DIGITAL TRANSFORMATION AND ADMINISTRATION INNOVATION

Citation: Asadi Chaharrahe Gashin, S., Malek Hosseini, R., Bagherifard, K., Yaghoobian, S. H., & Ranjbar, H. (2026). Multi-Objective Safe Energy Management in Smart Buildings Using LSTM-Based Load Forecasting and Soft Actor-Critic Reinforcement Learning. *Digital Transformation and Administration Innovation*, 4(1), 1-23.

Received date: 2025-09-08 Revised date: 2025-12-17 Accepted date: 2025-12-20 Published date: 2026-01-01

Multi-Objective Safe Energy Management in Smart Buildings Using LSTM-Based Load Forecasting and Soft Actor-Critic Reinforcement Learning

Saber Asadi Chaharrahe Gashin[1], Razieh Malek Hosseini[2]*, Karamollah Bagherifard [3], Seyed Hadi Yaghoobian[2], Hamideh Ranjbar[2]

1. Department of Computer Engineering, Yas.C., Islamic Azad University, Yasuj, Iran
2. Assistant Professor, Department of Computer Engineering, Yas.C., Islamic Azad University, Yasuj, Iran
3. Associate Professor, Department of Computer Engineering, Yas.C., Islamic Azad University, Yasuj, Iran

*Correspondence: malekhoseini.r@iau.ac.ir

Abstract

This study aimed to develop and evaluate an integrated LSTM–SAC–Safety Layer framework for predictive, multi-objective, and constraint-aware energy management in smart buildings, with simultaneous consideration of energy cost, peak demand, renewable-energy utilization, thermal comfort, and operational safety. A quantitative computational and experimental design was employed using real-world residential energy-consumption data from the Pecan Street Dataport, comprising 1,185 household identifiers and measurements recorded at 15-minute intervals throughout 2023. A stacked Long Short-Term Memory (LSTM) network was developed for short-term load forecasting, and its predictions were incorporated into the state space of a continuous-action Soft Actor–Critic (SAC) reinforcement-learning agent. A Safety Layer projected generated actions onto the feasible operating region before execution. Performance was evaluated over an independent 55-day test period under four scenarios: conventional rule-based control, standard SAC, LSTM–SAC without the Safety Layer, and the complete LSTM–SAC–Safety framework. Daily energy costs were additionally subjected to paired-samples t tests and Wilcoxon signed-rank tests. Inferential analysis demonstrated that the proposed LSTM–SAC–Safety framework produced statistically significant improvements in daily energy-cost performance compared with the competing reinforcement-learning configurations. The comparison between the proposed framework and LSTM–SAC yielded $t = 4.82$, $p < .001$, while the proposed framework also differed significantly from standard SAC ($p < .001$). The Wilcoxon signed-rank analysis corroborated the parametric findings, indicating that the observed superiority was not attributable to random variation across the 55 test days and remained statistically significant at the 99% confidence level. Integrating short-term LSTM forecasting with continuous SAC decision-making and an explicit Safety Layer provides a statistically supported and operationally reliable approach to smart-building energy management, enabling predictive control while maintaining physical feasibility and balancing economic, environmental, and occupant-comfort objectives.

Keywords: Home Energy Management System (HEMS); Deep Reinforcement Learning; Soft Actor–Critic (SAC); Long Short-Term Memory (LSTM); Safety Layer; Load Forecasting; Smart Buildings; Energy Optimization.

Copyright: © 2026 by the authors. Published under the terms and conditions of Creative Commons Attribution-NonCommercial 4.0 International (CC BY-NC 4.0) License.

1. Introduction

The residential and building sectors have become increasingly important targets for energy-efficiency improvement because of their substantial contribution to electricity demand, their growing dependence on electrically driven heating and cooling systems, and the rapid penetration of distributed energy resources, battery storage, electric vehicles, and smart appliances. In conventional residential environments, electricity consumption is often managed reactively, with limited consideration of short-term demand fluctuations, time-varying tariffs, renewable generation, equipment constraints, or occupant comfort. The transition toward smart buildings has therefore created a need for home energy management systems (HEMS) capable of continuously monitoring energy flows, predicting near-future conditions, coordinating controllable devices, and making operational decisions that balance economic, technical, environmental, and comfort-related objectives. Recent developments in the Internet of Things (IoT), advanced sensing, smart metering, artificial intelligence, and distributed control have accelerated this transition by providing high-resolution information regarding household demand, environmental conditions, equipment states, and renewable-energy availability (Mazhar et al., 2026). At the same time, the increasing complexity of residential energy ecosystems has made conventional rule-based scheduling progressively less adequate, particularly when systems must simultaneously coordinate photovoltaic generation, batteries, electric vehicles, heating, ventilation, and air-conditioning loads under uncertain operating conditions.

One of the earliest requirements for intelligent energy management is reliable characterization and prediction of electricity demand. Residential load profiles contain pronounced temporal dependencies, nonlinear variations, occupancy-driven changes, and short-duration peaks that are difficult to represent using static or purely deterministic models. Deep recurrent neural networks have demonstrated considerable potential for learning these temporal structures and improving electricity-consumption forecasting in both residential and commercial buildings (Rahman et al., 2018). Among recurrent architectures, Long Short-Term Memory (LSTM) networks are especially suitable because their gated memory mechanism enables them to retain relevant information over extended sequences while reducing problems associated with vanishing gradients. Accurate short-term load forecasting can provide an energy management controller with anticipatory information regarding expected demand, allowing decisions concerning battery charging, load shifting, thermal control, and grid interaction to be made before critical events occur rather than after them. This capability is important because actions taken at one point in time can materially influence energy costs, battery state of charge, peak grid demand, and occupant comfort during subsequent periods.

Closely related to load forecasting is the problem of obtaining detailed information about appliance-level consumption. Non-intrusive load monitoring (NILM) has emerged as an important approach for disaggregating aggregate household electricity measurements into appliance-specific patterns without requiring dedicated sensors for every device. Deep neural approaches to NILM demonstrated that useful appliance information can be extracted directly from whole-house electricity signals (Kelly & Knottenbelt, 2015). More recently, the integration of NILM with deep reinforcement learning has been proposed as a means of enabling HEMS to use richer information about device behavior when scheduling flexible loads and coordinating distributed energy resources (Katoch et al., 2025). Such developments indicate that future HEMS architectures increasingly depend on data-driven representation of both current and prospective energy states rather than on predefined operating schedules alone.

Although prediction improves situational awareness, forecasting by itself does not determine how a building should respond to future conditions. Energy management is fundamentally a sequential decision-making problem because the consequences of present actions extend over time. Reinforcement learning (RL) provides a natural framework for this setting by enabling an agent to learn control policies through repeated interaction with an environment. Early applications of deep reinforcement learning to smart-home energy management demonstrated that an agent can learn to coordinate energy-consuming devices and storage systems without relying on a fully specified analytical model of household behavior (Yu et al., 2019). At larger scales, entropy-based collective multi-agent reinforcement learning has also been investigated for coordinating multiple homes and distributed resources while addressing scalability and interactions among agents (Yang et al., 2019). These studies established the potential of learning-based control to move residential energy management beyond fixed heuristics and toward adaptive policies that respond to changing system states.

The effectiveness of RL in HEMS has subsequently been demonstrated across a range of objectives. Deep reinforcement learning has been applied to home energy management control to reduce energy use and enhance the utilization of photovoltaic

Copyright: © 2027 by the authors. Published under the terms and conditions of Creative Commons Attribution-NonCommercial 4.0 International (CC BY-NC 4.0) License.

generation while maintaining thermal comfort (Lissa et al., 2021). Real-time DRL strategies have further shown the feasibility of coordinating smart-home resources under changing prices, renewable generation, and operational conditions (Wei et al., 2023). Online HEMS based on Q-learning and Deep Q-learning have similarly illustrated that learning agents can adapt their decisions through continuous interaction with residential energy environments (İzmitligil & Karamancıoğlu, 2024). Deep reinforcement learning has also been used specifically in the context of time-of-use electricity prices and real-time energy-storage control, reinforcing the importance of temporally adaptive strategies when retail tariffs and storage states vary throughout the day (Xiong, 2024). Together, these developments support the use of RL as a flexible control paradigm, while also demonstrating that the design of the state space, action space, and reward function strongly determines the resulting policy.

A major development in recent HEMS research has been the integration of forecasting models with reinforcement learning. When an RL agent observes only the current operating state, its decisions may remain predominantly reactive and may fail to anticipate upcoming demand peaks, photovoltaic production changes, or costly tariff periods. Incorporating predicted information can transform the control problem from reactive optimization into anticipatory management. Real et al. integrated load forecasting with deep reinforcement learning for optimizing photovoltaic-battery systems, demonstrating the relevance of forecast-informed control for coordinating renewable generation and energy storage (Real et al., 2024). Transformer-based Soft Actor-Critic learning has likewise been investigated for HEMS to capture temporal relationships more effectively within continuous control tasks (Ludolfinger, Zinsmeister, et al., 2023). Recurrent Soft Actor-Critic has also been examined for demand-response problems, where temporal memory is particularly valuable under partial observability and time-dependent system behavior (Ludolfinger, Perić, et al., 2023). These approaches suggest that embedding future-oriented information into the decision process can materially improve the ability of RL agents to manage intertemporal trade-offs.

Among continuous-control RL algorithms, Soft Actor-Critic (SAC) is especially attractive for building energy management because many control actions, such as battery charging power, HVAC setpoints, or grid exchange levels, are naturally continuous rather than discrete. SAC uses an actor-critic architecture and augments the expected return with an entropy term, encouraging sufficient exploration while improving policy robustness. However, effective application of SAC to HEMS requires careful state representation and reward construction. Multi-step prioritized deep reinforcement learning has been proposed to improve learning efficiency by incorporating information from multiple future transitions and prioritizing more informative experiences (X. Wu et al., 2024). Data-driven DRL frameworks have also been developed to account for uncertain household parameters, emphasizing that residential systems must operate under considerable variability in user behavior, equipment characteristics, and environmental conditions (Ren et al., 2024). These findings underscore the importance of using learning architectures that can accommodate uncertainty while maintaining stable decision-making.

A second major challenge concerns personalization and heterogeneity among households. Consumption patterns, thermal preferences, appliance ownership, occupancy schedules, and available distributed-energy resources vary considerably among users. A single universal control policy may therefore perform suboptimally when applied across heterogeneous households. Personalized federated deep reinforcement learning has been proposed to combine shared learning across multiple residences with individualized adaptation while limiting the need to centrally pool raw household data (Gao et al., 2023). More recently, adaptive HEMS frameworks augmented by large language models have been developed to incorporate dynamically expressed user preferences into reinforcement-learning policies, thereby extending personalization beyond fixed numerical preferences toward more flexible user-system interaction (Du et al., 2026). These developments indicate that intelligent HEMS are evolving from generic optimization systems toward adaptive decision-support platforms that consider household-specific operating conditions and user preferences.

