IJECBE (2025), **3**, **1**, 43–68
Received (4 December 2024) / Revised (20 January 2025)
Accepted (21 January 2025) / Published (30 May 2025)
https://doi.org/10.62146/ijecbe.v3i1.98
https://ijecbe.ui.ac.id
ISSN 3026-5258

International Journal of Electrical, Computer and Biomedical Engineering

RESEARCH ARTICLE

Grid Import Optimization with Adaptive Deep Reinforcement Learning for PV-Battery Systems

Romi Naufal Karim* and Budi Sudiarto

Department of Electrical Engineering, Faculty of Engineering, Universitas Indonesia, Depok, Indonesia
*Corresponding author. Email: rominaufalkarim@gmail.com

Abstract

This article explores the application of Deep Reinforcement Learning (Deep RL) to optimize energy management in photovoltaic (PV) and battery systems. The framework introduces innovations such as Rule-Based Action Smoothing for maintaining system stability, Proximal Policy Optimization (PPO) Multi-House Training to generalize across diverse energy usage patterns, and Post-Controller Integration to address real-time operational issues. Although the dataset originates from Ireland, the model is adapted to Indonesia's dual-tariff system and energy regulations. Simulation results demonstrate significant cost reductions, achieving up to 76.61% in stable scenarios and 7.03% in high-variability environments. While RL outperforms load following (LF) in complex scenarios, its efficiency remains limited in stable conditions, highlighting the need for further optimization of the RL framework. Despite these challenges, the methodology demonstrates flexibility and resilience in leveraging renewable energy to reduce costs and improve system efficiency. Future improvements, such as refining reward functions, addressing over-discharge, and integrating seasonal data, will further enhance the model's applicability across a broad range of scenarios. This scalable approach supports Indonesia's renewable energy goals and provides insights for intelligent energy systems in residential contexts.

Keywords: Deep Reinforcement Learning, Proximal Policy Optimization, Energy Management, Photovoltaic and Battery Systems

1. Introduction

In the last few decades, there has been a tremendous expansion in using renewable energy sources, mainly because of the dire need to address global warming due to carbon emissions. International efforts like the Paris Agreement have brought forth the necessity of restricting the increase in global temperatures to 1.5°C and achieving net-zero emissions by the year 2050 [1]. The Intergovernmental Panel on Climate Change stresses that these goals are crucial to preventing disastrous consequences due to climate change. These have made governments worldwide develop policies to speed up the transition into renewable energy, with large targets and financial incentives directed toward capacity in renewable energies.

The Indonesian government has targeted that, by 2025, renewable energy will make up 23% of the country's energy mix, under Government Regulation No. 79/2014 [2]. That has given a spur to applying photovoltaic systems with battery storage for residential and industrial sectors to improve energy efficiency and lessen dependency on the electrical grid. The steep drop in lithium-ion battery costs, driven by the rapid electrification of the vehicle market, has made them a viable option for integrating renewable energy sources [3][4]. These reductions bring benefits in increasing techno-economic viability through longer running times, high efficiencies, and larger energy densities. In turn, photovoltaic battery systems are increasingly more affordable, allowing houses to save on energy expenses due to better self-consumption.

While inherently laden with potential, photovoltaic battery systems (PV-battery systems) face many hurdles in the area of energy management, especially in the aspect of minimizing grid imports. Ineffective energy management may lead to irregular grid imports, high energy costs, and inefficient battery utilization. In this regard, artificial intelligence is one of the most important players and an up-and-coming tool in developing intelligent control systems for PV-battery integration. Reinforcement learning (RL) is one of the most promising approaches in artificial intelligence since it copes well with large data sets and finds optimal policies in stochastic scenarios [5]. Unlike classical control methods, the RL approach does not explicitly require information about the environment or a definition of predetermined decision-making structures. Therefore, it is adequate for dynamic and uncertain systems, such as PV – battery setups in which solar production and consumption profiles present erratic variations. Moreover, RL does not need a precise mathematical model and can learn directly from empirical data.

Numerous studies have leveraged RL for the management of PV-battery systems. Elshazly et al. [6] proposed a reinforcement learning framework utilizing Proximal Policy Optimization (PPO) to manage house battery charging within smart grids in a single-agent multi-environment system. Their approach improved grid stability, fairness, and customer satisfaction by dynamically optimizing power allocation. On a community scale, Xiong et al. [7] applied Deep Q-Networks (DQN) to achieve more efficient energy scheduling than conventional methods such as round-robin and first-come-first-serve, reducing operational costs while enhancing local resource utilization.

On the other hand, Xia et al. developed actor-critic reinforcement learning models to enhance energy consumption in commercial buildings by factoring in human input for better energy efficiency, comfort, and indoor air quality [8]. Ali et al. applied Q-Learning for optimizing battery management in dairy operations with variable renewable energy production, showcasing reinforcement learning's versatility [9]. Additionally, Almughram et al. [10] developed a RL system integrating V2H (Vehicle-to-House), stationary batteries, and PV panels for residential energy management. Combining fuzzy Q-learning and deep learning, their approach reduced energy costs and improved self-consumption under ToU (Time of Use) and RTP (Real Time Price) tariff.

Real et al. [11] introduced an optimization framework based on RL incorporating Deep RL with load forecasting for PV-battery systems. Their approach significantly reduced energy costs and grid dependency, leveraging accurate load predictions to enhance RL model performance. However, dependency on accurate load forecasting introduces complexity, as prediction errors can negatively impact RL decision-making. Kang et al. [12] developed an RL-based scheduling model for a residential building with a PV system and BESS using PPO. Their model effectively traded off self-sufficiency against peak load reduction, outperforming traditional TD3 and SAC algorithms under South Korea's energy tariff structure.

In addition, Härtel and Bocklisch [13] proposed a PPO-based RL framework for PV-battery storage systems that achieved cost-efficient power allocation without requiring explicit load forecasting. Their approach simplified system complexity compared to traditional methods like Model Predictive Control (MPC) while maintaining high performance. However, similar to Kang et al. and Real et al., their work was limited to single-house training, raising questions about scalability and adaptability to diverse energy consumption profiles. Building on these efforts, Qi et al. [14] introduced EnergyBoost, an RL-based framework combining MPC and Advantage Actor-Critic (A2C) to manage house batteries. When implemented on a low-cost Raspberry Pi device, their system demonstrated significant energy cost reductions.

While these works highlight the flexibility and potential of RL in energy management, they also reveal shared limitations. For instance, Real et al. [11] approach relies heavily on accurate forecasting but lacks validation for handling sudden and extreme demand fluctuations. Similarly, although achieving cost-efficient energy scheduling, both Kang et al. [12] PPO-based model and Härtel and Bocklisch [13] PPO framework do not address mechanisms for recovering from RL system failures, such as sub-optimal decision-making or unstable training processes. These shared limitations could affect their robustness and reliability under highly dynamic or uncertain conditions.

Previous literature has demonstrated the potential of RL in optimizing energy management. However, most studies tend to focus on single-building scenarios or specific regulatory frameworks, limiting their applicability to a broader range of conditions. This study addresses these limitations by developing a generalized RL model that can operate in previously unseen environments (such as different houses) without retraining. Although the model is explicitly designed for Indonesia's dual-tariff system and local regulations as a proof of concept, its underlying methodology is crafted to be

transferable, ensuring adaptability across various regulatory and operational contexts. The multi-house training approach enhances the model's generalization ability by learning from diverse energy consumption patterns.

To bridge these gaps, this study introduces several key innovations:

- Rule-Based Action Smoothing: This innovation improves system performance stability by preventing sudden changes, thus promoting smoother battery operations and extending battery life. It also enhances reliability, defined in this study as the system's ability to maintain consistent and stable battery operations by avoiding abrupt fluctuations in battery levels. This ensures a steady energy flow and stable performance under varying energy demand and supply conditions, focusing solely on normal operations without considering fault scenarios.
- PPO Multi-House Training: Rather than basing the PPO training on single-building profiles from earlier models, the present work utilizes multi-residence training to include various energy consumption behaviors. This increases the flexibility of a reinforcement learning agent to fit different conditions and is, therefore, generalizable to heterogeneous residential areas.
- Post-Controller Integration: To deal with the challenge of real-time variability, a PostController is integrated to complement the RL agent. This mechanism ensures stability by dynamically responding to demand spikes or system failures—a critical feature absent in many existing models.

