Centralized Multi-Agent SOC Control for Battery Health Using Proximal Policy Optimization in EVs

Armin Lotfy, *Student member, IEEE,* Mohamad Alzayed, *member, IEEE*, Hicham Chaoui, *Senior member, IEEE,* *Loïc Boulon, Senior member, IEEE*

***Abstract*—Lithium-ion batteries (LIBs) have garnered significant attention due to their expanding use in various applications, including electric vehicles (EVs) and smart grids. To meet the diverse requirements of these applications, LIB cells are configured in different architectures, such as multiple cell/module/pack which are arranged in series and parallel configurations. In series configurations, a state of charge (SOC) balancing system is essential to ensure uniform SOC levels across all cells. For battery electric vehicles (BEVs), which rely solely on LIBs as their energy storage system (ESS), maximizing the ESS capacity is crucial for extending the driving range. SOC balancing is a key strategy to achieve optimal utilization of ESS capacity in EVs. This paper presents a model-free cooperative multi-agent control framework designed to regulate and balance the SOC of lithium-ion battery (LIB) cells in EVs during real-time driving operations. The proposed method utilizes a series architecture comprising three LIB cells, each equipped with a buck-boost converter and a proportional-integral (PI) controller, controlled by a reinforcement learning (RL) agent. The Proximal Policy Optimization (PPO) algorithm is used as the RL agent in this multi-agent framework, where each PPO agent independently manages the SOC of a corresponding battery cell based on observed data. During the training phase, all PPO agents work collaboratively to balance the SOCs of the LIB cells, thereby preventing interruptions in EV performance. The effectiveness of the proposed approach is demonstrated by comparing its performance with single-agent methods such as PPO, Soft Actor-Critic (SAC), and Twin Delayed Deep Deterministic Policy Gradient (TD3), as well as with other multi-agent methods. The results show that the proposed method performs better than the existing approaches, indicating its potential for superior performance.**

***Index Terms*—Active Cell Balancing, Proximal Policy Optimization, Centralized Training with Decentralized Execution, State of Charge, State of Health.**

I. Introduction

The surge in electric vehicles (EVs) adoption offers a promising solution to contemporary transportation challenges while promoting sustainability [1]. EVs encompass a diverse spectrum, including battery electric vehicles (BEVs), hybrid electric vehicles (HEVs), fuel cell electric vehicles (FCEVs), plug-in electric vehicles (PIEVs), and plug-in hybrid electric vehicles (PIHEVs). A common thread binding these categories is their reliance on battery energy storage systems (BESSs). Given the escalating demand for EVs, the ongoing enhancement of BESS is imperative to expand their practical driving range. This enhancement significantly bolsters the public's perception of EVs, contributing to heightened acceptance. Moreover, it represents a crucial advancement in the ongoing competition between EVs and conventional internal combustion vehicles [2]. Various battery types find application in EVs, encompassing Lead-acid, Ni-MH, Ni-Cd, and lithium-ion batteries (LIBs) [3]. LIBs showcase their superior attributes through higher C-rates, elevated energy densities compared to alternative battery varieties, minimal self-discharge, and less weight [4]. Although LIBs offer several advantages, they also encounter certain limitations, including challenges related to overcharging, over-discharging, and intricate and nonlinear characteristics. Taking these factors into account, the necessity of a Battery Management System (BMS) for LIBs becomes necessary [5]. BMS monitors various facets of the battery, including temperature, charging and discharging procedure, as well as cutoff voltage and current [6].

Battery packs utilized in EVs are composed of numerous battery cells organized in combinations of both series and parallel configurations. This configuration is essential to provide the requisite power output and attain the desired storage capacity [7]. Due to manufacturing variations, each battery cell possesses distinct traits that result in varied behaviors across cells. These disparities arise from production steps like mixing, coating, slitting, stacking, winding, and packaging in the battery manufacturing process [8], [9], [10]. Cell balancing is one of the suggested methods to mitigate the effect of these manufacturing variations, which is explored less than the other aspects of BMS [6], [11]. The cell balancing methods based on controlled variables are divided into voltage control, state of charge (SOC) control, and voltage control [12]. Cell balancing techniques based on SOC aims to equalize the SOC levels of distinct battery cells through active [13], [14] and passive methods [15], [16], [17]. Achieving a balanced SOC state is pivotal for LIBs due to the critical concerns of overcharging and over-discharging, both of which can detrimentally impact this battery type. Consequently, the controller needs to be capable of adjusting the SOC of battery packs to prevent these issues [16]. The controller must also maintain the SOC of the battery at a maximum level since the battery's overall capacity is determined by the minimum SOC of its individual cells, a measure taken to prevent battery overcharging. Addition-

A. Lofty, M. Alzayed and H. Chaoui are with Intelligent Robotic and Energy Systems Research Group (IRES), Department of Electronics, Carleton University, Ottawa, ON, K1S 5B6, Canada. (e-mail: armin.lotfy@carleton.ca; mohamad. alzayed@carleton.ca; hicham.chaoui@carleton.ca

Loïc Boulon is with the Hydrogen Research Institute, University of Québec at TroisRivières, Trois-Rivières, QC G9A 5H7, Canada, (e-mails: Loic.Boulon@uqtr.ca).

ally, SOC has a crucial role during the discharging process, with the minimum battery SOC set to avert over-discharging. Both overcharging and over-discharging can inflict irreversible damage on the battery [18]. Another crucial aspect is the controller's capability to adapt to environmental conditions, as real-world conditions are seldom ideal and continuously changing. Factors such as variations in driver behavior due to different conditions further exacerbate the challenge, making adaptation to uncertainty a critical aspect of controller design.

Passive cell balancing is a widely used technique to dissipate excess stored battery charge by converting it into heat through incorporated resistors [19], [20], [21]. The primary benefits of this approach lie in its uncomplicated design and cost-effectiveness. However, its primary drawbacks include energy wastage, leading to reduced battery capacity, an essential consideration in EVs [22], [23]. Moreover, the produced heat requires cooling, which consumes energy and adversely impacts neighboring battery cells. Unlike the passive cell balancing approach, the active cell balancing (ACB) method redistributes surplus stored energy to another battery cell, elevating its SOC while converting the excess energy into heat [23], [24], [25]. ACB is more effective compared to passive cell balancing, which holds significance in the context of EVs. However, it comes with greater costs and necessitates a controller for monitoring and managing the energy transfer between battery cells [25], [26]. Another method that has recently attracted attention is the use of reconfigurable battery packs [27]. [28] utilizes an ESS composed of multiple LIB cells that can dynamically adjust the battery pack's topology through the activation of controllable switches, depending on various scenarios. These adjustments enable the battery to maintain equilibrium among the SOCs of the LIB cells. However, a drawback of this method is that the incorporation of numerous controllable switches can significantly increase the overall cost of the battery pack. A re-configurable battery pack design is introduced in [29] that offers advantages such as high-speed balancing and fault tolerance, while eliminating the need for additional components. However, a notable disadvantage is that as the number of switches increases, the energy losses due to switching also rise while this effect is neglected. Moreover, SOC balancing occurs only when the ESS is not supplying power to the load. A research gap is discernible in the existing literature pertaining to cell balancing methodologies, with a particular dearth of attention directed toward the cell balancing controller. In the majority of scholarly works, this aspect of cell balancing has garnered limited focus. A prevalent choice for the cell balancing controller is the Proportional Integral (PI) controller, which has been favored due to its simplicity and effectiveness. However, it is noteworthy that the PI controller is knowledge-based, and as such, the results it yields are sub-optimal. Furthermore, PI controllers are not designed for adaptive operation. While the literature offers various optimization techniques for PI controllers [30], [31], a substantial portion of these methods grapple with issues of computational complexity and often fail to achieve optimal outcomes, particularly given the inherent complexity and non-linearity of battery systems.

To address these challenges, Deep Reinforcement Learning (DRL) methods present a promising solution. DRL methods are particularly advantageous for online operations because they are capable of managing the nonlinear and complex behaviors characteristic of batteries. Additionally, DRL techniques are highly adaptable to different scenarios, making them especially suitable for battery management applications [32]. Another significant benefit of DRL methods is their ability to learn and adapt policies based on interactions with the environment, thereby eliminating the need for a predefined model of the system. DRL methods have been applied in various contexts, such as reducing battery charging time [33], balancing SOC with controller multilevel converters [34], and solving optimization problems in battery swapping-charging system (BSCS) topology [35].

The Proximal Policy Optimization (PPO) algorithm has garnered considerable interest in recent years as a RL-based method. The PPO algorithm has been applied across various domains within the realm of EVs, such as [36], [37], [38], [39], [40], [41]. PPO, grounded in the policy gradient methodology, aims to enhance the policy of the network during interaction with the environment and utilizes stored experiences in batches to overcome shortcomings identified in previous algorithms such as Deep Q-network (DQN), Trust Region Policy Optimization (TRPO), and Actor Critic with Experience Replay (ACER), ultimately enhancing training speed [42]. These challenges include issues related to hyper-parameter tuning sensitivity, reduced sample efficiency, and simplified implementation [42]. The PPO algorithm leverages the Actor-Critic architecture comprising two Deep Neural Networks (DNNs), namely the policy and value networks. Being rooted in Model-free methods, PPO operates without prior knowledge of environment characteristics, dynamically improving its policy through interaction with the environment—an essential aspect in real-world applications like battery balancing, where controllers must contend with environmental uncertainties. Another key feature of the PPO algorithm is simultaneous updating of the policy and value networks, rendering it a promising candidate for online control. This capability enables PPO to continuously enhance its policy without interrupting its interaction with the environment.

The rapid advancements in reinforcement learning (RL) algorithms have fueled growing interest in Multi-Agent Reinforcement Learning (MARL), especially due to its ability to handle multi-objective tasks in complex, real-time environments. A key advantage of MARL lies in its capacity to enable multiple agents to interact within a shared environment simultaneously, with each agent autonomously learning an optimal course of action. This feature is particularly valuable in dynamic and intricate applications that demand simultaneous management and optimization of multiple objectives, such as energy management systems (EMS) in electric vehicles (EVs) and smart grids. For instance, [43] introduced a model predictive control (MPC)-based RL approach for battery management in smart grids. The method employs a least squares temporal difference (LSTD) learning algorithm, a variant of temporal difference (TD) learning, to optimize battery utilization based on predicted costs from the MPC model while simultaneously improving the MPC parameters.

Similarly, [44] proposed a MARL-based EMS for smart grids, utilizing predictive modeling to distribute power efficiently across various sectors. In [45], a MARL controller was developed to regulate load frequency and manage renewable energy sources such as photovoltaic (PV) systems, fuel cells, and wind turbines within a smart grid. MARL has also found significant applications in vehicular energy systems. Studies such as [46] and [47] showcased MARL-based EMS implementations in plug-in hybrid electric vehicles (PHEVs) and hybrid electric vehicles (HEVs), achieving energy efficiency improvements of up to 23.5% and 15.8%, respectively. In another notable example, [48] employed MARL to optimize the charging and discharging schedules of lithium-ion battery (LIB) packs, substantially extending the lifespan of individual cells. Additionally, [49] proposed a MARL-based operational management system for battery swapping and charging stations (BSCS), effectively managing both the charging process and the distribution of battery packs.

