Battery Energy Storage Control Using Reinforcement Learning

Elliott Basso
College of Science and Engineering
James Cook University
Townsville, Australia
elliott.basso@my.jcu.edu.au

Yang Du
College of Science and Engineering
James Cook University
Cairns, Australia
yang.du@jcu.edu.au

***Abstract*— With the increasing adoption of solar PV installations in Australian households, the availability of cheap renewable power during the day has surged. However, the challenge lies in rising electricity prices during morning and evening peak-consumption times. This project assessed the feasibility and profitability of using a Reinforcement Learning (RL) controller in a Battery Energy Storage System (BESS) to make cost-effective decisions by purchasing power when it's inexpensive and selling when it's costly. MATLAB/Simulink is used to create a BESS simulation model integrated into the electricity market, with the RL agent trained using normalized observation data and a reward function. Benchmarking demonstrated the RL controller's consistent outperformance of the timer-based controller in various market scenarios, emphasizing its adaptability and profitability advantages, particularly in volatile markets.**

Keywords*— *Battery Energy Storage System, Reinforcement Learning, Matlab/Simulink

I. INTRODUCTION

In recent years, with the rapidly growing awareness and adoption of solar PV installations on household roofs across Australia, it has led to having an increase of renewable power cheaply available throughout the daytime [1]. However, the major challenge that comes with the great adoption of solar generation is the surging electricity prices during the evening peak-consumption times. This is caused by the absence of solar generation in the evenings when people are typically coming home and turning on air-conditioning and cooking appliances, creating a 24 hour load graph phenomenon known as a "duck curve".

This daily supply and demand challenge created by the non-dispatchable renewable sources of energy is only going to become more significant as the number of new solar installations have continued trending slightly upwards since 2015, with the Australian total solar PV capacity effectively doubling between 2020 and 2023.

An apparent strategy to address this challenge would be to store the non-dispatchable energy generated during the cheaper off-peak times and dispatch it during higher-priced peak demand times. Thanks to advancements in battery storage technology, it has become significantly more affordable for homes and businesses to install Battery Energy Storage Systems (BESSs) with industry forecasts predicting it to become at least 20% cheaper within the decade [2]. The rapid adoption of dispatchable BESSs by consumers and industry, as well as the increasing government investment in large-scale batteries across the country [3] has highlighted the need for effective and efficient Energy Management System (EMS) to control these dispatchable energy resources (DERs).

The means of integrating renewable energy sources (RESs) and DER methods into the power grid at both the local-scale and grid-scale levels gave birth to the Micro-Grid (MG) concept. [4] The MG has many hierarchical layers and methods of energy management and control, but the most economically focused method is the tertiary control level. Tertiary control is considered to be part of the host grid and not the MG itself [4], it optimises the overall grid operation based on merits of interest, mostly efficiency and economics [5], with one of the main controllable options being DERs. Thus, with efficiency and economic interests being the focus, the tertiary controller needs an EMS to determine the best time to charge and discharge the DERs, a process known as optimisation. [6]

It has been often stated that due to the intermittency and volatility of RESs, their rising use, and integration into the power grid: traditional mathematical means of forecasting the supply, demand, and electricity pricing has become increasingly difficult and often error prone [6, 7] making it tedious for EMSs to determine the optimum times to charge and discharge dispatchable BESSs. With more RESs coming online, the price fluctuation challenge is only going to become more substantial. For EMSs to remain as efficient and economically competitive as possible, a better method of price forecasting and dispatch control needs to be determined.

In the past few years, numerous papers have appeared investigating the applications for Reinforcement Learning (RL) in the optimisation of battery storage [8, 9]. RL is a goal-oriented machine learning technique, which makes it perfectly suited for economic-based control problems [10]. Recent studies have applied RL agents to EMSs as proof-of-concept showing mixed levels of success. Some studies show that it can make economically beneficial decisions, but doesn't outperform traditional controllers, to other studies showing RL agents outperforming traditional controllers [10-12]. Studies which show RL underperforming typically conclude with the researchers believing that poor shaping of the reward criteria for the RL model is the cause for economically underperforming EMSs.