Interpretability is another emerging requirement. Deep RL controllers may provide effective actions, yet their internal decision processes are often difficult for users, operators, or engineers to understand. This limitation can reduce trust in automated energy-management systems, particularly when algorithmic decisions directly influence user comfort, energy expenditures, or equipment operation. Explainable reinforcement-learning HEMS based on differentiable decision trees have been explored as a means of making learned policies more transparent while retaining strong control performance (Gokhale et al., 2024). Human activity recognition has also been combined with deep reinforcement learning so that smart-home energy

Copyright: © 2027 by the authors. Published under the terms and conditions of Creative Commons Attribution-NonCommercial 4.0 International (CC BY-NC 4.0) License.

management can respond more explicitly to occupant behavior and activity states (Z. Wu et al., 2024). These approaches highlight the broader transition from optimization alone toward human-centered and interpretable energy-management systems in which control quality must be accompanied by greater contextual awareness.

The increasing complexity of smart buildings also creates important reliability and safety requirements. An RL agent may discover actions that appear advantageous in terms of its reward function but violate physical limits of batteries, converters, HVAC systems, electric-vehicle chargers, or grid-interconnection equipment. Consequently, high reward values do not necessarily imply operational feasibility. Safe deployment requires explicit mechanisms for enforcing limits related to state of charge, charging and discharging rates, power exchange, thermal comfort, voltage, and other system constraints. Reviews of reinforcement-learning approaches for smart-home energy systems with storage emphasize the importance of technical constraints, battery behavior, and suitable control structures when translating learning algorithms into practical HEMS applications (Pinthurat et al., 2024). The broader literature on power-electronic converters also demonstrates the importance of condition monitoring and reliable operation of critical electrical components, including DC-link capacitors used in battery-charger applications (Jafari et al., 2025). From this perspective, an energy-management strategy should not merely optimize cost but must also prevent control actions that could compromise equipment integrity or system reliability.

This concern becomes particularly important when renewable-energy resources and prosumer technologies are integrated into buildings. Photovoltaic generation is inherently variable, batteries introduce intertemporal constraints and degradation considerations, and bidirectional power exchanges make the operating environment more complex. Digital-twin approaches supported by deep learning, big-data analytics, and deep reinforcement learning have been introduced to represent prosumer systems and assess control strategies in a more comprehensive virtual environment (Ali, 2025). Such frameworks point toward increasingly sophisticated HEMS architectures in which prediction, simulation, optimization, and real-time control are closely integrated. The broader convergence of DRL, IoT, distributed renewable resources, and smart-building infrastructure similarly reflects the need for intelligent controllers capable of handling high-dimensional data and rapidly changing operating states (Mazhar et al., 2026).

Despite these advances, several important gaps remain. Many studies optimize energy cost or demand response but do not simultaneously incorporate short-term forecasting, continuous control, explicit physical safety, renewable-energy utilization, and thermal comfort within a single closed-loop architecture. Forecasting modules are sometimes evaluated separately from the control policy, even though a forecast is useful only insofar as it improves subsequent decisions. Similarly, reinforcement-learning policies may be optimized without a dedicated mechanism for preventing infeasible actions. Reward penalties can discourage undesirable behavior, but they do not necessarily guarantee that every action satisfies hard physical constraints. This distinction is particularly important during exploration, when an RL agent may temporarily produce actions outside acceptable operating ranges. A dedicated safety mechanism that projects or modifies raw actions before execution can therefore complement reward-based learning by providing direct operational protection.

Another unresolved issue concerns multi-objective optimization. Residential energy management cannot be evaluated only through financial savings. Excessive peak demand can increase stress on the distribution network; aggressive load shifting may degrade batteries; maximizing photovoltaic self-consumption can conflict with tariff-based operation; and energy-cost minimization can compromise thermal comfort if the reward function is improperly balanced. Accordingly, HEMS design requires a multi-objective reward formulation capable of capturing economic cost, peak demand, local renewable-energy utilization, occupant comfort, constraint compliance, and battery degradation simultaneously. The relative importance of these objectives may vary across applications, making normalization and weighting essential design considerations. A framework that couples forecast-informed states with a continuous SAC controller and an independent Safety Layer offers a logical means of balancing these competing requirements.

Recent work has progressively moved toward this integrated direction. Forecast-informed photovoltaic-battery optimization, uncertainty-aware DRL, human-activity-aware control, personalized federated learning, explainable RL, and adaptive preference modeling each address important aspects of modern HEMS (Du et al., 2026; Gao et al., 2023; Gokhale et al., 2024; Real et al., 2024; Ren et al., 2024; Z. Wu et al., 2024). However, these strands of research also illustrate the fragmentation of current solutions: temporal prediction, continuous control, user adaptation, interpretability, and operational safety are often addressed independently rather than through a unified decision pipeline. The combination of LSTM-based

Copyright: © 2027 by the authors. Published under the terms and conditions of Creative Commons Attribution-NonCommercial 4.0 International (CC BY-NC 4.0) License.

short-term load forecasting with SAC offers a mechanism for incorporating anticipated demand directly into the agent's state representation, while a Safety Layer can enforce feasibility before actions reach the physical HEMS environment. In such a configuration, the predictive model does not artificially increase the reward merely because its forecasts are accurate; instead, forecast information has value only when it guides the SAC agent toward better operational decisions. This design more clearly separates prediction quality, control intelligence, and physical safety.

Accordingly, the aim of this study is to develop and evaluate an integrated LSTM–SAC–Safety Layer framework for multi-objective smart-building energy management that combines short-term load forecasting, continuous reinforcement-learning control, explicit enforcement of physical constraints, and simultaneous optimization of energy cost, peak grid demand, renewable-energy self-consumption, thermal comfort, and operational safety.

2. Methods and Materials

This article presents a framework for modeling and managing energy at the household level based on a Home Energy Management System (HEMS). As an intelligent system, HEMS is responsible for monitoring, coordinating, and managing the energy consumption of different devices within a residential unit. By using information on consumption patterns, technical equipment constraints, and operating conditions, it can determine an optimal energy-consumption schedule. Within this framework, HEMS operation is not necessarily dependent on distributed generation resources such as photovoltaic panels and may focus solely on demand-side management and optimal scheduling of household loads. Accordingly, the following sections first introduce the structure of the system under study and its constituent components; then describe the mathematical model of household loads, operational constraints, and objective function(s); and finally present the proposed optimization method for obtaining an optimal consumption pattern. The proposed architecture consists of six main components: 1. input data and preprocessing; 2. an LSTM-based load-forecasting model; 3. an SAC-based decision-making agent; 4. the Safety Layer and constraint enforcement; 5. the HEMS environment; and 6. a multi-objective reward function.

The overall system process is: data preprocessing → LSTM forecasting → state-space construction → SAC decision-making → Safety Layer → HEMS execution → performance evaluation → reward computation → agent update, as illustrated graphically in Figure 1.



Figure 1. Overall process of the energy management system

At each time step, historical information and the current system state are received and preprocessed. The LSTM model then predicts the future load. The predicted value, together with the current system variables, forms the SAC state vector. Based on this state, the SAC agent generates a raw action. Before implementation, the raw action is examined by the Safety Layer and modified when necessary. The safe action is executed in the HEMS environment, after which the next system state and performance indicators are obtained. Finally, the multi-objective reward is calculated and the resulting transition is stored in the replay buffer. The following subsections describe the different components of the proposed method, including specification and preprocessing of input data, load forecasting with the LSTM network, construction of the SAC state space, formulation of the reinforcement-learning problem, the action space and Actor model, mapping of the Critic networks, definition of the soft objective in SAC, creation of the Safety Layer and physical constraints, formulation of the multi-objective reward function, definition of the HEMS environment, and the training and evaluation algorithm together with the flowchart and pseudocode of the proposed method. In the first step, the input data (Table 1) include current consumption, solar generation, ambient temperature, humidity, electricity price, occupant presence, time of day, day of week, battery state, and other relevant variables, which are specified and preprocessed. When 15-minute data are used, each day contains 96 time samples. Therefore, the 24-hour historical window is defined as follows:

$$L = 24 \times 4 = 96$$

Copyright: © 2027 by the authors. Published under the terms and conditions of Creative Commons Attribution-NonCommercial 4.0 International (CC BY-NC 4.0) License.

Table 1. Baseline input data

Variable	Symbol	Description
Current consumption	P_t^{load}	Building power consumption at time$t$
Solar generation	P_t^{PV}	Power generated by the photovoltaic system
Ambient temperature	T_t^{out}	Outdoor building temperature
Humidity	H_t	Ambient relative humidity
Electricity price	π_t	Energy tariff at time$t$
Occupant presence	O_t	Occupancy status
Time of day	h_t	Temporal position within the day
Day of week	d_t	Daily and weekly consumption pattern
Battery state	SOC_t	Battery state of charge

In the next stage, load forecasting is performed using the LSTM network. The LSTMinput sequence at timetis defined as follows:

$$X_t = [x_{t-L+1}, x_{t-L+2}, \dots, x_t]$$

If the number of input variables is equal to Fthen:

$$X_t \in \mathbb{R}^{L \times F}$$

For a 24-hour window with 15-minute resolution:

$$X_t \in \mathbb{R}^{96 \times F}$$

:The network objective is to estimate the load at the next time step

$$\hat{P}_{t+1}^{load} = f_\theta(X_t)$$

where θ is the parameter set of the LSTM network.

For each time step, the LSTM cell uses the following equations:

$$f_t = \sigma\left(W_f x_t + U_f h_{t-1} + b_f\right)$$

$$i_t = \sigma(W_i x_t + U_i h_{t-1} + b_i)$$

$$\tilde{c}_t = tanh(W_c x_t + U_c h_{t-1} + b_c)$$

$$c_t = f_t \odot c_{t-1} + i_t \odot \tilde{c}_t$$

$$o_t = \sigma(W_o x_t + U_o h_{t-1} + b_o)$$

$$h_t = o_t \odot tanh\,(c_t)$$

In these relations, f_{t}, i_{t}and o_{t} are the forget, input, and output gates, respectively. In addition, c_{t} is the cell state and h_{t}is the hidden state of the network.

The proposed architecture consists of two stacked LSTM layers. The network output is then passed to fully connected layers. The mean squared error loss function is used for network training, and MAE, RMSE, and MAPE are used as forecasting-performance metrics. In calculating MAPE, samples for which the actual value is zero should be excluded or replaced by a small value (Figure 2).

