Journal of Building Engineering 50 (2022) 104165

Contents lists available at ScienceDirect

Journal of Building Engineering

journal homepage: www.elsevier.com/locate/jobe

# Applications of reinforcement learning for building energy efficiency control: A review

Qiming Fu [a,b,1], Zhicong Han [a,b,1], Jianping Chen [b,c,*], You Lu [a,b], Hongjie Wu [a], Yunzhe Wang [a,b]

[a] School of Electronic and Information Engineering, SuZhou University of Science and Technology, Suzhou, Jiangsu, 215009, China
[b] Jiangsu Province Key Laboratory of Intelligent Building Energy Efficiency, Suzhou University of Science and Technology, Suzhou, Jiangsu, 215009, China
[c] School of Architecture and Urban Planning, SuZhou University of Science and Technology, Suzhou, Jiangsu, 215009, China

ARTICLE INFO

*Keywords:*
Reinforcement learning
Intelligent buildings
Energy consumption

ABSTRACT

The wide variety of smart devices equipped in modern intelligent buildings and the increasing comfort requirements of occupants for the environment make the control of intelligent buildings important and complex. Reinforcement learning, as a class of control techniques in machine learning, has been explored for its potential in the field of intelligent building control. Reinforcement learning methods applied to intelligent buildings can effectively reduce energy consumption. In this paper, we classify reinforcement learning algorithms and analyze the control problems that each algorithm is suitable for solving. In addition, we review the reinforcement learning methods applied to control and manage buildings, outline the problems and future directions of reinforcement learning applications in intelligent buildings, and give our suggestions for researchers who want to use reinforcement learning methods to solve control problems in this field.

## 1. Introduction

Building energy consumption has dramatically increased due to climate issues, people's increasing functional demand of buildings and so on. Building energy consumption contributes to the largest part of energy consumption around the world, which is as high as 40% in U.S. and E.U [1]. Building energy consumption mainly comes from Heating, Ventilation and Air Conditioning (HVAC) systems, water heating equipment, lifts and lighting systems, etc. [2]. These systems consume large amounts of energy to meet the occupants' usage needs, which not only increases energy costs but also increases greenhouse gas emissions, so it is urgent to develop effective building energy control strategies. Nowadays, modern intelligent buildings are often designed to save energy while ensuring comfort and safety. To achieve this goal, buildings should be properly designed, constructed and controlled with suitable building materials and intelligent facilities which usually depend on the related intelligent control methods. Moreover, building energy consumption mainly comes from energy consumption equipments in intelligent buildings, and developing effective control strategies for these equipments can significantly save energy. Because of too many different kinds of building energy consumption equipments, there needs to consider many types of energy consumption control objects in intelligent buildings, and the control objects are often complex,

* Corresponding author. Jiangsu Province Key Laboratory of Intelligent Building Energy Efficiency, Suzhou University of Science and Technology, Suzhou, Jiangsu, 215009, China.
*E-mail address:* alan@usts.edu.cn (J. Chen).
[1] These authors have contributed to this work equally.

https://doi.org/10.1016/j.jobe.2022.104165
Received 18 October 2021; Received in revised form 24 January 2022; Accepted 1 February 2022
Available online 11 February 2022

which makes it not easy to develop an effective control strategy. Simple control strategies are easy to deploy, but the control effect is not good, while excellent control strategies often require more resources. Thus, developing buildings control strategies is becoming more and more necessary, and need to make adaptations to climate change, occupant demand and society (i.e., urban) development.

Many traditional building control methods have been explored by researchers to reduce energy consumption in buildings. The first one is rule-based feedback control, which is a kind of typical methods. This method can meet the requirements of the occupants and saves energy at the same time. However, it requires a predetermined plan for control, so the method is not flexible. The second one is prescriptive and feed-based reactive control. This method is simple and effective, but not optimal, since it does not consider the prediction information and fails to meet specific building and climate conditions [3]. The third one is model predictive control (MPC). MPC has been studied to reduce building energy consumption, such as [4] using MPC for demand response control, which can reduce energy consumption and save costs. But MPC requires an accurate model. The use of MPC for building control requires modeling for each specific building, but this modeling process is too costly and unavoidable at present, and there is no relevant commercial software that can easily derive a model for MPC design [5]. The fourth one, which is based on the optimal control theory, has also been tried to achieve energy savings in buildings [6]. However, this approach is demanding on the model, and overly pursues superior control performance at the same time, which may neglect some necessary practical factors (e.g., comfort, cost, etc.) in practical building applications. Neither can these control methods learn a control strategy suitable for the building environment on their own, nor can they adapt to changes in the status of individual building equipment during building use.

Different from these methods, reinforcement learning (RL) can learn a control strategy that is appropriate for the building environment based on information about various states of buildings. RL is a branch of machine learning. RL has an agent, which learns what to do to map situations to actions, and its goal is to maximize a cumulative numerical reward signal. As the agent interacts with the environment, it must discover which action obtains the most reward by trial and error. RL has been used in intelligent buildings and shows great potential in building energy consumption savings. Over the years, more and more RL methods have been successfully applied to solve building energy consumption problems. However, as RL is a specialized field in machine learning (which is a field in computer science), researchers in building energy management and control field may not be familiar with and articulate the specific characteristics and advantages of various RL methods. This paper thus provides a state-of-the-art review of RL methods, which discusses the advantages and disadvantages of different methods, and presents their applications in buildings energy management and control. The main contributions of this paper are:

- This paper classifies the algorithms of reinforcement learning and analyzes the suitability of each algorithm for problem solving.
- This paper reviews papers on the application of reinforcement learning in the field of intelligent buildings.
- This paper provides a reference for the subsequent selection of reinforcement learning algorithms to solve practical problems.

The outline of this paper is as follows: Section 2 describes reinforcement learning, Section 3 introduces the application of reinforcement learning in intelligent buildings, and Section 4 discusses the limitations of reinforcement learning in the field of intelligent buildings and the future directions. Section 5 concludes the paper.

## 2. Reinforcement learning

### 2.1. Markov decision process

Applying RL to intelligent buildings requires describing the environment as Markov Decision Processes (MDPs). MDP is a classical formalization of sequential decision making. MDP is a framework for learning from interactions to achieve goals. In this interactive learning process, the agent learns and makes decisions, and the rest of the things with which the agent interacts are called the environment. The agent chooses actions to act on the environment, which then provides the agent with new states by giving feedback on the agent's actions, rewarding the agent, and moving to the next state [7]. Fig. 1 can explain this progression more clearly.

A MDP can be defined as a tuple $< S,A,T,R,\gamma >$, where $S$ is the collection of environmental states, $A$ is the collection of rewards that the agent receives from the environment, $T$ is the probability that the agent takes an action $a \in A$ to move to the next state, and $R$ is the reward that the agent receives after the state transition, $\gamma$ is a discount factor, $\gamma \in [0,1]$. At the discrete time $t(t = 0,1,2,3,\ldots)$, the agent observes the state $s_t \in S$, and takes an action $a_t \in A$, receives a numerical reward $r_{t+1} \in R$, and the state $s_t \in S$ moves to a new state, $s_{t+1} \in S$. The interactive decision sequence can be represented as follows:

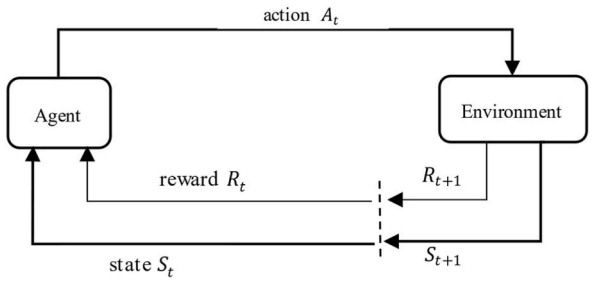


**Fig. 1.** Interaction between agent and environment in a Markov decision process.

$$S_0,\ A_0,\ R_1,\ S_1,\ A_1,\ R_2,\ S_2,\ A_2,\ R_3, \ldots \tag{1}$$

In a MDP, the transition probability determines the environment dynamics, which means the probability of each possible value for $S_t$ and $R_t$ only depends on the immediately preceding state and action, $s_{t-1}$ and $a_{t-1}$, and have no relation with the earlier states. Each state contains various information about the agent's interaction with the environment, we call this character as Markov property. If in a stochastic process, each state possesses Markov property, we call the stochastic process a Markov process. If the reward is considered in the Markov process, the process becomes a Markov reward process, and then the behavioral choices of individuals are taken into account, the process becomes a Markov decision process. In a MDP, each state has this property.

We use MDP to model the environment to solve real-world problems using RL methods, including the intelligent building scenario that we will present later. In modeling, we map the states, actions, and rewards of the environment in the real-world problems we need to solve to the definitions in the MDP, thus using RL methods to solve real-world problems in intelligent buildings.

### 2.2. Exploration and exploitation

In order to obtain the optimal strategy, the relationship between exploration and exploitation needs to be handled when solving RL problems, so that the maximum cumulative return can be obtained. Always adopting random strategies for exploration leads to experiencing a large number of uncertain strategies, which results in low cumulative returns obtained; always using the existing optimal strategies to decide on actions may lead to missing the global optimal strategies due to the lack of exploration of state space.

For the problem of exploration and exploitation in RL, we generally carry a stochastic strategy of greedy exploration, i.e., $\varepsilon$-greedy, where $\varepsilon \in [0, 1]$. $\varepsilon$ ensures that the agent has a probability of $1 - \varepsilon$ to select the highest value action among the existing optimal policies, while ensuring that there is a probability of $\varepsilon$ to randomly select actions to explore the state space, with a decaying as it continues to explore until a lower fixed exploration rate.

### 2.3. Policies and value functions

In RL, policy refers to the behavior of an agent, which is a mapping from states to actions. Through a policy $\pi$, the agent is able to decide itself to take actions in different states. The policy completely defines the behavior of the agent.

RL is essentially the maximization of the reward signal, so that can obtain an optimal policy. If only the instantaneous reward signal is maximized, it will result in choosing only the one with the largest reward from the action space each time, which becomes a greedy policy. Therefore, in order to characterize the current reward value as the maximum including the future (the maximum total reward from the current moment until the state reaches the goal), the value function is constructed to describe this variable. The expression is as follows:

$$v_\pi(s) = E_\pi\left[R_{t+1} + \gamma R_{t+2} + \gamma^2 R_{t+3} + \ldots \middle| S_t = s\right] \tag{2}$$

where $\gamma \in [0, 1]$ is the discount factor, which is to reduce the impact of future rewards on the current action. We call the function $v_\pi(S)$ the state-value function for policy $\pi$.

Similarly, we define the value of taking action $a$ in state $s$ under a policy $\pi$, denoted $q_\pi(s)$, as the expected return starting from s, taking the action a, and thereafter following policy $\pi$:

$$q_\pi(s, a) = E_\pi\left[R_{t+1} + \gamma R_{t+2} + \gamma^2 R_{t+3} + \ldots \middle| S_t = s, A_t = a\right] \tag{3}$$

Solving a RL task means figuring out a strategy that allows for substantial gains in the long run. The optimal policy $\pi_*$ is achieved by evaluating the optimal action-value function:

$$q_*(s, a) = max q_\pi(s, a) \tag{4}$$

Eq. (4) are usually solved by updating the Bell-man optimality equation:

$$q_*(s, a) = r(s, a) + \gamma \sum_{s' \in s} p(s' | s, a) \max_{a'} q_*(s', a') \tag{5}$$

Thus, RL algorithms can be classified into two classes, value-based and policy-based algorithms.

### 2.4. Algorithms of reinforcement learning

#### 2.4.1. Value-based algorithms

Value-based methods learn a policy according to value functions. Among traditional RL value-based methods, the representative is tabular-based methods. In these methods, the agent's policy is determined by its Q table, each action has its value in its Q table. In the table, state and action are used as two indicators, and the action in each state corresponds to a Q value in the table. One of the tabular-based methods is Q-learning. Q-learning was proposed by Watkins in 1992 [8], during the state $s \in S$, the agent takes an action $a \in A$, the agent receives a reward $r \in R$ from the environment, then updates the Q table based on Eq. (6).

$$Q(S_t, A_t) \leftarrow Q(S_t, A_t) + \alpha[R_{t+1} + \gamma \, max \, Q(S_{t+1}, a) - Q(S_t, A_t)] \tag{6}$$

Another tabular-based method is SARSA (State Action Reward State Action). In 1996, R.S. Sutton formally put forward the concept of SARSA [9], defined by:

$$Q(S_t, A_t) \leftarrow Q(S_t, A_t) + \alpha[R_{t+1} + \gamma \, Q(S_{t+1}, A_{t+1}) - Q(S_t, A_t)] \tag{7}$$

In Eq. (6) and Eq. (7), $\alpha \in [0, 1]$ is the learning rate, $\gamma \in [0, 1]$ is the discount factor.

