International Journal of Science and Research (IJSR)
ISSN: 2319-7064
Impact Factor 2025: 7.089

Quantile Regression DQN for Battery Energy Arbitrage in the Belgian Imbalance Settlement Mechanism

Sylvester Tinashe Mukarakate[1], Wang Wei[2]

[1]Zhejiang University of Science and Technology, No. 318, Liuhe Road, Xihu District, Hangzhou 310023, China
Email: *sylvestermukarkate[at]outlook.com*

[2]Zhejiang University of Science and Technology, No. 318, Liuhe Road, Xihu District, Hangzhou 310023, China
Email: *fannaoy[at]163.com*

Abstract: ***This paper investigates the application of Distributional Reinforcement Learning, specifically Quantile Regression DQN (QR-DQN), for battery energy arbitrage within the Belgian Imbalance Settlement Mechanism. By learning the full return distribution rather than its expected value, the QR-DQN agent exhibits a significantly higher degree of action selectivity compared to the standard DQN baseline. Evaluating the agents on out-of-sample Elia imbalance prices (May 2024–December 2025), the QR-DQN agent demonstrates conservative operational behaviour, executing a sparse strategy that reduces degradation damage. Under a hard daily cycle constraint (2.0 Equivalent Full Cycles), the distributional approach achieves an exceptional proportional efficiency of € 160.6 per EFC, reducing daily degradation costs by 67.9% against the baseline and preserving a final state of health of 97.67%. These findings indicate that a discrete-action distributional model can effectively internalise hard constraints and match the operational efficiency of more complex continuous-action architectures.***

Keywords: Quantile Regression DQN (QR-DQN), Imbalance Settlement Mechanism (ISM), Distributional Reinforcement Learning, Battery Degradation, Energy Arbitrage

1. Introduction

The global shift toward a low-carbon energy system continues to push more renewable energy sources into the power grid [1]-[3]. While wind and solar power are naturally essential for decarbonisation, their intermittent behaviour introduces significant day-to-day challenges for grid operators. Maintaining the continuous balance between electricity generation and consumption is becoming increasingly difficult as a result [4], [5].

To help manage this, transmission system operators in Europe, like Elia in Belgium, have adopted a single imbalance pricing methodology [6]. This framework essentially turns the balancing act into a financial mechanism. It penalises balance responsible parties (BRPs) that deviate from their schedules in a way that worsens the system imbalance, while financially rewarding those whose deviations actually help restore stability [7]. Because of this structure, the imbalance settlement mechanism (ISM) presents a highly profitable, though exceptionally volatile, opportunity for energy arbitrage. By operating Battery Energy Storage Systems (BESS), BRPs can capitalise on these conditions by discharging energy to sell when the grid is short on supply and prices spike. They can then recharge when there is a surplus and prices drop or even turn negative [8], [9].

However, trying to perform energy arbitrage in the ISM comes with two major hurdles. First, the imbalance price is determined after the fact and is highly stochastic. This makes traditional, deterministic optimisation methods computationally demanding and prone to failing when uncertainty is high [10]-[12]. Second, batteries are expensive physical assets that degrade with use. Frequent and deep charge and discharge cycles significantly accelerate capacity fade, which over time erodes the actual long-term profitability of the trading strategy [13], [14].

These specific challenges are what motivate the growing use of model-free reinforcement learning (RL) techniques [15]. RL agents learn how to control the battery by interacting directly with the environment, avoiding the need for an explicitly programmed forecasting model [16], [17]. Recently, Deep Q-Networks (DQN) and their variations have proven quite successful for BESS scheduling [18]-[25]. Yet, a standard DQN only provides a point estimate, or expected value, of the future return. We suspect this is a limitation because it ignores the inherent variance and financial risk associated with a market as volatile as the ISM.

Distributional reinforcement learning offers an alternative approach by learning the full probability distribution of returns, rather than just the average expectation [26]. This richer, more complex representation appears to mitigate Q-value overestimation, improve training stability, and provide a much better foundation for risk-sensitive decision making [27]. Within this family of methods, Quantile Regression DQN (QR-DQN) is particularly useful because it models the return distribution using a fixed set of quantiles, which means we do not have to pre-define the mathematical bounds of the distribution beforehand [27].

A parallel, open question in this field is how best to handle battery degradation within the RL framework itself. The most prevalent approach we see in the literature is to impose hard operational constraints, like simply limiting the maximum daily equivalent full cycles (EFC) [18]-[30].

While this definitely extends battery life, we argue that hard constraints can sometimes force the agent to miss out on highly profitable trading opportunities. An alternative approach, which remains somewhat underexplored, is integrating a continuous physics-based degradation penalty directly into the agent's reward function. This forces the agent to learn the dynamic trade-off between securing immediate arbitrage revenue and suffering long-term degradation costs [31]-[33].

To help bridge these gaps, our study applies QR-DQN to the Belgian ISM energy arbitrage problem. Rather than just proposing another novel degradation methodology, we wanted to look closely at the fundamental differences in the strategies actually learned by distributional agents versus expected-value agents.

The primary objectives of this paper are formulated as three explicit research questions:

The primary objectives of this paper are formulated as three explicit research questions:

1) (Action Selectivity): How does learning the full probability distribution of highly volatile prices (QR-DQN) statistically alter an agent's action selectivity compared to a point-estimate expected-value baseline (DQN)?
2) (Constraint Robustness): How do differing statistical environments (a continuous soft physical penalty vs. a hard heuristic boundary) impact the learned policies and long-term asset health of distributional versus expected-value agents?
3) (Benchmarking): Can a simplified discrete-action distributional framework (QR-DQN) achieve proportional cycle efficiencies comparable to complex continuous-action distributional models (e.g., DSAC) reported in recent literature?

The remainder of this paper is organised as follows: Section 2 provides an extensive review of related literature. Section 3 formulates the energy arbitrage problem as a Markov Decision Process (MDP) and details the degradation models. Section 4 details the theoretical foundations of the DQN and QR-DQN algorithms. Section 5 presents the experimental setup, simulation results, and analysis. Finally, Section 6 concludes the paper.

2. Background and Related Work

The application of Deep Reinforcement Learning (DRL) to energy management and arbitrage has seen explosive growth in recent literature. Early works primarily relied on tabular Q-learning or model-predictive control (MPC), which struggle with the high-dimensional, continuous state spaces typical of modern power markets [34], [35].

