Transfer learning applied to DRL-Based heat pump control to leverage microgrid energy efficiency

Smart Energy 3 (2021) 100044

Contents lists available at ScienceDirect

Smart Energy

journal homepage: www.journals.elsevier.com/smart-energy

Transfer learning applied to DRL-Based heat pump control to leverage microgrid energy efficiency

Paulo Lissa [a, b, *], Michael Schukat [a], Marcus Keane [a, b, c], Enda Barrett [a]

[a] College of Science and Engineering, National University of Ireland, Galway, Ireland
[b] Informatics Research Unit for Sustainable Engineering (IRUSE) Galway, Ireland
[c] Ryan Institute, National University of Ireland Galway, Ireland

ARTICLE INFO

Article history:
Received 8 April 2021
Received in revised form
11 August 2021
Accepted 16 August 2021
Available online 17 August 2021

Keywords:
Heat pump
Deep reinforcement learning
Transfer learning
Autonomous control
Demand response
Microgrid

ABSTRACT

Domestic hot water accounts for approximately 15% of the total residential energy consumption in Europe, and most of this usage happens during specific periods of the day, resulting in undesirable peak loads. The increase in energy production from renewables adds additional complexity in energy balancing. Machine learning techniques for heat pump control have demonstrated efficacy in this regard. However, reducing the amount of time and data required to train effective policies can be challenging. This paper investigates the application of transfer learning applied to a deep reinforcement learning-based heat pump control to leverage energy efficiency in a microgrid. First, we propose an algorithm for domestic hot water temperature control and PV self-consumption optimisation. Secondly, we perform transfer learning to speed-up the convergence process. The experiments were deployed in a simulated environment using real data from two residential demand response projects. The results show that the proposed algorithm achieved up to 10% of savings after transfer learning was applied, also contributing to load-shifting. Moreover, the learning time to train near-optimal control policies was reduced by more than a factor of 5.

1. Introduction

Domestic hot water (DHW) accounts for approximately 15% of the total residential energy consumption in Europe according to a report from Eurostat [1]. It comprises water heating for uses other than space heating, such as showers or taps, and most of its usage happens during specific periods of the day, such as early in the morning and evenings, resulting in undesirable peak loads. Meanwhile, studies from the International Energy Agency [2] show that energy production from renewables are increasing all the time, representing almost 20% of the total generation in the EU, and it is expected to expand to 32% by 2030 as stated by the European Commission [3]. Energy from renewable sources, such as PV and wind, rely on weather conditions, and coordinating demand and production is one of the challenges arising from this change in the energy matrix. This new scenario is part of the smart energy system concept, which includes energy efficiency and better management of resources in an integrated manner. It is relevant in terms of achieving a strategy of 100% renewable envisaged for the future, as explored by Mathiesen et al. [63].

Demand Response (DR) solutions, which have been commonly applied in the industrial and commercial sectors, have started to involve residential homes, as can be seen in recent projects around the world, such as the EU Horizon 2020 projects RESPOND [4] and REACT [5]. DR solutions empower consumers to play a significant role in grid operation, by reducing or shifting loads during peak load, in response to any incentives or time-based rates [6]. The DR actions can be performed manually or automated, through the use of new technologies like the Internet-of-Things (IoT). Electric heat pumps are very suitable for DR because of their high electricity consumption and flexibility. In a common residential application, heat pumps can be used for heating, ventilating, and air conditioning (HVAC), and DHW applications. The first aims to provide indoor thermal comfort and air quality, while the second improves the user's life in specific aspects, such as shower and hot water taps. From the control strategy perspective, electric heat pumps can be managed and turned on at different times of the day without affecting the needs of the end-user. They are an alternative option to oil or gas-based systems, thus contributing to the reduction of carbon emissions. Choosing the

* Corresponding author. NUI Galway, Galway, IT401, Ireland.
E-mail address: paulo.lissa@nuigalway.ie (P. Lissa).

https://doi.org/10.1016/j.segy.2021.100044

right time to use electric heat pumps is crucial in terms of cost savings, which should prioritize periods where electricity price is low or on-site energy production is high. Moreover, the water tank capacity in residences is limited and there are heat losses over the day, hence water heating has to be performed more than one time to keep the temperature at the desired setpoint, increasing the need for a more complex control algorithm.

The utilisation of heat pump systems combined with renewables has been object of study in the past years. Hedegaard et al. [61] stated that individual heat pumps and heat storage could play an important role in the integration of wind power. Their research shows that heat pumps can increase wind power utilisation, thus reducing the excess electricity production by up to 8% and 19%, for systems with and without heat storage, respectively. Protopapadaki and Saelens [62] also studied the interaction of heat pumps and renewables, but with a focus on solar generation. Results from their simulations confirmed that thermal storage offers the ability to shift electricity demand, leveraging the consumption of PV energy generation.

Rule and schedule-based approaches are the most common heat pump control methods, which allow users to set a specific temperature setpoint for the water tank or pre-program the time to activate the system, respectively. Nonetheless, the use of adaptative control solutions is gaining popularity due to their capacity of improving energy and cost savings while also allowing for load shifting capabilities, for instance considering optimisation by local renewable self-consumption and dynamic tariffs. Machine learning (ML) adaptative methods have demonstrated efficacy in this domain, as proven by authors in Refs. [7–12], but there are some issues such as the lack of data available or the amount of time needed to train optimal policies.

Some DHW controllers are based on reinforcement learning (RL), which is a sub-field of ML. Here the controller agent learns by interacting with the environment, thus improving and learning near-optimal policies according to the previous experiences after a large number of trials, where no previous knowledge is needed. For a particular state, the agent receives a reward after performing an action, which indicates whether the chosen action was good or not. For instance, in DHW control the agent would be penalized if the tank temperature drops to less than a specified setpoint, or it would be rewarded if water was heated when the energy price at that specific time is low. The RL agent target is to achieve a cost minimization while ensuring the specified parameters of temperature over time.

Achieving optimal control policies may require a huge number of trials, where the learning agent will be exposed to a number of conditions and verify the consequences after performing different actions. Transfer learning (TF) techniques in contrast have been showing efficiency in reducing learning time, by reusing knowledge from another agent. For example, if an agent already has experience in a particular state, the hypothesis is that this can be shared with a second agent in a similar environment, hence avoiding the need of visiting the same state again. This was proved by authors in Ref. [13], where a significant speed-up in learning times was achieved in an HVAC control system. Apart from the building efficiency domain, TF has also demonstrated efficacy in other domains, such as computer vision [14–16], games [17,18] and 3D environments [19,20].

Considering that some houses from a specific community are equipped with similar heat pumps and PV systems, hence sharing common environmental and heating dynamics, this paper investigates the use of transfer learning to a deep reinforcement learning (DRL)-based heat pump control to leverage energy efficiency in a microgrid, focusing on DHW. The thermal constraints of the hot water tank will not differ much between the houses, as they are built using the same material and insulation levels. Also, external temperatures will not significantly vary among the houses, and PV production will be also similar if the houses are geographically located close to one another, although some small variations can happen depending on the position of the installation. The main contributions of this paper include:

- The development of a model-free DRL algorithm for a DHW tank temperature control for a number of houses in a microgrid, aiming to reduce energy consumption from the grid by optimizing the usage of PV energy production and enhancing financial savings through a Time-of-Use tariff scheme.
- The utilisation of transfer learning techniques for a DRL-based DHW control, allowing for much faster learning rates, and comparison of its convergence speed with models without transfer learning.

The rest of this paper is organized as follows: *Related work*, which shows the related works regarding energy management systems and relevant transfer learning approaches. *Problem formulation* provides information about the environment structure and the principles of ML applied in this research. *Experimental setup* explains the variables and agent's definition for the test cases, including also the algorithms. The *Results* section presents all the relevant outputs of our experiments, showing the performance of the algorithm and comparing the methods. Finally, the *Conclusions and future works* section recaps the main points of the paper, introducing ideas for future work.

2. Related work

This section provides an overview of residential smart control for demand response, showing the current control methods and also detailing recent works in this subject, with a focus on water heating and reinforcement learning. Moreover, some examples of transfer learning applied to HVAC control and other domains are presented, aiming to show the advantages that this technique can bring when applied to the DHW domain.