There is limited literature addressing the application of RL methods for SOC balancing in lithium-ion battery packs. For instance, [50] introduced an RL-based SOC balancing approach employing an Actor-Critic architecture. While the method demonstrated improved performance compared to traditional techniques, it has some limitations. Specifically, the absence of battery pack SOH as an observation parameter overlooks critical aspects of battery degradation, which can impact the longevity and efficiency of the battery pack. Additionally, the extra charge transfer between cells inherent to this method leads to increased capacity loss, resulting in reduced overall efficiency. Furthermore, the selected RL framework is susceptible to the overestimation problem, which could compromise the reliability of the solution in practical applications. In [51], an RL-based approach was proposed for SOC balancing in Dynamic Reconfigurable Batteries (DRBs) by considering battery cell voltages. While this method attempts to equalize cell voltages, the architecture necessitates a large number of switches, significantly increasing the cost and complexity of the battery system. Moreover, the method does not consider battery health parameters, such as SOH, leaving a crucial aspect of long-term system performance unaddressed. Similarly, [52] proposed a Deep Q-Network (DQN) for SOC balancing in DRBs, but this method focuses exclusively on the discharging process, neglecting the charging dynamics critical for balanced and efficient energy management. Lastly, [53] introduced a single-cell balancing approach for extended-range electric drives; however, this work also fails to incorporate SOH considerations, a key parameter for maintaining the health and reliability of the battery pack over extended use. The lack of attention to battery health parameters, such as SOH, across these studies highlights a critical gap in the existing literature. Battery health directly influences the capacity retention, thermal stability, and overall lifespan of lithium-ion batteries. Without incorporating SOH into the observation space, RL-based methods risk suboptimal decision-making that may accelerate degradation and undermine the long-term efficacy of the system. Therefore, a comprehensive approach that integrates SOC balancing with SOH considerations is essential for advancing the field and achieving both high performance and sustainability in battery management systems.

A MARL architecture was put out by the authors as part of their prior research to optimize SOC EMS in order to improve the driving range of BEVs. Building on this framework, the current work presents a novel approach to state-of-health (SOH) balancing for LIB cells, taking into account the Multi-Agent Proximal Policy Optimization (MAPPO) architecture.

The proposed method in this paper focuses on maintaining equilibrium of the battery cells in terms of both SOC and SOH during driving conditions, ensuring that the vehicle's normal driving performance is not affected. To the best of the authors' knowledge, this is among the first studies to use a MARL framework for SOC balancing in such a setting, which fills a significant gap in the field. The integration of the SOC and SOH balancing method into the vehicle's typical operating cycle, which includes several charging and discharging phases, is another noteworthy contribution of this work. This guarantees that the balancing process continues consistently and adaptively during the vehicle's operation, improving the battery's longevity and performance. The technique in question is noteworthy for its ability to optimize battery health and efficiency without necessitating any dedicated downtime or disruptions to the BEV's normal operation.

In light of these considerations, the proposed method offers the following contributions:

- The proposed method introduces a unique approach to SOC balancing by adjusting the drawn or injected current from individual battery cells rather than transferring excessive charge between them. This strategy minimizes degradation, improves battery health, and reduces the number of charge-discharge cycles, thereby enhancing the longevity of the battery pack.
- This paper presents a cost-effective method for rapid SOC balancing with minimal additional hardware requirements. By eliminating the need for hardware modifications or auxiliary systems, the proposed approach significantly reduces the overall cost of battery pack design, making it a practical and scalable solution for real-world applications.
- The proposed method integrates key battery parameters, including voltage, current, temperature, and SOH, to achieve simultaneous balancing of SOC and SOH. This approach ensures that all critical criteria are maintained, delivering a comprehensive solution that optimizes battery performance without compromising operational stability or efficiency.
- The method employs a Multi-Agent Proximal Policy Optimization (MAPPO)-based controller to achieve dynamic SOC balancing while mitigating risks associated with overcharging and over-discharging. This ensures continuous and adaptive battery management during normal BEV driving cycle, enhancing both the reliability of the vehicle and the lifespan of the battery system.
- The system environment is modeled as a Centralized Partially Observable Markov Process (CPOMP) and a Multi-Agent Markov Decision Process (MMDP), enabling the MAPPO framework to process diverse and adaptable information. This structured environment ensures ro-

bust training of the multi-agent reinforcement learning (MARL) architecture by providing a comprehensive representation of the partially observable system states and enabling the agents to learn and adapt effectively in dynamic scenarios.
- Within the CPOMP framework, a custom-designed reward function aligns with the application's specific requirements and ensures all necessary conditions are met for effective agent training. This tailored reward structure plays a pivotal role in guiding the proposed method towards achieving the desired objectives, thereby ensuring optimal performance and learning convergence within the CPOMP-based MAPPO architecture.

The remainder of this manuscript is organized as follows: Section II provides a detailed explanation of the primary methodology utilized in this study. Section III presents the validation results derived from implementing the proposed method. A comparative analysis of these experimental results is conducted in Section IV. Finally, Section V concludes the paper by summarizing the key findings and suggesting potential directions for future research.

II. Proposed Method

The presented multi-agent ACB method is delineated comprehensively in this section, providing an in-depth exploration of its three principal components. These constituent sections of the proposed ACB can be outlined as follows:

- The developed framework for the proposed method establishes a scenario in which the performance of the proposed controller can be validated. This scenario involves situations where the current demanded by the EV varies. The proposed method ensures that the requested current is supplied while simultaneously balancing the SOC of the LIB cells, which is the primary objective of the proposed approach.
- The core component of the proposed method is a multi-agent-based controller, which comprises several Proximal Policy Optimization (PPO) reinforcement learning agents. These agents are trained collaboratively through interactions with the environment. This section provides a detailed description of the controller's architecture, the observation and action signals, as well as the specific reward function designed for this application.

A. Model

This paper employs an BEV model, which is a modified version of the EV model from the IEEE VTS Motor Vehicles Challenge 2021 [54] Fig. 1. The model employed in this study is developed using the Energetic Macroscopic Representation (EMR), a graphical formalism specifically designed to systematically organize system models for control purposes, as detailed in [55]. EMR has been extensively utilized for the modeling, control, and energy management of complex systems due to its structured and intuitive approach. This model incorporates two three-phase AC electrical machines (EMs) positioned on the front and rear axles of an EV, establishing a dual-motor configuration. The primary energy storage system (ESS) consists of a LIB pack, which serves as the principal energy source. The electrical machines are driven by dedicated inverters, with their DC buses interconnected in parallel with the lithium-ion battery to ensure seamless energy flow. Fig. 2 presents the EMR representation of the lithium-ion battery. This graphical depiction provides a comprehensive understanding of the interactions between the battery and other system components, thereby enabling the formulation of robust and efficient control strategies.

The model features a dual-motor architecture and a LIB pack as the primary energy storage systems (ESSs) as shown in . Further details of the model's characteristics are provided in Table I. According to Table I, the model uses 6 LIB cells in parallel and 90 LIB cells in series. The modification to the model involves distributing the required current to accommodate only three LIB cells. This is achieved by dividing the current by the number of parallel cells and then multiplying the result by 3, corresponding to the number of LIB cells in the modified model. The current demand for the modified EV is determined based on the driving cycle depicted in Fig. 3 that illustrates the employed New European Driving Cycle (NEDC) and the corresponding current demand for the modified EV.

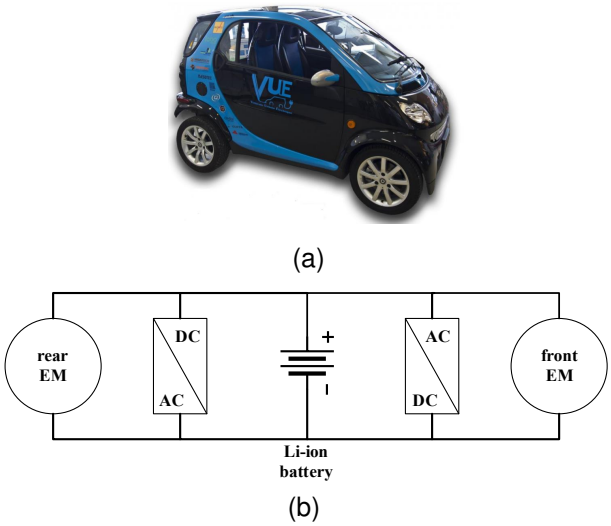

Fig. 1. (a) Real picture of the employed EV and (b) the structure of electric drives in the employed EV. Both figures are derived from the study presented in [54]

Using an appropriate environment for training DRL methods is crucial to ensure the agent receives sufficient and diverse data. Since DRL methods refine their policies through interactions with the environment, a lack of adequate or varied training scenarios can negatively impact their performance. To address this issue, the proposed driving cycle is designed to create a diverse range of situations necessary for effective training.

The proposed model incorporates three pre-processing units, which are primarily responsible for normalizing the input data to each agent in real-time. This normalization step is vital to prevent any single observation, especially those with larger value ranges, from disproportionately influencing the model's learning process. To achieve this, the Standardization method

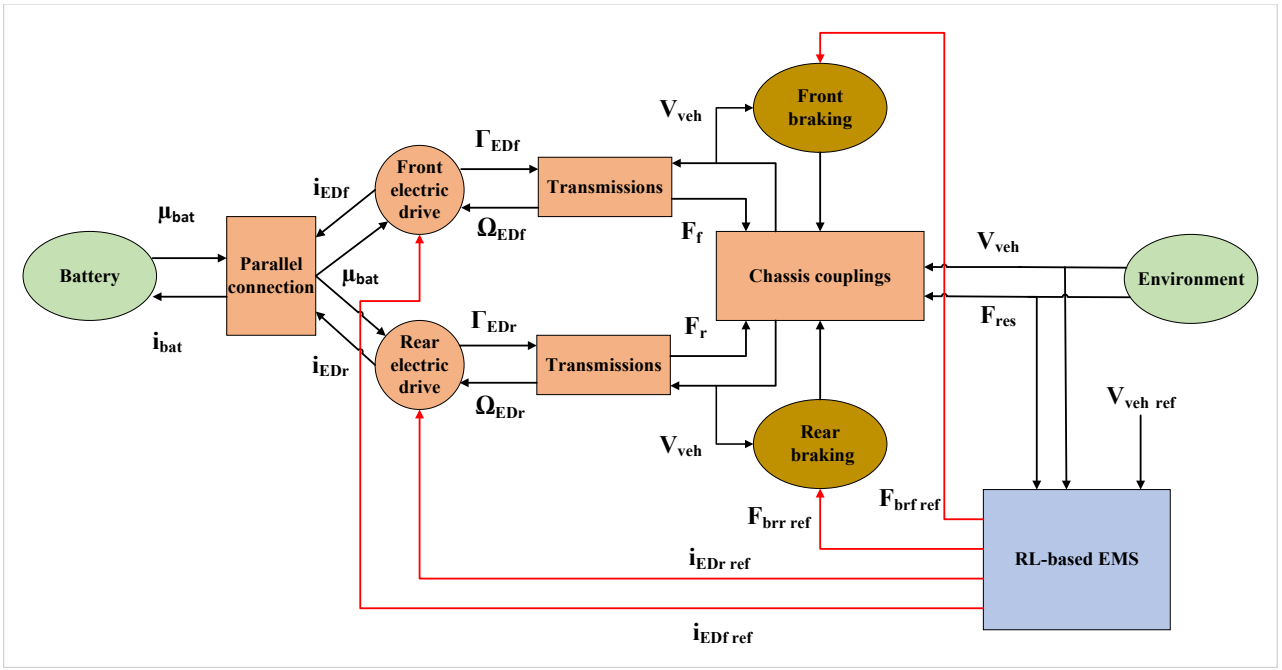

Fig. 2. The structure of the studied model [7]

