Data Driven Energy Management of Residential PV-Battery System Using Q-Learning

Krishna Baberwal[1], Anshul Kumar Yadav[2], Vikash Kumar Saini[3], Ravita Lamba[4], Rajesh Kumar [5]
[1,4,5]Department of Electrical Engineering, Malaviya National Institute of Technology Jaipur, India.
[2]Central Electronics Engineering Research Institute, Council of Scientific & Industrial Research, Pilani, India.
[3]Center for Energy and Environment, Malaviya National Institute of Technology Jaipur, India.

2023 IEEE International Conference on Recent Advances in Systems Science and Engineering (RASSE) | 979-8-3503-4167-6/23/$31.00 ©2023 IEEE | DOI: 10.1109/RASSE60029.2023.10363461

***Abstract*—Data-driven energy management of residential PV-battery systems using Q-learning offers several benefits, including optimal energy consumption, integration of renewable energy, improved grid stability, cost savings, and flexibility. These advantages contribute to the efficient and sustainable operation of residential energy systems and support the transition towards a cleaner and more resilient energy future. This research focuses on making a violation free, automated energy management system for residential loads using a model free reinforcement learning (RL) algorithm. The objective is to minimize the energy consumption of the system by leveraging the capabilities of the Photovoltaic (PV) system, battery storage, and home load. The energy management problem formulates and describes the state space, action space, and reward structure for Q-learning. This approach learns an optimal policy for energy management based on historical data and feedback from the system. A comprehensive reward function is proposed to ensure a proper battery energy utilization policy. The Australian household PV profile and load curve over a 24-hour horizon with an interval of half an hour are used to examine the performance of the proposed method.**

***Index Terms*—Residential Load, Q-learning, Energy management system, PV-Battery energy storage system.**

NOMENCLATURE

Indices	
t	Time step
Parameters	
SoC_{min}	Minimum value of SoC (%)
η_C	Efficiency (charging)
η_D	Efficiency (discharging)
SoC_{max}	Maximum value of SoC (%)
γ	Discounted reward
S_t	State space
A_{st}	Action space
α	Learning rate
P_L	Load power (kW)
P_{PV}	PV generated power (kW)
S_0	Initial state
λ	On-the-spot grid tariff for G2C
θ	Selling price discounted factor (AUS $)
E_{cap}	Battery capacity (kW)
Variables	
EC	Cost paid to the grid (AUS $)
P_{BC}	Battery charging power (kW)
P_{BD}	Battery discharging power (kW)
P_{G2C}	Power buy from the grid (kW)
P_{C2G}	Power sell to the grid (kW)
P_{BESS}	Energy supplied by the battery (kWh)
P_{grid}	Net energy exchange (kWh)
P_a	Transition probability
R_t	Reward function

I. INTRODUCTION

Residential PV-battery systems provide homeowners with the ability to store excess energy generated by their PV system [1]. One of the significant benefits of this system is the ability to have backup power during grid outages. Residential PV-battery systems can lead to cost savings by reducing reliance on the grid and avoiding peak-time electricity rates [2]. Household owners can store excess energy in energy storage systems during off-peak hours and use it during peak demand periods, reducing the need for additional power generation from fossil fuel sources and promoting the use of clean, renewable energy [3].

The operation and technical constraints associated with residential PV battery systems are complex system installation and operation, maintenance difficulties, efficient utilization of energy storage, low battery protection levels, Cost and size limitations, and battery technology limitations [4]. Similarly, intermittency of energy generation, integration with the existing grid, grid stability, and location specificity are some operation challenges of residential renewable microgrids. The change in the physical structure of the end user is controlled by the development of distributed energy technologies, a combination of technological advancements, supportive policies, and innovative financing models [5]. By investing in renewable energy infrastructure, promoting energy independence, utilizing energy storage systems, providing incentives and subsidies, and utilizing advanced control systems, the adoption of renewable energy can overcome the challenges of residential renewable microgrids and contribute to a more sustainable and resilient energy future.