2.1. Residential smart control for demand response

Demand response solutions can be applied in different types of residential assets, such as HVAC, DHW, smart appliances (e.g., dishwasher and washing machines), and electric vehicles. There is a variety of control methods for comfort and energy management, which can sometimes use renewable energy production, batteries, or dynamic prices as part of the optimisation process. The focus of this research is on RL model-free methods, where the agent controller does not need complete information about the environment, as it attempts to directly approximate a control policy through environmental interactions. Model-free methods are more adaptive to variations in the environment, likely to perform better than model-based when approximating the missing model is a complex task. Other methods out of the reinforcement learning domain include fuzzy control [21,22], and model predictive control [23–25], that can be very accurate, but sometimes costly and time-consuming as detailed models of the building are required.

Vázquez-Canteli and Nagy [26], in their review about reinforcement learning for demand response, identified more than one hundred studies across different applications in which only approximately 15% were about water heating (DHW, thermal storage, heat pump, and district heating). Their research shows that Q-Learning is the most popular learning method in the energy and buildings context, which can also be seen in the reviews from Mason and Grijalva [27] and Han et al. [28]. Their studies also found that the utilisation of Deep Q-learning has been increasing since 2015, mainly because its advantage of handling large state-action space with more efficiency, hence reducing learning time. Considering that in this

research we will be handling multiple devices (heat pumps and PVs) from a number of houses and also thinking about future expansion of this study, we chose a DRL approach combined with transfer learning techniques to enhance learning speeds.

The efficacy of reinforcement learning applied to DHW and residential PV production self-consumption can be seen in research carried out by De Somer et al. [9] and later on by Soares et al. [29], where the fitted Q-iteration control algorithm achieved on average 14% and more than 20% increase in self-consumption, respectively, compared to a thermostat control method. Both works used a model-based reinforcement learning approach, where the transition function of the control strategy was learned from historical data. In complex environments, approximating the missing model by learning the transition function from the historical data will not be adaptive to changes beyond the scope. For instance, weather conditions may vary from year to year, so the control strategy learnt from a data set from the previous year may not represent the new reality. Al-jabery et al. [10] applied Q-Learning for a number of different domestic electric water heaters models, targeting to reduce energy bills by using the Time-of-Use (ToU) tariff scheme as part of the optimisation process. Similarly, Ruelens et al. [8] reached a 15% reduction in the cost of energy reduction using fitted Q-Learning and dynamic prices. In another work, Ruelens et al. [42] applied a model-free batch RL algorithm using a Monte Carlo and fitted Q-iteration approach in different experiments involving an electric water heater and heat-pump thermostat. The results showed that RL can be an option to model-based controllers, with a decrease in the total electricity cost of up to 19%. Kazmi et al. [7] applied DRL to a set of 32 houses, reducing energy consumption used for water heating by approximately 20%. Kazmi et al. [43] applied a multi-agent RL approach to accelerate learning of a large-scale pilot consisting of 50 thermostatically controlled loads from different houses, where the proposed solution can potentially achieve savings up to 300 kWh annually per household (30% of reduction). Finally, Chen et al. [44] proposed an RL control method for DHW and space heating, using a Differentiable Model Predictive Control (MPC) policy, encoding domain knowledge on planning and system dynamics. The agent was pre-trained on historical data and once the controller behavior was learnt, a policy gradient algorithm was applied, reaching 6.6% of energy savings and 16.7% of cooling demand reduction.

Recent works, also about deep reinforcement learning applied to heat pumps, but not with a focus on DHW, also showed the efficiency of such algorithms in this domain. For instance, research articles [45–56], which are mostly from the HVAC and space heating domain, use DRL approaches with different targets. For instance, Vázquez-Canteli et al. [50] considered a multi-agent approach for a number of connected buildings, which reduced daily peak load by 15% by coordinating different loads. Kurte et al. [48] and Christensen et al. [54] used a price-responsive model as part of the optimisation, where the agent had to prioritize control actions when the cost of electricity was low.

Regarding convergence time, Yang et al. [30] proposed an RL approach with tabular Q-learning and batch Q-learning, applied to a heat pump control combined with a photovoltaic-thermal solution, outperforming the rule-based model by 10% after the third year of simulation. Patyn et al. [31] compared different neural architectures models for reinforcement learning applied to heat pump control, which outperformed standard thermostat controller and shift loads after 20–25 days. This is one of the motivations of our research: we aim to apply TF to reduce learning time, thus achieving near-optimal policies faster and bringing additional savings to users.

2.2. Transfer learning for reinforcement learning

Transfer learning refers to the use of experience learnt by performing one task to improve another related, but different, task [32]. The objective is to increase convergence speed, as instead of learning from the beginning without prior knowledge, the agent integrates value function estimates from other agents to perform the tasks, achieving better results faster as a result of the knowledge sharing. Taylor and Stone [32] describes in their review five different types of transfer learning approaches, where our research is classified as part of the allowed tasks differences group, which is applied to methods that have the same state variables and actions. For instance, in our experiments, the agent's representation of the world, which includes PV energy production metering, temperature sensing, and heat pump actuation all remain the same, while states values and/or actions can change. Furthermore, according to da Silva and Costa [33], our TF method is categorized as Inter-Agent transfer, in which knowledge is shared with another agent with a very similar characteristic, and this agent is able to merge the transferred knowledge with its own experience.

Although TF studies have been carried out and have demonstrated success in other domains, to the best of our knowledge this is the first time it has been applied to a model-free DRL-based heat pump control in a microgrid, leveraging PV self-consumption and reducing energy through ToU tariffs. In recent work from a close domain, Lissa et al. [13] proposed an RL algorithm which can transfer HVAC control polices between agents from different geographic locations. The results showed that the learning time to train near-optimal control policies was reduced by more than a factor of 6, compared to models without TF. Also in the HVAC field, Xu et al. [34] applied TF to a DRL-based HVAC control, transferring knowledge from different zones, equipment and locations, thus reducing the training time, energy cost, and temperature violations. Zhang et al. [55] proposed an RL approach with TF to reduce the training cost of an RL control policy for smart homes with different appliances and user preferences, showing that TL can effectively reduce the training time of a new policy if the new home is similar to the benchmark home. Finally, authors in Refs. [35,36] also explored TF, but with focus on building predictive models, predicting energy consumption, temperature and humidity.

3. Problem formulation

3.1. Test environment

The experiments were deployed in a simulated environment using real data from two residential demand response projects, RESPOND [4] and REACT [5], in which one of their case studies was realized in the Aran Islands, Galway, Ireland. Energy supply in small islands can be challenging due to its geographic characteristics, hence the community relies on energy imported through submarine cables or fossil fuels. Both can be very costly, the first is expensive to deploy or upgrade if needed, and the second is not easy for logistical reasons and it is also not environmentally friendly, due to its carbon emissions. Some of the houses in the island have individual PV energy systems, which helps to reduce the grid load. However, a smart controller is needed to coordinate demand and production, which is the aim of RESPOND [4] and also subject of study of REACT [5] in a number of islands across Europe.

Our study considers as reference five houses that have been recently upgraded in the Aran Islands, all equipped with an 8.5 kW Mitsubishi Electric Ecodan heat pump along with a PV panel array consisting of 8 panels, with a total nominal power of 2 kWp. For each house, the heat pump is connected to a 200-liter cylinder (tank) which is used to store hot water. The main characteristics of the cylinder can be seen in Table 1, extracted from the manufacturer's document [37]. These features rule the dynamics of the simulations in terms of water heating and temperature losses.

The standard DHW mode operates automatically based on the upper and lower limits set by the user. For instance, if the user sets the maximum temperature to 50 °C and a maximum temperature drop of 10 °C the hot water is automatically heated once the tank temperature drops to 40 °C, until it achieves 50 °C again, where this cycle restarts [38]. This rule-based method will be used as a comparison during the experiments.