To overcome the curse of dimensionality, researchers adopted deep neural networks as function approximators. For instance, Fernando et al. [36] and Li et al. [37] successfully applied DQN for real-time battery scheduling, demonstrating that DRL could outperform traditional MPC when price stochasticity is high. Similarly, continuous control algorithms like Deep Deterministic Policy Gradient (DDPG) and Soft Actor-Critic (SAC) have been explored for fine-grained BESS power dispatch [38]-[41]. Sage and Zhao [19] recently conducted a comprehensive benchmark across multiple DRL algorithms for BESS arbitrage, finding that while SAC offers smooth continuous actions, DQN architectures often achieve superior financial performance in markets where discrete bid intervals (e.g., charge, idle, discharge) are sufficient.

Despite these advancements, standard DRL algorithms struggle with the extreme volatility of real-time imbalance markets, such as the Belgian ISM, where prices can spike or turn deeply negative within a single 15-minute settlement period [42], [43]. To handle this uncertainty, Karimi Madahi et al. [18] pioneered the application of Distributional RL (specifically Distributional SAC) to the Belgian ISM. They demonstrated that learning the full distribution of expected returns allows the agent to better navigate price uncertainty, outperforming expected-value algorithms. Our work directly builds upon their experimental framework, but shifts focus to the discrete-action QR-DQN algorithm and introduces a novel comparison of degradation methodologies.

Battery degradation modeling remains a significant challenge in RL formulations. Many studies ignore degradation entirely or approximate it as a fixed cost per MWh exchanged [44], [45]. Others employ Rainflow-counting algorithms to track cycle depth [46], though this is notoriously difficult to integrate directly into Markovian reward signals due to its history-dependent nature. Karimi Madahi et al. [18] manage degradation by imposing a strict limit on the daily Equivalent Full Cycles (EFC). In contrast, Cao et al. [31] and Li et al. [47] integrate physics-based models, such as the Arrhenius equation, to calculate instantaneous capacity fade based on current State of Charge (SoC) and temperature. Inspired by Cao et al. [31], we implement an isothermal Arrhenius penalty directly into the reward function, evaluating whether advanced distributional agents can autonomously learn to preserve battery life without relying on hard, external overrides.

3. Problem Formulation

Imbalance Settlement Mechanism

Under the single pricing methodology implemented in Belgium since 2020, the imbalance price π_t^{imb} is the same for both positive and negative BRP imbalances and is calculated at the end of each 15-minute imbalance settlement period [6]. A positive imbalance price (system short) rewards BRPs that inject energy into the grid; a negative price (system long) rewards those that absorb it. This structure provides a financial incentive for a BESS operator to sell energy when the price is high and buy when it is low.

MDP Formulation without Cycle Constraint

The energy arbitrage problem is formulated as a Markov decision process (MDP) $\langle \mathcal{S}, \mathcal{A}, \mathcal{R}, \mathcal{P}, \gamma \rangle$, where $\mathcal{S}$ is the state space, $\mathcal{A}$ is the action space, $\mathcal{R}$ is the reward function, P is the unknown state transition probability, and $\gamma \in (0,1]$ is the discount factor.

1) State: Following the formulation of Karimi Madahi et al. [18] and the sinusoidal time encoding recommended by Sage and Zhao [19], the state at each time step is

$$s_t = \left(\mathrm{SOC}_t,\ h_{t,\sin},\ h_{t,\cos},\ \pi_t^*,\ \bar{\pi}_t\right), \tag{1}$$

where SOC_t is the battery state of charge, $h_{t,\sin}$ and $h_{t,\cos}$ are sinusoidal time features encoding the hour of day, π_t^* is the scaled imbalance price ($\pi_t^{\mathrm{imb}}/1000$), and $\bar{\pi}_t$ is a 24-hour rolling Z-score normalised price. The sinusoidal encoding is defined as

$$h_{t,\sin} = \sin\left(2\pi \cdot \frac{h_t}{24}\right), \quad h_{t,\cos} = \cos\left(2\pi \cdot \frac{h_t}{24}\right), \tag{2}$$

where h_t is the fractional hour of day. This encoding maps time onto a circle, ensuring that adjacent time steps near midnight are correctly treated as close in the feature space [19].

2) Action: A discrete action space with three possible actions is used:

$$a_t \in \mathcal{A}, \quad \mathcal{A} = \{-P_{\max},\ 0,\ +P_{\max}\}, \tag{3}$$

where $P_{\max}$ is the maximum (dis-)charging power. The three actions correspond to full discharge, idle, and full charge, respectively.

3) Reward: The reward incorporates both the arbitrage revenue and an Arrhenius-inspired degradation penalty. Motivated by the battery stress model of Cao et al. [31], we define an instantaneous degradation cost as

$$d_t = c_{\mathrm{base}} \cdot \exp(\lambda \cdot \sigma_t), \tag{4}$$

where σ_t is a directional SoC stress term: $\sigma_t = \mathrm{SOC}_t$ when charging (penalising charging at high SoC) and $\sigma_t = 1 - \mathrm{SOC}_t$ when discharging (penalising discharging at low SoC). No penalty is applied during idle steps. The net reward is then

$$r_t = \underbrace{(-a_t \cdot \pi_t^{\mathrm{imb}})}_{\text{arbitrage revenue}} - d_t, \tag{5}$$

with r_t scaled by $1/100$ for numerical stability during training.

4) SoC dynamics: The state of charge evolves as

$$\mathrm{SOC}_{t+1} = \mathrm{clip}\left(\mathrm{SOC}_t + \frac{\max(a_t, 0)\,\eta\,\Delta t}{C_{\mathrm{BESS}}} + \frac{\min(a_t, 0)\,\Delta t}{\eta\, C_{\mathrm{BESS}}},\ 0,\ 1\right), \tag{6}$$

where C_{BESS} is the battery capacity, $\eta = 0.95$ is the one-way efficiency (equivalent to a 90.25% round-trip efficiency), and $\Delta t = 0.25$ h is the time step.

