of power split hybrid electrical vehicle transmission**", IEEE International Conference on Fuzzy Systems : 940-947, 2014.

[2] S. Ziaeinejad, Y. Sananse, A Mehrizi, "**Fuel cell based auxiliary power unit: Ems. Sizing and current**

estimator based controller", IEEE Trans. on Vehicular Tech 65: 4826-4835, 2016.

[3] A. Tashakori Abkenar, A. Nazari, S. D. G. Jayasinghe, A. Kapoor, and M. Negnevitsky, **"Fuel cell power management using genetic expression programming in all-electric ships"**, IEEE Transactions on Energy Conversion: 779-787, 2017.

[4] Rui Wang , Srdjan M. Lukic, **" Dynamic programming technique in hybrid electric vehicle optimization"**, IEEE International Electric Vehicle Conference: 01-08, 2012.

[5] H. Ali Borhan , Ardalan Vahidi , Anthony M. Phillips, Ming L. Kuang , Ilya V. Kolmanovsky. **"Predictive energy management of a power-split hybrid electric vehicle"**, American Control Conference: 3970-3976, 2009.

[6] Hassan Khalil, **"Nonlinear System analysis"**, Pearson Publication, 2002

[7] K. Ettihir, L. Boulon, and K. Agbossou, **"Energy management strategy for a fuel cell hybrid vehicle based on maximum efficiency and maximum power identification"**, IET Electrical Systems in Transportation 6(4): 261-268, 2016.

[8] S. Mane, F. Kazi, and N. M. Singh, **"Fuel cell and ultra-capacitor based hybrid energy control using ida-pbc methodology"**, In International Conference on Industrial Instrumentation and Control (ICIC): 879-884, 2015.

[9] A. Henni-A. Abo M. Wack M.Ayad, M. Becherif , **"Energy management of a fuel cell and supercapacitors by passivity-based control and sliding mode control"**, Power journal 2(4): 1-7, 2011.

[10] P. Thounthong, S. Pierfederici, J. P. Martin, M. Hinaje, and B. Davat, **"Modeling and control of fuel cell/supercapacitor hybrid source based on differential flatness control"**, IEEE Transactions on Vehicular Technology 59(6): 2700-2710, 2010.

[11] J Snoussi, S. Ben Elghali, R. Outbib, M.F. Mimouni , "**Sliding mode control for frequency-based energy management strategy of hybrid Storage System in vehicular application"**, International Symposium on Power Electronics, Electrical Drives, Automation and Motion (SPEEDAM): 1109-1114, 2016.

[12] Chao-Ming Lee, Shin-Han Han, Chen-Hong Zheng, and We-Song Lin, **" Power split of fuel cell/ultracapacitor hybrid power system by backstepping sliding mode control"**, In IPEC 2012 Conference on Power Energy: 538-543, 2012.

[13] Daming, Z., AI-Durra, A., Fei, G., Simoes, M.G, "**Online energy management strategy of fuel cell hybrid electric vehicles based on data fusion approach"**, Journal of Power Sources 366 :278–29, 2017.

[14] Zheng, C., Xu, G., Xu, K., Pan, Z., Liang, Q , "**An energy management approach of hybrid vehicles using traffic preview information for energy saving**", Energy Convers. Manag., (105) : 462–470, 2015.

[15] Liu, T., Zou Y., Liu D., Sun F, **"Reinforcement learning-based energy management strategy for a hybrid electric tracked vehicle"**, Journal of Energies (8): 7243–7260, 2015.

[16] Heeyun Lee, Changbeom Kang, Yeong Park, Namwook K, "**Online data-driven energy management of a hybrid electric vehicle using model-based Q-Learning**", IEEE Access (8) : 84444-84454, 2020.

[17] Hu, Y., Li, W., Xu, H., Xu, G, "**An online learning control strategy for hybrid electric vehicle based on fuzzy Q-learning**", Energies (8): 11167–1118, 2015.

[18] Zhongping Yang, Feiqin Zhu, Fei Lin, " **Deep-Reinforcement-Learning-Based Energy Management Strategy for Supercapacitor Energy Storage Systems in Urban Rail Transit**", 22,(2): 1150-1160 , 2021.

[19] M. Sehang , Yonghua Z , Hamido Fujit , " **Deep reinforcement learning with reference system to handle constraints for energy-efficient train control**", Information Sciences Elsevier, 570: 708-721, 2021.

[20] Hyunsoo Lee, Seok-Youn Han, Keejun Park, Hoyoung Lee and Taesoo Kwon , " **Real-Time Hybrid Deep Learning-Based Train Running Safety Prediction Framework of Railway Vehicle**" , Machines MDPI:. 1-18, 2021.

[21] Kai Deng , Yingxu Liu, Di Hai , Hujun Peng, " **Deep reinforcement learning based energy management strategy of fuel cell hybrid railway vehicles considering fuel cell aging**", Energy conversion and management (251): 1-8, 2022.

[22] A. Saadi , M. Becherif , D. Hissel , H.S. Ramadan, " **Dynamic modeling and experimental analysis of PEMFCs: A comparative study**", International Journal of Hydrogen Energy, V 42(2): 1544-1557, 2017.

[23] Zekeriya Ender Eger, "**Reinforcement learning based energy management strategy for fuel cell hybrid vehicles**", Sabanci University: 1- 56, 2022.

[24] Xiaowei Guo, Teng Liu, Bangbei Tang, Xiaolin Tang, Jinwei Zhang, Wenhao Tan, Shufeng Jin , "**Transfer Deep Reinforcement Learning-enabled Energy Management Strategy for Hybrid Tracked Vehicle**", arXiv:2007.08690: 1-11, 2020.

[25] Yogesh E. Wankhede, Sheetal Rana, Faruk Kazi , "**SoC Estimation of Battery in FCHEVs Using Reformulated Constrained Unscented Kalman Filter**", 1st International Conference on Sustainable Technology for Power and Energy Systems (STPES): 1-6, 2022.

[26] Yuecheng Li , Hongwen He, "**Deep Reinforcement Learning-based Energy Management for Hybrid Electric Vehicles**", Springer Cham : 1-123, 2022.