Optimize Battery Control: A Multi-Objective Evolutionary Ensemble Reinforcement Learning Approach

Jingwei Hu[1] , **Kai Xie**[2] , **Zheng Fang**[1] , **Xiaodong Li**[1] , **Junchi Yan**[3] and **Zhihong Zhang**[1,2*]
[1]School of Informatics, Xiamen University
[2]Institute of Artificial Intelligence, Xiamen University
[3]Dept. of CSE & School of AI & MoE Key Lab of AI, Shanghai Jiao Tong University

Abstract

The Dynamically Reconfigurable Battery (DRB) systems, which use high-speed power electronic switches to dynamically adjust battery interconnections in real-time, are critical to the performance of the battery pack. Traditional battery management strategies often fail to address multi-objective optimization, leading to imbalanced performance and inadequate energy utilization. To enhance decision-making across multiple objectives, an Evolutionary Ensemble Reinforcement Learning (EERL) framework is proposed in this paper. This framework incorporates evolutionary algorithms to associate ensemble learning, thus improving reinforcement learning (RL) performance. It decomposes a complex objective into multiple sub-objectives, each optimized independently, while incorporating diverse performance metrics into the correlation stage to derive the Pareto optimal solution. The EERL can efficiently mitigate potential adverse effects such as short circuits, disconnections, and reverse charging, thereby effectively reducing capacity differences among various batteries. Simulations and real-world testing demonstrate that the proposed approach overcomes the issue of local optima entrapment in multi-objective optimization scenarios. In a real-world system, an 11.08 % increase in energy efficiency is observed compared to existing approaches.

1 Introduction

Lithium-ion battery energy storage systems are extensively utilized due to their high energy density, lack of memory effect, and long cycle life [Lu *et al.*, 2023b]. These systems serve as flexible and adjustable solutions for power charging and discharging, enabling the temporal and spatial conversion of energy to meet the demands of various scenarios [Matos *et al.*, 2019].

A battery system typically comprises multiple cells to meet a broad range of application requirements [Dai *et al.*, 2021; Deng *et al.*, 2020]. However, when these cells are assembled into a battery pack, the overall performance and energy utilization of the pack are significantly influenced by the consistency and aging issues of individual cells. Inherent variations in production processes and operational conditions may result in inconsistent performance among the cells within a battery pack [Lin and Ci, 2017]. This inconsistency tends to exacerbate over time with usage. The degradation of individual cells diminishes the overall capacity of the battery pack and poses safety risks, thereby shortening its operational lifespan [Turksoy *et al.*, 2020]. As illustrated in Fig. 1(a), a fixed topology cannot rectify cell inconsistencies. Consequently, the study of *battery equalization technology* is crucial.

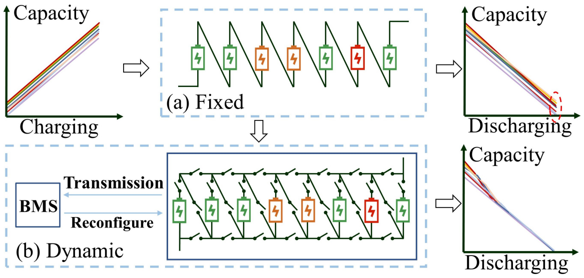


Figure 1: In a fixed topology, energy is constrained by the weakest battery, while the DRB system, controlled by the Battery Management System (BMS), enables complete energy release.

Currently, battery equalization strategies rely on additional energy consumption through component connections [Ismail *et al.*, 2017] or energy transfer via conversion modules [Park *et al.*, 2023]. However, the introduction of an extra equalization discharge flux leads to a significant reduction in energy efficiency [Cui *et al.*, 2022]. To address this issue, Dynamically Reconfigurable Battery (DRB) systems, as depicted in Fig. 1(b), have been proposed. These systems consist of interconnected cells controlled by high-speed power electronic switches. The system dynamically modulates interconnections in real-time by adjusting the series-parallel configuration, aligning with load demands and battery cell conditions, ultimately optimizing the performance [Yang *et al.*, 2022].

The number of switches in a DRB system affects the diversity of configurations it can support [Cui *et al.*, 2022]. More switches increase adaptability in battery management but add complexity to control strategies. With more switches, the risk of short circuits or disconnections grows, highlight-

[*]Corresponding Author: zhihong@xmu.edu.cn

ing the need for methods to balance performance and safety among the exponential growth of connection configurations. Designing adaptive control strategies is challenging due to the complex interplay of requirements and constraints. The combination of different targets as rewards often fails to achieve optimal performance, making reward engineering a significant challenge.

Conventional strategies often fail to improve overall system performance. While they may achieve energy equalization, they usually cannot well address challenges such as connectivity hazards, stability issues, and power losses. In DRB systems, ensuring the safety of all topological configurations is crucial and optimizing performance involves balancing the energy of the battery pack while avoiding additional adverse effects, which is framed as a multi-objective optimization problem within the decision-making process. Scalarizing multi-objective using intricate reward engineering techniques may lead to convergence at a local optimum. Futhermore, existing multi-objective reinforcement learning (RL) strategies are typically restricted by specific frameworks [Yang *et al.*, 2019], limiting their applicability to tasks involving multi-dimensional continuous action spaces, and often fail to ensure power system security. The reliance on pre-defined weight indices is also inadequate for reward settings involving complex interdependencies. To address these challenges, we propose an Evolutionary Ensemble Reinforcement Learning (EERL) framework.

The proposed framework aims to improve model performance on specific objectives by employing ensemble learning. The evolutionary algorithm prevents the imbalanced development of agents on corresponding objectives. By utilizing non-dominated sorting with multiple objectives, the model can focus on individual sub-tasks while ensuring a synergistic enhancement in model performance across multiple dimensions. Furthermore, a buffering mechanism for the agents, avoids the direct replacement of the actor. This refinement strengthens the agent's robustness during the training phase. In applying the DRB systems, this approach targets diverse requirements into several sub-tasks, thereby enhancing the system's performance through targeted learning. The main highlights of the framework are as follows.