In the literature, multiple optimization techniques and machine learning algorithms are considered which have been implemented to minimize the cost of energy system [6]–[8]. In [9] author proposed a heuristics-based solution and other

priority rules. The direct search method has been utilized for energy management of PV microgrids, including start-up, maintenance, and operating cost constraints [10]. The purchasing and selling costs of energy from the local utilities are also incorporated. In [11], the control logic-based MPC technique is utilized to improve microgrid performance and reliability. Advanced machine learning algorithms, such as reinforcement learning, also find applications in residential energy management. Appropriate operation strategies have been proposed for residential battery energy [7], whereby battery energy does not depend on predicted PV power generation. In addition, RL-based approaches, i.e., Q-learning [12], Proximal policy optimization (PPO) [13], Deep Deterministic Policy Gradient (DDPG), and Trust Region Policy Optimization (TRPO) [14] have been used to obtain the economic cost of energy consumption in the residential community. In [13] Q-learning is utilized for home energy management. Similarly, in [15] authors have done optimal power scheduling using the same Q-learning in a smart energy building. In [16], [17], the performance of Q-learning is good, but the action state is limited by 2. The other action, such as consumer-to-grid (C2G) is solved by a trading algorithm.

The paper aims to improve the residential PV-Battery system performance and minimize energy consumption costs using the data-driven model. The RL-based Q-learning has been utilized to optimize energy and reduce the net power consumption of the system. Q- learning has adaptive optimization capabilities that are able to handle complex environments, optimal control of battery charging and discharging, integration of multiple objectives, and flexibility in different system configurations. It offers a data-driven approach to optimizing energy consumption and improving the overall performance of residential systems. The main contribution of this work is as follows:

1) Design an efficient energy management strategy that maximizes the consumption of PV power generated by the photovoltaic system while minimizing reliance on the grid and reducing electricity costs for residential users.
2) Data-driven Q-learning is suggested to optimize the energy consumption of the residential PV battery system.

The remaining paper is structured as follows: Section II defines the EMS framework using the proposed algorithm, The mathematical formulation of the Q-learning algorithm in Section IV. Section V illustrated the results and discussion.

II. Problem Formulation

This section describes the problem formulation of energy management systems including power, energy, and state of charge constraints. The problem solution is addressed through the development of optimization models considering control strategies of PV battery systems. The basic framework of the residential grid energy management using the proposed algorithm is shown in Fig. 1. It is assumed that firstly demand meets through generated PV power if the excess remaining energy can be used to charge the battery or feed to the grid for profit. In the non-solar hours, the demand is met by the battery, if the battery could not able to satisfy the remaining load demand then the remaining demand meet from the grid. The Q-learning agent is used as an EMS controller for decision-making. The explanation of Q learning is presented in Section III. The main aim of energy management control is to minimize the electricity utilization of households.

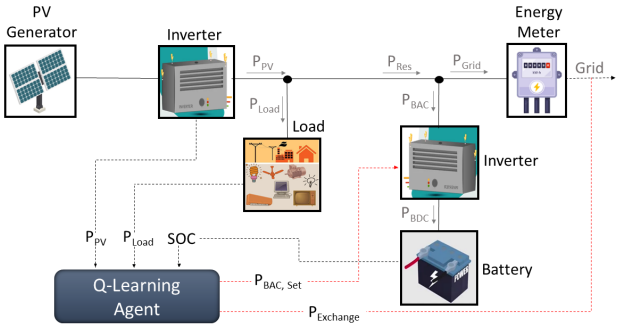

Fig. 1: PV-Battery energy storage system with Q-learning agent

The cost function for a grid-connected residential home can be given by Eq. (1).

$$TC = min \sum_{1}^{48} EC_t \tag{1}$$

where EC_t is the cost paid to the grid operator for energy consumption from the grid. The cost associated with using PV is neglected. Grid is a state in formulation, it can either supply to load when PV-BESS does not meet their demand (Grid to consumer (G2C)) or buy from the consumer if the consumer has excess energy (consumer to Grid (C2G)).

The cost function after installing the PV-battery system is presented by Eq. (2).

$$EC_t = P_{G2C}(t)\lambda_{(g,t)} - P_{C2G}(t)\theta\lambda_{(g,t)} \tag{2}$$

where θ is 0.8, P_{G2C} is power buy from the grid, P_{C2G} indicates the power sell to the grid, $\lambda_{(g,t)}$ is indicating the on-spot grid electricity price $(\$/kWh)$.