Our final test environment is a python application deployed from a detailed building model. Sensors were installed in the houses to measure indoor temperature °C, total electricity consumption (kWh), total PV production (kWh) and heat pump electricity consumption (kWh). The construction characteristics obtained from a site survey, such as dimensions and *u*-values, along with data collected from each of the sensors, were used as a reference for developing a detailed and calibrated white-box model, using the Integrated Environmental Solution Virtual Environment (IESVE) software. Further details on each of the *u*-values and measurements can be found in Ref. [53]. The tank dynamics of Table 1 are also included as part of the environment. Finally, we added historical data from the PV sensors of the residences with a 15-min resolution, acquired in May 2020. In summary, the test environment reads the current state every 15 min and estimates the next DHW tank temperature values, considering the action taken by the RL algorithm, developed with TensorFlow, and the dynamics established previously.

3.2. Markov decision process and reinforcement learning

The main goal of a reinforcement learning algorithm is the solution of the Markov Decision Process (MDP). Its foremost property states that given the present, the future is independent of the past. It means that the current state concisely represents all the history, hence there is no need to keep previous records. Board games, such as chess or checkers are a good example of this property, in which the current configuration of pieces on the board at any stage of the match is sufficient for a player to make a decision on the next move. The MDP is composed of 4 components:

- Environment state space (S).
- Total action space (A).
- Probability distributions of state transitions ($p(s_{t+1}|s_t,\ a_t)$).
- Probability distribution governing the rewards received ($q(s_{t+1}|s_t,\ a_t)$).

Overall, an MDP task can be discretized into time periods, where at each period t the agent occupies a state $s_t \in S$, and then chooses an action a_t from the set of all possible actions within the current state. The execution of the selected action induces an environmental state s_{t+1} and results in the allocation of a reward $R(s_t,\ a_t)$. The state transition probability $p(s_{t+1}|s_t,\ a_t)$ governs the probability that the agent will transition to state s_{t+1} as a result of choosing a_t in s_t. $q(s_{t+1}|s_t,\ a_t)$ represents the expected reward received by the agent after transitioning from state s_t to s_{t+1} by executing action a_t. Solving an MDP results in the output of a policy, which is a mapping from states to actions guiding the agent's decisions over the entire learning period.

Table 1
DHW tank characteristics.

Domestic Hot Water tank	
Volume	200 L
Material	Duplex 2304 stainless (EN10088)
Time to raise DHW tank temp 15–65 °C	22.75 min
Time to reheat 70% of DHW tank to 65 °C	17.17 min
Heat loss	1.91 kWh/24 h

Specific problems where the model is complete and fully observable can be solved through traditional techniques such as dynamic programming. However, the vast majority of real-world problems cannot be fully observed, as part of the environment may be unknown. To solve this issue, the possible solutions are to estimate the model by using statistical methods (Model-based Reinforcement Learning) or by directly estimating its value function or policy (Model-Free Reinforcement Learning), where learners attempt to directly approximate a control policy through environmental interactions [7–13]. Buildings and heat pumps can be modelled as an MDP if the complete information about the environment is available. However, in case of some missing information, dynamic programming cannot be applied to generate an optimal or near-optimal policy. As an option, the missing model may be either approximated by a Model Based Reinforcement Learning algorithm, or directly have its policy or value function estimated by a model-free Reinforcement Learning technique, such as Q-Learning [40]. This method is part of the temporal difference methods and is capable of making predictions incrementally and in an online way. For each time where a state is non-terminal, the update rule in Equation (1) is calculated.

$$Q(s_t, a_t) \leftarrow Q(s_t, a_t) + \alpha[r_{t+1} + \gamma\ max_a Q(s_{t+1}, a) - Q(s_t, a_t)] \tag{1}$$

Actions are selected according to the policy π and approximations of $Q^{\pi}(s_t, a_t)$ are calculated after each time interval. To balance exploration and exploitation and achieve the best results, an action selection policy, for instance $\varepsilon - greedy$, can be selected. The $\varepsilon - greedy$ strategy chooses the best action from the policy most of the time, however the agent explores a certain amount of the time ruled by ε. After a number of trials, the agent should converge to a near-optimal operation, which means it will choose the actions that leads to the highest reward. In our experiments, a near-optimal policy means that the agent achieved the comfort standards of a rule-based controller and managed to do this at the lowest electricity cost possible, when PV production is high and/or the tariff is low. The importance of future rewards is governed by the discount factor γ, situated in a range from 0 to 1. Values close to 1 assigns a greater weight to it, while a value close to 0 considers only the most recent rewards. Finally, α is a value lower than 1 that represents the learning rate of value estimates over the learning process.

Overall, tabular Q-Learning methods need to interact with the environment and revisit states a number of times in order to learn a good policy. The learning time to converge to an a near-optimal policy depends on the state-space size, where for each additional state or action included, the problem size can increase exponentially, which is known as the curse of dimensionality. This effect of this issue can be mitigated against by replacing the Q-table with a function approximator, using for example an artificial neural network to estimate the Q-values using Equation (1). Mnih et al. [60], followed by Wei et al. [41], used this method, where the Q-value estimates for all outputs can be calculated by performing one forward pass (inference) in the neural network. The loss function is calculated as the mean squared error between the target Q-value and the inferred output of the neural network, and the weights of the neural network are updated using a gradient descent method. As this method has a built-in back-propagating optimizer for the learning process, the learning rate α is no longer needed in Equation (1). Hence, both $Q(s_t, a_t)$ terms cancel each other, resulting in the simplified Equation (2).

$$Q(s_t, a_t) \leftarrow r_{t+1} + \gamma\ max_a Q(s_{t+1}, a) \tag{2}$$

For the purposes of this work we utilize this method, called Deep Q-learning, as it can manage larger state-action spaces and bring

more possibilities of increasing learning speed by estimating values.

4. Experimental setup

This section describes the 3 experiments proposed in our research, showing their algorithms and reward structure. Overall, all of them have the same common target, which is to achieve financial savings by using the heat pump to heat the hot water tank when the cost of electricity is low, which happens either when PV production is high, or the tariff applied at that specific time is reduced. As the controller has also to keep the tank temperature between the specified setpoints set by the user, this is a multi-criteria learning RL problem. Starting with the electricity cost (Ec), we defined it as the cost to perform an action at a time t, calculated using Equation (3). In summary, it is the difference between the energy consumed by the heat pump and PV production, multiplied by the tariff at a specific time.

$$Ec = \left(HP_{power} - PV_{prod}\right) * Tariff \tag{3}$$

For instance, let's assume that a heat pump consumes on average 2 kWh when heating the water and the tariff cost is 0.0915 EUR/kWh for off-peak (from midnight to 9 a.m.) and 0.185 EUR/kWh for other periods [39]. Considering also the average PV production per hour of May 2020 from one of the houses, Fig. 1 shows the average electricity cost per hour to activate the heat pump for DHW purposes. The line represents the energy to be imported from the grid considering heat pump on (total HP consumption minus PV production) times the cost of electricity at the specific hour. Note that in this example the best time to use the heat pump is from midnight to 9 a.m. (reduced tariff) and from 1 p.m. to 6 p.m. (high PV production). In theory, the aim of the control algorithm in this regard is to operate within these ranges. This example shows the monthly average; however, the near-optimal operation times will vary across the days, hence the algorithm will need to adapt itself according to the circumstances.