Q-learning and SARSA are effective for problems with small and discrete state. However, when the number of states becomes much larger, the tabular-based methods are no longer suitable. To solve problems with the above characteristics, the Google-deepmind team proposed the Deep Q-Network algorithm [10] in 2015. Deep Q-Network (DQN) algorithm is an improvement of the Q-learning algorithm, it uses neural networks instead of Q table. DQN uses a convolutional neural network (CNN) to estimate the value function so that it can solve the problem of high dimensional problems and continuous space problems. Other value-based methods such as Double DQN, Prioritized-DQN, Dueling-DQN etc. are the extensions of DQN.

DQN has been applied to intelligent buildings by many literatures to reduce building energy consumption, here we explain the algorithmic principle of DQN, and the related research will be expressed in Chapter 3. DQN uses the experience replay mechanism in the training process. The transfer sample $e_t = (s_t, a_t, r_t, s_{t+1})$ obtained during the interaction between agent and environment, then the samples are stored in the playback memory unit $D = \{e_1, \ldots, e_t\}$ after a certain time step. For training, small batches of transfer samples are randomly selected at a time from $D$, and update the network parameters $\theta$ by using the stochastic gradient descent algorithm, so that can reduce the correlation of samples. The structure of DQN is shown in Fig. 2.

There are two networks in DQN, one is the Main network and the other is the Target network. They have the same structure but have different parameters. Main network outputs the current Q value, $Q(s,a|\theta_i)$, and Target network outputs target Q value, $Q(s,a|\theta_i^-)$. After each certain number of iterations, the parameters of the Main network are copied to the Target network, thus completing the learning process once. By minimizing the mean square error between the current Q value and the target Q value, we can update the parameters in networks. The loss function is as follows:

$$L(\theta_i) = E_{s,a,r,s'}\left[(Y_i - Q(s,a|\theta_i))^2\right] \tag{8}$$

Taking partial derivatives of the parameter $\theta$, the following gradient is obtained,

$$\nabla_{\theta_i} L(\theta_i) = E_{s,a,r,s'}\left[(Y_i - Q(s,a|\theta_i))\nabla_{\theta_i} Q(s,a|\theta_i)\right] \tag{9}$$

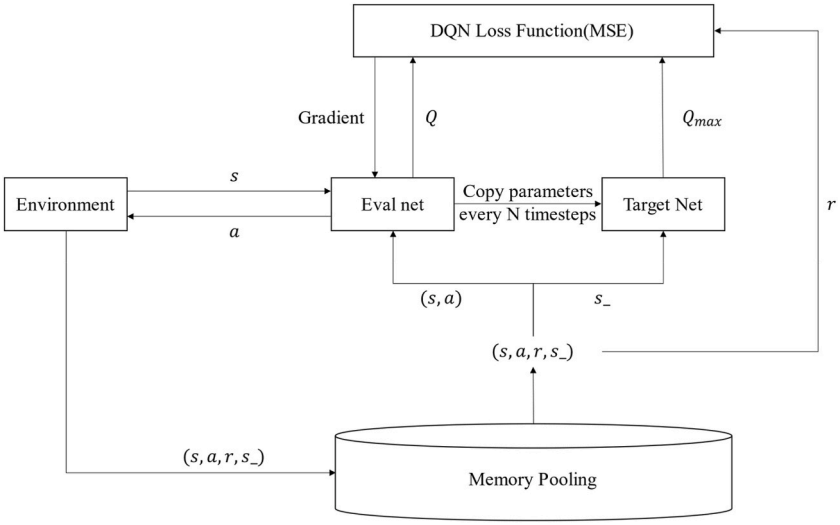


**Fig. 2.** The structure of DQN.

The way of updating the parameters after a certain number of iterations enables the target Q value to remain constant over a period of time, which reduces the correlation between the current Q value and the target Q value, thus improving the algorithm's stability.

For solving problems with discrete states and discrete action spaces, value-based methods can be competent, but for solving the problem of continuous state action space, policy-based methods perform better.

### 2.4.2. Policy-based algorithms

Policy-based methods learn policy directly instead of value functions, it has its own function approximator. Compared with value-based methods, policy-based methods have better convergence, can efficiently handle problems of large and continuous action spaces, and can learn stochastic policies. Policy-based methods updated its function approximator according to the gradient of expected reward, with respect to the policy parameters $\theta$.

$$J(\theta) = E_{\pi \sim p_\theta(\tau)}[r(\tau)] \tag{10}$$

where $r(\tau)$ is the total reward for the interaction between the agent and the environment in the trajectory $\tau$. $p_\theta(\tau)$ describes the probability of obtaining a particular $\tau$ from a dynamic environment at a fixed $\theta$. Thus, the method of finding the optimal $J$ can be translated into solving the maximization problem using gradient ascent for a set of parameters $\theta$.

because you are maximizing this fuctipon

Policy-based methods can be applied to solve continuous state RL problems, and have better convergence and stability. The three main categories include classical policy gradient [11], Trust Region Policy Optimization (TRPO) [12] and deterministic policy gradient (DPG) [13]. Classical strategy gradient updates the strategy parameters by calculating the total expected reward of the strategy, using the gradient about the strategy parameters, and eventually converges to the optimal strategy after multiple iterations.

The parameter update equation of the strategy gradient is as follows:

$$\theta_{new} = \theta_{old} + \alpha \nabla_\theta J \tag{11}$$

where $\alpha$ is the update step size and $J$ is the reward function. It is difficult to choose the appropriate update step size in the classical strategy gradient method, and the choice of step size directly affects the effect of strategy learning. Sutton et al. proposed the TPRO [11] method, which constrains the update step size to a certain range and decomposes the reward function of the new strategy into the reward function of the old strategy and the other terms. As long as the other terms in the new strategy are greater than or equal to zero, the reward function corresponding to the new strategy is guaranteed to be monotonically non-decreasing, and thus the strategy will not deteriorate.

Actor-critic (AC) algorithms combine value-based and policy-based methods to learn both value functions and policies. In recent years, some algorithms such as DPG [13] and deep deterministic policy gradient (DDPG) [14], Asynchronous Advantage Actor-Critic (A3C) [15] are representatives of the AC algorithms. The actor-critic algorithms have a small variance in the estimation of the value function, high sample utilization, and fast training speed of the algorithm.

The DDPG algorithm also has a wide range of applications in the papers we researched, mainly for solving problems with continuous state, action spaces. DDPG is a deep policy gradient algorithm based on the AC framework, where the AC framework is shown in Fig. 3.

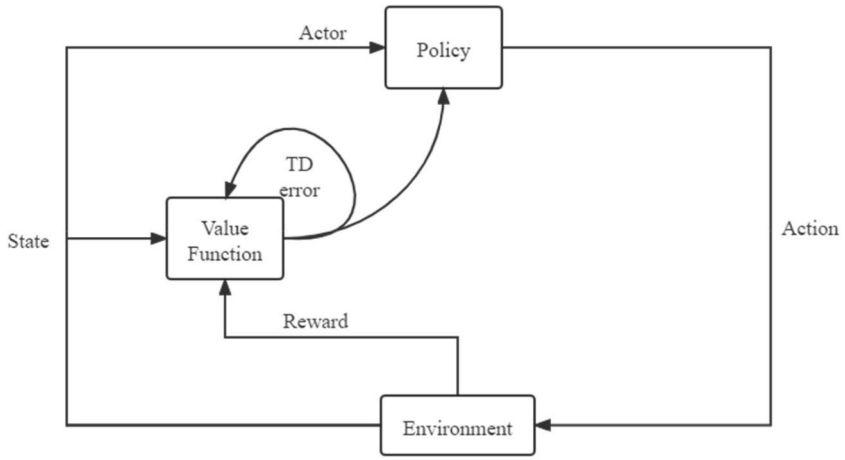


**Fig. 3.** The structure of Actor-Critic.

DDPG uses deep neural networks with parameters $\mu$ and Q to represent the deterministic policy $a = \pi(s|\theta^{\mu})$ and the value function $Q(s, a|\theta^{Q})$, respectively, and the objective function is defined as follows:

$$J(\theta^{\mu}) = E_{\theta^{\theta}}\left[r_1 + \gamma r_2 + \gamma^2 r_3 + \ldots\right] \tag{12}$$

It has been shown in Ref. [10] that the gradient of the objective function with respect to $\theta^{\mu}$ is equivalent to the expected gradient of the Q-value function with respect to $\theta^{\mu}$, and then according to the deterministic strategy $a = \pi(s|\theta^{\mu})$, stochastic gradient descent method is used to optimize the objective function.

$$\frac{\partial J(\theta^{\theta})}{\partial \theta^{\theta}} = E_s\left[\frac{\partial Q(s, a|\theta^{Q})}{\partial \theta^{\theta}}\right] = E_s\left[\frac{\partial Q(s, a|\theta^{Q})}{\partial_a}\frac{\partial_{\pi}(s|\theta^{\theta})}{\partial \theta^{\theta}}\right] \tag{13}$$

DDPG inherits the target network of DQN and uses the method of updating the value network in DQN to update the critic network, at which time the gradient information is as follows:

$$\frac{\partial L(\theta^{Q})}{\partial \theta^{Q}} = E_{s, a, r, s' \sim D}\left[\left(\left(r + \gamma Q\left(s', \pi(s'|\widehat{\theta}^{\mu})|\widehat{\theta}^{Q}\right)\right) - Q(s, a|\theta^{Q})\right)\frac{\partial Q(s, a|\theta^{Q})}{\partial \theta^{Q}}\right] \tag{14}$$

where $\widehat{\theta}^{\mu}$ and $\widehat{\theta}^{Q}$ denote the parameters of the target strategy network and the target value network respectively. DDPG uses the empirical playback mechanism to obtain training samples from playback memory unit $D$, and passes the gradient information about the action by the Q-value function from the critic network to the actor network, and updates the parameters of the strategy network along the direction of boosting the Q-value according to Eq. (14).

The algorithms are classified as shown in Fig. 4:

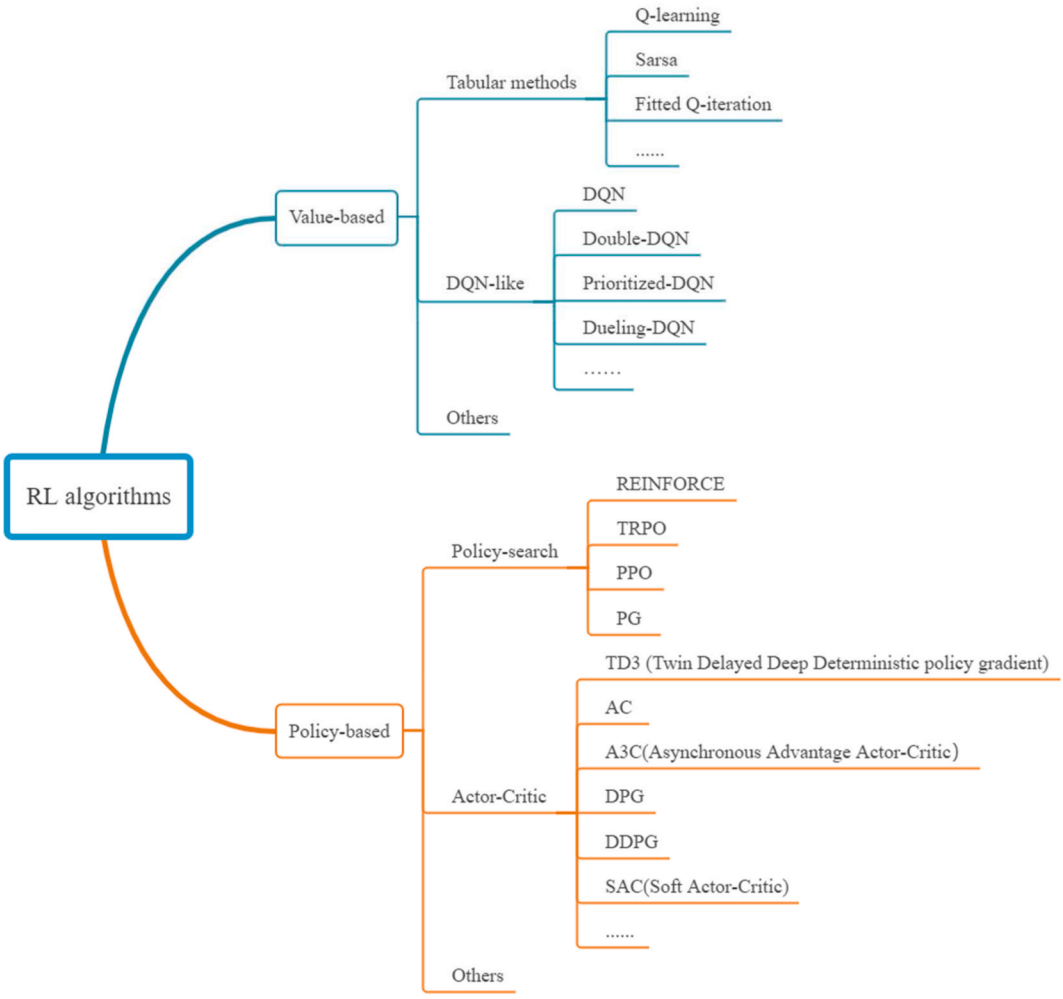


**Fig. 4.** Classification structure of the algorithms.

**Table 1**
Mainstream RL algorithms.

| Algorithm category | | Algorithm | Applicable scenarios | Considerations for method selection in application scenarios |
|---|---|---|---|---|
| Value-based | | Q-learning<br>SARSA<br>DQN-like | Discrete Action Space | 1 They have a relatively faster convergence rate.<br>2 Q-learning and SARSA are preferred for scenarios with a small and discrete state space.<br>3 DQN-like methods are preferred for scenarios with a large or continuous state space. |
| **Policy-based** | **Policy search** | REINFORCE<br>TRPO<br>PPO | Discrete/continuous action space | 1 They have a relatively slower convergence rate.<br>2 They can handle scenarios with a large or continuous action space.<br>3 PPO can provide a more stable training relatively. |
| | **Actor-critic** | AC<br>A2C<br>A3C<br>DPG<br>DDPG | Discrete/continuous action space | 1 They have a relatively slower convergence rate.<br>2 They can handle scenarios with a large or continuous action space.<br>3 DDPG is only available for continuous action spaces.<br>4 A2C and A3C can provide a faster training relatively. |

The mainstream RL algorithms are shown in Table 1.
Some of these algorithms have been applied to reduce building energy consumption, which we will analyze in Section 3.