- To the best of our knowledge, EERL is the first approach to apply RL to multi-objective policy control in DRB systems.
- The EERL framework effectively addresses local optima arising from the multi-objective scalarization in existing RL methods.
- An enhanced method for updating evolutionary RL is presented, expediting the optimization process of a multi-objective battery system.
- A real prototype is developed for testing, as well as rigorous simulations of the DRB systems.

2 Related Work

2.1 Dynamically Reconfigurable Battery

The battery energy storage system typically comprises multiple individual cells interconnected via a static framework. During operation, discrepancies in voltage, internal resistance, and capacity among these cells tend to accumulate and exacerbate [Morstyn *et al.*, 2015]. An effective structure that mitigates these disparities through dynamic connections between different batteries is known as the DRB systems [Lin *et al.*, 2018]. A battery system, denoted by $(b_1, b_2, \ldots, b_{n_c})$, is represented by n cells. The state of charge (SOC) of cells can be described as $(soc_1, soc_2, \ldots, soc_{n_c})$. The initial value of SOC is related to factors such as production conditions and usage environment. We define the current SOC for each cell in the DRB systems as follows:

$$SOC^t = SOC^{t-1} + c \cdot s \cdot \Delta SOC^{t-1} \tag{1}$$

Where t denotes the current moment, Δ denotes the consumption of SOC during the time period. The variable c represents the charging and discharging case, with 1 indicating charging and -1 indicating discharging, and s denotes the on-off state of the battery as controlled by the switches, with 0 indicating the battery is off and 1 indicating the battery is connected to the system operation.

The battery system encompasses a variety of topologies. The SOC of different batteries can be maintained in a relatively balanced state through continuous adjustments. Researchers have explored various techniques to achieve battery pack equalization. Initial studies concentrated on static series and parallel configurations [Alvarez-Diazcomas *et al.*, 2020; Samanta and Chowdhuri, 2021]. In certain researches, battery equalization has been achieved through the integration of resistors, capacitors, and additional circuit components. Despite the ease of implementation, this approach incurs higher energy consumption and significant heat generation [Alvarez-Diazcomas *et al.*, 2020]. To mitigate the substantial reactive power losses, researchers have employed power electronic switches for battery pack equalization [Jiang *et al.*, 2023]. Some researchers have determined the optimal configuration of the system through dynamic planning, path search [Lin and Ci, 2017; He *et al.*, 2019]. However, these models are simplified to alleviate computational complexity, focusing on a single objective such as energy or voltage difference. These strategies neglect a system-level perspective that incorporates multiple objectives. Several researchers have explored the application of RL in controlling DRB systems [Yang *et al.*, 2023]. However, the inherent complexity of training and configuring RL frameworks has largely limited their application to single-switch topologies in DRB systems [Lu *et al.*, 2024], precluding direct multi-switch control without reducing the number of actions. Furthermore, these approaches often rely on a limited set of features, which undermines their ability to optimize overall system performance.

2.2 Multi-Objective Reinforcement Learning

The RL has demonstrated its effectiveness in tackling continuous decision-making problems across various fields, such as robot training [Dalal *et al.*, 2021; Akalin and Loutfi, 2021], workshop scheduling [Zhang *et al.*, 2020; Zhang *et al.*, 2022], and urban planning [Lin *et al.*, 2024; Peng *et al.*, 2021]. By formulating problems as Markov Decision Processes (MDPs), RL models sequential decision-making, where actions of the agent influence subsequent states and future rewards [Van Otterlo and Wiering, 2012]. The optimal control strategy is derived through iterative processes. Traditional methods often struggle with constructing accurate and efficient mathematical models for optimization problems [Bengio *et al.*, 2021]. In contrast, RL can overcome these limitations by learning optimal strategies through interaction data [Nguyen *et al.*, 2020a]. For improving the DRB systems performance, RL can explore the impact of battery factors on performance in complex scenarios and maximize rewards through agent decisions.

In practical scenarios, decision-making involves multi-objective optimization, where rewards are represented as vectors instead of scalars [Nguyen *et al.*, 2020b]. Unlike single-objective RL tasks, where the focus is on maximizing the cumulative reward, multi-objective tasks necessitate the consideration of conflicts and constraints among multiple objectives, thereby amplifying the complexity of the decision-making process [Pirotta *et al.*, 2015]. Some researchers have attempted to combine multiple objectives into a single scalarized objective [Prabhakar *et al.*, 2022], but this approach can result in reaching a local optimum, which impedes the ability of the agent to make effective trade-offs [Van Moffaert and Nowé, 2014]. Some researchers have addressed multi-objective requirements by learning distinct state-hidden representations tailored to specific task preferences [Shu *et al.*, 2024]. Others have explored linear scalarization based methods to identify the convex set of Pareto front [Alegre *et al.*, 2023; Chen *et al.*, 2019]. However, in many real-world RL applications, specific preference or linear scalarization are unavailable or difficult to specify, posing significant challenges for effective training and deployment.

3 Methodology

3.1 Problem Definition

To achieve equalization of the DRB system, a reasonable action a^* needs to be obtained. A topological control set $A = \{a_1, a_2, \ldots, a_m\}$ is defined for the DRB system, where m is obtained as an exponent of the number of cells. This problem is formulated as a sequential decision-making process, where the agent interacts with the environment to pursue multiple objectives, which are jointly represented by the reward function R and constrained by a set of constraints C. The goal is to find an action $a^* \in A$ that maximizes the reward while softly satisfying the constraints:

$$a^* = \arg\max_{a_{se} \in A} R(a_{se}, \text{Env}(s, a_{se})) \quad \text{subject to } C(a_{se}) \geq 0 \tag{2}$$

where a_{se} represents the selected action, s represents the state after executing the action, and Env is the environment obtained based on the action and state. However, as the number of strategies in DRB system increases exponentially with the number of cells, it is challenging to choose a reasonable topological pattern among the exponential level of paths. The complex task design causes the agent to fall into local optima during learning. Especially in the scenarios with strict safety conditions, the agents can not take into account multiple objectives well, which makes the performance unstable and prone to serious safety hazards. As mentioned in equation 2 with the constraints C should be satisfied to the greatest extent possible. It represents several conditions such as security, equilibrium, stability, and multiple parallel.