The objective of this paper is to develop an EMS via the utilization of a RL model to effectively control a battery energy storage system in the most economical way possible. A variety of reward criteria will be trialled on different RL agents in a MATLAB Simulink environment of a small scale microgrid, then benchmarked to determine the most effective configuration which will then be compared to a traditional timer-based heuristic EMS. These reward criteria will be tailored to incentivise RL agents to maximise profits from exported variable-priced electricity.

II. REINFORCEMENT LEARNING

In this section, RL agent is implemented into the EMS. The hyperparameters and reward criteria will be tuned to encourage the RL agent to optimise for the minimal cost of

2023 International Future Energy Electronics Conference (IFEEC) | 979-8-3503-3988-8/23/$31.00 ©2023 IEEE | DOI: 10.1109/IFEEC58486.2023.10458579

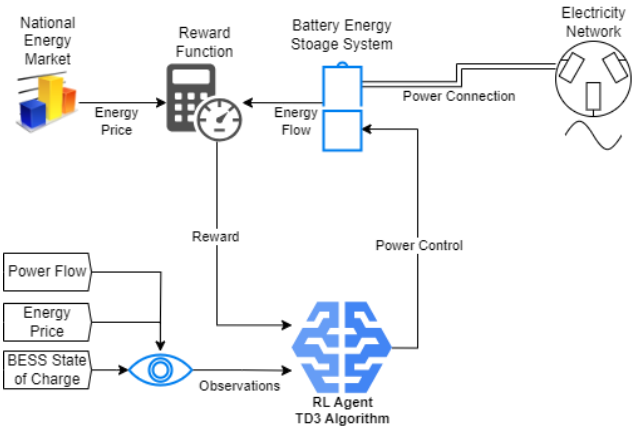


Fig. 1 Microgrid Model for Economic MPC

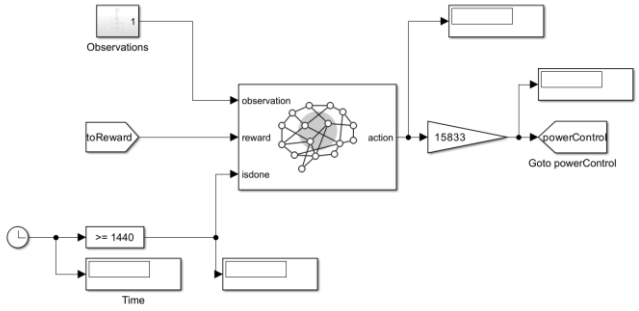


Fig. 2 Simulink Block Diagram of RL Agent and it's ports.

electricity throughout the day. After the agent has been tuned in its best performing configuration, it's performance will be compared with the benchmark controllers in the initial Simulink Microgrid model used for training. Then finally, the best performing RL agent will be compared to the benchmark controller regarding their performance in a dynamic environment of stable and varying live market data, simulating a realistic energy market environment.

The overall objective of this paper is to build a RL agent to manage a BESS in a way that leads to a sizable profit. This is to be done by training an RL agent to operate a BESS within a Simulink environment that replicates the real-world electricity market.

As the RL process is realised to be computationally demanding and slow, a more streamlined Simulink model was developed to simulate the integration of a BESS with the electricity market, as shown in Fig. 1. This model accurately meters grid energy consumption and production based on market prices.

The other controllers for which the RL implementation is to be benchmarked against will use an identical model of the BESS and electricity market implementation in a separate Simulink file.

A. General RL Agent Block Configuration

To interact with an RL agent in Simulink, it requires the placement and configuration of an RL Agent block in the Simulink environment which can be seen in Fig. 2.

The inputs of the RL Agent block are the observations, the reward value, and the '*isdone*' flag which indicates to the RL manager that it needs to end the episode/current simulation. In this case, it's configured to end the episode when it reaches the end of the simulated day.