1.1 Reinforcement Learning

Reinforcement Learning (RL) is a subfield of machine learning where an agent learns to make decisions by interacting with an environment, aiming to maximize cumulative rewards over time. The process relies on trial and error, where the agent receives feedback in the form of rewards or penalties based on its actions [15]. Unlike supervised learning, which requires labeled data, RL learns from sequential interactions, making it well-suited for dynamic and complex environments such as energy management, robotics, and game playing.

The RL framework is typically formalized as a Markov Decision Process (MDP) characterized by:

- State (s), represents the current condition of the environment.
- Action (a), The set of all decisions the agent can take.
- Reward (r), feedback received by the agent based on the quality of its action.
- Policy (π), a strategy that maps states to actions.

The agent's goal is to learn a policy (π^*), that maximizes the expected cumulative reward over time, considering the discount factor (γ) for future rewards and to maximize the expected cumulative reward, R_t, defined as:

$$R_t = \sum_{k=0}^{\infty} \gamma^k r_{t+k+1} \tag{1}$$

That RL process is effectively illustrated in Figure 1, which provides a clear overview of the interaction between the agent and the environment [16]. RL is

a cyclic process in which an agent learns how to make optimal decisions while interacting with the environment. At every iteration, it observes the current state (s_t) of the environment, encapsulating its present conditions. With this, the agent makes an action (a_t) guided by policy (π) this means a general framework for behavior. This action's execution affects the environment, triggering a transition into a new state (s_{t+1}) and generating a reward (γ_t) that informs how good the action actually is. This reward becomes a feedback system for the agent to improve its policy iteratively. Through repetition of this interaction, an agent adjusts its strategy for maximizing the overall reward through a careful balancing of short-term and long-term gains. Through a series of interactions, the agent learns an optimal policy (π^*) that allows it to navigate the environment and successfully achieve its long-term objectives.

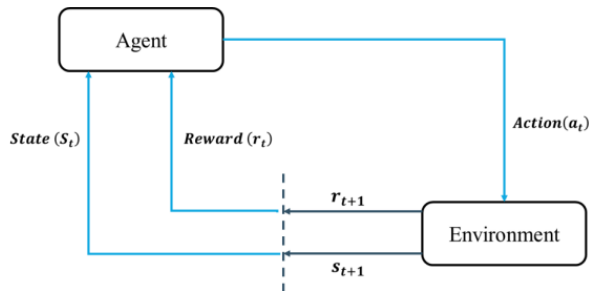

Figure 1. Reinforcement Learning Process: Interaction Between Agent and Environment

1.2 Deep Reinforcement Learning

Deep Reinforcement Learning (Deep RL) develops and extends the fundamental ideas in traditional RL by bringing the RL methodologies together with the deep neural networks, thereby giving an effective solution to handle big state and action spaces and their management. Compared with the conventional RL approaches using explicit representations and hand-engineered features, Deep RL uses neural networks as universal approximators, enabling agents to generalize to enormous and continuous environments. This would eventually allow Deep RL to scale up and solve complicated real-world tasks, from robotics to energy optimization and game playing. The classic example of DQN, as introduced by Mnih et al. in [17], was the first successful attempt at approximating the Q-value function using convolutional neural networks, playing Atari games at human-level performance— beating the so-called "curse of dimensionality" behind large state spaces.

While traditional RL has focused on learning policies or value functions from tabular representations, Deep RL automated feature extraction from raw inputs, significantly reducing the dependence on domain expertise. The last has been especially true for tasks needing end-to-end learning, where an agent directly learns from raw sensory data without pre-processing. With its ability to automatically extract features, Deep Reinforcement Learning adapts to dynamic and changing environments, which leads to significant developments in areas such as autonomous systems and renewable energy.

Deep RL contains a few algorithms: value-based, policy-based, and actor-critic. Valuebased methods, like DQN, approximate the optimal action-value function for the agents to select actions that maximize expected rewards. Policy-based approaches, such as Proximal Policy Optimization (PPO), update policies directly to learn effectively and obtain consistent performance in continuous action spaces [18]. Actor-critic architectures, as represented by Deep Deterministic Policy Gradient (DDPG), incorporate both value-based and policyoriented components; an actor is used for proposing actions, and a critic is used for estimating the value of proposed actions, making it suitable for high-dimensional environments with continuous actions [19].

This study has chosen PPO as it balances exploration and exploitation effectively, which is critical in managing dynamic and stochastic energy environments. Exploration enables the RL agent to test and discover new strategies for handling variations in energy consumption and PV production. At the same time, exploitation ensures that the agent applies learned policies to achieve stable and efficient energy management. By integrating PPO, this study addresses the need for adaptability across diverse residential energy consumption patterns, ensuring the system can reliably optimize battery operations while minimizing grid import costs under varying conditions.

2. Preliminary Framework and Data Analysis

2.1 Proposed System Framework

An overview of the proposed system architecture is depicted in Figure 2. In this paper, the controller is designed as a hybrid framework consisting of a proposed Deep Reinforcement Learning (Deep RL), especially the PPO model, and a Post-Controller for risk management. The Deep RL model serves as the primary decision-making unit, optimizing energy flow by deciding actions such as charging (P_t^{Ch}) or discharging (P_t^{Dis}) the battery and importing power from the grid (grid import ((P_t^{Grid}))). The model operates based on inputs such as PV production (P_t^{PV}), residential load demand (L_t), and the current battery level.

The Post-Controller is designed to complement the RL decisions, ensuring the system's reliability and robustness. It guarantees that the demand of the house load (L_t) is always met, even if the decisions taken by the RL model are insufficient to supply the required power. In such cases, the Post-Controller independently sources the extra power from the grid without influencing the RL model's decision-making process. This complementary nature of the Post-Controller provides an additional layer of security by mitigating the risks of suboptimal decisions in a dynamic and uncertain environment.

While the initial system framework addresses the core challenges of energy flow optimization and reliability, this study introduces a critical enhancement: improving the generalization capability of the Deep RL model. Unlike the traditional approach, which trains the model on a single house, this study will simultaneously train the Deep RL model using data from multiple houses. Each house, from House-1 to House-n, represents a unique environment with different energy consumption patterns, photovoltaic production, and battery characteristics. It then trains on the various datasets for each house so that the Deep RL model learns a generalized policy that can adapt when there are changes in energy needs, irregularities in PV production, or

dynamics within battery behavior. Through interaction with various situations, the framework allows the Deep RL model to develop a strong and generalized policy that can adapt to changes in energy needs and environmental conditions.

After training, the Deep RL model is tested on a house that was not in the training dataset. This testing phase evaluates the model's ability to generalize its learned policy to unseen environments. The key evaluation metrics are minimizing grid import costs, maintaining stable battery operations, and adapting effectively to dynamic energy demands. This will ensure the practical success of the model in real-world deployment scenarios, reducing the need for retraining when applied to other houses. The proposed framework addresses the key challenges in energy optimization and paves the way for future integration of renewable energy sources in smart houses to make the energy ecosystem more sustainable and resilient.

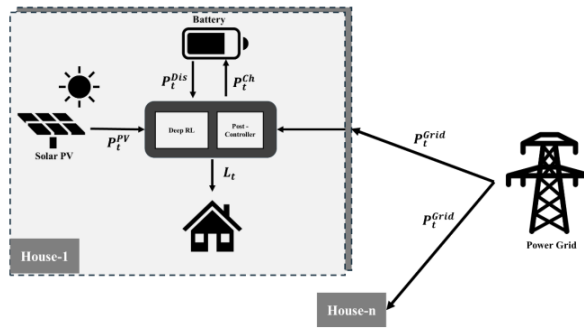

Figure 2. The System Architecture

2.2 Data Profile and Preprocessing

The dataset used in this study originates from the StoreNet Project, which collected highresolution electricity usage data (1-minute intervals) from residential users in an energy community in Ireland throughout 2020 [20]. This dataset includes key variables such as active power usage, solar panel (PV) electricity production, energy exchange with the grid, battery charging/discharging activities, and state of charge (SOC). Ten houses are equipped with PV-battery systems, each featuring a 3.3 kW Sonnen Battery, making the dataset suitable for analyzing house energy dynamics and PV-battery optimization.

To ensure data quality and suitability for analysis, several preprocessing steps were applied:

- Duplicate Removal: Duplicate entries are removed to maintain data integrity.
- Handling Missing Values: Missing values are filled using the column mean for highly variable house consumption data. Meanwhile, more stable variables, such as battery level, charging power, and discharging power, were forward-filled or interpolated as they are less prone to variability.
- Outlier Detection and Clipping: For house consumption, an outlier clipping process was applied using the Interquartile Range (IQR) method with a factor

of 2.5. This factor is chosen to filter extreme noise or measurement errors while retaining naturally high consumption values relevant to energy modeling (e.g., peak usage during heavy appliance operation). Previous studies, such as Jurj et al. [21], used a standard factor of 1.5 for moderate outliers in energy datasets; however, the higher factor in this study accounts for the inherent variability in house data.