3.2 EERL Framework

The decomposition model of ensemble learning in RL is initially introduced, followed by an explanation of how evolutionary algorithms integrate the decomposition model to solve multi-objective problems. In principle, most RL methods can be incorporated into this framework. In the DRB experiments, the RL policy employs the Soft Actor-Critic (SAC) algorithm exclusively. This choice is based on the ability of SAC to integrate effectively with a variational autoencoder (VAE) for efficient selection within large-scale discrete action spaces [Hu *et al.*, 2025]. Unlike conventional settings, the sparse rewards in this large action space lead to severely degraded performance of standard RL models. A reparameterization strategy is used in the VAE, which is the same as the sampling strategy of SAC, as illustrated in the following equation:

$$z = \mu + \sigma \cdot \epsilon \tag{3}$$

where μ and σ denote the mean and variance of the output from the network. ϵ denotes the random variable drawn from the standard normal distribution. z denotes the obtained latent variable.

With the current RL framework, direct policy optimization by combining all objectives and using composite rewards can result in sub-optimal control strategies. Conflicting or inconsistent objectives may hinder the agents from finding a optimal solution in complex tasks. We propose a decomposition and correlation approach, illustrated in Fig. 2. The complex task is decomposed into multiple sub-tasks, executed in randomized form on different agents, and trained using ensemble learning. The evolutionary algorithm maintains a population for updating the randomly selected sub-target agent. The best-performing agent is refined through crossover and mutation processes.

The method decomposes a complex task into multiple simple sub-tasks. Each sub-task is trained by an independent agent that focuses on learning the optimal policy for the simple goal, as shown in the following equation:

$$\pi_i = \arg\max_{\pi} E\left[\sum_{t=0}^{\infty} \gamma^t r_t^i \mid \pi\right] \tag{4}$$

where π denotes the policy, and E denotes the expectation over trajectories. The r_t^i represents the reward acquired for sub-task i at time step t and γ is the discount factor. The decomposition helps to reduce the complexity of the problem, allowing each agent to learn and optimize more efficiently. This ensemble learning approach in the case of inputs with the same state, each agent will update its network parameters according to the corresponding objective. Therefore, the concept of evolution is applied to construct a model for updating evolutionary algorithms in a composite. During training, the population considers each objective as a criterion of superiority and inferiority. For each sub-task i agent π_i is associated with multiple objectives $O = \{o_1, o_2, \ldots, o_n\}$, and the Pareto frontier of different population (including alternative actor and real actor) are computed, which allows the agent

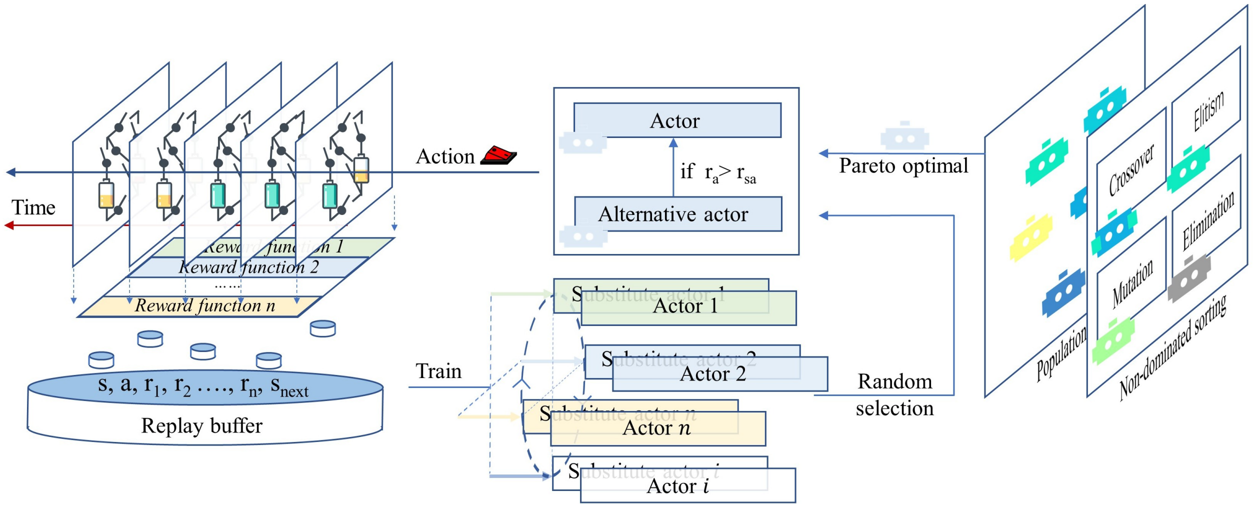


Figure 2: EERL training process. A randomly selected agent performs the task in each round. Ensemble learning leverages multiple reward functions, each corresponding to a specific objective, to reward and train the agents. An evolutionary algorithm integrates with agent selection in ensemble learning, identifying the current optimal agent based on Pareto optimality.

performing sub-tasks to be associated to multiple objectives during the updating process. A single objective is used as a criterion to judge whether the alternative actor is better than the real actor. If $o_i(\pi_i^{alt}) > o_i(\pi_i)$ then $\pi_i \leftarrow \pi_i^{alt}$, which is used as a substitution requirement. The alternative actor and the best individual in the population are then obtained by non-dominated sorting. We define two individuals π_i and π_j, if $O(\pi_i) \geq O(\pi_j)$ and existence of $o_l(\pi_i) > o_l(\pi_j)$ which represent dominance relation. The absence of direct substitution of real actors is intended to introduce a buffer mechanism capable of managing the balance between exploration and exploitation during policy updating. Storing the best-performing agent as an alternative actor instead of immediately using it ensures superior performance after updates and prevents retaining poorly updated or accidental individuals, thereby improving efficiency.