The consumer-to-grid tariff is modeled using $\lambda_{(g,t)}$ with discounted factor, θ i.e. known as the selling price. The constraints of Eq. (2) are as follows:

$$P_L(t) = P_{PV}(t) + P_{BESS}(t) + P_{grid}(t) \tag{3}$$

$$0 \leq P_{grid}(t) < P_{grid\ max} \tag{4}$$

The battery has two states, either Charge or Discharge. In the feasibility analysis of residential PV-Battery systems,

the optimal BESS size for the installed PV system is 7 kW considered as presented in [18]. The battery constraints are as follows:

$$SoC_{min} \leq SoC(t) \leq SoC_{max} \tag{5}$$

The value of SoC_{min} is 20% while SoC_{max} is 100% considered for this study. The updated present state of energy of storage is indicated by Eq. (9).

$$E_{min} = SoC_{min} * E_{cap} \tag{6}$$

$$E_{min} \leq E_{(i,t)} \leq E_{cap} \tag{7}$$

$$\sum_{(i=1)}^{N} E_{(i,t)} \leq E_{max} \tag{8}$$

$$E_{(i,t)} = E_{(i,(t-\Delta t))} + (\eta_C * P_{(i,t)}^{BC}) - \left(\frac{P_{(i,t)}^{BD}}{\eta_D}\right) \tag{9}$$

III. Problem Solution

The energy management problem defined in section II is solved by Q-learning. In this approach, an environment is present with an agent who takes appropriate action to get the optimal reward. In this context, an agent is a learner and decision-maker. In a present state, an agent takes action in the environment, by which the state will update, and the agent gets into a new state in which it again interacts with the environment to continuously update its state to get the best reward value. To derive an optimal policy that maximizes the expected value of discounted rewards is the overarching goal of the learning approach.

The energy management of the PV-BESS is viewed as a Markov decision process. The MDP process is represented by the four-factor (S, A, P_a, R_a, γ) in this study, where A is action space, and S is state space. The transition probability indicated by $P_a(s_{t+1}|s_t, a_t) \rightarrow [0, 1]$ is the transition probability from the current state to the next state when executing the action. Reward function indicates by R_a. γ represents the discount factor. To solve this MDP with Q-learning in the presence of the environment, execution of the current policy of the Q-learning agent takes place.

$$\pi^*(s_t) = \arg\max_{\pi} \mathbb{E}\left[\sum_{k=0}^{\infty} \gamma^k r_{t+k+1} \mid \pi\right] \tag{10}$$

where $\mathbb{E}$ and π^* denote the expected value and the optimal policy, respectively.

Q-learning is an off-policy model, which works on the Markovian decision framework [19], [20]. It continuously enhances the evaluation quality of a specific action in a specific state at every time step [21]

The purpose of this framework is to select the optimal set of actions given the current state. To accomplish this, it may come up with a set of rules of its own. To maximize discounted reward, the agent finds the optimal policy. The discounted reward received in s steps for the future is less by the γ factor. The γ varies between γ^s $(0 < \gamma < 1)$.
The Q-lookup table/state-action pair is used, which tells how good an action A is for a particular state S for a policy [22]. The State (S), Action (A), and Reward (R) are as follows for energy management.

A. State

The present state of an agent in its environment is a state. The Bellman equation is used to determine the value of a particular state. From this, it can be estimated how well it is for the acting agent to be in that position. The highest optimum value is given by the optimum position. A discount rate, which determines its importance to the current position for our agent to find the next position. The learning rate controls the learning aspect of the acting agent. The Q-learning equation derived from Bellman's equation is presented in Eq. (11).

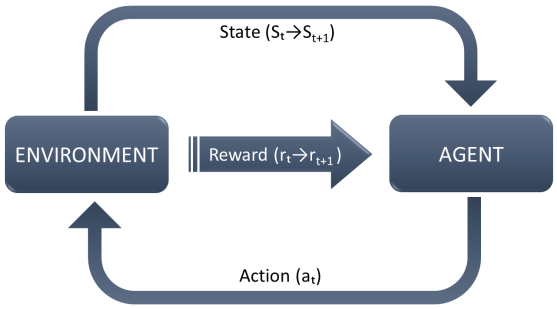

Fig. 2: Q-learning components

$$Q(S_t, A_t) \leftarrow Q(S_t, A_t) + \alpha\left[R_{t+1} + \gamma \max_a Q(S_{t+1}, a) - Q(S_t, A_t)\right] \tag{11}$$

$$S_t = [(P_{Load} - P_{PV}),\ SoC\,] \tag{12}$$

B. Action

In the presence of the environment, the RL agent has to take appropriate actions to update the state, so in this article, the agent considers four actions, namely, charging and discharging of batteries and energy supplied to the grid or PV availability. Based on the energy drawn from the grid. At each time step t the action pool (A_{st}) consists as presented in Eq. (13).