- Resampling to 30-Minute Intervals: To reduce resolution while maintaining temporal granularity, data is resampled into 30-minute intervals. Numerical columns were averaged, while categorical variables were forward-filled to ensure consistency.

As shown in Table 1, house consumption data demonstrates a wide range of variability, with median consumption values spanning from 2.42 kW (House 1) to 24.35 kW (House 10). The maximum value observed, 89.38 kW in House 10, represents high-power usage events typical in real-world scenarios, such as operating heavy appliances. A higher IQR factor of 2.5 was applied to maintain these significant consumption patterns while effectively mitigating extreme anomalies caused by noise or measurement errors.

In this study, the simulation is assumed to be conducted in Indonesia to adjust data processing and electricity tariff schemes to the local context. Although the maximum demand of 89.38 kW exceeds the typical single-phase house capacity in Indonesia (approximately 14 kVA), this dataset was selected to provide a broad range of variability for training the RL model. The goal is to develop a generalized control policy capable of handling diverse consumption profiles, including those that may be less common in local contexts. Consequently, while real-world implementations in Indonesia are expected to have lower maximum loads, the model benefits from exposure to these high-load scenarios during training, improving its adaptability. Moreover, in our experiments, we validate the model on multiple houses—particularly those with low correlation to the training set—to assess how effectively the RL agent adapts to 'unfamiliar' consumption patterns.

To ensure our cost calculations accurately reflect local electricity pricing in Indonesia, we determine the cost of electricity imported from the network based on the Peak Load Time (PLT) and Off-Peak Load Time (OPLT) tariff schemes. According to our resampled data, the PLT tariff is set at IDR 1,035.78 per kW (with a factor of K equal to 2), while the OPLT tariff is IDR 517.89 per kW [22]. The PLT tariff applies during peak hours, particularly from 18:00 to 23:00 local time, whereas the OPLT tariff applies during off-peak hours.

The dataset was systematically filtered and divided into training and testing subsets to evaluate the proposed framework. Statistical properties and correlation values of energy consumption patterns across houses guided the training and testing data selection process. These correlation values are visualized in the heatmap presented in Figure 3, while descriptive statistics for each house are detailed in Table 1. The analysis and dataset partitioning specifically focus on energy consumption data from the summer months (June to August), as this period includes peak PV production and captures diverse house energy consumption behaviors critical for the study.

The grouping of training and testing sets in this study is based on the energy

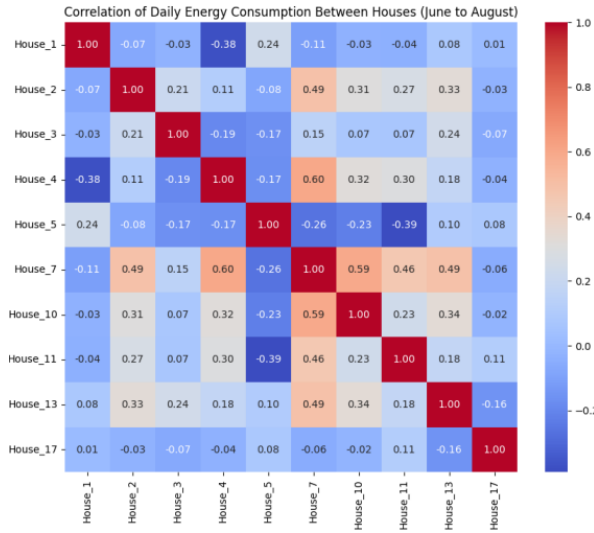

Figure 3. Correlation of Daily Energy Consumption Between Houses

consumption patterns analyzed from the statistics in Table 1 and the correlations shown in Figure 3. Houses 2, 7, 10, and 11 were selected as the training set because they exhibit diverse energy usage patterns and have moderate to strong correlations with each other. In this context, 'moderate to strong correlations' means that the energy consumption patterns in these houses are related, which can help the model learn more effectively, as indicated in Figure 3.

- House 10 (mean: 30.58 kW, std: 15.05 kW) demonstrates high average consumption with significant daily fluctuations, making it ideal for teaching the Deep RL model to handle dynamic and challenging scenarios.
- House 7 (mean: 24.21 kW, std: 10.36 kW) adds moderately variable energy usage patterns, balancing consistency and variability.
- House 11 (mean: 23.20 kW, std: 6.12 kW) contributes relatively stable energy consumption with lower variability, ensuring predictable patterns in the training set.
- House 2 (mean: 18.11 kW, std: 5.04 kW) introduces stable and moderately varying patterns, complementing the diversity within the training set.

These houses provide a representative dataset with varying energy consumption and variability degrees, allowing the Deep RL model to generalize effectively across different scenarios.

Table 1. Energy Consumption Statistics of Houses (June to August)

	House 1	House 2	House 3	House 4	House 5	House 7	House 10	House 11	House 13	House 17
count	92.00	92.00	92.00	92.00	92.00	92.00	92.00	92.00	92.00	92.00
mean	4.35	18.11	19.04	33.79	11.33	24.21	30.58	23.21	11.15	6.55
std	3.04	5.04	2.80	13.73	2.74	10.36	15.05	6.12	1.64	1.94
min	1.07	9.77	10.83	7.95	6.52	3.94	11.63	6.98	8.34	2.93
25%	1.90	14.21	17.34	26.22	9.78	18.19	20.26	20.22	10.21	5.21
50%	2.43	17.08	18.85	32.93	11.14	21.27	24.78	24.35	10.87	6.49
75%	6.99	21.37	20.18	39.36	12.63	27.54	38.20	26.94	11.75	7.78
max	12.06	33.97	26.69	75.57	18.53	54.54	89.38	34.11	18.01	11.78

For testing, Houses 1, 4, 13, and 17 were selected as candidates because they exhibit unique and contrasting patterns that are not prominently represented in the training dataset:

- House 1 (mean: 4.35 kW, std: 3.04 kW) represents low and stable energy consumption.
- House 17 (mean: 6.55 kW, std: 1.94 kW) provides a similarly low consumption profile with even less variability.
- House 13 (mean: 11.15 kW, std: 1.64 kW) demonstrates a moderate but highly stable profile.
- House 4 (mean: 33.79 kW, std: 13.73 kW) is included due to its exceptionally high and variable energy consumption, which provides a challenging and extreme case for the model to evaluate.

Referring to Figure 3, these test houses have low correlations with the training houses (most values below 0.1), making them ideal for evaluating the model's ability to generalize to unseen scenarios. This division ensures the Deep RL model is trained on a broad range of realistic patterns while rigorously testing its adaptability and performance in handling novel energy consumption behaviors, including extreme conditions such as those exhibited by House 4.

3. Decision-Making and Control Framework

3.1 Grid Import Minimization and Energy Balancing

The energy management problem in residential PV-battery systems in this study involves optimizing the operation of energy storage to minimize grid energy costs while ensuring that house energy demand is consistently met. This is a dynamic decision-making problem affected by uncertainties in solar PV production, house energy consumption, and electricity prices. In this study the system operates under the assumption that no power is exported to the grid. All surplus energy from the PV will be stored in the battery or reduced to meet the load demand if the battery is full. Therefore, at timestep t the system must make decisions regarding:

- How much power to charge or discharge the battery
- How much power to import from the grid to meet the house load demand

The optimization process aims minimize the total grid energy cost (C_{total}) over a time horizon (T) by efficiently managing energy flows among the PV system, battery storage, and house demand, while ensuring adherence to operational constraints. Since grid export is not allowed, the cost is given by:

$$C_{total} = \sum_{t=t_i}^{T} P_t^{Grid} \lambda_t \tag{2}$$

The total energy cost is calculated based on the power imported from the grid (P_t^{Grid}) at each timestep (t) and the corresponding electricity tariff (λ_t). Here t_i represents the initial timestep, and T marks the end of the time horizon.