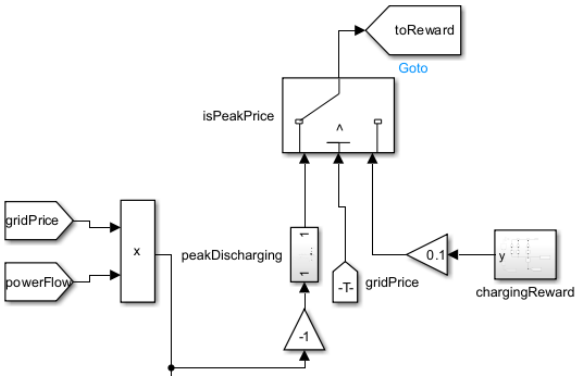


Fig. 3 Simulink Block Diagram of RL Agent's Overall Reward Function

The output is a number within limits that are set in the environment configuration prior to simulating or training the RL Agent. In this case, its configured to output a number between -1 and 1 (being a percentage), which is then passed through a gain block of 15,833 to simulate the minute-by-minute maximum wattage output of a half-scale Tesla Megapack. This output action signal was then forwarded to the power control of the battery subsystem.

The data inputs and action outputs are scaled to be as close to the -1 to 1 range as possible. This was done to standardise the overall data range of the RL agent in an attempt to speed up and stabilise learning.

B. RL Agent Reward Function Configuration

The RL Agent relies on a numerical reward provided by the reward function to gauge whether its actions are suitable based on the circumstances in the environment. This numerical reward represents the desirability of its current actions within the environment. This will in-turn guide the agent to learn optimal behaviours by seeking to maximise cumulative rewards (profit) over time.

The overall reward function built in Simulink for the RL Agent can be seen above in Fig. 3. The reward function is based on the current electricity price and the charging/discharging behaviour.

In the centre of Fig. 3's diagram, the '*isPeakPrice*' switch will switch based on the peak price value, determined by using the 19th out of the 20 quantiles derived from the electricity price forecast.

At the lower left corner of Fig. 3's diagram, the grid price is multiplied by the current power flow to/from the battery. This result is then transformed through a -1 scale to convert the negative value, linked to power discharge from the battery, into a positive value representing earnings. These earnings are then fed into the 'peak Discharging' subsystem. The transformed values are then routed to the '*isPeakPrice*' switch, which forwards them to the RL Agent when operating during peak electricity prices.

Conversely, near the lower right corner of the diagram in Fig. 3, the 'charging Reward' subsystem scales down the reward value to the desired level using a 0.1 gain block, which is then passed to the '*isPeakPrice*' switch to be forwarded to the RL Agent when operating outside of peak prices.

C. Determining the Optimal RL Agent Algorithm

The MATLAB RL Designer tool was used to create, train, and simulate RL agents in the Simulink environment. Before

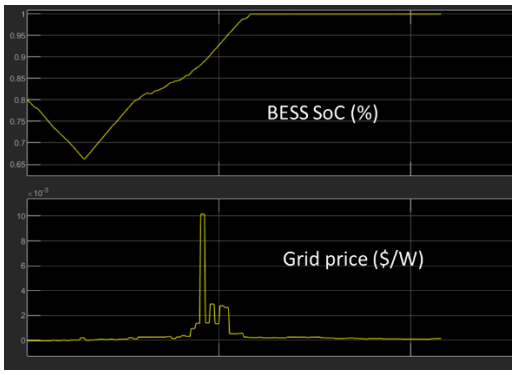


Fig. 4 BESS SOC (top) and Electricity Price (bottom) VS Time After 1900 DDPG Training Episodes.

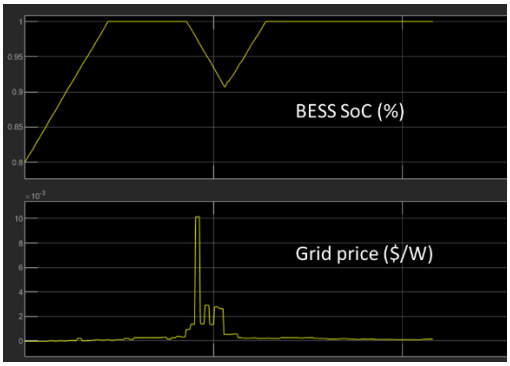


Fig. 5 BESS SOC (top) and Electricity Price (bottom) VS Time After 765 TD3 Training Episodes.

