

RESEARCH ARTICLE

Energy-efficient virtual sensor-based deep reinforcement learning control of indoor CO₂ in a kindergarten

Patrick Nzivugira Duhirwe, Jack Ngarambe, Geun Young Yun*

Department of Architectural Engineering, Kyung Hee University, Yongin 17104, Republic of Korea

Received 20 June 2022; received in revised form 4 October 2022; accepted 21 October 2022



KEYWORDS

Indoor air quality;
Indoor CO₂ control;
Machine learning;
Virtual sensor;
Deep reinforcement
learning

Abstract High concentrations of indoor CO₂ pose severe health risks to building occupants. Often, mechanical equipment is used to provide sufficient ventilation as a remedy to high indoor CO₂ concentrations. However, such equipment consumes large amounts of energy, substantially increasing building energy consumption. In the end, the issue becomes an optimization problem that revolves around maintaining CO₂ levels below a certain threshold while utilizing the minimum amount of energy possible. To that end, we propose an intelligent approach that consists of a supervised learning-based virtual sensor that interacts with a deep reinforcement learning (DRL)-based control to efficiently control indoor CO₂ while utilizing the minimum amount of energy possible. The data used to train and test the DRL agent is based on a 3-month field experiment conducted at a kindergarten equipped with a heat recovery ventilator. The results show that, unlike the manual control initially employed at the kindergarten, the DRL agent could always maintain the CO₂ concentrations below sufficient levels. Furthermore, a 58% reduction in the energy consumption of the ventilator under the DRL control compared to the manual control was estimated. The demonstrated approach illustrates the potential leveraging of Internet of Things and machine learning algorithms to create comfortable and healthy indoor environments with minimal energy requirements.

© 2022 Higher Education Press Limited Company. Publishing services by Elsevier B.V. on behalf of KeAi Communications Co. Ltd. This is an open access article under the CC BY-NC-ND license (<http://creativecommons.org/licenses/by-nc-nd/4.0/>).

* Corresponding author.

E-mail address: gyyun@khu.ac.kr (G.Y. Yun).

Peer review under responsibility of Southeast University.

1. Introduction

People are estimated to spend about 90% of their time in indoor spaces (Klepeis et al., 2001). Consequently, there has been an increasing need to provide sufficient indoor air quality (IAQ) that promotes health and the general well-being of building occupants. Carbon dioxide (CO₂) is a primary element in the assessment of IAQ (Morawska et al., 2021; Satish et al., 2012) – high CO₂ concentration levels are an indicator of poor indoor environments that could potentially be harmful to human health and lead to reduced productivity and deterioration of occupant well-being (Jacobson et al., 2019).

Pioneering studies have linked low-level exposure (e.g., 700–1000) to changes in the respiratory movement amplitude, increased in the flow of peripheral blood, and reduced functioning state of the cerebral cortex (Azuma et al., 2018). Such changes induced by exposure to CO₂, even in moderate concentrations, have also been observed in experimental studies. For instance, Satish et al. observed that decision-making capabilities subtly diminish when indoor concentration levels change from 600 ppm to 1000 ppm but that the hindrance in decision-making capabilities are significantly large when the concentration levels reach about 2500 ppm (Satish et al., 2012). Other experimental studies have linked low-to-moderate CO₂ concentration levels to reduced psychomotor performance (Allen et al., 2016, 2019), sick building syndrome (Redlich et al., 1997), and an array of other health conditions (Azuma et al., 2018). Recent studies have also shown that increased CO₂ concentration could amplify the rate at which airborne diseases are transmitted in indoor spaces – the current SARS – CoV-2 responsible for the recent global pandemic is a good example (Lewis, 2021).

To limit the effects of CO₂ concentrations on building occupants, maximum permissible thresholds of CO₂ are often dictated. In South Korea, for instance, the maximum permissible CO₂ level is set to 1000 ppm (Hwang et al., 2018). Similar thresholds have been dictated in Canada (Nathansan, 1993) and Japan (National Research Council, 1981). Keeping indoor CO₂ levels below dictated thresholds is critically important and has several implications regarding occupant health and well-being. This is more so in specific buildings such as educational facilities as they are often densely and frequently occupied and even perhaps more important for schools catering to small children (e.g., kindergartens) as they may be much more susceptible to the health risks associated with poor IAQ.

Sufficient air ventilation is the principal method of controlling indoor CO₂ concentrations and ensuring that it remains below dictated thresholds. While natural ventilation is possible in certain cases, mechanical equipment is required most of the time, especially in metropolitan cities where outdoor air contains high concentrations of atmospheric pollutants (Leung, 2015). The important role of mechanical equipment in maintaining safe indoor CO₂ levels coupled with the desperate and essential need for building managers/owners to reduce energy consumption often leads to high indoor CO₂ concentrations or unnecessary energy consumption; IAQ is foregone in an attempt to save energy or vice-versa. Consequently, the issue be-

comes an optimization problem that revolves around maintaining CO₂ levels below a certain threshold while utilizing the minimum amount of energy possible.

In the past, conventional control methods such as proportional integrated derivative (PID) and rule-based controls were employed in indoor environment control studies (Ryzhov et al., 2019). Although extensively applied in the field, such methods are, in certain cases, suboptimal given the complexity in modeling the components of indoor environments, particularly the non-linearities associated with controlling heating, ventilation, and air conditioning (HVAC) systems. For example, PID controllers become unstable when control gains are improperly chosen (Dounis and Caraiscos, 2009), and rule-based controls may not capture the non-linearities in indoor CO₂ concentrations resulting in suboptimal outcomes (Salsbury, 2005). Advanced modeling theories such as those based on model predictive control (MPC) have also been widely employed to provide cost-effective optimization of indoor environments (Ryzhov et al., 2019) – here, a dynamic system is understood through a series of mathematical constraints and optimizes the future system trajectories of the system based on established models of the said system. System models of indoor environments that are sufficiently extensive are not always available, and the optimal sequence of control signals which is computed at each control timestep is computationally expensive, limiting MPC applications. Moreover, interactively coupling real building data with predictive modeling to improve learning is often impossible under the MPC approach.

As real building data becomes easily available through the Internet of Things (IoT) systems and highly calibrated sensors, some reports have drawn attention to the potential usefulness of data-driven agent-based intelligent control systems, particularly reinforcement learning (RL), in providing sufficient IAQ (Yang et al., 2021). Such agents provide continuous learning to improve control systems and usually require no existing models of the control problem at hand. However, the application of RL to provide cost-effective solutions to poor indoor air quality has not been extensively explored in the literature. Moreover, only a few studies have demonstrated the importance of advanced control techniques using actual field experimental data. The present study thus had two central objectives. The first objective was to employ deep reinforcement learning (DRL) and assess its potential in reducing the energy consumption of a heat recovery ventilator (HRV) while at the same time maintaining indoor CO₂ concentration levels below the maximum permissible level of 1000 ppm. The second objective was to illustrate the use of a virtual sensor as an alternative performance feedback platform when training the DRL control scheme. This is particularly important because previous studies have often used simulation tools as the feedback platform to train DRL control schemes. However, simulating real environments is computationally expensive and is associated with other complexities inherent in modeling tasks. Consequently, the use of a virtual sensor avoids the high computation load and complexities associated with simulated environments during DRL training. To the best of our knowledge, this is the first study that utilized a trained machine learning (ML) model as the basis for the performance feedback system to the

DRL-based control agent. The present study contributes to the small but increasing number of studies showcasing the suitability of autonomous smart agent controls to improve building performance.

The rest of the article is arranged as follows: Section 2 briefly discusses the theory of RL and the elements of RL pertinent to the present study, Section 3 describes the case study space, indoor CO₂ predictions, and the development of a virtual sensor, Section 4 presents and discusses the obtained results, while Sections 5 and 6 discuss future research and provide conclusive remarks, respectively.

2. Reinforcement learning

2.1. Theory

RL is a subcategory of ML algorithms in which the learning system, also known as the agent, learns how to interact with its environment to reach a defined objective (Glorennec, 2000; Sutton and Barto, 2018). In RL problems, the agent learns how to map its actions to the changing environment to maximize a scalar reward signal, also known as the reinforcement. RL problems are closed-loop problems in an essential way because the agent's actions have an influence on its future actions. Contrary to other ML algorithms, in RL, the agent is not told which actions it should take, instead, it should find those actions by trial and error, and these actions taken may not only affect the contiguous rewards, but also influence all subsequent rewards. The above properties are the main discerning features of RL problems.

RL is different from supervised learning. In supervised learning, the learning system is provided with labeled examples to train on by a supervisor and learned how to map inputs to the labels. The goal of the learning system is to generalize its results when presented with data that was not part of the training dataset. RL is also different from unsupervised learning, where the learning system is given an unlabeled dataset, and the goal is to discover hidden structures or patterns. Whilst on the might of RL as a type of unsupervised learning as it does not depend on examples of correct actions to take, RL problems are about maximizing the total cumulative rewards instead of discovering hidden patterns. Discovering hidden patterns may be part of the agent's learning process but by itself does not contribute to cumulative reward maximization. Thus, RL is considered as a third ML subcategory, besides supervised learning and unsupervised learning, possibly other subcategories as well.

