Journal of Data Acquisition and Processing
ISSN: 1004-9037
https://sjcjycl.cn/
DOI: 10.5281/zenodo.777912

MULTI-ZONE COMMERCIAL/MARKET HVAC CONTROL STRATEGY BASED ON REINFORCEMENT LEARNING ALGORITHM MODELS

Ganesh Murade[1*], Bhanu Pratap Soni[2], Ankit Kumar Sharma[3]
[1*,3]Dept. of EE, University of Engineering & Management, Jaipur, Rajasthan
[2]SEEE, Fiji National University, Fiji
***Corresponding Author: - Ganesh Murade**
[*]Dept. of EE, University of Engineering & Management, Jaipur, Rajasthan

Abstract:
The recent increase of renewable energy technology in the building industry is predicted to cut fossil fuel usage while increasing the complexity of heating, ventilation, and air conditioning (HVAC) system design and control. Due to population increase and ongoing efforts to raise living standards, research into energy saving and sustainability has grown popular. In both commercial and commercial buildings, the installation of central HVAC systems is generally on the rise. The HVAC systems in commercial buildings have been looked at as a potential demand response resource. The complexity of creating warm-powerful models and the vulnerability related with both occupant driven heat loads and climatic projections make it far from easy to improve commercial air cooling across the board. In this research, we develop a perfect control philosophy for a commercial cooling system with several zones using a sharp sans model substantial RL technique termed the profound deterministic arrangement slope (DDPG), determined to diminish energy costs without adversely influencing solace levels for building tenants. Through unsupervised, ongoing contact with a simulated building environment, the utilized deep RL-based method gathers information. When the DDPG-based HVAC control strategy is compared to the linear-based HVAC control strategy, the converge may be reduced by 56%, and when the DDPG-based HVAC control strategy is compared to the linear reinforcement model, the converge may be reduced by 15%. Also mean steps required for DDPG RL model and Linear RL model is 9.9 and 115.3 respectively.

Keywords: Actor-critic learning, deep deterministic policy gradient (DDPG), deepreinforcement learning (deep RL), multi-zone commercial HVAC. Linear RL, Reinforcement Learning.

1. INTRODUCTION

It is estimated that 41% of India's yearly energy consumption [1] comes from the building sector. This is why the issue of temperature regulation in buildings has received so much focus [2] [3] in recent times. The primary objective is to find a happy medium between low energy use and high levels of comfort for office workers. The actions of occupants and a lack of information regarding their feelings of comfort introduce uncertainty into the warm elements of consumed constructing spaces [4, 5] and the method involved with deciding the best control execution standards. A squeezing new issue is the means by which to integrate these points of interest into the creation interaction. However, the occupants' building is a highly layered, stochastic system with continuous and discrete states, and the building's dispersed state detectability is complicated by the sensors' varying requirements. As a result, developing a

scaleable control design framework is essential. Propelled by late improvements in RL [6] [7], the essential objective of this study is to survey the adequacy of RL in tenant intelligent structure temperature control issues.

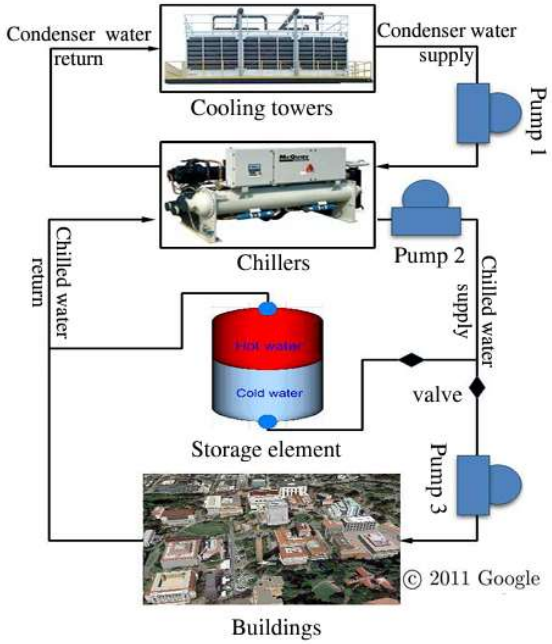

Figure 1: Schematic of a cooling system. [2]