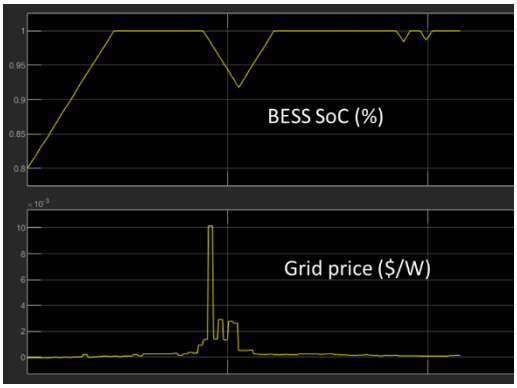


Fig. 6 BESS SOC (top) and Electricity Price (bottom) VS Time After 3500 TD3 Training Episodes.

any long-term training and fine-tuning of the RL Agent's reward function can take place, it is important to determine an efficient RL Algorithm with effective hyperparameters.

MATLAB's RL Designer tool provide a range of algorithms to choose from when first creating an RL agent. Algorithms were tested in the order recommended by the MATLAB documentation for continuous action space environments, starting with DDPG being the simplest compatible agent, followed by TD3, PPO, and SAC, which are then followed by TRPO.

The first RL algorithm trialled was a Deep Deterministic Policy Gradient (DDPG) Agent. Following an extended training period of 1900 episodes, the agent returned to making undesirable actions as depicted in Fig. 4, discharging during periods of low prices and charging when prices were high.

The Twin-Delayed Deep Deterministic (TD3) Policy Gradient Agent was trialled in the training environment using the same testing parameters as the previous DDPG algorithm. After its first 765 episodes of training, the agent exhibited highly desirable action-taking behaviour. Fig. 5 demonstrates that the agent immediately charges the battery to full and discharges the battery when the price spike occurs, to then revert to recharging the battery again once the power price falls back to normal levels. As the RL agent progressed past 3500 episodes of training, its actions grew more responsive to the increasing electricity prices. This is evident in Fig. 6, showing it discharging slightly quicker with more certainty when compared to its previous performance shown in Fig. 5.

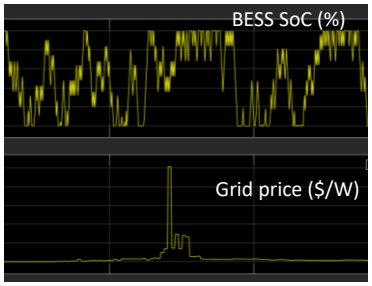
As shown in Fig. 7, the three other algorithm options being the PPO, SAC, and TRPO agents all produced unsatisfactory results after day-long training times. The PPO agent's actions remained constantly sporadic throughout training. The SAC was very slow to train with itself and TRPO agents consistently making poor economic choices and never improving. Due to this, these agents were swiftly ruled out from consideration.

After trialling all agents within the recommended algorithm, the TD3 algorithm remained as the preferred choice for the final EMS RL agent.

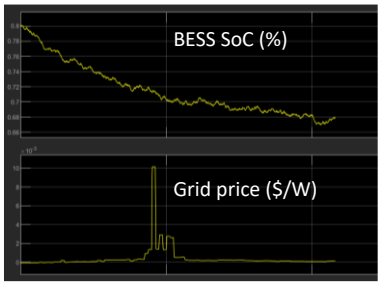
D. Developing the Final RL Agent

After selecting TD3 algorithm, extended training has been carried out. The sensitivity of the model has been fine tuned to best catch the price variation so the profit can be maximised. Fig. 8 shows the process of finalising the RL agent.