The combination of the two strategies enables agents to focus on a single objective while being constrained by others, ensuring excellence in one sub-task and satisfactory performance in others. When the training is complete, the process of using is shown in Fig. 3. The actions generated by different agents are scored by all the critics. By summation, the highest rated action is output. Each critic performs an empirical replay evaluation before the task, linearly mapping the evaluation range to $[\alpha, \beta]$, where α and β are the minimum and maximum values respectively, so that evaluation values of different criteria can be compared. We perform the conversion according to the following linear mapping formula:

$$score_{c_i} = \alpha + \frac{\left(c_i - c_i^{\min}\right) \cdot (\beta - \alpha)}{c_i^{\max} - c_i^{\min}} \tag{5}$$

where $score_{c_i}$ denotes the transformed value of critic, c_i denotes the initial value obtained from critic, and $c_i^{\min}$ and $c_i^{\max}$ denote the minimum and maximum values in the additional evaluations. The conversion function can be used to obtain the evaluation value in the same benchmark.

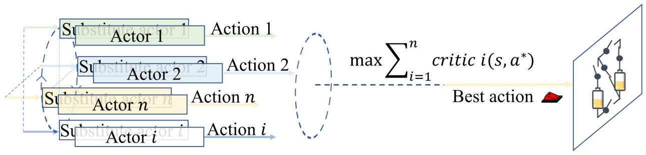


Figure 3: EERL testing process. The actions generated by different agents are scored by all critics outputting the most appropriate action through normalized comparison.

3.3 RL in the DRB System

To better describe the MDPs of the DRB system to illustrate specific applications, the RL framework is further illustrated. The DRB system in Fig. 1(b) allows for different connections through the opening and closing of switches, which include legal and illegal structures. The switching change in the DRB topology is modelled as a MDPs with quaternion (s, a, p, R), where s is the state, a is the action, p is the state transition probability derived from system dynamics (implicitly defined), and R is the reward. The specific settings are as follows:

State: The state encompasses the agent's perception of the battery module, including the capacities of individual batteries, external load demands, battery count, and their structural characteristics.

Action: The action variable is a 6-dimensional continuous variable ranging from 0 to 1. It represents a hidden state that controls switches in the system after training with the VAE. The system includes $4n_c - 3$ switches, leading to 2^{4n_c-3} potential discrete actions that must be evaluated. Directly managing an exponential array of actions using a neural network is impractical due to the high dimensionality of the space. Consequently, a VAE is introduced to efficiently transform the learned hidden variables. It represents a hidden state that controls switches in the system after training with the VAE.

Algorithm 1 Sub-tasks Rewards.

Input: SOC_{t-1}, SOC_t, $switch_{t-1}$, $switch_t$
Output: $r1, r2, r3$

```
if Legal then
    if Satisfy the load then
        $r_{parallel} \leftarrow$ if $parallel$ then
                if $all(I) \geq 0$ then $1$ else $0.2$ else $0.6$
        if ($\Delta SOC_{t-1} < 1$ and $\Delta SOC_t < 1$) or ($\Delta SOC_{t-1} > \Delta SOC_t$) then
            $r1 \leftarrow 1$
            $r2 \leftarrow sum(switch_{t-1} = switch_t)/num\_switch$
            $r3 \leftarrow r_{parallel}$
        else
            $r1, r2, r3 \leftarrow 0.05$
        end if
    else
        $r1, r2, r3 \leftarrow -0.05$
    end if
else
    $r1, r2, r3 \leftarrow -1$
end if
```

Reward: The reward signifies whether the agent's actions align with the task objectives. Our primary aim is to manage DRB systems so that battery capacity differences diminish as they meet external load demands. We also aim to maximize parallel connections of batteries due to the rate-capacity effect, where discharging all cells together yields greater capacity than partial discharge. Additionally, we strive for the last controlled action to closely resemble the current action being executed, which improves system stability. The existence of correlations and exclusions between multiple requirements is complex and needs to be decomposed. The original task reward is decomposed into $r1, r2, r3$ in Alg. 1.

Following the above analysis, the RL-based DRB system exhibit a comprehensive structure, defined by the establishment of states, actions, and rewards.

4 Results

4.1 Environment and Setting

The DRB system is developed and experimented on SIMULINK. The system comprises eight cells and twenty-nine switches, with each cell modeled according to a predefined discharge profile [Zhu *et al.*, 2013]. Our simulations provide the current, voltage, and SOC of the cells, achieving series and parallel connections by controlling the switches. The environment supports interaction with Python. Furthermore, tests are conducted in real-world scenarios.

In the experiments, fifty training iterations are conducted for each discharge process, with each discharge consisting of 150 steps. We establish three sub-tasks with set rewards and employ an evolutionary algorithm to maintain a population of twelve individuals. In each training round, 256 sample points are selected, with the discount factor set to 0.99, the learning rate to 3e-4, and the entropy coefficient to 0.12. In the real-world application scenario, the INR 18650-20R model battery, which has a rated capacity of 2000 mAh, is used. The training duration is 160 hours, while the inference time is approximately 1.00 ms. Both the inference time and the hardware execution time are within the millisecond range, indicating that the model's inference is exceptionally efficient. The time required for the microcontroller to transmit the collected data back to the host computer from 57.86 to 87.23 ms, which is considered reasonable.

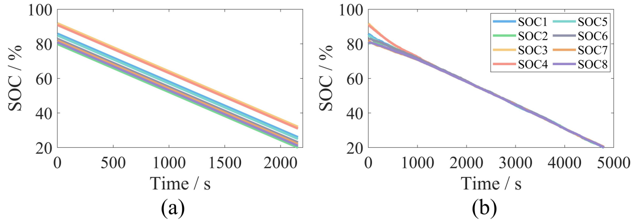


Figure 4: The SOC discharges are analyzed for batteries under various configurations. (a) operate within a fixed topology, (b) utilize EERL within the DRB system.

4.2 System Operation Effect

To evaluate the performance of the proposed framework, the batteries are placed in a fixed topology and the enhanced framework. As depicted in Fig. 4, the framework demonstrates superior performance on the DRB system, as it accounts for the total battery capacity to achieve equalization.