## 3. Reinforcement learning algorithms for building energy efficiency control

### *3.1. Search methods and results*

RL itself is a means to find the optimal strategy, and RL applied to the building field to find the optimal energy saving strategy can help reduce building consumption. RL has been applied to many scenarios in the building field, such as HVAC systems, and has also been proven to be effective in the area of building energy saving. In this section, we review recent papers on the application of RL to the building field.

We used *Google Scholar* and *Connected papers* for literature search. When searching, we do not restrict the publisher, and we search for relevant papers using keywords ("reinforcement learning" or "deep reinforcement learning"), intelligent building, energy conversation and so on. We have selected papers from the search result that use reinforcement learning or deep reinforcement learning for building control, all of which are examples of successful applications of RL algorithms to reduce building energy consumption in recent ten years. HVAC systems are the main energy consumption parts in intelligent buildings, many RL algorithms have been applied to save energy consumption of HVAC systems, and have achieved good results. The RL algorithms and application scenarios we investigated are shown in the appendix part.

According to our classification of RL methods, the comparison of the volume of applied papers by year for the two types of algorithms is shown in Fig. 5.

As shown in Fig. 5, the application of RL algorithms in intelligent buildings have been relatively less until 2010. While after 2010, RL algorithms were gradually applied to this field, and this trend also showed a yearly growth. In particular, the proposal of DQN in 2015 brought this growing trend to a peak. Due to the introduction of deep learning in RL, DQN can even surpass human beings in

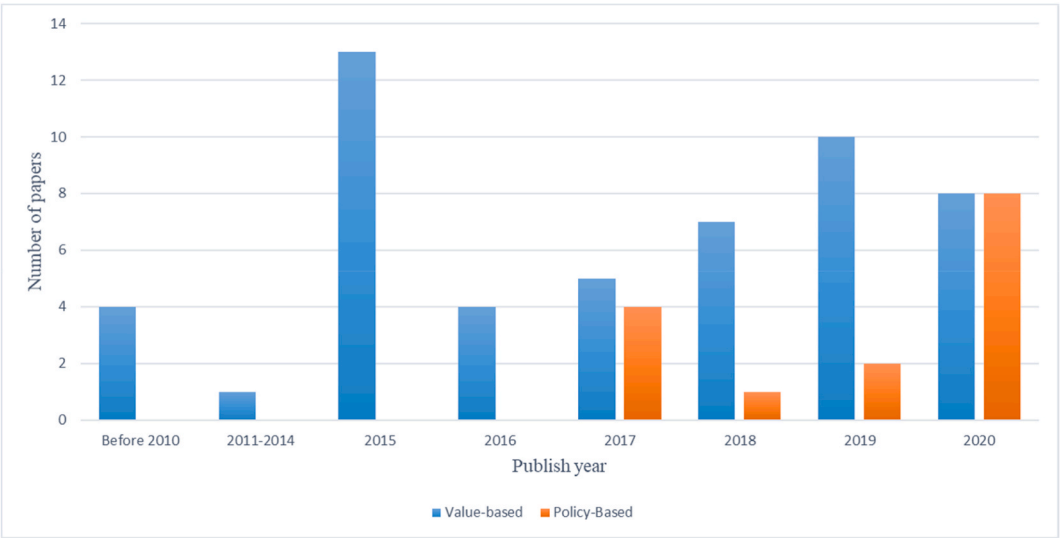


**Fig. 5.** The number of papers on "value-based algorithms" and "policy-based algorithms" from "before 2010" to the present.

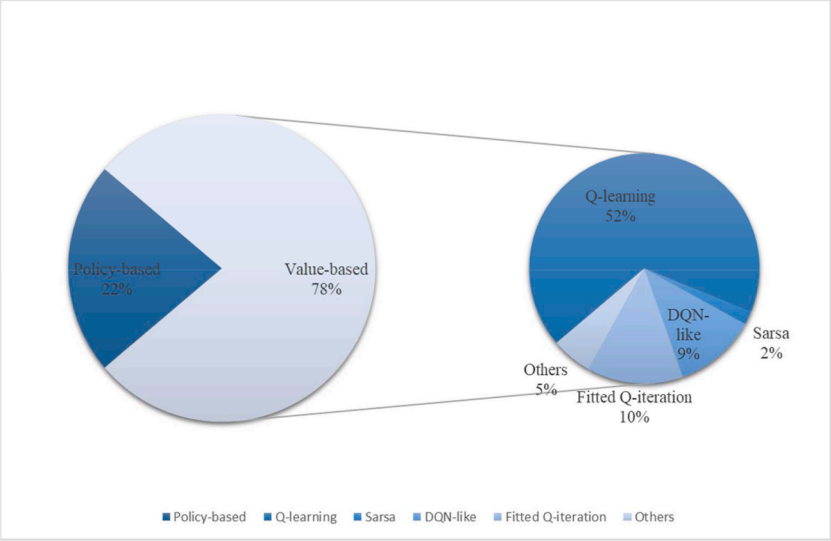


**Fig. 6.** The proportion of each algorithm in the value-based class of methods.

many complex games, which demonstrate its powerful control capabilities and deep potentials. Since the introduction of deep reinforcement learning, more and more researchers have also tried to apply RL to intelligent building control, and this trend has been growing steadily as seen in our research, which shows the potential of RL. In our research, value-based algorithms have been used since before 2010, and show a year-by-year growth trend after 2010, which is one of reasons why value-based algorithms account for a larger proportion of RL algorithms (as shown in Fig. 6). Policy-based algorithms started to be used in our research from 2017 and grow gradually, which is related to the development of policy-based algorithms (DPG (2014), DDPG (2016)).

### 3.2. *Value-based algorithms for building energy efficiency control*

In this section we will discuss the application of the methods of the value-based class in intelligent buildings, where the ratio of each method in this class is shown in Fig. 6.

Owing to the earlier applications of value-based algorithms in Fig. 5, we can find that the proportion of value-based algorithms are relatively large, which contributes to 78% in Fig. 6. By contrast, policy-based algorithms only account for 22%.

Value-based algorithms are suitable for practical problems with small and discrete action space. When solving problems in continuous action space, they seem to be powerless. However, researchers can use discretization to overcome this drawback, which expands the application scope of value-based algorithms (e.g., the control set temperature interval is 23°C–26 °C, and after discrete, the set points are 23 °C, 24 °C, 25 °C, and 26 °C), but compared with policy-based algorithms, the discretization may lead to a sub-optimal control policy.

Among value-based algorithms, Q-learning is the most classical algorithm, which contributes 52% applications of value-based algorithms as shown in Fig. 6. For control problems with small-scale state space, Q-learning is easy to understand and deploy relatively. However, it is difficult to solve problems through Q-table, when the state space is large or actually continuous. Function approximation is a common way to deal with this shortcoming. For example, DQN-like algorithms, proposed after 2015, use deep neural network to be the function approximator for representing the Q-table, which can tackle some complex practical problems with large or continuous state space. According to our research, DQN-like algorithms contribute 9% applications in intelligent buildings over the past 5 years.

#### 3.2.1. *Tabular-based methods*

In tabular-based methods, Q-learning is more commonly used. The Q-learning algorithm is suitable for solving problems where the action state space is discrete and relatively small. The selection of Q-learning actions depends on the update of the Q-table, and if the action state space is large, it will lead to a very large Q-table, which in turn will lead to a degradation in the performance of the algorithm. However, Q-learning is the preferred method for many problems, and if you encounter a large Q-table you can consider using a neural network instead of a Q-table, which is the method of the DQN-like. Problems that are usually solved using the Q-learning algorithm require that the state action space of the problem be discretized, or that the continuous variables in it be discretized.

Li et al. [16] control the HVAC system in order to control energy consumption and ensure comfort. In view of the problem of slow convergence rate in RL, they put forward multi-grid Q-learning to achieve a better convergence strategy. Lork et al. [17] used a Bayesian convolutional neural network combined with data from all rooms to construct a temperature and air conditioning power prediction model to reduce uncertainty. This model is then adapted to individual rooms and the temperature set point is controlled using Q-learning to achieve a balance between comfort and energy savings. Barrett et al. [18] use Bayesian inference to predict room occupancy, Q-learning in RL to learn the control strategy of HVAC. The results show a 10% energy savings over programmable control

methods. Lu et al. [19]. develop statistical thermal comfort models with various machine learning algorithms. They performed a data-driven simulation of a comfort-based temperature setpoint control system using Q-learning. Regardless of the initial temperature setting, Q-learning-based temperature control does achieve a comfortable temperature range for occupants. Sun et al. [20] develop an innovative event-based approach within the lagrangian relaxation framework. Q-learning is used to address the problem that optimal policies may change over time. The results show that the approach can be able to maintain similar levels of energy consumption and human comfort as the time-based approach, but with significantly less computational effort and faster response to events. Chen et al. [21] use Q-learning to make optimal control decisions for HVAC and window systems. The results show that RL can reduce energy consumption and time to thermal discomfort.

Al-Jabery et al. [22] proposed Q-learning based demand-side management techniques for domestic electric water heaters, and simulation experiments showed that the energy cost consumed by domestic electric water heaters was reduced by about 26%.

Lu et al. [23] describe the dynamic pricing problem as a discrete finite MDP and use Q-learning to implement decisions on electricity pricing. Simulation results show that the algorithm can improve the profitability of electricity suppliers while reducing the energy cost of consumers, achieving a win-win situation and balancing the supply and demand in the electricity market. Rocchetta et al. [24] develop a RL framework based on Q-learning for Optimal management of grid operation and maintenance. They use Artificial Neural Networks (ANN) tools to replace the Q table of Q-learning, so that the method can be extended to realistic problems with large continuous state spaces. The results show that the framework outperforms the expert-based solution. Henze [25] proposes a Q-Learning based adaptive optimal control algorithm for grid-independent photovoltaic systems. The results show that the Q-learning-based strategy outperforms the traditional control strategy in terms of cost function. Naghibi-Sistani et al. [26] propose a Q-learning based bidding strategy for electricity prices according to temperature changes to help market participants find the optimal offer strategy. Wen [27] describes the fully automated energy management system rescheduling problem as a RL problem in order to solve the demand response problem for buildings. Theoretically practicable with any RL algorithm, the paper uses Q-learning for simulation and the results show that their rescheduling strategy outperforms the baseline strategy. Ding [28] proposes a Hidden Mode Markov Decision Process (HM-MDP) model in smart homes for real-time decision making of customers to maximize their profits using a Q-learning based approximate dynamic programming algorithm for online decision making. Experimental results show that the algorithm outperforms greedy and stochastic algorithms. Etim et al. [29] proposed three energy management strategies based on power pinch point analysis to eliminate demand compliance problems and random variations of renewable energy in hybrid energy storage systems. The third of these methods is based on Q-learning, and this method responds best in all performance metrics species. The paper states that Q-learning is preferred when dealing with conditions with large errors and no information about the type of uncertainty. T [30] proposed a generalized model of the renewable residential load dispatch or load distribution problem (LCP) considering stochastic regenerative sources, stochastic rates, and stochastic critical loads, and applicable to any tariff type, and used Q-learning to solve the LCP. Simulation experiments verify the effectiveness and scalability of the algorithm.