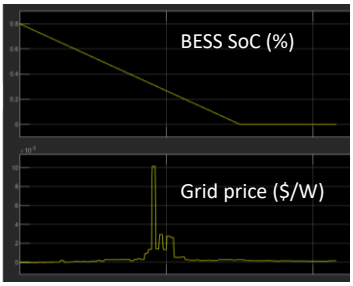
The TD3 agent was re-loaded into the RL Designer tool, where the Gaussian noise options were adjusted from their default values to 0.4 for the Exploration Model and 0.5 for the Target Policy Smoothing Model. These settings control the agent's level of randomness when exploring new actions and its ability to learn from accumulated rewards. Additionally,



1000 PPO Training Episodes
Proximal Policy Optimization (PPO) Agent



500 SAC Training Episodes
Soft Actor-Critic (SAC) Agent



2000 TRPO Training Episodes
Trust Region Policy Optimization (TRPO) Agent

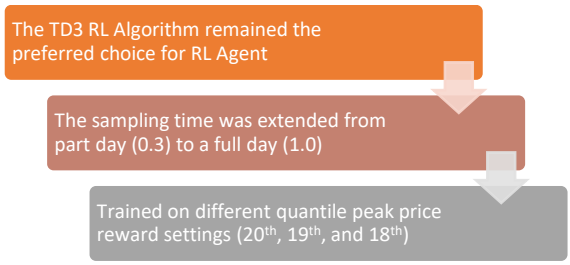


Fig. 8 Block diagram for developing the final RL agent.

Fig. 9 Q20, Q19, and Q18 quantile values Peak Price Performance Results.

the agent's sample time was increased to 1.0, enabling it to learn from complete days through 1440-minute simulated episodes.

Initially, the RL agent was trained using a peak power price threshold set at the 20th quantile value derived from the forecasted power price. After 2000 episodes of training, the Q20 agent exhibited deliberate and profitable behavior by charging the battery during periods of low electricity prices and discharging it during significant price spikes, resulting in a $1700 profit.

Subsequently, the peak price threshold was shifted to the 19th quantile to increase sensitivity to minor price spikes. This adjustment significantly improved the agent's responsiveness, leading to a $1874 profit, demonstrating its ability to capitalize on smaller price fluctuations compared to the Q20 agent.

However, when the peak price threshold was further reduced to the 18th quantile, the Q18 agent exhibited counterproductive behavior, keeping the battery fully charged and ignoring legitimate price spikes. This was likely due to oversensitivity in the reward function, causing the agent to become desensitized to minor price fluctuations, negatively impacting its performance.

When comparing the performance and price sensitivity of the Q20, Q19, and Q18 quantile values, it is clear that the performance of the Q19 agent, with its 19th quantile price threshold, exhibits the most optimal actions which consequently yields the highest generated profits out of the three options. As shown in Fig. 9, for the comparison of the 19th and 20th Quantile peak price reward values the 20th Quantile agent actions are very deliberate and desirable; however, the minor morning price peak is missed. The 19th

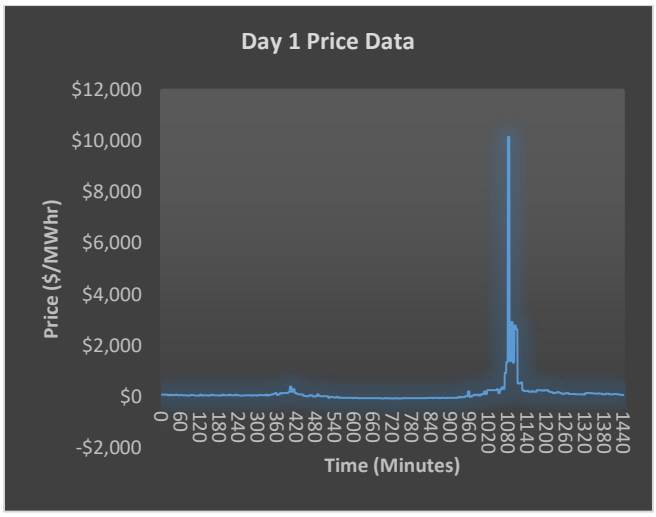


Fig. 10 Price vs Time of Major Evening Peak Price Data.