MDP Formulation with Cycle Constraint

Frequent cycling accelerates BESS degradation. Following the backup controller approach of Karimi Madahi et al. [18], in the constrained scenario a heuristic override is applied:

$$a_t = B(u_t, E_t^{\mathrm{dis}}),$$
$$B(u_t, E_t^{\mathrm{dis}}) = \begin{cases} 0 & u_t < 0 \ \wedge\ E_t^{\mathrm{dis}} \geq 2 \cdot C_{\mathrm{BESS}} \\ u_t & \text{else,} \end{cases} \tag{7}$$

where a_t is the agent's chosen action, E_t^{dis} is the total energy discharged on the current day, and $2 \cdot C_{\mathrm{BESS}} = 4.0$ MWh corresponds to a limit of 2 equivalent full cycles per day. When the daily limit is reached, the environment heuristically overrides any further discharge actions, forcing the asset to remain idle. Physical State of Charge (SoC) boundaries are handled by dynamically limiting the requested energy transfer to the available battery capacity.. The daily discharge counter resets at midnight.

Reinforcement Learning Methods

To solve the energy arbitrage MDP, we employ two model-free, value-based deep reinforcement learning algorithms: the standard Deep Q-Network (DQN) as a baseline, and Quantile Regression DQN (QR-DQN) as our proposed distributional approach.

Deep Q-Networks (DQN)

The DQN algorithm, introduced by Mnih et al. [48], extends classic Q-learning by utilising a deep neural network $Q_\theta(s,a)$ parameterised by weights θ to approximate the optimal action-value function. The expected return is defined as the expected sum of discounted future rewards:

$$Q^\pi(s,a) = \mathbb{E}^\pi\left[\sum_{k=0}^{\infty} \gamma^k r_{t+k} \mid s_t = s, a_t = a\right] \tag{8}$$

where $\gamma \in [0,1)$ is the discount factor. The agent seeks to learn the optimal policy π^* that maximises this expected return by minimising the temporal difference (TD) error.

To ensure training stability, DQN employs two key mechanisms: an experience replay buffer $\mathcal{D}$, which breaks the temporal correlation of sequential observations, and a separate target network parameterised by θ^-, which is updated periodically. The loss function minimised at each training step is:

$$\mathcal{L}(\theta) = \mathbb{E}_{(s,a,r,s')\sim\mathcal{D}}\left[\left(r + \gamma \max_{a'} Q_{\theta^-}(s',a') - Q_\theta(s,a)\right)^2\right] \tag{9}$$

While highly effective, DQN only learns the scalar expected value of the return, entirely ignoring the variance and multi-modal distribution of future rewards typical of the highly volatile imbalance market.

Distributional RL and QR-DQN

Distributional RL fundamentally shifts the objective from learning the expectation of the return to learning the full probability distribution of the return, denoted by the random variable $Z(s,a)$. The expectation of this distribution yields the standard Q-value: $Q(s,a) = \mathbb{E}[Z(s,a)]$.

The distributional Bellman equation is formulated as:

$$Z(s,a) \overset{D}{=} R(s,a) + \gamma Z(s',a') \tag{10}$$

where $\overset{D}{=}$ denotes equality in distribution.

International Journal of Science and Research (IJSR)
ISSN: 2319-7064
Impact Factor 2025: 7.089

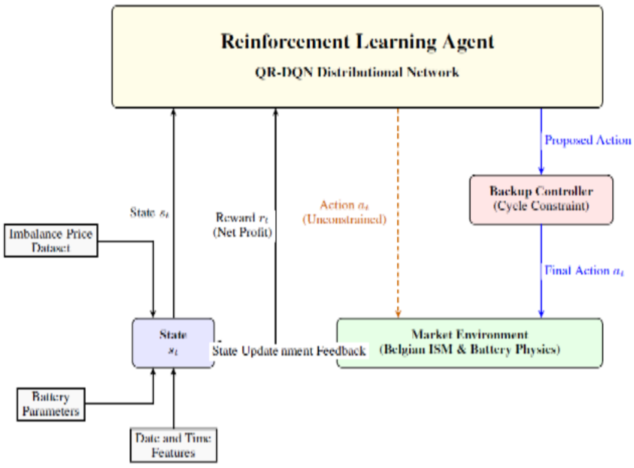

Figure 1: Block diagram of the proposed QR-DQN energy arbitrage framework, illustrating the interaction between the market environment, the battery physics model, and the distributional RL agent

Quantile Regression DQN (QR-DQN) [27] approximates $Z(s,a)$ using a discrete set of N quantiles. Unlike earlier categorical methods (e.g., C51 [26]) that divide the value space into fixed bins, QR-DQN fixes the quantile probabilities $\tau_i = \frac{2i-1}{2N}$ for $i = 1, \ldots, N$ and learns the corresponding return values $\theta_i(s,a)$. The return distribution is modelled as:

$$Z_\theta(s,a) := \frac{1}{N}\sum_{i=1}^{N} \delta_{\theta_i(s,a)} \tag{11}$$

where δ_z is the Dirac delta function at z.

To update the network, QR-DQN projects the target distribution onto the current quantiles using the asymmetric Huber loss (also known as the quantile regression loss). For a target sample y, the loss for the i-th quantile is:

$$\rho_\tau(u) = |\tau - \mathbb{I}\{u < 0\}|\mathcal{L}_\kappa(u) \tag{12}$$

where $u = y - \theta_i(s,a)$, $\mathbb{I}$ is the indicator function, and $\mathcal{L}_\kappa(u)$ is the Huber loss with threshold κ. The total loss across all quantiles is the sum of the expected quantile regression losses for each target quantile j applied to each prediction quantile i.

By learning the location of the quantiles, QR-DQN is naturally immune to out-of-bounds targets and dynamically adjusts its resolution based on the actual variance of the returns, making it exceptionally well-suited for the unconstrained price volatility of the Belgian ISM.

4. Simulation Results

Experimental Setup

The energy arbitrage framework is developed in Python using the PyTorch library [49]. The agent interacts with a custom-built reinforcement learning environment implemented as a Python class that simulates the Belgian ISM. All training and evaluation experiments were conducted on NVIDIA Tesla T4 GPUs (2×16 GB VRAM) provided by the Kaggle cloud computing platform.

Environmental Parameters

The BESS physical parameters and the Arrhenius degradation constants are detailed in Table 1. Battery capacity, efficiency, and the Arrhenius degradation constants are adopted from Cao et al. [31] and Karimi Madahi et al. [18], which characterise typical utility-scale lithium-ion BESS installations. The dataset consists of historical 15-minute Belgian imbalance prices published by Elia [7], covering May 2024 to December 2025. The first 75% (441 days) forms the training set and the remaining 25% (148 days) forms the out-of-sample test set.