2.2. Markov Decision Processes and the Bellman equation

Markov Decision Processes (MDPs) are the fundamental mathematical formalism of any RL problem (van Otterlo and Wiering, 2012). They are composed of (i) states set (S_t), which describe any information available to the agent at time t , (ii) actions set (A_t), which are the decisions chosen by the agent to induce changes in the environment at time t , (iii) transition probability (p) from one state to

another, and (iv) the reward function (R_t), which defines the objective of the RL problem. The Markov property describes that the state of the environment at $t + 1$ only depends on the state and actions at t , and the environment dynamic changes can be defined by equation (1) (Sutton and Barto, 2018).

$$p(s_{t+1}, r|s_t, a) = Pr\{R_{t+1} = r, S_{t+1} = | S_t, A_t \} \quad (1)$$

To maximize the total cumulative rewards, the agent consistently chooses the correct actions and avoids the wrong ones, responding to the environment changes at each time step. This process of choosing actions is known as the policy. Thus, the process that yields the highest rewards is referred to as the optimal policy. The optimal policy is computed by learning the value functions. The value functions inform the agent of the goodness of taking an action a while in the state s . The Bellman equation (Bellman, 1952) is used to estimate these value functions, and the value of choosing a particular action in a particular state is computed based on the immediate rewards and the future discounted rewards that depend on the discounting factor (γ). Therefore, the optimal value function (q_*), computed by the Bellman optimality equation (Bellman, 1966; Sutton and Barto, 2018), yields the optimal policy that the agent follows to obtain the highest cumulative rewards, which is given by equation (2).

$$q_* = \sum_{s_{t+1}, r} p(s_{t+1}, r|s_t, a) = \left[r + \gamma \max_a q_*(s_{t+1}, a) \right] \quad (2)$$

2.3. Double Q learning

Q-learning (Aryana et al., 2021) is a popular model-free RL algorithm and widely used to find the optimal solution for MDPs. To find the solution of any MDP, Q-learning uses a Q-table that records Q-values associated with each action taken. The update rule for Q-learning value function $Q(s, a; \theta_t)$ for taking an action A_t in a state S_t and evaluating the immediate reward R_{t+1} and next state S_{t+1} is shown in equation (3) where α is the learning rate.

$$\theta_{t+1} = \theta_t + \alpha \left[\left(R_{t+1} + \gamma \max_a Q(S_{t+1}, a; \theta_t) \right) - Q(S_t, A_t; \theta_t) \right] \cdot \nabla_{\theta_t} Q(S_t, A_t; \theta_t) \quad (3)$$

Q-learning has two major drawbacks: (i) it is computationally expensive to maintain the Q-table with an increased number of actions and states, and (ii) it suffers from poor performance caused by an overestimation of action values when solving a stochastic MDP. To solve these drawbacks, a deep Q-learning (DQN¹) (Mnih et al., 2015) algorithm was developed. The DQN uses a deep neural network (DNN) (LeCun et al., 2015) in place of a Q-table and takes the current state as input to provide a vector of possible estimated action values. Moreover, there was an introduction of the experience replay and a target network with parameters θ^- to increase the DQN performance. The update rule of a DQN is shown in equation (4). The DQN also

¹ DQN: deep Q-learning.

showed the same overestimation issue as Q-learning. This overestimation behavior is caused by the maximization function over calculated action values, which tends to choose large action values over low action values (van Hasselt et al., 2016).

$$\theta_{t+1} = \theta_t + \alpha \left[\left(R_{t+1} + \gamma \max_a Q(S_{t+1}, a; \theta_t^-) \right) - Q(S_t, A_t; \theta_t) \right] \cdot \nabla_{\theta_t} Q(S_t, A_t; \theta_t) \quad (4)$$

Double Q-learning was proposed as a solution for the overestimation problem (van Hasselt et al., 2016). Instead of using the same value in the maximization function to both pick and assess an action, the double Q-learning learns two value functions, and each update provides two distinct sets of weights, θ for action selection (as in equation (3)) and θ' for action evaluation as shown in equation (5). These two sets of weights play an important role in determining the policy and its values.

$$\theta_{t+1} = \theta_t + \alpha \left[\left(R_{t+1} + \gamma Q(S_{t+1}, \operatorname{argmax}_a Q(S_{t+1}, a; \theta_t); \theta_t') \right) - Q(S_t, A_t; \theta_t) \right] \nabla_{\theta_t} Q(S_t, A_t; \theta_t) \quad (5)$$

This study utilized the double Q-learning as the control algorithm for indoor CO₂ because of its capabilities of handling stochastic environments with complex interactions, involving trade-offs between the maintenance of adequate indoor CO₂ and minimization of ventilation energy consumption. These complexities are affected by the number of occupants and activities of the day, which call for a sophisticated and robust algorithm. Therefore, the double Q-learning was a viable solution for our environment MDPs. Interested readers are referred to (Arulkumaran et al., 2017; Kiumarsi et al., 2018; Yang et al., 2020) for more details on the aforementioned algorithms.

3. Methods

Fig. 1 illustrates the stages involved in the research framework. Each stage is detailed in the subsequent sections.

3.1. Site and data collection

The experimental site was a daycare center located in Seoul, Republic of Korea (37° 21' 57.13"N 126° 57' 32.69"E). The center hosted children between the ages of one and seven years, from Monday to Saturday, and the working hours were from 07:30 a.m. to 07:30 pm.

An HRV equipped with sensors to record indoor and outdoor air conditions was installed in one of the rooms of the daycare center. The floor plan of the room and the position of the HRV are shown in Fig. 2. The sensors recorded CO₂, PM2.5, PM10, temperature, and relative humidity of indoor and outdoor air. Fig. 3 shows the position of the sensors in the installed HRV. Moreover, measurements related to the operation status (On or Off), air supply

mode (Low, Medium, and High), and ventilation energy were recorded. All the real-time data were recorded in intervals of 10 seconds and saved in an established Dynamo database on Amazon web services and later synchronized to 30 minute intervals. Table 1 elaborates the recorded variables and Table 2 shows manufacturer details of the installed HRV.

3.2. Estimating energy saving potential by an HRV

HRVs are types of air conditioning systems that utilize air-to-air exchangers and recover heat or coolness from the stale exhaust air to the fresh supply air. This air-to-air exchange also helps remove odor, excess moisture, and air contaminants by maintaining thermal comfort while conserving energy.

The energy consumed by the unducted HRV (ISO 16494: 2014, 2014) employed in our study was tested and calculated following the ISO16494 (ISO 16494: 2014, 2014). The total consumed energy is expressed in terms of the net supply airflow (Q_{SAnet}), the coefficient of energy (COE), the coefficient of performance (COP), and the effective work (EW). Fig. 4 shows the airflows through an HRV and equations (6)–(11) (ISO 16494: 2014, 2014) summarise the total energy derivation. Interested readers may refer to (ISO 16494: 2014, 2014) for more details.

As the ventilation system installed in the kindergarten was a HRV, the presented equations were primarily used in estimating the overall energy consumption of the system while accounting for the energy reduced by the usage of a heat exchanger.

$$Q_{SAnet} = Q_1 - Q_2 \quad (6)$$

$$COE = \frac{|qm_{2,net}(h_2 - h_1)| \times 1000}{P_{in}} \quad (7)$$

$$EW = P_{in} \times (COE - 1) \quad (8)$$

$$E_{total,HRV} = \frac{Q_{SAnet}}{COP} - \frac{EW}{COE} + P_{in} \quad (9)$$

where Q is the airflow (m³/s), qm is the net supply mass flow rate (kg/s), h is the enthalpy of the air (kJ/kg of dry air), and P_{in} is the input power of the ventilator (W).

If the heat exchanger is not turned on, the total energy consumed is:

$$E_{total} = \frac{qm_2 \times |h_1 - h_3|}{COP} \quad (10)$$

The total saved energy using the HRV is, therefore:

$$E_{saved} = E_{total} - E_{total,HRV} \quad (11)$$

3.3. Deep reinforcement learning (DRL) for indoor CO₂ control

3.3.1. Overall process flow

Fig. 5 illustrates the design of the developed virtual sensor coupled with DRL for indoor CO₂ control. The role of the virtual sensor is to emulate environment changes,

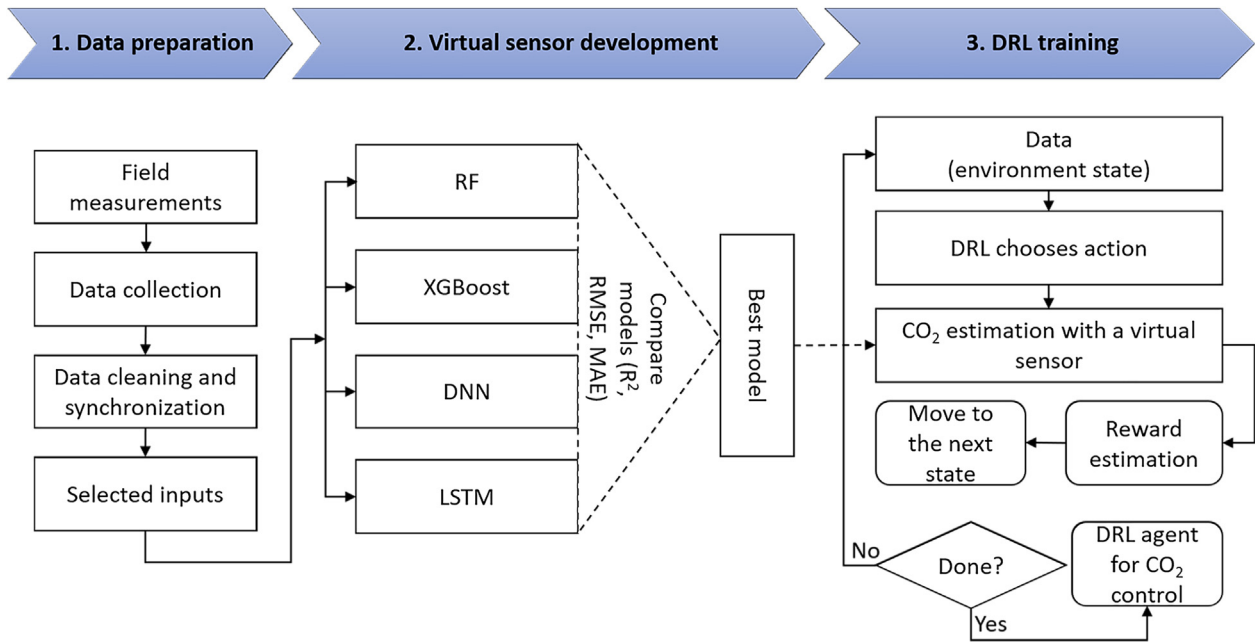


Fig. 1 Research framework.

in this case, the changes in indoor CO₂ in the kindergarten. The DRL part is responsible for maintaining indoor CO₂ below 1000 ppm by making decisions and forwarding these decisions to the virtual sensor for evaluation. The overall workflow during the design is described in the following steps – as subtly discussed in the introduction, this unique approach of combining a virtual sensor and DRL reduces the time, modeling, and computation complexities, consequently turning the focus on the DRL control development which facilitates

the development of various configurations of the double Q-learning.

- (i) The virtual sensor module allocates memory for holding all the information related to changes in the environment after each action is taken.
- (ii) The agent reads a single observation from the record holding the training data and provides the best action accordingly.