$$A_{st} = \begin{bmatrix} P_{BC} \\ P_{BD} \\ P_{C2G} \\ P_{G2C} \end{bmatrix} \tag{13}$$

C. Reward

Reward is an expression used to guide the agent learning process towards its goal. It helps with the selection of the correct action for the objective of the algorithm. Reward (R_t) has been divided into two parts. The first reward framework deals with the immediate reward received by performing the action (A_{st}). The second reward framework evaluates the overall performance of the agent based on the violation. The reward expression is presented in Eq. (14).

$$R_t = \begin{cases} (P_{G2C}(t)Prc_g(t)) - P_{C2G}(t)\theta Prc_g(t) \\ \quad +P_{BESS}Prc_g(t) + 5 \\ \\ -3 \quad , \quad violation > 0 \end{cases} \tag{14}$$

If there is no violation in 48 step time, then the reward is incremented as:

$$R_t = R_t + 10 \tag{15}$$

The constant values present in the reward function are perfected via the method of trial and error. In equation 14, the term P_{BESS} encourages the acting agent to maintain an optimal battery level for future loads. With a high value of $\lambda_{(g,t)}$, the agent will receive a higher reward when discharged at that particular time.

D. Simulation material

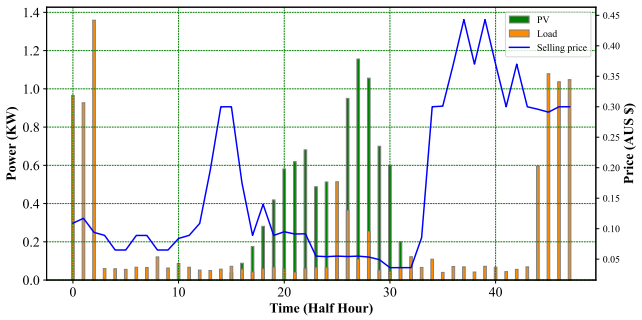

Fig. 3: Load and PV profiles with selling price over one day

TABLE I: Simulation Parameters

Hyperparameters	Selected Value	Values
Epsilon	ϵ	0.1
Learning rate	α	0.5
Discount factor	γ	0.9
Time step	t	1
Battery capacity	E_b	7.0 kW
Initial SoC	SoC_0	1
Selling price discounted factor	θ	0.80

1) Simulation data: In this study, the data is taken from Australian Energy Market Operator price data and Australian Grid. PV-load has been utilized for the year 2013. The data has 1,048,576 samples. Load and PV profiles, along with the grid price of energy over one day, are demonstrated in Fig 3. The Time of Use (ToU) grid tariff has been considered at $25^{th} July$ 2023. The 20000 simulation episodes have been set, and a stopping mechanism has been in place if no violation occurs for 3 continuous iterations.

2) Simulation parameters: As the Q-learning outcome is highly sensitive to parameters α and γ, it is of great importance to select parameters. The value of epsilon was set to 0.1, so the chances of selecting an action suitable for S_t are greater after a few episodes. The simulation parameters are presented in Table I.

3) Simulation methodology: The pseudocode of the proposed problem solution using Q-learning is presented in Algorithm 1. The goal of the acting agent is to maximize the long-term episode reward using an optimal policy from an initial state S_0.

Algorithm 1 Pseudocode of Problem Solution

```
Initialize the q-table with state-action pair Q(S, A).
Initialize the parameters γ, α, ϵ.
while Stopping criteria not fulfilled do
    Establish the Home Energy Environment.
    for t = 1 to 48 do
        Get Current State, S_t = P_Load, P_PV, SoC
        Choose an action from the available actions utilizing the greedy policy(s).
        In the environment, carry out the chosen action and note the reward R_t and the new state S_{t+1}.
        Update the q-table for the chosen action.
    end for
    Calculate the Overall episode reward.
    Exit Home Energy Environment.
end while
```

IV. Results and Discussions

This section presents the simulation results of the proposed problem by applying the Q-learning algorithm.