In fixed topology operation, all cells are connected in series, and each cell experiences the same current. As a result, disparities in the SOC among cells persist. When the weakest cell reaches its lower threshold, the remaining energy of the other cells cannot be harnessed. With a fixed topology, there is no significant change in SOC differences through sustained discharging, which would be more pronounced with a completely different battery attributes. Applying our proposed method to the DRB system resulted in a gradual reduction of SOC variance during operation. It is reduced from 12 % at the beginning to less than 1 %, specifically 0.427 %. The amount of energy released from each cell also improves from 4.356 Wh to 4.701 Wh upon the implementation of EERL, resulting in a significant improvement of 7.920 %.

4.3 Modular Enhancement

To evaluate the improvement offered by the proposed method in the baseline, we train the baseline and the model without employing evolutionary algorithms to bootstrap the various sub-tasks. The baseline employs the SAC algorithm as the RL framework without incorporating additional modules. Building on this, ensemble learning is integrated to manage various sub-tasks by simplifying complex goals. Finally, we enhance the system by optimizing and adjusting these sub-objectives through evolutionary learning, allowing the overall update process to consider multiple objectives. Six tests are conducted with randomized initial values for each SOC between 80% and 95%. The average results for different objectives are presented in Table 1. Note that higher values indicate better performance for energy and safe paralleling, whereas lower values are required for switch operations and illegal connections.

Significant improvements in the EERL framework are observed. The average energy released per battery increased

Algorithm	Baseline	Ensemble	Proposed
Energy / Wh	4.847E+00	4.910E+00	**4.913E+00**
operations / Times	7.379E+00	6.374E+00	**6.303E+00**
Safe Paralleling	8.060E-01	8.688E-01	**8.847E-01**
illegal	3.736E-03	0.000E+00	**0.000E+00**

Table 1: Detailed comparison of additions to different modules. The average results of multiple experiments are shown.

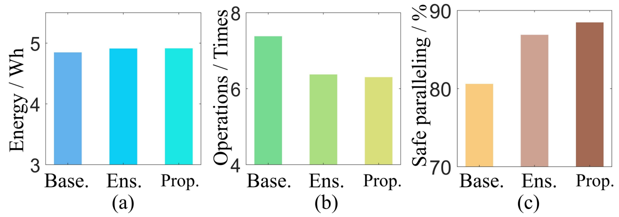


Figure 5: Comparison of average results across methods: Baseline (Base.), Ensemble (Ens.), and Proposed (Prop.). (a) – (c) represent the energy released by each cell, the number of actuated switches per decision, and the percentage of safe paralleling over total runs.

from 4.847 Wh to 4.913 Wh compared to the baseline, indicating more efficient utilization of the battery. The average number of switching actions per decision is reduced by 14.581 %, indicating the agent achieves battery balancing at a smaller action cost. During operation, the DRB system should be powered by more cells due to the rate capacity effect. By training the agent achieves safe parallel demand in 88.470 % of the cases during a discharge. Crucially, the discharging process may result in actions that still violate connection constraints due to the complexity of hidden variables. Illegal connections can poses a significant risk to systems. The proposed framework includes a pre-screening feature to ensure system safety by selecting actions that are both safe and have a high critic score. It reduces the original 0.374 % violation percentage to 0 % and ensures the safety, which is a huge improvement in the performance of the DRB system.

Compared to RL with the added ensemble strategy, EERL is more advantageous in several metrics. The table 1 demonstrates that the model with ensemble learning shows improvement over the baseline and has an advantage in individual metrics. This improvement arises because decomposing complex goals allows the agent to perform better on individual objectives. However, the absence of constraints from the evolutionary algorithm on multiple goals results in less pronounced improvements than EERL. Fig. 5 provides a more intuitive view, illustrating that EERL outperforms both the baseline and RL with ensemble learning.

To demonstrate the effectiveness of the improved evolutionary algorithm, both the basic and the improved evolutionary algorithms are trained within the EERL model. The basic achieves the optimal population of individuals by comparing non-dominated sorting results and using the best individuals as agents to perform tasks. We implemented a buffer mechanism, where optimal individuals are placed in alternative regions for observation. The replacement of the executing agent is carried out only after the network is updated and the optimal are confirmed. The reward boosts for the three sub-tasks during training are shown in Fig. 6. While both schemes perform well when fully trained, the improved evolutionary algorithm reaches the optimal result more quickly during the training process compared to the basic. The improved scheme avoids retaining poorly updated and accidentally excellent individuals.

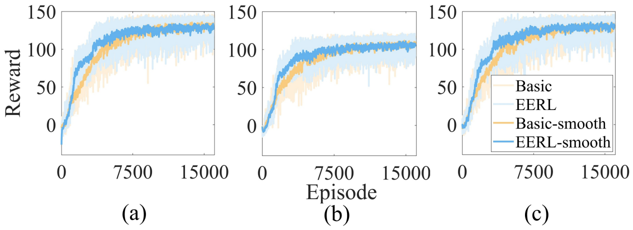


Figure 6: Rewards for different sub-tasks during training. (a) - (c) represent training with different rewards as described in Alg. 1.

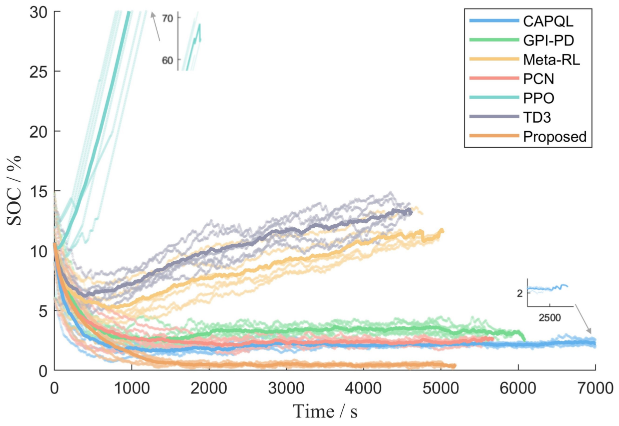


Figure 7: Variation in SOC of batteries under different algorithms.