TABLE I
THE UTILIZED MODEL CHARACTERISTICS

Parameter Name	Value
Vehicle mass	870
Velocity gain	0.0011
Front area (m^2)	2.4
Drag coefficient	0.35
The density of air (20°C)	1.223
Cell nominal capacity (Ah)	2.4
Cell equivalent capacity (F)	8280
Cell nominal voltage (V)	3.7
Open circuit voltage at 100%	4.15 V
Terminal resistance at 100%	63 mΩ
Cell initial temperature (°C)	44
Cell maximum temperature	50
Cell minimum temperature	−20
Battery cells in parallel configuration	6
Battery cells in series configuration	90
Air temperature	44
Gearbox ratio front axis	3
Gearbox ratio rear axis	4
Wheel radius (m)	0.38
Front transmission gain	7.8947
Rear transmission gain	10.5263

TABLE II
CHARACTERISTICS OF THE BATTERY MODEL USED

Parameter Name	Value
Cell Nominal Voltage (V)	4.07
Cell nominal capacity (Ah)	12
Internal cell Resistance ω	0.0396
Number of cells in series	3
Number of cells in parallel	1
Initial cell1 SOC %	$89 \pm \epsilon$
Initial cell2 SOC %	$85 \pm \epsilon$
Initial cell3 SOC %	$80 \pm \epsilon$

is used for data normalization, as illustrated in Eq. (1).

$$z = \frac{x - \mu}{\sigma} \tag{1}$$

where x represents the input data, μ is the mean of the data, and σ is the standard deviation. By standardizing the input data, the training process becomes more efficient, reducing the overall training time while improving performance.

The battery model used in this paper, as previously mentioned, consists of three LIB cells modeled using the Equivalent Circuit Model (ECM) [7], [54]. All the characteristics of this battery model are detailed in Table II. The primary reason for choosing this approach is that the proposed controller does not require an accurate model, which is a significant advantage of DRL algorithms. Additionally, this method reduces the complexity and computational cost of the model, thereby directly enhancing the training speed. Another benefit of the proposed method is the scalability of the controller, allowing the model to be applied across different categories of EVs with varying battery capacities. This adaptability is a result of the inherent properties of RL methods. The overall structure of the model used is depicted in Fig. 4. It is essential to point out that the battery's voltage and capacity match those stated in the reference exactly [54]. The suggested method's flexibility and scalability allow it to be applied to various battery architectures and configurations. The initial value of each LIB cell consists of a constant value and a random value. The primary rationale behind this architecture is to enhance the diversity and uncertainty of the model, thereby increasing the complexity of the environment. This added complexity allows for more robust testing and evaluation of the system's performance under a wider range of conditions. The proposed multi-agent-based ACB system processes several normalized observation signals, as discussed in the previous section. These observation signals vary for each agent, with each agent receiving signals specific to the battery it controls. The observation signals typically include the total requested current, the SOC of the LIB cell, and the cell voltage. The optimal order of these observation signals was determined through extensive

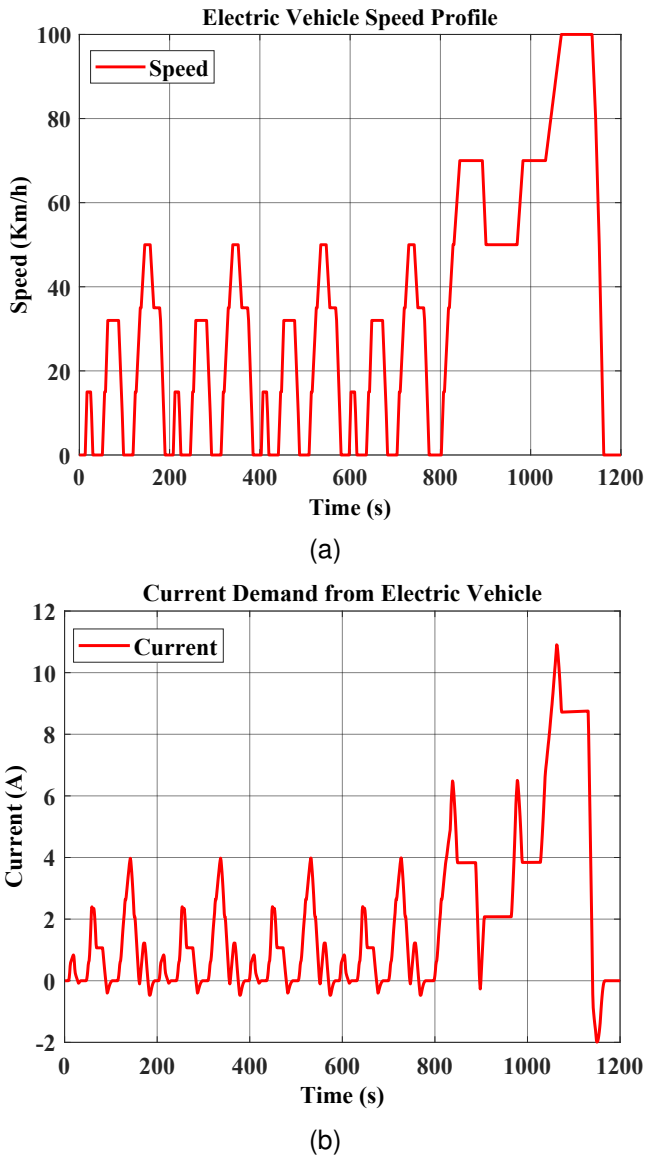

Fig. 3. (a) The Adopted New European Driving Cycle (NEDC) and (b) rescaled current demand profile of the BEV, both referenced from [7], [54].

trial and error to ensure the selection of the most effective signals. Choosing the correct observation signals is crucial to provide relevant information while minimizing noise and excluding irrelevant data. After processing these signals, the agents generate appropriate action signals that dictate the charging or discharging current for each LIB cell. These action signals are fed into a proportional-integral (PI) controller, which determines the required pulse-width modulation (PWM) signals for the buck-boost converter. Finally, the output of each converter is connected to a DC link. The ranges for the observation and action signals will be detailed in the following sections. The proposed method utilizes a Coulomb counting approach within the ACB system [56]. This approach is designed to optimize our methodology by reducing the training duration. Since DRL methods may not always rely on accurate models and incorporating such models can increase computational costs, employing Coulomb counting helps to improve training efficiency. This method calculates the SOC of the LIB cell based on variations in current, as detailed in Eq. (2).

$$\mathrm{SOC}[\%] = \mathrm{SOC}_{\mathrm{init}} - \frac{\int I\,dt}{\mathrm{C}_{\mathrm{nom}}} \times 100\%. \tag{2}$$

where SOC_{init}, C_{nom} represent initial SOC, the nominal capacity of the LIB cell. It is essential to note that the temperature of the battery pack is managed by the EV model itself. Therefore, the proposed controller does not need to take this parameter into account.

B. Battery Health Estimation

The proposed method offers improvements in SOH balancing by introducing a mechanism to reduce computational costs while incorporating model uncertainty. Specifically, to achieve this, a random value, denoted as ϵ, ranging between 0 and 0.02, is added to the initial capacity of each battery cell. This range is selected based on the fact that the total variation during the process remains between 0 and 0.001, ensuring that the introduced randomness aligns with the natural fluctuations in battery capacity. The initial capacity of each LIB cell is determined using the formulation provided in Eq. (3), which accounts for these variations in order to enhance the robustness and reliability of the balancing process.

$$C_i = C_{init} - \gamma \sum_{t=1}^{K} (\mathrm{I(t)}) \tag{3}$$

here, C_i and C_{init} represent the current and initial capacities of the LIB, respectively, while γ denotes a random variable that varies across multiple training iterations but remains constant within each individual iteration. By adding unpredictability across repeats and ensuring consistency within a single iteration, this technique improves the training process' robustness. Based on the computed capacity, as specified in Eq. (4), the SOH for each battery is estimated. By considering the beginning capacity as the nominal capacity of each LIB cell, this formulation establishes an indicator for evaluating the SOH throughout the operational cycle [57].

$$\mathrm{SOH}(\%) = \left(\frac{C_i}{C_{init}}\right) \times 100 \tag{4}$$

C. Methodology of the Proposed Approach

1) PPO Agent: The current study benefits a PPO-based controller, which offers distinct advantages over its counterparts such as TRPO, DQN, and ACR. These advantages include heightened stability, improved sample efficiency, reduced complexity, and streamlined implementation [42]. PPO leverages the Actor-Critic architecture, featuring a neural network (NN) dedicated to the value function (Critic network) and another NN for policy (Actor network). As previously mentioned, in the PPO framework, both neural networks undergo simultaneous updates. PPO is grounded in Policy Gradient (PG) methods, which aim to directly enhance the agent's policy through iterative interactions with the environment. While PG-based methods offer notable advantages such as superior convergence compared to value-based approaches