Jiang [31] uses Q-learning to find an optimal discharge strategy for Vanadium Redox Battery in a residential community microgrid to reduce cost. The simulation results verify its effectiveness in reducing energy consumption. Xin [32] proposes an interconnection topology and Q-learning to optimize the coordination between different energy storage systems in a microgrid. The results show that the method is able to coordinate the discharge cycles of different battery systems to obtain better system efficiency. Wei et al. [33] proposed a new dual iterative Q-learning algorithm to solve the problem of optimal management and control of batteries in a smart housing environment to obtain an optimal management scheme for residential energy systems. There are two iterations, external iteration and internal iteration, the internal iteration minimizes the total cost per cycle and the external iteration is to achieve the

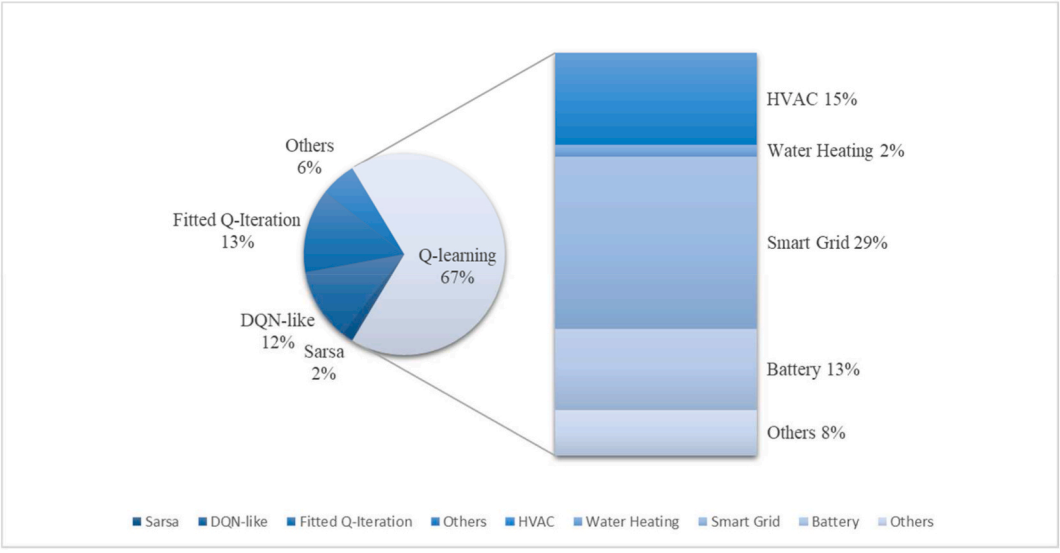


**Fig. 7.** Proportion of Q-learning in each application scenario.

optimal Q-function. Jiang [34] uses Q-learning to find an optimal discharge strategy for vanadium redox battery to reduce energy cost considering power price, wind power and load demand stochastic factors.

Liu et al. [35] use the empirical wavelet transform (EWT), three deep neural networks and RL to perform wind prediction, where RL uses Q-learning, which is used to combine the prediction results derived from the prediction models built by the three deep neural networks to give the final results. The experimental results show that reinforcement-based learning is not only effective in integrating the three deep networks, but also gives accurate results in all cases.

Q-learning is widely used in various scenarios. By analyzing the problem requirements of each scenario, problems with discrete state action spaces can be directly matched with Q-learning algorithms, and discrete means can also be used to match Q-learning algorithms. Q-learning algorithms are widely used because they are relatively simple. In this regard, we also did the statistics of the proportion of Q-learning algorithm application scenarios in the papers we collected, and the results are shown in Fig. 7.

As shown in Fig. 7, the application of Q-leaning reaches 67% of all value-based algorithms, and its application areas cover almost all the fields as we investigated, especially smart grids and HVAC systems. A possible explanation for this phenomenon might be that, for these problems, the state and action spaces are small or discrete, or the action space can be discretized easily, so the Q-table is easy to be built. For example, in Ref. [23], dimensions of state space and action space are 2d and 1d, respectively. The state space consists of two types of consumers' demands (before and after knowing the price of electricity). And the action is just the electricity pricing. In Ref. [36], for the temperature control, the state space has three dimensions, including clothing insulation, indoor air temperature and relative humidity, and the action space is composed of temperature setpoint (15°C–31 °C with interval 1 °C).

Compared to Q-learning, the Sarsa algorithm has been applied relatively less in our research. Gracia et al. [37] used the Sarsa ($\lambda$) algorithm to control a ventilated curtain wall with Phase change materials (PCM) to determine the operating schedule in its air chamber. The algorithm uses a simple isothermal model to determine the optimal schedule for charging the PCM. Experimental results show that effective energy savings can be achieved under different test climatic conditions.

Another tabular-based method is fitted Q-iteration, which obtains the optimal policy by solving the value function of the agent state-action pair.

Brida M et al. [38] use fitted Q-Iteration to control the operation of battery storage devices in a microgrid to improve the utilization of energy generated by the local PV system, thereby reducing the cost of electricity and dependence on the local grid. F Ruelens et al. [39] control domestic water heater clusters with fitted Q-Iteration to plan water heater clusters with minimum electricity cost. Simulation results show that the method is able to reduce daily electricity costs within a reasonable learning period of 40–45 days compared to a stagnant loop controller. F Ruelens et al. [40] used fitted Q-Iteration to find a control strategy for electric water heaters and applied an autoencoder network to find a compact feature representation of sensor measurements, and the experimental results showed that the method was able to reduce the total energy consumption by 15% over the thermostatic controller for a 40-day indoor experiment. Somer et al. [41] uses a model-based fitted Q-Iteration method with practical experiments on six dwellings in order to improve the self-consumption of local PV production to optimize the buffer heating cycle for domestic hot water. The experimental results showed that the method significantly improved the self-consumption of PV production. Asa et al. [42] achieve the goal of maximizing the self-consumption of PV production by optimizing the heat pump duty cycle for domestic plus hot water and the charge/discharge cycle of the storage battery. They use a model based fitted Q- Iteration method to solve the sequential decision problem, which enables the domestic hot water consumption to be 100% provided by PV production in summer and 50% in winter.

The application examples of the above tabular-based methods cover common systems in building control. Researchers often firstly try to use tabular-based methods to solve the related problems. The tabular-based methods are relatively easy to implement and the control effect is relatively not too bad, which is also the reason why researchers are willing to adopt or experiment with it. In the above examples, the state and action space of the control object are relatively small, and the action space is also discrete. In the face of continuous state or action space, a common way is to use discretization methods to deal with them, so that the problems can be solved by using tabular-based methods. It is worth mentioning that Q-learning algorithm is the most commonly used algorithm in this class and covers the most problem scenarios.

According to the research result of tabular-based algorithms, we recommend this type of algorithm for practical problems with small action space and small-scale state space. Continuous state or action can be pre-processed by discretization method to match this type of algorithms. In addition, Q-learning algorithm can be considered first when this type of algorithms is considered.

#### *3.2.2. DQN-like methods*

The DQN series of algorithms is an improved version of the Q-learning algorithm and is able to cope with more complex problems. The DQN series is able to solve high latitude problems and continuous space problems compared to Q-learning. However, the problems that the DQN series methods can solve are still oriented to discrete action space, and the DQN series algorithms are difficult to solve continuous action problems, such as vehicle driving.

Ahn et al. [43] use DQN to achieve model-free optimal control equilibrium between different HVAC systems. The results prove that DQN not only reduces energy consumption, provides model-free optimal control, but also balance the control of different energy consumers in a building. Gupta et al. [44] make heating control decisions for intelligent buildings and proposed DQN-based heating controllability to minimize energy costs while improving thermal comfort in intelligent buildings. Simulation experiments using real data were conducted in the study, and the experimental results showed that the controller was able to reduce energy consumption by 5%–12% while improving thermal comfort by 15%–30%, showing excellent performance over conventional thermostat controllers. Brandi et al. [36] use DQN to control the water supply temperature set point of the heating system terminal unit, which can achieve a heating energy saving ranging between 5% and 12%. Yoon [45] developed a thermal comfort prediction model, combined with DQN, to achieve an optimal control strategy that can guarantee minimum energy consumption.

In the above application examples, the state space of the problem becomes large enough or continuous, which makes the algorithms of tabular-based methods no longer applicable. DQN-like methods use deep neural network as function approximator to solve this kind of problem. In our research, although the application examples of the DQN-like methods are less applied compared to the tabular-based methods, the DQN-like methods will be more widely used with the development of deep RL and the increasing complexity of building control problems.

According to our research result of DQN-like methods, we suggest using DQN-like methods when the practical problem action space is discrete, but the state space becomes large enough or continuous.

### 3.2.3. Case study

Paper [46] belongs to the application of Q-learning in HVAC. Q-learning algorithm is used to control the cooling system, which consists of the same chillers, the same cooling towers, and the same cooling water pumps. The Q-learning algorithm itself is a model-free RL algorithm, and the paper uses the Q-learning algorithm for the characteristics of the Q-learning algorithm and the needs of the problem. The COP (coefficient of performance) of the chiller plant has a direct impact on the operation of the chilled water system and on the energy consumption of the HVAC system. The paper investigates the optimal control method for the chilled water system, using Q-learning to optimize the system COP.

**State:** Combination of ambient wet bulb temperature and discretized system cold load.

These two variables were chosen as states because they are not affected by system operation and are important for chilled water systems. To accommodate the Q-learning algorithm, the authors discretized them, with the wet bulb temperature discretized into an integer (e.g., 24 °C) within the temperature interval (e.g., 23 °C–29 °C). The system cooling load is discretized into 0.5, 0.6, …, k CC (cooling capacity) for a system with k chillers based on the cooling capacity of each chiller (e.g., the cooling load of 0.57 CC is discretized into 0.6 CC).

**Action:** Frequency setting for pumps and cooling tower fans.

The action selection needs to be within the reasonable range allowed by the hardware device, and the frequency accuracy of the action selection is 1hz, which is also discrete.

Reward:

$$System\ COP = \frac{CL_{system}}{P_{chillers} + P_{cwps} + p_{towers}} \tag{15}$$

where $CL_{system}$ is the system cooling load (kW), $P_{chillers}$ is the total power of all chillers (kW), $P_{cwps}$ is the total power of all cooling water pumps (kW), and $P_{towers}$ is the total power of all cooling towers (kW). The goal is to make this reward value as large as possible, i.e., the sum of the powers in the denominator is small, which means that energy savings are achieved.

The system uses two agents for controlling the cooling pumps and cooling towers, both trained using the Q-learning algorithm, and the update principle by Eq. (1), where the meaning of Q-value is the value of COP in this paper, $\alpha$ was set to 0.9. $\gamma$ was set to 0.01 because in this study, the pump and tower agent action did not affect the next state, and at each time step, it was only necessary to focus

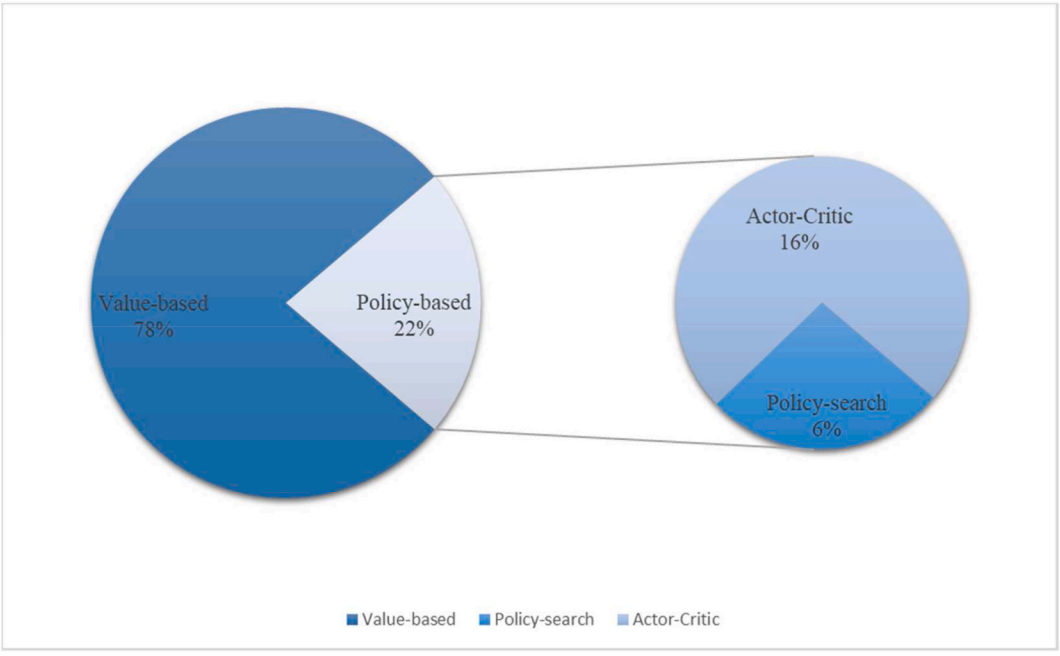


**Fig. 8.** The proportion of each algorithm in the policy-based class of methods.

on how to maximize the current long-term reward rather than the total reward.

This study conducted simulation experiments based on real-world data comparing a basic controller, a local feedback controller, a model-based controller, and a model-free controller based on RL. Three months of simulation experiments showed that the RL-based model-free controller was able to save 11% of the system energy, more than the 7% saved by the local feedback controller. The model-based controller has enough sensors and data to obtain 14% system energy savings, but the RL-based model-free controller not only provides satisfactory energy savings, but also does not require too many sensors and is more suitable for engineering practice than the model-based controller.

In this study, the authors used Q-learning to obtain optimal control of the cooling water system. The Q-learning method addresses the problem of discrete action states and does not require a complete model, which matches the needs of the cooling water system in this study. Since the Q-learning method requires a discrete action state space, the authors also discretize the states and control actions in the cooling water system to match the Q-learning algorithm for the characteristics of the method.