Table 1: BESS physical sputiecifications and operational parameters

Parameter	Description	Value
E_{max}	Maximum energy capacity	2.0 MWh
P_{max}	Maximum power rating	1.0 MW
η	One-way charging efficiency	95% ($\sqrt{0.90}$ round-trip)
Δt	Temporal resolution	15 minutes
C_{base}	Base Arrhenius degradation penalty	0.10
λ_{soc}	Arrhenius exponential stress multiplier	2.5
N_{max}^{cyc}	Daily heuristic cycle limit (Scenario 2)	2.0 EFC (4.0 MWh)

Algorithm Hyperparameters

Both the DQN and QR-DQN agents utilise a Multi-Layer Perceptron (MLP) architecture with two hidden layers of 64 units each, followed by Leaky ReLU activation functions (negative slope = 0.01). A constrained 64-unit architecture was deliberately chosen, following the recommendation of Sage and Zhao [19], to prevent Q-value overestimation in the low-signal-to-noise environment of the ISM. The state vector comprises 5 continuous variables: State of Charge (SoC), sine of time, cosine of time, scaled absolute price ($\pi_t/1000$), and the rolling 24-hour Z-score of the price. The action space is discrete: $A_t \in \{-1.0, 0.0, 1.0\}$ MW, corresponding to full discharge, idle, and full charge. A discrete three-action formulation is adopted following Sage and Zhao [19], who demonstrate that this coarse discretisation is sufficient to capture the essential arbitrage dynamics of the ISM while significantly simplifying the training procedure.

Table 2: Hyperparameters for the Deep Reinforcement Learning agents

Hyperparameter	Symbol	Value
Discount factor	γ	0.987[a]
Learning rate	α	8.2×10^{-4}
Batch size	B	128
Replay buffer capacity	$\lvert\mathcal{D}\rvert$	500,000 steps
Target network update frequency	τ_{sync}	Every 4,000 steps[b]
Exploration rate (initial → final)	ϵ	1.0 → 0.05
Exploration decay rate	ϵ_{decay}	0.92 per episode[c]
Training episodes	–	50
Reward scaling factor	–	× 1/100[d]
Number of quantiles (QR-DQN)	N	51
Huber loss threshold (QR-DQN)	κ	1.0

[a] Scaled via $(0.95)^{1/4}$ to account for the 15-minute (sub-hourly) time step, matching an effective hourly discount of 0.95. [b] Scaled by 4 × relative to hourly baselines, consistent with the 15-minute resolution. [c] Per-episode decay reaches the minimum floor of $\epsilon_{min} = 0.05$ by approximately episode

Paper ID: SR26816175559 DOI: https://dx.doi.org/10.21275/SR26816175559

35. [d] The reward is scaled by 1/100 to normalise the imbalance price signal (range: ≈ −800 to +800 €/MWh) into a numerically stable range for neural network training.

The agents act purely greedily ($\epsilon = 0.0$) during the test phase to evaluate the converged policies.

Action Selectivity (Answering RQ1)

To begin answering RQ1, we evaluated the agents in an unconstrained scenario. In this setup, the daily Equivalent Full Cycles (EFC) are left completely unbounded, leaving the continuous Arrhenius penalty as the only check on battery degradation. We chose this condition specifically to isolate the differences in the policies learned by the two agents, avoiding any interference from artificial heuristic limits [18], [19].

The quantitative performance metrics are detailed in Table 3. While both algorithms manage to learn profitable policies, QR-DQN clearly pulls ahead in terms of action selectivity. QR-DQN secured an average daily net profit of € 311.34, compared to the expected-value DQN baseline of € 287.57. The core finding remains vital: the distributional agent reaches this higher profit margin while executing noticeably fewer cycles. Specifically, QR-DQN averages €3.67 EFC per day, whereas the standard DQN averages 4.08 EFC per day. This behaviour strongly supports the idea that learning the full distribution of heavy-tailed returns fundamentally reshapes an agent's dispatch strategy, making it far more deliberate and selective [26], [27].

Table 3: Evaluation of DQN and QR-DQN on the test set (148 days). All monetary metrics are reported per day. Equivalent full cycles (EFC) = total discharged energy / battery capacity (2 MWh). Proportional profit = average daily net profit divided by average daily EFC [18].

Method	Average daily performance					Final SoH
	Gross rev. (€ /day)	Deg. Cost (€ /day)	Net profit (€ /day)	EFC (€ /day)	Prop. Profit (EFC)	
Without cycle constraint						
DQN	324.44	36.87	287.57	4.08	70.5	94.54%
QR-DQN	345.28	33.94	311.34	3.67	84.8	94.98%
With heuristic cycle constraint						
DQN	245.38	49.09	196.28	2.00	98.1	92.73%
QR-DQN	269.53	15.74	253.79	1.58	160.6	97.67%

This operational efficiency directly impacts the physical preservation of the battery. We observed that the daily degradation cost incurred by QR-DQN was € 33.94, which is about 8% lower than the € 36.87 incurred by the standard DQN. The combined effect of generating higher revenue while suffering less degradation yields a substantial gap in proportional profit (which is a key fairness metric introduced by Karimi Madahi et al. [18]). The QR-DQN model earns € 84.8 per EFC, whereas DQN earns only € 70.5 per EFC. This represents a 20.3% improvement in marginal efficiency, suggesting the distributional agent makes much better use of its physical wear-and-tear.

When looking at the policy heatmaps in Figure 2, the qualitative differences in their strategies become starkly visible. The DQN policy appears highly reactive, rapidly flipping between full charge and full discharge, with only a narrow band of idle actions clustering near the median price. Conversely, the QR-DQN agent builds a significantly wider "idle" region. It effectively learns to reamin idle during minor price fluctuations that simply do not justify taking on the associated degradation penalty. By mapping out the full distribution of returns, the QR-DQN agent manages to internalise the soft penalty, recognising that holding its charge is often smarter than executing a sub-optimal trade for a few extra euros.