4.4 Algorithm Comparison

A comparison is made for the application of current algorithms to the DRB system. Traditional RL strategies, PPO [Schulman *et al.*, 2017], and TD3 [Fujimoto *et al.*, 2018] are compared to evaluate whether they can learn effectively with different sampling methods. Different multi-objective optimization algorithms are applied for the multi-objective optimization nature of the model, such as CAPQL [Lu *et al.*, 2023a], GPI-PD [Alegre *et al.*, 2023], Meta-RL [Chen *et al.*, 2019], PCN [Reymond *et al.*, 2022]. Fig. 7 shows the variation in SOC the operation of the DRB system for the basic models and multi-objective models. Determining whether the battery energy can be fully discharged is the most important criterion for evaluating the quality of a battery pack. Each trained model tests six times using different initial values. Light colors indicate raw results and dark colors indicate average results.

The SAC demonstrates effective learning in equalizing the DRB system, as anticipated. In contrast, PPO and TD3 algorithms fail to converge efficiently, resulting in limited reward improvement. The different sampling approach from VAE in other RL methods is the essential reason for the inability to

Algorithm	CAPQL	GPI-PD	Meta-RL	PCN	PPO	TD3	Proposed
Energy / Wh	4.806E+00	4.640E+00	4.313E+00	4.824E+00	2.429E+00	4.314E+00	**4.913E+00**
operations / Times	4.060E+00	1.299E+01	1.112E+01	6.911E+00	**6.933E-01**	1.157E+01	6.303E+00
Safe Paralleling	2.943E-01	1.801E-01	1.184E-01	7.417E-01	1.974E-01	2.103E-01	**8.847E-01**
illegal	1.036E-02	5.252E-02	6.624E-02	5.156E-03	0.000E+00	5.975E-02	**0.000E+00**

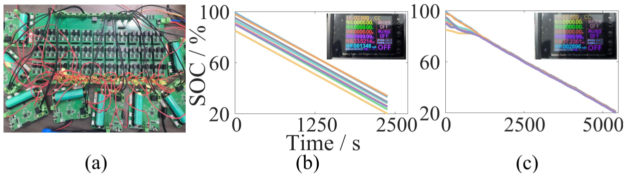
Table 2: Detailed comparison of different algorithms. The average results of multiple experiments are shown.

train effectively. SAC employs a reparameterization strategy where the sampling is the same as the VAE, and the reparameterization introduces tractable randomness that can be efficiently optimized by the back-propagation algorithm. PPO and TD3 employ probabilistic strategies and direct perturbations, respectively. Equalization is a fundamental requirement in DRB systems. However, basic algorithms struggle to meet this requirement, limiting their ability to optimize other objectives effectively. For Meta-RL, the training outcomes are suboptimal. Although a meta-learning model is incorporated into RL, the artificially defined weights increase the search space without enabling effective training under highly complex reward conditions. CAPQL converges to local optima during learning and fails to satisfy load response requirements in each decision-making step. Additionally, the discharge process does not align with specified external demands and consistently operates at low voltage levels, resulting in an unreasonably extended runtime. While GPI-PD and PCN demonstrate effective training, the SOC discrepancy between different cells remains approximately 2%, leading to reduced energy. Incorporating weights or preferences into the comparison algorithm and learning the Pareto frontier expands the learning space and increases task complexity, which may result in less effective decision-making within the current domain compared to the basic scheme without weights. Details of the comparison algorithms are provided in Table 2.

The proposed model demonstrates the highest energy release. The sampling method causes the PPO algorithm to fail in convergence, leading to repetitive execution of the same action. As illustrated in Fig. 7, the SOC difference expands rapidly. After excluding models that are not effectively trained, the proposed achieves maximum energy release with minimal actions and the highest percentage of safe parallelism. Notably, the proposed model ensures operational safety by avoiding illegal connections entirely. In contrast, the comparison algorithm shows occasional illegal connections despite effective training, posing significant safety risks.

4.5 Real-World Scenario Testing

In this section, a real DRB system is constructed for practical evaluation. Ensuring absolute safety during testing in real environments is crucial, as the comparison algorithm poses risks of battery shorts and incorrect series-parallel connections. Consequently, practical tests are conducted on fixed topology and the EERL model. The field-effect transistor switches, with their exceptional performance, enable rapid execution of control decisions. Decisions are made based on current and voltage feedback from the BMS, alongside the SOC values derived using the ampere-hour integration method. Control of the DRB system is executed by operating the array of circuit switches through control signals generated by the agent.



(a) (b) (c)

Figure 8: The SOC discharges in the Real-World Scenario. (a) represents established the real prototype, (b) operate within a fixed topology, (c) utilize EERL within the DRB system.

The actual hardware and experimental SOC curves are shown in Fig. 8. In real-world testing, the measured voltages during the run indicated that different batteries discharged at varying rates. To create a more pronounced difference among the battery packs at the outset, we discharged the batteries to different SOC. As illustrated in the Fig. 8(b), The feedback SOC difference ranges from 13.995 % to 13.910 % as all cells are continuously discharged in series. This indicates that most batteries do not reach the lower discharge threshold. The display connected to the load shows that the total discharged energy is 33.521 Wh. As illustrated in the Fig. 8(c), although there is a significant initial difference in SOC among the batteries, this disparity gradually decreases as the system operates. The feedback data indicates that the initial SOC difference is 13.967 %, which decreased to 1.078 % by the end of the run. The system discharged a total of 37.236 Wh, as observed on the monitor connected to the load. This represents a significant improvement of 11.083 %. The operation of a real energy system confirms the practicality of the proposed model.

5 Conclusion

In this paper, control strategies for the DRB system are investigated, which can be effectively applied to electric vehicles, uninterruptible power supplies, and the terraced utilization of retired batteries. By splitting the complex task and correlating different sub-tasks, we solve the problem of multi-objective control falling into local optima in traditional RL and achieve the performance improvement of the DRB system and battery pack equalization while ensuring safety. It is demonstrated that the EERL is capable of addressing the multi-objective policy control problem and performs well across several metrics. In addition, we conduct model testing in real scenarios to verify that the proposed model is sufficiently practical. For future work, the model is intended to be applied to high-capacity energy storage systems, with deeper problems in DRB systems analyzed for further optimization.