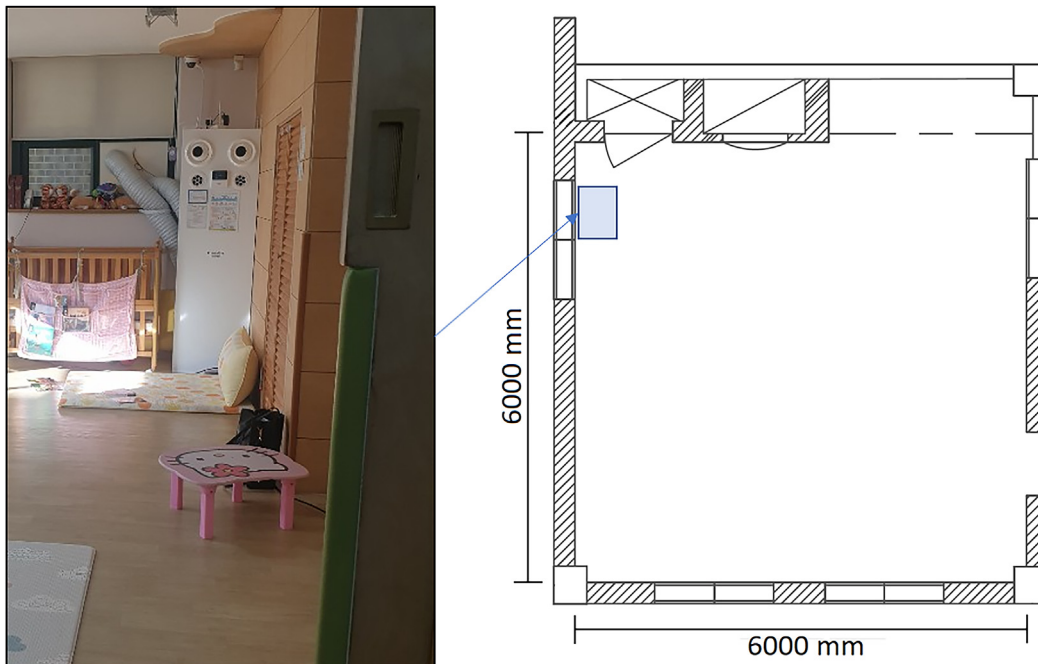


Fig. 2 Floor plan of the case study kindergarten area.

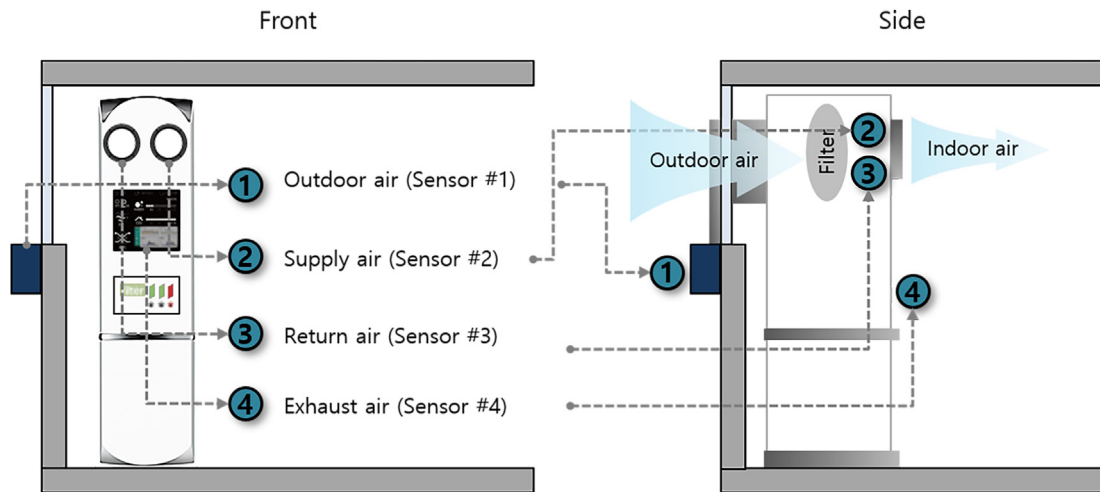


Fig. 3 Location of sensors in the HRV.

Table 1 Details of the recorded variables.

Variable	Unit	Device	Range (Accuracy)
CO ₂	Parts per million (ppm)	MH-Z19B	400–2000 (±30 ppm)
PM2.5	Micrograms/cubic meter (µg/m ³)	PM2008M	0–100 µg/m ³ (±10 µg/m ³)
PM10	Micrograms/cubic meter (µg/m ³)	PM2008M	0–100 µg/m ³ (±10 µg/m ³)
Temperature	Degree Celsius (°C)	SHTC1	–30–100 °C (±0.3 °C)
Relative humidity	Percentage (%)	SHTC1	0–100% (±3%)
Power consumption	Watt (W)	Built-in meter	–
Ventilation mode	–	Built-in meter	–

Table 2 Manufacturer information of the installed HRV.

Model	HRD-EG400S		
Power	220 V × 60 Hz		
Ventilation mode	High	Medium	Low
Air volume (m ³ /h)	400	250	150
Power consumption (W)	160	75.5	41
Noise (dB)	50 or less		
Effective heat transfer efficiency (%)	Heating	71	
	Cooling	59	
Filter	Dust (2) + Medium (1) + HEPA (1)		
Weight (kg)	70		
Product size (mm)	545 (Width) × 420 (Depth) × 1780 (Height)		
Option	Deodorization filter		

- (iii) The action chosen in step (ii) is forwarded to the virtual sensor to assess the changes in indoor CO₂ from which rewards are provided. These changes are appended to the memory created in (i).
- (iv) Based on how indoor CO₂ has changed, the rewards are provided, and the agent adjusts the learning parameters accordingly.
- (v) The agent reads the next record in line as in step (iii), and the workflow continues until all the records have been updated.

3.3.2. Virtual sensor to estimate indoor CO₂

The general idea of a virtual sensor is to provide an estimate of a parameter based on the information extracted from other measurements that the parameter of interest depends on. Virtual sensors, therefore, provide a solution to the limitations (time, costs, complexities) of physical sensors or simulation-based estimations through ML processing. To overcome the high computational loads and modeling complexities associated with computer simulation tasks in estimating indoor CO₂, the performance feedback

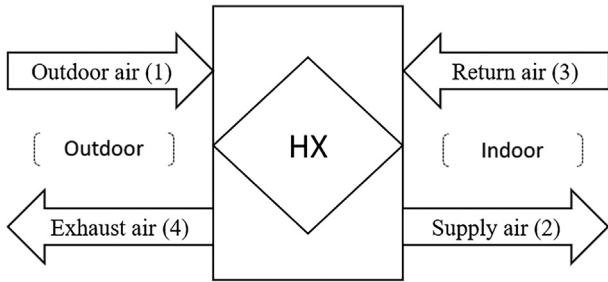


Fig. 4 Airflows for the HRV

XGBoost is an advanced gradient boosting tree algorithm through which accurate predictions are achieved from a combination of weak learners added using the gradient descent method (Chen and Guestrin, 2016). The learning is conducted in a sequential fashion by developing several models and concentrating on those observations in the data that are difficult to predict, also known as the boosting approach. Each additional model is solely added to improve the accuracy of the previous model based on the objective function (Friedman, 2002; Schapire, 1990). XGBoost has several advantages over regular gradient boosting tree al-

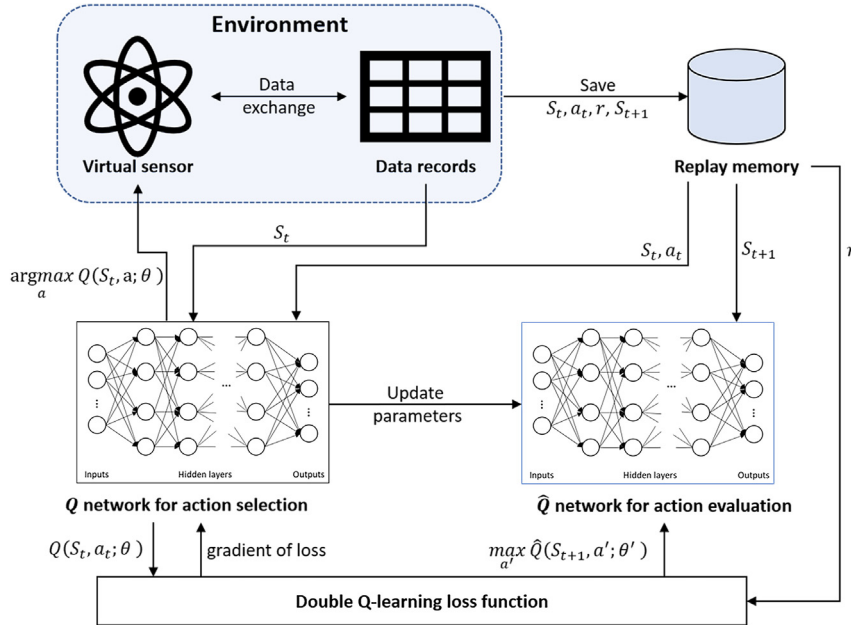


Fig. 5 Virtual sensor coupled with DRL design approach.

to the developed control agent was based on a data-driven ML virtual sensor. Several ML algorithms were employed, and their performances were evaluated to choose the best algorithm to couple with DRL. These algorithms are the random forest (RF), extreme gradient boosting (XGBoost), deep neural networks (DNN), and long short-term memory (LSTM) networks.

RF is an algorithm developed by Breiman (2001), and it employs classification and regression trees (CART²) to make predictions. These trees are constructed randomly by selecting a subset of the training dataset with replacement. This is also known as bagging. The other subset not selected is used by the algorithm for the internal cross-validation process to provide an accurate evaluation. The training and cross-validation subsets are known as in-bag and out-of-bag sets. New trees are created independently, without pruning, based on the out-of-bag error and the final results are estimated from the aggregation of all trees' estimates. Interested readers may refer to (Biau and Scornet, 2016; Breiman, 2001) for more details.

gorithms and has gained popularity among ML practitioners. These advantages include the regularization methods (Lasso and Ridge) to avoid overfitting/underfitting, parallel processing for faster computations, high flexibility in customizing the objective function, capabilities in handling missing values, tree pruning for early stopping, and cross-validation mechanism. Readers are referred to (Chen and Guestrin, 2016) for more information.