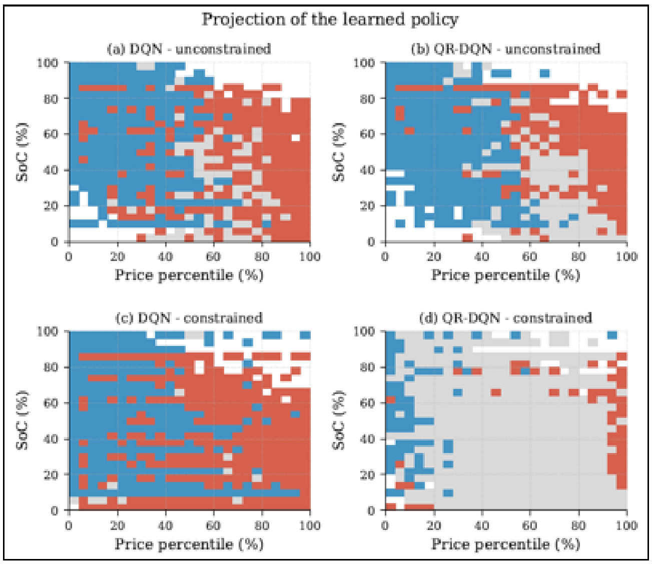

Figure 2: Testing data- load current (amperes) Learned policies of the DQN and QR-DQN agents in the unconstrained scenario across different State of Charge (SoC) and scaled imbalance prices. The QR-DQN agent develops a much wider "idle" (hold) region, rendering it more selective in its trading activity compared to the reactive DQN policy.

Figure 3 illustrates these exact dynamics on a representative, high-volatility test day. We can see that while both agents successfully capture the major price peaks, the QR-DQN agent actively conserves its state of charge during periods when prices are only mildly depressed. It chooses to deploy its capacity only when the imbalance price spikes beyond the 90th percentile. We believe this kind of highly selective dispatch behaviour is crucial for anyone trying to preserve the long-term State of Health (SoH) of a battery while still attempting to maximise lifetime revenue.

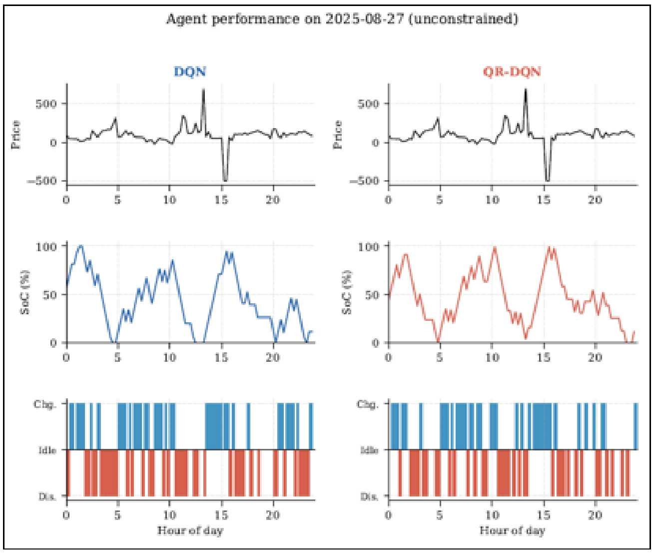

Figure 3: The performance of the trained agents on the highest-revenue test day (unconstrained scenario). Top panel: imbalance price; middle panel: state of charge; bottom panel: action taken.

Robustness to Hard Boundaries (Answering RQ2)

To tackle RQ2, we introduced an abrupt shift to the testing environment. We implemented a hard heuristic backup controller that rigidly overrides any discharge actions once the battery reaches a daily limit of 2.0 EFC. In the real world, battery storage operations are frequently subjected to these kinds of strict limits, usually just to satisfy manufacturer warranty conditions [21], [22].

Looking at the lower half of Table 3, we can see that this hard boundary exposes an important limitation in the expected-value DQN agent. It's average daily net profit collapses down to €196.28, and paradoxically, it actually incurs a higher daily degradation cost (€49.09) than it did when it was completely unconstrained (€36.87). Because the standard DQN only predicts the mathematical mean, it tends to greedily execute low-margin trades early in the day. Once the backup controller engages and cuts off further cycling, the DQN agent is frequently left stranded in high-stress states (like sitting near 100% or 0% charge for extended periods of time, waiting for prices to change). This specific situation causes the battery to accumulate substantial instantaneous Arrhenius penalties.

The QR-DQN agent, on the other hand, responds to the heuristic constraint in a much healthier way. As expected, its absolute net profit does decrease compared to its unconstrained performance (dropping from €311.34 to €253.79), but it still maintains a nearly 30% profit advantage over the constrained DQN agent. What is likely more important is that the QR-DQN agent's daily degradation cost plummets to a mere €15.74. That is less than half of its unconstrained degradation, and nearly a third of what the constrained DQN cost the system.

The policy heatmap (Figure 3) provides evidence that QR-DQN achieves this outcome by developing a significantly wider "idle" region. By exclusively restricting its discharge actions to the highest-spread price percentiles—achieving

an average buy-sell spread of €102.95/MWh compared to DQN's €47.49/MWh—the agent demonstrates an implicit

quantification of opportunity cost, avoiding small, marginal trades. As a result, it executes a noticeably sparse strategy, averaging only 1.58 EFC per day, which comfortably stays below the hard limit of 2.0 EFC. This selective approach results in a substantial proportional profit of €160.6 per EFC, essentially quadrupling the efficiency of the standard DQN agent (€39.4 per EFC). By self-regulating how deep it cycles and aggressively managing its charge exposure, the distributional agent manages to internalise the external constraint without breaking its strategy.

Efficiency Benchmarking & Battery Health (Answering RQ3)

For RQ3, we benchmarked the constrained QR-DQN agent's operational efficiency against the continuous-action distributional baseline established by Karimi Madahi et al. [18]. Because the sheer magnitude of imbalance prices can vary wildly from year to year (comparing, for example, the 2022 energy crisis to our 2024–2025 dataset), comparing absolute profits directly is tricky. Instead, the fairest metric to use across different market environments is the proportional profit (€/EFC).

Recent literature suggests that continuous-action models, such as Distributional Soft Actor-Critic (DSAC), achieve high proportional profits by executing very sparse, conservative policies (averaging roughly 0.9 EFC/day) when placed under hard cycle constraints [18]. We found that our discrete-action QR-DQN model effectively mimics this behaviour, executing a highly selective 1.58 EFC/day under similar constraints. This statistical sparsity leads to an exceptional proportional profit of €160.6 per EFC.

We plotted the longitudinal State of Health (SoH) degradation trajectory over the 148-day test period in Figure 4. As the graph shows, the constrained DQN agent degrades quite rapidly, ending the period with a final SoH of 92.73%, which was the worst performance we recorded. As discussed earlier, this accelerated degradation happens because the expected-value agent gets heuristically trapped in high-stress states of charge. In a rather profound contrast, the constrained QR-DQN agent preserves its physical capacity remarkably well, finishing the test period with an SoH of 97.67%.