Acknowledgements

This research is supported by the National Natural Science Foundation of China (62176227); Fundamental Research Funds for the Central Universities (20720210047).

References

[Akalin and Loutfi, 2021] Neziha Akalin and Amy Loutfi. Reinforcement learning approaches in social robotics. *Sensors*, 21(4):1292, 2021.

[Alegre *et al.*, 2023] Lucas N. Alegre, Ana L. C. Bazzan, Diederik M. Roijers, Ann Nowé, and Bruno C. da Silva. Sample-efficient multi-objective learning via generalized policy improvement prioritization. In *Proceedings of the 22nd International Conference on Autonomous Agents and Multiagent Systems (AAMAS 2023)*, 2023.

[Alvarez-Diazcomas *et al.*, 2020] Alfredo Alvarez-Diazcomas, Adyr A Estévez-Bén, Juvenal Rodríguez-Reséndiz, Miguel-Angel Martínez-Prado, and Jorge D Mendiola-Santíbañez. A novel rc-based architecture for cell equalization in electric vehicles. *Energies*, 13(9):2349, 2020.

[Bengio *et al.*, 2021] Yoshua Bengio, Andrea Lodi, and Antoine Prouvost. Machine learning for combinatorial optimization: a methodological tour d'horizon. *European Journal of Operational Research*, 290(2):405–421, 2021.

[Chen *et al.*, 2019] Xi Chen, Ali Ghadirzadeh, Mårten Björkman, and Patric Jensfelt. Meta-learning for multi-objective reinforcement learning. In *2019 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pages 977–983. IEEE, 2019.

[Cui *et al.*, 2022] Haoyong Cui, Zhongbao Wei, Hongwen He, and Jianwei Li. Novel reconfigurable topology-enabled hierarchical equalization of lithium-ion battery for maximum capacity utilization. *IEEE Transactions on Industrial Electronics*, 70(1):396–406, 2022.

[Dai *et al.*, 2021] Haifeng Dai, Bo Jiang, Xiaosong Hu, Xianke Lin, Xuezhe Wei, and Michael Pecht. Advanced battery management strategies for a sustainable energy future: Multilayer design concepts and research trends. *Renewable and Sustainable Energy Reviews*, 138:110480, 2021.

[Dalal *et al.*, 2021] Murtaza Dalal, Deepak Pathak, and Russ R Salakhutdinov. Accelerating robotic reinforcement learning via parameterized action primitives. *Advances in Neural Information Processing Systems*, 34:21847–21859, 2021.

[Deng *et al.*, 2020] Youjun Deng, Yongxi Zhang, Fengji Luo, and Yunfei Mu. Operational planning of centralized charging stations utilizing second-life battery energy storage systems. *IEEE Transactions on Sustainable Energy*, 12(1):387–399, 2020.

[Fujimoto *et al.*, 2018] Scott Fujimoto, Herke Hoof, and David Meger. Addressing function approximation error in actor-critic methods. In *International conference on machine learning*, pages 1587–1596. PMLR, 2018.

[He *et al.*, 2019] Liang He, Linghe Kong, Yu Gu, Cong Liu, Tian He, and Kang G Shin. Extending battery system operation via adaptive reconfiguration. *ACM Transactions on Sensor Networks (TOSN)*, 15(1):1–21, 2019.

[Hu *et al.*, 2025] Jingwei Hu, Xinjie Li, Xiaodong Li, Zhensong Hou, and Zhihong Zhang. Optimizing reinforcement learning for large action spaces via generative models: Battery pattern selection. *Pattern Recognition*, 160:111194, 2025.

[Ismail *et al.*, 2017] Kristian Ismail, Asep Nugroho, Sunarto Kaleg, et al. Passive balancing battery management system using mosfet internal resistance as balancing resistor. In *2017 International Conference on Sustainable Energy Engineering and Application (ICSEEA)*, pages 151–155. IEEE, 2017.

[Jiang *et al.*, 2023] Bowen Jiang, Junfei Tang, Yujing Liu, and Luca Boscaglia. Active balancing of reconfigurable batteries using reinforcement learning algorithms. In *2023 IEEE Transportation Electrification Conference & Expo (ITEC)*, pages 1–6. IEEE, 2023.

[Lin and Ci, 2017] Ni Lin and Song Ci. Toward dynamic programming-based management in reconfigurable battery packs. In *2017 IEEE Applied Power Electronics Conference and Exposition (APEC)*, pages 2136–2140. IEEE, 2017.

[Lin *et al.*, 2018] Ni Lin, Song Ci, Dalei Wu, and Haifeng Guo. An optimization framework for dynamically reconfigurable battery systems. *IEEE Transactions on Energy Conversion*, 33(4):1669–1676, 2018.

[Lin *et al.*, 2024] ChungYi Lin, Shen-Lung Tung, Hung-Ting Su, and Winston H Hsu. Enhancing sustainable urban mobility prediction with telecom data: A spatio-temporal framework approach. In *Proceedings of the Thirty-Third International Joint Conference on Artificial Intelligence, IJCAI-24*, pages 7340–7348, 2024.

[Lu *et al.*, 2023a] Haoye Lu, Daniel Herman, and Yaoliang Yu. Multi-objective reinforcement learning: Convexity, stationarity and pareto optimality. In *The Eleventh International Conference on Learning Representations*, 2023.

[Lu *et al.*, 2023b] Jiahuan Lu, Rui Xiong, Jinpeng Tian, Chenxu Wang, and Fengchun Sun. Deep learning to estimate lithium-ion battery state of health without additional degradation experiments. *Nature Communications*, 14(1):2760, 2023.

[Lu *et al.*, 2024] Chenlei Lu, Dongji Xuan, Shengnan Liu, Jiaqi Tan, Haoqin Hu, Zehao Kang, and Liqu Lin. Active equalization control method for battery pack based on double-dqn. *Journal of Energy Storage*, 88:111361, 2024.

