Sigma J Eng Nat Sci, Vol. 44, No. 2, pp. 880–893, April, 2026

Sigma Journal of Engineering and Natural Sciences
Web page info: https://sigma.yildiz.edu.tr
DOI: 10.14744/sigma.2025.00051

Research Article

Multi-zone commercial market building HVAC control strategy based on reinforcement learning algorithm models

Ganesh MURADE[1,*], Bhanu Pratap SONI[2], Ankit Kumar SHARMA[1]

[1]University of Engineering & Management, Jaipur, Rajasthan, 700160, India
[2]SEEE, Fiji National University, 1544, Fiji

ARTICLE INFO

Article history
Received: 16 July 2024
Revised: 04 October 2024
Accepted: 05 March 2025

Keywords:
Actor-critic Learning; Deep Deterministic Policy Gradient (DDPG); Deep Reinforcement Learning (Deep RL); Multi-Zone Commercial HVAC; Linear RL; Reinforcement Learning

ABSTRACT

The optimization of residential heating, ventilation, and air conditioning (HVAC) systems is crucial for reducing energy consumption and maintaining user comfort. As urbanization increases, there is a demand for smart buildings with energy-consuming appliances. Researchers are developing HVAC control strategies, especially for commercial buildings with complex load patterns. This paper presents a novel model-free deep reinforcement learning (RL) approach. RL allows systems to learn and adapt to individual occupant preferences, creating customized comfort using the deep deterministic policy gradient (DDPG) algorithm to design an optimal control strategy for multi-zone residential HVAC systems. DDPG aims to reduce energy costs while maintaining occupant comfort. The DDPG reinforcement learning technique is applied to control a multi-zone commercial HVAC system, aiming to minimize energy costs while maintaining occupant comfort. The DDPG method achieves a 56% faster convergence time compared to linear HVAC control methods and a 15% improvement over linear reinforcement learning models. The mean steps required for DDPG and linear RL models are 9.9 and 115.3, respectively.

Cite this article as: Murade G, Soni BP, Sharma AK. Multi-zone commercial market building HVAC control strategy based on reinforcement learning algorithm models. Sigma J Eng Nat Sci 2026;44(2):880–893.

INTRODUCTION

It is estimated that 41% of India's annual energy consumption [1] is attributed to the building sector. Consequently, building temperature regulation has garnered significant attention [2,3] recently. The primary aim is to strike a balance between minimizing energy usage and ensuring high comfort levels for office workers. However, uncertainty arises due to occupants' behavior and insufficient data on their comfort preferences, impacting decision-making regarding optimal control strategies [4,5]. Integrating these considerations into the design process poses a pressing challenge. Occupant behavior adds complexity to building systems, characterized by continuous and discrete states, further complicated by variable sensor requirements. Hence, developing a scalable control framework is imperative. Drawing inspiration from recent

***Corresponding author.**
*E-mail address: ganesh.murade26@gmail.com

This paper was recommended for publication in revised form by Editor-in-Chief Ahmet Selim Dalkilic

Published by Yıldız Technical University Press, İstanbul, Turkey

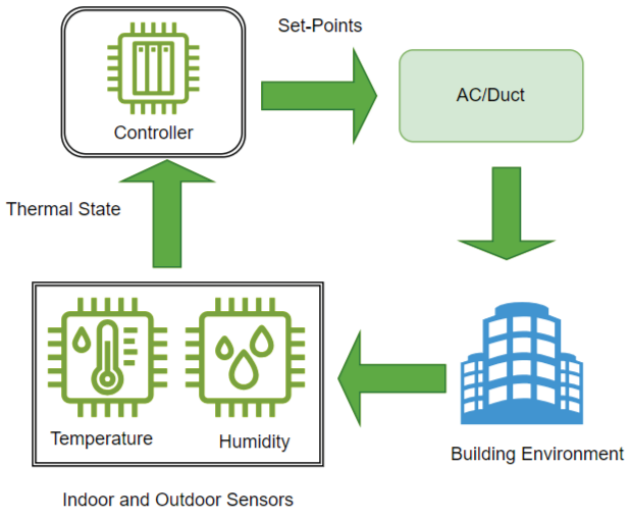

Figure 1. The design of the thermal control system and Controller set Points (sensors).

advancements in RL [6,7], this study aims to evaluate the efficacy of RL in addressing intelligent temperature control in building environments.

The air conditioning system, depicted in Figure 1, currently stands as the most sophisticated device for maintaining a comfortable temperature within a structure. Additionally, efficient demand-side energy management measure, aid in reducing peak loads and stabilizing overall system performance [8]. Several studies have been conducted to determine the optimal methods for optimizing HVAC control systems to enhance efficiency.

A primal-dual algorithm is used in [9] to determine the energy supplier's pricing strategy and the HVAC operating states that are most beneficial to the user, within the context of a load prediction error model for HVAC energy management. For responsive commercial HVAC demand scheduling a day in advance, another study uses a regression method [10]. According to [11], a hierarchical control method might allow the HVAC system to serve as the principal frequency regulator for the bulk system. In [12], an optimization method based on the work of Yuri Lyapunov is presented for managing HVAC loads without requiring estimates of price or temperature fluctuations. A distributed Tran's active control market approach for HVAC systems in commercial buildings is presented in [13] to demonstrate the efficiency of HVAC systems during peak shaving and load shifting (Fig. 2). Using an adaptive control system, Andres-Chicote et al. [11], managed the building's HVAC system to maximize both occupant thermal comfort and the building's energy efficiency. Giuseppe et al. [12], proposed a model prescient control and hereditary calculation-based streamlining structure to diminish warming energy utilization and related inconvenience. Wei and others introduced an information-driven technique for controlling air conditioning frameworks with changing air volume giving profound support to learning. Deng et al. [14], proposed a unique HVAC control system that actively detects changes in the building's environment by using DQN, resulting in significant energy savings.

All of the previously mentioned methodologies fall under the umbrella term of "model-based techniques," which need logical arrangement tool compartments for pragmatic runtime of the executives as well as displaying the particular warm elements of the air conditioning while at the same time considering the impacts of the encompassing climate. Since the structure and hardware models should be acclimated to a given structure to deliver the right

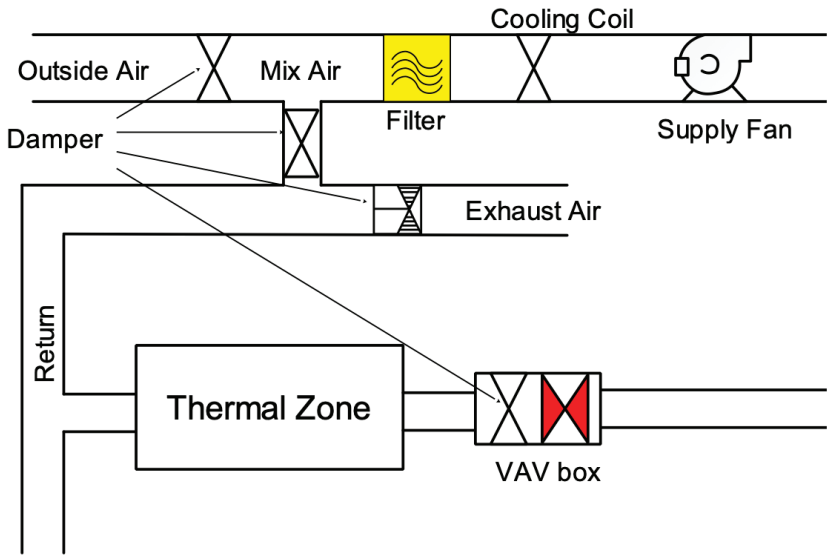

Figure 2. Schematic of the HVAC air distribution system.

discoveries, the model-based approaches might experience the ill effects of estimation mistakes (for instance, building model incorrectness) and figuring failures. This presents a significant obstacle to the widespread application of model-based methods.