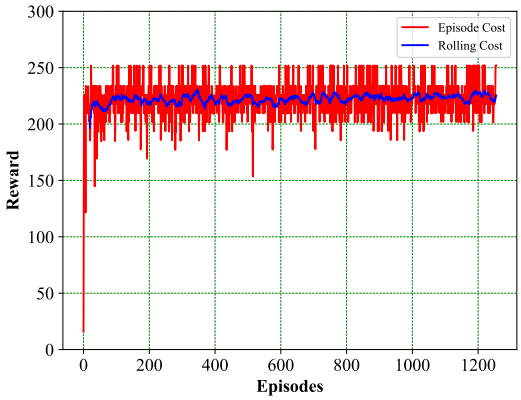

Fig. 4: Reward in context of cost

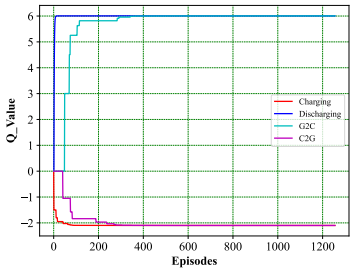

Fig. 5: Q-value regarding different action

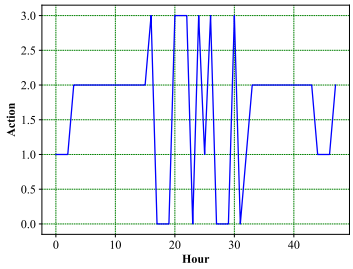

Fig. 6: Action space of algorithm

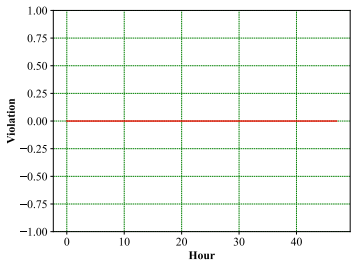

Fig. 7: Violation per hour of algorithm

The Fig. 4 shows the reward and the rolling reward over the training period. Since the minimum exploration probability ϵ, and system parameters (e.g. pv, load, SoC) vary in each episode results in the episode reward fluctuation within a small range. To show the changing trend of rewards more clearly, the graph provides the rolling average value of the past 10 episodes. This assessment shows the learning progress of the acting agent over the training episodes based on the reward it receives. The goal is to learn from the environment without getting penalized.

The agent underwent a learning phase to understand the environment before reaching the final episode (800). During this period, the agent received relatively low and fluctuating rewards. However, between episodes 800 and 1200, the rolling reward became more stable, indicating that the agent gained a better understanding of the environment. After episode 1255, a stopping mechanism was activated to halt the simulation. At this point, convergence is achieved as the acting agent selects the best action for a given S_t. The agent's learning progress has been recorded in the Q-table, which stores information about the likelihood of the agent taking specific actions for each state, forming state-action pairs. To generate the states during the simulation, the researchers opted for a dynamic approach rather than predefined states. This approach offered a significant advantage by drastically reducing the total number of states. As a result, the simulation became more efficient and practical for real-world applications.

Fig. 5 depicts the graphical interpretation of the state-action pair for the initial state S_0 (1.0, 1.0) in the context of the residential PV-battery system. In this state, there is an additional load of 1.0 kW, and the battery is charged to its maximum value (100%). After 1255 episodes, the acting agent reached the conclusion that the optimal actions for this particular state are discharging and grid-to-consumer (G2C) transfer. Initially, the acting agent favoured using the BESS to meet the additional load requirements. However, through the learning process, it was later discovered that both actions (C2G and G2C) have similar impacts on the state due to the low price (P_{G2C}) and the moderate load level. The agent effectively learned from the rewards accumulated during the learning process, enabling it to identify the correct actions with fewer episodes. Furthermore, the agent learned that the actions of G2C and C2G transfers are incorrect and result in penalties. The combination of positive rewards and penalties facilitated the agent's ability to identify the correct actions more efficiently. As a result, the agent decided to discard the actions of charging and C2G, which signifies a clear understanding of the physical significance of additional PV power generation, extra load, battery operation, and grid interactions. A similar state-action pair is established for each existing state. The corresponding actions at each time step up to $t = 47$ are illustrated in Fig. 6. The acting agent selects the action of charging (P_{BC}) and grid consumption (P_{C2G}) when excess PV power is available, and it chooses discharging (P_{BD}) and grid-to-consumer (P_{G2C}) action when there is an additional load. The agent consistently adheres to the correct actions, showcasing its proficiency in accurately identifying suitable actions for each given state. The violation graph is also presented in Fig. 7.