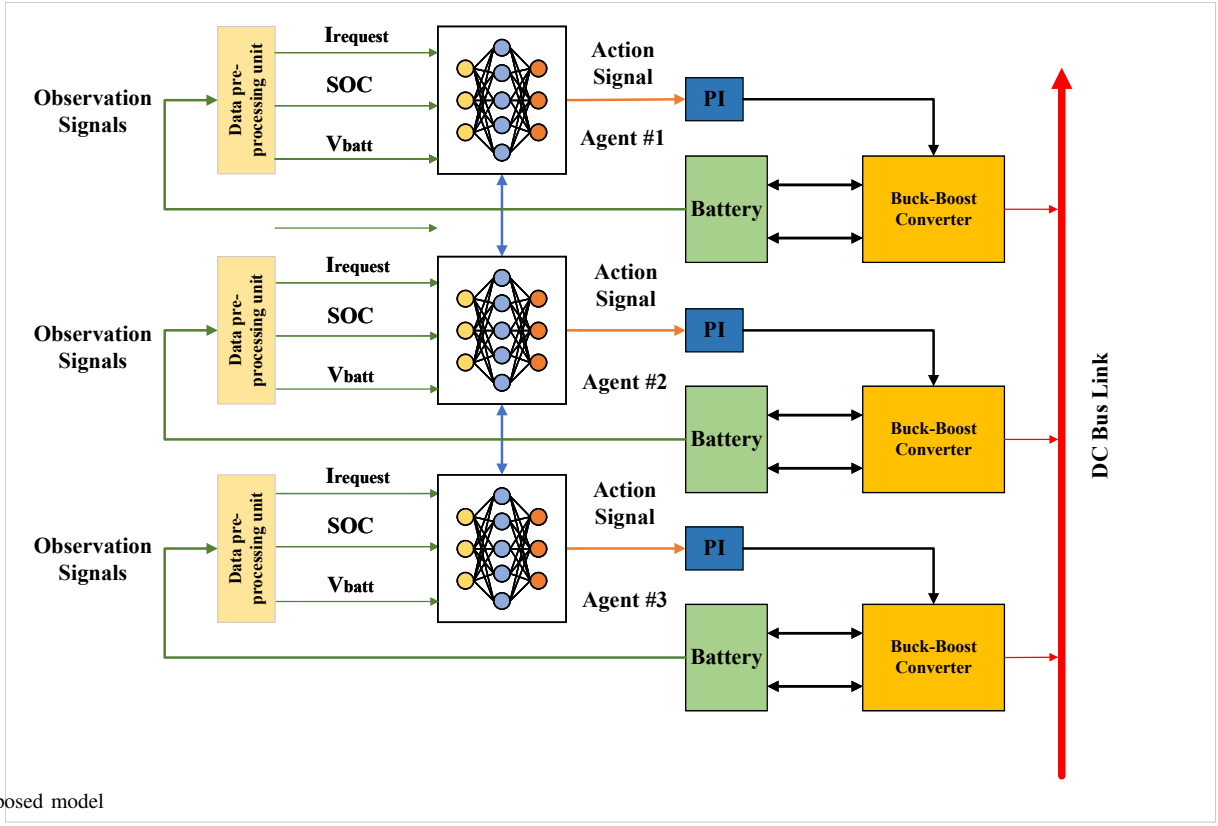

Fig. 4. Proposed model

and efficacy in managing high-dimensional state spaces, they are also associated with computational overheads, training intricacies, and variance issues. To mitigate these challenges, the PPO algorithm was proposed by [42]. The PG formula is refined utilizing a gradient estimator, as delineated in Eq. (5), derived through the differentiation of Eq. (6).

$$\hat{g} = \hat{\mathbb{E}}_t \left[\nabla_\theta \log \pi_\theta(a_t|s_t) \hat{A}_t \right] \tag{5}$$

$$L^{\text{PG}}(\theta) = \hat{E}_t \left[\log \pi_\theta(a_t|s_t) \hat{A}_t \right] \tag{6}$$

here, $\hat{\mathbb{E}}_t$ represents the expectation value, π_θ denotes a stochastic policy, and $\hat{A}_t$ stands for an estimator of the advantage function, as defined in Eq. (7). The expectation value $\hat{\mathbb{E}}_t$ is computed by iterative interacting with the environment and storing the state, action, reward, and next state in finite batches under the current policy (π_θ). Simultaneously, the advantage function provides insight to the PPO agent on the effectiveness of completed actions. With data saved in batches, advantage estimations are calculated as $\hat{A}_1, \hat{A}_2, \hat{A}_3, \ldots, \hat{A}_T$. Subsequently, the policy is refined by maximizing the objective function, as depicted in Eq. (8), utilizing the policy gradient of the advantage function, as illustrated in Eq. (9).

$$A_\pi(s,a) = Q_\pi(s,a) - V_\pi(s) \tag{7}$$

$$L_{\text{Clip}}(\theta) = \hat{E}_t \left[\min \left(r_t(\theta)\hat{A}_t, \right. \right. \\ \left. \left. \text{clip}\left(r_t(\theta), 1-\varepsilon, 1+\varepsilon\right)\hat{A}_t \right) \right] \tag{8}$$

$$\begin{aligned} \nabla L(\theta) &= E'_\pi \left[r_t(\theta) A'_\pi(s,a) \nabla \ln \pi_\theta(a|s) \right] \\ &= E'_\pi \left[\nabla r_t(\theta) A'_\pi(s,a) \right] \end{aligned} \tag{9}$$

where, $r_t(\theta)$ stands for the probability ratio, defined as per Equation (Eq. (10)), while ϵ denotes the hyperparameter. The clip function is invoked to confine the learning ratio within the interval $[1-\epsilon, 1+\epsilon]$. The objective behind employing the clip function is to ascertain an optimal learning rate conducive to fostering stability in the actions generated by the PPO algorithm.

$$r_t(\theta) = \frac{\pi_\theta(a_t|s_t)}{\pi_{\theta_{\text{old}}}(a_t|s_t)} \tag{10}$$

In the proposed methodology, the SOC of LIB cells is considered a dynamic variable that must be balanced across all cells in the battery system. This balance is influenced by the current load demand, as well as the SOC and voltage of each cell. As discussed in the previous section, each PPO agent receives observation signals that include the requested current (I_{request}), the SOC, and the voltage of the battery. The agent then determines the appropriate action signal, which corresponds to the specific current allocated to each LIB cell. The characteristics of the proposed agents, along with the observation and action ranges, are detailed in Table III. The sampling frequency is configured to 1 Hz, and the observation range is determined following the normalization process. The proposed method allows the agent to discharge each battery cell at a rate of up to $1C$ and charge at a rate of up to $0.5C$. These rates are specifically chosen to mitigate the risk of damage to the battery cells. The lower and upper bounds of the observation signals are defined as the minimum and maximum values of the normalized observation data. Similarly, the action signal bounds are determined by the maximum permissible charge and discharge currents for each battery cell, ensuring the agent

TABLE III
ATTRIBUTES OF PPO AGENTS

Parameter Name	Value
Number of episodes	1000
Sample Time (s)	1
Discount factor	0.99
Experience horizon	1200
Entropy loss weight	0.1
Mini batch size	300
Clip factor	0.2
Number of hidden neurons	200
Observation lower range	-0.8
Observation upper range	3.5
Action range	$[-6:12]$

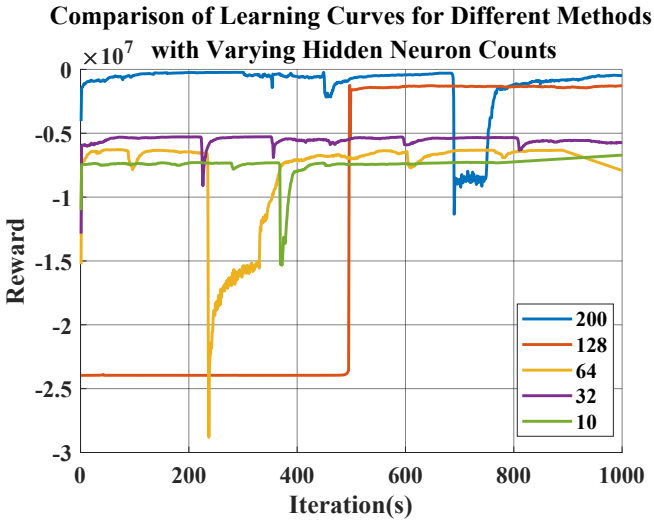

Fig. 5. Comparison of Learning Curves for Different Numbers of Hidden Layers

can bypass a specific cell while satisfying the requested current without compromising the cell's capacity or health. This configuration grants the agent increased flexibility to navigate and address more complex scenarios within the observation space, thereby improving its ability to optimize decision-making under intricate and dynamic operational conditions. The proposed method utilizes 200 neurons in the hidden layers, a value determined through an empirical trial-and-error process. This selection represents the minimum number of neurons required to achieve the maximum attainable value of the reward function. To illustrate the impact of varying hidden layer architectures, the learning curves for different network configurations with varying neuron counts are presented in Fig. 5.

2) Reward Function: The design of the reward function is another critical aspect in developing DRL agents, as it significantly influences the convergence behavior of the agent. In light of this, the proposed method employs a modified reward function tailored specifically for this application, with a focus on achieving the primary objectives during the training phase, as illustrated in Eq. (16). The proposed reward function is divided into two components: local and global. The global component calculates a penalty based on the overall system performance, particularly the discrepancy between the requested current and the total current provided by all LIB cells. This component ensures that all agents work collaboratively to achieve the collective goal. However, the local components assess the penalty for each individual LIB cell by calculating the discrepancy between the SOC and SOH of each cell, taken with their respective average values of SOC and SOH as shown in Eqs. (14) and (15).

$$\delta_{\text{global}} = (\text{I}_{\text{requested}} - |I_{\text{bat}_1} + I_{\text{bat}_2} + I_{\text{bat}_3}|)^2 \quad (11)$$

$$\overline{SOC} = \frac{SOC_1 + SOC_2 + SOC_3}{3} \quad (12)$$

$$\overline{SOH} = \frac{SOH_1 + SOH_2 + SOH_3}{3} \quad (13)$$

$$\delta_{local_{SOC}} = \sum_{i=1}^{3} (\overline{SOC} - \text{SOC}_{\text{bat}_i})^2 \quad (14)$$

$$\delta_{local_{SOH}} = \sum_{i=1}^{3} \exp(\overline{SOH} - \text{SOH}_{\text{bat}_i})^2 \quad (15)$$

$$R = -\left(\alpha \times R_{\text{global}} + \beta \times R_{\text{local}_{\text{SOC}}} + \zeta \times R_{\text{local}_{\text{SOH}}}\right) - P_V - P_t \quad (16)$$

here, $\overline{SOC}$, $\overline{SOH}$, R, P_V, and P_t denote the average values of the SOC and SOH, the cumulative reward corresponding to each action, and the penalty imposed for exceeding the predefined voltage and temperature thresholds, respectively. The primary reason for applying a penalty only when exceeding the predefined threshold is to provide the agent with greater flexibility in exploring and identifying the optimal solution. The hyperparameters α, β, and ζ are critical in ensuring the convergence of the proposed method. The values of these hyperparameters were determined through iterative experimentation, aiming to appropriately balance the weights in response to variations in the terms of the reward function. This fine-tuning process is essential for optimizing the performance and stability of the method.