The air conditioning framework as shown in figure 1 is currently the most involved device for keeping a structure at an agreeable temperature. Through effective demand-side energy management measures, it also helps reduce peak loads and stabilize system-wide functioning [8]. Numerous papers have been written on how to best optimize HVAC control systems to maximize efficiency.

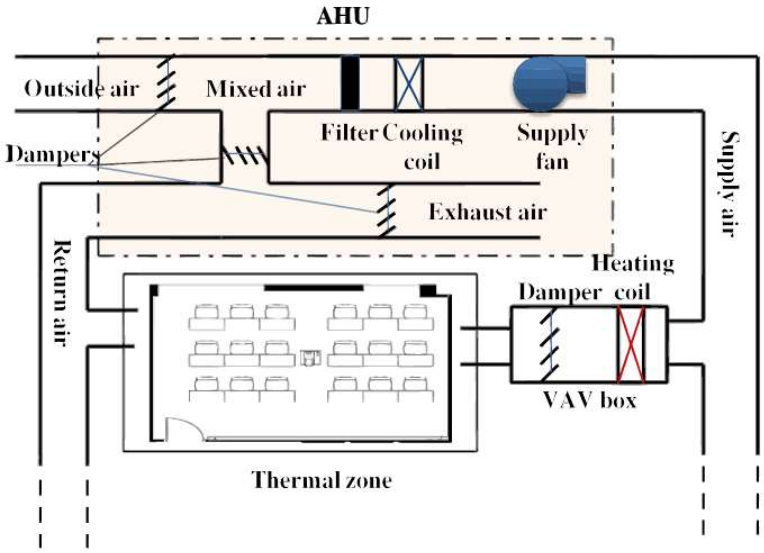

Figure 2: Schematic of the HVAC air distribution system.[33]

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. Using an adaptive control system, Sampath et al. [30] managed the building's HVAC system to maximize both occupant thermal comfort and the building's energy efficiency. Giuseppe et al. [31] proposed a model prescient control and hereditary calculation based streamlining structure to diminish warming energy utilization and related inconvenience. Wei and others [33] introduced an information driven technique for controlling air conditioning frameworks with changing air volume in view of profound support learning. Deng et al. [34] 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 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 right 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 mean time, 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 to the field of force frameworks 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 include 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 [32]. 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, the linear reinforcement is used in [25].

All of these studies show that using deep RL techniques to optimise the HVAC temperature management strategy is more successful than using custom-built benchmarks. In order 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:

- We have 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 used 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.
- We show that 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.

The structure of present research paper are as above: In Section II, we present the detailing of the air conditioning control issue; We briefly discuss the two typical deep RL methods, the linear reinforcement and DDPG methods, in Section III; The DDPG method's simulation results are presented in Section IV, plus comparison with the linear reinforcement; and in Section V, we draw conclusions and end the paper.

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

2.1 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.

2.1.1 How Does A Multi-Zone HVAC System Work?

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.

2.2 The Markov Decision Process (MDP) is being used to the HVAC control issue.

In this subsection, we present a Markov Choice Cycle (MDP) detailing of 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 Markov decision process is a controlled stochastic process satisfying the Markov property with costs as signed to state transitions. A Markov decision problem is a Markov decision process together with a performance criterion. A solution to a Markov decision problem is a policy, mapping states to actions, that (perhaps stochastically) determines state transitions to minimize the cost according to the performance criterion. Markov decision problems (MDPs) provide the theoretical foundations for decision-theoretic planning, reinforcement learning, and other sequential decision-making tasks of interest to researchers and practitioners in artificial intelligence and operations research. MDPs employ dynamical models based on well-understood stochastic processes and performance criteria based on established theory in operations research, economics, combinatorial optimization, and the social sciences. It would seem that MDPs exhibit special structure that might be exploited to expedite their solution. In investment planning, for example, often the initial state is known with certainty (the current price for a stock or commodity) and as a result the set of likely reachable states (future prices) and viable investment strategies in the near-term future is considerably restricted.