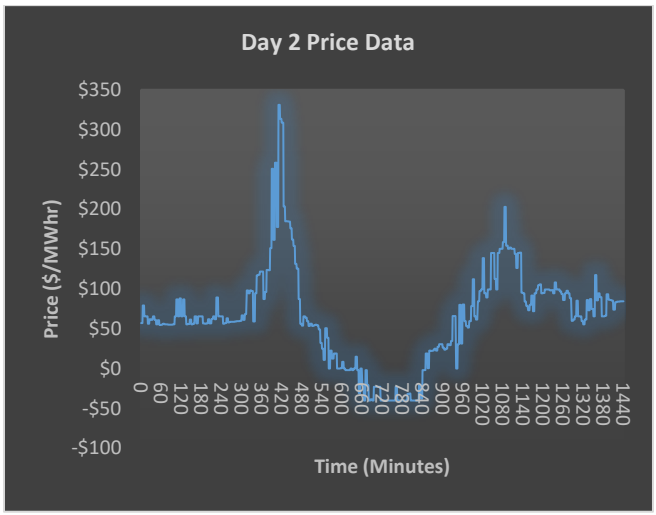


Fig. 11 Price vs Time of Larger Morning and Smaller Evening Peak Price Data.

Quantile agent is significantly more sensitive to minor power price spikes, as shown by the discharging patters in the lower graph. This extra sensitivity resulted in a greater profit of $1874 over the prior $1700.

To see if making the reward function even more sensitive to peak prices, the 18th Quantile setting was attempted. It is evident that the agent completely ignored any spike in power price and kept the battery fully charged resulting in missed profits. Due to this fact, the Q19 agent has been selected as the RL Controller for the benchmarking process.

III. Simulation Results

This section will assess the benchmarking process and results of the two BESS controllers and discuss the effectiveness of each controller based on the circumstances of the intended application and electricity market behaviour.

This section compares the RL controller and Heuristic controller using two different days of energy market data: one of which has a significant afternoon peak as shown in Fig. 10 and the other having a larger morning peak and a smaller one in the afternoon as shown in Fig. 11.

Fig. 12 Comparing the power control behaviour of the RL Agent with the Heuristic time-based controller, Day 1 price data.

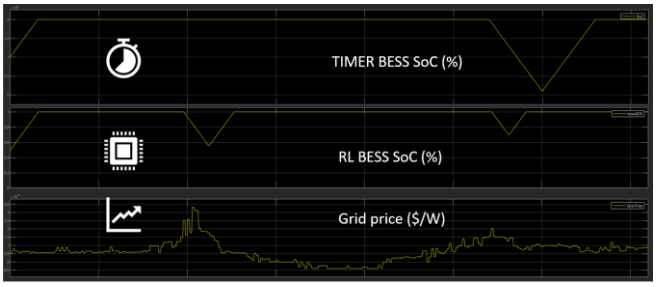


Fig. 13 Comparing the power control behaviour of the RL Agent with the Heuristic time-based controller, Day 2 price data.

TABLE I BENCHMARKING RL VS TIMER-BASED CONTROL METHOD

Market Price Data	RL Profit ($)	Timer Profit ($)
Regular Market Day 1	**$1,874**	$1,744
Regular Market Day 2	**$91**	-$2

Both controllers were benchmarked and compared using standard energy market scenarios, involving power price peaks in morning and evening times of the day.

The Day 1 market data shown in Fig. 10 is the market behaviour of a day with a substantially large price peak in the evening time. When comparing the power control behaviour of the RL Agent with the Heuristic time-based controller, as shown in Fig. 12, the RL agent effectively responds to the minor morning peak by selling power to the grid based on its learned behaviours and rewards from previous training. Whereas the Heuristic controller only discharges between its set timer of 6-8pm, which coincides with the peak evening consumption times. Additionally, during the afternoon peak time, the RL controller reacts to minor price drops by allowing the battery to recharge slightly before further discharging when the power price increases again, as evidenced by the slight fluctuations in its charging/discharging behaviour.

When comparing the end of day 1 profit of the Heuristic controller to be $1744, against the profit of the RL controller being $1874, it is shown that the reactivity of the RL controller to the price fluctuations of the electricity market proves to be beneficial from a profit standpoint when compared to a pre-determined time-based charge/discharge controller like the Heuristic controller which misses peaks outside of the set time.