Moving to tank temperature control, the agent's performance is measured according to the difference of the actual tank temperature and the lower temperature or maximum temperature in the action zone, as can be seen in Equation (4). The action zone ranges from the minimum temperature set by the user and the average of the upper and lower setpoints. For example, if the minimum and maximum tank temperatures are defined as 40 °C and 50 °C, respectively, the action zone range will be from 40 °C to 45 °C. The reason for limiting the action zone is to avoid exceeding the maximum temperature setpoint, as the heating process happens considerably fast (see Table 1). On the other hand, as the tank is well-insulated, heating losses are slow, hence an action zone of 5 °C gives the algorithm enough flexibility to perform actions.

$$\Delta T = \begin{cases} |T_{actual} - T_{zone\_min}|, & if\ T_{actual} < T_{zone\_min} \\ |T_{actual} - T_{zone\_max}|, & if\ T_{actual} > T_{zone\_max} \\ 0, & otherwise \end{cases} \tag{4}$$

The timestep is discretized to 15-min periods, where the system analyses the current state, performs an action, and moves to another state. The states are independent, as they only rely on the heat pump and PV production conditions at the current time, hence following the Markov property. The state-space S comprises the PV production, tank temperature and hour of the day. After performing an action, the environment will move to a new state and receive a reward r_t, which is 0 if the action is water heating "on" and the temperature is below T_{zone_min} or water heating is "off" and the temperature is above T_{zone_max}. For all other possibilities, r_t is calculated based on Equations (3) and (4). A constant σ was added to balance the targets of temperature setpoints and economic savings, resulting in Equation (5).

$$r_t = -(Ec + \Delta T * \sigma) \tag{5}$$

The number of tank temperatures and PV production inputs may vary depending on the algorithm and the number of houses included, as detailed in the next subsections. In a simplified way, the proposed DRL control approaches follow the steps of Algorithm 1. The difference among them is basically the environment parameters, for instance, if houses are treated individually or as a group, and the way agents access the policy and previous experiences.

ALGORITHM 1: Generic DRL Heat Pump Control

Initialize Environment parameters
Initialize $Q(s, a)$ arbitrarily
  Repeat (for each episode)
  Initialize s
  repeat
    Choose a from s using policy derived from $Q(\epsilon - greedy)$
    Take action (a)
    Update environment states (s_{t+1})
    Calculate reward (r)
    $Q(s_t, a_t) \leftarrow r_{t+1} + \gamma\, max_a Q(s_{t+1}, a)$
    $s \leftarrow s_{t+1}$
  until s is terminal
end

4.1. Independent learners

The aim of this first experiment is to deploy the individual heat pump control DRL agent, which will be used as a base for comparisons in later stages. It considers that each of the houses has its own PV, heat pump and control system, operating without having any communication with other houses in the community, as can be seen in Fig. 2. The neural network (NN) designed to estimate the Q-values is similar to the one presented by Wei et al. [41], from the HVAC domain. It is composed of an input layer, two hidden layers, and an output layer. The input values are tank temperature, PV production, and hour of the day, but first, they are normalized on a scale from 0 to 1 (min-max normalization). This helps the algorithm to achieve a more stable learning process, as it does not have to work with numbers with different dimensions. As an example, tank temperature values are lower than one hundred degrees, while PV production can go up to two thousand W. Then, there are two hidden layers, where softmax is the activation function chosen, followed by the output layer with possible actions (turn on or off DHW). The losses are calculated by the mean squared difference between the current output and the ideal target values. Furthermore, we use a Gradient Descent Optimizer, which adjusts weights in the neural network to minimize these losses. The proposed algorithm uses an $\varepsilon - greedy$ policy to improve efficiency. Finally, rewards are calculated individually for each of the houses following Equation (5).

4.2. Independent learners with transfer learning

The second experiment consists of having individual learners in each of the houses, similarly to the independent learners' experiment, but now the agents can share their knowledge among the

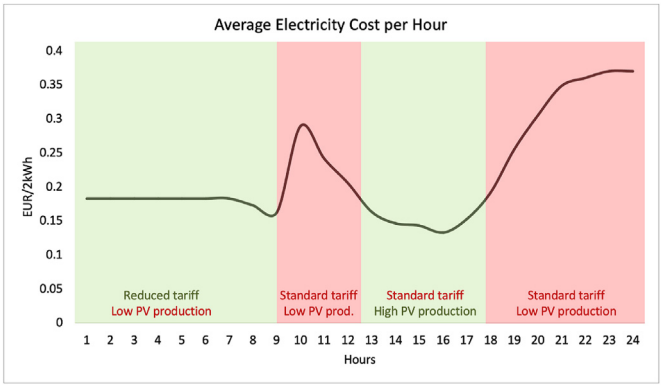

Fig. 1. Average electricity cost (may 2020).

other houses in the community, as illustrated in Fig. 3. The NN structure, normalization process and rewards follow the same approach as 4.1 and Algorithm 1. As the equipment available in the houses is the same, the agents update their respective policy in each timestep but also take into account the neighbors' policy. For instance, if an agent from a specific house faces the same state that was previously visited by any other house, it will either choose the same action performed (exploit) or explore, if the agent decides to take another action. Similar to the previous experiment, the possible actions for each agent are to turn on or turn off DHW. Next, it will receive a reward and update its own policy according to the new experience learnt. Similarly, the agents from other houses will follow this process. The idea is to reduce the number of trials a single agent has to perform to learn near-optimal policies by using the knowledge obtained from the agent that already visited some particular state. This concept is called parallel transfer learning [57,58], where the knowledge to be shared between agents does not need to wait until the end of the process to be available. Similarly, da Silva and Costa [59] proposed that agents can advise each other in a multi-agent system composed of simultaneously learning agents.

4.3. Global learner

The third experiment explores a different solution, where a global agent is used to control all heat pumps in the community, as presented in Fig. 4. The general idea follows Algorithm 1, but now there is only one NN that receives all tank temperatures. Also, as the energy production is very similar among the houses (PV systems have the same capacity and are very close geographically), we considered the average of the PV production as input instead of having individual inputs for each of the PV systems. In this new model, for each timestep the agent reads all tank temperatures and the PV average at once, and takes an action that can be turning on only one of the heat pumps or keeping them all off. The rewards are now calculated considering the summation of temperature difference values (Equation (4)) and the electricity cost (Equation (3)). These changes impact on the setpoint and savings balance, hence σ will have a different value to keep a good performance in both. In summary, the global agent has a higher number of inputs and outputs compared to the previous models, but control actions cannot be done in parallel. In theory, as the tank temperature drop is slow, the agent can manage the time slots of activation and keep the performance without violating the lower setpoint temperature.

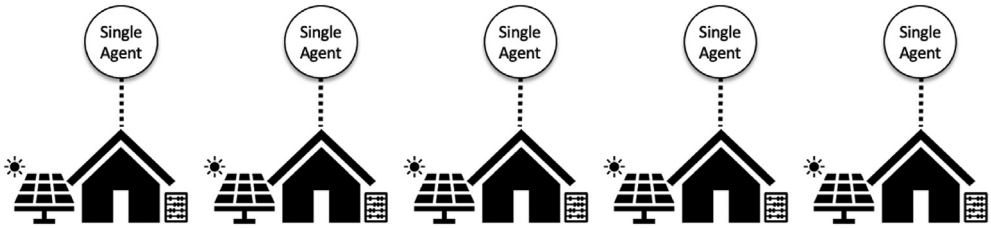

Fig. 2. Independent Learners topology.

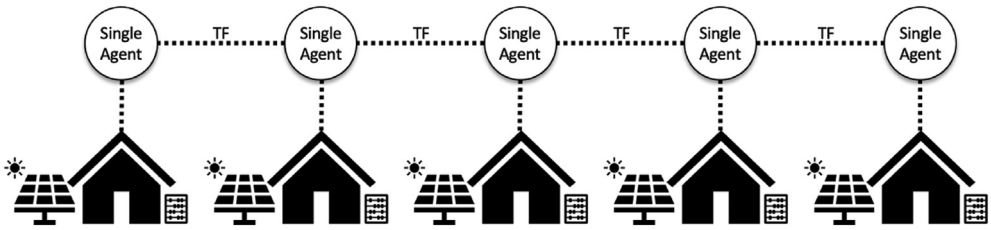

Fig. 3. Independent Learners with Transfer Learning topology.

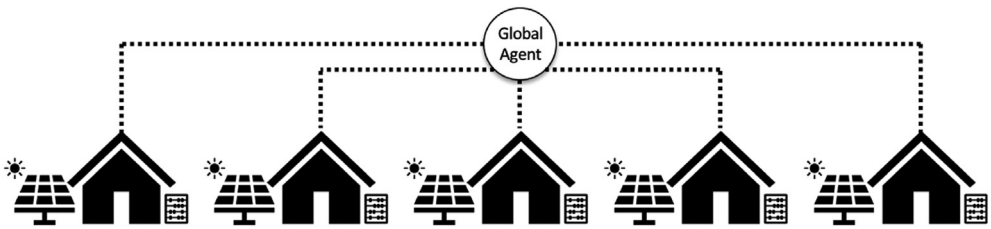

Fig. 4. Global Learner topology.