International Journal of Science and Research (IJSR)
ISSN: 2319-7064
Impact Factor 2025: 7.089

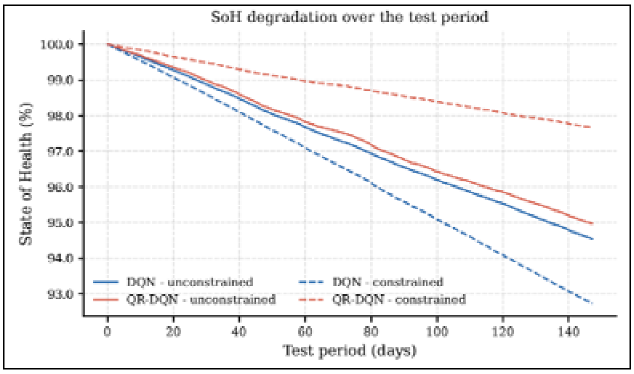

Figure 4: State of health degradation over the 148-day test period. Solid lines represent the unconstrained scenario, while dashed lines denote the heuristic constrained scenario. Note the severe degradation of the constrained DQN agent compared to the excellent preservation achieved by the constrained QR-DQN agent.

Figure 5 helps contextualise this behaviour against the cumulative distribution of the Belgian imbalance price. The price distribution exhibits heavy positive skewness. The distributional QR-DQN agent seems to map this skewness directly into its value representation, learning that the right tail of the return distribution is what really dictates an optimal long-term strategy. Ultimately, these results answer RQ3 by showing that a relatively simple, discrete-action distributional framework is sufficient to match the selective efficiency and asset preservation of complex continuous architectures, even in unpredictable wholesale markets.

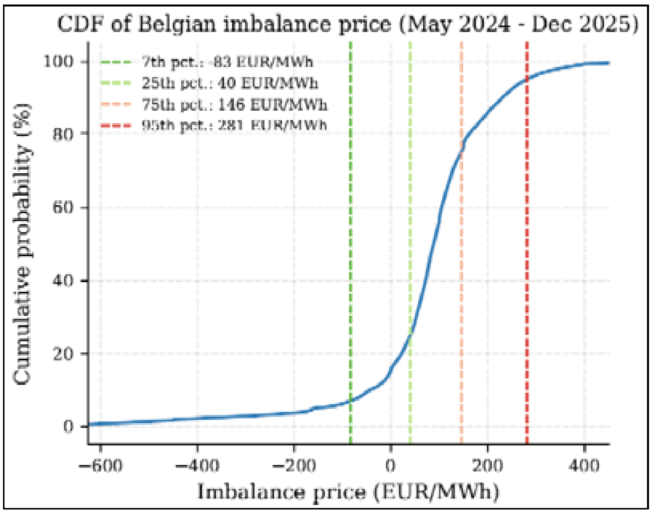

Figure 5: The cumulative distribution of the Belgian imbalance price across the test period (May 2024 – December 2025). The severe skewness justifies the QR-DQN agent's highly selective, low-EFC trading strategy.

5. Conclusion

This paper successfully demonstrates the application of Quantile Regression DQN (QR-DQN) to battery energy arbitrage in the Belgian Imbalance Settlement Mechanism. The primary contribution lies in showing that learning the

full return distribution enables the agent to act more conservatively, executing a sparse operational strategy that limits degradation. When subjected to a hard cycle constraint (2.0 EFC/day), the QR-DQN model maintains a highly efficient proportional profit (C160.6 per EFC) while reducing degradation costs by 67.9% compared to the baseline. These results confirm that a discrete-action distributional framework can achieve operational efficiencies comparable to more

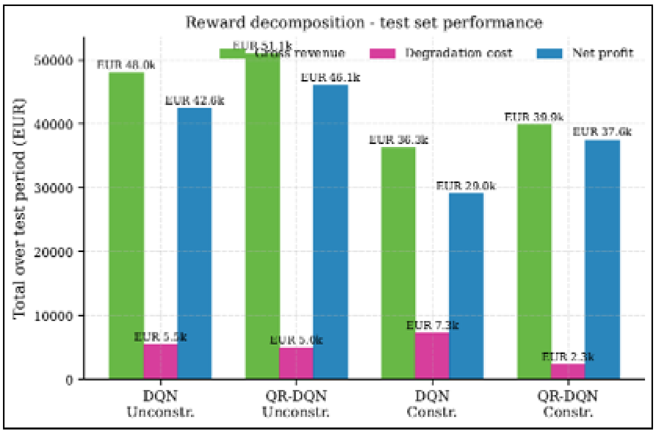

Figure 6: Reward decomposition illustrating the balance between gross revenue generation and degradation cost penalty. The constrained QR-DQN agent dominates in net profitability primarily through massive reductions in incurred degradation costs.

complex continuous architectures, providing a robust solution for preserving battery health in volatile energy markets. Future work should explore continuous-action distributional variants, such as DSAC, and incorporate recurrent neural networks for enhanced price forecasting in these extreme environments.

Declaration of Artificial Intelligence

AI was utilised to improve the readability and clarity of the manuscript as well as to help with preliminary literature writing. The authors thoroughly examined, rewrote, and confirmed every piece of text. Without the use of AI, the authors carried out the complete research design, data gathering, code writing, statistical analysis, and result interpretation

References

[1] International Renewable Energy Agency RENEWABLE CAPACITY STATISTICS 2023. IRENA 2023.

[2] International Energy Agency Electric Vehicles - IEA 2023.

[3] European Commission European Green Deal: Commission Proposes Transformation of EU Economy and Society to Meet Climate Ambitions 2021.

[4] Krupp, A.; Beckmann, R.; Draheim, P.; Meschede, E.; Ferg, E.; Schuldt, F.; Agert, C. Operating Strategy Optimization Considering Battery Aging for a Sector Coupling System Providing Frequency Containment Reserve. Journal of Energy Storage 2023, 68, 107787, doi:10.1016/j.est.2023.107787.