DNN (Montavon et al., 2018), also known as deep feed-forward neural networks or multilayer perceptron is a DL algorithm developed to overcome the generalization drawbacks of traditional ML algorithms. DNN uses the backpropagation approach to learn non-linear patterns existing in the data. They are called feed-forward due to the unidirectional flow of information from the input layer to the output layer passing through hidden layers and the model's output is not forwarded back to itself. The non-linear mappings are learned using the activation function (Nwankpa et al., 2018), and inputs weights are adjusted by the optimizer (Ruder, 2017) to minimize the loss. DNN has gained popularity with big data availability where traditional ML failed and with the advances in computing

² CART: classification and regression trees.

resources. Interested readers may refer to (Goodfellow et al., 2016; Montavon et al., 2018) for detailed information.

LSTM networks were developed by (Hochreiter and Schmidhuber, 1997) as an improvement of traditional recurrent neural networks which face vanishing or exploding gradient issues in the presence of long-term dependencies. They are composed of memory blocks capable of storing past information from time-series data. Each LSTM unit has three types of gates that govern the flow of information in the model. These are the forget gate, the input gate, and the output gate. The forget gate determines the information from the past cell state to be retained; the input gate updates the cell state by conditioning the current information to be stored. The information from the input gate is forwarded to the output gate which determines the portion of information to be sent to the next layer or output cell. Interested readers may refer to (Goodfellow et al., 2016; Hochreiter and Schmidhuber, 1997) for detailed information.

The inputs for each algorithm were the time of the day in minutes, the outdoor CO₂, the ventilation mode, and previous indoor CO₂ (See Table 3). For receding timesteps, the models relied on the previous 1-h values (two receding timesteps) to predict the indoor CO₂ values 30 min ahead. Hyperparameters that demonstrated to affect the performance for each algorithm were manually tuned to obtain the best models. For the RF model, the tuned hyperparameters were the maximum level or depth (max_depth) of each tree, the minimum number of observations required for a node split (min_sample_split), the maximum number of nodes in a tree (max_leaf_nodes), the maximum number of trees (n_estimators), and the fraction of the training set given to each tree (max_samples). For the XGBoost, the hyperparameters were the maximum level or depth (max_depth) of each tree, the learning rate (eta), and the percentage of training variables to choose while creating a new tree (col_sample_bytree), the percentage of observations to choose while creating a new tree (subsample), and Ridge regression (λ) as the regularization parameter. For DNN and LSTM, the hyperparameters were the number of hidden layers, the number of neurons in each layer, the activation function, the optimizer, the batch size, and the learning rate.

3.3.3. Controlling indoor CO₂ with DRL

3.3.3.1. Environment states. The sensors attached to the ventilation system recorded variables that provided the IAQ status at the kindergarten. The variables of interest were the time of the day in minutes and outdoor CO₂ levels. These recorded variables, along with indoor CO₂ provided by the virtual sensor in response to the ventilation mode chosen, represented the state of the environment for the designed control agent. The time of the day was an important variable as it provided information about peak and low activity hours, such as playtime and sleep time, respectively, that are likely to result in increased indoor CO₂ levels. Thus, with various states, the double Q-learning agent learned all the expectations about indoor CO₂ changes.

3.3.3.2. Control actions. Indoor CO₂ can only be reduced through ventilation by exchanging indoor contaminated air

Table 3 Variables used for indoor CO₂ virtual sensor.

Inputs (current and past values)		Output (30 min ahead)
Variable	Range	
Time of the day (in minutes)	0–1410	Indoor CO ₂
Outdoor CO ₂	400–711.5 [ppm]	
Ventilation mode	0, 1, 2, 3	
Indoor CO ₂	400–2098.3 [ppm]	

with clean outdoor air with low CO₂ concentrations. Consequently, for this study, the controlling variable (agent's actions) was the ventilation rate composed of discrete set of values $V = \{0, 1, 2, 3\}$ where 0 is off, 1 for low volume ventilation rate, 2 for medium volume ventilation rate, and 3 for high volume ventilation rate. Please refer to Table 2 of section 3.1 for details on air volume associated with these ventilation rates.

3.3.3.3. Reward function. The reward function was designed to balance the trade-off between indoor CO₂ concentration levels and ventilation energy consumption. The approach was to reward the agent with small rewards every time the action taken yielded indoor concentration levels below 1000 ppm and impose a heavy penalty when the action taken resulted in indoor CO₂ concentration levels above the limit. This approach was necessary to avoid the cobra effect in RL (Itri et al., 2019). That is, avoiding that our reward function does not make the problem worse instead of solving it.

To include the notion of ventilation energy consumption, the reward shaping technique (Grzes and Daniel, 2008; Laud, 2004) was used. This technique provides a means to incorporate domain knowledge, which is the power consumed by the ventilation fan for each ventilation mode. Equation (12) describes our employed reward function. It was designed as such to alleviate the need for assigning weight factors manually.

$$\text{Reward} = \begin{cases} 1 - \text{scaled ventilation energy,} \\ \quad (\text{if indoor CO}_2 \leq 1000 \text{ ppm}); \\ -20 - \text{scaled ventilation energy,} \\ \quad (\text{if indoor CO}_2 > 1000 \text{ ppm}). \end{cases} \quad (12)$$

3.3.3.4. Training the DRL agent. The DRL agent was trained to recognize possible indoor air CO₂ changes and take actions accordingly. The training process starts with the exploration phase. In this phase, the agent gathers necessary information on how to potentially change the environment by trying random actions. Here, at each time, the agent records in the memory the state before executing an action, the action for that state, the state after executing the action, and the obtained reward. This phase permits the agent to improve its current knowledge about each action to make more adequate decisions in the future.

After gathering enough experience, the training process moves forward with the exploitation phase.

In the exploitation phase, the agent gets the state from the environment and picks the best corresponding action. The agent uses the knowledge gained and transitions from exploration to exploitation following a probability ϵ of choosing a random action. ϵ is also referred to as the greedy factor, and exploiting an action that has yielded a good cumulative reward is based on the probability of $1 - \epsilon$ of that action. As the training proceeds, the probability ϵ decreases also and the agent learns the best policy that maximizes the overall cumulative rewards. ϵ never reaches 0, and this implies that there will always be a chance of picking a random action along the training course. In this study, the agent was trained on two months' data (December 2020 and February 2021) and tested on one-month data (January 2021).

4. Results and discussion

4.1. Temporal variations in indoor CO₂ concentrations and HRV usage

Fig. 6 illustrates typical temporal CO₂ concentration variations within the kindergarten. The figure shows that daily CO₂ concentration levels tended to vary from slightly below 500 ppm to about 2000 ppm. The lowest concentrations were primarily observed in the early morning hours (e.g., 8 a.m.) and late evening hours (e.g., after 5 p.m.). Peak indoor CO₂ concentrations (e.g., approximately 2000 ppm) were seen around mid-day. Additionally, the CO₂ levels tended to be mostly below the maximum permissible threshold (e.g., 1000 ppm) in the early mornings. However, they gradually increased past the threshold between midday and early afternoon before decreasing in the late evenings. The observed CO₂ patterns were somewhat expected and can be explained by the occupancy schedule and the activities within the kindergarten – the school activities roughly began in the mid-morning, and the increase in occupancy (i.e., as children report to the classroom) explains the gradual increase in CO₂ concentrations. As the space houses children, daily routine activities such as lunch and playtime are conducted inside the same room, drastically increasing indoor CO₂ concentrations during lunchtime. Occupancy is also likely to increase during lunch hours due to the presence of extra personnel (e.g., cooks). The figure also shows the fan operation mode during the day. Three interesting observations can be deduced from the CO₂ concentration trend and the fan operation mode.

First, it is observed that the fan was either off (i.e., Mode 0) or operated at the maximum (i.e., Mode 3) – this tends to indicate that, if the fan is operated, it was mostly operated at the maximum capacity. Operating the fan at the maximum capacity has potentially significant implications in terms of consumed energy and perhaps points to the main limitation of manual control and the need for efficient automated control; unnecessary energy consumption.

Secondly, observing the CO₂ concentration patterns showed instances when the CO₂ concentrations were above

the maximum allowable level, yet the fan was not set to operation. For example, as seen in the figure, at peak CO₂ levels (i.e., 2000 ppm), which was often achieved at midday, the fan is off and perhaps the reason for the considerably high CO₂ levels. This further points to the importance of automated controls for IAQ control as manual control relies mainly on the occupants' assessment of IAQ, which is likely inaccurate as it is based mainly on the perception of the environment and subjective assessment rather than the quantifiable objective assessment of the environment.

Thirdly, it was observed that even when the ventilator was used, there was a considerable lag between the instant the fan was switched on and when the CO₂ concentrations reduced to acceptable levels (i.e., below 1000 ppm). For instance, we estimated a 30 min to 2 h lag between the time the ventilator was turned on and when the CO₂ concentration reduced to acceptable levels. This observation points to the importance of preemptive control of CO₂ to always maintain acceptable levels.

4.2. Modeling CO₂ concentration levels and the basis for the virtual sensor

As discussed in section 2, multiple ML algorithms to forecast CO₂ concentration levels were developed, thus forming the basis for our virtual environment that provided performance feedback to the double Q-learning agent. Four prediction models were initially developed and tested based on four different cutting-edge ML algorithms categorized into tree-based (i.e., RF and XGBoost) and DL-based (i.e., DNN and LSTM). The models were optimized by determining important hyperparameters and tuning said parameters to obtain the best-performing model variants. Table 4 shows the determined hyperparameters and the inherent elements that provided the best model performance.