To achieve the objective of cost minimization, the optimization problem is subject to several constraints. First, the total energy supply must equal the house load demand at every time step. Constraint in Equation (3) ensures that energy supply and demand are balanced, accounting for all sources and sinks in the system:

$$P_t^{Grid} + P_t^{V} + P_t^{Dis} \geq +L_t + P_t^{ch} \tag{3}$$

The values of P_t^{Ch} and P_t^{Dis} represent the charging and discharging rates of the battery, respectively. In this study the $P_t^{Ch} = -P_t^{Dis}$ is imposed to reflect the operational characteristics of the battery to ensures that the battery cannot simultaneously charge and discharge at the same time, maintaining the physical feasibility of the system. Furthermore, P_t^{Ch} and P_t^{Dis} are limited the maximum charging and discharging capacities of the battery, as determined by its specifications.

To further ensure the operational safety and efficiency of the battery system, the energy stored in the battery at any time step (E_t^{bat}) is constrained within a safe range between 10% and 90% of its maximum capacity (E_{max}^{bat}):

$$0.1\, E_{max}^{bat} \leq E_t^{bat} \leq 0.9\, E_{max}^{bat} \tag{4}$$

This restriction prevents overcharging and deep discharging, which could lead to premature battery degradation and reduced system reliability. While (a_t) represents the intended action by the agent, the actual change in E_t^{bat} . Specially a_t may not directly translate to energy flow due to safety limits (10–90% of battery capacity). The actual energy change is calculated as the difference in battery level ($\Delta E_t^{bat} = E_{t+1}^{bat} - E_t^{bat}$) after considering these constraints.

$$E_{t+1}^{bat} = \begin{cases} E_t^{bat} + \eta_{ch} a_t \ \ if \ \ a_t > 0 \\ E_t^{bat} + \frac{a_t}{\eta_{Dis}} \ \ if \ \ a_t < 0 \\ E_t^{bat} \ \ if \ \ a_t = 0 \end{cases} \tag{5}$$

Equation (5) is used to calculate E_t^{bat} for the next timestep iteration based on the action taken by the RL agent. The charging and discharging efficiencies are represented by $\eta_{ch}$and η_{Dis}, respectively. These parameters are needed to ensure that energy transitions are realistic and within the battery system's physical bounds.

3.2 Reinforcement Learning-Driven Control Design

The control scheme proposed in this study utilizes PPO as the core decision-making algorithm for optimizing battery charging and discharging. This approach addresses the dynamic and stochastic nature of energy consumption and PV production. By formulating the problem as a Markov Decision Process (MDP), PPO learns optimal energy management policies through direct interaction with the energy system, ensuring effective and reliable control under varying conditions [15].

The MDP formulation in this study is adapted from the framework introduced by Real et al.[11], previously applied to PV-battery systems to define transition dynamics and grid import rules. This MDP structure remains fundamental in describing the relationships among PV production, energy consumption, and energy management decisions. However, this study introduces some important modifications to enhance the stability and efficiency of the system, as detailed below:

3.2.1 State Representation Refinement

The state space has been reduced from eight variables to seven, as shown in Equation (6). This refinement eliminates redundancy, such as the load prediction variable, making the control system computationally more efficient. The retained variables include PV production (P_t^{PV}), load demand (L_t), grid import (P_t^{Grid}), current and previous battery levels (E_t^{bat} , E_{t-1}^{bat}), electricity tariff (λ_t), and the battery energy change ($\Delta E_t^{bat} = E_{t+1}^{bat} - E_t^{bat}$).

$$S_t = \left\{ L_t, P_t^{PV}, P_t^{Grid}, E_t^{bat}, E_{t-1}^{bat} \Delta E_t^{bat}, \lambda_t \right\} \tag{6}$$

This reduction ensures the model focuses on essential dynamics while eliminating the need for predictive input data, streamlining computational demands.

3.2.2 Action Space and Transition Dynamics

The action space comprises decisions for charging ($a_t > 0$) , discharging ($a_t < 0$) and idle ($a_t = 0$). These actions directly affect E_t^{bat} and P_t^{Grid} , governed by transition dynamics. At each time step t the agent observes the current state (S_t) and selects an action (a_t) based on its probabilistic policy. The minimum change of a_t value is determined by the minimum incremental value i, given that the applied actions are discrete. The action updates the battery energy level and adjusts the grid import there depending on charging or discharging efficiency if is a deficit or surplus of energy (δ_t):

$$\delta_t = P_t^{PV} - L_t \tag{7}$$

The system then transitions to the next state (S_{t+1}), reflecting updated variables, while providing a reward signal to guide learning. This process allows the RL agent to iteratively learn optimal policies under stochastic environmental conditions, effectively adapting to uncertainties in PV production and load demand.

3.2.3 Grid Import Rules

The grid import rules are adapted from Real et al.[11] to ensure the agent appropriately adjusts P_t^{Grid} based on its chosen a_t and δ_t. For charging actions ($a_t > 0$), grid import is determined by:

$$P_t^{Grid} = \begin{cases} a_t - \delta_t \; if \; a_t > \delta_t \\ 0 \; otherwise \end{cases} \tag{8}$$

For discharging ($a_t < 0$), grid import is updated based on battery constraints (E_t^{bat}):

$$P_t^{Grid} \begin{cases} |\delta_t a_t| \; if \; E_t^{bat} > 0.1E_{max}^{bat} and \delta_t \; < \; 0 \; and \; a_t < \delta_t \\ 0 \; if \; E_t^{bat} > 0.1E_{max}^{bat} \; and \; \delta_t < 0 and a_t < \delta_t \\ max(|\delta_t| - |\eta_{Dis}E_{t-1}^{bat}|, |\eta_{Dis}E_{t-1}^{bat}|) if \; E_t^{bat} < 0.1E_{max}^{bat} \; and \delta_t \leq 0 \\ E_t^{bat} if E_t^{bat} < 0.1E_{max}^{bat} and \; \delta_t \geq 0 \end{cases} \tag{9}$$

3.2.4 Reward Function Adjustment

The reward function in this study is designed to guide the RL agent towards achieving practical energy management objectives, such as minimizing grid imports, maximizing the utilization of PV production, preserving battery health, and ensuring system stability. By providing context-sensitive feedback based on the dynamics of charging, discharging, and idle actions, the reward function shapes the agent's learning process to align with real-world operational priorities. At each time step, the total reward (R_{tot}) is computed as:

$$R_{tot} = w_1 R_{ch} + w_2 R_{Dis} + w_3 R_{Idle} + w_4 R_{Grid} \tag{10}$$

Here w_1, w_2 w_3, and w_4 are weights reflecting the priorities of different operational objectives. The highest weight, $w_4 = 0.5$ is assigned to minimize reliance on grid imports, particularly during high-tariff peak hours. The other weights are set as$w_1 = 0.3$, $w_2 = 0.3$, and $w_3 = 0.2$, determined through trial and error to balance the competing objectives effectively.

The charging reward is designed to encourage efficient energy storage while penalizing overcharging and charging during deficits. For actions corresponding to charging ($a_t > 0$), the reward is given as:

$$R_{ch} = \begin{cases} -(\frac{1}{P_{Dis}}) if E_t^{bat} > 0.9E_{max}^{bat} \\ 1.5\frac{1}{P_{ch}}\frac{\delta_t}{P_{ch}} if E_t^{bat} \leq 0.9 \; E_{max}^{bat} \; and \; \delta_t > 0 \; and \; \delta_t \leq 0.1 \; E_{max}^{bat} \\ 3\frac{1}{P_{ch}}\frac{\delta_t}{P_{ch}} if E_t^{bat} \leq 0.9 \; E_{max}^{bat} \; and \; \delta_t > 0 \; and \; \delta_t > 0.1E_{max}^{bat} \\ -\frac{\delta_t}{P_{ch}} \; otherwise \end{cases} \tag{11}$$

From equation above, a significant penalty is applied to discourage overcharging, which can degrade battery health. Small surpluses ($\delta_t \leq 0 \; .1 \; E_{max}^{bat}$). Larger surpluses ($\delta_t > 0 \; .1 \; E_{max}^{bat}$) receive higher rewards, than lower surplus ($\delta_t > 0 and \delta_t \leq 0.1 \; E_{max}^{bat}$), prioritizing efficient utilization of PV production, while deficits ($\delta_t < 0$) and in any other condition incur penalties to discourage grid energy use for charging.