3) Multi- agent Architecture: A sophisticated framework entitled MARL allows several agents to interact with each other in addition to a shared environment. In contrast to single-agent reinforcement learning, wherein the agent trains by simply interacting with the environment, multi-agent learning involves simultaneous interactions with the environment and other agents. MARL is categorized into cooperative, competitive, or mixed settings, depending on the nature of agent interactions during training [58]. In a cooperative setting, all agents work together to achieve a common objective, striving for optimal collective performance. Conversely, in a competitive setting, agents are driven by individual goals, often trying to outperform their counterparts [59]. Within a shared environment, agents in MARL can either be homogeneous, sharing similar policies and goals, or heterogeneous, with distinct objectives and strategies. In the proposed study, we adopt a homogeneous MARL setting, where all agents pursue a common goal. This environment is characterized by partial observability, as agents have access only to certain variables, such as the required current and the SOC and voltage of the batteries.

The recommended method utilized PPO RL agent as Multi-Agent PPO (MAPPO) concurrently benefit from the advantages of MARL and PPO RL agent. The proposed architecture employs a Centralized Training with Decentralized Execution (CTDE) framework for training MAPPO agents [60]. This approach is chosen because it enables agents to learn coordinated policies that optimize overall performance, specifically by balancing the SOCs of battery cells in real-time. The agents must collectively meet the general objective of providing the required current for a BEV while operating in a noisy environment and handling multiple objectives. CTDE is particularly effective for this application because it allows each agent to learn a policy that maximizes its expected return by considering the policies of other agents. This is crucial given the dynamic nature of the problem, which involves a noisy environment and requires simultaneous control of multiple objectives. MAPPO methods excel in such Real-Time Strategy (RTS) environments, where multiple agents interact with the environment simultaneously to manage various goals. This collaborative yet decentralized approach enhances the robustness and flexibility of the system, making it well-suited for complex, real-world applications. Based on the previous discussion, the policy update function for the PPO agent can be adapted as outlined in Eq. (17). This modification is crucial to account for the interactions among agents, ensuring that the behavior of each PPO agent is influenced by the actions and strategies of the other agents within the environment. By incorporating this modification, the updated policy function better captures the interdependencies and dynamic interactions among agents, which is essential for optimizing overall performance in a multi-agent reinforcement learning setting.

$$L_{\mathrm{i}}^{\mathrm{CLIP}}(\theta_i) = \mathbb{E}_{s,a_i,a_{i'}\sim D}\Bigg[\min\Big(r_i(\theta_i)\hat{A}_i(s,a_i,a_{i'}), \mathrm{clip}(r_i(\theta_i),1-\epsilon,1+\epsilon)A^i(s,a_i,a_{i'})\Big)\Bigg] \quad (17)$$

In the context of the proposed new architecture, the advantage and value function has been revised to accommodate the specific requirements of a multi-agent system, as demonstrated in Eqs. (18) and (19).

$$A^i(s,a_i,a_{i'}) = Q^i(s,a_i,a_{i'}) - V_i(s) \quad (18)$$

$$L_i^{\mathrm{VF}}(\phi_i) = \mathbb{E}_{s\sim D}\left[\left(V_{\phi_i}(s) - \hat{R}_i\right)^2\right] \quad (19)$$

here, $Q_i(s,a_i,a_{i'})$ represents the Q-value for agent i, which depends on both its own action, a_i, and the actions of the other agents, $a_{i'}$. The value function, $V_i(s)$, denotes the expected return for agent i starting from state s under its current policy. Additionally, $V_{\phi_i}(s)$ is the parameterized value function for agent i, and $\hat{R}_i$ is the estimated reward function for agent i.

In the CTDE framework, it is essential not only to update each agent's policy and value functions individually but also to compute the joint Q-value and value function. These computations are necessary to accurately capture the interactions among agents and are defined as shown in Eqs. (20) and (21).

$$Q^i(s,a_i,a_{-i}) = \text{Centralized Critic}(s,a_i,a_{i'};\theta_{\text{critic}}) \quad (20)$$

$$V_i(s) = \text{Centralized Value Function}(s;\phi_{\text{critic}}) \quad (21)$$

One of the primary components of the CTDE framework is a Centralized Critic, which is intended to improve the process of training several agents from each other in a cooperative setting, thereby enhancing overall performance. It facilitates the evaluation and improvement of agent policies by utilizing comprehensive joint information during the training process [61]. Unlike decentralized agents, the Centralized Critic has access to the complete state of the environment, including all agents' actions and states. This characteristic allows the MAPPO architecture to simultaneously evaluate each agent's performance independently and from an extensive point of view. By doing so, the Centralized Critic effectively computes the advantage function, which is critical for determining the discrepancy between expected and actual rewards, thereby guiding the agents' policy updates toward optimal behavior. This centralized approach to training allows agents to learn more efficiently and adapt to complex, multi-agent environments which is a necessary term in this application.

III. Validation and Discussion

In Section III, the validation results of the suggested approach are described in detail, and their comparison with the outcomes of the most popular RL-based methods in single agent and MARL architectures is done. The primary motivation is to demonstrate the suggested method's improved performance and to compare the outcomes with the proposed strategy. The hardware utilized to implement the recommended approach is a PC with an Intel Core i9-13900K with 128GB RAM and an NVIDIA GeForce RTX 3090 GPU. The suggested method is executed in MATLAB/Simulink. The comparison includes the PPO, Soft Actor Critic (SAC), and Twin-delayed deep deterministic policy gradient (TD3) algorithms within a single-agent framework, while in the MARL architecture, SAC and TD3 are chosen for evaluation. These algorithms were selected due to their superior performance in a noisy environment relative to the other approaches. Each proposed method and the other methods are trained for 1000 iterations to provide enough time to find the optimal solution. The proposed method is evaluated against other approaches using two performance metrics: Mean Square Error (MSE) and Mean Absolute Error (MAE) as shown in Eqs. (22) and (23). These metrics provide a quantitative basis for comparing the accuracy and effectiveness of the methods.

$$\mathrm{MSE} = \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2 \quad (22)$$

$$\mathrm{MAE} = \frac{1}{n}\sum_{i=1}^{n}|y_i - \hat{y}_i| \quad (23)$$

The experimental results are presented in Fig. 6, which demonstrates the effectiveness of the proposed method in achieving its intended objectives. Fig. 6a emphasizes the method's capability to deliver the requested current while simultaneously managing other tasks, ensuring continuous and uninterrupted current supply. Furthermore, the balancing of SOC and SOH is illustrated in Fig. 6b and Fig. 6c, respectively. These results

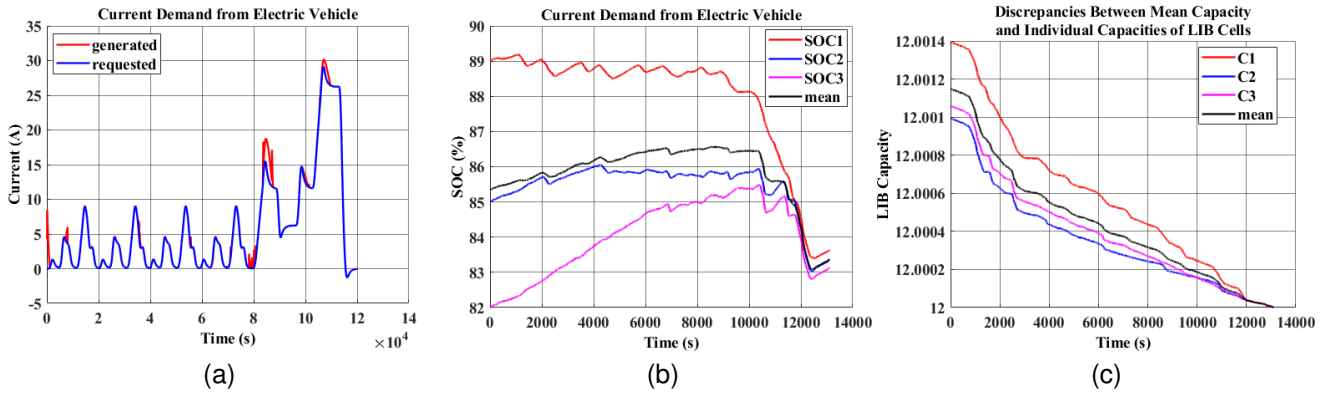

Fig. 6. (a) Comparison of the requested current versus the current supplied by the proposed method. (b) SOC balancing process. (c) capacity discrepancy during the balancing process.

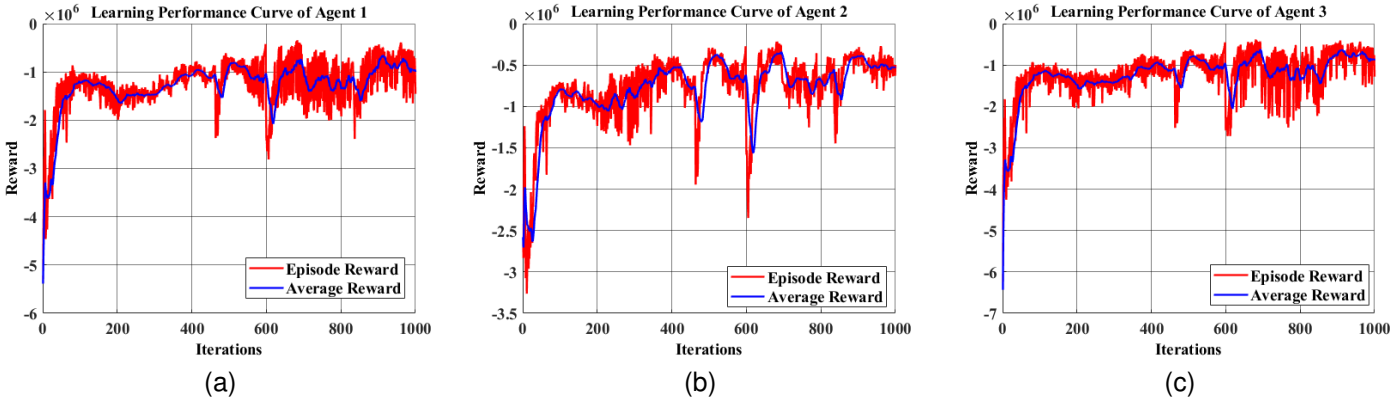

Fig. 7. Learning curves showing the reward progression for (a) Agent 1, (b) Agent 2, and (c) Agent 3 during training.

highlight the robustness of the proposed approach in maintaining optimal energy management and preserving battery health, thereby showcasing its comprehensive and efficient design.

An important accomplishment of the suggested approach is its ability to achieve this goal without producing any surplus current, as seen in Fig. 6b. Fig. 6b illustrates the process of balancing SOC, with an initial differential of 7% between SOC arrangements. The validation results concludes with a cumulative error of 0.51%, therefore emphasizing the efficacy of the approach in attaining SOC equilibrium.