Furthermore, among the four considered ML algorithms, the XGBoost algorithm best predicted CO₂ concentration levels both on the training dataset ($R^2 = 0.998$) and test dataset ($R^2 = 0.992$) (see Table 5). The better performance of XGBoost than the other considered ML algorithms was also supported by lower predictive errors compared to the other algorithms; training phase (RMSE = 2.891, MAE = 1.575) and testing phase (RMSE = 4.008, MAE = 3.846). Table 5 shows the comparative performance of the developed models. This finding is commensurate with previous studies comparing the predictive performance of ensemble tree algorithms, particularly XGBoost and deep-learning-based algorithms such as the DNN and LSTM on tabulated data (Shwartz-Ziv and Armon, 2022; Zamani Joharestani et al., 2019). The better performance of XGBoost than the other algorithms, particularly DNN and LSTM, might be the fact that tree-based algorithms are deterministic (i.e., the parameters are fit to guide the flow of information) while DL algorithms are probabilistic (i.e., the parameters are fit to transform the given inputs and indirectly guide the activations of subsequent neurons). Thus, the inherent simplicity of XGBoost models provides automatic feature selection and reduces their selection bias, in turn improving their performance (Shwartz-Ziv and

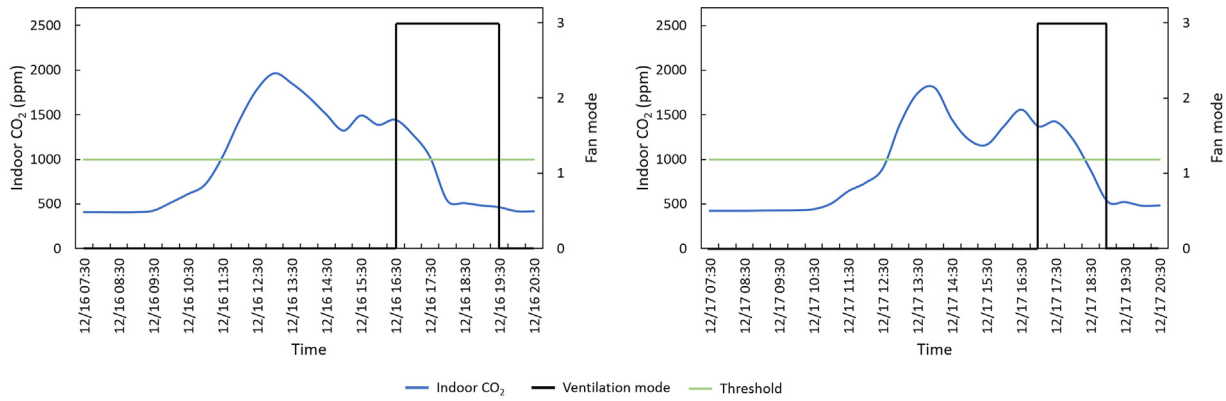


Fig. 6 Temporal variations in indoor CO₂ concentrations and HRV usage.

Table 4 Best hyperparameters for the developed models.

Model	Hyperparameter	Best value
RF	max_depth	5
	min_sample_split	2
	max_leaf_nodes	100
	n_estimators	50
	max_samples	0.8
XGBoost	max_depth	10
	eha	0.01
	colsample_bytree	0.8
	subsample	0.8
DNN	λ	1
	Hidden layers and neurons	[64, 64, 32, 32]
	Activation	[ReLU, Tanh, ReLU, ReLU]
	Optimizer	RMSPProp
	Batch size	32
LSTM	Learning rate	0.001
	Hidden layers and neurons	[128, 64, 32, 32]
	Activation	[ReLU, ReLU, ReLU, Swish]
	Optimizer	Adam
	Batch size	32
	Learning rate	0.001

Table 5 Performance results for the developed models.

Algorithm	Train			Test		
	R ²	RMSE	MAE	R ²	RMSE	MAE
RF	0.939	18.407	6.508	0.818	25.667	16.828
XGBoost	0.998	2.891	1.575	0.992	4.008	3.846
DNN	0.987	6.726	5.531	0.986	7.638	5.608
LSTM	0.990	5.103	4.552	0.989	4.553	4.581

compared to other standard algorithms, XGBoost was chosen as the basis for our virtual environment.

4.3. CO₂ concentration control under manual control and DRL control

Fig. 7 compares hourly variations in CO₂ concentration levels with manual control of the ventilator (i.e., controlled by the occupants) and when the ventilator is controlled using the trained DRL agent. The patterns in daily CO₂ concentration levels are shown for a 5-week period of the field experiment (i.e., the period used in the agent testing phase). As seen in the figure, the CO₂ concentration was somewhat below the maximum permissible levels (i.e., 1000 ppm) most of the time regardless of the control methods. However, there were some particular instances when CO₂ concentration levels shot to considerably high amounts under manual control while the levels are maintained below acceptable levels when DRL is the main element of ventilator control. For example, On Day 1 of Week 1 (See Fig. 7a), the CO₂ concentration levels under manual control tended to increase gradually passed the acceptable threshold at around noon, reached the peak concentration amount (i.e., 2000 ppm) at around 15:30 before decreasing gradually to below 1000 ppm at around 17:30. For this same period, it is observed that the DRL control manages to keep the CO₂ concentration levels in acceptable ranges. This capability of DRL control to maintain acceptable CO₂ concentration levels at peak times and when on the contrary, manual control seems to fail is again observed on Day 6 of Week 3 (See Fig. 7n), Day 4 of Week 4 (Fig. 7r), Day 5 of Week 4 (Fig. 7s) and Day 4 of Week 5 (Fig. 7x).

The general observation implies that DRL maintains CO₂ levels within acceptable ranges all the time while manual control tends to fail in certain instances, particularly peak times. This is an obvious observation with a couple of possible explanations. First, CO₂ is a tasteless, colorless, and odorless gas, making it difficult for occupants to determine when the concentration levels have increased significantly and even more challenging to quantify the extent of deviation from acceptable ranges; this limits their timing for the manual control of the ventilator leading to poor IAQ. Secondly, CO₂ levels under manual control tended

Armon, 2022). Based on the illustrated better performance of XGBoost than the other algorithms on the employed dataset and the discussed reports from previous studies attesting to the robustness of XGBoost on tabulated data

to increase mainly in the early afternoons to late afternoons, which is most likely the busiest time in kindergartens. For example, we observed that lunch meals were served in the classroom, which significantly increased the number of occupants as the food servers were in the class together with the students and teachers – the increased temporal occupancy rate explains the relatively sharp increase in CO₂ concentration while the increased activity, perhaps, points to the lack of active measures to reduce CO₂ levels (i.e., turning on the ventilator). These reasons, particularly in pertinence to CO₂ control, showcase the necessity for advanced control techniques in indoor environments and are attested to in similar previous studies (Kumar et al., 2016). For instance, DRL manages to keep CO₂ concentration levels within acceptable ranges by following the employed policies and attempting to maintain a desirable state of the environment. The results are commensurate with multiple studies that demonstrate the use of agent-based intelligent control systems such as DRL for the optimal control of indoor environments (Dounis and Caraiscos, 2009).

4.4. Ventilator mode state under manual control and DRL control

Fig. 8 shows hourly mode settings under the manual control and DRL control – the ventilator could be operated in 3 modes; 1 (low mode), 2 (medium mode), 3 (high mode), and 0 when the ventilator was off. The state of the ventilator usage is shown for the entire period of the experiment.

From the figure, we can visualize the temporal interactions of the occupants with the ventilator and whether their actions had useful outcomes in terms of IAQ. For example, we observed that on Day 1 of Week 1 (Fig. 8a), and under manual control, the ventilator was mostly off for the first hours of the day and only switched on at around 14:30 – this observation potentially explains the increase in CO₂ concentration levels observed in Fig. 7 under manual control for the period between 11:30 and 17:30. On the contrary, from the same figure (Fig. 8a), we see that under DRL control, the ventilator is switched on at 12:30 and the mode of the ventilator quickly increases to 3 (the maximum level); by taking this action, the agent can maintain the CO₂ concentration levels below permissible levels for the entire afternoon (see Fig. 7a for comparisons). Similar patterns are seen across the other days during the experiment, further highlighting the usefulness of properly trained DRL agents compared to manual control in improving IAQ.

Furthermore, looking at the ventilator mode across the entire field experiment period, it is observed that in most cases, the ventilator was turned on in the early morning at the beginning of the day and left on until the end of the day (see, for example, Fig. 8k and 8l). Moreover, there were instances when the ventilator was turned on past 18:30 while there was little, or no occupancy expected in the room. This is an interesting observation with vast implications on building energy consumption and an illustration of how some aspects of occupant behavior could impact energy use in buildings. This complex interaction between occupants and their buildings or subsystems of their buildings and the subsequent influence on energy use is well broadly explored in

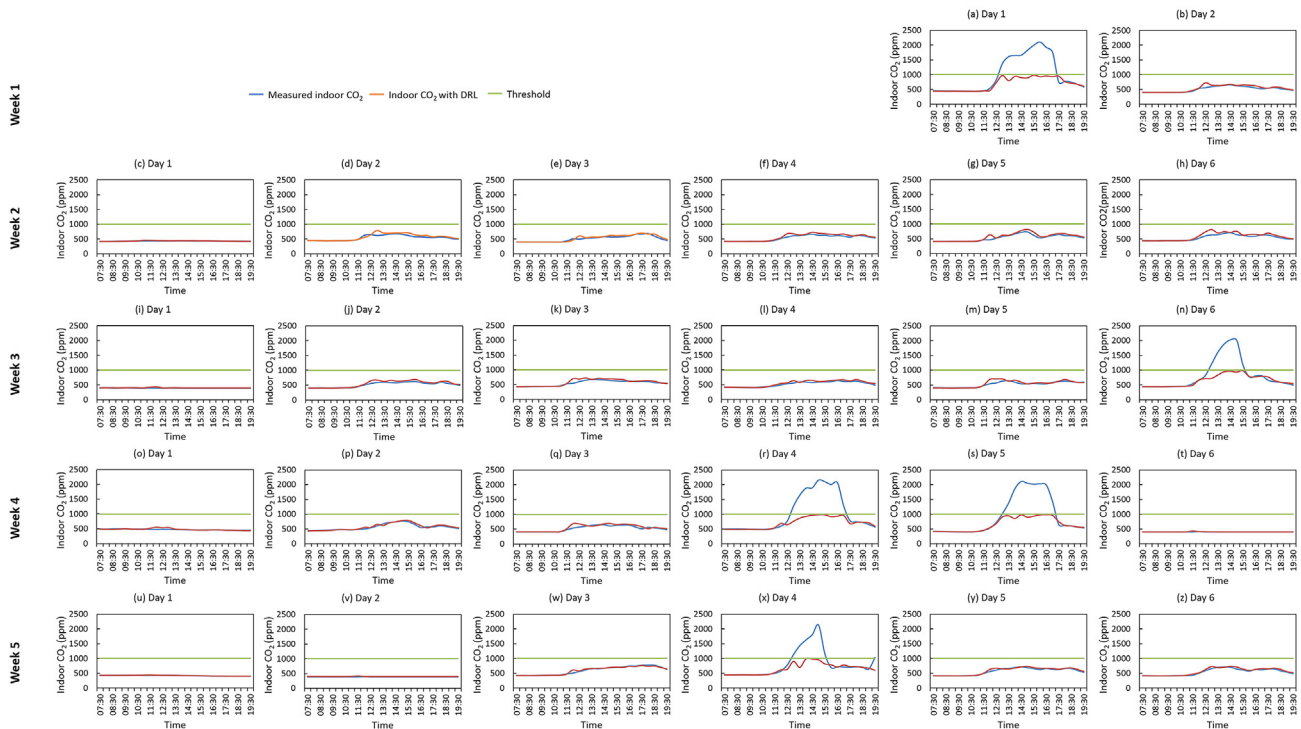


Fig. 7 Hourly indoor CO₂ variations.

the literature (Chen et al., 2021; Zhang et al., 2018). Similarly, the observations from this study provide empirical evidence on the role of occupant behavior, particularly in the proper manipulation of ventilators to maintain sufficient IAQ and increase energy conservation.