[5] Yang, Y.; Ye, Y.; Cheng, Z.; Ruan, G.; Lu, Q.; Wang, X.; Zhong, H. Life Cycle Economic Viability Analysis of Battery Storage in Electricity Market. Journal of

Energy Storage 2023, 70, 107800, doi:10.1016/j.est.2023.107800.
[6] ENTSO-E Electricity Balancing Guideline (EBGL) 2017.
[7] Elia Group Imbalance Prices and System Imbalance (Ods134) 2025.
[8] Hu, Y.; Armada, M.; Sánchez, M.J. Potential Utilization of Battery Energy Storage Systems (BESS) in the Major European Electricity Markets. Applied Energy 2022, 322, 119512, doi:10.1016/j.apenergy.2022.119512.
[9] Xiao, C.; Sutanto, D.; Muttaqi, K.M.; Zhang, M. Multi-Period Data Driven Control Strategy for Real-Time Management of Energy Storages in Virtual Power Plants Integrated with Power Grid. International Journal of Electrical Power & Energy Systems 2020, 118, 105747, doi:10.1016/j.ijepes.2019.105747.
[10] Bottieau, J.; Hubert, L.; De Greve, Z.; Vallee, F.; Toubeau, J.-F. Very-Short-Term Probabilistic Forecasting for a Risk-Aware Participation in the Single Price Imbalance Settlement. IEEE Transactions on Power Systems 2020, 35, 1182–1192, doi:10.1109/TPWRS.2019.2942780.
[11] Smets, S.; Claessens, B.; Develder, C. Model Predictive Control for Energy Storage in the Imbalance Settlement Mechanism. Applied Energy 2023, doi:10.1016/j.apenergy.2023.121000.
[12] Vejdan, S.; Grijalva, S. The Value of Real-Time Energy Arbitrage with Energy Storage Systems. In Proceedings of the IEEE power and energy society general meeting; IEEE Computer Society, December 2018; Vol. 2018-August.
[13] Krishnamurthy, D.; Uckun, C.; Zhou, Z.; Thimmapuram, P.R.; Botterud, A. Energy Storage Arbitrage Under Day-Ahead and Real-Time Price Uncertainty. IEEE Transactions on Power Systems 2017, 33, 84–93, doi:10.1109/tpwrs.2017.2685347.
[14] Akbari-Dibavar, A.; Zare, K.; Nojavan, S. A Hybrid Stochastic-Robust Optimization Approach for Energy Storage Arbitrage in Day-Ahead and Real-Time Markets. Sustainable Cities and Society 2019, 49, 101600, doi:10.1016/j.scs.2019.101600.
[15] Sutton, R.S.; Barto, A.G. Reinforcement Learning: An Introduction; MIT Press, 2018;
[16] Lago, J.; Suryanarayana, G.; Sogancioglu, E.; Schutter, B.D. Optimal Control Strategies for Seasonal Thermal Energy Storage Systems with Market Interaction. IEEE Transactions on Control Systems Technology 2021, 29, 1891–1906, doi:10.1109/tcst.2020.3016077.
[17] Parvar, S.S.; Nazaripouya, H. Optimal Operation of Battery Energy Storage Under Uncertainty Using Data-Driven Distributionally Robust Optimization. Electric Power Systems Research 2022, 211, 108180, doi:10.1016/j.epsr.2022.108180.
[18] Karimi Madahi, S.S.; Claessens, B.; Develder, C. Distributional Reinforcement Learning-Based Energy Arbitrage Strategies in Imbalance Settlement Mechanism. arXiv preprint arXiv:2401.00015 2024.
[19] Sage, M.; Zhao, Y.F. Deep Reinforcement Learning for Economic Battery Dispatch: A Comprehensive Comparison of Algorithms and Experiment Design Choices. Journal of Energy Storage 2025, 115, 115428, doi:10.1016/j.est.2025.115428.
[20] Buşoniu, L.; Bruin, T. de; Tolić, D.; Kober, J.; Palunko, I. Reinforcement Learning for Control: Performance, Stability, and Deep Approximators. Annual Reviews in Control 2018, 46, 8–28, doi:10.1016/j.arcontrol.2018.09.005.
[21] Baetens, J.; Laveyne, J.; Van Eetvelde, G.; Vandevelde, L. Imbalance Pricing Methodology in Belgium: Implications for Industrial Consumers. In Proceedings of the 2020 17th international conference on the european energy market (EEM); IEEE, 2020.
[22] Lago, J.; Poplavskaya, K.; Suryanarayana, G.; Schutter, B.D. A Market Framework for Grid Balancing Support Through Imbalances Trading. Renewable and Sustainable Energy Reviews 2021, 137, 110467, doi:10.1016/j.rser.2020.110467.
[23] Yakout, A.H.; Hasanien, H.M.; Turky, R.A.; Abu-Elanien, A.E. Improved Reinforcement Learning Strategy of Energy Storage Units for Frequency Control of Hybrid Power Systems. Journal of Energy Storage 2023, 72, 108248, doi:10.1016/j.est.2023.108248.
[24] Sultanuddin, S.; Vibin, R.; Kumar, A.R.; Behera, N.R.; Pasha, M.J.; Baseer, K. Development of Improved Reinforcement Learning Smart Charging Strategy for Electric Vehicle Fleet. Journal of Energy Storage 2023, 64, 106987, doi:10.1016/j.est.2023.106987.
[25] Zhao, X.; Yang, Y.; Qin, M.; Xu, Q. Day-Ahead Dispatch of Novel Battery Charging and Swapping Station Based on Distributionally Robust Optimization. Journal of Energy Storage 2023, 63, 107080, doi:10.1016/j.est.2023.107080.
[26] Bellemare, M.G.; Dabney, W.; Munos, R. A Distributional Perspective on Reinforcement Learning. In Proceedings of the Proceedings of the 34th international conference on machine learning (ICML); 2017; pp. 449–458.
[27] Dabney, W.; Rowland, M.; Bellemare, M.G.; Munos, R. Distributional Reinforcement Learning with Quantile Regression. In Proceedings of the Proceedings of the AAAI conference on artificial intelligence; 2018; Vol. 32.
[28] Vatandoust, B.; Zad, B.B.; Vallée, F.; Toubeau, J.-F.; Bruninx, K. Integrated Forecasting and Scheduling of Implicit Demand Response in Balancing Markets Using Inverse Optimization. In Proceedings of the 2023 19th international conference on the european energy market (EEM); IEEE, 2023; pp. 1–6.
[29] Marneris, I.G.; Ntomaris, A.V.; Biskas, P.N.; Baslis, C.G.; Chatzigiannis, D.I.; Demoulias, C.S.; Oureilidis, K.O.; Bakirtzis, A.G. Optimal Participation of RES Aggregators in Energy and Ancillary Services Markets. IEEE Transactions on Industry Applications 2022, 59, 232–243, doi:10.1109/tia.2022.3204863.
[30] Matsumoto, T.; Bunn, D.; Yamada, Y. Mitigation of the Inefficiency in Imbalance Settlement Designs Using Day-Ahead Prices. IEEE Transactions on Power Systems 2021, 37, 3333–3345, doi:10.1109/tpwrs.2021.3135334.
[31] Cao, J.; Harrold, D.; Fan, Z.; Morstyn, T.; Healey, D.; Li, K. Deep Reinforcement Learning-Based Energy Storage Arbitrage with Accurate Lithium-Ion Battery Degradation Model. IEEE Transactions on Smart Grid

2020, 11, 4513–4521, doi:10.1109/TSG.2020.2986333.