In general, notions of time, action, and reachability in state space are inherent characteristics of MDPs that might be exploited to produce efficient algorithms for solving them. It is important that we understand the computational issues involved in these sources of structure to get some idea of the prospects for efficient sequential and parallel algorithms for computing both exact and approximate solutions. A Markov decision process describes the dynamics of an agent interacting with a stochastic environment. Given an initial state or distribution over states and a sequence of actions, the Markov decision process de scribes the subsequent evolution of the system state over a (possibly infinite) sequence of times referred to as the

stages of the process. This paper focuses on the infinite-horizon case, in which the sequence of stages in infinite.
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, Tlower(t), should be reported together with the current outside temperature, Tout(t), and the current inside temperature, Tin,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.

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

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:

$$r(t)=-\omega_c\sum_{t'=t-\Delta t}^{t}\lambda^{retail}(t')E_{HVAC}(t')-\omega_p\sum_{t'=t-\Delta t}^{t}c^{penalty}(t') \quad (2)$$

The HVAC system's energy cost is represented by the first component in Eq. (2), where λ retail(t), EHAVC(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.

$$c^{penalty}(t')=\begin{cases}1, for\ T_{in}(t')<T_{lower}(t')-T_{th}\\0, elsewise\end{cases} \quad (3)$$

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, multi-objective limit is provoked by relegating various loads to the two goals, proposed by ωc and ωp 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 of 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 manage 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.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.

3. A DDPG-BASED MULTI-ZONE HVAC CONTROL STRATEGY

3.1 Methods in deep reinforcement learning: a quick overview

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 based RL assesses the Q worth of a state-movement pair (s,a), which is the all out compelled reward starting from performing activity an 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. 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 3 depicts a typical DNN layout for RL regression.

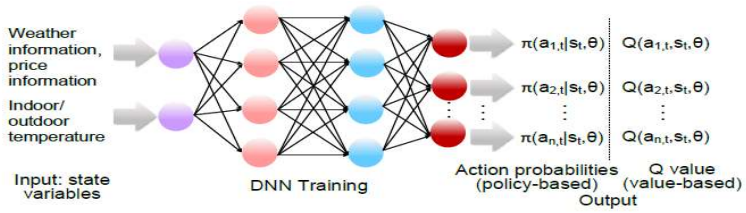

Figure 3: Approximating functions in RL using a DNN framework.

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 realising, 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.

3.3 The DDPG allows for constant temperature and humidity regulation in a building

1) The DDPG: A Brief Overview

The DDPG technique was developed for the exclusive purpose of dealing with issues involving continuous variables. Rather than the Straight Assistance or DPG, where the DNN creates the Q values or development probabilities of all potential exercises for the master to pick, the designation "deterministic" in the DDPG indicates that there is only one output from the DNN, which is not fixed in stone. Since there is just one such device, the action space may be continuously defined.

The fact that the DDPG combines Linear Reinforcement with the DPG is an additional benefit. The DDPG utilizes two brain organizations: the entertainer organization, which is responsible for assembling the DPG, and the critic network, which is in charge of putting together the linear reinforcement. The roles they play are outlined 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 (Direct Support). Taking everything into account, the entertainer network is answerable for choosing acts, while the pundit network assesses the picked activity.