In the meantime, the outcome of Alpha Go [14] shows the headway that has been made in AI advancements like profound learning and support learning. In [15], the creator presents a wide outline of the possible utilizations of AI in the field of force framework research. In [16], an illustration of how AI-based technology is beginning to be implemented in industrial applications in actual control centers, the primary known control-room utilization of man-made intelligence-driven circulated includes choice for an enormous, certifiable power matrix is shown.

Specifically, profound Reinforcement learning (RL), which consolidates a profound brain organization (DNN) with RL, has collected critical interest as of late for its capability to successfully address exceptionally complex control and improvement issues across many aspects. Two unique ways to deal with upgrading the energy the board strategies for HEVs are twofold Q learning [17] and ceaseless profound deterministic approach inclination (DDPG). Reinforcement learning (RL), a machine learning method that has emerged in recent years, features self-learning and online learning capabilities. RL is an information driven control approach since, utilizing the "activities and prizes" component, it can achieve the versatile improvement of regulators in any event, when no control framework models are free. In [18], the no concurrent advantage entertainer pundit is utilized to decide the most practical times for the activity of an organization of decentralized power plants. In [19], a profound Q learning approach is created to help mass power framework upkeep direction. In [20], a protected profound RL technique is investigated to get the ideal control plan of the dynamic dissemination network considering voltage level cutoff points. To do this, we add a secure layer to the standard actor-network, which helps to ensure that voltage limitations are not breached.

A portion of the primary endeavors on the air conditioning framework control issue zeroed in on the most proficient method to best utilize the strong profound RL technique to boost both energy and monetary effectiveness. A convolutional neural network (CNN) approximates the state-activity esteem capability in [21] to more readily catch the spatial and worldly connections in the info state information. In [22], scientists investigate the utilization of a profound strategy slope (DPG) way to deal with dealing with a wide assortment of responsive necessities. Utilizing an actor-critic approach, [23] focuses on optimizing HVAC's thermal comfort and energy consumption. In [24], an advantageous actor-critic-based HVAC control framework is created for a comprehensive building energy model. To accomplish appropriate control balance across various HVAC systems, linear reinforcement is used [25].

All of these studies show that using deep RL techniques to optimize the HVAC temperature management strategy is more successful than using custom-built benchmarks. To reduce the search space, previous studies frequently discretize continuous HVAC control activities like HVAC set point or air flow rate. At low granularities or without combining action spaces, discretization may achieve satisfactory performance. However, when the action space is high-dimensional, as when controlling HVAC in a building with multiple zones, it encounters the exponential explosion issue. The algorithm's performance suffers as a result, and deep RL technique training necessitates more simulations. When the action space is high-dimensional, like when controlling HVAC in a building with several zones, however, it runs into the problem of exponential growth. As a consequence, the algorithm performance degrades and more simulations are required for training deep RL techniques.

As a result of the above worries, this study additionally employs the DDPG technique to fine-tune the continuous thermal management strategy of commercial HVAC. Here is a quick rundown of what this study adds to the body of knowledge:

- Designed benchmark scenarios without RL to illustrate that the implemented DDPG may give larger monetary gains while keeping consumer comfort.
- To demonstrate the superiority of the DDPG technique in managing the never-ending activity space, which is more typical in a plethora of verifiable scenarios, we lead a comprehensive correlation between it and the widely used direct assistance strategy.
- The optimum HVAC control technique is a well-trained deep RL approach with good generalizability and adaptability that can handle a wide range of pricing signals and physical constraints.

Research paper insights in-depth exploration of the air conditioning control problem. A brief overview of two prominent deep reinforcement learning approaches. The paper also emphasizes the linear reinforcement method and simulation results of the DDPG method, along with a comparative analysis against the linear reinforcement method. The paper underscores the importance of the linear reinforcement method for its simplicity and ease of implementation.

PROBLEM FORMULATION FOR MULTI-ZONE CLIMATE CONTROL IN COMMERCIAL MARKET SETTINGS

Commercial buildings, such as malls, offices, and supermarkets, contain distinct climate zones with different heating and cooling requirements. Controlling multi-zone efficiently is a challenge due to factors like occupancy fluctuations, weather variations, building thermal properties, and energy costs. Multi-zone climate control systems aim to enhance energy efficiency while maintaining comfort and regulatory standards. Multi-zone HVAC (Heating,

Ventilation, and Air Conditioning) systems are designed to regulate the temperature and air quality across different zones within a building. Section 2.1. gives the brief information about the various issues related to implementation of HVAC in multizone.

Introduction to the HVAC Multi-Zone Control Issue

In this analysis, we focus on a multi-floor commercial/market structure. The HVAC system has a setpoint that may be changed to regulate the interior temperature in each zone individually. In addition to the "Cooling" and "Heating" settings, an "Auto" setting is also available for the HVAC system. In view of the room's ongoing temperature and the client's chosen temperature edge, the air conditioning framework will consequently switch among cooling and warming when the indoor regulator is in "Auto" mode. To keep individuals agreeable, the warming, ventilation, and cooling framework will turn on when within temperature decreases beneath the setpoint. We will zero in on the situation in which all zones require warming without forfeiting over-simplification. The motivation behind central air framework control is to give an agreeable inside climate at the most reduced conceivable energy consumption.

Multi-Zone HVAC System

A multi-zone HVAC system divides your house into smaller regions or zones, allowing you to independently manage the temperature in each. You could simply establish a distinct zone for each level or even separate zones for each room in the home, depending on the arrangement of your home and your individual wants and needs.

No matter how many zones you have, each one will have its own thermostat that solely measures and regulates the temperature in that zone. This allows you to control the temperature for that zone while leaving the rest of the building alone. Because all of the air still flows via the same air handler and ductwork, the only true constraint is that you cannot have one zone set to heating and another set to cooling.

Dampers, which are simply metal plates situated inside the ductwork that may open and close to control the airflow to each zone, are used in a multi-zone system. This is analogous to blocking the supply vents to a specific room or area of the house. The distinction is that sealing the vents simply prevents air from entering that room, implying that air will continue to flow via that branch of the ductwork.

This is normally not suggested since it might cause a pressure imbalance inside the HVAC system, affecting how well and efficiently the system warms or cools. It also puts more strain on your furnace or air conditioner, which can contribute to premature wear and tear on its components. Multi-zone HVAC systems are not affected by pressure imbalance difficulties since the damper cuts off the whole branch or section of ductwork rather than merely turning it off at the end of the branch as happens when the vents are closed.

Despite the fact that each zone has its own thermostat, the system is still managed by a single central control panel. When a zone requires hot or cold air, the control panel opens the damper for that zone, allowing air to flow to the zone. When the zone achieves the desired temperature, the thermostat in that zone sends a signal to the central control panel, which closes the zone's damper. Each zone may have a single damper or many dampers that govern the airflow to that zone, depending on the size and structure of the zones.

Application of Markov Decision Process (MDP) in HVAC Control Issue

In this subsection, we present a Markov Choice Cycle (MDP) detailing the multi-zone Business/Market air conditioning the board issue that will be tended to in Area 3 utilizing a without model profound RL-based approach. The current interior temperature, according to the enhanced warm elements model of air conditioning in [26], is only associated with previous state boundaries, such as the indoor temperature at the previous time span, and is independent of the indoor temperature at certain other time spans. Since climate control may be represented as a restricted Markov process, the RL method might be used to find optimal configurations.