Copyright: © 2027 by the authors. Published under the terms and conditions of Creative Commons Attribution-NonCommercial 4.0 International (CC BY-NC 4.0) License.

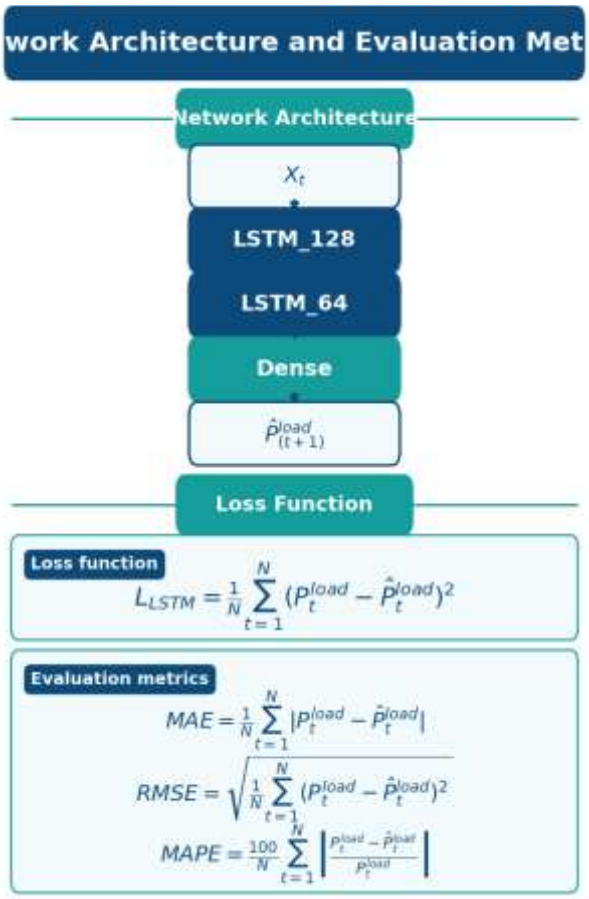

Figure 2. Network architecture and evaluation metrics

The SAC state space is then constructed as follows. In the proposed framework, the state vector includes current variables and forecast information:

$$s_t = \left[P_t^{load}, \hat{P}_{t+1}^{load}, P_t^{PV}, S, O, C_t, \pi_t, E, V_t, O_t T_t h_t d_t\right]^T$$

With multiple forecast steps, the forecast vector can be extended as follows:

$$= \left[P_t^{load}, \hat{P}_{t+1}^{load}, \ldots, \hat{P}_{t+H}^{load}, P_t^{PV}, S, O, C_t, \pi_t, EV_t O_t T_t\right]^T$$

The key point is that the LSTM output does not directly determine the reward. Its influence follows the path below:

$$\hat{P}_{t+1}^{load} \rightarrow s_t \rightarrow a_t \rightarrow r_t$$

Thus, load forecasting is indirectly reflected in the final reward through its effect on the agent's decision.

Reinforcement learning, another major component of the proposed framework, is modeled as a Markov decision process:

$$\mathcal{M} = (\mathcal{S}, \mathcal{A}, P, R, \gamma)$$

where:

$\mathcal{S}$;state space :$\mathcal{A}$;action space :P ;environment transition function :R;reward function :γ.discount factor :

At each time step, the agent observes state s_{t} and selects action a_{t}. After the action passes through the Safety Layer, the final action is applied in the HEMS environment:

$$s_t \rightarrow a_t^{raw} \rightarrow a_t^{safe} \rightarrow (s_{t+1}, r_t)$$

The objective of the agent is to maximize the discounted cumulative return:

$$J(\pi) = \mathbb{E}_\pi\left[\sum_{t=0}^{T-1} \gamma^t r_t\right]$$

Copyright: © 2027 by the authors. Published under the terms and conditions of Creative Commons Attribution-NonCommercial 4.0 International (CC BY-NC 4.0) License.

The action space includes control commands for controllable equipment. Depending on the system structure, these actions may include the following:

$$a_t = \left[P_t^{bat}, P_t^{EV}, u_t^{HVAC}, u_t^{shiftable}\right]^T$$

which respectively denote battery power, electric-vehicle charging power, the HVAC command, and the command for shiftable loads.

In SAC, the Actor models a stochastic policy:

$$a_t \sim \pi_\phi(a_t \mid s_t)$$

First, a variable sampled from a Gaussian distribution is defined:

$$u_t \sim \mathcal{N}\left(\mu_\phi, (s_t)\sigma_\phi^2(s_t)\right)$$

The output is then bounded using the hyperbolic tangent function:

$$a_t = \tanh(u_t)$$

For equipment with different operating ranges, a scaled transformation is applied:

$$a_{t,i} = a_i^{min} + \frac{\tanh(u_{t,i}) + 1}{2}\left(a_i^{max} - a_i^{min}\right)$$

To reduce overestimation error, SAC uses two independent Critic networks:

$$Q_{\psi_1}(s_t, a_t)$$

$$Q_{\psi_2}(s_t, a_t)$$

The input to the Critic network is formed by concatenating the state and action:

$$z_t = [s_t, a_t]$$

The proposed Critic architecture consists of three fully connected layers with 256, 256, and 128 neurons:

$$h_1 = ReLU(W_1 z_t + b_1)$$

$$h_2 = ReLU(W_2 h_1 + b_2)$$

$$h_3 = ReLU(W_3 h_2 + b_3)$$

$$Q_{\psi_i}(s_t, a_t) = W_4 h_3 + b_4$$

In addition to the reward, SAC incorporates policy entropy into the objective function:

$$J(\pi) = \mathbb{E}\left[\sum_t \gamma^t \left(r_t + \alpha\mathcal{H}\left(\pi(\cdot \mid s_t)\right)\right)\right]$$

Policy entropy is defined as:

$$\mathcal{H}\left(\pi(\cdot \mid s)\right) = -\mathbb{E}_{a\sim\pi}[\log \pi\,(a \mid s)]$$

In this relation, αis the temperature coefficient or entropy weight. Increasing entropy promotes exploration and prevents premature convergence of the policy.

The TD target for the Critic network is defined as follows:

$$y_t = r_t + \gamma\left[\min_{j=1,2} Q_{\bar{\psi}_j}(s_{t+1}, a_{t+1}) - \alpha \log \pi_\phi\,(a_{t+1} \mid s_{t+1})\right]$$

where:

Copyright: © 2027 by the authors. Published under the terms and conditions of Creative Commons Attribution-NonCommercial 4.0 International (CC BY-NC 4.0) License.

$$a_{t+1} \sim \pi_\phi(\cdot \mid s_{t+1})$$

The Critic loss function is:

$$J_Q(\psi_j) = \mathbb{E}\left[\frac{1}{2}\left(Q_{\psi_j}(s_t, a_t) - y_t\right)^2\right]$$

The Actor objective function is written as:

$$J_\pi(\phi) = \mathbb{E}_{s_t, a_t}\left[\alpha \log \pi_\phi\left(a_t \mid s_t\right) - \min_{j=1,2} Q_{\psi_j}(s_t, a_t)\right]$$

The raw action generated by the Actor may be incompatible with the actual equipment state. Therefore, in the proposed framework, the raw action is projected onto the set of feasible actions before execution:

$$a_t^{safe} = P_{\mathcal{A}(s_t)}(a_t^{raw})$$

A general form of this projection is:

$$a_t^{safe} = \arg\min_a \frac{1}{2}\|a - a_t^{raw}\|_2^2$$

subject to:

$$a \in \mathcal{A}(s_t)$$

In this case, the Safety Layer selects the feasible action closest to the action generated by the Actor.

The main constraints in the proposed framework are presented in Table 2.

Table 2. Main constraints in the proposed framework

Title		Relation
Battery state-of-charge constraint		$SOC^{min} \le SOC_t \le SOC^{max}$
Battery power limit		$-P_{dis}^{max} \le P_t^{bat} \le P_{ch}^{max}$
Battery dynamics	Charging	$SOC_{t+1} = SOC_t + \frac{\eta_{ch} P_t^{ch} \Delta t}{E_{bat}}$
	Discharging	$SOC_{t+1} = SOC_t - \frac{P_t^{dis} \Delta t}{\eta_{dis} E_{bat}}$
Power balance		$P_t^{grid} + P_t^{PV} + P_t^{dis} = P_t^{load} + P_t^{ch} + P_t^{EV}$
Thermal comfort constraint		$T^{min} \le T_t^{in} \le T^{max}$

The reward function must combine the conflicting objectives of energy management within a unified structure. The proposed function is defined as follows:

$$r_t = -\lambda_c C_t - \lambda_p P_t^{peak} + \lambda_r R_t^{RE} - \lambda_{comf} D_t^{comfort} - \lambda_v V_t - \lambda_b D_t^{battery}$$

In this relation, C_{t}: energy cost;P_{t}^{peak}: peak-demand index;R_{t}^{RE}: renewable-energy utilization;D_{t}^{comfort}: deviation from the comfort range;V_{t}: magnitude of constraint violation;D_{t}^{battery}: battery-degradation index; and λ: objective weighting coefficients.

Table 3. Corresponding relationships in the proposed framework

Title	Relation	Description
Energy cost	$C_t = \pi_t E_t^{grid}$	π_t Electricity price andE_t^{grid} energy received from the grid
Total energy cost over the time horizon	$C_{total} = \sum_{t=1}^{T} \pi_t E_t^{grid}$	

Copyright: © 2027 by the authors. Published under the terms and conditions of Creative Commons Attribution-NonCommercial 4.0 International (CC BY-NC 4.0) License.

Peak load	$P^{peak} = \max_t P_t^{grid}$	
Peak penalty	$R_t^{peak} = -\lambda_p P^{peak}$	
Renewable-energy utilization	$E_t^{RE,use} = \min(P_t^{PV}, P_t^{load})\Delta t$	$R_t^{RE} = \lambda_r E_t^{RE,use}$
User comfort	$D_t^{comfort} = \max(0, T_t - T^{max}) + \max(0, T^{min} - T_t)$	
Constraint violation (the Safety Layer is expected to substantially reduce this (component	$V_t = \sum_{k=1}^{K} v_{k,t}$	

The HEMS environment models the relationship between the control action and the physical state of the system. This environment may include a battery, a PV system, the power grid, HVAC, controllable loads, and measurement equipment.

At each time step, the environment follows the sequence shown in Figure 3: receiving the safe action; applying the action to the equipment; computing consumption, generation, and power exchange; updating the battery and equipment states; computing the next state; extracting performance indicators; and calculating the reward.