The algorithm first obtains state information for the external environment, such as temperature and retail price value, as shown in Fig. 4. In addition to the state information, the algorithm

receives a job ID from the external environment, which is a 0-1 binary variable indicating whether the situation is cooling or heating. The task ID is a crucial indicator of the current task that the actor is working on. The state parameters will then be normalized. Normalization is required because the state characteristics of the two jobs can differ greatly. For example, in the cooling scenario, the outdoor temperature is substantially higher than in the heating scenario. Data that is not normalised might cause algorithm divergence. The concatenation of the normalised 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 behaviour network, while the target network generates a target value for the behaviour network to learn, which is analogous to labelled data in supervised learning. The target network aids in the stabilization of the training process. The DDPG algorithm has four neural networks in total. Figure 2 also shows the structure of each neural network.

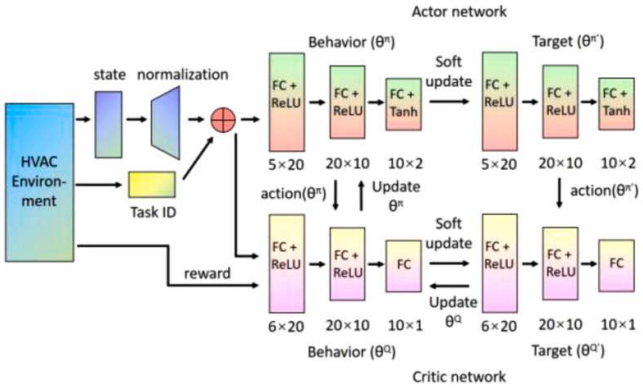

Figure 4. Multi-task DDPG for multi-zone HVAC control

What's more, like the Straight Support approach, two brain organizations — an objective organization and a conduct organization — are built for both the entertainer organization and the pundit network in the DDPG. This implies there are a sum of four brain organizations. Utilizing the objective organization assists keep the calculation's union with steadying. The DDPG calculation is depicted in additional profundity in the following segment.

2) Algorithm DDPG for Optimal HVAC Control Design

In [27], we find the foundations for the specialised version of DDPG that we propose here. 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.

Prior to initiating their respective target networks, we randomly initialize the actor network and the critic network, both of which are neural networks. 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. In order to encourage the algorithm to further investigate the chosen action, some noise is included.

During the full t period, the specified action is carried out in the environment while the outcome states and rewards are monitored. For the purpose of training 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 are used to update the parameters of the actor network and the behaviour 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.

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. 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):

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

The actor network □(*s*;□□) is used to provide the solution a(i)(t) in Eq. The chain rule is utilized to work out the Q worth's slope according to.While the conduct network is continually being refreshed, the objective pundit organization and target entertainer organization's boundaries,□□□ and □□, are changed at a slower pace. This more gradual update serves to strengthen the reliability of the learning process.

4. 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.

4.1 Modelling and simulation setting

Real-world weather data is used for training and testing in a two-zone house HVAC model [28] 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), Setpt(i) z (t), r(i)(t), s(i)(t + Δt)). 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.

4.2 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.

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

Algorithm	**DDPG**		**Linear RL**
	critic network	**actor network**	
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	ω_c : 10, ω_p : 1		

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. Despite the fact that the setpoint is a consistent variable, there is dependably a scope of it by and by 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]. Setptz = T_{lower} + ΔT·(y_{out} + 1), 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 ºC, 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. A vector of 25 Q values, each of which represents a unique set of actions, is what you get when you run Linear RL.

4.3 Effectiveness of Real-Time Climate Control

1) The DDPG and Linear RL algorithms have reached convergence.

In this section, we provide the average gains made after each training session in the DDPG and the linear RL. 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 does the Linear RL approach. The linear RL model's output is higher than the DDPG RL model's, and after 10,000 episodes, the combination of actions has not been thoroughly explored, resulting in a lower average return. The algorithm used for present research work are as follows:

```
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
agent.test(env, nb_episodes=num_episodes, visualize=False, nb_max_ep
if __name__ == "_main__":
```

Figure 4 shows that the Converge of Linear RL and DDPG RL model for all episodes of different zones. From the figure 5 it is observed that the Converge of Linear RL model is very less as compared to DDPG model who's Converge is 100 % all episodes.