A controlled stochastic process with the Markov property with costs assigned to state changes is referred to as a Markov decision process. A Markov decision problem consists of a Markov decision process and a performance criterion. A policy, mapping states to actions, that (possibly stochastically) determines state transitions to reduce the cost according to the performance criterion is a solution to a Markov decision problem. Markov decision problems (MDPs) provide the theoretical underpinnings for decision-theoretic planning, reinforcement learning, and other sequential decision-making tasks of interest to artificial intelligence and operations research academics and practitioners. MDPs use dynamical models based on well-understood stochastic processes and performance criteria based on known theory in operations research, economics, combinatorial optimized performance, and the social sciences. It appears that MDPs have a unique structure that can be used to accelerate their resolution. In investment planning, for example, the beginning state is frequently known with certainty (the present price of a stock or commodity), limiting the range of likely achievable states (future values) and possible investment plans in the near-term future.

In general, MDPs have inherent properties such as time, action, and reachability in state space that can be used to create efficient algorithms for solving them. It is critical that we grasp the computational challenges involved in these sources of structure in order to understand the potential for efficient sequential and parallel algorithms that compute both exact and approximate answers. The dynamics of an agent interacting with a stochastic environment are described by a Markov decision process. The Markov decision process describes the following development of the

system state over a (potentially infinite) sequence of times referred to as the stages of the process, given a starting state or distribution of states and a sequence of actions. This study focuses on the infinite-horizon scenario, when the stage sequence is unlimited.

The four fundamental parts of a MDP are the state (s), the activity (a), the probability of changing to another state (p), and the award (r). These four elements are at play when a multi-zone HVAC control issue arises in a commercial or industrial setting:

The user's minimum acceptable temperature, $T_{lower}(t)$, should be reported together with the current outside temperature, Tout(t), and the current inside temperature, T_{in},z(t), for each zone z. Fourth, the retail price is the retail(t) in this time step.

Keep in mind that the user's minimum acceptable degree of discomfort is a time-dependent component of the condition. This is because we expect that HVAC occupants have varying comfort needs throughout the day. This makes sense, since the interior temperature's comfort range may be adjusted to save money on utility bills when no one is home throughout the day. When the home is used in the evenings and on weekends, the temperature may be returned to a more comfortable setting.

In order to achieve the preheating impact of HVAC, the current retail price is also included in the state parameters. Setting the air conditioning framework's setpoint to a moderately large number while the retail cost of energy is low takes into consideration early warming of within climate, saving money on energy costs that would otherwise be incurred as the temperature outside drops.

$$\text{HVAC} = \begin{cases} 1, if\ Tin\ (t) < setpoint - deadband \\ 0, if\ Tin\ (t) > setpoint \\ remain\ at\ the\ current\ status, elsewise \end{cases} \quad (1)$$

In this review, we exclusively center on the warming utilization of the central air model. The dead band in Eq. (1) is the temperature range outside of which the thermostat will not switch between the on and off positions, hence preventing rapid cycling. In Eq. (1), when the interior temperature is above the user-specified comfort level, we can see that the HVAC system is turned on when the temperature drops below that level.

- Reward: During the control period, the cost of energy usage is added to the cost of comfort deterioration, which is defined as follows:

$$r(t) = -\omega_c \sum_{t=t-\Delta t}^{t} \lambda^{retail}(t)\, E_{HVAC}(t) - \sum_{t=t-\Delta t}^{t} c^{penalty}(t) \quad (2)$$

The HVAC system's energy cost is represented by the first component in Eq. (2), where $\lambda^{retail}(t)$, $E_{HAVC}(t)$, and t, respectively, denote the Δt control interval, power consumption, and retail price; the subsequent term addresses the punishment for client solace infringement, still up in the air as follows.

$$c^{penalty}(t) = \begin{cases} 1, for\ T_{in}\ (t) < T_{lower}\ (t) - T_{in} \\ 0, elsewise \end{cases} \quad (3)$$

T_{th} has a low value, which is a threshold used in Eq. (3). If the magnitude of the temperature deviation is less than T_{th}, it is disregarded. The dead band in the HVAC system makes it difficult to maintain the desired interior temperature at all times. There's wiggle room for the thermostat inside the house because of the threshold.

Since the prize incorporates the energy cost and the punishment, the multi-objective limit is provoked by relegating various loads to the two goals, proposed by ω_c and ω_p in Eq. (2). The reduction of r(t) is a clear goal of air conditioning warm control, which is the cost of energy consumed in addition to the penalty incurred throughout the control cycle, to the bare minimum: r(t): ΣNT t=1r(t). As a result, a multi-step decision-making dilemma arises from the need for a long-term control approach to forestall the effects of unknowable future conditions.

Considering the abovementioned, this study utilizes the sans model profound RL way to deal with managing the imperceptibility inborn in the multi-zone Business/ Market air conditioning control issue. To be effective, the sans model RL approach requires no earlier information on the climate or state changes. It learns from the results of its decisions and exchanges information with its surroundings over time to improve its decision-making process. In this manner, mistakes made in forecasting due to unknown variables, as well as those made in measuring the thermal mass of a structure, may be avoided. In the following paragraphs, we will go into further depth about the deep RL approach.

DDPG-BASED MULTI-ZONE HVAC CONTROL STRATEGY

Deep deterministic policy gradient (DDPG) is an RL algorithm that integrates deep learning with both policy-based and value-based approaches. It is particularly effective in multi-zone HVAC management, where it improves energy performance, maintains ideal comfort conditions, and adapts to environmental variations. Section 3.1. insights the various methods in deep reinforcement learning.

Methods in Deep Reinforcement Learning

The RL approach is an AI procedure for further developing MDP dynamic systems. The MDP-specified reward serves as an input for the RL algorithm's evolutionary process. If the reward for continuing in the present direction is high enough, the algorithm will continue to look there, and vice versa. When there are time limitations or a hidden state space involved in a decision, the RL approach excels.

The two most prevalent types are policy-based and value-based RL techniques. The approaches to activity assessment are where the two methodologies diverge. Esteem

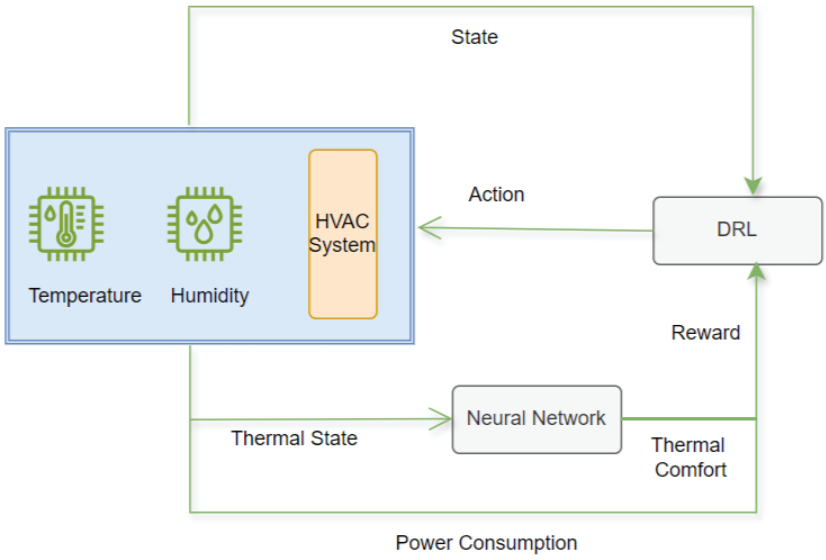

Figure 3. The flow of the building control system using DNN for Thermal comfort and DRL for thermal control.

based RL assesses the Q worth of a state-movement pair (s, a), which is the all-out compelled reward starting from performing activity at state s, though procedure-based RL produces probabilities of all potential activities at the ongoing status and chooses the activity with the most noteworthy probability.