5. Results

Several tests with different parameters were performed to determine the best configuration for balancing energy savings and tank temperature matters. The final NN architecture is composed of 8 neurons in each of the hidden layers, with $\alpha = 0.0001$, $\gamma = 0.95$ and $\varepsilon = 0.05$, where these values are fixed during the whole experiment for a fair comparison of learning speed. First, we removed *Ec* and selected $\sigma = 1$ in Equation (5), to certify that the algorithm was achieving a similar performance to the rule-based method. Next, we added *Ec* and started to tune the value of σ, in which the final value was defined as 100. Values close to zero mean low savings, while high values affect the tank temperature as the algorithm will start to focus only on savings and possibly will not turn on the equipment even if the temperature is lower than the expected setpoint. The final control algorithms were analyzed in terms of control efficiency, energy savings, and convergence speed.

5.1. Control's efficiency

After selecting and balancing the algorithm parameters, simulations were performed for the independent and global learner. The agent's goal was to achieve a near-optimal policy by interacting with the environment, the details about the number of episodes needed and convergence speed is found in subsection 5.3. The first metric to evaluate the control's efficiency was to certify the agent's ability to keep the tank temperature in an acceptable range, so we selected one random month (episode) after the learning process achieved the near-optimal policy. After achieving stability, both independent learners' methods (with and without transfer learning) have the same result, hence this analysis considers only one independent and one global agent, both with 5 houses. Fig. 5 summarizes the tank temperatures for each of the methods.

The independent learners presented a slightly more uniform temperature distribution, as each agent had to handle only their own tank temperature. Their average temperatures were around 45 °C, with a few exceptions where it achieved up to 60 °C. This can represent, for instance, a moment when PV production is high and the agent decides to heat the water creating a buffer, which will consequently postpone later actions. The overall performance of the global agent was similar, as managing 5 tank temperatures seems to be a more challenging task. In this control strategy, rewards of the five houses are accumulated, and sometimes the agent can prioritize one house over another, as long as the net rewards are low, and this can result in more temperatures out of the expected range.

The second criteria for a control's efficiency analysis consist of evaluating if the control algorithm can optimize energy usage by performing actions when the electricity cost is lower, as presented in Fig. 1. Similarly to the first test, we selected a month after a near-optimal policy was achieved, then we verified the accumulated distribution of the loads over the hours for the two DRL algorithms (independent and global) and also compared them with a rule-based controller, as can be seen in Fig. 6. The average PV production of May 2020 is also plotted in the chart as a reference for a quick understanding of the best control periods.

The results show that the independent learner (green line) was able to perform water heating actions in the periods of low energy price (midnight to 9 a.m.) or high PV production (noon to 5 p.m.), and that brings benefits for both end-users and utilities. In a simple manner, the first one gets a reduction in the electricity bill, while the second will have a load reduction during the peak hours. According to EirGrid [64], the Irish national grid operator, the energy peak in Ireland happens at 5 p.m., which matches with one of the periods where energy imported from the grid is lower in the independent learner experiment. On a large scale, it could help even more the utility, by avoiding unnecessary upgrades in the only connection point with the main island [5].

Due to the tank dynamics, on average, the heat pump needs to perform a control action more than once a day to keep the temperature in between the setpoints. The agent must coordinate actions in such a way the next action will also be in periods of low energy cost. For instance, although hours 2–4 a.m. have a low cost, performing an action in that period may result in a next action to be done in the period of high cost 9 a.m. to noon due to a temperature drop below the setpoint. In that case, it would be better to anticipate the action, making the next action fall before 9 a.m. (first low-cost period), or delay it to 5–7 a.m., so the next action would probably be performed in the next low-cost period (high PV production). The global agent has a similar distribution of the rule-based control, having the loads well-distributed over the hours, and one of the reasons is the difference in the process flow. In the independent learners' approach, actions can be performed in parallel in each 15-min timestep, while the global agent only allows one heat pump activation per timestep.

5.2. Energy savings

In the energy savings analysis, we compared the amount of energy consumed from the grid by the independent and global

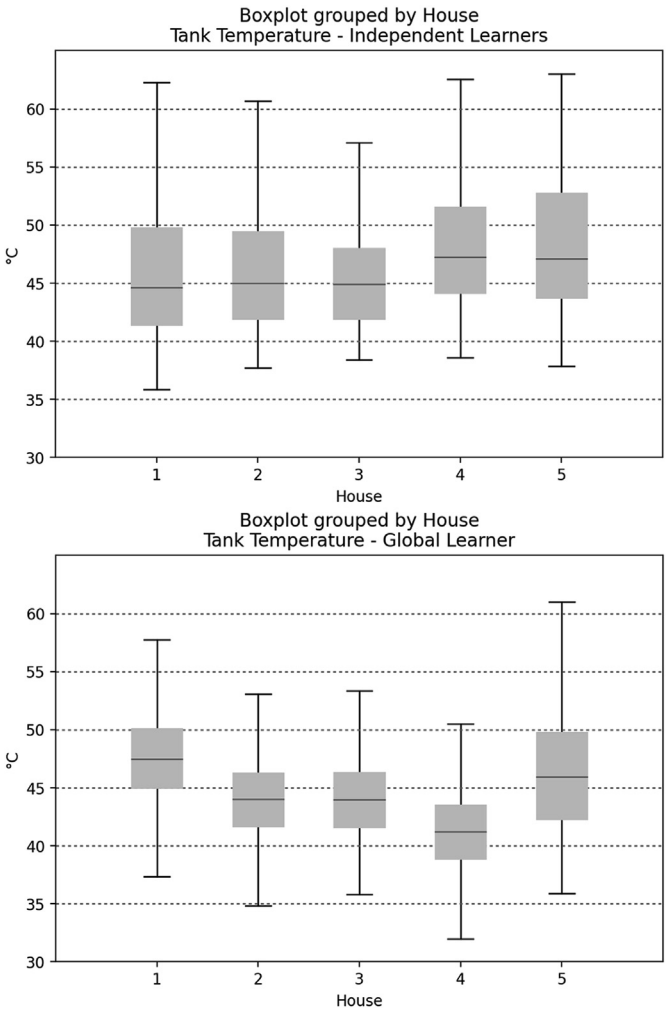

Fig. 5. Tank temperatures demonstration.

learners against the rule-based method, our reference model. Overall, the independent learner performed better in all the houses, achieving up to 9.55% of reduction. The global learner reached almost 5% of reduction in the best case, but the average of this model was just around 2%. The individual information for each model and house can be found in Table 2. Without considering reductions from the PV production, the gross average energy consumption presented across the DRL-based models was reduced by 2% for most of the houses. Only house 2 presented an increase of 1%, but this fact does not mean a real increase in terms of final savings, as the actions of this specific house could have happened in periods of low-cost tariff or high PV production, resulting in a net value could be lower. In summary, as the comfort element was the first control target, the average gross energy consumption was very similar to the rule-based model, but the difference is the time energy is consumed.

Regarding financial savings, we verified the energy bought from the grid for each of the periods, low-cost and high-cost tariffs. Both DRL-based models achieved an average of 5% compared to the rule-based, with a maximum reduction of 7.46%. This maximum is lower than the 9.55% found for energy savings and the main reason is the ToU tariff chosen, which has a long period of low-price energy (9 h a day). Considering that heat pumps have to turn on more than once a day for water heating, there is a big probability that one of the cycles will be within the low-cost tariff, even operating through a rule-based model.

5.3. Convergence speed

The convergence speed test aims to verify if transfer learning helped the algorithm to achieve a near-optimal policy faster. First, we evaluated the performance of an independent learner (subsection 4.1), where we identified the number of episodes necessary to learn the control strategy, as can be seen in Fig. 7. Each episode represents the whole month of May 2020, and the agent took around 11 episodes to reach the plateau, mainly because the tank heating losses are slow. Overall, it is necessary to execute only two actions on average for water heating in an entire day, hence visiting the same set of states can take a large number of trials even using estimates from the NN. Secondly, we performed an experiment to assess the independent learner with transfer learning applied (subsection 4.2), considering a different number of houses sharing experiences in each of the tests ($H = \{1, 2, 3, 4, 5\}$). The idea is to compare if rewards are reduced even from early episodes, thus achieving a near-optimal control policy faster than letting a single agent learn by itself.