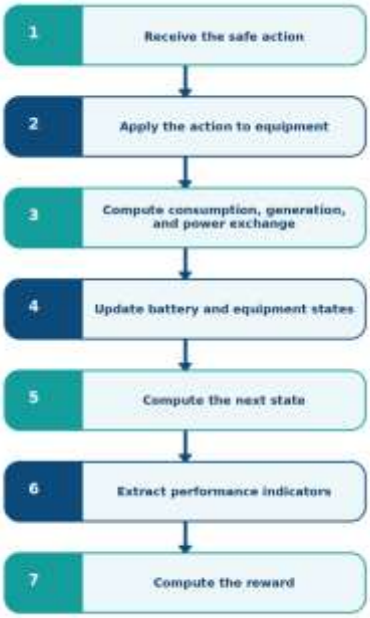

Figure 3. Sequential process of the HEMS environment

The environment transition function can be expressed generally as:

$$(s_{t+1}, r_t) = \mathcal{E}(s_t, a_t^{safe}, \xi_t)$$

whereξ_t.may represent uncertainty arising from user behavior, weather conditions, and forecasting error

In the proposed method, the LSTM does not directly participate in calculating the reward function. Instead, by forecasting the short-term load, it provides forward-looking information to the SAC agent. The LSTM output is used to construct the state space, and SAC selects the appropriate action based on this state. After application of the Safety Layer and execution of the action in the HEMS environment, the consequences of the decision are evaluated in terms of energy cost, peak demand, renewable-energy utilization, user comfort, and technical constraints, and the multi-objective reward is calculated. Therefore, the relationship between the LSTM and the reward function is indirect and follows the path: load forecasting → state construction → SAC decision → action execution → performance evaluation → reward computation.

The training procedure forms a completeLSTM–SAC :loop with the following general mapping

$$X_{t-L+1:t}\hat{P}_{t+1}, h_t$$

Then:

Copyright: © 2027 by the authors. Published under the terms and conditions of Creative Commons Attribution-NonCommercial 4.0 International (CC BY-NC 4.0) License.

$$\left[\hat{P}_{t+1}, h_t, P_t, P, V_t, S, OC_t Price_t \ldots\right] \rightarrow SAC$$

and:

$$SAC \rightarrow a_t^{raw} \rightarrow Safety\ Layer \rightarrow a_t \rightarrow HEMS \rightarrow R_t, s_{t+1}$$

and during the learning stage:

$$(s_t, a_t, r_t, s_{t+1}) \rightarrow Actor/Critic\ Update$$

This cycle is repeated throughout the training episodes. The overall agent-training structure is consistent with the sequence "state observation → forecasting → action selection → safety control → action execution → reward reception → next state → network update." The complete LSTM–SAC loop should be defined so that the load forecast produced by LSTM directly enters SAC decision-making, after which the outcome of the decision is returned to the agent through the HEMS environment as a multi-objective reward. Figure 4 presents the overall algorithm of the proposed method.

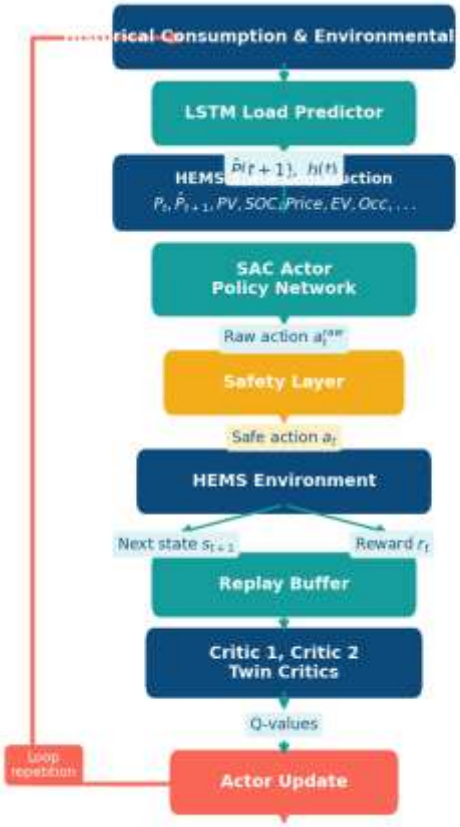

Figure 4. Comprehensive closed-loop flowchart of intelligent building energy management (HEMS): integration of LSTM-based load forecasting, continuous-policy control with SAC, enforcement of physical constraints through the Safety Layer, storage of experiences in the replay buffer, and gradient-based updating

The LSTM–SAC–Safety training process, shown in Figure 5, receives historical consumption data, environmental conditions, and equipment states as inputs and intelligently produces an optimal energy-management policy as its output.

Copyright: © 2027 by the authors. Published under the terms and conditions of Creative Commons Attribution-NonCommercial 4.0 International (CC BY-NC 4.0) License.

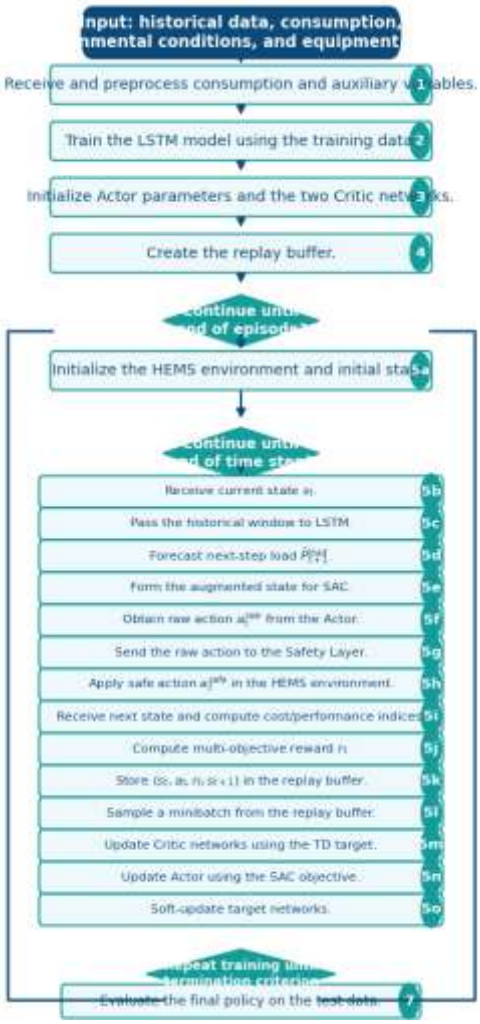

Figure 5. Details of the LSTM–SAC–Safety training process

3. Findings and Results

To implement and evaluate the proposed method, the Pecan Street Dataport dataset was used. This dataset contains real residential energy-consumption data collected through a smart-energy project in Austin, Texas, and includes 1,185 household identifiers. The characteristics of the dataset and the experimental configuration are presented in Table 4.

Table 4. Dataset Description

Item	Description
Data Recording Period	January 1, 2023 to December 31, 2023 (one complete calendar year, 365 days).
Time Resolution	15 minutes (96 time steps per day).
Total Number of Samples	35,040 time steps (365 × 96).
Data Splitting	
Training Data (Train):	70% (24,528 samples - days 1 to 255)
Validation Data (Validation)	15% (5,280 samples - days 256 to 310)
Final Test Data (Test)	15% (5,232 samples - days 311 to 365)

Table 5 presents the equipment specifications, tariff structure, and physical constraints of the HEMS.

Table 5. Equipment Specifications, Tariff Structure, and Physical Constraints of the HEMS

Parameter	Numerical Value	Unit	Description
Nominal Battery Capacity (E_{bat})	10.0	kWh	Total residential energy-storage capacity (Li-ion)

Copyright: © 2027 by the authors. Published under the terms and conditions of Creative Commons Attribution-NonCommercial 4.0 International (CC BY-NC 4.0) License.

Permissible State-of-Charge Range ($[\boldsymbol{S}, \boldsymbol{OC^{min}SOC^{max}}]$)	$[0.20, 0.90]$	-	Prevention of deep discharge and battery overcharging
Maximum Charging and Discharging Power ($\boldsymbol{P_{ch}^{max}}, \boldsymbol{P_{dis}^{max}}$)	3.5	kW	Battery inverter power limit
Charging and Discharging Efficiency ($\boldsymbol{\eta_{ch}}, \boldsymbol{\eta_{dis}}$)	0.95	-	Round-trip battery power-conversion efficiency
Nominal PV Capacity ($\boldsymbol{P_{PV}^{peak}}$)	5.0	kWp	Rooftop photovoltaic system
Electric-Vehicle Battery Capacity ($\boldsymbol{E_{EV}}$)	24.0	kWh	EV battery pack with maximum charging power 3.7 kW
Grid-Exchange Limit ($\boldsymbol{P_{grid}^{max}}$)	8.0	kW	Maximum permitted household-meter power
Occupant Thermal-Comfort Range ($[\boldsymbol{T^{min}}, \boldsymbol{T^{max}}]$)	$[21.0, 25.0]$	°C	Indoor HVAC comfort setpoint
Time-of-Use Electricity Tariff (ToU Pricing)	Off-peak: 0.08; Mid-peak: 0.15; Peak: 0.32	kWh	23:00-07:00 (off-peak); 07:00-17:00 (mid-peak); 17:00-23:00 (peak)

The LSTM forecasting model and the SAC agent used in this implementation are specified in Table 6.

Table 6. Specifications of the LSTM Forecasting Model and SAC Agent

Section	Parameter	Value
LSTM Structure		
	Input Window Length ($\boldsymbol{L}$)	96 steps (previous 24 hours)
	Layer Dimensions	Stacked: first layer 128 and second layer 64 cells
	Learning Rate	1×10^{-3} (Adam optimizer)
	Batch Size / Number of Epochs	Batch Size: 64, Epochs: 100 (with Early Stopping)
SAC Structure		
	Network Architecture: Actor and Critic	3 dense layers with dimensions $[256, 256, 128]$ neurons (ReLU activation)
	Network Learning Rates	Actor: 3×10^{-4}, Critic: 3×10^{-4}
	Discount Factor ($\boldsymbol{\gamma}$)	0.99
	Soft-Update Coefficient ($\boldsymbol{\tau}$)	0.005
	Replay Buffer Capacity	1,000,000 transitions
	Entropy Temperature Coefficient ($\boldsymbol{\alpha}$)	Adaptive (automatically tuned toward target $\mathcal{H}_{target} = -\dim(\mathcal{A})$)
	Multi-Objective Reward-Function Weights ($\boldsymbol{\lambda_i}$)	
	Energy-Cost Weight ($\boldsymbol{\lambda_c}$)	1.00
	Grid Peak-Shaving Weight ($\boldsymbol{\lambda_p}$)	0.45
	Local Renewable-Energy Utilization Weight ($\boldsymbol{\lambda_r}$)	0.30
	Thermal-Comfort Weight ($\boldsymbol{\lambda_{comf}}$)	0.50
	Constraint-Violation Penalty Weight during Training ($\boldsymbol{\lambda_v}$):	2.50
	Battery-Degradation Weight ($\boldsymbol{\lambda_b}$):	0.15

The performance of the energy management system is evaluated using total energy cost, peak grid power, cost reduction relative to the baseline method, peak reduction, local utilization of PV energy, the number of technical-constraint violations, maintenance of user comfort, the number and depth of battery cycles, agent-training stability, and decision-making computation time. To compare the proposed method with the baseline method, cost reduction is calculated according to the corresponding equation, and peak reduction is likewise calculated using its corresponding equation.

$$Reduction_{cost}(\%) = \frac{C_{baseline} - C_{method}}{C_{baseline}} \times 100$$

$$Reduction_{peak}(\%) = \frac{P_{baseline}^{peak} - P_{method}^{peak}}{P_{baseline}^{peak}} \times 100$$

Copyright: © 2027 by the authors. Published under the terms and conditions of Creative Commons Attribution-NonCommercial 4.0 International (CC BY-NC 4.0) License.