4.5. Ventilator energy usage under manual control and DRL control

Fig. 9 shows hourly energy consumption by the ventilator during the experimental period, both under manual control and under the proposed DRL agent. The figure shows that the ventilator energy consumption differs significantly under the two control mechanisms. For most of the days, the hourly energy consumption of the ventilator is maintained at 15 Wh from the early mornings to late evenings. On the contrary, under DRL control, energy consumption is mostly 0 Wh except for a few instances where it sharply increases to 15 Wh around midday. This observation was somewhat expected and is backed by our previous discussions on the state of the ventilator during the experiment period; the ventilator was mostly switched on under manual control and off under DRL control hence the observed differences in the consumed energy.

It is worth highlighting that there were a few instances when hourly energy consumption was higher under the DRL control scheme than manual control. Taking Day 1 of Week 1 as an example (see Fig. 10), the energy consumption is considerably higher under the DRL control scheme than manual control. Between 11:30 and 18:30. An explanation for such instances can be derived from the CO₂

concentration levels during the same time - it can be deduced that CO₂ concentration levels were abnormally high during the said period (i.e., 11:30–18:30). Consequently, the agent turns on the ventilator to the highest mode to ensure the CO₂ concentration levels are kept below 1000 ppm. Similar instances were observed through the experimental period; see, for example, Day 4 of Week 4 (Fig. 9r) and Day 4 of Week 5 (Fig. 9x) – the same explanation applies to these similar cases and is essentially because the DRL policy is designed such that it prioritizes IAQ over energy conservation in circumstances where it is impossible to achieve both sufficient IAQ and low energy use.

While, in a few instances discussed above, the DRL control scheme resulted in higher hourly energy consumption than the manual control scheme, the overall daily energy consumption was in general considerably lower under the DRL control scheme than the manual scheme (See Fig. 11). Similar observations were seen in weekly energy consumption reductions (see Table 6). For instance, it was estimated that the employment of the DRL control scheme would reduce the weekly energy consumption of the ventilator by –8%, 92%, 77%, 28%, and 58% in Week 1, Week 2, Week 3, Week 4 and Week 5 respectively. The negative reduction in Week 1, in which only two days were considered, is due to our earlier discussed scenarios where the DRL agent maximized the ventilator usage (Mode 3) to keep the rapidly increasing CO₂ levels from surpassing the dictated threshold. Furthermore, considering the monthly energy consumption of the ventilator under both control schemes, the estimated monthly energy consumption

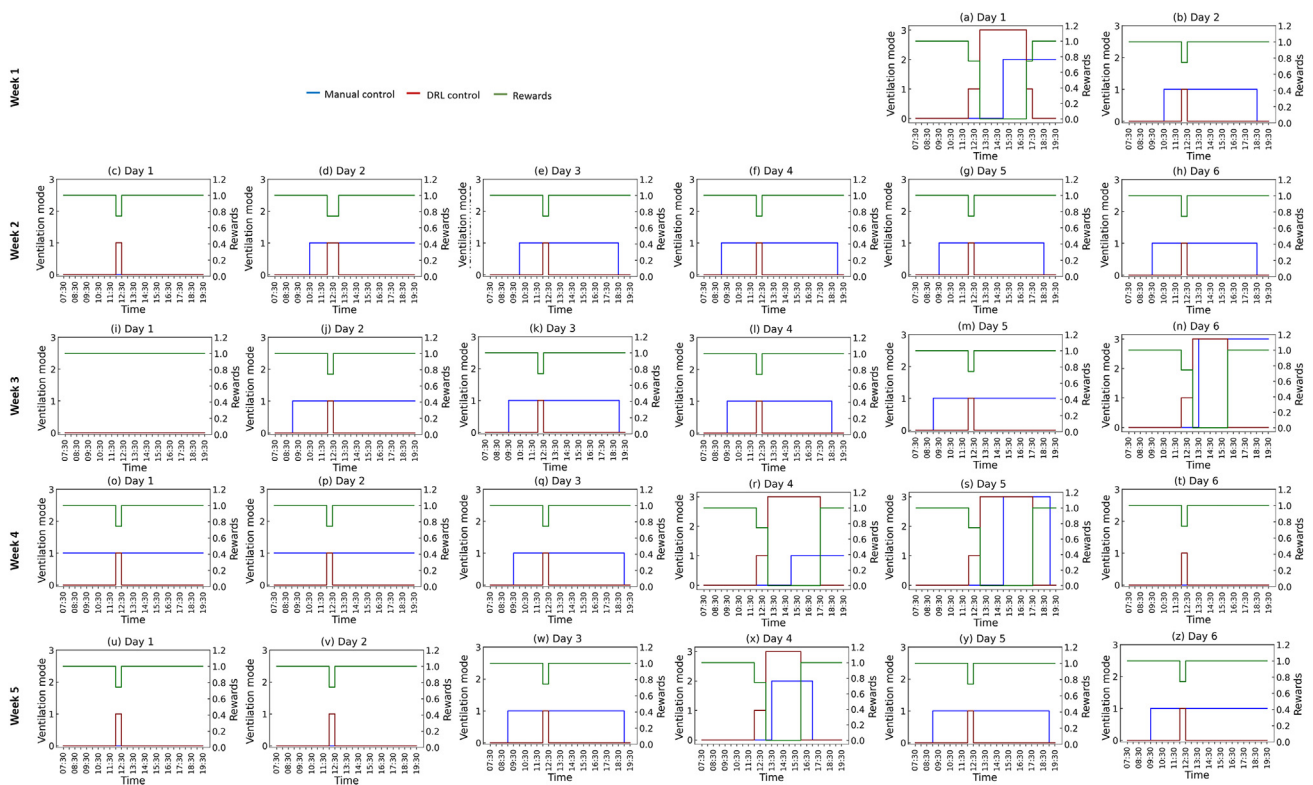


Fig. 8 Hourly ventilation mode setting.

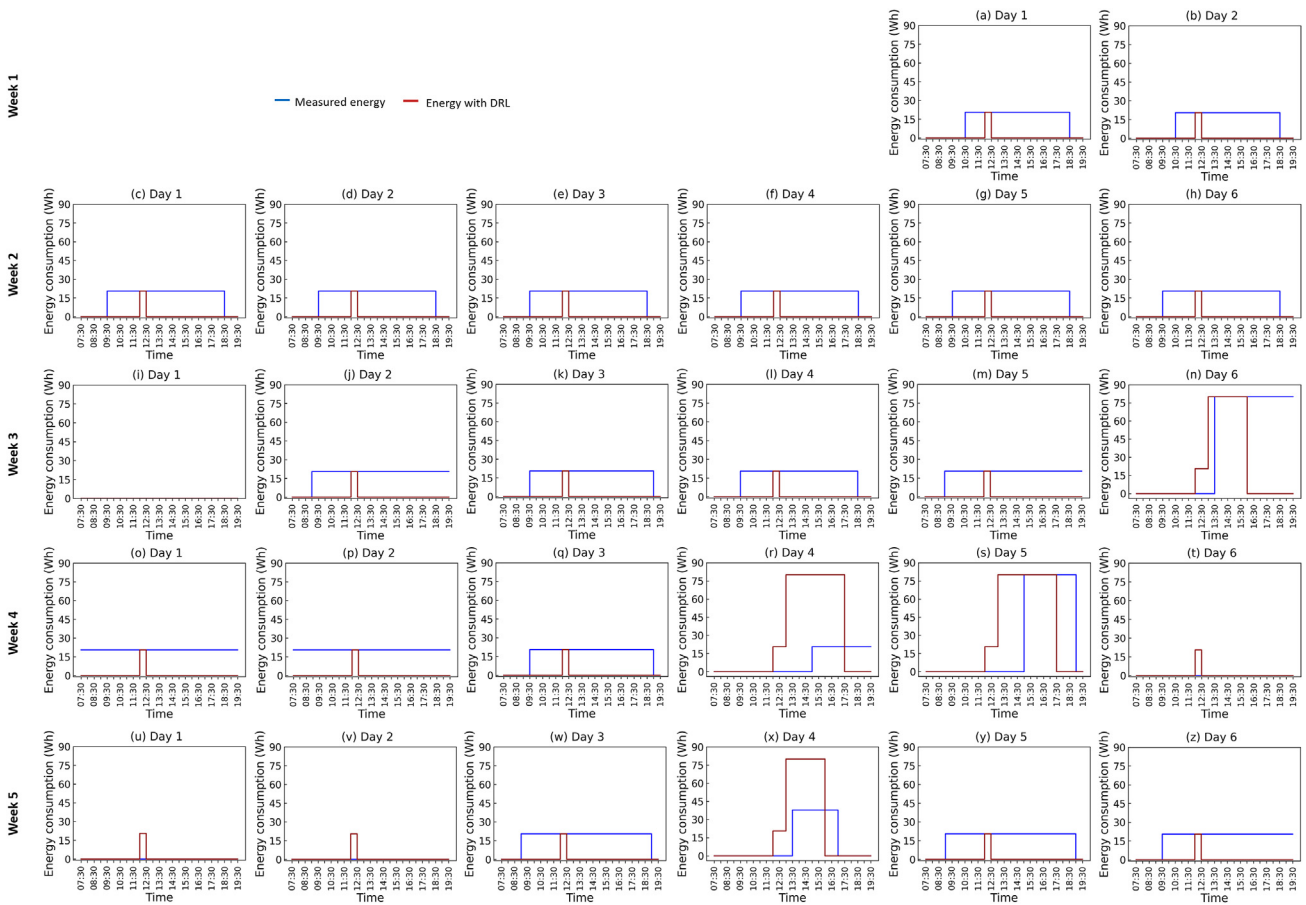


Fig. 9 Hourly energy consumption.