In summary, the algorithm convergence speed increases gradually according to the number of houses sharing experiences. Thus, the best result can be found in $H = 5$, where the total reward is considerably lower since episode 0 compared to the other experiments. In this scenario, the learning time to train near-optimal control policies was reduced by more than a factor of 5. This improvement is achieved mainly because agents are most of the time at different states, hence their knowledge is complementary to each other. Considering 4 houses ($H = 4$), the time was reduced by a third. Although the experiment with three houses ($H = 3$) presented a lower initial total reward compared to $H = 2$, both achieved stability in episode 8.

Besides improving convergence speed, reducing total rewards in the initial episodes also helps the agent to present an initial better behavior. For instance, it will partially perform the right water heating actions following its policy instead of just guessing randomly. The difference between the total rewards of the agent without TF (reference) and the agent with TF is known as Jumpstart [32,33]. In our experiment, the Jumpstart also increased gradually following the number of houses included.

6. Conclusions and future works

This work presented a new DRL algorithm with a transfer learning approach for a DRL-based heat pump control to leverage energy efficiency in a microgrid. The main objective was to control the heat pump tank temperature at the lowest energy cost possible, by optimizing PV self-consumption and taking advantage of the Time-of-Use tariff. Moreover, transfer learning was applied to speed up the learning process, hence achieving a satisfactory

Table 2
Energy savings.

	Rule-based	Independent DRL-based		Global DRL-based	
House	Ref. Avg. (kWh)	Total (kWh)	Reduction	Total (kWh)	Reduction
#1		29.86	−5.12%	31.15	−1.02%
#2		28.57	−9.23%	31.03	−1.40%
#3	31.47	29.86	−5.13%	30.31	−3.69%
#4		28.46	−9.55%	29.97	−4.77%
#5		28.93	−8.06%	32.12	2.06%

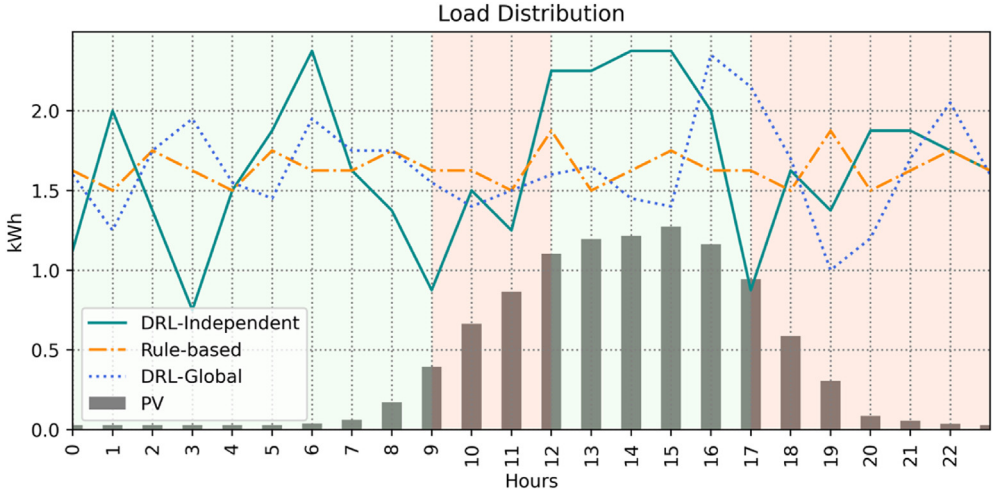

Fig. 6. Load distribution per control type.

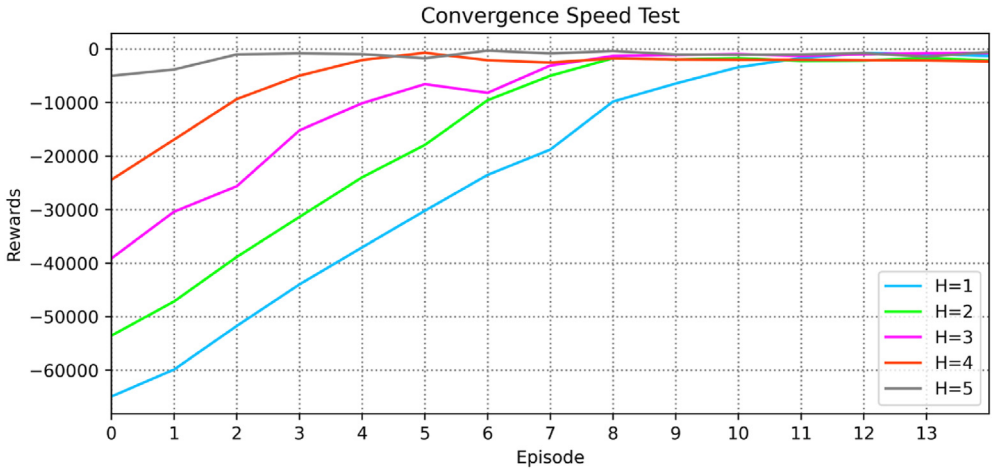

Fig. 7. Convergence speed.

performance since early trials. A total of three different DRL topologies were proposed and tested, having their results compared to a rule-based reference model. The energy savings of almost 10% achieved by the TF DRL algorithm is similar to the values found by Refs. [9,30]. In addition to that, our TF approach reduced the time to learn near-optimal policies by more than a factor of 5, considering the whole cluster of houses. This performance is similar to the one achieved by the Q-Learning model with TF proposed by Ref. [13] in the HVAC domain.

Our test case was based on a real-life project [4], located in the Aran Islands, Galway, Ireland. Although we selected five houses, the island has around 450 dwellings with similar characteristics that could be part of a project in future works, totalling a significant amount of load. The proposed algorithm will be improved to incorporate space heating, and other types of targeted optimisation, such as the demand curve of the neighbourhood. About the Time-of-Use tariff chosen, new tests with a shorter period of low cost will be carried out to better quantify energy savings and algorithm accuracy. Our research showed how transfer learning can speed-up convergence in DQNs. New experiments will be deployed to understand the benefits of transfer learning in other RL methods, including, but not limited to, soft-actor-critic (SAC) and tabular methods. Furthermore, the algorithm will be enhanced with experience replay, which was not considered in the current version to isolate the transfer learning performance, both versions will be compared in future work. Finally, the adaptability of the agent and transfer learning will be tested considering heat pump with different dynamics.

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.

Acknowledgements

This research work was funded by the European Union under the RESPOND project with Grant agreement No. 768619 and the REACT project with Grant agreement No. 824395.