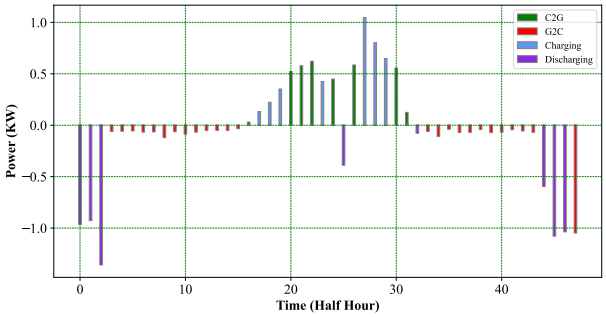

Fig. 8: State space

The simulation results related to the residential environment are presented in Fig. 8. Additionally, the SoC of the battery illustrates in Fig. 9. Initially, a decrease in SoC can be observed as no PV power is available. The battery is seen to supply the net load up to $t = 3$. As the load decreases from $t = 4$, the acting agent switches to G2C mode, utilizing grid power to meet the load demands. With the availability of PV power, the acting agent shifts to charging mode first and later to C2G mode, selling PV power to the grid to generate a feasible income. Throughout these actions, the agent ensures that a sufficient amount of battery SoC remains for future use. Such

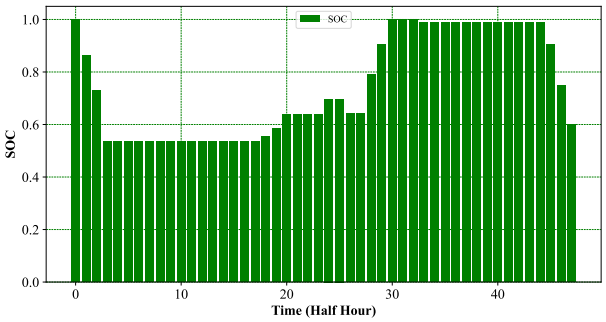

Fig. 9: SoC of PV-Battery system

behavior is promoted through reward engineering. As the load significantly reduces from $t = 33$, the acting agent utilizes grid power to save consumer costs and gradually switches to battery operation as the load increases. The reward function ensures the proper utilization of the installed BESS, as evidenced in Fig. 9. The total power purchased from the grid is 2.579 kW, while the total power sold to the grid is 3.479 kW. The overall cost after $t = 47$ time steps is 0.4422 Aus$, demonstrating that the algorithm is proactive during the decision-making process.

V. Conclusion

In this study, the Q-learning algorithm was applied to a residential PV-BESS model to automate the energy management process. The simulation considered four actions, ensuring the practicality of the residential model. The acting agent successfully learned to make correct decisions based on a two-step reward function, achieving proficiency within 800 episodes. The agent's actions appropriately responded to dynamic grid traffic, effectively utilizing the battery during high load periods and relying on the grid during significantly low load periods. Consequently, by the end of the simulation, the energy sold to the grid at discounted prices exceeded the energy purchased. The learning approach employed effectively utilized available infrastructure like PV and batteries to reduce costs. In the future, the model could be further enhanced by integrating battery degradation considerations, providing more accurate results while considering the battery's health. Additionally, the model could be extended to include trading at the community level, allowing for more comprehensive energy management and collaboration among residential users.

References

[1] V. K. Saini, R. Kumar, A. S. Al-Sumaiti, and B. Panigrahi, "Uncertainty aware optimal battery sizing for cloud energy storage in community microgrid," *Electric Power Systems Research*, vol. 222, p. 109482, 2023.

[2] G. Zhang, W. Hu, D. Cao, Z. Zhang, Q. Huang, Z. Chen, and F. Blaabjerg, "A multi-agent deep reinforcement learning approach enabled distributed energy management schedule for the coordinate control of multi-energy hub with gas, electricity, and freshwater," *Energy Conversion and Management*, vol. 255, p. 115340, 2022.

[3] T. Bocklisch, "Hybrid energy storage approach for renewable energy applications," *Journal of Energy Storage*, vol. 8, pp. 311–319, 2016.

[4] V. K. Saini, A. Seervi, R. Kumar, A. Sujil, M. A. Mahmud, and A. S. Al-Sumaiti, "Cloud energy storage based embedded battery technology architecture for residential users cost minimization," *IEEE Access*, vol. 10, pp. 43685–43702, 2022.