The discharging reward encourages the agent to use stored energy efficiently during energy deficits while avoiding deep discharges or unnecessary discharging during surpluses. For actions corresponding to discharging ($a_t < 0$), the reward is given as:

$$R_{Dis} = \begin{cases} \frac{1}{P_{Dis}} if E_t^{bat} < E_{max}^{bat} \\ 1.5|1 - \frac{\delta_t}{P_{Dis}}| if\ \delta_t < 0 \text{ and not peak hour} \\ 1.5|1 - \frac{\delta_t}{P_{Dis}}| + 0.5 if\ \delta_t < 0 \text{ and peak hour} \\ 1\frac{\delta_t}{P_{Dis}} if\ \delta_t > 0 \text{ and} \delta_t \leq 0.1 E_{max}^{bat} \\ 2\frac{\delta_t}{P_{Dis}} if\ \delta_t > 0.1\ E_{max}^{bat} \end{cases} \tag{12}$$

During energy deficits, the reward scales inversely with the deficit magnitude, ensuring the agent prioritizes significant shortfalls. Additional rewards (+0.2) during peak hours align the agent's behavior with real-world cost-saving goals. Penalties for discharging during surpluses are scaled proportionally to prevent wasteful battery usage. For Idle actions ($a_t = 0$), the reward (R_{Idle}) is given by:

$$R_{Idle} = \begin{cases} 1\ if\ \delta_t = 0 \text{ or } \delta_t \geq 0 \\ 0.5|\delta_t|\frac{1}{P_{Dis}} if\ \delta_t < 0 \\ 0 \text{ otherwise} \end{cases} \tag{13}$$

In balanced conditions and when a surplus is available, agents receive a reward to increase stability and avoid unnecessary adjustments. Then, in a deficit situation, a penalty is applied to motivate agents to actively address the energy shortage, thereby reducing their dependence on the grid. The grid import penalty discourages reliance on external energy sources, particularly during high-tariff peak hours, It is expressed as:

$$R_{Grid} = -log(1 + P_t^{Grid}\lambda_t) \tag{14}$$

The grid import penalty is part of the objective function (Equation (2)). Logarithmic scaling is used to gradually increase the penalty as grid usage and tariff rates rise. This method prevents excessive penalties for small imports while ensuring a proportional penalty that reflects the dynamic nature of energy costs. Using logarithms in the penalty function allows for a more controlled increase in response to higher energy consumption and tariff rates, which helps avoid disproportionately large penalties when grid import increases slightly.

3.2.5 Rule-Based Action Smoothing

One of the major contributions of this work is the incorporation of a rule-based action smoothing mechanism, which was not implemented in the framework of Real et al.'s framework [11]. This mechanism will address the abrupt fluctuations in charging and discharging actions often found in RL-generated policies to have smoother transitions in realworld implementations and operational reliability. The smoothing mechanism incorporates adaptive rules to refine the charging ($a_t > 0$) and discharging ($a_t < 0$) actions based on the battery level (E_t^{bat}), energy balance (δ_t), and previous actions (

a_t–1). Adaptive rules for charging actions. For charging actions ($a_t > 0$), the following rules are applied:

$$a_t = \begin{cases} min(1.5a_t, \eta_{ch}P_{ch}) if\, \delta_t > 0.7P_{ch} \\ min(max(0.05, \frac{\delta_t}{P_{ch}})a_t, \eta_{ch}P_{ch}) if\, \delta_t < 0.2P_{ch} \\ min(a_t, \eta_{ch}P_{ch}) if\ E_t^{bat} \leq 0.2\ E_{max}^{bat} and\ \delta_t \geq 0.5P_{ch} \\ min(0.5a_t, \eta_{ch}P_{ch})\ if\ 0.5\ E_t^{bat} \leq E_t^{bat} \leq E_t^{bat} \end{cases} \quad (15)$$

When the energy surplus exceeds 70% of the maximum charging rate (P_{ch}), the charging action is accelerated by capping to $\eta_{ch}\ P_{ch}$. Conversely, for minimal energy surplus ($\delta_t < 0\ .2\ P_{ch}$), charging is moderated by scaling a_t proportionally to the surplus. Intermediate battery levels (50% to 90% of E_{max}^{bat}) further reduce charging rates for operational stability. At critically low battery levels ($E_t^{bat} \leq 0.2\ E_{max}^{bat}$), charging is boosted if the energy surplus is sufficient ($\delta_t^{bat} \geq 0\ .5\ P_{ch}$). Adaptive rules for discharging actions. For discharging actions ($a_t < 0$), the following rules ensure stability:

$$a_t = \begin{cases} 0\ if\ \delta_t > 0 \\ min(a_t, \eta_{Dis}P_{Dis})\ if E_t^{bat} \geq 0.7\ E_{max}^{bat}\ and\ \delta_t < 0.5\ P_{Dis} \\ min(0.5a_t, \eta_{Dis}P_{Dis})\ if\ 0.3\ E_t^{bat} \leq E_t^{bat} \leq 0.6\ E_{max}^{bat} \end{cases} \quad (16)$$

Discharging is avoided entirely if there is an energy surplus to prevent unnecessary use of stored energy. When the battery is near full capacity $E_t^{bat} \geq 0.7\ E_{max}^{bat}$ and the energy deficit ($\delta_t < 0\ .5\ P_{Dis}$) substantial, the discharging rate is capped to $\eta_{Dis}P_{Dis}$ to meet demand efficiently. For moderate battery levels (30% to 60% of E_{max}^{bat}) discharging is scaled down to maintain system balance while preserving battery health. Action smoothing with inertia and damping. Once the adaptive rules are applied, the action (a_t) smoothed by incorporating inertia (θ) and damping ($\wr$) factors to ensure gradual transitions and reduce abrupt changes:

$$a_t = \theta a_{t-1} + (1 - \theta)a_t \quad (17)$$

$$a_t = a_t \wr \quad (18)$$

The inertia factor (θ) blends the current action with the previous action (a_{t-1}) creating a smoother transition between time steps. The damping factor ($\wr$) further adjusts the smoothed action to enhance stability and reliability under varying environmental conditions. In this study, $\theta = 0.6$ and $\wr = 0.9$ were determined through a process of trial and error. These values were found to provide a balance between stability and responsiveness, ensuring practical operational reliability while maintaining the flexibility of RL-generated policies.

3.3 Post-Controller Design

The Post-Controller is introduced as an ancillary mechanism aimed to ensure the load demand (L_t) is always satisfied. The controller operates independently of the RL model and turns ON to draw power from the grid when the offered energy is

inadequate to meet the demand. he Post-Controller implementation is based on the principle of energy balance introduced in Equation (3), ensuring the total supplied energy is not less than L_t . The energy gap (G_t), representing the energy supply shortage, is defined as:

$$G_t = (P_t^{Grid} + P_t^{PV} - P_t^{ch} + |P_t^{Dis}|) - L_t \tag{19}$$

Where, actual power P_t^{ch} and P_t^{Dis} are determined by changes in the battery level, (ΔE_t^{bat}) as expressed in Equation (20) and (21). To guarantee that L_t is consistently satisfied, the Post-Controller imports additional energy from the grid equivalent to the calculated energy gap. This would activate the Post-Controller to supplement the RL-calculated grid imports by adding the absolute value of the energy gap, $|G_t|$), only when the energy gap is negative.

$$P_t^{ch} = \frac{\Delta E_t^{bat}}{\eta_{ch}} if\, \Delta E_t^{bat} > 0 \tag{20}$$

$$P_t^{Dis} = \Delta E_t^{bat} \eta_{Dis}\ if\ \Delta E_t^{bat} < 0 \tag{21}$$

4. Experimental Simulation Framework

The RL environment in this study was implemented using the OpenAI Gym framework [23], leveraging Python programming. The computational setup utilized for the simulation consisted of a laptop with an AMD Ryzen 3 5300U processor (8 CPU cores at 2.6 GHz) and 12 GB RAM. The hyperparameters for the PPO model, presented in Table 2, were carefully fine-tuned to achieve optimal performance over a 7-day horizon with a 30-minute timestep resolution.

Table 2. Hyperparameter Setting of PPO

Parameter	Value	Description
Policy Network	MlpPolicy	The neural network architecture used to approximate the policy
Seed	42	Ensures reproducibility by initializing a fixed random number generator
Learning Rate	1×10^{-5}	Controls the step size for updating the neural network weights
Discount Factor	0.99	Determines the importance of future rewards in the decision-making process
Timesteps per Update	336	Number of time steps collected before performing a policy update
Batch Size	112	Number of samples used in each gradient update to improve learning stability
Clip Range	0.1	Limits the magnitude of policy changes to ensure training stability
Value Function Coefficient	0.5	Scales the contribution of the value function loss in the total loss
Gradient Norm Clipping	0.5	Prevents large gradient updates by capping the gradient norm
Entropy Coefficient	0.01	Promotes exploration by encouraging policy diversity

The PPO model was trained using data that had been resampled to a 30-minute resolution to accurately reflect real-world energy usage patterns. The primary dataset and its variables, outlined in Table 3, were used to configure the energy management environment and simulate realistic scenarios.