References

[1] Eurostat. Energy consumption in households. 2018. https://ec.europa.eu/eurostat/statistics-explained/index.php/Energy_consumption_in_households. [Accessed 15 June 2020].
[2] International Energy Agency (IEA). Renewables 2017: analysis and forecasts to 2022. 2017. https://doi.org/10.1787/re_mar-2017-en.
[3] European Commission. 2030 climate & energy framework. https://ec.europa.eu/clima/policies/strategies/2030_en. [Accessed 15 June 2020].
[4] RESPOND. Integrated demand response solution towards energy positive neighborhoods. 2020. http://project-respond.eu. [Accessed 10 November 2020].
[5] REACT. Renewable energy for self-sustainable island communities. 2020. https://react2020.eu. [Accessed 10 November 2020].
[6] Mortaji H, Ow Siew Hock, Moghavvemi M, Almurib HAF. Smart grid demand response management using internet of things for load shedding and smart-direct load control. In: 2016 IEEE industry applications society annual meeting, portland, OR; 2016. p. 1–7. https://doi.org/10.1109/IAS.2016.7731836.
[7] Kazmi Hussain, Mehmood Fahad, Lodeweyckx Stefan, Driesen Johan. Gigawatt-hour scale savings on a budget of zero: deep reinforcement learning based optimal control of hot water systems. Energy 2017;144. https://doi.org/10.1016/j.energy.2017.12.019.
[8] Ruelens F, Claessens BJ, Quaiyum S, De Schutter B, Babuška R, Belmans R. Reinforcement learning applied to an electric water heater: from theory to practice. IEEE Transactions on Smart Grid July 2018;9(4):3792–800. https://doi.org/10.1109/TSG.2016.2640184.
[9] De Somer O, Soares A, Vanthournout K, Spiessens F, Kuijpers T, Vossen K. Using reinforcement learning for demand response of domestic hot water buffers: a real-life demonstration. In: 2017 IEEE PES innovative smart grid technologies conference Europe (ISGT-Europe), torino; 2017. p. 1–7. https://doi.org/10.1109/ISGTEurope.2017.8260152.
[10] Al-jabery K, Xu Z, Yu W, Wunsch DC, Xiong J, Shi Y. Demand-side management of domestic electric water heaters using approximate dynamic programming. IEEE Trans Comput Aided Des Integrated Circ Syst May 2017;36(5):775–88. https://doi.org/10.1109/TCAD.2016.2598563.
[11] Patyn C, Reymond M, Rădulescu R, Deconinck G, Nowé A. Reinforcement learning for demand response of domestic household appliances. In: Adaptive learning agents 2018 proceedings; 2018. p. 1–7.
[12] Patyn C, Peirelinck T, Deconinck G, Nowe A. Intelligent electric water heater control with varying state information. In: 2018 IEEE international conference on communications, control, and computing technologies for smart grids (SmartGridComm), aalborg; 2018. p. 1–6. https://doi.org/10.1109/SmartGridComm.2018.8587453.
[13] Lissa P, Schukat M, Barrett E. Transfer learning applied to reinforcement learning-based HVAC control. SN COMPUT. SCI. 2020;1:127. https://doi.org/10.1007/s42979-020-00146-7.
[14] Kasthurirangan Gopalakrishnan, Khaitan Siddhartha K, Choud- hary Alok, Agrawal Ankit. Deep Convolutional Neural Networks with transfer learning for computer vision-based data-driven pave- ment distress detection. Construct Build Mater 2017;157:322–30.
[15] Hoo-Chang Shin, Roth Holger, Gao Mingchen, Lu Le, Xu Ziyue, Isabella Nogues, Yao Jianhua, Mollura Daniel J, Summers Ronald M. Deep convolutional neural networks for computer aided detection: CNN architectures, dataset characteristics and transfer learning. IEEE Trans Med Imag 2016;35:1285–98.
[16] Long Mingsheng, Wang Jianmin, Ding Guiguang, Sun Jia-Guang, SYu Philip. Transfer feature learning with joint distribution adap- tation. IEEE Int Conf Comput Vis 2013;2013:2200–7.
[17] Bianchi Reinaldo AC, Celiberto Luiz A, Santos Paulo E, Matsuura Jackson P, Ramon Lopez, de Mantaras Ramon Lopez. Transferring knowledge as heuristics in reinforcement learning: a case- based approach. Artif Intell 2015;226:102–21.