In our research, Q-learning algorithm is relatively simple to implement and is often used in the control of HVAC systems. In this case, the action space of the target control object is relatively small (cooling water pump control frequency is between 35Hz and 50Hz, and cooling tower control frequency is between 30Hz and 50Hz, both with a frequency interval of 1Hz) and the state space can be discretized, so the control policy can be achieved by using Q-learning. The authors used Q-learning algorithm for the actual control of the cooling water system in HVAC, which is consistent with our research results and recommendations. The authors discretized the state space, which is a common way that we have mentioned in order to make the Q-learning algorithm more applicable. The control results achieved in this case are second only to the model-based control method, and the actual implementation costs are also effectively controlled.

### 3.3. Policy-based algorithms for building energy efficiency control

In this section, we will discuss the application of policy-based class of methods in intelligent buildings, where the proportion of each method in this class is shown in Fig. 8.

Since policy-based algorithms have been applied to intelligent buildings recently (as shown in Fig. 5), its applications account for a smaller proportion compared to the value-based algorithms, just 22% in our research. From the perspective of policy optimization, policy-based algorithms can be divided into two categories, namely policy-search and actor-critic algorithms. As shown in Fig. 8, applications of policy-search algorithms account just 6% in our research. Policy-search algorithms take relatively long time for policy evaluation, which is costly in practical application. Actor-Critic algorithms combine the value functions and the approximated policy representation to reduce the time of policy evaluation, and thus are more efficient for practical applications. In our research, its application reaches 16% of all the algorithms we investigated.

In Section 3.1, we mentioned that value-based algorithms are easier to implement than policy-based algorithms. However, when solving complex control problems with large or continuous action space, the restriction of value-based algorithms often leads to a decrease in control performance. Some researchers simplified the problem to make value-based algorithms more applicable, but this way is a compromise at the expense of control performance, just as leading to a suboptimal policy. By contrast, policy-based algorithms are more effective and have excellent performance in complex control problems.

#### 3.3.1. Policy search methods

The method of policy search class parameterizes the policy to obtain the optimal policy by optimizing the parameters, which has better performance and convergence for solving the problem of continuous actions.

PPO algorithm is able to strike a balance between sampling efficiency, algorithm performance, and implementation. The PPO algorithm tries to compute a new policy in each iteration that minimizes the loss function, and the policy does not differ much from the original policy, with good experimental results.

Chen et al. [47] proposed the GNU-RL method based on the PPO algorithm to control HVAC and tested its deployment in the simulated environment Energy Plus and in the real world. The experimental results showed that the method was able to save 6.6% of energy in the simulated environment compared to the best RL results in the same environment, while ensuring comfort. In the real world, the method was able to save 16.7% of cooling requirements compared to existing controllers. Azuatalam [48] used PPO-Clip to control the entire building to meet demand response, simulation results show that considerable energy savings can be achieved using RL while maintaining an acceptable level of thermal comfort.

In the above application examples, the problem action space is relatively large or continuous, and can be solved by using policy search methods, but its application is relatively less because of its complexity compared with actor-critic methods. According to our research, policy search methods can handle practical problems with relatively large or continuous action space, but they are not the first choice.

#### 3.3.2. Actor-critic methods

The Actor-Critic class of algorithms combines value-based and policy-based approaches to learn both policy and value functions, and also combines the advantages of both algorithms. It has better performance for solving complex building control problems.

Similar to the DQN method, the DPG method uses a neural network to output the probabilities of all possible actions. Mocanu et al. [49] applied deep RL in order to optimize the scheduling of electrical equipment in residential buildings, using two algorithms, DQN and DPG, and experimentally demonstrated that both methods can achieve energy cost minimization.

DDPG can be used to solve problems with continuous state action space, which is specially designed for solving problems with continuous variables. DQN outputs the Q values of all actions, DPG outputs the probabilities of all actions, and DDPG directly outputs a

deterministic action, so the output action space can also be continuous.

Yan et al. [50] apply DDPG to generate an optimal control strategy for a multi-zone residential HVAC system, it can greatly reduce energy consumption while ensuring comfort. The DDPG-trained agent is able to intelligently balance different optimization objectives with generalization ability and adaptability to unknown environments. Liu et al. [51] combine Autoencoder (AE) algorithm with DDPG algorithm. They use AE to extract the high-level features of the DDPG state space, and thus propose a method for short-term prediction of energy consumption of HVAC systems based on DDPG, which has better prediction performance than the traditional method.

Kou et al. [52] describe the optimal voltage control problem as a constrained Markov decision process, where both the state space and action space are continuous, for which the DDPG is used to determine the optimal control behavior.

Odonkor [53] uses DDPG to learn continuous charge/discharge strategies for shared batteries in building clusters. The experimental results show that the method can be adapted to different building clusters and can solve the ongoing task of charging/discharging shared batteries.

Wang [54] designs a model-free actor-critic RL controller using long and short term memory (LSTM) networks combined with RL which aims to optimize thermal comfort and energy consumption. The experimental results showed that the RL controller improved thermal comfort by an average of 15% and energy efficiency by an average of 2.5%.

In the above application examples, the action space of the problem is large or continuous. Actor-critic methods combine the advantages of value-based and policy-based methods to solve such problems and achieve excellent control. Although these algorithms are relatively difficult to implement, they are able to achieve excellent control effect in solving such complex scale problems, while other methods, such as tabular-based methods, policy search methods and so on, maybe only can achieve a sub-optimal control policy, or even fail to achieve any one.

According to our research result of actor-critic methods, we suggest that actor-critic algorithms should be given priority when solving practical problems with large or continuous action space.

### 3.3.3. Case study

Paper [50] uses the DDPG algorithm to learn control strategies for intelligent multi-zone residential HVAC with the goal of reducing energy consumption while ensuring comfort. The DDPG algorithm is used to solve continuous problems and is able to solve problems with high-dimensional action spaces. Only the heating function of the HAVC system is considered in the study, i.e., the implementation of a continuous thermal control strategy for residential HVAC. The DDPG approach has better performance corresponding to the continuity problem, so it is adopted.

**State:** (1) current outdoor temperature $T_{out}(t)$; (2) current indoor temperature $T_{in,z}(t)$ for all the zones z; (3) the lower bound of the user comfort level $T_{lower}(t)$; (4) retail price $\lambda^{retail}(t)$, where t is the current time step.

**Action:** The setpoint $Setpt_z(t)$ for the zone z, where the setpoints in each region are continuous variables.

**Reward:** A combination of energy costs and comfort violation costs for control intervals, which can be described as follows:

$$r(t) = -w_c \sum_{t'=t-\Delta t}^{t} \lambda^{retail}(t') E_{HVAC}(t') - w_p \sum_{t'=t-\Delta t}^{t} C^{penalty}(t') \tag{16}$$

This reward is a multi-objective function which consists of two items, the former being the energy consumption of the HVAC system and the latter being the penalty for violating the user's comfort, where $\lambda^{retail}(t')$ is the retail price, $E_{HVAC}(t')$ is the power consumption, and $\Delta t$ is the control interval. To balance these two objectives, two weighting factors $w_c$ and $w_p$ are added, setting the ultimate goal of this reward to obtain the lowest total energy consumption plus the smallest penalty. The user comfort violation can be calculated as follows:

$$rC^{penalty}(t') = \begin{cases} 1, for\ T_{in}(t') < T_{lower}(t') - T_{th} \\ 0, elsewise \end{cases} \tag{17}$$

$T_{th}$ is a threshold with a small value, which allows violation levels less than $T_{th}$, but penalty terms above $T_{th}$ will be activated.

The authors matched the parameters in their intelligent multi-zone residential HVAC heating control problem with the DDPG algorithm, and the algorithm flow can be found in paper [50]. The DDPG algorithm is capable of solving complex problems with continuous action spaces, and the properties of this problem fit the properties of the DDPG method, so the DDPG algorithm is adopted.

In this study, the authors compare the results of the DDPG method with the DQN method for this same problem, where both methods have the same states, actions, and rewards. According to the properties of the DQN method, state space can be continuous, but the action space is discrete, so the authors discretized the actions in steps of 0.5 °C. The authors compare 2 benchmark cases with DDPG, DQN. In the benchmark cases, the rule-based case has the lowest cost because it is based on the price result; the fixed setpoint case has no temperature violation but the highest energy cost. Both DDPG and DQN are both able to achieve a balance between price and temperature violation, but DDPG has lower energy cost and less temperature violation than DQN, and energy consumption can be reduced by 15%. Comfort violations are reduced by 79%.

According to our research, we recommend using policy-based algorithms when solving problems with large action space or continuous action space. In this case, the state space dimension of the problem is high and the action is a continuous temperature set point, the authors use DDPG in policy-based algorithms to solve the problem, which is consistent with our recommendation. The relatively large dimensionality of the state space in the case makes the tabular algorithms no longer applicable, and the authors use a deep neural network as a function approximator instead, which is a common alternative in the papers we researched. In order to

compare the results with the DDPG algorithm, the authors also conducted experiments using DQN, but the action (temperature set points) was discrete, since DQN cannot output continuous actions. In the experimental results, both the DQN and DDPG algorithms meet the control requirements, but the DDPG algorithm achieves better control than DQN, and this result is in line with the expectations of our research.

## 4. Discussion

The application of RL to reduce building energy consumption can provide significant building energy or cost savings and other benefits compared to traditional methods. With the development of RL algorithms, many new approaches are emerging, such as deep reinforcement learning (DRL), and multi-agent reinforcement learning (MARL), which are very promising to advance building management and control. This section will discuss the advantages and limitations of RL algorithms to provide guidance for researchers or engineers regarding the applications of RL in buildings filed. In addition, some limitations of RL applications and future directions are also outlined.

### *4.1. Advantage*

This paper reviews papers that apply RL methods to intelligent buildings. According to our research, applying RL to intelligent buildings can help reduce building energy consumption and achieve energy savings. The exploratory and exploitative nature of RL makes it unnecessary to build a complete system model, and it can demonstrate superior performance and efficiency over traditional control methods when dealing with intelligent buildings control problems with uncertain information.

### *4.2. Limitations*

Applying RL to intelligent buildings inevitably has some problems. Paper [55] using DQN and DDPG to control refrigerant control parameters also points out that there are still many problems to be investigated when applying RL to real buildings. For example, the training time, the modeling of the environment, the difference between the simulated environment and the real environment, the cost of applying to the real environment, etc. These are all obstacles to the application of RL to building control.

#### *4.2.1. Cost*

RL itself is a process of learning by trial and error, so RL agents will perform actions during the training phase to obtain results that do not meet our expectations, or even the opposite of what we expect to achieve. If RL is deployed directly in a real building for training and learning, it will bring great inconvenience to occupants and even bring damage to the equipment, resulting in uncontrollable actual costs and easily bring security problems. For example, paper [52] applies RL to smart grids, where the security of the grid is largely and directly related to the cost issue. The paper proposes a security exploration approach to constrain the operation of active distribution grids. In the paper, a security layer is composed directly on top of the participant network of the DDPG, which predicts the change of the constrained state and thus limits the violation of the working operation of the active distribution grid.

To reduce unnecessary cost, many current studies, which use simulation platforms to deploy RL for training learning, take simulation experiments to derive experimental results. Paper [56] mentioned that modeling and simulation are effective methods for solving complex system problems. For example, in the paper [36], the simulation environment for the experiments relies on the Building Control Virtual Test Bed (BCVTB) and the external interface of EnergyPlus to perform the simulations. Paper [57] introduces a new simulation environment, merged by the building energy simulator CitySim and the machine learning library TensorFlow. This simulation environment enables researchers to study new learning control algorithms to verify the robustness of the method and various applications in the building environment.

Applying RL to real buildings, how to eliminate the control signals that may appear in various stages of RL that bring undesirable consequences and reduce unnecessary costs is a key issue for future implementation of RL.

#### *4.2.2. Training time*

RL training controllers require a lot of training time to achieve the desired performance. The dimensionality of the state-action space in RL, the complexity of the control problem, the arithmetic power of the computer, and the trial and error of RL are all factors that limit the training time for RL. The application of RL to real buildings will again lead to a rise in the real cost.

In paper [58], the authors mention that adding prior knowledge to the learning process can speed up the training due to the fact that prior knowledge increases the agent's knowledge of the environment and reduces the spontaneous exploration of the environment. Thus, the training time is reduced. Adding prior knowledge to the learning process is an effective way to reduce training time.

For the application of RL methods in real intelligent buildings, how to reduce the time of RL training is also a key issue that cannot be ignored.

### *4.3. Future directions*

The application of DRL to intelligent buildings is one of the inevitable trends in the future. With the upgrading and updating of information collection devices in intelligent buildings, the amount of data in intelligent buildings is bound to increase, and the complexity of smart building control problems rises. DRL algorithms are more prominent in dealing with problems with very complex state-action spaces. As the computing power of computers continues to increase, the training period for DRL decreases accordingly, and thus its feasibility increases. In our review, there have been a number of papers using DRL techniques for control of intelligent buildings. Paper [59] used three DRL methods to predict building energy consumption for the case of an office building, and the results demonstrated the potential of DRL techniques in the field of building energy consumption prediction.