The main RL strategy utilizes a DNN combined with RL. Figure 3 depicts a flow of the building control system using a deep neural network for thermal comfort and Deep Reinforcement learning for thermal control. In significant RL, the DNN is utilized as a backtracking device for foreseeing the Q regard (in accordance with the worth-based RL strategy) or the development likelihood (in accordance with the method-based RL approach). Figure 4 depicts a typical DNN layout for RL regression.

To perform significant level control for extremely confounded circumstances, for example, those with persistent state space or activity space, without the plain limits is the major advantage of the profound RL approach over the customary RL technique. In contrast to traditional Q realizing, which uses an actual Q table to record all possible activity levels, advanced RL creates a more rounded relapse model. On account of constant control, this summed up relapse model gives more strong and versatile strategies against obscure circumstances. Straight Support will act as a substitute for esteemed based profound RL strategies, and the DPG approach will act as a representation of strategy based profound RL techniques. Then, at that point, the DPG strategy, a ceaseless control technique that consolidates the previously mentioned approaches, will be described in depth as a means of determining an ideal multi-zone Commercial/Market HVAC management strategy.

DDPG for Constant Temperature and Humidity Regulation in a Building

The DDPG (Deep Deterministic Policy Gradient) technique was specifically designed to address challenges involving continuous variables. Unlike the Straight Assistance or DPG methods, where the Deep Neural Network (DNN) generates Q-values or probabilities for all possible actions

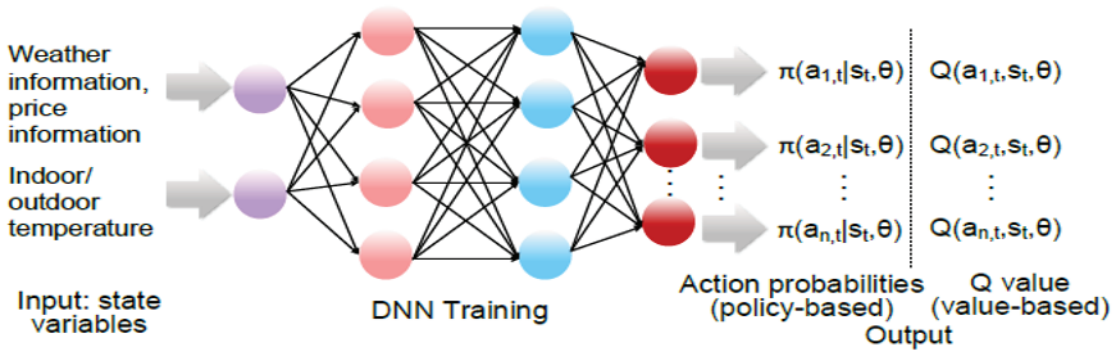

Figure 4. Approximating functions in RL using a DNN framework.

for the agent to choose from, the term "deterministic" in DDPG highlights that the DNN produces a single, definitive output rather than a set of options. This unique characteristic allows the action space to be defined continuously, facilitating effective handling of continuous control tasks. A key advantage of DDPG is its integration of Linear Reinforcement with DPG. The Deep Deterministic Policy Gradient (DDPG) algorithm is a reinforcement learning method that employs two neural networks working together to solve continuous action space problems. The functions they serve are described below.

The entertainer network gets the present status and plays out a deterministic activity; the current status and the action made by the performer network are dealt with into the savant association, in which the value of the state-action pair, Q, is sent. Utilizing this new Q esteem, the entertainer organization's settings will be changed. The entertainer organization's misfortune capability is the greatest squared blunder (MSE) of the Q esteem, as per DPG rationale, while the pundit organization's misfortune capability is the mean squared mistake (MSE) of the Q esteem. Taking everything into account, the entertainer network is answerable for choosing acts, while the pundit network assesses the picked activity [26-30].

As illustrated in Figure 5, the algorithm initially acquires state information for the external environment, such as temperature and retail price value. The program also receives a task ID from the external environment, which is a 0-1 binary variable that indicates whether the situation is cooling or heating. The task ID is an important indicator of the current task on which the actor is working. The state parameters will then be normalized. Normalization is required because the state characteristics of the two jobs can differ greatly [31-32]. For example, in the cooling scenario, the outdoor temperature is substantially higher than in the heating scenario. Data that is not normalized might cause algorithm divergence [33]. The concatenation of the normalized state parameters with the task ID is then delivered to the deep neural networks. The DDPG method employs two types of neural networks: the actor-network and the critic network. The actor-network generates HVAC control actions, while the critic network computes the Q value as an evaluation of the chosen action. There is also a behaviour network and a target network for both the actor-network and the critic network. The control action is produced by the behavior network, while the target network generates a target value for the behavior network to learn, which is analogous to labelled data in supervised learning. The target network aids in the stabilization of the training process [34,35]. The DDPG algorithm has four neural networks in total. Actor Network, Critic Network, Target Actor Network, and Target Critic Network. To improve action selection, the actor network updates itself using policy gradients based on state inputs. The critic network calculates

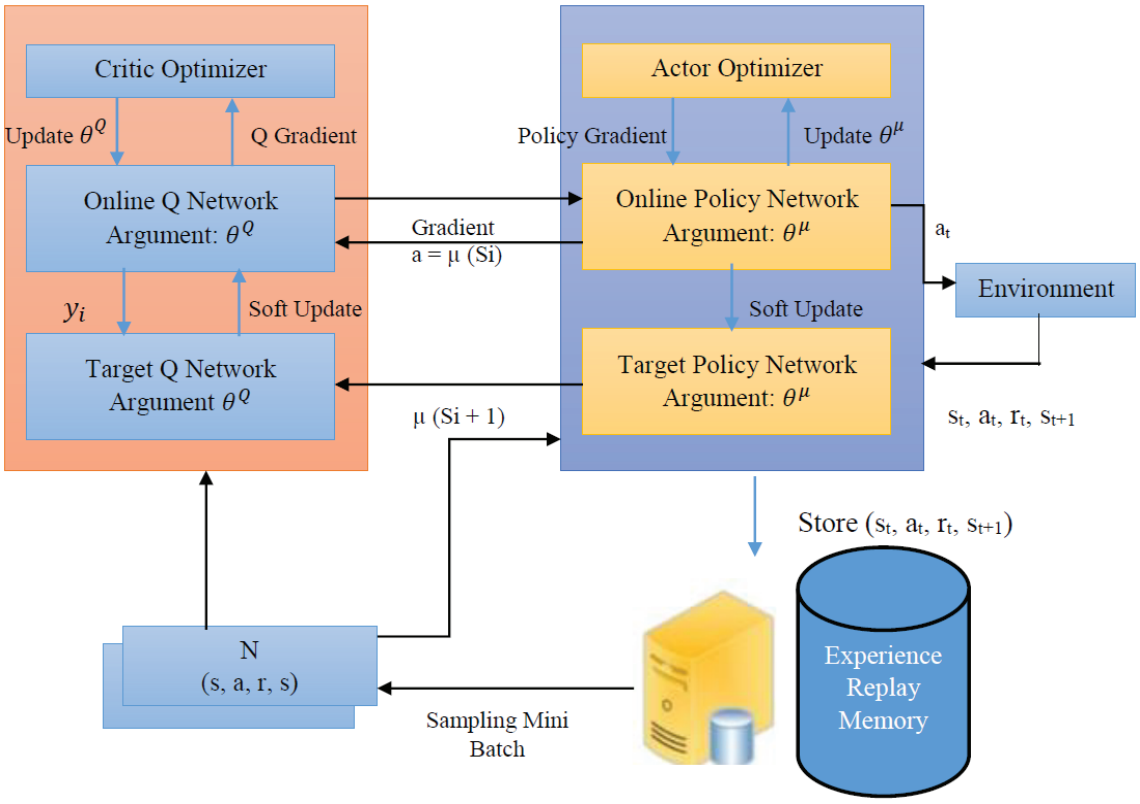

Figure 5. Multi-task DDPG for multi-zone HVAC control.

the Q-value, representing the expected reward for a specific state-action pair [36-38]. Designed for stability, the target actor network is a delayed version of the main actor network with slow updates. The target critic network acts as a backup of the primary critic, updating gradually to prevent fluctuations in Q-values.