To evaluate the proposed model, defined as the complete configuration combining SAC, a multi-objective reward function, and a Safety Layer, four scenarios were implemented as shown in Table 7 and their results were evaluated. These scenarios were designed to examine the effect of forecast information on decision-making; the effect of the SAC algorithm relative to the baseline method; the effect of the multi-objective reward function on balancing competing objectives; the effect of the Safety Layer on reducing constraint violations; and the extent to which final performance depends on each component.

Table 7. Implementation Scenarios for the Proposed Method

Scenario	Description
Baseline	Baseline method without the complete proposed architecture
SAC	SAC without LSTM forecasting
LSTM–SAC	LSTM forecasting included in the SAC state space
Complete Model	LSTM + SAC + multi-objective reward function + Safety Layer

To independently evaluate the Safety Layer, the constraint-violation rate before and after application of the layer is also compared.

$$Violation\ Reduction(\%) = \frac{V_{raw} - V_{safe}}{V_{raw}} \times 100$$

After implementation of the proposed model according to the algorithm, the accuracy of the LSTM load-forecasting model and the performance of the LSTM model on the independent test data (the final 55 days of the year) were calculated, and the results are presented below.

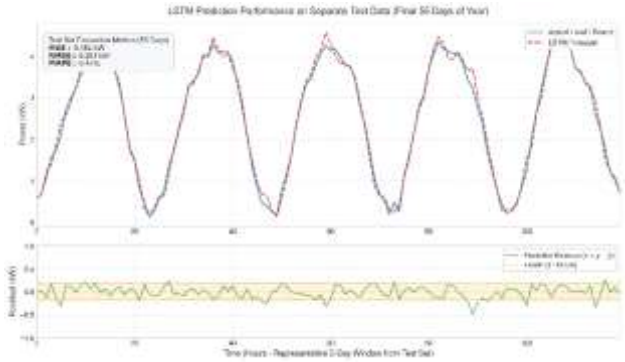

Figure 6. Validation and Performance Evaluation of the LSTM Forecasting Model

The high degree of overlap between the LSTM prediction curve and the actual values indicates the network's strong ability to extract recurring daily patterns and nonlinear behavior. The recorded mean percentage error MAPE = 6.42% indicates excellent accuracy for building-energy engineering and load-management applications; an error below 10% in load/generation forecasting is regarded as highly satisfactory. The MAE = 0.184 kW and RMSE = 0.261 kW values show that prediction error is very small at most time steps. The closeness of RMSE to MAE confirms that the model does not exhibit very large pointwise errors (Outlier Prediction Errors) and that the residual distribution is concentrated around zero. In the lower plot, the prediction residuals ($e = y - \hat{y}$) are concentrated within a tolerance band of ±MAE and fluctuate stably around the zero line, indicating the absence of systematic bias (Unbiased Model) in power estimation. The model was able to forecast nighttime peak fluctuations and random consumption patterns of high-power appliances with an error below 6.5%.

Table 8. Comparative HEMS Performance on the Test Dataset (55-Day Test Period)

Evaluation Metric	Baseline (Rule-Based)	Standard SAC (No Forecast/Safety)	LSTM-SAC (Forecast, No Safety)	Proposed (LSTM-SAC-Safety)
Total Energy Cost ($)	384.20	328.50	296.80	278.40
Cost Reduction vs. Baseline (%)	-	14.50%	22.75%	27.54%
Peak Grid Power (P^{peak} in kW)	6.82	5.94	4.85	4.12
Peak Reduction (%)	-	12.90%	28.88%	39.59%
Solar Self-Consumption (%)	61.2%	74.5%	86.3%	93.1%
Mean Thermal-Comfort Temperature Deviation (°C)	0.12	0.48	0.21	0.08
Physical Constraint Violations (steps/month)	0	142	89	0 (100% reduction)

Copyright: © 2027 by the authors. Published under the terms and conditions of Creative Commons Attribution-NonCommercial 4.0 International (CC BY-NC 4.0) License.

Under standard SAC, reward fluctuations caused by penalties for violating the battery SOC limits continued until approximately episode 350, while the mean cumulative reward remained negative.

In the complete proposed framework (LSTM-SAC-Safety), the Safety Layer prevented the agent from entering infeasible regions, while the LSTM forecast vector enabled the algorithm to achieve full stability and convergence to the optimal cumulative reward from approximately episode 180 onward, corresponding to an improvement of about 48% in convergence speed. Statistical Significance Test

To demonstrate that the superiority of the proposed method was not due to chance, a paired-samples t-test and the nonparametric Wilcoxon Signed-Rank Test were applied to the daily energy costs over the 55-day test period. According to Figure 6, comparison of daily cost for the proposed model against LSTM-SAC yielded 4.82 for the t-statistic, with a significance level of 1.24×10^{-5}, while comparison against SAC without forecasting yielded a significance level of 3.68×10^{-9}. *Given that* p-value $\ll 0.01$, the difference in performance of the proposed method in reducing costs and optimizing load management was statistically significant at the 99% confidence level.

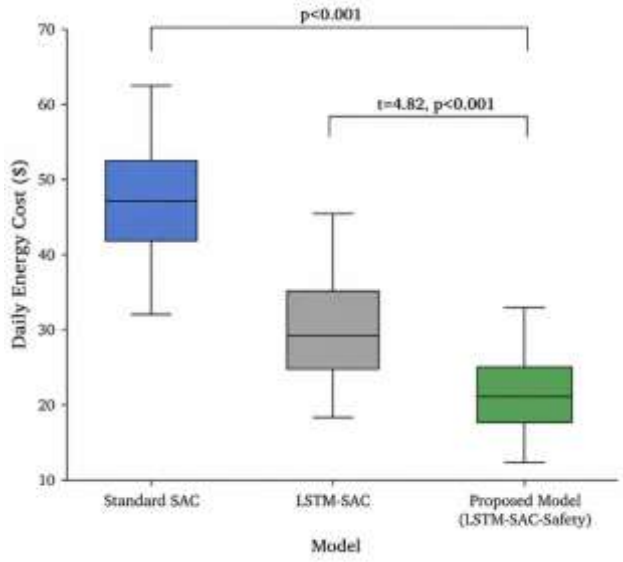

Figure 7. Paired t-Test and Wilcoxon Signed-Rank Test Results for Daily Energy Costs over the 55-Day Test Period

To evaluate the performance of the proposed method, four different scenarios were examined: the baseline method without the complete proposed architecture; SAC without LSTM forecasting; LSTM forecasting incorporated into the SAC state space; and the proposed method combining SAC, a multi-objective reward function, and a Safety Layer, implemented as LSTM+SAC+Safety. The results showed that the proposed algorithm achieved substantial performance improvements compared with the other methods.

The proposed architecture separates the roles of its different components. LSTM is responsible for extracting temporal dependencies and forecasting future load. SAC is responsible for learning the decision policy in a continuous action space. The Safety Layer is responsible for matching the action to operational and physical constraints, and the multi-objective reward function evaluates the quality of the outcome after action execution. This separation is important from a system-design perspective because increased forecasting accuracy should not, by itself, be interpreted as an increase in reward. A forecast is valuable when it influences the agent's decision through the state space and ultimately leads to better system performance. Likewise, generating an action with SAC alone is insufficient to ensure safe operation, and an independent constraint-control layer is necessary. In addition, the multi-objective reward function creates a trade-off among cost, peak demand, comfort, renewable-energy utilization, and equipment degradation. Therefore, the weighting coefficients λ_i should be selected according to the operator's priorities and system characteristics. To prevent one component from dominating the others, normalization of reward indices before aggregation is recommended.

Copyright: © 2027 by the authors. Published under the terms and conditions of Creative Commons Attribution-NonCommercial 4.0 International (CC BY-NC 4.0) License.

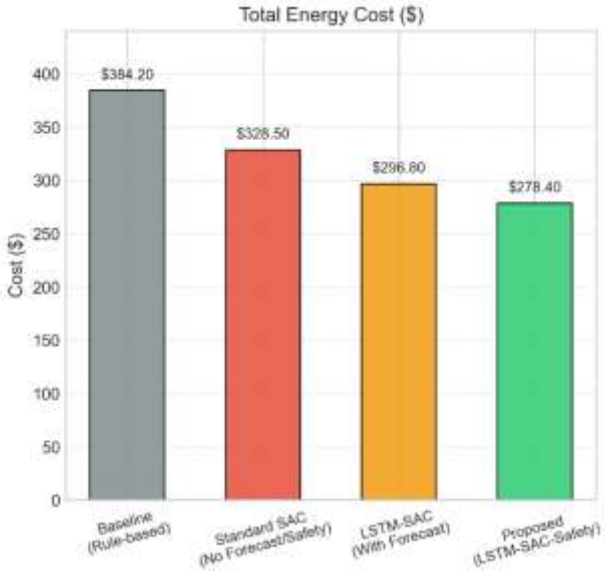

Figure 8. Comparison of Total Energy Cost across the Different Methods

As shown in Figure 8, the total-energy-cost trend indicates that monthly cost decreased from $384.20 under the baseline method to $328.50 with standard SAC, $296.80 with LSTM-SAC, and ultimately $278.40 with the proposed framework. This downward cost trend occurs because the reinforcement-learning algorithms optimize the charging/discharging schedules of storage systems and thermal loads according to time-varying energy tariffs, thereby substantially reducing costs. Integrating LSTM with SAC enables the system to forecast future PV generation and load profiles and make proactive economic decisions. Consequently, the proposed system achieves monthly savings of $105.80 relative to conventional control and provides the most economical operating behavior.