Table 3. Dataset Variables

Parameter	Value
λ_t (electricity price) IDR/kW	1,035.78 (18.00 – 23.00), 517.89 (otherwise)
P_t^{Ch} and P_t^{Dis} (charging and discharging rate) kW	1.65
η_{Ch} and η_{Dis} (charging and discharging efficiency)	95%
E_{max}^{bat} (battery capacity) kW	3.3
i (Minimum charging/discharging incremental value) kW	0.01

4.1 Training Phase

The training procedure involves randomizing houses to expose the RL agent to various energy management scenarios. Each episode begins with randomly selecting one of four residences, each with a unique energy consumption and PV production profile. The training is designed to cover 30 complete cycles of each house in the dataset (30 episode per house), ensuring a thorough understanding of each house's energy dynamics. After each episode, the setting is reset, and a new residence is randomly chosen for the next episode. By frequently encountering a wide range of conditions, the agent develops a strong and durable strategy capable of managing the dynamic and stochastic nature of real-world energy systems.

4.2 Testing Phase

The testing phase will assess the PPO model's adaptability to conditions not present in its training set. This is a crucial evaluation of the agent's capacity to adapt to previously unseen houses with varying energy usage and PV production patterns. Unlike the training phase, which exposes the model to a range of dwellings, testing isolates the agent's performance under unexpected situations—mirroring real-world deployment scenarios .

Testing leverages the PPO algorithm's stochastic nature to navigate the inherent uncertainty in PV production and energy consumption. As Bao et al. [24] highlight, stochastic settings accurately represent the unpredictability and fluctuation of renewable energy sources. Unlike deterministic models, the PPO framework learns directly from past data, enabling the system to effectively navigate complex, real-world energy dynamics.

While a rule-based mechanism is used to smooth RL-generated actions for operational stability, the policy remains stochastic. This ensures the agent's ability to explore and adapt flexibly to fluctuating energy conditions. The integration of stochastic policy production with deterministic smoothing guarantees a practical balance between adaptability and operational reliability. The testing process, repeated ten times for each configuration, mitigates the inherent randomness of stochastic policies. These iterations capture variations in policy outcomes, providing a comprehensive

view of the system's performance. The results of the grid import costs are the average values of ten tests. The battery level is represented by the 10th simulation (the last simulation).

5. Result and Discussion

5.1 Simulation Result

Simulations are conducted in this study in order to validate the effectiveness of RL agents in optimizing energy management over houses with diversified energy consumption profiles. These simulations were applied to houses with characteristics different from those in the training dataset, as described in data profile and preprocessing section. In this case, the grid import initial was assumed as zero, and the battery level is initialized with the first value from the dataset.

Table 4. Summary of Average Grid Import and Energy Cost Across Houses (June – August)

	Avg. Grid Import (kW)			Avg. Energy Cost (IDR)		
	Baseline	RL	LF	Baseline	RL	LF
House 1	6.83	2.44	0.90	5395.86	1261.95	609.91
House 4	24.78	29.08	23.64	16197.37	15057.95	15875.75
House 13	7.41	6.54	4.64	4178.57	3388.48	2731.59
House 17	2.97	2.37	0.71	1834.25	1228.06	554.46

Table 5. Summary of Grid Import and Energy Cost Reduction Across Houses (June – August)

	Grid Import Reduction (%)		Energy Cost Reduction (%)	
	RL	LF	RL	LF
House 1	64.32	86.78	76.61	88.70
House 4	-17.34	4.60	7.03	1.99
House 13	11.68	37.35	18.91	34.63
House 17	20.21	75.96	33.05	69.77

Table 4 and Table 5 compare the baseline with the results of RL implementation and simple load following model (LF). The baseline is the result of the original power recording from the dataset without any further changes applied. However, the tariff is adjusted to Indonesian-based to ensure consistency in the cost-value calculation across all models. Therefore, the method of controlling the battery charging action in the baseline is unknown.

Meanwhile, the applied load following works to control battery charging with a simple rule: when there is an energy surplus, the excess energy from PV production is used to charge the battery until it reaches the maximum capacity (90% of the total battery capacity). Conversely, when there is an energy deficit, where consumption exceeds PV production, the battery will be lost to meet energy needs. However, energy release is limited to the minimum battery capacity (10% of the total capacity). If the battery cannot fully meet the energy needs, the energy shortage will be imported from the electricity grid. In this condition, the battery does not charge. In addition, if

PV production remains the same as consumption, there is no charging or releasing of energy from the battery, so the battery status remains at the same level. As shown in Table 4 and Table 5, the RL model can reduce energy costs due to power imports from the grid in each house compared to the baseline. However, RL's performance is still inferior to the next load, resulting in a more significant overall cost reduction.

As shown in Table 4 and Table 5, the RL model can reduce energy costs due to power imports from the grid in each house compared to the baseline. However, RL's performance is still inferior compared to the next load, which results in a more significant overall cost reduction. House 4, however, presents a unique case. Its high average energy consumption of 33.79 kW and very high variability, with a standard deviation of 13.73 kW, challenge simple energy management strategies. However, the RL model's adaptability is impressive, demonstrating superior energy cost efficiency compared to the load following in this scenario.

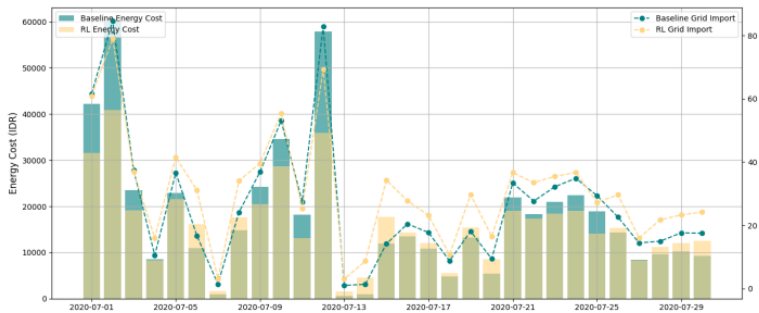

Figure 4. House 4: Daily Comparison of Energy Costs and Grid Imports: Baseline vs RL Agent

In this case, even though the level of power import from the grid is higher than the baseline, at 17.34%, the RL model still manages to generate a lower final cost. This underscores the RL model's ability to make more optimal decisions, particularly in utilizing Off-Peak Load Time (OPLT) and minimizing power imports during Peak Load Time (PLT). The 5.04% efficiency difference between RL and load following in House 4 further demonstrates the potential of RL in handling more complex scenarios compared to simple rule methods like load following. Figure 4 visually represents how the RL agent in House 4 can reduce energy costs in July. In that month, RL was able to cut energy costs by 9.8% despite an additional 15.07% grid import compared to the baseline.

This study found that one of the main weaknesses of the RL model is its tendency to trigger over-discharge, especially in houses with low energy consumption and stable variability, such as House 1, House 13, and House 17. This pattern is seen from the frequent power discharge exceeding the load demand, even when the load has been met. As a result, the battery capacity reaches the minimum limit faster, so the system must rely on power imports from the grid to meet the energy demand at the next timestep. This pattern is analyzed from House 17, with the results in Figure 5 showing the battery level, Figure 6 showing the total available energy, and Figure 7 representing the charging and discharging time.

In contrast, the LF method shows a more efficient power discharge because it is only carried out according to actual needs.