[Matos *et al.*, 2019] Catarina R Matos, Júlio F Carneiro, and Patrícia P Silva. Overview of large-scale underground energy storage technologies for integration of renewable energies and criteria for reservoir identification. *Journal of Energy Storage*, 21:241–258, 2019.

[Morstyn *et al.*, 2015] Thomas Morstyn, Milad Momayyezan, Branislav Hredzak, and Vassilios G Agelidis.

Distributed control for state-of-charge balancing between the modules of a reconfigurable battery energy storage system. *IEEE Transactions on Power Electronics*, 31(11):7986–7995, 2015.

[Nguyen *et al.*, 2020a] Thanh Thi Nguyen, Ngoc Duy Nguyen, and Saeid Nahavandi. Deep reinforcement learning for multiagent systems: A review of challenges, solutions, and applications. *IEEE transactions on cybernetics*, 50(9):3826–3839, 2020.

[Nguyen *et al.*, 2020b] Thanh Thi Nguyen, Ngoc Duy Nguyen, Peter Vamplew, Saeid Nahavandi, Richard Dazeley, and Chee Peng Lim. A multi-objective deep reinforcement learning framework. *Engineering Applications of Artificial Intelligence*, 96:103915, 2020.

[Park *et al.*, 2023] Kyung-Hwa Park, Minsu Lee, and Gun-Woo Moon. A new phase shift full bridge dc/dc converter with integrated inter-module battery equalization circuit (ibec). *IEEE Transactions on Transportation Electrification*, 2023.

[Peng *et al.*, 2021] Ningyezi Peng, Yuliang Xi, Jinmeng Rao, Xiangyuan Ma, and Fu Ren. Urban multiple route planning model using dynamic programming in reinforcement learning. *IEEE Transactions on Intelligent Transportation Systems*, 23(7):8037–8047, 2021.

[Pirotta *et al.*, 2015] Matteo Pirotta, Simone Parisi, and Marcello Restelli. Multi-objective reinforcement learning with continuous pareto frontier approximation. In *Proceedings of the AAAI conference on artificial intelligence*, volume 29, 2015.

[Prabhakar *et al.*, 2022] Prakruthi Prabhakar, Yiping Yuan, Guangyu Yang, Wensheng Sun, and Ajith Muralidharan. Multi-objective optimization of notifications using offline reinforcement learning. In *Proceedings of the 28th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, pages 3752–3760, 2022.

[Reymond *et al.*, 2022] Mathieu Reymond, Eugenio Bargiacchi, and Ann Nowé. Pareto conditioned networks. In *Proceedings of the 22nd International Conference on Autonomous Agents and Multiagent Systems (AAMAS)*, 2022.

[Samanta and Chowdhuri, 2021] Akash Samanta and Sumana Chowdhuri. Active cell balancing of lithium-ion battery pack using dual dc-dc converter and auxiliary lead-acid battery. *Journal of Energy Storage*, 33:102109, 2021.

[Schulman *et al.*, 2017] John Schulman, Filip Wolski, Prafulla Dhariwal, Alec Radford, and Oleg Klimov. Proximal policy optimization algorithms. *arXiv preprint arXiv:1707.06347*, 2017.

[Shu *et al.*, 2024] Tianye Shu, Ke Shang, Cheng Gong, Yang Nan, and Hisao Ishibuchi. Learning pareto set for multi-objective continuous robot control. In *Proceedings of the Thirty-Third International Joint Conference on Artificial Intelligence, IJCAI-24*, pages 4920–4928, 2024.

[Turksoy *et al.*, 2020] Arzu Turksoy, Ahmet Teke, and Alkan Alkaya. A comprehensive overview of the dc-dc converter-based battery charge balancing methods in electric vehicles. *Renewable and Sustainable Energy Reviews*, 133:110274, 2020.

[Van Moffaert and Nowé, 2014] Kristof Van Moffaert and Ann Nowé. Multi-objective reinforcement learning using sets of pareto dominating policies. *The Journal of Machine Learning Research*, 15(1):3483–3512, 2014.

[Van Otterlo and Wiering, 2012] Martijn Van Otterlo and Marco Wiering. Reinforcement learning and markov decision processes. In *Reinforcement learning: State-of-the-art*, pages 3–42. Springer, 2012.

[Yang *et al.*, 2019] Runzhe Yang, Xingyuan Sun, and Karthik Narasimhan. A generalized algorithm for multi-objective reinforcement learning and policy adaptation. *Advances in neural information processing systems*, 32, 2019.

[Yang *et al.*, 2022] Feng Yang, Fei Gao, Baochang Liu, and Song Ci. An adaptive control framework for dynamically reconfigurable battery systems based on deep reinforcement learning. *IEEE Transactions on Industrial Electronics*, 69(12):12980–12987, 2022.

[Yang *et al.*, 2023] Yikun Yang, Jiarui He, Chunlin Chen, and Jingwen Wei. Balancing awareness fast charging control for lithium-ion battery pack using deep reinforcement learning. *IEEE Transactions on Industrial Electronics*, 71(4):3718–3727, 2023.

[Zhang *et al.*, 2020] Cong Zhang, Wen Song, Zhiguang Cao, Jie Zhang, Puay Siew Tan, and Xu Chi. Learning to dispatch for job shop scheduling via deep reinforcement learning. *Advances in Neural Information Processing Systems*, 33:1621–1632, 2020.

[Zhang *et al.*, 2022] Yi Zhang, Haihua Zhu, Dunbing Tang, Tong Zhou, and Yong Gui. Dynamic job shop scheduling based on deep reinforcement learning for multi-agent manufacturing systems. *Robotics and Computer-Integrated Manufacturing*, 78:102412, 2022.

[Zhu *et al.*, 2013] Cong Zhu, Xinghu Li, Lingjun Song, and Liming Xiang. Development of a theoretically based thermal model for lithium ion battery pack. *Journal of Power Sources*, 223:155–164, 2013.