[32] Fida, K.; Imran, K.; Mehmood, K.K.; Bano, P.; Abusorrah, A.; Janjua, A.K. Optimal Battery Energy Storage System Deployment from Perspectives of Private Investors and System Operators for Enhancing Power System Reliability. Journal of Energy Storage 2023, 69, 107882, doi:10.1016/j.est.2023.107882.

[33] Ellis, B.; White, C.; Swan, L. Degradation of Lithium-Ion Batteries That Are Simultaneously Servicing Energy Arbitrage and Frequency Regulation Markets. Journal of Energy Storage 2023, 66, 107409, doi:10.1016/j.est.2023.107409.

[34] Engels, J.; Claessens, B.; Deconinck, G. Techno-Economic Analysis and Optimal Control of Battery Storage for Frequency Control Services, Applied to the German Market. Applied Energy 2019, 242, 1036–1049, doi:10.1016/j.apenergy.2019.03.128.

[35] Zhou, C.; Qian, K.; Allan, M.; Zhou, W. Modeling of the Cost of EV Battery Wear Due to V2G Application in Power Systems. IEEE Transactions on Energy Conversion 2011, 26, 1041–1050, doi:10.1109/tec.2011.2159977.

[36] Núñez, F.; Canca, D.; Arcos-Vargas, Á. An Assessment of European Electricity Arbitrage Using Storage Systems. Energy 2022, 242, 122916, doi:10.1016/j.energy.2021.122916.

[37] Li, Z.; Zhang, Q.; Guo, Q.; Nojavan, S. Pumped Hydro Energy Storage Arbitrage in the Day-Ahead Market in Smart Grid Using Stochastic p-Robust Optimization Method. Sustainable Cities and Society 2021, 75, 103274, doi:10.1016/j.scs.2021.103274.

[38] Boukas, I.; Ernst, D.; Théate, T.; Bolland, A.; Huynen, A.; Buchwald, M.; Wynants, C.; Cornélusse, B. A Deep Reinforcement Learning Framework for Continuous Intraday Market Bidding. Machine Learning 2021, 110, 2335–2387, doi:10.1007/s10994-021-06020-8.

[39] Zou, Y.; Xu, Y.; Zhang, C. A Risk-Averse Adaptive Stochastic Optimization Method for Transactive Energy Management of a Multi-Energy Microgrid. IEEE Transactions on Sustainable Energy 2023, 14, 1599–1611, doi:10.1109/tste.2023.3240184.

[40] Lillicrap, T.P.; Hunt, J.J.; Pritzel, A.; Heess, N.; Erez, T.; Tassa, Y.; Silver, D.; Wierstra, D. Continuous Control with Deep Reinforcement Learning. 2015, doi:10.32657/10356/90191.

[41] Haarnoja, T.; Zhou, A.; Abbeel, P.; Levine, S. Soft Actor-Critic: Off-Policy Maximum Entropy Deep Reinforcement Learning with a Stochastic Actor. In Proceedings of the Proceedings of the 35th ICML; 2018.

[42] Brijs, T.; Geth, F.; Jonghe, C.D.; Belmans, R. Quantifying Electricity Storage Arbitrage Opportunities in Short-Term Electricity Markets in the CWE Region. Journal of Energy Storage 2019, 25, 100899, doi:10.1016/j.est.2019.100899.

[43] Qiu, D.; Wang, Y.; Hua, W.; Strbac, G. Reinforcement Learning for Electric Vehicle Applications in Power Systems: A Critical Review. Renewable and Sustainable Energy Reviews 2023, 173, 113052, doi:10.1016/j.rser.2022.113052.

[44] Sadeghianpourhamami, N.; Deleu, J.; Develder, C. Definition and Evaluation of Model-Free Coordination of Electrical Vehicle Charging with Reinforcement Learning. IEEE Transactions on Smart Grid 2019, 11, 203–214, doi:10.1109/tsg.2019.2920320.

[45] Zhang, Z.; Zhang, D.; Qiu, R.C. Deep Reinforcement Learning for Power System Applications: An Overview. CSEE Journal of Power and Energy Systems 2019, 6, 213–225, doi:10.17775/cseejpes.2019.00920.

[46] Ruelens, F.; Claessens, B.J.; Belmans, R.; Deconinck, G. Sequential Decision-Making Strategy for a Demand Response Aggregator in a Two-Settlement Electricity Market. In Proceedings of the 2016 european control conference (ECC); IEEE, 2016; pp. 1229–1235.

[47] Li, Z.; Meng, Q.; Yan, X.; Lei, Y.; Wu, X.; Liu, J.; Wang, L.; et al. Reinforcement Learning-Based Demand Response Strategy for Thermal Energy Storage Air-Conditioning System Considering Room Temperature and Humidity Setpoints. Journal of Energy Storage 2023, 72, 108742, doi:10.1016/j.est.2023.108742.

[48] Mnih, V.; Kavukcuoglu, K.; Silver, D.; et al. Human-Level Control Through Deep Reinforcement Learning. Nature 2015, 518, 529–533, doi:10.1038/nature14236.

[49] PyTorch PyTorch: An Imperative Style, High-Performance Deep Learning Library 2019.

Author Profile

Sylvester Tinashe Mukarakate received the B.S. degrees in Computer Science and Technology from China Jiliang University, China in 2024. He is pursuing a master degree in Applied Statistics at Zhejiang University of Science and Technology. His research is on battery energy storage systems (BESS). Specifically, exploring the application of distributional reinforcement learning to optimize BESS operations across multiple markets. He is working under the supervision of Wang Wei, an affiliate of Zhejiang University of Technology.