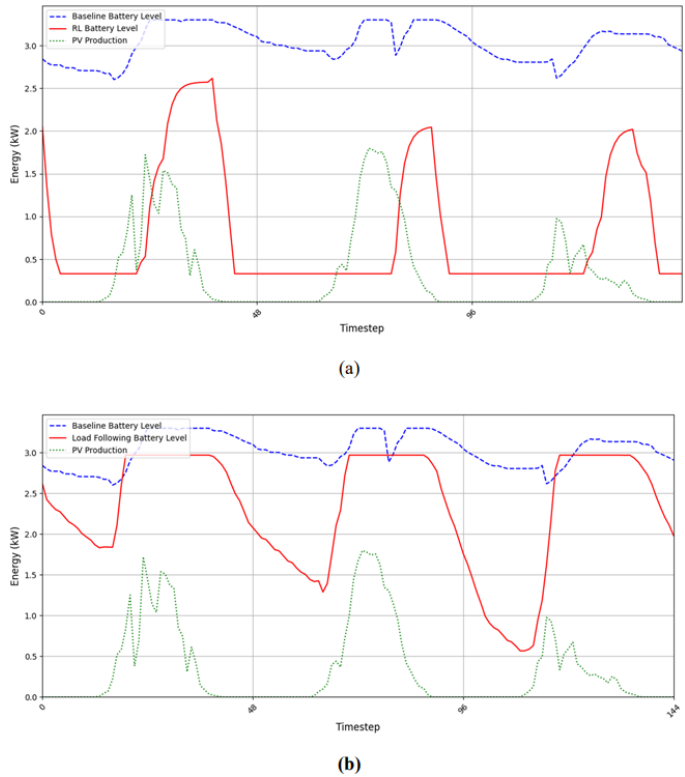

Figure 5. House 17: Battery Levels and PV Production: a) RL & b) LF

One of the causes of this over-discharge is the lack of priority in battery capacity management in the RL reward function. This Overcharge phenomenon can be seen in Figure 6 and Figure 7 before the 48th timestep (1 day). The RL model focuses more on reducing the overall energy cost without considering the impact of excessive battery discharge. From Figure 6, it can also be seen that the designed Post-Controller works well, namely maintaining the load to be met, which is represented by the gray shading. On the other hand, the RL model does not have a validation mechanism that ensures that the battery power released is comparable to the load requirements. This causes energy waste that not only reduces efficiency but also results in suboptimal grid usage.

Table 6 shows that the efficiency of the RL model varies depending on the characteristics of the house. In House 4, which has high consumption and large variability, RL can surpass the load following efficiency even after 10 training episodes, reflecting good adaptation to the complex environment. In contrast, in houses with stable consumption, such as House 13 and House 17, the improvement in RL efficiency tends to be stagnant or insignificant, indicating the limited adaptation of RL to environments

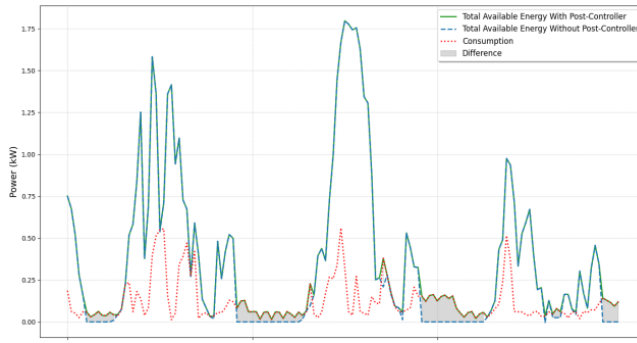

Figure 6. House 17 RL Total Available Energy vs Consumption

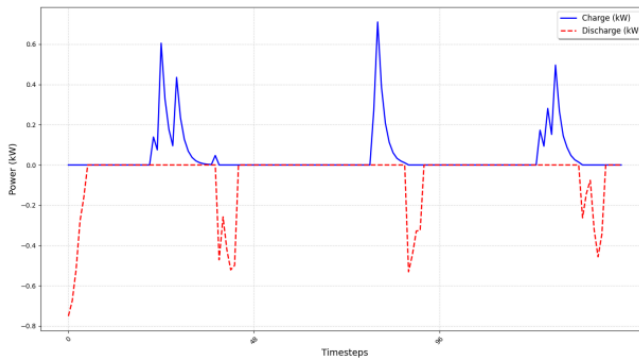

Figure 7. House 17 RL Charging and Discharging Power

with low variability. Computational constraints limit the training to 30 episodes, so the full potential of RL has not been fully explored in this study.

Table 6. Impact of Episode Count on Grid Import and Energy Cost Reduction

Episode Count	Grid Import Reduction (%)				Energy Cost Reduction (%)			
	House				House			
	1	4	13	17	1	4	13	17
10	64.04	-22.23	12.22	18.90	76.43	3.16	19.41	31.94
20	64.55	-19.54	12.49	20.32	76.76	5.29	19.65	33.14
30	64.32	-17.34	11.68	20.21	76.61	7.03	18.91	33.05

The reward trend in Figure 8 shows a gradual increase during training as the number of episodes increases. This trend is consistent with the results in Table 6, which show an increase in the efficiency of the RL model in most scenarios, especially for houses with more complex energy consumption patterns, such as House 4. This finding suggests the potential for the RL model to continue improving with additional training, although there is still variation in performance for houses with more stable characteristics.

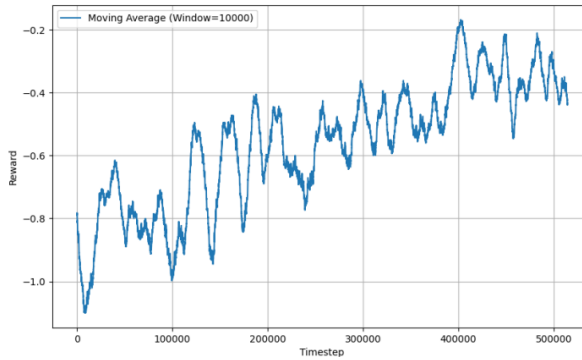

Figure 8. Moving Average Reward Trend of the RL Model (30 Episode Training)

5.2 Discussion

This study develops a general RL model through multi-house training, allowing direct application without retraining. This approach is relevant in the real-world context, especially in developing countries such as Indonesia, which is in the early stages of energy transition. With this model, PV-battery customers without an energy consumption history can use the optimal control system directly. This study found that RL's efficiency is still lower than simple methods such as load following (LF) in specific scenarios. However, RL can surpass LF if further optimization is carried out. Adding training data is more representative of testing data if the goal is to create a more targeted model. Because in real-world applications, training with data similar to the target will be more effective and better. However, because the primary purpose of this study is to investigate the effect of applying the RL model to unfamiliar homes (not yet recognized by the model), this was not done by selecting testing data with a low correlation with training (0.1). In addition, further development of the RL model is also needed, such as improvements and the use of other RL model architectures that need to be explored, as well as improvements to rewards and rules in RL that can further pressure the model to avoid over-discharge (a problem that occurred in this study).

The advantages of RL are still visible in houses with high consumption and large variability (House 4), where RL outperforms LF with its flexibility in adjusting energy strategies based on consumption and tariff dynamics. In addition, rule-based action smoothing helps keep battery charging and discharging smooth, protecting battery health and life. An example can be seen in Figure 9, where even though House 4's consumption variability is high, the battery level remains stable within the limits of 10–90% of maximum capacity. In addition, this study also validates RL's ability to meet loads that are not explicitly stated in previous studies and safety measures when RL fails to meet the load, namely with the Post-Controller.

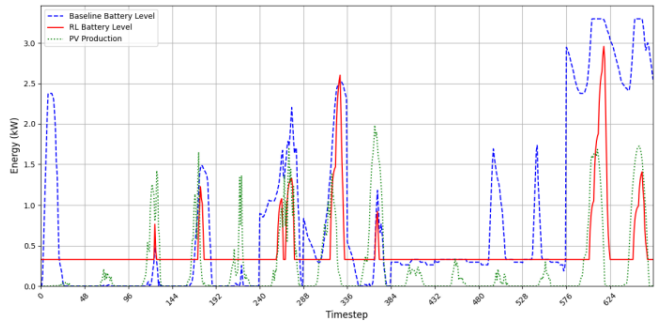

Figure 9. House 4: Battery Levels and PV Production

This study uses energy consumption data focused on the summer season (June to August). This selection is based on the high PV production during the summer and the diversity of household energy consumption patterns that better reflect the real challenges in energy management. However, energy consumption and PV production characteristics in other seasons—such as the rainy season or seasons with lower solar radiation—are significantly different from summer. Integrating datasets from various seasons in future studies will be an important step in evaluating the generalization ability of RL models to more complex seasonal fluctuations.

In addition, local regulations that limit energy management options also influence challenges in optimizing energy. One example is the inability of customers in Indonesia to export surplus energy to the grid because current regulations prohibit the sale of energy from customers to the electricity network [25]. This causes surplus PV energy only to be used for battery charging or to be unused, which directly limits the potential efficiency of the system. In this context, testing RL models on data from various seasons can also provide additional insights into managing surplus energy during periods of low consumption or fluctuating PV production.