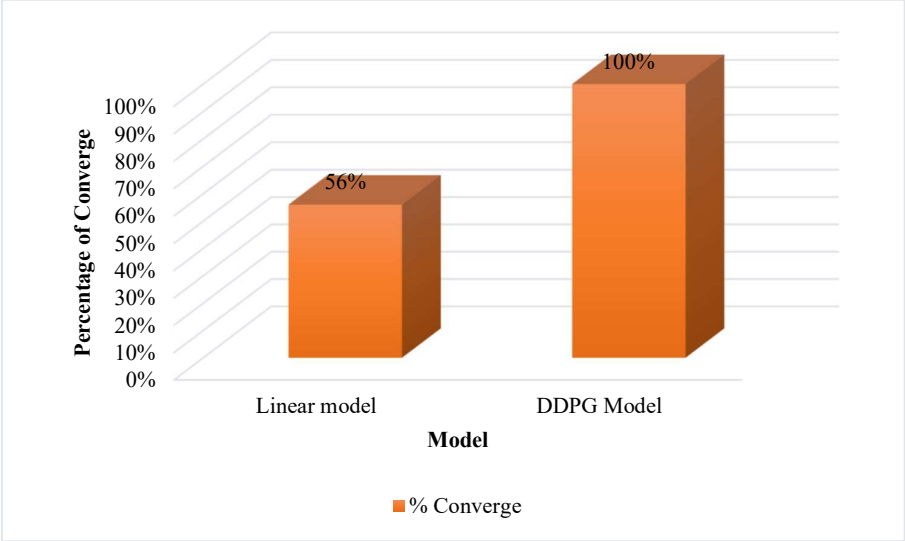

Figure 5: Converge of Linear RL and DDPG RL model

2) Computational Per Unit Efficiency

Using the real-world data from [34], the DDPG RL agent is trained and then used to 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.

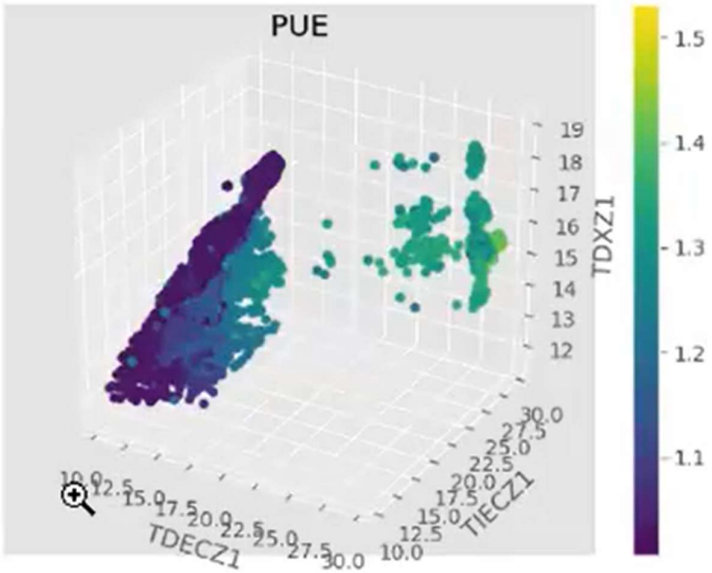

Figure 6: Per Unit efficiency of DDPG for different zones

Above figure 6 shows that the Per Unit Efficiency of DDPG RL model for Different zones such as TDECZ1, TIECZ1 and TDXZ1. The light green color in above figure 3 shows that the higher Per Unit Efficiency for most of the episodes which is lies in between 1.3 to 1.4.