The training curves for all agents are presented in Fig. 7, demonstrating the performance progression over time. This figure highlights the PPO agent's capability to refine its policy through continuous interaction with the environment. The following analysis aims to compare and evaluate the final differences between the mean SOC value and the SOC values of individual LIB cells across various architectures. Table IV comprehensively presents the final SOC values for all battery cells across the implemented methods, alongside the corresponding discrepancies between them. This table highlights the effectiveness of the proposed method compared to alternative approaches. Specifically, the mean SOC values are calculated as shown in Eq. (12), and the resulting discrepancies are analyzed to illustrate the performance differences. Notably, the table underscores the advantages of the multi-agent architecture. While the single-agent PPO method exhibits the poorest performance, its multi-agent counterpart demonstrates the most significant improvement, showcasing the superiority of the proposed multi-agent framework in achieving SOC balancing. Fig. 9 presents a comparative analysis of the final capacities achieved by different methods. This figure highlights the variations in performance across the approaches, offering insights into how each method impacts the final capacity of the battery cells. The results provide a clear indication of the effectiveness of the proposed method in optimizing capacity, especially when compared to alternative techniques. It is noteworthy that the MATD3 method demonstrates comparable performance in terms of capacity discrepancies. However, achieving this level of performance requires significantly longer training time compared to the other methods. The learning curves for different network architectures are presented in Fig. 8, highlighting the effectiveness of the proposed method in comparison to alternative approaches.

In order to assess the generalization and scalability of the proposed method, its performance is tested on three battery cells with varying degradation levels, each affecting capacity. These degradation factors are set at 0.009, 0.001, and 0.011, respectively. This test scenario is independent of the training procedure, meaning the proposed method encounters these conditions for the first time. The resulting performance is illus-

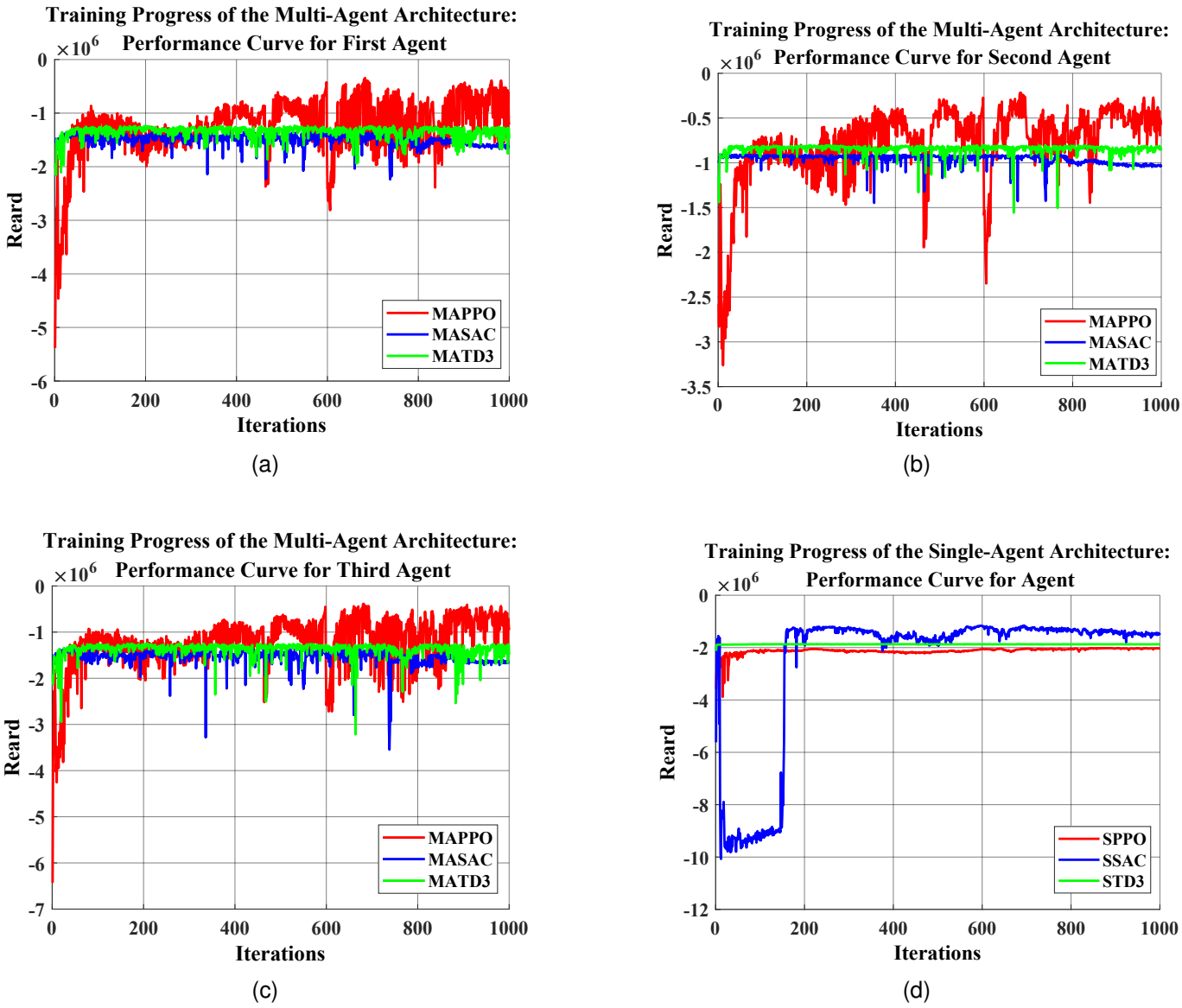

Fig. 8. Reward progression depicted through learning curves for: (a) Agent 1 in the multi-agent architecture, (b) Agent 2 in the multi-agent architecture, (c) Agent 3 in the multi-agent architecture, and (d) the single-agent scenario.

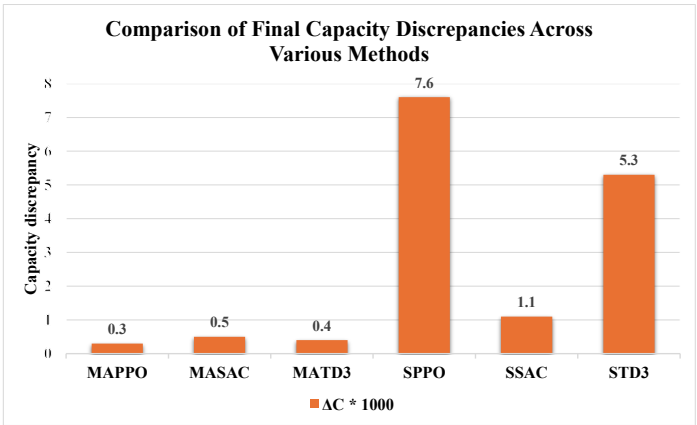

Fig. 9. Comparison of Final Capacity Discrepancies Across Various Methods

trated in Fig. 10. By showcasing these differences, the figure underscores the ability of the proposed approach to deliver more consistent and efficient outcomes in managing battery capacity under various conditions. A final comparative analysis is conducted to assess the performance of the proposed method against other architectural approaches. This comparison, illustrated in Fig. 11, demonstrates the clear superiority of the proposed method across all evaluation metrics. The results indicate that the proposed method successfully meets all of the predefined criteria, a level of performance that competing methods fail to achieve. Specifically, the comparison focuses on the methods' ability to provide the required current, and the results in Fig. 11 underscore the effectiveness of the proposed method in fulfilling this demand. While both the SAC and TD3 methods exhibit some ability to track the required current, they fall short in adhering to the SOC balancing criteria. In

TABLE IV
COMPARISON OF FINAL SOC DISCREPANCIES ACROSS VARIOUS METHODS

Method	**SOC$_1$ (%)**	**SOC$_2$ (%)**	**SOC$_3$ (%)**	**$\overline{\text{SOC}}$ (%)**	**δSOC (%)**
Proposed Method	83.1394	83.3618	83.6343	83.3785	0.5116
MASAC	88.6877	84.0708	77.8519	83.5368	11.3698
MATD3	88.6306	85.0010	81.5470	85.0595	7.1421
Single PPO	78.6459	67.1349	89.9912	78.5907	22.9115
Single SAC	83.6953	80.3586	79.1988	81.0842	5.2221
Single TD3	86.3318	82.9798	76.0769	81.7962	11.4385

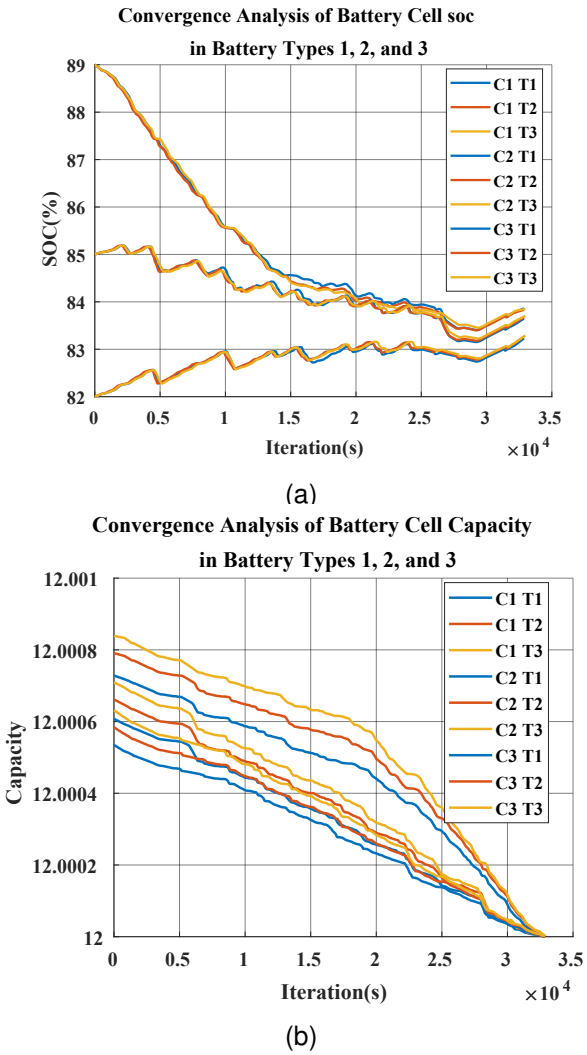

Fig. 10. Performance comparison of the proposed method across different battery cell types.