In contrast, the market data for Day 2, shown in Fig. 11, is the market behaviour of a day with a larger morning peak and comparatively smaller afternoon peak. The RL agent's power control behaviour on Day 2, shown in Fig. 13, mirrors the same discharging behaviour as observed on Day 1 data from Fig. 12 during periods of higher power prices. In comparison, due to the Heuristic controller limited to discharging power only between 6-8pm, it misses the morning peak prices entirely and only discharges during the evening peak electricity hours.

At the end of day 2, the Heuristic controller incurred a cost of $2. In comparison, the RL controller achieved profit of $90, which highlights the advantage the RL agent's responsiveness to the market price gives it over the traditional time-based Heuristic controller. It demonstrates the RL controller's ability to consistently make economically favourable decisions regardless of the daily market activity and changing price patterns. The simulation results for the profit has been summarized in Table I.

IV. CONCLUSION

This paper explored the feasibility of a RL agent serving as an EMS controller to make the most profitable decisions in electricity markets, including volatile ones. Through experiments with a simulated BESS, the study found that the RL agent consistently outperformed traditional timer-based heuristic control methods, effectively adapting to market dynamics and leading to higher profits. Particularly in volatile markets, the RL agent's flexibility in responding to price fluctuations resulted in significant profit advantages over the heuristic controller. These findings underscore the adaptability and effectiveness of the RL agent in managing energy consumption and profit generation for BESSs, positioning it as a valuable tool for optimizing energy use in dynamic energy markets.

REFERENCES

[1] Z. Liu and Y. Du, "Evolution towards dispatchable PV using forecasting, storage, and curtailment: A review," Electric Power Systems Research, vol. 223, p. 109554, 2023.

[2] A. W. F. Wesley Cole, Chad Augustine, "Cost Projections for Utility-Scale Battery Storage: 2021 Update," National Renewable Energy Laboratory2021, Available: https://www.nrel.gov/docs/fy21osti/79236.pdf.

[3] D. S. Miles, "Supercharging Queensland's future as the battery capital," ed, 2023.

[4] A. Mohammed, S. S. Refaat, S. Bayhan, and H. Abu-Rub, "AC Microgrid Control and Management Strategies: Evaluation and Review," IEEE Power Electronics Magazine, vol. 6, no. 2, pp. 18-31, 2019.

[5] L. Meng, E. R. Sanseverino, A. Luna, T. Dragicevic, J. C. Vasquez, and J. M. Guerrero, "Microgrid supervisory controllers and energy management systems: A literature review," Renewable and Sustainable Energy Reviews, vol. 60, pp. 1263-1273, 2016/07/01/ 2016.

[6] H. Shayeghi, E. Shahryari, M. Moradzadeh, and P. Siano, "A Survey on Microgrid Energy Management Considering Flexible Energy Sources," Energies, vol. 12, no. 11, p. 2156, 2019.

[7] X. Chen, Y. Du, E. Lim, L. Fang, and K. Yan, "Towards the applicability of solar nowcasting: A practice on predictive PV power ramp-rate control," Renewable Energy, vol. 195, pp. 147-166, 2022.

[8] R. Subramanya, S. A. Sierla, and V. Vyatkin, "Exploiting Battery Storages With Reinforcement Learning: A Review for Energy Professionals," IEEE Access, vol. 10, pp. 54484-54506, 2022.

[9] X. Chen et al., "Robust Proactive Power Smoothing Control of PV Systems Based on Deep Reinforcement Learning," IEEE Transactions on Sustainable Energy, 2023.

[10] X. Xiao, "Reinforcement Learning Optimized Intelligent Electricity Dispatching System," Journal of Physics: Conference Series, vol. 2215, no. 1, p. 012013, 2022/02/01 2022.

[11] P. Graham, "Control of Residential Battery Charge Scheduling using Machine Learning," 2019.

[12] J. M. Specht and R. Madlener, "Deep reinforcement learning for the optimized operation of large amounts of distributed renewable energy assets," Energy and AI, vol. 11, p. 100215, 2023/01/01/ 2023.