3) Comparison of the DDPG with the Linear RL 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

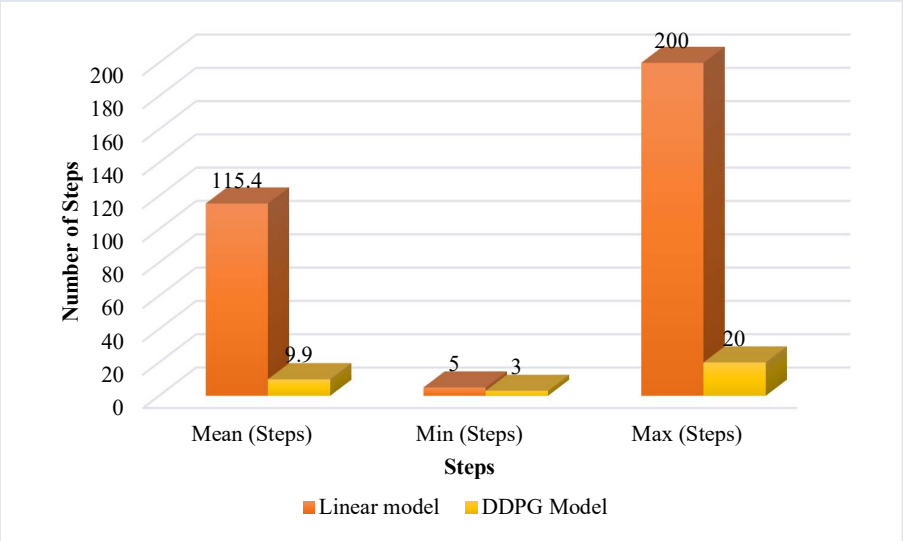

Figure 7: Mean step required for Convergence of Sample for both the models

After step of Convergence we count the mean number of step required for converge of the episode and rewards. The mean number of step 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 7 shows that the proper difference between numbers of steps taken by each selected model.

4) Temperature of selected zone

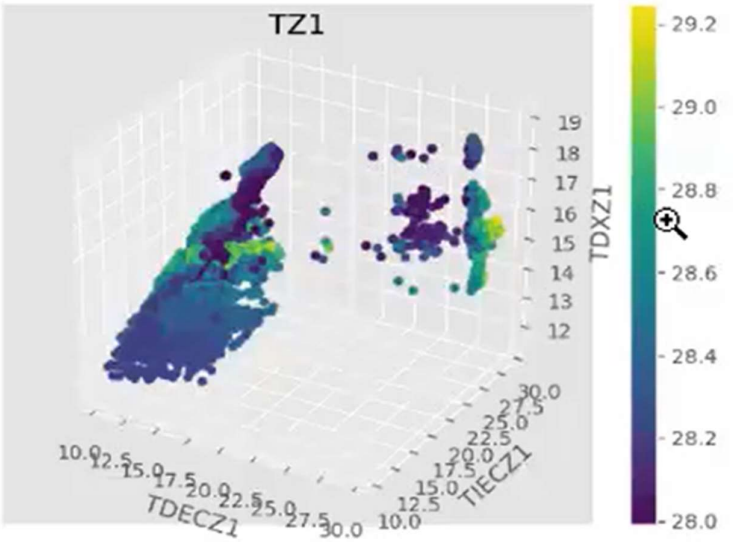

Figure 8: Temperature of TZ1

Temperature range in Zone 1 and Zone 2 after the application of designed model for controlling temperature of Commercial/Market building as shown in Figure 7 and Figure 9. In zone one control temperature lies between 28.4 ℃ to 28.8℃, besides in zone 2 this temperature lies in between 20 ℃ to 30 ℃. This minimization of temperature possible by DDPG RL algorithm model.