under the manual control scheme totaled up to 8.846 kWh and 3.696 kWh under the DRL control scheme. This finding indicates that employing DRL control mechanisms in this case study kindergarten would result in 58% reductions in

the monthly energy consumed by the ventilator. This finding is not particularly new and reiterates review reports from Dounis and Caraiscos (Dounis and Caraiscos, 2009) regarding the use of advanced control techniques for

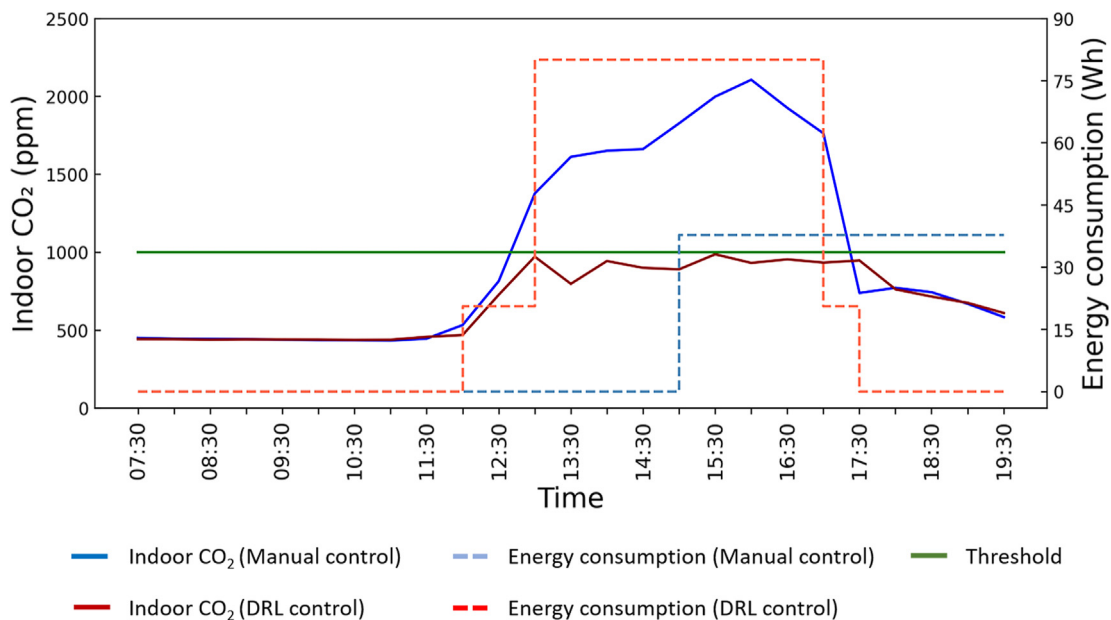


Fig. 10 DRL control energy consumption in response to changing indoor CO₂.

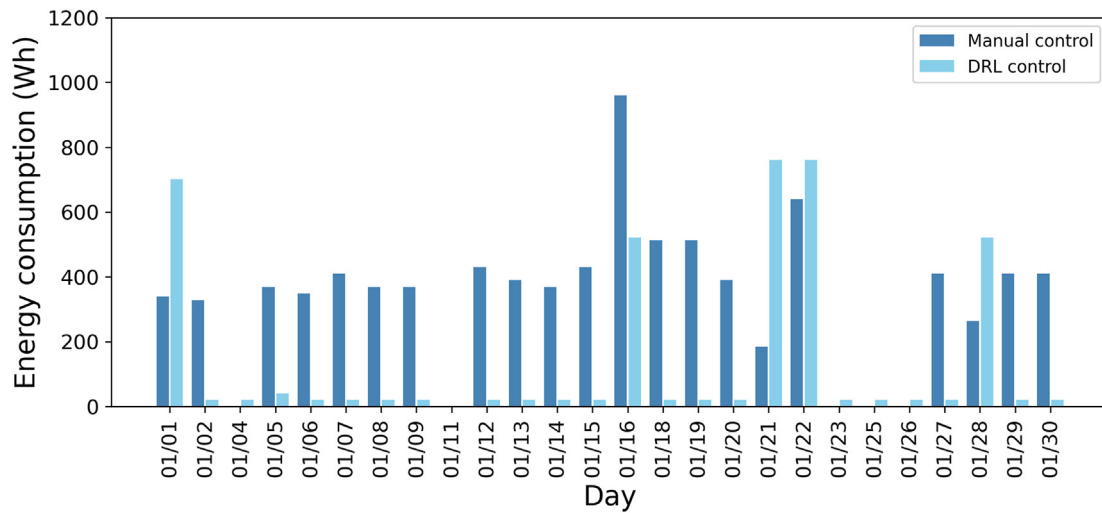


Fig. 11 Daily energy consumption.

energy conservation purposes. However, only a few studies have demonstrated the importance of advanced control techniques via real experiments. The present study adds to the small but increasing number of studies showcasing the suitability of autonomous smart agent controls to improve building performance.

4.6. DRL parameters and policy

There were various parameters involved in obtaining the optimal policy in DRL. These included the architecture of the neural networks used to estimate the value function, the discount factor to balance how the agent cares about immediate and future rewards, the learning rate, the size of the experience replay buffer, ϵ to control the exploration and exploitation, and how many episodes or rounds the agent was trained. Tuning for the best values for these parameters was done manually. The neural network consisted of two hidden layers with 64 units each. The discount factor was 0.9 and gave more importance to future rewards as indoor CO₂ could not be reduced with immediate actions (ventilation), rather than with a well-planned ventilation strategy. The learning rate was 0.01, the buffer size of 50 kilobytes, and the minimum ϵ for the ϵ -greedy policy was 0.1. The agent was trained with 300,000 timesteps.

Table 6 Weekly ventilation energy consumption under the manual control and DRL control schemes.

Period	Manual control (Wh)	DRL control (Wh)	Energy saved by DRL (%)
Week 1	667.75	722	−8
Week 2	1865.5	143.5	92
Week 3	2579.5	603	77
Week 4	2239	1604	28
Week 5	1494.25	623.5	58
Total (month)	8846	3696	58

5. Limitations and future research

The data used to train and test the virtual sensor and DRL-based agent was collected from one kindergarten. This did not allow us to evaluate the performance of developed control system (i.e., DRL agent coupled with an XGBoost-based virtual sensor) on other types of buildings. Future research may consider using diverse data that span different types of buildings. It is worth noting that the observed energy savings are also particularly influenced by the occupant behavior. It is possible that the proposed DRL control scheme might not achieve similar performance if installed in completely new environments where occupant behavior significantly differs from that in the studied kindergarten. Moreover, another scope for future research is to evaluate the applicability of transfer learning (Pinto et al., 2022) for virtual sensors and DRL.

One other major limitation of the current study is that the performance of the proposed DRL control scheme is not comparatively assessed against other common control methods such as the MPC, and PID. A comparison of these conventional techniques to the proposed DRL control scheme would further highlight the usefulness of advanced ML control techniques in providing conducive indoor environments with minimum energy consumption and is worth exploring in future research.

6. Conclusion

This study proposes a demand-controlled ventilation for indoor CO₂ control using a virtual sensor and a double Q-learning control agent. In particular, the developed approach performs intelligent preemptive ventilation to maintain indoor CO₂ below 1000 ppm with minimal ventilation energy consumption, compared to the existing manual control method. In a nutshell, the study achieves two significant outcomes. First, the proposed DRL-based control scheme showed significant energy savings compared to the manual control that existed in the studied kindergarten; a 58% energy reduction was observed. The proposed approach

thus provides a unique solution to tackle the shortcomings of the existing control methods such as control lags in response to changing indoor CO₂ and unnecessary energy consumption. Second, the developed virtual sensor could accurately model the variations in CO₂ concentrations; the predictive performance of the virtual sensor showed a predictive performance with R² of 0.99. Furthermore, the developed virtual sensor was also successfully employed as the feedback platform during the training of the DRL-based control scheme. This showcases a new approach to training DRL agents via trained virtual sensors and avoids the need for simulated environments which are often computationally expensive and whose performance are largely dependent on factors beyond the mechanisms of the simulation tools (e.g., the simulation process and skills of the person performing the simulations).

Declaration of competing interests

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.

Acknowledgement

This work was supported by the National Research Foundation of Korea (NRF) grant funded by the Korea government (MSIT) (No. 2020R1A2C1099611).