Similar to the Straight Support approach, the Deep Deterministic Policy Gradient (DDPG) algorithm utilizes two types of neural networks: the actor network (policy network) and the critic network (Q-value network). Each of these networks is paired with a corresponding target network, resulting in a total of four neural networks. The inclusion of target networks helps stabilize the training process by providing more consistent and reliable updates, which improves the convergence of the algorithm.

The DDPG calculation is depicted in additional profundity in the following segment

Algorithm DDPG for Optimal HVAC Control Design

Similar to Linear Reinforcement, an actor-network is constructed using the DDPG method to choose a deterministic activity. The DDPG calculation utilized is made sense of more meticulously underneath. Before initiating their respective target networks, we randomly initialize the actor network and the critic network, both of which are neural networks [39]. The system state is initialized at the start of each cycle, and the current actor-network is used as the setpoint for selecting an HVAC control action [40]. In order to encourage the algorithm to further investigate the chosen action, some noise is included [41]. During the full t period, the specified action is carried out in the environment while the outcome states and rewards are monitored. To train an algorithm, the state transition (s(t), Setptz(t), r(t), s(t + t)) is recorded in a replay buffer. Once enough transitions have been gathered, a random subset of transitions is used to update the parameters of the actor-network and the behavior network. The assumption of independence and uniform distribution in the learning model may be preserved by the use of random selection, which can sever temporal connections between transitions. The efficiency with which the transitions may be used is further improved by the fact that they can be sampled several times [42].

In response to the loss functions, adjustments are made to the Q and of the neural network. The mean squared mistake (MSE) between the objective Q respect and the genuine Q respect is utilized to portray the blunder-making ability of the predominant scholarly union. The objective Q esteem is the sum of the continuing prize and a limited Q esteem from the objective pundit organization "Q," where is the contrast error, for the new control stretch "t + t," where is the discount variable [40]. Subsequent to deciding the misfortune capability, the Q conduct pundit organization's boundaries are changed utilizing the angle. The pace of learning is denoted by ηQ.

The actor network's loss function is designed to optimize the quality factor (Q):

$$\max \frac{1}{M} \sum_{i=1}^{M} Q\left(s^{(i)}(t); a^{(i)}(t), \theta^{Q}\right) \mid a^{(i)}(t) = \pi s^{(i)}(t); \theta^{\pi}) \quad (5)$$

The actor-network Q(s;θ^{Q}) is used to provide the solution an (i)(t) in Eq. The chain rule is utilized to work out the Q worth's slope according to. While the conducting network is continually being refreshed, the objective pundit organization and target entertainer organization's boundaries, θ^{Q} and θ^{π}, are changed at a slower pace. This more gradual update serves to strengthen the reliability of the learning process. The actor network is updated using policy gradients, influenced by the critic's Q-values. The critic network is updated using the TD error, learning to predict accurate Q-values. Both networks gradually improve through these updates, with target networks stabilizing the process. Through this iterative process, both the actor and critic networks become better at selecting actions and evaluating them, leading to improved decision-making and long-term rewards [43,44].

PROBLEM FORMULATION

Energy management in smart homes has become increasingly important due to rising energy costs and the need for sustainable practices. This research focuses on two primary objectives: cost minimization and efficiency in energy management. By leveraging time-of-day tariffs and various optimization algorithms, this study aims to develop strategies that reduce energy costs and enhance efficiency without compromising the residents' comfort.

Rating Performance

Here, we prove the benefits of the DDPG method by simulating it with real-world data and comparing it to the linear reinforcement-based discrete control method and the benchmark cases. This is done so as to show that the DDPG method is the superior choice for multizone Commercial/ Market HVAC systems.

Modelling And Simulation Setting

Real-world weather data is used for training and testing in a two-zone house HVAC model from 1/1/2021 to 31/12/2021 for 8 different zones (TOUT, ITUZ1, ITUZ2, TCWZ2, TCH2, TDICZ1, TIECZ1, and TDXZ1). The goal of using such a dynamic pricing sequence is to test the deep RL agent's ability to decode market signals and adapt its control tactics accordingly. It is also believed that the user's minimum acceptable degree of comfort fluctuates 144 times every day.

The RL agent's control interval is set at 10 minutes Δt = 10.

Since we only care about how the HVAC system affects heating, we chose the November weather data for training. Each training session is considered one episode. Each episode will provide 24 transitions of the form:

$$(s^{(i)}(t), \text{Setpt (i) z (t)}, r^{(i)}(t), s^{(i)}(t + \Delta t)) \quad (6)$$

Table 1. The DDPG and linear RL algorithms both make use of a DNN-based framework

Algorithm	DDPG		Linear RL
	Critic network	Actor network	
	Air Conditioner	1500	6
Input Size	(1,7)	(1,5)	(1,5)
Number of Secret Levels	2	2	2
the thickness of each covert layer	(7,20), (20,10)	(5,20), (20,10)	(5,20), (20,10)
Quantity of results	(1)	(2)	(25)
The secret layer's activation function	ReLU	ReLU	ReLU
Rate of learning (η)	0.001	0.01	0.01
Payoff significance	$\omega_{c:}$ 10, ω_p: 1		

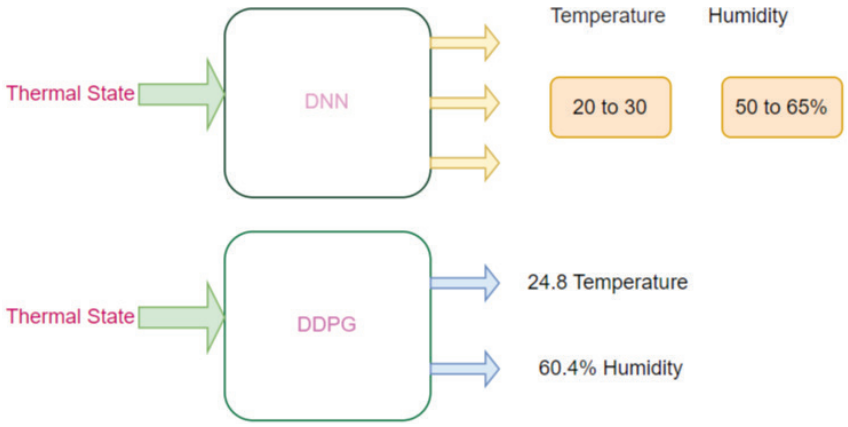

Figure 6. A comparison of the thermal control strategies, DDPG and DNN.

The RL agent learns from 10,000 simulated events. After the RL agent has been trained, it will be used in further tests under varying environmental circumstances.

Deep RL's DNN Architecture Design

The DDPG's actor and critic network architecture is laid out in full in Table 1. Linear reinforcement learning's architecture is included for reference, as well. The present configurations of the DDPG model and the linear reinforcement learning model are the best of many potential outcomes gained via trial and error.