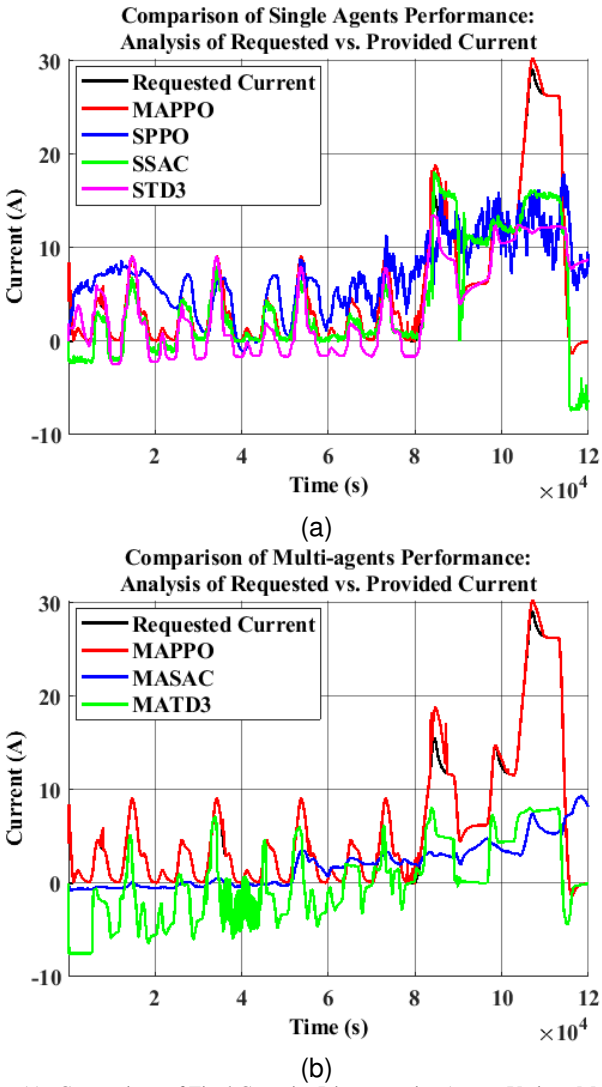

Fig. 11. Comparison of Final Capacity Discrepancies Across Various Methods

contrast, the proposed method not only maintains SOC balance but also consistently provides the requested current without compromising battery health or operational efficiency. This comprehensive evaluation highlights the robustness and adaptability of the proposed approach, particularly in comparison to existing architectures.

IV. CONCLUSION

This paper introduces a novel MAPPO architecture aimed at balancing the SOC and SOH of LIB cells, while simultaneously providing the required current during the operation of a BEV. The key innovation of the proposed approach lies in the use of a CTDE architecture, designed to achieve SOC balancing under variable current conditions, distinct from traditional methods that focus primarily on the battery charging process.

The validation results demonstrate that the proposed method effectively reduces the disparities in both SOC and SOH among the LIB cells, leading to improved overall battery longevity. This improvement is critical for maintaining the long-term performance and health of LIB packs in BEVs, making the proposed method a significant advancement in battery management systems.

For the future work, further enhancements can be achieved by increasing the number of learning iterations, which would refine the model's performance. Additionally, expanding the number of RL agents could enable the proposed SOC and SOH balancing strategy to account for additional factors, such as temperature, which plays a crucial role in battery degradation. Addressing temperature variations alongside SOC and SOH would further enhance the effectiveness of the method and contribute to mitigating battery degradation more comprehensively.

REFERENCES

[1] M. Ehsani, Y. Gao, S. Longo, and K. Ebrahimi, *Modern electric, hybrid electric, and fuel cell vehicles*. CRC press, 2018.

[2] Q. Ouyang, W. Han, C. Zou, G. Xu, and Z. Wang, "Cell balancing control for lithium-ion battery packs: A hierarchical optimal approach," *IEEE Transactions on Industrial Informatics*, vol. 16, no. 8, pp. 5065–5075, 2020.

[3] C. Iclodean, B. Varga, N. Burnete, D. Cimerdean, and B. Jurchiş, "Comparison of different battery types for electric vehicles," in *IOP conference series: materials science and engineering*, vol. 252, no. 1. IOP Publishing, 2017, p. 012058.

[4] Z. Gao, H. Xie, X. Yang, W. Niu, S. Li, and S. Chen, "The dilemma of c-rate and cycle life for lithium-ion batteries under low temperature fast charging," *Batteries*, vol. 8, no. 11, p. 234, 2022.

[5] K. See, G. Wang, Y. Zhang, Y. Wang, L. Meng, X. Gu, N. Zhang, K. Lim, L. Zhao, and B. Xie, "Critical review and functional safety of a battery management system for large-scale lithium-ion battery pack technologies," *International Journal of Coal Science & Technology*, vol. 9, no. 1, p. 36, 2022.

[6] M. Uzair, G. Abbas, and S. Hosain, "Characteristics of battery management systems of electric vehicles with consideration of the active and passive cell balancing process," *World Electric Vehicle Journal*, vol. 12, no. 3, p. 120, 2021.

[7] A. Lotfy, H. Chaoui, M. Kandidayeni, and L. Boulon, "Enhancing energy management strategy for battery electric vehicles: Incorporating cell balancing and multi-agent twin delayed deep deterministic policy gradient architecture," *IEEE Transactions on Vehicular Technology*, pp. 1–16, 2024.

[8] X. Han, L. Lu, Y. Zheng, X. Feng, Z. Li, J. Li, and M. Ouyang, "A review on the key issues of the lithium ion battery degradation among the whole life cycle," *ETransportation*, vol. 1, p. 100005, 2019.

[9] M. A. Hannan, F. Azidin, and A. Mohamed, "Hybrid electric vehicles and their challenges: A review," *Renewable and Sustainable Energy Reviews*, vol. 29, pp. 135–150, 2014.

[10] L. Lu, X. Han, J. Li, J. Hua, and M. Ouyang, "A review on the key issues for lithium-ion battery management in electric vehicles," *Journal of power sources*, vol. 226, pp. 272–288, 2013.

[11] T. Duraisamy and D. Kaliyaperumal, "Active cell balancing for electric vehicle battery management system," *International Journal of Power Electronics and Drive Systems*, vol. 11, no. 2, p. 571, 2020.

[12] Z. B. Omariba, L. Zhang, and D. Sun, "Review of battery cell balancing methodologies for optimizing battery pack performance in electric vehicles," *IEEE Access*, vol. 7, pp. 129 335–129 352, 2019.

[13] Z. Zhang, L. Zhang, L. Hu, and C. Huang, "Active cell balancing of lithium-ion battery pack based on average state of charge," *International Journal of Energy Research*, vol. 44, no. 4, pp. 2535–2548, 2020.

[14] A. Samanta and S. Chowdhuri, "Active cell balancing of lithium-ion battery pack using dual dc-dc converter and auxiliary lead-acid battery," *Journal of Energy Storage*, vol. 33, p. 102109, 2021.

[15] W. C. Lee, D. Drury, and P. Mellor, "Comparison of passive cell balancing and active cell balancing for automotive batteries," in *2011 IEEE Vehicle Power and Propulsion Conference*. IEEE, 2011, pp. 1–7.

[16] S. Wang, C. Fernandez, Y. Chunmei, Y. Fan, C. Wen, D.-I. Stroe, and Z. Chen, *Battery system modeling*. Elsevier, 2021.

[17] D. Thiruvonasundari and K. Deepa, "Optimized passive cell balancing for fast charging in electric vehicle," *IETE Journal of Research*, vol. 69, no. 4, pp. 2089–2097, 2023.

[18] J. Chen, C. Buhrmester, and J. Dahn, "Chemical overcharge and overdischarge protection for lithium-ion batteries," *Electrochemical and Solid-State Letters*, vol. 8, no. 1, p. A59, 2004.

[19] R. Di Rienzo, M. Zeni, F. Baronti, R. Roncella, and R. Saletti, "Passive balancing algorithm for charge equalization of series connected battery cells," in *2020 2nd IEEE International Conference on Industrial Electronics for Sustainable Energy Systems (IESES)*, vol. 1. IEEE, 2020, pp. 73–79.

[20] M. Bowkett, K. Thanapalan, T. Stockley, M. Hathway, and J. Williams, "Design and implementation of an optimal battery management system for hybrid electric vehicles," in *2013 19th International Conference on Automation and Computing*. IEEE, 2013, pp. 1–5.

[21] Y. Chen, X. Liu, H. K. Fathy, J. Zou, and S. Yang, "A graph-theoretic framework for analyzing the speeds and efficiencies of battery pack equalization circuits," *International Journal of Electrical Power & Energy Systems*, vol. 98, pp. 85–99, 2018.

[22] A. Pröbstl, S. Park, S. Narayanaswamy, S. Steinhorst, and S. Chakraborty, "Soh-aware active cell balancing strategy for high power battery packs," in *2018 Design, Automation & Test in Europe Conference & Exhibition (DATE)*. IEEE, 2018, pp. 431–436.

[23] K. Friansa, I. N. Haq, E. Leksono, N. Tapran, D. Kurniadi, and B. Yuliarto, "Battery module performance improvement using active cell balancing system based on switched-capacitor boost converter (s-cbc)," in *2017 4th International Conference on Electric Vehicular Technology (ICEVT)*. IEEE, 2017, pp. 93–99.

[24] W. Diao, N. Xue, V. Bhattacharjee, J. Jiang, O. Karabasoglu, and M. Pecht, "Active battery cell equalization based on residual available energy maximization," *Applied energy*, vol. 210, pp. 690–698, 2018.

[25] J. Cao, N. Schofield, and A. Emadi, "Battery balancing methods: A comprehensive review," in *2008 IEEE Vehicle Power and Propulsion Conference*. IEEE, 2008, pp. 1–6.

[26] Y. Chen, X. Liu, H. K. Fathy, J. Zou, and S. Yang, "A graph-theoretic framework for analyzing the speeds and efficiencies of battery pack equalization circuits," *International Journal of Electrical Power & Energy Systems*, vol. 98, pp. 85–99, 2018.

[27] Q. Ouyang, N. Ghaeminezhad, Y. Li, T. Wik, and C. Zou, "A unified model for active battery equalization systems," *arXiv preprint arXiv:2403.03910*, 2024.

[28] Y. Weng and C. Ababei, "Ai-assisted reconfiguration of battery packs for cell balancing to extend driving runtime," *Journal of Energy Storage*, vol. 84, p. 110853, 2024.

[29] H. Huang, A. M. Ghias, P. Acuna, Z. Dong, J. Zhao, and M. S. Reza, "A fast battery balance method for a modular-reconfigurable battery energy storage system," *Applied Energy*, vol. 356, p. 122470, 2024.

[30] S. B. Joseph, E. G. Dada, A. Abidemi, D. O. Oyewola, and B. M. Khammas, "Metaheuristic algorithms for pid controller parameters tuning: Review, approaches and open problems," *Heliyon*, 2022.

[31] M. G. Abdolrasol, M. Hannan, S. S. Hussain, and T. S. Ustun, "Optimal pi controller based pso optimization for pv inverter using spwm techniques," *Energy Reports*, vol. 8, pp. 1003–1011, 2022.

[32] G. Dulac-Arnold, N. Levine, D. J. Mankowitz, J. Li, C. Paduraru, S. Gowal, and T. Hester, "Challenges of real-world reinforcement learning: definitions, benchmarks and analysis," *Machine Learning*, vol. 110, no. 9, pp. 2419–2468, 2021.