6. Conclusion

This study demonstrates the effectiveness of reinforcement learning (RL) agents in improving energy management for houses with diverse consumption profiles. The RL agent achieved cost reductions, particularly in challenging scenarios with high variability. In particular, House 4, characterized by a high average energy consumption of 33.79 kW and high variability (standard deviation: 13.73 kW), reduced energy expenditure by 7.03% while maintaining battery levels within the recommended 10–90% capacity range. In this scenario, RL outperformed the load following (LF) method, showcasing its superior adaptability in dynamic environments where LF's simple fixed rules struggle to respond effectively to fluctuations.

In more stable contexts, such as House 1, the RL agent reduced energy expenses by 76.61%, proving its ability to adapt across changing house profiles. However, in cases like House 13 and House 17, which exhibit stable consumption with minimal variability, the load following (LF) method generally provided comparable or even better performance due to its simplicity and direct approach, which aligns well with

stable energy patterns. This highlights RL's potential while also emphasizing the need for further optimization in low-variability scenarios.

The key contributions of this study are:

- Rule-Based Action Smoothing: This was an important method to ensure stable battery performance, ensured stable battery performance by preventing abrupt variations and protecting battery life. The RL agent was able to charge the batteries gradually during the Off-Peak Load Time (OPLT) and discharge them at controlled rates during Peak Load Time (PLT), as shown in Figure 5 and Figure 9. Strategies like these reduced stress on the system while improving the use of surplus photovoltaic (PV) energy.
- PPO Multi-House Training: The RL agent trained on all houses displayed robust generalization. For stable consumption houses, such as House 17 and House 13, there was a persistent reduction of 33.05% and 18.91% of the energy cost, respectively. However, these results highlight that while RL adapts well across diverse profiles, it requires further refinement to fully optimize performance in stable conditions where LF often provides competitive results with less computational complexity.
- Post-Controller Integration: The addition of a post-controller introduced a holistic solution to address real-time energy shortfalls experienced by the RL agent, as depicted in Figure 6. This feature mitigates failures in the RL model by supplementing grid imports during critical moments, further stabilizing the system.

This study highlights the potential of reinforcement learning (RL) while identifying areas for improvement, particularly in stable energy consumption scenarios where load following (LF) still outperforms RL. Future work will focus on refining the RL architecture and reward functions to enhance performance in low-variability settings, making RL more competitive against LF. To further improve generalization, future studies will incorporate datasets from different seasons to investigate the effects of seasonal variations on RL performance. This approach will evaluate the model's ability to adapt to annual fluctuations in PV production and consumption profiles. Furthermore, future work will explore grid export scenarios, where surplus energy can be sold back to the grid. This addition could significantly improve the economic viability of RL-based systems, particularly in regulatory environments that allow for dynamic grid interactions.

References

[1] Intergovernmental Panel on Climate Change (IPCC). *Global Warming of 1.5°C: IPCC Special Report on Impacts of Global Warming of 1.5°C above Pre-industrial Levels in Context of Strengthening Response to Climate Change, Sustainable Development, and Efforts to Eradicate Poverty*. Cambridge: Cambridge University Press, 2022. DOI: 10.1017/9781009157940.

[2] Indonesia. *Indonesia Long-Term Strategy for Low Carbon and Climate Resilience 2050 (Indonesia LTS-LCCR 2050)*. Unpublished report. 2020.

[3] F. Echevarría Camarero et al. "Profitability of Batteries in Photovoltaic Systems for Small Industrial Consumers in Spain under Current Regulatory Framework and Energy Prices". In: *Energies* 16.1 (Jan. 2023). DOI: 10.3390/en16010361.

[4] BloombergNEF. *Battery Pack Prices Fall to an Average of $132/kWh, But Rising Commodity Prices Start to Bite*. Accessed: Dec. 09, 2024. 2024. URL: https://about.bnef.com/blog/battery-pack-prices-fall-to-an-average-of-132-kwh-butrising-commodity-prices-start-to-bite.

[5] J. Quer and E. Ribera Borrell. "Connecting stochastic optimal control and reinforcement learning". In: *Journal of Mathematical Physics* 65.8 (Aug. 2024). DOI: 10.1063/5.0140665.

[6] A. A. Elshazly et al. "Reinforcement Learning for Fair and Efficient Charging Coordination for Smart Grid". In: *Energies* 17.18 (Sept. 2024). DOI: 10.3390/en17184557.

[7] B. Xiong et al. "Optimizing electricity demand scheduling in microgrids using deep reinforcement learning for cost-efficiency". In: *IET Generation, Transmission and Distribution* 17.11 (June 2023), pp. 2535–2544. DOI: 10.1049/gtd2.12866.

[8] S. Xia et al. "A multi-task deep reinforcement learning-based recommender system for co-optimizing energy, comfort, and air quality in commercial buildings with humans-in-the-loop". In: *Data-Centric Engineering* 5 (Nov. 2024). DOI: 10.1017/dce.2024.27.

[9] N. Ali et al. "A reinforcement learning approach to dairy farm battery management using Q learning". In: *Journal of Energy Storage* 93 (July 2024). DOI: 10.1016/j.est.2024.112031.

[10] O. Almughram, S. Abdullah ben Slama, and B. A. Zafar. "A Reinforcement Learning Approach for Integrating an Intelligent Home Energy Management System with a Vehicle-to-Home Unit". In: *Applied Sciences* 13.9 (May 2023). DOI: 10.3390/app13095539.

[11] A. C. Real et al. "Optimization of a photovoltaic-battery system using deep reinforcement learning and load forecasting". In: *Energy and AI* 16 (May 2024). DOI: 10.1016/j.egyai.2024.100347.

[12] H. Kang et al. "Reinforcement learning-based optimal scheduling model of battery energy storage system at the building level". In: *Renewable and Sustainable Energy Reviews* 190 (Feb. 2024). DOI: 10.1016/j.rser.2023.114054.

[13] F. Hartel and T. Bocklisch. "Minimizing Energy Cost in PV Battery Storage Systems Using Reinforcement Learning". In: *IEEE Access* 11 (2023), pp. 39855–39865. DOI: 10.1109/ACCESS.2023.3267978.

[14] B. Qi, M. Rashedi, and O. Ardakanian. "EnergyBoost: Learning-based control of home batteries". In: *Proceedings of the 10th ACM International Conference on Future Energy Systems (e-Energy)*. Association for Computing Machinery, 2019, pp. 239–250. DOI: 10.1145/3307772.3328279.

[15] R. S. Sutton and A. G. Barto. *Reinforcement Learning: An Introduction*. The MIT Press, 2020.

[16] A. A. Elshazly et al. "Reinforcement Learning for Fair and Efficient Charging Coordination for Smart Grid". In: *Energies (Basel)* 17.18 (Sept. 2024). DOI: 10.3390/en17184557.

[17] V. Mnih et al. "Human-level control through deep reinforcement learning". In: *Nature* 518.7540 (2015), pp. 529–533. DOI: 10.1038/nature14236.

[18] J. Schulman et al. *Proximal Policy Optimization Algorithms*. http://arxiv.org/abs/1707.06347. 2017.

[19] T. P. Lillicrap et al. *Continuous control with deep reinforcement learning*. http://arxiv.org/abs/1509.02971. 2015.

[20] R. Trivedi et al. "Comprehensive Dataset on Electrical Load Profiles for Energy Community in Ireland". In: *Scientific Data* 11.1 (Dec. 2024). DOI: 10.1038/s41597-024-03454-2.

[21] D. I. Jurj et al. "Custom outlier detection for electrical energy consumption data applied in case of demand response in block of buildings". In: *Sensors* 21.9 (May 2021). DOI: 10.3390/s21092946.

[22] Kementerian Energi dan Sumber Daya Mineral. *Tarif Tenaga Listrik Triwulan I Tahun 2021*. https://www.esdm.go.id/assets/media/content/content-tarif-tenaga-listrik-tw-i-2021.pdf. 2021.

[23] G. Brockman et al. *OpenAI Gym*. http://arxiv.org/abs/1606.01540. 2016.

[24] G. Bao and R. Xu. "A Data-Driven Energy Management Strategy Based on Deep Reinforcement Learning for Microgrid Systems". In: *Cognitive Computation* 15.2 (Mar. 2023), pp. 739–750. DOI: 10.1007/s12559-022-10106-3.

[25] Kementerian Energi dan Sumber Daya Mineral Republik Indonesia. *Peraturan Menteri Energi dan Sumber Daya Mineral Republik Indonesia Nomor 2 Tahun 2024 tentang Pembangkit Listrik Tenaga Surya Atap*. Jan. 2024. URL: https://jdih.esdm.go.id/common/dokumen-external/Permen%20ESDM%20Nomor%202%20Tahun%202024.pdf.