References

- Allen, J.G., MacNaughton, P., Cedeno-Laurent, J.G., Cao, X., Flanagan, S., Vallarino, J., Rueda, F., Donnelly-McLay, D., Spengler, J.D., 2019. Airplane pilot flight performance on 21 maneuvers in a flight simulator under varying carbon dioxide concentrations. *J. Expo. Sci. Environ. Epidemiol.* 29, 457–468.
- Allen, J.G., MacNaughton, P., Satish, U., Santanam, S., Vallarino, J., Spengler, J.D., 2016. Associations of cognitive function scores with carbon dioxide, ventilation, and volatile organic compound exposures in office workers: a controlled exposure study of green and conventional office environments. *Environ Health Perspect* 124, 805–812. <https://doi.org/10.1289/ehp.1510037>.
- Arulkumaran, K., Deisenroth, M.P., Brundage, M., Bharath, A.A., 2017. Deep reinforcement learning: a brief survey. *IEEE signal process. Mag.* 34, 26–38. <https://doi.org/10.1109/MSP.2017.2743240>.
- Aryana, K., Gaskins, J.T., Nag, J., Stewart, D.A., Bai, Z., Mukhopadhyay, S., Read, J.C., Olson, D.H., Høglund, E.R., Howe, J.M., Giri, A., Grobis, M.K., Hopkins, P.E., 2021. Interface controlled thermal resistances of ultra-thin chalcogenide-based phase change memory devices. *Nat. Commun.* 12, 774. <https://doi.org/10.1038/s41467-020-20661-8>.
- Azuma, K., Kagi, N., Yanagi, U., Osawa, H., 2018. Effects of low-level inhalation exposure to carbon dioxide in indoor environments: a short review on human health and psychomotor performance. *Environ. Int.* 121, 51–56. <https://doi.org/10.1016/j.envint.2018.08.059>.
- Bellman, R., 1966. Dynamic programming. *Science* 153, 34–37. <https://doi.org/10.1126/science.153.3731.34>.
- Bellman, R., 1952. On the theory of dynamic programming. *Proc. Natl. Acad. Sci. USA* 38, 716–719. <https://doi.org/10.1073/pnas.38.8.716>.
- Biau, G., Scornet, E., 2016. A random forest guided tour. *Test* 25, 197–227. <https://doi.org/10.1007/s11749-016-0481-7>.
- Breiman, L., 2001. Random forests. *Mach. Learn.* 45, 5–32.
- Chen, S., Zhang, G., Xia, X., Chen, Y., Setunge, S., Shi, L., 2021. The impacts of occupant behavior on building energy consumption: a review. *Sustain. Energy Technol. Assessments* 45, 101212. <https://doi.org/10.1016/j.seta.2021.101212>.
- Chen, T., Guestrin, C., 2016. XGBoost: a scalable tree boosting system. In: *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. Presented at the KDD '16: the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. ACM, San Francisco California USA, pp. 785–794. <https://doi.org/10.1145/2939672.2939785>.
- Dounis, A.I., Caraiscos, C., 2009. Advanced control systems engineering for energy and comfort management in a building environment—a review. *Renew. Sustain. Energy Rev.* 13, 1246–1261. <https://doi.org/10.1016/j.rser.2008.09.015>.
- Friedman, J.H., 2002. Stochastic gradient boosting. *Comput. Stat. Data Anal.* 38, 367–378. [https://doi.org/10.1016/S0167-9473\(01\)00065-2](https://doi.org/10.1016/S0167-9473(01)00065-2).
- Glorennec, P.Y., 2000. Reinforcement Learning: an Overview 19.
- Goodfellow, I., Bengio, Y., Courville, A., 2016. *Deep Learning*. MIT press.
- Grzes, Marek, Daniel, Kudenko, 2008. Plan-based reward shaping for reinforcement learning. In *2008 4th International IEEE Conference Intelligent Systems* 2, 10–22. <https://doi.org/10.1109/IS.2008.4670492>.
- Hochreiter, S., Schmidhuber, J., 1997. Long short-term memory. *Neural Comput.* 9, 1735–1780.
- Hwang, S.H., Roh, J., Park, W.M., 2018. Evaluation of PM10, CO₂, airborne bacteria, TVOCs, and formaldehyde in facilities for susceptible populations in South Korea. *Environ. Pollut.* 242, 700–708. <https://doi.org/10.1016/j.envpol.2018.07.013>.
- ISO 16494: 2014, 2014. Heat Recovery Ventilators and Energy Recovery Ventilators — Method of Test for Performance. International Organization for Standardization.
- Itri, J.N., Bruno, M.A., Lalwani, N., Munden, R.F., Tappouni, R., 2019. The incentive dilemma: intrinsic motivation and workplace performance. *J. Am. Coll. Radiol.* 16, 39–44. <https://doi.org/10.1016/j.jacr.2018.09.008>.
- Jacobson, T.A., Kler, J.S., Hernke, M.T., Braun, R.K., Meyer, K.C., Funk, W.E., 2019. Direct human health risks of increased atmospheric carbon dioxide. *Nat. Sustain.* 2, 691–701.
- Kiumarsi, B., Vamvoudakis, K.G., Modares, H., Lewis, F.L., 2018. Optimal and autonomous control using reinforcement learning: a survey. *IEEE Transact. Neural Networks Learn. Syst.* 29, 2042–2062. <https://doi.org/10.1109/TNNLS.2017.2773458>.
- Klepeis, N.E., Nelson, W.C., Ott, W.R., Robinson, J.P., Tsang, A.M., Switzer, P., Behar, J.V., Hern, S.C., Engelmann, W.H., 2001. The National Human Activity Pattern Survey (NHAPS): a resource for assessing exposure to environmental pollutants. *J. Expo. Sci. Environ. Epidemiol.* 11, 231–252. <https://doi.org/10.1038/sj.jea.7500165>.
- Kumar, P., Martani, C., Morawska, L., Norford, L., Choudhary, R., Bell, M., Leach, M., 2016. Indoor air quality and energy management through real-time sensing in commercial buildings. *Energy Build.* 111, 145–153. <https://doi.org/10.1016/j.enbuild.2015.11.037>.
- Laud, Adam Daniel, 2004. *Theory and application of reward shaping in reinforcement learning*. University of Illinois at Urbana-Champaign.
- LeCun, Y., Bengio, Y., Hinton, G., 2015. Deep learning. *Nature* 521, 436–444. <https://doi.org/10.1038/nature14539>.
- Leung, D.Y.C., 2015. Outdoor-indoor air pollution in urban environment: challenges and opportunity. *Front. Environ. Sci.* 2. <https://doi.org/10.3389/fenvs.2014.00069>.
- Lewis, D., 2021. The challenges of making indoors safe. *Nature* 22–25.

- Mnih, V., Kavukcuoglu, K., Silver, D., Rusu, A.A., Veness, J., Bellemare, M.G., Graves, A., Riedmiller, M., Fidjeland, A.K., Ostrovski, G., Petersen, S., Beattie, C., Sadik, A., Antonoglou, I., King, H., Kumaran, D., Wierstra, D., Legg, S., Hassabis, D., 2015. Human-level control through deep reinforcement learning. *Nature* 518, 529–533. <https://doi.org/10.1038/nature14236>.
- Montavon, G., Samek, W., Müller, K.-R., 2018. Methods for interpreting and understanding deep neural networks. *Digit. Signal Process.* 73, 1–15. <https://doi.org/10.1016/j.dsp.2017.10.011>.
- Morawska, L., Allen, J., Bahnfleth, W., Bluyssen, P.M., Boerstra, A., Buonanno, G., Cao, J., Dancer, S.J., Floto, A., Franchimon, F., Greenhalgh, T., Haworth, C., Hogeling, J., Ison, C., Jimenez, J.L., Kurnitski, J., Li, Y., Loomans, M., Marks, G., Marr, L.C., Mazzearella, L., Melikov, A.K., Miller, S., Milton, D.K., Nazaroff, W., Nielsen, P.V., Noakes, C., Peccia, J., Prather, K., Querol, X., Sekhar, C., Seppänen, O., Tanabe, S., Tang, J.W., Tellier, R., Tham, K.W., Wargocki, P., Wierzbicka, A., Yao, M., 2021. A paradigm shift to combat indoor respiratory infection. *Science* 372, 689–691. <https://doi.org/10.1126/science.abg2025>.
- Nathansan, T., 1993. *Indoor Air Quality in Office Buildings: A Technical Guide*.
- National Research Council, 1981. *Indoor Pollutants*.
- Nwankpa, C., Ijomah, W., Gachagan, A., Marshall, S., 2018. *Activation Functions: Comparison of Trends in Practice and Research for Deep Learning*. arXiv, 1811.03378 [cs].
- Pinto, G., Wang, Z., Roy, A., Hong, T., Capozzoli, A., 2022. Transfer learning for smart buildings: a critical review of algorithms, applications, and future perspectives. *Adv. Appl. Energy* 5, 100084. <https://doi.org/10.1016/j.adapen.2022.100084>.
- Redlich, C.A., Sparer, J., Cullen, M.R., 1997. Sick-building syndrome. *Lancet* 349, 1013–1016.
- Ruder, S., 2017. *An Overview of Gradient Descent Optimization Algorithms*. arXiv, 1609.04747 [cs].
- Ryzhov, A., Ouerdane, H., Gryazina, E., Bischi, A., Turitsyn, K., 2019. Model predictive control of indoor microclimate: existing building stock comfort improvement. *Energy Convers. Manag.* 179, 219–228. <https://doi.org/10.1016/j.enconman.2018.10.046>.
- Salsbury, T.I., 2005. A survey of control technologies in the building automation industry. *IFAC Proc.* Vol. 38, 90–100. <https://doi.org/10.3182/20050703-6-CZ-1902.01397>.
- Satish, U., Mendell, M.J., Shekhar, K., Hotchi, T., Sullivan, D., Streufert, S., Fisk, W.J., 2012. Is CO₂ an indoor pollutant? Direct effects of low-to-moderate CO₂ concentrations on human decision-making performance. *Environ Health Perspect* 120, 1671–1677. <https://doi.org/10.1289/ehp.1104789>.
- Schapire, R.E., 1990. The strength of weak learnability. *Mach Learn* 5, 197–227.
- Shwartz-Ziv, R., Armon, A., 2022. Tabular data: deep learning is not all you need. *Inf. Fusion* 81, 84–90. <https://doi.org/10.1016/j.inffus.2021.11.011>.
- van Hasselt, H., Guez, A., Silver, D., 2016. Deep reinforcement learning with double q-learning, in: *Proceedings of the AAAI Conference on Artificial Intelligence*.
- Sutton, R.S., Barto, A.G., 2018. *Reinforcement Learning: an Introduction*. MIT press.
- van Otterlo, M., Wiering, M., 2012. Reinforcement learning and markov decision Processes. In: *Wiering, M., van Otterlo, M. (Eds.), Reinforcement Learning, Adaptation, Learning, and Optimization*. Springer Berlin Heidelberg, Berlin, Heidelberg, pp. 3–42. https://doi.org/10.1007/978-3-642-27645-3_1.
- Yang, T., Zhao, L., Li, W., Wu, J., Zomaya, A.Y., 2021. Towards healthy and cost-effective indoor environment management in smart homes: a deep reinforcement learning approach. *Appl. Energy* 300, 117335. <https://doi.org/10.1016/j.apenergy.2021.117335>.
- Yang, T., Zhao, L., Li, W., Zomaya, A.Y., 2020. Reinforcement learning in sustainable energy and electric systems: a survey. *Annu. Rev. Control* 49, 145–163. <https://doi.org/10.1016/j.arcontrol.2020.03.001>.
- Zamani Joharestani, M., Cao, C., Ni, X., Bashir, B., Talebiesfandarani, S., 2019. PM_{2.5} prediction based on random forest, XGBoost, and deep learning using multisource remote sensing data. *Atmosphere* 10, 373. <https://doi.org/10.3390/atmos10070373>.
- Zhang, Y., Bai, X., Mills, F.P., Pezzey, J.C.V., 2018. Rethinking the role of occupant behavior in building energy performance: a review. *Energy Build.* 172, 279–294. <https://doi.org/10.1016/j.enbuild.2018.05.017>.