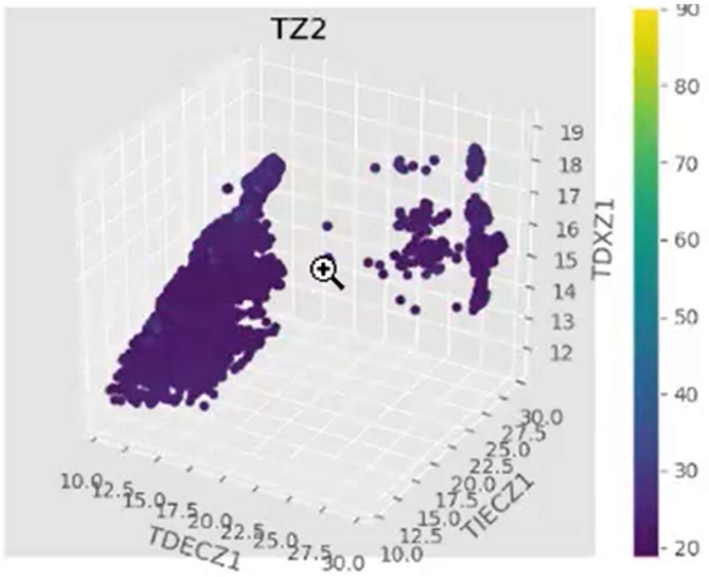

Figure 9: Temperature of TZ2

3. CONCLUSION

In this study, the DDPG RL technique is used to regulate a commercial/market HVAC system with many zones in order 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:
1) 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;
2) 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.

REFERENCES

1. U.S. Dept. Energy, 2011 Buildings Energy Data Book, Washington, DC,USA, 2011.
2. Y. Ma and A. Kelman., "Predictive control for energy efficient buildings with thermal storage," *IEEE Control system magazine,* vol. 32, p. 44–64, 2012.
3. A. I. Dounis and C. Caraiscos., "Advanced control systems engineering for energy and comfort management in a building environment–A review," *Renewable and Sustainable Energy Reviews,* vol. 13, p. 1246–1261, 2009.
4. A. Aswani and N. Master., "Reducing transient and steady state electricity consumption in hvac using learning-based model-predictive control," *Proceedings of the IEEE,,* vol. 100, p. 240–253, 2012.
5. J. R. Dobbs and B. M. Hencey., "Model predictive hvac control with online occupancy model," *Energy and Buildings,* vol. 82, p. 675–684, 2014.
6. V. Mnih and K. Kavukcuoglu et al., "Human-level control through deep reinforcement learning," *Nature,* vol. 518, p. 529–533, 2015.
7. D. Silver and A. Huang et al., "Mastering the game of go with deep neural networks and tree search," *nature,* vol. 529, p. 484–489, 2016.
8. Kuruganti T and Zandi H., "A distributed energy management approach for residential demand response," in *In: 2019 3rd International Conference on Smart Grid and Smart Cities (ICSGSC).*, Berkeley, California, US;, 2019.
9. Hu G and Spanos CJ., "Energy management considering load operations and forecast errors with application to HVAC systems," *IEEE Trans Smart Grid,* vol. 9, pp. 605-14., 2016.
10. Erdinc O et al., "End-user comfort oriented dayahead planning for responsive residential HVAC demand aggregation considering weather forecasts.," *IEEE Trans Smart Grid,* vol. 8, pp. 362-72, 2016.