[33] Y. Yang, J. He, C. Chen, and J. Wei, "Balancing awareness fast charging control for lithium-ion battery pack using deep reinforcement learning," *IEEE Transactions on Industrial Electronics*, vol. 71, no. 4, pp. 3718–3727, 2023.

[34] D. Karnehm, S. Pohlmann, and A. Neve, "State-of-charge (SoC) balancing of battery modular multilevel management (bm3) converter using q-learning," in *2023 IEEE Green Technologies Conference (GreenTech)*. IEEE, 2023, pp. 107–111.

[35] Y. Liang, Z. Ding, T. Zhao, and W.-J. Lee, "Real-time operation management for battery swapping-charging system via multi-agent deep reinforcement learning," *IEEE Transactions on Smart Grid*, vol. 14, no. 1, pp. 559–571, 2022.

[36] M. H. Lipu, M. Hannan, T. F. Karim, A. Hussain, M. H. M. Saad, A. Ayob, M. S. Miah, and T. I. Mahlia, "Intelligent algorithms and control strategies for battery management system in electric vehicles: Progress, challenges and future outlook," *Journal of Cleaner Production*, vol. 292, p. 126044, 2021.

[37] J. Jin, S. Mao, and Y. Xu, "Optimal priority rule enhanced deep reinforcement learning for charging scheduling in an electric vehicle battery swapping station," *IEEE Transactions on Smart Grid*, 2023.

[38] Z. Zhang, T. Zhang, J. Hong, H. Zhang, and J. Yang, "Energy management strategy of a novel parallel electric-hydraulic hybrid electric vehicle based on deep reinforcement learning and entropy evaluation," *Journal of Cleaner Production*, vol. 403, p. 136800, 2023.

[39] H. Chen, G. Guo, B. Tang, G. Hu, X. Tang, and T. Liu, "Data-driven transferred energy management strategy for hybrid electric vehicles via deep reinforcement learning," *Energy Reports*, vol. 10, pp. 2680–2692, 2023.

[40] X. Tang, J. Chen, T. Liu, Y. Qin, and D. Cao, "Distributed deep reinforcement learning-based energy and emission management strategy for hybrid electric vehicles," *IEEE Transactions on Vehicular Technology*, vol. 70, no. 10, pp. 9922–9934, 2021.

[41] D. Qiu, Y. Wang, W. Hua, and G. Strbac, "Reinforcement learning for electric vehicle applications in power systems: A critical review," *Renewable and Sustainable Energy Reviews*, vol. 173, p. 113052, 2023.

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

[43] A. B. Kordabad, W. Cai, and S. Gros, "Multi-agent battery storage management using mpc-based reinforcement learning," in *2021 IEEE Conference on Control Technology and Applications (CCTA)*. IEEE, 2021, pp. 57–62.

[44] M. Afrasiabi, M. Mohammadi, M. Rastegar, and A. Kargarian, "Multi-agent microgrid energy management based on deep learning forecaster," *Energy*, vol. 186, p. 115873, 2019.

[45] D. Yu, H. Zhu, W. Han, and D. Holburn, "Dynamic multi agent-based management and load frequency control of pv/fuel cell/ wind turbine/ chp in autonomous microgrid system," *Energy*, vol. 173, pp. 554–568, 2019. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0360544219302890

[46] M. Hua, C. Zhang, F. Zhang, Z. Li, X. Yu, H. Xu, and Q. Zhou, "Energy management of multi-mode plug-in hybrid electric vehicle using multi-agent deep reinforcement learning," *Applied Energy*, vol. 348, p. 121526, 2023. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0306261923008905

[47] Y. Wang, Y. Wu, Y. Tang, Q. Li, and H. He, "Cooperative energy management and eco-driving of plug-in hybrid electric vehicle via multi-agent reinforcement learning," *Applied Energy*, vol. 332, p. 120563, 2023. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0306261922018207

[48] Y. Sui and S. Song, "A multi-agent reinforcement learning framework for lithium-ion battery scheduling problems," *Energies*, vol. 13, no. 8, p. 1982, 2020.

[49] Y. Liang, Z. Ding, T. Zhao, and W.-J. Lee, "Real-time operation management for battery swapping-charging system via multi-agent deep reinforcement learning," *IEEE Transactions on Smart Grid*, vol. 14, no. 1, pp. 559–571, 2022.

[50] Z. Tang, X. Yang, and Y. Feng, "A novel reinforcement learning balance control strategy for electric vehicle energy storage battery pack," *International Journal of Low-Carbon Technologies*, vol. 19, pp. 1968–1980, 2024.

[51] S. Baccari, M. Tipaldi, and V. Mariani, "Deep reinforcement learning for cell balancing in electric vehicles with dynamic reconfigurable batteries," *IEEE Transactions on Intelligent Vehicles*, 2024.

[52] F. Yang, F. Gao, B. Liu, and S. Ci, "An adaptive control framework for dynamically reconfigurable battery systems based on deep reinforcement learning," *IEEE Transactions on Industrial Electronics*, vol. 69, no. 12, pp. 12 980–12 987, 2022.

[53] D. Flessner, J. Chen, and G. Xiong, "Reinforcement learning-based event-triggered active-battery-cell-balancing control for electric vehicle range extension," *Electronics*, vol. 13, no. 5, p. 990, 2024.

[54] B.-H. Bao-Huy Nguyn, J. P. F. Trovão, S. Jeme, L. Boulon, and A. Bouscayrol, "IEEE VTS motor vehicles challenge 2021 - energy management of a dual-motor all-wheel drive electric vehicle," in *2020 IEEE Vehicle Power and Propulsion Conference (VPPC)*, 2020, pp. 1–6.

[55] A. Bouscayrol, B. Davat, B. De Fornel, B. Francois, J. P. Hautier, F. Meibody-Tabar, and M. Pietrzak-David, "Multi-converter multi-machine systems: application for electromechanical drives," *The European physical journal applied physics*, vol. 10, no. 2, pp. 131–147, 2000.

[56] K. S. Ng, C.-S. Moo, Y.-P. Chen, and Y.-C. Hsieh, "Enhanced coulomb counting method for estimating state-of-charge and state-of-health of lithium-ion batteries," *Applied energy*, vol. 86, no. 9, pp. 1506–1511, 2009.

[57] B. Jiang, Y. Liu, and J. Tang, "Lithium-ion battery state of health estimation with recurrent convolution neural networks," in *11th International Conference on Power Electronics, Machines and Drives (PEMD 2022)*, vol. 2022, 2022, pp. 479–484.

[58] A. Oroojlooy and D. Hajinezhad, "A review of cooperative multi-agent deep reinforcement learning," *Applied Intelligence*, vol. 53, no. 11, pp. 13 677–13 722, 2023.

[59] H. Ryu, H. Shin, and J. Park, "Cooperative and competitive biases for multi-agent reinforcement learning," *arXiv preprint arXiv:2101.06890*, 2021.

[60] Y. Zhou, S. Liu, Y. Qing, K. Chen, T. Zheng, Y. Huang, J. Song, and M. Song, "Is centralized training with decentralized execution framework centralized enough for marl?" *arXiv preprint arXiv:2305.17352*, 2023.

[61] X. Lyu, Y. Xiao, B. Daley, and C. Amato, "Contrasting centralized and decentralized critics in multi-agent reinforcement learning," *arXiv preprint arXiv:2102.04402*, 2021.

Armin Lotfy (Student Member, IEEE) received the M.Sc. degree in Electrical Engineering from the Iran University of Science and Technology, Tehran, Iran, in 2019. In 2021, he commenced his Ph.D. studies in the Department of Electronics at Carleton University, Ottawa, Canada, where he became a member of the Intelligent Robotic and Energy Systems (IRES) Research Group. His research is centered on the application of advanced machine learning methodologies—including deep learning and reinforcement learning—to address complex challenges in energy systems. His primary interests include the development of intelligent energy management strategies for battery electric vehicles (BEVs), optimization of battery and hybrid energy storage systems, and the integration of predictive and adaptive control algorithms to improve the efficiency, reliability, and sustainability of energy-related technologies.

Mohamad Alzayed (M'20) received the Ph.D. degree in Electrical and Computer Engineering from Carleton University, Ottawa, ON, Canada. His career has bridged both academia and industry with more than 14 years of experience in control and electrical energy resources, grid management and efficiency, electrical design and contracting, and project management. Since 2019, he has been a member of the Intelligent Robotic and Energy Systems (IRES) research group, Department of Electronics, Carleton University, and joined the Laboratory of Signal and System Integration (LSSI), Department of Electrical and Computer Engineering, the University of Quebec at Trois-Rivieres, in 2022. His teaching/research activities focus on intelligent control of electric machines and power converters for energy systems, real-time simulations, and power systems. He is a member of IEEE and a Guest Editor of several journals.

Hicham Chaoui (Senior Member, IEEE) received the Ph.D. degree in electrical engineering from the University of Quebec, Trois-Rivières, QC, Canada, in 2011. From 2007 to 2014, he held various engineering and management positions with Canadian Industry. He is currently an Associate Professor with Carleton University, Ottawa, ON, Canada. He is a Registered Professional Engineer in the Province of Ontario. His career has spanned both academia and industry in the field of control and energy systems. His scholarly work has resulted in more than 200 journal and conference publications. He was a recipient of the Best Thesis Award, the Governor General of Canada Gold Medal Award, the Carleton's Research Excellence Award, the Early Researcher Award from Government of Ontario, and the Top Editor Recognition from the IEEE Vehicular Technology Society. He is an associate editor of several IEEE journals.

loïc boulon (Senior Member, IEEE) received the master's degree in electrical and automatic control engineering from the University of Lille, Lille, France, in 2006, and the Ph.D. degree in electrical engineering from the University of Franche-Comté, Besançon, France. Since 2010, he has been a Professor with Université du Québec (UQTR), Trois-Rivières, QC, Canada, attaining the rank of Full Professor in 2016, and the Deputy Director of the Hydrogen Research Institute since 2019. His research interests include modeling, control, and energy management of multiphysics systems, with interests spanning hybrid electric vehicles, energy and power sources, such as fuel cells, batteries, and ultracapacitors. He has authored more than 140 peer-reviewed papers in international journals and conferences and has delivered over 40 invited talks worldwide. In 2022, he was recognized as one of the top 10 most prolific authors globally in "Proton Exchange Membrane Fuel Cell (PEMFC)" research and ranked in the top 20 for "Plug-in Hybrid Vehicles," as identified by Elsevier SciVal. In 2015, he was the General Chair of the IEEE Vehicular Power and Propulsion Conference in Montréal, Canada. He is VP-Motor Vehicles Vehicles for the IEEE Vehicular Technology Society. He founded the "International Summer School on Energetic Efficiency of Connected Vehicles" and the "IEEE VTS Motor Vehicle Challenge." He holds the prestigious Canada Research Chair in Energy Sources for Vehicles of the Future and is the Director of the Réseau Québécois sur l'Énergie Intelligente.