The scalar estimated Q value is output by the critic network in the DDPG method, whereas the actor network outputs the setpoint for each zone after receiving the vector of state variables and the vector of action variables as inputs.

Temperature and humidity are continuous variables, hence DDPG can be used for continuous action control directly, whereas DNN requires action space discretization. Although the setpoint is a consistent variable, there is dependably a scope for it to keep clients agreeable. Accordingly, the entertainer organization's result layer utilizes tanh as the actuation capability, restricting the result to the stretch [-1,1].

$$Setpt_z = T_{lower} + \Delta T \cdot (y_{out} + 1) \quad (7)$$

Yields the genuine setpoint, where y_{out} is the entertainer organization's result and ΔT is the set point's upper reach. ΔT is set to 2oC in the reenactment. As a result, the DDPG RL model chose a value for the setpoint that is between $[T_{lower}, T_{lower} + 2]$.

The outputs of the Linear RL algorithm are identical to the inputs. With a stage size of 0.5 ^{0}C, we discretize the setpoint space to meet the requirements of Linear RL's discrete action space. This means the 2-zone HVAC system has 25 possible configurations, with 5 actions each zone. Figure 6 shows a comparison of the thermal control strategies, DDPG and DNN. Temperature and humidity are continuous variables, hence DDPG can be used for continuous action control directly, whereas DNN requires action space discretization. A vector of 25 Q values, each of which represents a unique set of actions, is what you get when you run Linear RL.

Effectiveness of Real-Time Climate Control

Real-time climate control refers to the dynamic adjustment of heating, ventilation, and air conditioning (HVAC) systems based on live data from environmental sensors, occupancy patterns, and external weather conditions.

THE DDPG AND RL LINEAR ALGORITHMS HAVE REACHED CONVERGENCE

The average performance improvements achieved after each training session were evaluated for both the deep deterministic policy gradient (DDPG) algorithm and the linear reinforcement learning (linear RL) approach.

```
def train(env, ENV_NAME, training_steps):
nb_actions = env.action_space.shape[0]
agent = build_agent(env.action_space.shape[0], env.observation_space
agent.compile(Adam(1r=.001, clipnorm=1.), metrics=['mae''])
agent.fit(env, nb_steps=training_steps, visualize=False, verbose=1, |
agent.save_weights('results/weights/ddpg_{} weights.h5f".format(ENV_|
agent.test(env, nb_episodes=1, visualize=False, nb_max_episode_steps
def test(env, ENV_NAME , num_episodesf):
nb_actions = env.action_space.shape[0]
agent = build_agent(envjaction_space.shape[0], env.observation_space
agent.compile(Adam(1r=.901, clipnorm=1.), metrics=['mae"])
agent.load_weights('results/weights/ddpg_{}_weights.h5f".format(ENV_I
```

Linear-based approaches and linear reinforcement models generally refer to traditional control methods or simpler reinforcement learning (RL) techniques where policies, value functions, or both are modeled using linear functions. These approaches are commonly used when the system dynamics are relatively simple. For DDPG, the gains demonstrated a non-linear progression, characterized by significant jumps in performance as the model fine-tuned its policy through exploration and exploitation of the action space. In contrast, the linear RL approach exhibited a more steady and incremental improvement pattern, reflecting the constraints of its simpler policy structure and limited adaptability. This contrast underscores the differences in learning dynamics and capability between the two methodologies. In the first few episodes, returns seem to be on average higher than in the last few. This happens because an arbitrary training day is chosen for each episode. On days with moderate exterior temperatures, the energy cost and penalty of exercise may be relatively low, and vice versa. As training progresses, however, more episodes are added, decreasing the average return. However, the DDPG RL approach often yields greater returns than the Linear RL approach. The linear RL model's output is higher than the DDPG RL models, and after 10,000 episodes, the combination of actions has not been thoroughly explored, resulting in a lower average return. The algorithm used for the present research work is as follows:

Figure 7 shows the convergence of Linear RL and DDPG RL model for all episodes of different zones. From Figure 7 it is observed that the convergence of the Linear RL model is much less as compared to the DDPG model whose converge is 100 % of all episodes.

Computational per unit efficiency

Using the real-world data the DDPG RL agent is trained and then used for 20 test days in January 2021 to provide the best possible HVAC management plan. The whole examination process takes no more than eight minutes. The software utilizes the free and open-source TensorFlow framework for deep learning and is developed in Python 3.6. The platform consists of a laptop equipped with 16.00 GB of RAM and a 2.8 GHz Intel®CoreTM i7- 7600U CPU.

Figure 8 shows the Per Unit Efficiency of the DDPG RL model for Different zones such as TDECZ1, TIECZ1, and TDXZ1. The light green color in the above figure shows the higher Per-Unit Efficiency for most of the episodes which lies between 1.3 to 1.4.

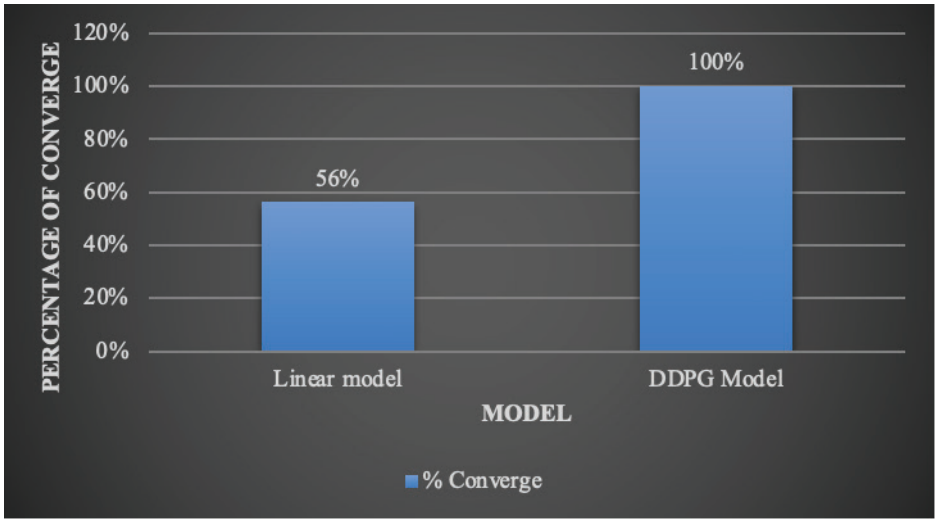

Figure 7. Converge of linear RL and DDPG RL model.

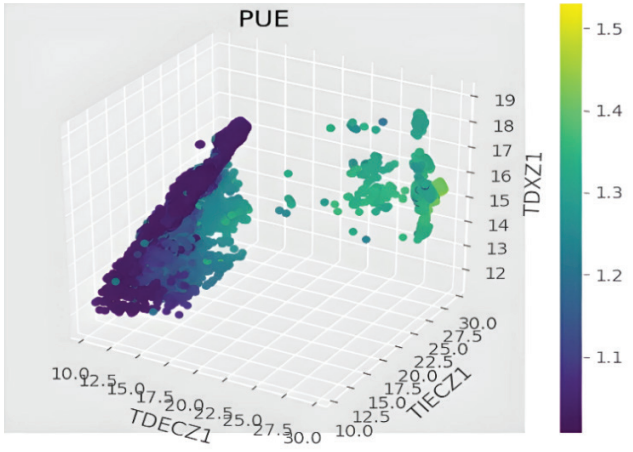

Figure 8. Per unit efficiency of DDPG for different zones.