MARL is also a future direction for the application of RL to intelligent buildings. In the heating control of paper [44], the authors performed centralized control of buildings (all buildings based on a single DQN controller) versus decentralized control (each heater is controlled independently and based on its own DQN controller). The decentralized control system is characterized by multi-agent. As the number of buildings and the differences in temperature set points between buildings increased, the decentralized control showed better performance than centralized control. Hussain Kazmi et al. [60] used MARL applied to thermostatic control of water heaters, and the control effect was also better than that of the single-agent system. In the smart grid application, Leo Raju et al. [61] use MARL, where each agent aims to improve the utilization of batteries and solar energy, ultimately in order to achieve the goal of reducing power consumption in the grid. Multiple individuals of the same class of objects in intelligent buildings are described as multi-agent, or individuals of different classes of objects are described as multi-agent, and generally there is competition, cooperation, or co-existence of competition and cooperation among these agents. Different RL algorithms can be used to train each agent, consider the communication between agents, and achieve the final control goal, such as controlling energy savings. When facing control problems with complex objects in intelligent buildings, MARL approaches have higher performance and efficiency compared to single-agent approaches.

Apart from building energy management and control, RL has been increasingly investigated in occupant related building control in recent years. Occupant behaviors have great impacts on the usage of building equipment and HVAC systems, and thus affecting the building energy consumption. It is particularly important to combine the behavior of building occupants and building control. For instance, Wang et al. [62] proposed an occupant-density-detection based approach for building ventilation control to reduce the risk of COVID-19 in densely occupied indoor environment. However, the behavior of building occupants is stochastic and complex in nature. Modeling and predicting the occupant behavior is challenging, making the occupant-based control difficult [63]. RL offers a model-free approach for the building control and management. Paper [64] provides a review of papers on the application of RL to combine the behavior of people in buildings. Considering its attractive features and promising application potentials, the study combing the RL with occupant behavior for building control needs more exploration and application, which is an important research direction in the future.

## 5. Conclusion

This paper presents a review of reinforcement learning algorithms applied to intelligent buildings. We find that current research on reinforcement learning algorithms applied to the field of intelligent buildings, the value-based class of reinforcement learning algorithms is used more often, among which the more prominent is the Q-learning algorithm. The policy-based methods become class of methods have been gradually applied since 2017, and are gradually applied to complex building control problems. Deep reinforcement learning algorithms including DQN, DDPG, etc. are also gradually applied to intelligent building control.

Reinforcement learning as a control method applied to intelligent buildings has better performance and efficiency than traditional control methods. We classify the algorithms of reinforcement learning and study the application of each algorithm in each category separately. In addition, we also analyze the characteristics of the algorithms, so as to provide a reference for the subsequent selection of reinforcement learning algorithms to solve practical problems. When selecting RL algorithms to solve control problems in different scenarios of intelligent buildings, we try to present some suggestions for consideration:

- For practical problems with small action space, we suggest using value-based algorithms. When the problem with small-scale state space and the action space is discrete or can be discretized, Q-learning algorithm can be considered, such as [19,23,46]. When the problem's state space becomes large enough or continuous, we suggest using function approximation methods, such as DQN-like algorithms [24,45].
- For practical problems with large action space or continuous action space, we suggest using policy-based algorithms, especially Actor-Critic algorithms. For this type of problems, policy-based algorithms can achieve excellent control results, such as [48,52,65,66].
- When solving practical problems with complex control scales and multiple control objects (e.g., building clusters, multi-building control), we suggest using multi-agent reinforcement learning algorithms to control different objects in complex systems separately to achieve the final control goal, such as [16,60,61,67].

Regarding the future direction of reinforcement learning applied to intelligent buildings, we summarize the following points:

- Deep reinforcement learning is more appropriate than existing reinforcement learning algorithms for solving more complex building control problems. With the improvement of computer arithmetic power and research on deep reinforcement learning, it is worth applying deep reinforcement learning to building control to solve real-world problems.
- Multi-agent reinforcement learning has better control efficiency and performance than single-agent reinforcement learning for building control problems, and multi-agent reinforcement learning applied to building control needs more recent research.
- The use of reinforcement learning combined with occupant behavior for building control is also a very promising application direction.

## Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

## Acknowledgment

This work was financially supported by National Key R&D Program of China (No. 2020YFC2006602), National Natural Science Foundation of China (No. 62072324, No. 61876217, No. 61876121, No. 61772357), University Natural Science Foundation of Jiangsu Province(No. 21KJA520005), Primary Research and Development Plan of Jiangsu Province (No. BE2020026).

## Appendix

| Category | Paper | Algorithm | Scenario | Control object/ control target | Saving effect/Conclusion | Year-Journal Name |
|---|---|---|---|---|---|---|
| Value-Based | [23] | Q-learning | Smart Grid | Grid Power Pricing | Improves the profitability of electricity providers, reduces energy costs for customers. | 2018-Applied Energy |
| | [16] | Q-learning (multi-grid Q-learning) | HVAC | Energy saving and comfort | It learns faster than traditional q-learning and maintains good performance in the early stages. | 2015-IEEE International Conference on Automation Science and Engineering |
| | [68] | Q-learning | Smart Grid | Residential energy storage control with grid pricing | Cost savings of up to 117.6% and 72% respectively under two tariff functions (Compared to baseline). | 2016-IEEE Transactions on Sustainable Energy |
| | [35] | Q-learning | Wind Power Forecast | Wind Power Forecast | Can effectively integrate the three deep networks and improve prediction accuracy. | 2020-Energy |
| | [69] | Q-learning | Smart Grid | Maximum power tracking for photovoltaic power generation | Quick response and close to optimal behavior without requiring any prior knowledge (Perturb and Observe (P&O)). | 2017-Renewable Energy |
| | [70] | Q-learning | Smart Grid | Photovoltaic array parameter extraction | It can improve the accuracy of the extracted modeling parameters and extract them faster. | 2020-Energy Conversion and Management |
| | [24] | Q-learning (ANN instead of Q table) | Smart Grid | Grid Operation and Maintenance | With good approximation capability, it outperforms expert-based solutions in grid O&M management. | 2019-Applied energy |
| | [25] | Q-learning | Smart Grid | Photovoltaic systems | Q-learning can be adapted to any load profile. | 2003-Journal of Solar Energy Engineering |
| | [17] | Q-learning (Bayesian convolutional neural network + Q-learning) | HVAC | The balance of energy efficiency and comfort in residential air conditioning | Under the same conditions, the proposed method consumes 19.89 kWh in 50 h (rule-based control: 20.63 kWh) with an average discomfort degree of only 1.44 °C/h (rule-based control: 1.56 °C/h). | 2020-Applied Energy |
| | [43] | DQN | HVAC | Balancing different energy consumers | DQN reduces total energy consumption by 15.7%, while keeping indoor $CO_2$ levels below 1000 ppm. (Compared to baseline operation) | 2019-Science and Technology for the Built Environment |
| | [26] | Q-learning | Smart Grid | Optimum Tariff Offer Strategy (Temperature changes correlate with electricity price changes) | Find the optimal quotation strategy using a small amount of information from the real market. | 2006-Energy Conversion and Management |
| | [60] | Monte Carlo with Exploring Starts – MCES (Distributed Multi-Agent) | Water Heating | Constant temperature load | Save 20% of the energy used to produce hot water for each household, and in theory the actual savings should be even greater. | 2019-Applied Energy |
| | [18] | Q-learning (Bayesian learning to predict room occupancy) | HVAC | Thermostatic temperature control | Can achieve cost savings of 10% (compared to programmable thermostats) while maintaining high occupant comfort standards. | 2015-Springer-Verlag New York |
| | [44] | DQN | HVAC | Heating control | DQN-based thermostats improve thermal comfort by | 2020-Journal of Building Engineering |

(*continued on next page*)

(*continued*)

| Category | Paper | Algorithm | Scenario | Control object/ control target | Saving effect/Conclusion | Year-Journal Name |
|---|---|---|---|---|---|---|
| | | | | | 15% ~ 30% in simulated environments and reduce energy costs by 5% ~ 12% (compared to conventional thermostats). | |
| | [38] | Fitted Q-iteration | Battery | Self-consumption of photovoltaic production | 19% difference in performance between the proposed method and the optimal controller. | 2017-Energies |
| | [71] | Fitted Q-Iteration (Reduce the maximum power injection from PV systems to the grid) | Smart Grid | PV system | Compared to the default controller, the FQI controller is closer to the optimum, and adding predictive information improves FQI performance by 7%. | 2016-IEEE International Energy Conference |
| | [37] | Sarsa(λ) | Thermal energy storage | Ventilated curtain wall with phase change materials | Energy savings can be obtained under different test climatic conditions. | 2015-Energy and Buildings |
| | [72] | Q-learning (double Q-learning) | Battery | Battery Optimization | The proposed method leads to a near-optimal cost performance (96.9%). | 2020-Applied Energy |
| | [19] | Q-learning | HVAC | Temperature Setpoint | Regardless of the initial temperature setting, the temperature control strategy always achieves a temperature range that is comfortable for the occupants. | 2019-Building and Environment |
| | [36] | DQN | HVAC | Water supply temperature set point | For DQN controllers controlling room temperature by static and dynamic deployment, 5–12% heating energy savings can be obtained. | 2020-Energy & Buildings |
| | [73] | Q-learning (deep transfer Q-learning) | Smart Grid | Grid supply and demand | Efficient use of prior knowledge leads to rapid acquisition of optimal solutions for new tasks. | 2017-Energy |
| | [39] | Fitted Q-Iteration | Water hearing | Domestic water heater cluster | In a stochastic environment (unknown tap water distribution), the method was able to reduce the power cost of a cluster of 100 electric water heaters within a learning time of 40–45 days. | 2020-Energy Conversion and Management |
| | [74] | Q-learning | Smart Grid | Conversion of electricity and natural gas | Can reduce residential customers' energy bills and peak electricity loads by 20% and 24%, respectively. | 2016-Sustainable Cities and Society |
| | [22] | Q-learning | Water heating | Demand-side side management | Q-learning consumes about 26% less energy cost. | 2016-IEEE Transactions on Computer-Aided Design of Integrated Circuits and Systems |
| | [61] | Q-learning (Coordinated Multi-Agent) | Smart Grid & Battery | Improved utilization of batteries and solar energy | The method achieves the lowest possible cost of power generation compared to traditional distributed optimization methods and single agent Q-learning. | 2015-Procedia Computer Science |
| | [31] | Q-learning | Battery | Battery discharge strategy | It effectively reduces energy costs while ensuring user satisfaction. | 2011-Energy Procedia |
| | [75] | Q-learning (Expanded Joint Action Learning (eJAL)) | Smart Grid | Grid Uncertainty | The eJAL + Q-learning has the most significant reduction in overload time within the expected comfort constraints. | 2018-IEEE Transactions on Industrial Informatics |
| | [76] | DQN (double Q-learning) | HVAC | Energy consumption of air conditioning systems, thermal | The proposed method reduces $CO_2$ concentration by 10% and energy consumption by | 2019-Building and Environment |

(*continued on next page*)

(*continued*)

| Category | Paper | Algorithm | Scenario | Control object/ control target | Saving effect/Conclusion | Year-Journal Name |
|---|---|---|---|---|---|---|
| | | | | comfort, indoor air quality | 4–5% compared to existing control systems. | |
| | [77] | Q-learning (Introduction of modal calculations) | Smart Grid | Maximum power point tracking for photovoltaic power generation | Compared with seven typical MPPT algorithms, this method can obtain the optimal solution with less frequent system power fluctuations. | 2019-Energy |
| | [78] | Q-learning | heat storage in buildings | Electrically driven chilled water system | It is confirmed that RL control is a feasible approach to derive an optimal control strategy for this particular problem. | 2007-Journal of Solar Energy Engineering |
| | [20] | Q-learning (Lagrangian relaxation frame) | HVAC | Energy consumption and comfort | Significantly reduced computational effort and faster response time to events. | 2015-IEEE Transactions on Automation Science & Engineering |
| | [79] | MARL Fuzzy Q-learning | Smart Grid | Energy from independent microgrids | Demonstrates the effectiveness of a single agent control system component and the effectiveness of a multi agent system to ensure power supply and improve system reliability. | 2018-Applied Energy |
| | [80] | Deep RL (Data-driven) | Water Hearting | Residential comfort and energy consumption. | Can reduce the energy consumption of hot water production by about 20% in practical scenarios without affecting the comfort of the occupants. | 2018-Energy |
| | [32] | Q-learning | Battery | Coordination between different energy storage systems in a microgrid | Better system efficiency can be achieved, and the method can be extended to handle many types of batteries. | 2015-IEEE Transactions on Smart Grid |
| | [46] | Q-learning | HVAC | Cooling water system | The proposed method saves 11% of the system energy. | 2020-Energy & Buildings |
| | [81] | Fitted Q-Iteration | Heating System | Thermostatic temperature control | The proposed method achieves a performance of more than 65% of the available optimization potential after 40–60 days of learning. | 2017-Energy & Buildings |
| | [49] | DQN&DPG | Smart Grid | Scheduling of electrical equipment | In terms of cost reduction for the 48 buildings, DQN reduced the peak by 9.6% and minimized costs by 14.1%. | 2017-IEEE Transactions on Smart Grid |
| | [82] | Q-learning | Building Monitoring | On-line adjustment of monitoring controllers | RL-based monitoring controllers can improve the performance of complex low-energy building systems, correct misinformation generated by inaccurate offline simulations, and compensate for incorrect specifications of uncomfortable costs. | 2010-Control Engineering Practice |
| | [21] | Q-learning | HVAC | Coordination of natural air and HVAC | The proposed method reduces HVAC system energy consumption by 13% and 23%, discomfort hours by 62% and 80%, and high humidity hours by 63% and 77%. | 2018-Energy & Buildings |
| | [27] | Q-learning | Building Energy Management | Automated energy management system | Enables energy management systems to self-start operations and allows users to initiate more flexible requests. | 2015-IEEE Transactions on Smart Grid |