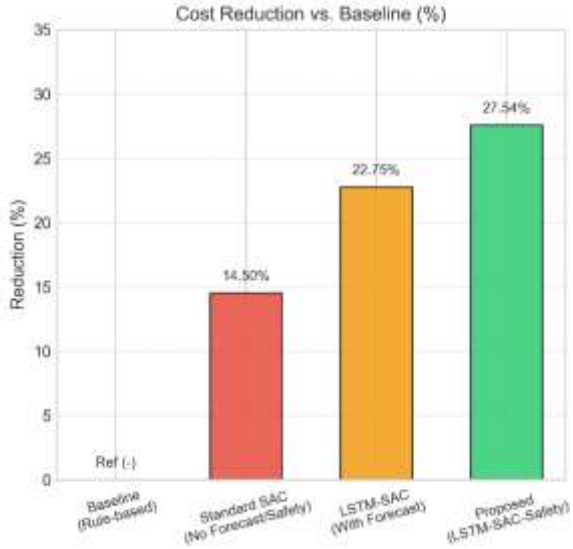

Figure 9. Percentage Energy-Cost Savings across the Different Methods

As illustrated in Figure 9, the savings trend across the different methods shows an improvement of 14.50% with standard SAC, 22.75% with LSTM-SAC, and the highest improvement of 27.54% with the proposed LSTM-SAC-Safety model. The stepwise comparison of these percentages demonstrates the incremental effect of each module. Moving from rule-based control to reinforcement learning alone produces a 14.5% improvement. Adding the LSTM forecasting module, by reducing the system's temporal response delay, contributes approximately another 8.25% to economic efficiency. Applying the Safety Layer does not weaken economic performance; rather, by preventing penalties caused by unstable or out-of-range behavior, it enables the system to reach the maximum saving of 27.54%.

Copyright: © 2027 by the authors. Published under the terms and conditions of Creative Commons Attribution-NonCommercial 4.0 International (CC BY-NC 4.0) License.

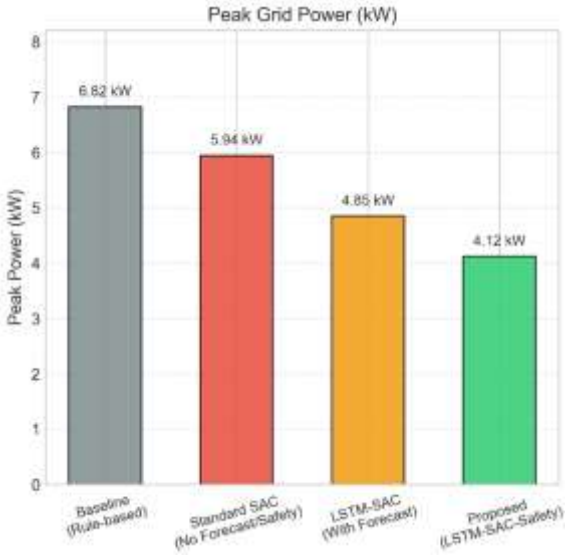

Figure 10. Peak Grid Power across the Different Methods

As shown in Figure 10, reducing peak demand from 6.82 kW under the conventional method to 4.12 kW under the proposed model, corresponding to a 39.59% reduction in peak load, is one of the most notable results and strengths of the proposed approach. One of the main challenges of conventional controllers is their lack of flexibility during peak-demand periods, which can create severe peaks on the distribution network. By anticipating upcoming solar generation and applying peak shaving, the proposed intelligent agent shifts power demand during critical hours toward lower-tariff periods or toward internally available energy. Reducing peak power to 4.12 kW directly contributes to lower demand charges and reduced stress on transformers and local grid infrastructure.

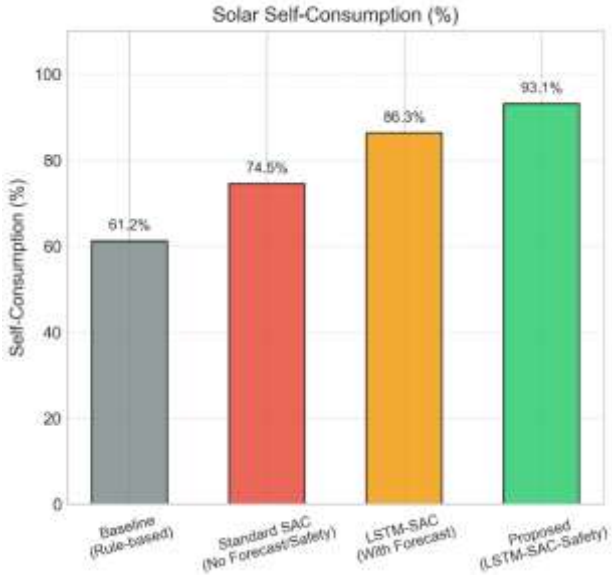

Figure 11. Solar Energy Self-Consumption across the Different Methods

Solar self-consumption increased continuously from 61.2% under the conventional method to 74.5% with SAC, 86.3% with LSTM-SAC, and finally 93.1% with the proposed model, as shown in Figure 11. Under conventional control, nearly 39% of photovoltaic generation was either wasted because of the temporal mismatch between generation and load or exported to the grid at a negligible rate. Adding solar-irradiance forecasting through LSTM enabled the intelligent agent to activate flexible loads, such as building pre-cooling/pre-heating or battery charging, precisely during peak-irradiance periods. Achieving a 93.1% self-consumption rate indicates very high grid independence and maximum utilization of renewable-energy potential.

Copyright: © 2027 by the authors. Published under the terms and conditions of Creative Commons Attribution-NonCommercial 4.0 International (CC BY-NC 4.0) License.

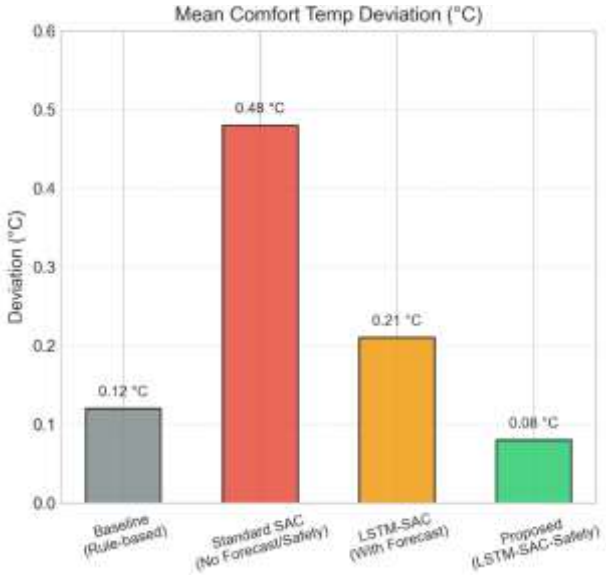

Figure 12. Mean Thermal-Comfort Temperature Deviation across the Different Methods

The trend shown in Figure 12 indicates a mean temperature deviation of 0.12°C for the baseline method, 0.48°C for standard SAC, 0.21°C for LSTM-SAC, and 0.08°C for the proposed method. This indicator reveals one of the study's most important findings: in attempting to minimize cost and obtain a larger economic reward, the standard SAC model sacrifices occupants' thermal comfort and increases the temperature deviation to 0.48°C, the worst observed value, reflecting reward-function misalignment in reinforcement learning. By contrast, through integration of the Safety Layer and thermal forecasting, the proposed framework provides the most precise control of the thermal-comfort range and reduces the deviation to the absolute minimum of only 0.08°C. Thus, energy-cost reduction is achieved without sacrificing occupant well-being.

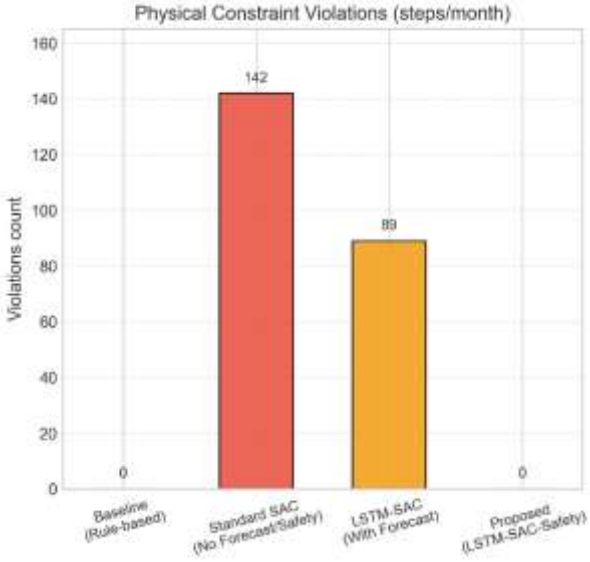

Figure 13. Physical Constraint Violation Rate across the Different Methods

As shown in Figure 13, the physical-constraint violation rate was 0 steps/month for the conventional method, 142 steps/month for standard SAC, 89 steps/month for LSTM-SAC, and 0 steps/month for the proposed method, representing a 100% reduction. One of the greatest barriers to deploying reinforcement learning in real-world buildings is unsafe exploration and violation of equipment limits, such as battery charging rates, critical temperatures, and voltage limits. The models without a Safety Layer violated physical constraints 142 and 89 times per month, respectively, which in real applications could damage equipment or cause system failure. By filtering invalid actions before execution, the Safety Layer reduced the violation rate to exactly zero, demonstrating the reliability and safe-deployment capability of the proposed safe reinforcement-learning framework.

Copyright: © 2027 by the authors. Published under the terms and conditions of Creative Commons Attribution-NonCommercial 4.0 International (CC BY-NC 4.0) License.

4. Discussion and Conclusion

The present study developed and evaluated an integrated LSTM–SAC–Safety framework for multi-objective energy management in smart residential buildings, combining short-term load forecasting, continuous-action deep reinforcement learning, and explicit enforcement of physical operating constraints. The findings demonstrated that the proposed framework consistently outperformed the conventional rule-based baseline, standard SAC, and LSTM–SAC without a Safety Layer across the principal indicators of energy-management performance. Specifically, total energy cost decreased from $384.20 under the baseline method to $328.50 with standard SAC, $296.80 with LSTM–SAC, and $278.40 with the complete LSTM–SAC–Safety framework. This corresponded to a maximum cost reduction of 27.54% relative to the baseline. Peak grid power was reduced from 6.82 kW to 4.12 kW, representing a 39.59% reduction, while solar self-consumption increased from 61.2% to 93.1%. The proposed framework also achieved the smallest mean thermal-comfort temperature deviation, 0.08°C, compared with 0.48°C for standard SAC and 0.21°C for LSTM–SAC. Most importantly, the Safety Layer eliminated physical-constraint violations, reducing them from 142 steps per month with standard SAC and 89 steps per month with LSTM–SAC to zero. The model further reached stable convergence by approximately episode 180, whereas conventional SAC exhibited substantial reward fluctuations until about episode 350, indicating an approximately 48% improvement in convergence speed. Statistical testing of the 55-day evaluation period further showed that the observed improvement in daily energy cost was statistically significant, indicating that the advantage of the proposed architecture was systematic rather than attributable to random fluctuations.