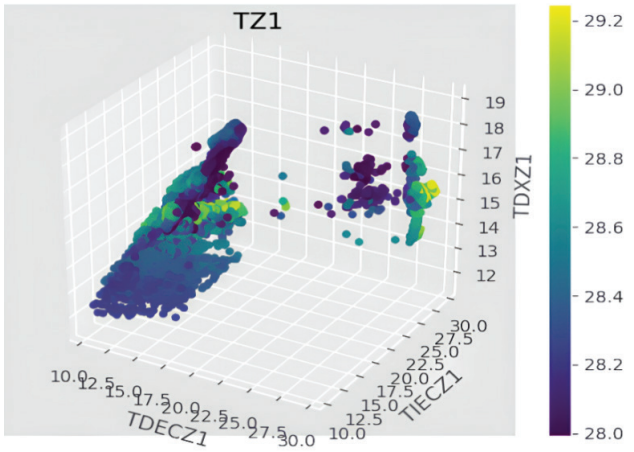

Figure 10. Temperature of TZ1.

Comparison of the DDPG with the linear RL model

After step of convergence, we count the mean number of steps required for converge of the episode and rewards. The mean number of steps taken by linear is 115.4 which is very higher that steps taken by DDPG model and it's also time consuming too. Mean steps taken by DDPG model is only 9.9. The mean and maximum step taken by both the model also mentioned in above Table 2. Figure 9 shows that the proper difference between numbers of steps taken by each selected model.

Temperature of selected zone

The temperature range in Zone 1 and Zone 2 after the application of a designed model for controlling the temperature of Commercial/Market buildings is shown in Figure 10 and Figure 11. In zone one control temperature lies between 28.4 °C to 28.8 °C, in zone 2 this temperature lies between 20 °C to 30 °C. This minimization of temperature is possible by the DDPG RL algorithm model.

Table 2. Comparison of the DDPG model with linear RL model

Type of model	Total samples	% Converge	Mean (Steps)	Min (Steps)	Max (Steps)
Linear model	10,000	56%	115.4	5.0	200.0
DDPG model	10,000	100%	9.9	3.0	20.0

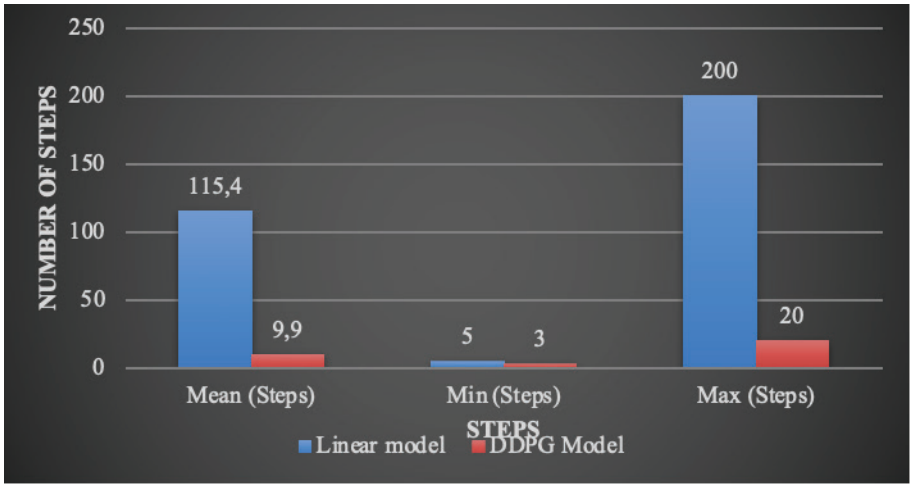

Figure 9. Mean step required for convergence of sample for both the models.

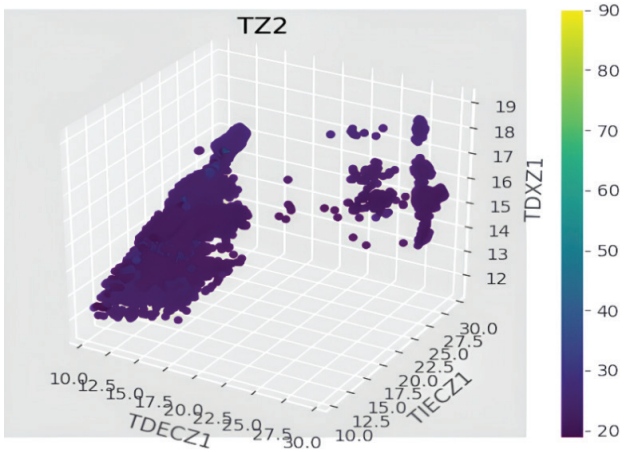

Figure 11. Temperature of TZ2.

CONCLUSION

In this study, the DDPG RL technique is used to regulate a commercial/market HVAC system with many zones to reduce energy costs without sacrificing comfort. Using DNNs, the DDPG can keep the temperature and humidity in the building at a constant level. The simulation results demonstrate that a well-trained DDPG RL expert can act wisely to adjust the various improvement targets and that it can also acquire speculation and flexibility to concealed climate, recommending its true capacity for future web-based applications in tackling MDP issues with stowed away data or with consistent pursuit space. The RL-based control method's stability will be further enhanced by research on two main areas in the future:

- The deep RL agent must first master the ability to adjust to a variety of seasonal conditions by automatically switching between cooling and heating modes of operation in order to provide HVAC customers with cost-effective management techniques over a year-long period;
- The deep RL specialist should have the option to learn a set point plan that takes client preferences into account more in order to provide HVAC management strategies that are more flexible. We can increase the deep RL agent's adaptability and resistance to the uncertainties that arise in real-world applications by exploring these two directions.
- The study highlights the potential of DRL, particularly the DDPG approach, in revolutionizing air conditioning control systems. While the linear reinforcement method serves as a valuable baseline, the superior performance of DDPG underscores the importance of advanced learning techniques in tackling complex, dynamic control problems. This research contributes to the growing body of knowledge aimed at developing energy-efficient and user-centric AC control solutions.

REFERENCES

[1] U.S. Dept. Energy. 2011 Buildings Energy Data Book. Available at: https://ieer.org/wp/wp-content/uploads/2012/03/DOE-2011-Buildings-Energy-DataBook-BEDB.pdf. Accessed on 17 Mar 2026.

[2] Erdinc O, Tascıkaraoglu A, Paterakis NG, Eren Y, Catalao JPS. End-User Comfort-Oriented Day Ahead Planning for Responsive Residential Hvac Demand Aggregation Considering Weather Forecasts. IEEE Xplore 2016;8:362–372. [CrossRef]

[3] Hao H, Corbin CD, Kalsi K, Pratt RG. Transactive Control of Commercial Buildings for Demand Response. IEEE Xplore 2016;32:774–783. [CrossRef]

[4] Silver D, Huang A, Maddison CJ, Guez A, Sifre L, Lanctot M, et al. Mastering the Game of Go with Deep Neural Networks and Tree Search. Nature 2016;529:484–489. [CrossRef]

[5] Huang T, Guo Q, Sun H. Distributed Computing Platform Supporting Power System Security Knowledge Discovery Based on Online Simulation. IEEE Xplore 2017;8:1513–1524. [CrossRef]

[6] Hua H, Qin Y, Hao C, Cao J. Optimal energy management strategies for energy Internet via deep reinforcement learning approach. Appl Energy 2019;239:598–609. [CrossRef]

[7] Rocchetta R, Bellani L, Compare M, Zio EE, Patelli E. A reinforcement learning framework for optimal operation and maintenance of power grids. Appl Energy 219;241:291–301. [CrossRef]