(*continued on next page*)

(*continued*)

| Category | Paper | Algorithm | Scenario | Control object/ control target | Saving effect/Conclusion | Year-Journal Name |
|---|---|---|---|---|---|---|
| | [33] | Novel dual iterative Q-learning | Battery | Optimal management and control of battery issues | The internal iteration minimizes the total cost of the electric load per period, while the external iteration allows the number of iterations to reduce the Q function convergence to the optimum. | 2015-IEEE Symposium on Adaptive Dynamic Programming and Reinforcement Learning |
| | [45] | DQN (a thermal comfort prediction model, combined with DQN) | HVAC | Control energy consumption | The proposed method reduces the total energy consumption of the variable refrigerant flow system by 32.2% (22 °C) and 12.4% (24 °C) compared to the fixed setpoint control. | 2019-Energy & Buildings |
| | [28] | Q-learning | Smart Grid | Optimal sequential decision making to maximize the user's own interests | Optimal sequential decisions can be made based on local fully observable information and environmental hidden information, with greater flexibility and adaptability and performance. | 2015-IEEE Green Energy & Systems Conference |
| | [40] | Fitted Q-Iteration | Water heating | Energy consumption | The proposed method is able to reduce the total energy cost of electric water heaters by 15% compared to thermostat controllers. | 2015-IEEE Transactions on Smart Grid |
| | [29] | Q-learning | Smart Grid | Energy management | Among the three proposed methods, RL combined with the adaptive method has the best performance. | 2019-Energy |
| | [58] | Q-learning (Tabular Q-learning and Batch Q-learning) | HVAC | Heating demand | All of RL controllers outperform rule-based controllers (RBC), meeting heating demand 100% of the time (97% for RBC) and maintaining optimal operating temperatures for heat pumps. | 2015-Applied Energy |
| | [83] | TD(λ)-Learning | Energy Storage Systems | Energy Storage Systems | The proposed method can achieve 59.8% energy saving and consumption reduction. | 2015-IEEE Consumer Communications and Networking Conference |
| | [30] | Q-learning | Smart Grid | Residential load dispatching \ load distribution | The use of the RL approach allows for efficient use of PV and the grid, and satisfies constraints that benefit consumers. | 2018-IEEE Systems Journal |
| | [34] | Q-learning | Battery | Battery discharge strategy | Q-learning is used to obtain the optimal discharge strategy that can effectively reduce energy costs while meeting user preferences. | 2015-IEEE Transactions on Smart Grid |
| | [84] | Q-learning (Monte-Carlo tree search + Q-learning) | Battery | the stochastic dispatch of battery energy storage systems | The algorithm asymptotically optimizes the dispatch policy and outperforms other algorithms. | 2019-Applied Energy |
| | [41] | Fitted Q-Iteration | Water Heating | Self-consumption of photovoltaic power generation | The proposed method enables a significant increase in the self-consumption of photovoltaic power compared to the default thermostat control. | 2018-IEEE |
| | [42] | Fitted Q-Iteration | Photovoltaic power | Maximize self-consumption of photovoltaic power | This method can significantly localize the instantaneous self-consumption of PV power. | 2019-Energy & Buildings |
| Policy-Based | [54] | AC (LSTM + AC) | HVAC | Optimize thermal comfort and energy consumption | The proposed approach improves thermal comfort by an average of 15% and energy | 2017-Processes |

(*continued on next page*)

(*continued*)

| Category | Paper | Algorithm | Scenario | Control object/ control target | Saving effect/Conclusion | Year-Journal Name |
|---|---|---|---|---|---|---|
| | | | | | efficiency by 2.5% compared to other strategies. | |
| | [47] | PPO | HVAC | HAVC control | Gnu-RL saves 16.7% in cooling requirements over existing controllers. | 2017-Energy |
| | [85] | AC | Smart Grid | Encourage staggered electricity consumption | The proposed algorithm can reduce the expected cost of users and the peak average ratio in the aggregated load by 28% and 13%, respectively. | 2017-IEEE Transactions on Smart Grid |
| | [55] | DQN&DDPG | HVAC | Operating parameters of central compressor chillers | Average weekly cost savings of 14%. | 2020-Energy & Buildings |
| | [52] | DDPG (Build a secure exploration approach) | Smart Grid | Optimal voltage control | 15% reduction in system losses compared to the uncontrolled case. | 2020-Applied Energy |
| | [53] | DDPG | Battery | Continuous charging and discharging of shared batteries | DDPG can automatically design effective control strategies for shared battery assets within a building complex, with the ability to minimize peak demand. | 2018-Journal of Mechanical Design |
| | [86] | DDPG | Thermal system | Superheat of the organic Rankine cycle (ORC) precisely under a transient heat source | Compared with the average error of 2.16 K for conventional control, the average error of reinforcement learning is only 0.19 K. | 2020-Applied Energy |
| | [65] | PPO | Energy Conversion | Wind Power Conversion | The proposed PPO-based renewable energy conversion algorithm can effectively reduce the operating cost of the system operator. | 2019-Energy Conversion and Management |
| | [66] | DDPG | Integration of electric and gas systems | Electrical gas system energy conversion | The algorithm improves the profitability of the system operator, reduces wind power curtailment, and effectively smooths the net load curve in real time. | 2020-Energy Conversion and Management |
| | [50] | DDPG | HVAC | Multi-zone residential HVAC systems | Can reduce energy costs by 15% and comfort violations by 79% | 2020-Applied Energy |
| | [51] | DDPG (Autoencoder (AE) extracts high-level features of DDPG state space, AE + DDPG) | HVAC | Short-term forecast of energy consumption | Compared with the popular supervisory model, the average absolute and root mean square errors of the AE-DDPG model were reduced by more than 22.46% and 25.96%, respectively. | 2019-International Journal of Refrigeration |
| | [67] | DDPG (Multi-Agent Deep Deterministic Policy Gradient, MADDPG) | Smart Grid | Power costs and grid stability | Able to reduce total power costs by 9.8% compared to not using a disaster recovery solution. | 2020-Applied Energy |
| | [87] | MADRL DDPG | HVAC | Energy Costs | Simulation results based on real traces show the effectiveness, robustness and scalability of MADRL. | 2020-IEEE Transactions on Smart Grid |
| | [49] | DQN&DPG | Smart Grid | Scheduling of electrical equipment | In terms of cost reduction for the 48 buildings, DPG reduced the peak by 26.3% and minimized costs by 27.4%. | 2017-IEEE Transactions on Smart Grid |
| | [48] | PPO-Clip | HVAC | Thermal comfort and energy efficiency | A maximum weekly energy reduction of 22% can be achieved compared to the baseline controller. | 2020-Energy and AI |

## References


[1] X. Cao, X. Dai, J. Liu, Building energy-consumption status worldwide and the state-of-the-art technologies for zero-energy buildings during the past decade, Energy Build. 128 (2016) 198–213.
[2] K. Mason, S. Grijalva, A review of reinforcement learning for autonomous building energy management, Comput. Electr. Eng. 78 (2019) 300–312.
[3] Z. Wang, T. Hong, Reinforcement Learning for Building Controls: the Opportunities and Challenges, 269, Applied Energy, 2020.
[4] Y. Wu, A. Mki, J. Jokisalo, R. Kosonen, B. Li, Demand response of district heating using model predictive control to prevent the draught risk of cold window in an office building, J. Build. Eng. 33 (3) (2021) 101855.
[5] M. Killian, M. Kozek, Ten questions concerning model predictive control for energy efficient buildings, Build. Environ. 105 (aug) (2016) 403–412.
[6] Ez, A. , Ma, A. , Dk, B. , Rs, A. , & Mm, A. . Energy saving potentials of a photovoltaic assisted heat pump for hybrid building heating system via optimal control. J. Build. Eng., 27(C), 100854-100854.
[7] R. Sutton, A. Barto, Reinforcement Learning:An Introduction, second ed., MIT Press, 2018.
[8] C. Watkins, Learning from Delayed Rewards, PhD thesis, Cambridge University, 1989.
[9] R.S. Sutton, Generalization in reinforcement learning : successful examples using sparse coarse coding, Neural Inform. Process. Syst. 8 (1996).
[10] M. Volodymyr, K. Koray, S. David, A.A. Rusu, V. Joel, M.G. Bellemare, et al., Human-level control through deep reinforcement learning, Nature 518 (7540) (2019) 529–533.
[11] R.S. Sutton, D. Mcallester, S. Singh, Y. Mansour, Policy gradient methods for reinforcement learning with function approximation, Submitted Adv. Neural Inform. Process. Syst. 12 (1999).
[12] J. Schulman, S. Levine, P. Moritz, M.I. Jordan, P. Abbeel, Trust region policy optimization, Comput. Sci. (2015) 1889–1897.
[13] S. David, L. Guy, H. Nicolas, D. Thomas, W. Daan, R. Martin, Deterministic policy gradient algorithms, in: Proceedings of the 30st International Conference on Machine Learning, 2014.
[14] T.P. Lillicrap, J.J. Hunt, A. Pritzel, N. Heess, T. Erez, Y. Tassa, et al., Continuous control with deep reinforcement learning, Comput. Sci. (2015).
[15] V. Mnih, Badia Adrià Puigdomènech, M. Mirza, A. Graves, K. Kavukcuoglu, Asynchronous Methods for Deep Reinforcement Learning, 2016.
[16] B. Li, L. Xia, A multi-grid reinforcement learning method for energy conservation and comfort of hvac in buildings, IEEE (2015) 444–449.
[17] C. Lork, W.T. Li, Y. Qin, Y. Zhou, T.K. Saha, An Uncertainty-Aware Deep Reinforcement Learning Framework for Residential Air Conditioning Energy Management, 276, Applied Energy, 2020.
[18] E. Barrett, S.P. Linder, Autonomous Hvac Control, a Reinforcement Learning Approach, ECML. Springer-Verlag New York, Inc, 2015.
[19] S. Lu, W. Wang, C. Lin, E.C. Hameen, Data-driven simulation of a thermal comfort-based temperature set-point control with ashrae rp884, Build. Environ. 156 (JUN) (2019) 137–146.
[20] B. Sun, P.B. Luh, Q.S. Jia, B. Yan, Event-based optimization within the Lagrangian relaxation framework for energy savings in hvac systems, IEEE Trans. Autom. Sci. Eng. 12 (4) (2015) 1396–1406.
[21] Y. Chen, L.K. Norford, H.W. Samuelson, A. Malkawi, Optimal control of hvac and window systems for natural ventilation through reinforcement learning, Energy Build. 169 (JUN) (2018) 195–205.
[22] K. Al-Jabery, Z. Xu, W. Yu, D.C. Wunsch, J. Xiong, Y. Shi, Demand-side management of domestic electric water heaters using approximate dynamic programming, IEEE Trans. Comput. Aided Des. Integrated Circ. Syst. 36 (5) (2017), 1-1.
[23] R. Lu, S.H. Hong, X. Zhang, A dynamic pricing demand response algorithm for smart grid: reinforcement learning approach, Appl. Energy 220 (JUN.15) (2018) 220–230.
[24] R. Rocchetta, L. Bellani, M. Compare, E. Zio, E. Patelli, A reinforcement learning framework for optimal operation and maintenance of power grids, Appl. Energy 241 (2019) 291–301.
[25] G.P. Henze, R.H. Dodier, Adaptive optimal control of a grid-independent photovoltaic system, J. Sol. Energy Eng. 125 (1) (2003).
[26] M.B. Naghibi-Sistani, M.R. Akbarzadeh-Tootoonchi, J.D. Bayaz, H. Rajabi-Mashhadi, Application of q-learning with temperature variation for bidding strategies in market based power systems, Energy Convers. Manag. 47 (11–12) (2006) 1529–1538.
[27] Z. Wen, D. O'Neill, H. Maei, Optimal demand response using device based reinforcement learning, IEEE Trans. Smart Grid 6 (5) (2015) 2312–2324.
[28] L. Ding, S.K. Jayaweera, Reinforcement learning aided smart-home decision-making in an interactive smart grid, in: Green Energy & Systems Conference, IEEE, 2015.
[29] N. Etim, D. Giaouris, C. Patsios, S. Papadopoulou, S. Gadoue, Reinforcement learning based adaptive power pinch analysis for energy management of stand-alone hybrid energy storage systems considering uncertainty, Energy 193 (2019) 116622.
[30] T. Remani, E.A. Jasmin, T. Ahamed, Residential load scheduling with renewable generation in the smart grid: a reinforcement learning approach, IEEE Syst. J. 13 (3) (2019) 3283–3294.
[31] B. Jiang, Y. Fei, Dynamic residential demand response and distributed generation management in smart microgrid with hierarchical agents, Energy Proc. 12 (39) (2011) 76–90.
[32] Q. Xin, A.N. Tu, M.L. Crow, Heterogeneous energy storage optimization for microgrids, IEEE Trans. Smart Grid 7 (3) (2015) 1453–1461.
[33] Q. Wei, D. Liu, G. Shi, L. Yu, G. Qiang, Optimal Self-Learning Battery Control in Smart Residential Grids by Iterative Q-Learning Algorithm, in: 2014 IEEE Symposium on Adaptive Dynamic Programming and Reinforcement Learning (ADPRL), IEEE, 2015.
[34] B. Jiang, Y. Fei, Smart home in smart microgrid: a cost-effective energy ecosystem with intelligent hierarchical agents, Smart Grid IEEE Trans. on 6 (1) (2015) 3–13.
[35] H. Liu, C. Yu, H. Wu, Z. Duan, G. Yan, A new hybrid ensemble deep reinforcement learning model for wind speed short term forecasting, Energy (2020), 202.
[36] S. Brandi, M.S. Piscitelli, M. Martellacci, A. Capozzoli, Deep reinforcement learning to optimise indoor temperature control and heating energy consumption in buildings, Energy Build. 224 (2020) 110225.
[37] A.D. Gracia, C. Fernández, A. Castell, C. Mateu, L.F. Ca Be Za, Control of a pcm ventilated facade using reinforcement learning techniques, Energy Build. 106 (2015) 234–242.
[38] M. Brida, R. Frederik, S. Fred, D. Geert, Battery energy management in a microgrid using batch reinforcement learning, Energies 10 (11) (2017) 1846.
[39] F. Ruelens, B.J. Claessens, S. Vandael, S. Iacovella, Demand response of a heterogeneous cluster of electric water heaters using batch reinforcement learning, in: Power Systems Computation Conference, IEEE, 2015.
[40] F. Ruelens, B. Claessens, S. Quaiyum, B.D. Schutter, R. Babuska, R. Belmans, Reinforcement learning applied to an electric water heater: from theory to practice, IEEE Trans. Smart Grid 99 (2015), 1-1.
[41] O.D. Somer, A. Soares, T. Kuijpers, K. Vossen, K. Vanthournout, F. Spiessens, Using Reinforcement Learning for Demand Response of Domestic Hot Water Buffers: a Real-Life Demonstration, IEEE, 2018.
[42] B. Asa, B. Dga, B. Fsa, B. Dea, C. Ods, B. Kva, Using Reinforcement Learning for Maximizing Residential Self-Consumption – Results from a Field Test - Sciencedirect, Energy and Buildings, 2017.
[43] K.U. Ahn, C.S. Park, Application of deep q-networks for model-free optimal control balancing between different hvac systems, Science and Technology for the Built Environment (2019) 1–16.
[44] A. Gupta, Y. Badr, A. Negahban, R.G. Qiu, Energy-efficient heating control for smart buildings with deep reinforcement learning, J. Build. Eng. (2020) 101739.
[45] Y.R. Yoon, H.J. Moon, Performance based thermal comfort control (ptcc) using deep reinforcement learning for space cooling, Energy Build. 203 (2019) 109420.
[46] A. Sq, A. Zl, B. Zla, A. Jl, C. Sl, X.L. D, Model-free control method based on reinforcement learning for building cooling water systems: validation by measured data-based simulation, Energy Build. 218 (2020).
[47] B. Chen, Z. Cai, Mario Bergés, Gnu-RL: A Precocial Reinforcement Learning Solution for Building HVAC Control Using a Differentiable MPC Policy. The 6th ACM International Conference, ACM, 2019.