11. Barooah P and Mathieu JL., "Ancillary services through demand scheduling and control of commercial buildings," *IEEE Trans Power System,* vol. 32, pp. 186-97, 2016.
12. Jiang T and Zou Y., "Online energy management for a sustainable smart home with an HVAC load and random occupancy," *IEEE Trans Smart Grid,* vol. 10, pp. 1646-59, 2017.
13. Hao H et al., "Transactive control of commercial buildings for demand response," *IEEE Trans Power System,* vol. 32, pp. 774-83, 2016.
14. Silver D and Huang A. et al., "Mastering the game of go with deep neural networks and tree search," *Nature,* vol. 529, pp. 484-9, 2016.
15. F. Li and Y. Du., "From AlphaGo to Power System AI: What Engineers Can Learn from Solving the Most Complex Board Game," *IEEE Power and Energy Magazine,* vol. 16, pp. 76-84, 2018.
16. G. Q. a. S. H. Huang T, "A Distributed Computing Platform Supporting Power System Security Knowledge Discovery Based on Online Simulation," *IEEE Trans Smart Grid,* vol. 8, pp. 1513-24, 2017.
17. Wu Y and Tan 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,* vol. 247, pp. 454-66, 2019.
18. Hua H et al., "Optimal energy management strategies for energy Internet via deep reinforcement learning approach," *Appl Energy,* pp. 598-609, 2019.
19. Rocchetta R and Bellani L. et al., "A reinforcement learning framework for optimal operation and maintenance of power grids," *Appl Energy,* vol. 241, pp. 291-301, 2019.
20. Kou P and Liang D., "Safe deep reinforcement learning-based constrained optimal control scheme for active distribution networks," *Appl Energy,* vol. 264, 2020.
21. Vrancx P and Ruelens F., "Convolutional neural networks for automatic state-time feature extraction in reinforcement learning applied to residential load control," *IEEE Trans Smart Grid,* vol. 9, pp. 3259-69, 2016.
22. Mocanu E and Mocanu DC et al., "On-line building energy optimization using deep reinforcement learning," *IEEE Trans Smart Grid,* vol. 10, pp. 3698-708., 2018.
23. Wang Y Velswamy K and Huang B., "A long-short term memory recurrent neural network based reinforcement learning controller for office heating ventilation and air conditioning systems," *Processes,* vol. 5, pp. 46-63, 2017.
24. Zhang Z and Chong A., "Whole building energy model for HVAC optimal control: A practical framework based on deep reinforcement learning.," *Energy and Buildings,* vol. 199, pp. 472-90, 2019.
25. Ahn KU and Park CS., "Application of deep Q-networks for model-free optimal control balancing between different HVAC systems.," *Science and Technology for the Built Environment,* vol. 26, pp. 61-74, 2019.
26. L. N., "An evaluation of the HVAC load potential for providing load balancing service," *IEEE Trans Smart Grid,* vol. 3, pp. 1263-70, 2012.

27. Lillicrap TP and Hunt JJ., "Continuous control with deep reinforcement learning," [Online]. Available: arXiv preprint arXiv:1509.02971.
28. Cui B and Munk J., "Building thermal model development of typical house in US for virtual storage control of aggregated building loads based on limited available information.," in *In: 30th International Conference on Efficiency, Cost, Optimisation, Simulation and Environmental Impact of Energy Systems*, San Diego, California, USA, 2017.
29. C. P. Research.. [Online]. Available: https://www.cleanpower.com/.
30. Salins S.S.; Kumar S.S.;."Performance Characterization of an Adaptive-Controlled Air Handling Unit to Achieve Thermal Comfort in Dubai Climate. Energy"**2023**, 273, 127186
31. Aruta G.; Ascione F.; "Optimizing Heating Operation via GA- and ANN-Based Model Predictive Control: Concept for a Real Nearly-Zero Energy Building." Energy Build. **2023**, 292, 113139.
32. Sutton, R.; Barto, A. Reinforcement Learning; MIT Press: Cambridge, MA, USA, 2022.
33. Wei, T.;Wang, Y.; Zhu, Q. Deep Reinforcement Learning for Building HVAC Control. In Proceedings of the 54th Annual Design Automation Conference 2017, Austin, TX, USA, 18 June 2017.
34. Deng, X.; Zhang, Y.; Qi, H. Towards Optimal HVAC Control in Non-Stationary Building Environments Combining Active Change Detection and Deep Reinforcement Learning. Build. Environ. **2022**, 211, 108680.