[18] Chen Tessler, Givony Shahar, Tom Zahavy, Mankowitz Daniel J, Mannor Shie. A deep hierarchical approach to lifelong learn- ing in minecraft. In: Proceedings of the thirty-first AAAI conference on artificial intelligence (AAAI17). AAAI Press; 2017. p. 1553–61.
[19] Chaplot DS, Lample G, Sathyendra KM, Salakhutdinov R. Transfer deep reinforcement learning in 3d environments: an empirical study. In: NIPS deep reinforcemente leaning workshop; 2016.
[20] Teh Y, Bapst V, Czarnecki WM, Quan J, Kirkpatrick J, Hadsell R, Heess N, Pascanu R Distral. Robust multitask reinforcement learning. In: Advances in neural information processing systems; 2017. p. 4496–506.
[21] Shepherd A, Batty W. Fuzzy control strategies to provide cost and energy efficient high quality indoor environments in buildings with high occupant densities. Build Serv Eng Technol 2003;24(1):35–45.
[22] Calvino F, La Gennusa M, Rizzo G, Scaccianoce G. The control of indoor thermal comfort conditions: introducing a fuzzy adaptive controller. Energy Build 2004;36(2):97–102.
[23] Wei T. Design and management for energy-efficient cyber-physical systems. UC: Riverside; 2018.
[24] Wei Tianshu, Chen Xiaoming, Li Xin, Zhu Qi. Model-based and data-driven approaches for building automation and control. 2018. p. 1–8. https://doi.org/10.1145/3240765.3243485.
[25] Patyn Christophe, Deconinck Geert. Electric water heater control through informed fitted Q-iteration. 2019. p. 1–5. https://doi.org/10.1109/ISGTEurope.2019.8905737.
[26] Vázquez-Canteli José R, Nagy Zoltán. Reinforcement learning for demand response: a review of algorithms and modeling techniques. Appl Energy 2019;235:1072–89. https://doi.org/10.1016/j.apenergy.2018.11.002. ISSN 0306-2619.
[27] Mason Karl, Grijalva Santiago. A review of reinforcement learning for autonomous building energy management. Comput Electr Eng 2019;78:300–12. https://doi.org/10.1016/j.compeleceng.2019.07.019. ISSN 0045-7906.
[28] Han Mengjie, May Ross, Zhang Xingxing, Wang Xinru, Pan Song, Yan Da, Jin Yuan, Xu Liguo. A review of reinforcement learning methodologies for controlling occupant comfort in buildings. Sustainable Cities and Society 2019;51:101748. https://doi.org/10.1016/j.scs.2019.101748. ISSN 2210-6707.
[29] Soares Ana, Geysen Davy, Fred Spiessens, Ectors Dominic, De Somer Oscar, Vanthournout Koen. Using reinforcement learning for maximizing residential self-consumption – results from a field test. Energy Build 2020;207:109608. https://doi.org/10.1016/j.enbuild.2019.109608. ISSN 0378-7788.
[30] Yang Lei, Nagy Zoltan, Goffin Philippe, Schlueter Arno. Reinforcement learning for optimal control of low exergy buildings. Appl Energy 2015;156:577–86. https://doi.org/10.1016/j.apenergy.2015.07.050. ISSN 0306-2619.
[31] Patyn C, Ruelens F, Deconinck G. Comparing neural architectures for demand response through model-free reinforcement learning for heat pump control. In: 2018 IEEE international energy conference (ENERGYCON), limassol; 2018. p. 1–6. https://doi.org/10.1109/ENERGYCON.2018.8398836.
[32] Taylor Matthew E, Stone Peter. Transfer learning for reinforcement learning domains: a survey. J Mach Learn Res 2009;10:1633–85.
[33] Silva Felipe, Costa Anna. A survey on transfer learning for multiagent reinforcement learning systems. J Artif Intell Res 2019;64. https://doi.org/10.1613/jair.1.11396.
[34] Xu Shichao, Wang Yixuan, Wang Yanzhi, Zheng O'Neill, Qi Zhu. One for many: transfer learning for building HVAC control. In: Proceedings of the 7th ACM international conference on systems for energy-efficient buildings, cities, and transportation (BuildSys '20). New York, NY, USA: Association for Computing Machinery; 2020. p. 230–9. https://doi.org/10.1145/3408308.3427617.
[35] Chen Yujiao, Tong Zheming, Zheng Yang, Samuelson Holly, Norford Leslie. Transfer learning with deep neural networks for model predictive control of HVAC and natural ventilation in smart buildings. J Clean Prod 2020;254:119866. https://doi.org/10.1016/j.jclepro.2019.119866. ISSN 0959-6526.
[36] Mocanu Elena, Nguyen Phuong H, Kling Wil L. Madeleine Gibescu, Unsupervised energy prediction in a Smart Grid context using reinforcement cross-building transfer learning. Energy Build 2016;116:646–55. https://doi.org/10.1016/j.enbuild.2016.01.030. ISSN 0378-7788.
[37] Mitsubishi Electric. Ecodan renewable heating technology data book, vol. 4; 2018. https://www.mitsubishi-les.info/database/servicemanual/files/201803_ATW_DATABOOK.pdf.
[38] Mitsubishi Electric. Mitsubishi electric pre-plumber cylinder with FTC4 control system, travellers lane. Hatfield, Hertfordshire, AL10 8XB, England: Mitsubishi Electric; 2018.
[39] Electric Ireland. Ally you need to know about the NightSaver meter. https://www.electricireland.ie/news/article/all-you-need-to-know-about-the-nightsaver-meter. [Accessed 10 November 2020].
[40] Watkins C. Learning from delayed rewards. England: Ph.D. dissertation, University of Cambridge; 1989.
[41] Wei T, Wang Y, Zhu Q. Deep reinforcement learning for building HVAC control. In: 2017 54th ACM/EDAC/IEEE design automation conference (DAC). Austin; 2017. p. 1–6.
[42] F. Ruelens, B. J. Claessens, S. Vandael, B. De Schutter, R. Babuška and R. Belmans, "Residential demand response of thermostatically controlled loads using batch reinforcement learning," in IEEE Transactions on Smart Grid, vol. 8, no. 5, pp. 2149-2159, Sept. 2017, doi: 10.1109/TSG.2016.2517211.
[43] Kazmi Hussain, Suykens Johan, Balint Attila, Driesen Johan. Multi-agent reinforcement learning for modeling and control of thermostatically controlled loads. Appl Energy 2019;238:1022–35. https://doi.org/10.1016/j.apenergy.2019.01.140. ISSN 0306-2619.
[44] Chen Bingqing, Cai Zicheng, Bergés Mario. Gnu-RL: a precocial reinforcement learning solution for building HVAC control using a differentiable MPC policy. In: Proceedings of the 6th ACM international conference on systems for energy-efficient buildings, cities, and transportation (BuildSys '19). New York, NY, USA: Association for Computing Machinery; 2019. p. 316–25. https://doi.org/10.1145/3360322.3360849.
[45] Zhao Huan, Zhao Junhua, Shu Ting, Pan Zibin. Hybrid-model-based deep reinforcement learning for heating, ventilation, and air-conditioning control. Frontiers in Energy Research. Vol. 8. 2021. DOI: 10.3389/fenrg.2020.610518.
[46] Gupta Anchal, Badr Youakim, Negahban Ashkan, Robin G, Qiu. Energy-efficient heating control for smart buildings with deep reinforcement learning. Journal of Building Engineering 2021;34:101739. https://doi.org/10.1016/j.jobe.2020.101739. ISSN 2352-7102.
[47] Taha Abdelhalim Nakabi, Toivanen Pekka. Deep reinforcement learning for energy management in a microgrid with flexible demand. Sustainable Energy, Grids and Networks 2021;25:100413. https://doi.org/10.1016/j.segan.2020.100413. ISSN 2352-4677.
[48] Kurte Kuldeep, Munk Jeffrey, Amasyali Kadir, Olivera Kotevska, Cui Borui, Kuruganti Teja, Zandi Helia. Electricity pricing aware deep reinforcement learning based intelligent HVAC control. In: Proceedings of the 1st international workshop on reinforcement learning for energy management in buildings & cities (RLEM'20). New York, NY, USA: Association for Computing Machinery; 2020. p. 6–10. https://doi.org/10.1145/3427773.3427866.
[49] Ding Xianzhong, Wan Du, Alberto E, Cerpa. MB2C: model-based deep reinforcement learning for multi-zone building control. In: Proceedings of the 7th ACM international conference on systems for energy-efficient buildings, cities, and transportation (BuildSys '20). New York, NY, USA: Association for Computing Machinery; 2020. p. 50–9. https://doi.org/10.1145/3408308.3427986.
[50] Vazquez-Canteli Jose R, Henze Gregor, Nagy Zoltan. MARLISA: multi-agent reinforcement learning with iterative sequential action selection for load shaping of grid-interactive connected buildings. In: Proceedings of the 7th ACM international conference on systems for energy-efficient buildings, cities, and transportation (BuildSys '20). New York, NY, USA: Association for Computing Machinery; 2020. p. 170–9. https://doi.org/10.1145/3408308.3427604.
[51] Zhang Chi, Kuppannagari Sanmukh R, Kannan Rajgopal, Prasanna Viktor K. Building HVAC scheduling using reinforcement learning via neural network based model approximation. In: Proceedings of the 6th ACM international conference on systems for energy-efficient buildings, cities, and transportation (BuildSys '19). New York, NY, USA: Association for Computing Machinery; 2019. p. 287–96. https://doi.org/10.1145/3360322.3360861.
[52] Ding Xianzhong, Wan Du, Alberto Cerpa. OCTOPUS: deep reinforcement learning for holistic smart building control. In: Proceedings of the 6th ACM international conference on systems for energy-efficient buildings, cities, and transportation (BuildSys '19). New York, NY, USA: Association for Computing Machinery; 2019. p. 326–35. https://doi.org/10.1145/3360322.3360857.
[53] Lissa Paulo, Deane Conor, Schukat Michael, Seri Federico, Keane Marcus, Barrett Enda. Deep reinforcement learning for home energy management system control. Energy and AI 2021;3:100043. https://doi.org/10.1016/j.egyai.2020.100043. ISSN 2666-5468.
[54] Christensen Morten Herget, Ernewein Cédric, Pierre Pinson. Demand response through price-setting multi-agent reinforcement learning. In: Proceedings of the 1st international workshop on reinforcement learning for energy management in buildings & cities (RLEM'20). New York, NY, USA: Association for Computing Machinery; 2020. p. 1–5. https://doi.org/10.1145/3427773.3427862.
[55] Zhang Xiangyu, Jin Xin, Tripp Charles, Biagioni David J, Graf Peter, Jiang Huaiguang. Transferable reinforcement learning for smart homes. In: Proceedings of the 1st international workshop on reinforcement learning for energy management in buildings & cities (RLEM'20). New York, NY, USA: Association for Computing Machinery; 2020. p. 43–7. https://doi.org/10.1145/3427773.3427865.
[56] Nagy A, Kazmi H, Cheaib F, Driesen J. Deep reinforcement learning for optimal control of space heating. 2018. arXiv preprint arXiv:1805.03777.
[57] Taylor A. Parallel transfer learning: accelerating reinforcement learning in multi-agent systems. Trinity College Dublin): Doctoral dissertation; 2016.
[58] Taylor Adam, Dusparic Ivana, Galván-López Edgar, Clarke Siobhán, Cahill Vinny. Transfer learning in multi-agent systems through parallel transfer. In: Proceedings of the 30th international conference on machine learning; 2013. p. 1–9.
[59] da Silva F, Costa A. Accelerating multiagent reinforcement learning through transfer learning. February. In: Proceedings of the AAAI conference on artificial intelligence. vol. 31; 2017. No. 1.
[60] Mnih V, Kavukcuoglu K, Silver D, et al. Human-level control through deep reinforcement learning. Nature 2015;518:529–33. https://doi.org/10.1038/nature14236.
[61] Hedegaard Karsten, Vad Mathiesen Brian, Lund Henrik, Heiselberg Per. Wind power integration using individual heat pumps – analysis of different heat storage options. Energy 2012;47(Issue 1):284–93. https://doi.org/10.1016/

j.energy.2012.09.030. ISSN 0360-5442.
[62] Protopapadaki Christina, Saelens Dirk. Heat pump and PV impact on residential low-voltage distribution grids as a function of building and district properties. Appl Energy 2017;192:268–81. https://doi.org/10.1016/j.apenergy.2016.11.103. ISSN 0306-2619.
[63] Mathiesen BV, Lund H, Connolly D, Wenzel H, Østergaard PA, Möller B, Nielsen S, Ridjan I, Karnøe P, Sperling K, Hvelplund FK. Smart Energy Systems for coherent 100% renewable energy and transport solutions. Appl Energy 2015;145:139–54. https://doi.org/10.1016/j.apenergy.2015.01.075. ISSN 0306-2619.
[64] EirGrid Group. http://www.eirgridgroup.com/. [Accessed 7 August 2021].