[48] D. Azuatalam, W.L. Lee, F.D. Nijs, A. Liebman, Reinforcement Learning for Whole-Building Hvac Control and Demand Response, Energy and AI, 2020, p. 100020.
[49] E. Mocanu, D.C. Mocanu, P.H. Nguyen, A. Liotta, M.E. Webber, M. Gibescu, et al., On-line building energy optimization using deep reinforcement learning, IEEE Trans. Smart Grid (2017), 1-1.
[50] D.A. Yan, B. Hz, B. Ok, B. Kk, B. Jm, B. Ka, et al., Intelligent Multi-Zone Residential Hvac Control Strategy Based on Deep Reinforcement Learning, 281, Applied Energy, 2020.
[51] T.A. Liu, C.A. Xu, Y.B. Guo, H.A. Chen, A novel deep reinforcement learning based methodology for short-term hvac system energy consumption prediction, Int. J. Refrig. 107 (2019) 39–51.
[52] P. Kou, D. Liang, C. Wang, Z. Wu, L. Gao, Safe Deep Reinforcement Learning-Based Constrained Optimal Control Scheme for Active Distribution Networks, 264, Applied Energy, 2020.
[53] P. Odonkor, K. Lewis, Automated design of energy efficient control strategies for building clusters using reinforcement learning, J. Mech. Des. 141 (2018).
[54] Y. Wang, K. Velswamy, B. Huang, A long-short term memory recurrent neural network based reinforcement learning controller for office heating ventilation and air conditioning systems, Processes 5 (4) (2017) 46.
[55] T. Schreiber, S. Eschweiler, M. Baranski, D. Müller, Application of two promising reinforcement learning algorithms for load shifting in a cooling supply system - sciencedirect, Energy Build. 229 (2020).
[56] J. Wang, J. Wu, Y. Che, Agent and system dynamics-based hybrid modelling and simulation for multilateral bidding in electricity market, Energy 180 (AUG.1) (2019) 444–456.
[57] R. José, a b C.S. U. Vázquez-Canteli, d Jérme Kmpf, a. Zoltán Nagy, Fusing tensorflow with building energy simulation for intelligent energy management in smart cities - sciencedirect, Sustain. Cities Soc. 45 (2019) 243–257.
[58] L. Yang, Z. Nagy, P. Goffin, A. Schlueter, Reinforcement learning for optimal control of low exergy buildings, Appl. Energy 156 (OCT.15) (2015) 577–586.
[59] Tao, L. A. , Zt, B. , Cx, A. , Hc, A. , & Zl, A. . Study on Deep Reinforcement Learning Techniques for Building Energy Consumption Forecasting - Sciencedirect. Energy and Buildings, 208.
[60] B. Hka, B. Js, A. Ab, B. Jd, Multi-agent reinforcement learning for modeling and control of thermostatically controlled loads, Appl. Energy 238 (2019) 1022–1035.
[61] L. Raju, S. Sankar, R. Milton, Distributed optimization of solar micro-grid using multi agent reinforcement learning, Procedia Comput. Sci. 46 (2015) 231–239.
[62] J. Wang, J. Huang, Z. Feng, S.J. Cao, F. Haghighat, Occupant-density-detection based energy efficient ventilation system: prevention of infection transmission, Energy Build. 240 (5) (2021) 110883.
[63] P. Anand, C. Deb, K. Yan, J. Yang, D. Cheong, C. Sekhar, Occupancy-based energy consumption modelling using machine learning algorithms for institutional buildings, Energy Build. (2021), https://doi.org/10.1016/j.enbuild.2021.111478.
[64] M. Han, J. Zhao, X. Zhang, J. Shen, Y. Li, The reinforcement learning method for occupant behavior in building control: a review, Energy Built Environ. 2 (2) (2020).
[65] A. Bz, A. Wh, C.A. Di, H.A. Qi, C.B. Zhe, B. Fb, Deep Reinforcement Learning–Based Approach for Optimizing Energy Conversion in Integrated Electrical and Heating System with Renewable Energy - Sciencedirect, Energy Conversion and Management, 2019, p. 202.
[66] B. Zhang, W. Hu, J. Li, D. Cao, F. Blaabjerg, Dynamic energy conversion and management strategy for an integrated electricity and natural gas system with renewable energy: deep reinforcement learning approach, Energy Convers. Manag. 220 (6321) (2020) 113063.
[67] R. Lu, Y.C. Li, Y. Li, J. Jiang, Y. Ding, Multi-agent deep reinforcement learning based demand response for discrete manufacturing systems energy management, Appl. Energy 276 (2020) 115473.
[68] Y. Wang, X. Lin, M. Pedram, A near-optimal model-based control algorithm for households equipped with residential photovoltaic power generation and energy storage systems, Sustain. Energy, IEEE Trans. on 7 (1) (2016) 77–86.
[69] P. Kofinas, S. Doltsinis, A.I. Dounis, G.A. Vouros, A reinforcement learning approach for mppt control method of photovoltaic sources, Renew. Energy 108 (AUG) (2017) 461–473.
[70] J. Zhang, Y. Liu, Y. Li, K. Ding, J. Wu, A reinforcement learning based approach for on-line adaptive parameter extraction of photovoltaic array models, Energy Convers. Manag. 214 (2020) 112875.
[71] T. Leurs, B.J. Claessens, F. Ruelens, S. Weckx, G. Deconinck, Beyond Theory: Experimental Results of a Self-Learning Air Conditioning Unit. 2016 IEEE International Energy Conference (ENERGYCON), IEEE, 2016.
[72] P. Wu, J. Partridge, R. Bucknall, Cost-effective reinforcement learning energy management for plug-in hybrid fuel cell and battery ships, Appl. Energy 275 (2020) 115258.
[73] X. Zhang, T. Bao, T. Yu, B. Yang, C. Han, Deep transfer q-learning with virtual leader-follower for supply-demand stackelberg game of smart grid, Energy 133 (2017).
[74] A. Sheikhi, M. Rayati, A.M. Ranjbar, Demand Side Management for a Residential Customer in Multi-Energy Systems, Sustainable Cities & Society, 2016, pp. 63–77.
[75] L.A. Hurtado, E. Mocanu, P.H. Nguyen, M. Gibescu, I.G. Kamphuis, Enabling cooperative behavior for building demand response based on extended joint action learning, IEEE Trans. Ind. Inf. 99 (2018), 1-1.
[76] W. Valladares, M. Galindo, J. Gutiérrez, W.C. Wu, C.C. Wang, Energy Optimization Associated with Thermal Comfort and Indoor Air Control via a Deep Reinforcement Learning Algorithm, 155, Building and Environment, 2019.
[77] X. Zhang, S. Li, T. He, B. Yang, T. Yu, H. Li, et al., Memetic Reinforcement Learning Based Maximum Power Point Tracking Design for Pv Systems under Partial Shading Condition, Energy, 2019.
[78] S. Liu, G.P. Henze, Evaluation of reinforcement learning for optimal control of building active and passive thermal storage inventory, J. Sol. Energy Eng. (2006).
[79] P. Kofinas, A.I. Dounis, G.A. Vouros, Fuzzy q-learning for multi-agent decentralized energy management in microgrids, Appl. Energy 219 (JUN.1) (2018) 53–67.
[80] H. Kazmi, F. Mehmood, S. Lodeweyckx, J. Driesen, Gigawatt-hour scale savings on a budget of zero: deep reinforcement learning based optimal control of hot water systems, Energy 144 (2018).
[81] B.J. Claessens, D. Vanhoudt, J. Desmedt, F. Ruelens, Model-free Control of Thermostatically Controlled Loads Connected to a District Heating Network, Energy & Buildings, 2017. S0378778817303353.
[82] Y. Zhen, A. Dexter, Online tuning of a supervisory fuzzy controller for low-energy building system using reinforcement learning, Control Eng. Pract. 18 (5) (2010) 532–539.
[83] C. Guan, Y. Wang, L. Xue, S. Nazarian, M. Pedram, Reinforcement Learning-Based Control of Residential Energy Storage Systems for Electric Bill Minimization, IEEE, 2015.
[84] Y. Shang, W. Wu, J. Guo, Z. Ma, W. Sheng, Z. Lv, et al., Stochastic Dispatch of Energy Storage in Microgrids: an Augmented Reinforcement Learning Approach, 261, Applied Energy, 2020.
[85] S. Bahrami, V. Wong, J. Huang, An online learning algorithm for demand response in smart grid, IEEE Trans. Smart Grid (2017), 1-1.
[86] X. Wang, R. Wang, M. Jin, G. Shu, H. Tian, J. Pan, Control of superheat of organic Rankine cycle under transient heat source based on deep reinforcement learning, Appl. Energy 278 (2020) 115637.
[87] L. Yu, Y. Sun, Z. Xu, C. Shen, X. Guan, Multi-agent Deep Reinforcement Learning for Hvac Control in Commercial Buildings, 2020 arXiv e-prints.