[8] Ma Y, Kelman A, Daly A, Borelli F. Predictive Control for Energy Efficient Buildings with Thermal Storage: Modeling, Stimulation, and Experiments. IEEE Xplore 2012;32:44–64. [CrossRef]

[9] Dounis AI, Caraiscos C. Advanced Control Systems Engineering for Energy and Comfort Management in a Building Environment–A Review. Renew Sustain Energy Rev 2009;13:1246–1261. [CrossRef]

[10] Aswani A, Master N, Taneja J, Culler D, Tomlin C. Reducing Transient and Steady State Electricity Consumption in Hvac Using Learning-Based Model-Predictive Control. IEEE Xplore 2012;100:240–253. [CrossRef]

[11] Manuel A-C, Tejero-Gonzalez A, Illera-Riesgo J, Salins SS. Integrated evaluation of energy and water use in air-cooled and adiabatic condensers – a case study in an office building. Energy Nexus 2026;22:100706. [CrossRef]

[12] Razzano G, Brandi S, Piscitelli MS, Capozzoli A. Rule extraction from deep reinforcement learning controller and comparative analysis with ASHRAE control sequences for the optimal management of Heating, Ventilation, and Air Conditioning (HVAC) systems in multizone buildings. Appl Energy 2025;381:125046. [CrossRef]

[13] Tianshu W, Wang Y, Zhu Q. Deep Reinforcement Learning for Building HVAC Control. Proceedings of the 54th Annual Design Automation Conference. New York: Association for Computing Machinery; 2017. p. 1–6. [CrossRef]
[14] Deng X, Zhang Y, He Q. Towards optimal HVAC control in non-stationary building environments combining active change detection and deep reinforcement learning. Build Environ 2022;211:108680.
[15] Dobbs JR, Hencey BM. Model Predictive Hvac Control with Online Occupancy Model. Energy Build 2014;82:675–684. [CrossRef]
[16] Mnih V, Kavukcuoglu K, Silver D, Rusu AA, Veness J, Graves A, et al. Human-Level Control Through Deep Reinforcement Learning. Nature 2015;518:529–533. [CrossRef]
[17] Kuruganti T, Kou X, Li F, Dong J, Starke M, Munk J. A distributed energy management approach for residential demand response. IEEE Xplore 2019. [CrossRef]
[18] Lin Y, Barooah P, Mathieu JL. Ancillary services through demand scheduling and control of commercial buildings, IEEE Xplore 2016;32:186–197. [CrossRef]
[19] Li F, Du Y. From AlphaGo to Power System AI: What Engineers Can Learn from Solving the Most Complex Board Game. IEEE Xplore 2018;16:76–84. [CrossRef]
[20] Wu Y, Tan H, Peng J, Zhang H, He H. Deep reinforcement learning of energy management with continuous control strategy and traffic information for a series-parallel plug-in hybrid electric bus. Appl Energy 2019;247:454–66. [CrossRef]
[21] Kou P, Liang D, Wang C, Wu Z, Gao L. Safe deep reinforcement learning-based constrained optimal control scheme for active distribution networks. Appl Energy 2020;264:1–12. [CrossRef]
[22] Mocanu E, Mocanu DC, Nguyen PH, Liotta A, Webber ME, Gibescu M. On-line building energy optimization using deep reinforcement learning. IEEE Xplore 2018;10:3698–3708. [CrossRef]
[23] Ahn KU, Park CS. Application of deep Q-networks for model-free optimal control balancing between different HVAC systems. Sci Technol Built Environ 2019;26:61–74. [CrossRef]
[24] Zhang Z, Chong A, Pan Y, Zhang C, Lam KP. Whole building energy model for HVAC optimal control: A practical framework based on deep reinforcement learning. Energy Build 2019;199:472–490. [CrossRef]
[25] Vrancx P, Ruelens F, Claessens BJ. Convolutional neural networks for automatic state-time feature extraction in reinforcement learning applied to residential load control. IEEE Xplore 2016;9:3259–3269. [CrossRef]
[26] Wang Y, Velswamy K, Huang B. A long-short term memory recurrent neural network-based reinforcement learning controller for office heating ventilation and air conditioning systems. Processes 2017;5:46–63. [CrossRef]
[27] Jiang T, Zou Y, Yu L. Online energy management for a sustainable smart home with an HVAC load and random occupancy. IEEE Xplore 2017;10:1646–1659. [CrossRef]
[28] Hu G, Spanos CJ, Ma K. Energy management considering load operations and forecast errors with application to HVAC systems. IEEE Xplore 2016;9:605–614. [CrossRef]
[29] Lu N. An evaluation of the HVAC load potential for providing load balancing service. IEEE Xplore 2012;3:1263–1270. [CrossRef]
[30] Sharma VK. Secret image scrambling and dwt-based image steganography using smoothing operation and convolution neural networks. J Discret Math Sci Cryptogr 2023;26:695–705. [CrossRef]
[31] Markad K, Lal A. Thermal post buckling analysis of smart SMA hybrid sandwich composite plate. Polym and Polym Compos 2021. [CrossRef]
[32] Cui B, Munk J, Jackson RK, Fugate DL, Starke M. Building thermal model development of typical house in US for virtual storage control of aggregated building loads based on limited available information. In: 30th International Conference on Efficiency, Cost, Optimisation, Simulation and Environmental Impact of Energy Systems; California, USA. 2017.
[33] Bulut M. The impact of the establishment of the cross-border balancing market on the integration of RES into the regional electricity grid. Int J Energy Stud 2024;309–329. [CrossRef]
[34] Lal A, Markad K. Static and dynamic nonlinear stability analysis of hybrid sandwich composite beam under variable inplane loads. J Mech Sci Technol 2021;35:3895–3908. [CrossRef]
[35] Pogaku N, Prodanovic M, Green TC. Modeling, analysis and testing of autonomous operation of an inverter-based microgrid, IEEE Xplore 2007;22:613–625. [CrossRef]
[36] Wu H, Liu X, Ding M. Dynamic economic dispatch of a micro-grid: Mathematical models and solution algorithm. Int J Elect Power Energy Sys 2014;63:336–346. [CrossRef]
[37] Choi J, Shin Y, Choi M, Park W, Lee W. Robust control of a micro-grid energy storage system using various approaches, IEEE Xplore 2018;2702–2712. [CrossRef]
[38] Kanif M, Vivek P, Kishor K. Experimental and Numerical Investigation of Nano-material based Structural Composite. Curved and Layered Structures 2024;11. [CrossRef]
[39] Gu S, Lillicrap T, Sutskever I, Levine S. Continuous deep Q-learning with model-based acceleration. Proceedings of the 33rd International Conference on Machine Learning; New York, USA. 2016.

[40] Kingma DP, Ba J. Adam: A method for stochastic optimization. arXiv 2015;1–15.

[41] Kuznetsova E, Li Y, Ruiz C, Zio E, Ault G, Bell K. Reinforcement learning for microgrid energy management. Energy 2013;59:133–146. [CrossRef]

[42] Lal A, Markad K. Influence of Dynamic Temperature Variation and Inplane Varying Loads over Post-Buckling and Free Vibration Analysis of Sandwich Composite Beam. Int J Comput Mater Sci Eng 2020;S2047684120500128. [CrossRef]

[43] Salpakari J, Lund P. Optimal and rule-based control strategies for energy flexibility in buildings with pv. Appl Energy 2016;161:425–436. [CrossRef]

[44] Rajabi H, Hu Z, Ding X, Pan S, Du W, Cerpa A. Modes: Multisensor occupancy data-driven estimation system for smart buildings. Proceedings of the Thirteenth ACM International Conference on Future Energy Systems; New York, USA. 2022;228–239. [CrossRef]