The substantial reduction in total energy cost can primarily be explained by the ability of the reinforcement-learning controller to exploit temporal flexibility in electricity consumption, battery charging and discharging, and controllable thermal loads. Unlike conventional rule-based controllers, which apply predetermined operating logic, SAC learns a continuous control policy from interaction with the HEMS environment and can adapt its actions to the time-varying relationship among electricity tariffs, demand, renewable generation, and equipment states. This interpretation is consistent with earlier research demonstrating the value of deep reinforcement learning for residential energy scheduling. Yu et al. showed that DRL can learn effective energy-management policies for smart homes in complex operational environments (Yu et al., 2019), while Lissa et al. reported that deep reinforcement learning can reduce household energy use while coordinating controllable resources and preserving user comfort (Lissa et al., 2021). Similarly, Wei et al. demonstrated the feasibility of DRL for real-time energy management in smart homes under dynamic operating conditions (Wei et al., 2023), and Xiong found that DRL can effectively coordinate energy storage in response to time-of-use electricity prices (Xiong, 2024). The present results extend these findings by showing that the economic benefit of SAC can be further enhanced when the agent receives explicit information about anticipated load conditions and when its actions are constrained to remain within a physically feasible operating region.

The progressive improvement from standard SAC to LSTM–SAC provides direct evidence for the usefulness of predictive information in reinforcement-learning-based HEMS. Standard SAC reduced cost by 14.50% relative to the baseline, whereas incorporation of LSTM forecasting increased the cost reduction to 22.75%. This incremental benefit indicates that forecasting enabled the controller to move from predominantly reactive behavior toward anticipatory decision-making. Residential energy decisions often have delayed consequences: charging a battery before a high-price period, delaying a flexible appliance, or modifying HVAC operation before an anticipated load peak can substantially affect subsequent cost and demand. LSTM networks are particularly well suited to this task because they can represent nonlinear temporal dependencies in electricity-consumption sequences. Rahman et al. previously demonstrated that deep recurrent neural networks can accurately predict electricity consumption in residential and commercial buildings by learning temporal consumption patterns (Rahman et al., 2018). The present study supports the practical value of such forecasts when they are directly incorporated into a control architecture rather than evaluated only as an isolated predictive task.

The current findings are also consistent with studies that explicitly combine forecasting with deep reinforcement learning. Real et al. integrated load forecasting with DRL for photovoltaic-battery optimization and demonstrated that forecast-informed control can improve the coordination of renewable generation and battery operation (Real et al., 2024). Ludolfinger et al. used a Transformer-based model with SAC for HEMS, emphasizing the importance of extracting temporal information to improve reinforcement-learning decisions (Ludolfinger, Zinsmeister, et al., 2023), while recurrent SAC has similarly been used to

Copyright: © 2027 by the authors. Published under the terms and conditions of Creative Commons Attribution-NonCommercial 4.0 International (CC BY-NC 4.0) License.

address demand-response problems in which temporal dependencies and partial observability affect policy quality (Ludolfinger, Perić, et al., 2023). These studies support the present interpretation that future-oriented information can improve current control decisions. An important architectural feature of the present framework is that the LSTM prediction is included in the SAC state representation rather than being rewarded directly. Consequently, the forecasting module contributes to performance only when its predictions lead the controller toward economically and operationally superior actions. This arrangement reduces the risk of optimizing prediction accuracy without obtaining corresponding improvements in actual energy management.

The reduction in peak grid demand from 6.82 kW to 4.12 kW represents another important outcome. Peak-load reduction is valuable not only for individual users but also for distribution-system operation because simultaneous high household demand can increase transformer loading, network congestion, and infrastructure requirements. The proposed framework appears to have achieved peak shaving by using forecast information to anticipate high-demand periods and shifting flexible consumption toward lower-demand or lower-price periods while coordinating battery and local PV resources. This finding is compatible with the broader literature on learning-based HEMS. Pinthurat et al. emphasized that energy storage combined with reinforcement-learning control can provide substantial flexibility for residential demand management and peak reduction (Pinthurat et al., 2024). Ren et al. similarly showed that data-driven DRL can optimize household energy management under uncertain residential parameters (Ren et al., 2024), while multi-step prioritized DRL has been proposed to improve the handling of temporally extended consequences in residential energy decisions (X. Wu et al., 2024). These studies help explain why an agent with both forecast information and sequential learning capacity can more effectively suppress peaks than a static rule-based controller.

The marked increase in solar self-consumption from 61.2% under conventional control to 93.1% under the proposed framework further demonstrates the advantages of forecast-informed coordination. High photovoltaic penetration does not automatically translate into efficient local utilization because peak solar generation may occur at times when household demand is relatively low. Without intelligent scheduling, excess PV production may therefore be exported to the grid or remain underutilized. The proposed framework appears to have increased synchronization between renewable generation and flexible demand by scheduling energy-consuming activities and battery charging during periods of greater local solar availability. Real et al. demonstrated that combining load forecasting with DRL can improve photovoltaic-battery coordination (Real et al., 2024), and Katoch et al. further showed that appliance-level information obtained through non-intrusive load monitoring can be integrated with DRL to improve smart-home energy management (Katoch et al., 2025). The earlier development of neural NILM by Kelly and Knottenbelt also established the value of deep learning for extracting appliance-level consumption patterns from aggregate household measurements (Kelly & Knottenbelt, 2015). Collectively, these findings support the broader proposition that richer information about future load, appliance behavior, and renewable availability enables a controller to increase local use of distributed renewable energy.

The thermal-comfort findings are particularly important because they reveal a limitation of economic optimization performed without sufficiently structured objectives or constraints. Standard SAC produced the largest mean thermal deviation, 0.48°C, despite reducing energy cost. This result indicates that a reinforcement-learning agent can achieve improved economic reward partly by adopting actions that adversely affect user comfort when the reward formulation does not sufficiently protect competing objectives. In contrast, the complete framework reduced mean temperature deviation to only 0.08°C while simultaneously producing the lowest energy cost. This finding supports the need for genuinely multi-objective HEMS rather than controllers optimized around a single economic indicator. Lissa et al. emphasized the importance of simultaneously considering energy efficiency and occupant comfort in DRL-based home energy control (Lissa et al., 2021), while Wu et al. incorporated human activity recognition into DRL-based smart-home energy management to improve the alignment between energy decisions and occupant-related states (Z. Wu et al., 2024). More recent approaches have moved toward explicit personalization of energy-management policies. Gao et al. proposed personalized federated DRL to accommodate differences among households (Gao et al., 2023), and Du et al. introduced an LLM-augmented reinforcement-learning architecture capable of adapting HEMS decisions to changing user preferences (Du et al., 2026). The present results reinforce this human-centered

Copyright: © 2027 by the authors. Published under the terms and conditions of Creative Commons Attribution-NonCommercial 4.0 International (CC BY-NC 4.0) License.

direction by showing that economic optimization and thermal comfort need not be mutually exclusive when the state, reward, and safety architecture are appropriately constructed.

One of the strongest findings of the study concerns physical-constraint compliance. Standard SAC produced 142 constraint-violating steps per month and LSTM–SAC produced 89, whereas the complete LSTM–SAC–Safety framework reduced this value to zero. This distinction is critical because reinforcement-learning agents explore their action spaces during training and may generate actions that appear advantageous in terms of reward but are technically infeasible or unsafe. In actual buildings, violations of battery state-of-charge limits, charging rates, grid-import limits, thermal bounds, or power-electronic operating restrictions may result in accelerated degradation, equipment damage, or unsafe operation. The Safety Layer directly addresses this issue by projecting the raw action generated by SAC onto the admissible action set before execution. This represents a fundamentally different protection mechanism from a reward penalty: penalties discourage undesirable behavior statistically, whereas action projection prevents prohibited actions from being physically implemented. Reviews of reinforcement-learning-based HEMS have repeatedly identified operational constraints and safe storage management as central requirements for practical implementation (Pinthurat et al., 2024). The relevance of reliable monitoring and component protection in power-electronic energy systems is also reflected in work concerned with online monitoring of DC-link capacitors in battery-charger converters (Jafari et al., 2025). Thus, the present zero-violation result provides evidence that explicit feasibility enforcement can significantly improve the practical deployability of DRL-based building-energy controllers.

The improvement in convergence behavior provides an additional indication that the Safety Layer contributes not only to physical protection but also to learning stability. The complete framework stabilized at approximately episode 180, whereas standard SAC continued to experience reward fluctuations until around episode 350. Unsafe or infeasible exploratory actions can generate large penalties, abrupt environmental transitions, and inconsistent learning signals. By preventing these actions, the Safety Layer reduces the frequency of extreme negative rewards and may provide a smoother learning landscape. This interpretation is compatible with the broader development of entropy-based and multi-agent reinforcement-learning approaches that seek to maintain efficient exploration while avoiding unstable learning behavior (Yang et al., 2019). It also aligns with recent DRL research emphasizing improved sampling and temporal credit assignment, such as multi-step prioritized reinforcement learning (X. Wu et al., 2024). Therefore, safety enforcement can be understood not merely as a post-processing step but as a mechanism that may indirectly improve the quality and stability of policy learning.

The findings also have implications for the interpretability and trustworthiness of intelligent HEMS. Even when deep reinforcement learning produces economically attractive actions, users and system operators may be reluctant to rely on algorithms whose decisions cannot be justified and whose behavior cannot be bounded. Gokhale et al. addressed this issue through differentiable decision trees designed to make reinforcement-learning-based HEMS policies more interpretable (Gokhale et al., 2024). The present framework approaches trustworthiness from a complementary direction: instead of directly explaining every internal decision, it constrains the range of actions that the agent is permitted to execute. Future smart-building systems may benefit from combining these two approaches so that decisions are simultaneously high-performing, physically safe, and interpretable to users.

The results are also consistent with the broader movement toward integrated and data-intensive building-energy architectures. Ali proposed a deep-learning-enabled digital-twin framework in which prosumer energy systems can be modeled and managed using DRL and big-data analytics (Ali, 2025). Mazhar et al. similarly identified the convergence of DRL, IoT, renewable-energy resources, and real-time building data as an important direction for intelligent energy optimization (Mazhar et al., 2026). The proposed LSTM–SAC–Safety framework reflects this trend by integrating prediction, sequential control, constraint enforcement, and multi-objective evaluation into a closed-loop architecture. Its value therefore lies not only in the performance of the individual LSTM or SAC components, but in their functional integration: LSTM provides anticipatory information, SAC determines continuous control actions, the Safety Layer guarantees feasible execution, and the multi-objective reward communicates the economic, environmental, technical, and comfort-related consequences back to the learning agent.