[5] S. Yelisetti, V. K. Saini, R. Kumar, and R. Lamba, "Energy consumption cost benefits through smart home energy management in residential buildings: An indian case study," in *2022 IEEE IAS Global Conference on Emerging Technologies (GlobConET)*, pp. 930–935, IEEE, 2022.

[6] R. Liemthong, C. Srithapon, P. K. Ghosh, and R. Chatthaworn, "Home energy management strategy-based meta-heuristic optimization for electrical energy cost minimization considering tou tariffs," *Energies*, vol. 15, no. 2, 2022.

[7] C. Guan, Y. Wang, X. Lin, S. Nazarian, and M. Pedram, "Reinforcement learning-based control of residential energy storage systems for electric bill minimization," in *2015 12th Annual IEEE Consumer Communications and Networking Conference (CCNC)*, pp. 637–642, 2015.

[8] W. Li, T. Logenthiran, V. Phan, and W. Woo, "Implemented iot-based self-learning home management system (shms) for singapore," *IEEE Internet of Things Journal*, vol. 5, pp. 2212–2219, June 2018.

[9] E. Handschin, F. Neise, H. Neumann, and R. Schultz, "Optimal operation of dispersed generation under uncertainty using mathematical programming," *International Journal of Electrical Power Energy Systems*, vol. 28, no. 9, pp. 618–626, 2006. Selection of Papers from 15th Power Systems Computation Conference, 2005.

[10] F. A. Mohamed and H. N. Koivo, "System modelling and online optimal management of microgrid using mesh adaptive direct search," *International Journal of Electrical Power Energy Systems*, vol. 32, no. 5, pp. 398–407, 2010.

[11] G. Bruni, S. Cordiner, V. Mulone, V. Sinisi, and F. Spagnolo, "Energy management in a domestic microgrid by means of model predictive controllers," *Energy*, vol. 108, pp. 119–131, 2016. Sustainable Energy and Environmental Protection 2014.

[12] F. Härtel and T. Bocklisch, "Minimizing energy cost in pv battery storage systems using reinforcement learning," *IEEE Access*, vol. 11, pp. 39855–39865, 2023.

[13] E. Kuznetsova, Y.-F. Li, C. Ruiz, E. Zio, G. Ault, and K. Bell, "Reinforcement learning for microgrid energy management," *Energy*, vol. 59, pp. 133–146, 2013.

[14] Y. Ji, J. Wang, J. Xu, and D. Li, "Data-driven online energy scheduling of a microgrid based on deep reinforcement learning," *Energies*, vol. 14, no. 8, 2021.

[15] S. Kim and H. Lim, "Reinforcement learning based energy management algorithm for smart energy buildings," *Energies*, vol. 11, no. 8, 2018.

[16] S. Yelisetti, V. K. Saini, R. Kumar, R. Lamba, and A. Saxena, "Optimal energy management system for residential buildings considering the time of use price with swarm intelligence algorithms," *Journal of Building Engineering*, vol. 59, p. 105062, 2022.

[17] G. Muriithi and S. Chowdhury, "Optimal energy management of a grid-tied solar pv-battery microgrid: A reinforcement learning approach," *Energies*, vol. 14, no. 9, p. 2700, 2021.

[18] U. Mulleriyawage and W. Shen, "Optimally sizing of battery energy storage capacity by operational optimization of residential pv-battery systems: An australian household case study," *Renewable Energy*, vol. 160, pp. 852–864, 2020.

[19] B. Singh, R. Kumar, and V. P. Singh, "Reinforcement learning in robotic applications: a comprehensive survey," *Artificial Intelligence Review*, pp. 1–46, 2022.

[20] P. Yadav, V. K. Saini, A. S. Al-Sumaiti, R. Kumar, *et al.*, "Intelligent energy management strategies for hybrid electric transportation," in *2023 IEEE IAS Global Conference on Renewable Energy and Hydrogen Technologies (GlobConHT)*, pp. 1–7, IEEE, 2023.

[21] R. S. Sutton and A. G. Barto, *Reinforcement learning: An introduction*. MIT press, 2018.

[22] C. J. Watkins and P. Dayan, "Q-learning," *Machine learning*, vol. 8, pp. 279–292, 1992.