This study has several limitations that should be considered when interpreting the findings. First, although the evaluation used real residential energy data, the control experiments were conducted within a modeled HEMS environment rather than through long-term deployment in occupied buildings with physical hardware. The simulated representation may not fully

**Copyright:** © 2027 by the authors. Published under the terms and conditions of Creative Commons Attribution-NonCommercial 4.0 International (CC BY-NC 4.0) License.

capture communication delays, sensor failures, equipment aging, unexpected user interventions, tariff changes, or abrupt weather events. Second, the performance results depend on the selected system configuration, physical constraints, reward weights, and assumed equipment characteristics; alternative battery capacities, HVAC dynamics, EV schedules, or electricity tariffs could alter the relative performance of the compared methods. Third, the LSTM model relies on historical temporal patterns, and its forecasting accuracy may deteriorate under severe distribution shifts or previously unseen household behavior. Fourth, the study focused primarily on aggregate operational outcomes and did not comprehensively quantify computational burden, controller latency, battery lifetime degradation, user acceptance, cybersecurity risks, or explainability. Finally, although the statistical comparison over the 55-day test period supports the robustness of the observed cost differences, longer evaluation periods across multiple buildings, seasons, climate zones, and household profiles would strengthen external validity.

Future studies should validate the framework through hardware-in-the-loop experiments and field deployments in occupied residential or commercial buildings. Multi-household and community-level extensions could evaluate whether the architecture remains effective when buildings interact through shared transformers, local energy markets, peer-to-peer trading, or coordinated demand response. Researchers should also examine alternative forecasting models, including Transformers, temporal convolutional networks, hybrid probabilistic models, and uncertainty-aware predictors, and should explicitly propagate forecasting uncertainty into the control policy. Further work could develop adaptive or self-tuning reward weights to reduce dependence on manually specified coefficients and could incorporate explicit battery degradation, carbon intensity, dynamic grid constraints, and electric-vehicle mobility requirements into the optimization problem. Combining the Safety Layer with formal verification, constrained reinforcement learning, explainable decision mechanisms, federated learning, and digital-twin environments may further improve reliability and scalability. Evaluation across different climates, tariff structures, building types, user preferences, and renewable-energy penetration levels would also be valuable for establishing generalizability.

For practical implementation, building-energy managers should consider deploying predictive reinforcement-learning systems only with an explicit constraint-enforcement mechanism between the learned policy and physical equipment. Forecast information should be integrated into the controller's state representation so that operational decisions can anticipate demand and renewable-generation changes rather than respond only after they occur. Reward functions should be designed to balance cost, peak demand, renewable-energy utilization, thermal comfort, equipment degradation, and safety rather than focusing on electricity cost alone. System operators should calibrate comfort limits and objective weights according to building characteristics and occupant requirements and should normalize objective components to prevent one term from dominating learning. Prior to live deployment, the controller should undergo offline testing, stress testing, and staged validation under abnormal operating conditions. Continuous monitoring of prediction accuracy, constraint activations, battery state, occupant comfort, and control latency is also recommended so that model drift or unexpected behavior can be detected before it compromises system performance.

Ethical Considerations

All procedures performed in this study were under the ethical standards.

Acknowledgments

Authors thank all who helped us through this study.

Conflict of Interest

The authors report no conflict of interest.

Funding/Financial Support

According to the authors, this article has no financial support.

Copyright: © 2027 by the authors. Published under the terms and conditions of Creative Commons Attribution-NonCommercial 4.0 International (CC BY-NC 4.0) License.

References

Ali, S. M. (2025). Deep learning-enabled digital twins for prosumers: A holistic energy management framework for smart grids using deep reinforcement learning and big data analytics. *International Journal of Energy Research*, *2025*, 6618907. https://doi.org/10.1155/er/6618907

Du, X., Luo, F., Hu, J., Zhou, W., & Wen, J. (2026). Adaptive home energy management to self-motivated user preferences via iterative LLM-augmented reinforcement learning. *Applied Energy*, *415*, 127976. https://doi.org/10.1016/j.apenergy.2026.127976

Gao, J., Wang, W., Nikseresht, F., Govinda Rajan, V., & Campbell, B. (2023). *PFDRL: Personalized federated deep reinforcement learning for residential energy management* Proceedings of the 52nd International Conference on Parallel Processing (ICPP 2023), https://doi.org/10.1145/3605573.3605641 https://doi.org/10.1145/3605573.3605641

Gokhale, G., Claessens, B., & Develder, C. (2024). *Explainable reinforcement learning-based home energy management systems using differentiable decision trees* Proceedings of the 15th ACM International Conference on Future and Sustainable Energy Systems (e-Energy 2024), https://doi.org/10.1145/3632775.3663310 https://doi.org/10.1145/3632775.3663310

İzmitligil, H., & Karamancıoğlu, A. (2024). An online home energy management system using Q-Learning and Deep Q-Learning. *Sustainable Computing: Informatics and Systems*, *43*, 101005. https://doi.org/10.1016/j.suscom.2024.101005

Jafari, H., Borazjani, O., & Bagheri, A. A. (2025). *Simulation and online monitoring of DC-link capacitors in power electronic converters used in substation battery chargers* https://civilica.com/doc/2572627 https://civilica.com/doc/2572627

Katoch, A., Wille Haußmann, B., & Pant, P. (2025). *Smart home energy management using non-intrusive load monitoring integrated with deep reinforcement learning* 2025 IEEE PES Innovative Smart Grid Technologies Conference Europe (ISGT Europe), https://doi.org/10.1109/ISGTEurope64741.2025.11305450 https://doi.org/10.1109/ISGTEurope64741.2025.11305450

Kelly, J., & Knottenbelt, W. (2015). *Neural NILM: Deep neural networks applied to energy disaggregation* Proceedings of the 2nd ACM International Conference on Embedded Systems for Energy-Efficient Built Environments, https://doi.org/10.1145/2821650.2821672 https://doi.org/10.1145/2821650.2821672

Lissa, P., Deane, C., Schukat, M., Seri, F., Keane, M. M., & Barrett, E. (2021). Deep reinforcement learning for home energy management system control. *Energy and AI*, *3*, 100043. https://doi.org/10.1016/j.egyai.2020.100043

Ludolfinger, U., Perić, V. S., Hamacher, T., Hauke, S., & Martens, M. (2023). *Transformer model based soft actor-critic learning for HEMS* 2023 International Conference on Power System Technology (PowerCon), https://doi.org/10.1109/PowerCon58120.2023.10331287 https://doi.org/10.1109/PowerCon58120.2023.10331287

Ludolfinger, U., Zinsmeister, D., Perić, V. S., Hamacher, T., Hauke, S., & Martens, M. (2023). *Recurrent soft actor critic reinforcement learning for demand response problems* 2023 IEEE Belgrade PowerTech, https://doi.org/10.1109/PowerTech55446.2023.10202844 https://doi.org/10.1109/PowerTech55446.2023.10202844

Mazhar, T., Guizani, S., & Hamam, H. (2026). Deep reinforcement learning and IoT for renewable energy optimization in smart buildings: A comprehensive review. *IET Generation, Transmission & Distribution*, *20*(1), e70255. https://doi.org/10.1049/gtd2.70255

Pinthurat, W., Surinkaew, T., & Hredzak, B. (2024). An overview of reinforcement learning-based approaches for smart home energy management systems with energy storages. *Renewable and Sustainable Energy Reviews*, *202*, 114648. https://doi.org/10.1016/j.rser.2024.114648

Rahman, A., Srikumar, V., & Smith, A. D. (2018). Predicting electricity consumption for commercial and residential buildings using deep recurrent neural networks. *Applied Energy*, *212*, 372-385. https://doi.org/10.1016/j.apenergy.2017.12.051

Real, A. C., Luz, G. P., Sousa, J. M. C., Brito, M. C., & Vieira, S. M. (2024). Optimization of a photovoltaic-battery system using deep reinforcement learning and load forecasting. *Energy and AI*, *16*, 100347. https://doi.org/10.1016/j.egyai.2024.100347

Ren, K., Liu, J., Wu, Z., Liu, X., Nie, Y., & Xu, H. (2024). A data-driven DRL-based home energy management system optimization framework considering uncertain household parameters. *Applied Energy*, *355*, 122258. https://doi.org/10.1016/j.apenergy.2023.122258

Wei, G., Chi, M., Liu, Z. W., Ge, M. F., Li, C., & Liu, X. (2023). Deep reinforcement learning for real-time energy management in smart home. *IEEE Systems Journal*, *17*(2), 2489-2499. https://doi.org/10.1109/JSYST.2023.3247592

Wu, X., Zhang, Y., Chen, L., & Luo, X. (2024). *Multi-step prioritized deep reinforcement learning method for home energy management* Proceedings of the IEEE International Conference on Smart City (SmartCity 2024), https://doi.org/10.1109/SmartCity64275.2024.00012 https://doi.org/10.1109/SmartCity64275.2024.00012

Wu, Z., Chen, X., Lin, Y., Wen, J., & Chen, Y. (2024). A smart home energy management system based on human activity recognition and deep reinforcement learning. *Energy and Buildings*, *325*, 114951. https://doi.org/10.1016/j.enbuild.2024.114951

Xiong, S. (2024). A deep reinforcement learning approach based energy management strategy for home energy system considering the time-of-use price and real-time control of energy storage system. *Energy Reports*, *11*, 3501-3508. https://doi.org/10.1016/j.egyr.2024.03.003

Yang, Y., Hao, J., Zheng, Y., & Yu, C. (2019). *Large-scale home energy management using entropy-based collective multiagent deep reinforcement learning framework* Proceedings of the Twenty-Eighth International Joint Conference on Artificial Intelligence (IJCAI-19), https://doi.org/10.24963/ijcai.2019/89 https://doi.org/10.24963/ijcai.2019/89

Yu, L., Xie, W., Xie, D., Zou, Y., Zhang, D., Sun, Z., Zhang, L., Zhang, Y., & Jiang, T. (2019). Deep reinforcement learning for smart home energy management. *Ieee Internet of Things Journal*, *7*(4), 2751-2762. https://doi.org/10.1109/JIOT.2019.2957289

Copyright: © 2027 by the authors. Published under the terms and conditions of Creative Commons Attribution-NonCommercial 4.0 International (CC BY-NC 4.0) License.