LUND UNIVERSITY

SCHOOL OF ECONOMICS AND MANAGEMENT

Household Energy Cost Optimization Using Deep Reinforcement Learning

Authors:
Anna Terekhova
Jade Fang

Lund University School of Economics and Management
Master's Degree Thesis 15 HE Credits, Spring 2022
Subject: DABN01 - Master Essay I
Program: Master of Science in Data Analytics and Business Economics
Supervisor: Krzysztof Podgórski

Acknowledgements

First, we wish to express our gratitude for our thesis supervisor Krzysztof Podgórski for his continuous support, feedback, and expertise. We were often steered right by Krys's wisdom and we thank him for his patience and focus. We would also like to thank our casemaker Sara Moricz for the generous amounts of time, effort, and support she has given us. Her positive attitude and true belief in us made it a blessing for us to work with Sensative.

Abstract

This thesis aims to address the rising energy costs by using IoT technology and reinforcement learning. We use historical sensor data to fit a deep reinforcement learning model that is capable of optimizing the control of a heating system in a way that minimizes energy costs, while maintaining a comfortable indoor temperature. This model-free approach uses neural networks to simulate the thermodynamic behavior of an existing building, making it more cost-effective than using building simulation software. Using the final Deep Q-Network model, a cost reduction of up to 25% was achieved.

Keywords: Energy optimization, deep reinforcement learning, sensor data, neural network, indoor heating, DQN, energy costs

Abbreviations

DRL - Deep reinforcement learning

DQN - Deep Q-Network

FFNN - Feed-forward neural network

IoT - Internet of Things

ML - Machine learning

RL - Reinforcement learning

RNN - Recurrent neural network

SMHI - Swedish Meteorological and Hydrological Institute (weather institute)

Acknowledgements 1

Abstract 2

Abbreviations 3

1. Introduction 5
1.1 Background 5
1.2 Formulating the Energy Cost Optimization Problem 6
1.3 Original Dataset 7
1.4 Research Design and Model Structure 8
1.5 Aim and Contribution to Knowledge 9
1.6 Results 9
1.7 Section Outline 9

2. Literature Review and Theoretical Framework 9
2.1 Data-gathering Methods 10
2.2 Model-based vs. Model-free Approaches 10
2.3 Performance Studies Based on Deep Reinforcement Learning 12

3. DRL Research Design and Methodology 15
3.1 Application of DRL on Energy Cost Optimization 15
3.2 Research Methodology 15
3.2.1 Overview 15
3.2.2 Data Collection and Selection 16
3.2.3 Neural Networks for Indoor Temperature Estimation 19
3.2.4 Custom Reinforcement Learning Environments 21
Environment and Agent 21
States and Actions 22
Episodes and Rewards 23
3.2.5 Deep Reinforcement Learning Setup 24
3.3 Source Critical Consideration 26

4. Empirical Results from DRL Models 27
4.1 Reward Analysis 27
4.2 Action Analysis 28
4.3 Indoor Temperature Analysis 29
4.4 Cost Analysis 31

5. Discussion and Critical Reflection 32

6. Conclusion 34

References 35

1. Introduction

1.1 Background

As climate change increases in urgency, the need for energy conservation has equally increased. In Sweden, around 40% of energy consumption by commodities goes to the residential and services sector, with 80% of that consumption being used primarily for heating (Energimyndigheten, 2021). Although electricity usage has declined since 2001 (Energimyndigheten, 2021), continuous efforts are needed to sustain that trend. In addition, despite the decline in electricity usage, as can be seen in Figure 1, energy prices are hitting new highs regularly, leading to the search for more stringent energy conservation methods in order to keep costs down for consumers. In Sweden, consumer energy prices change by the hour, leading to higher costs when demand is high, and vice versa. Due to this, energy costs for households could be reduced not only by conservation but also by using energy at optimal periods of time.

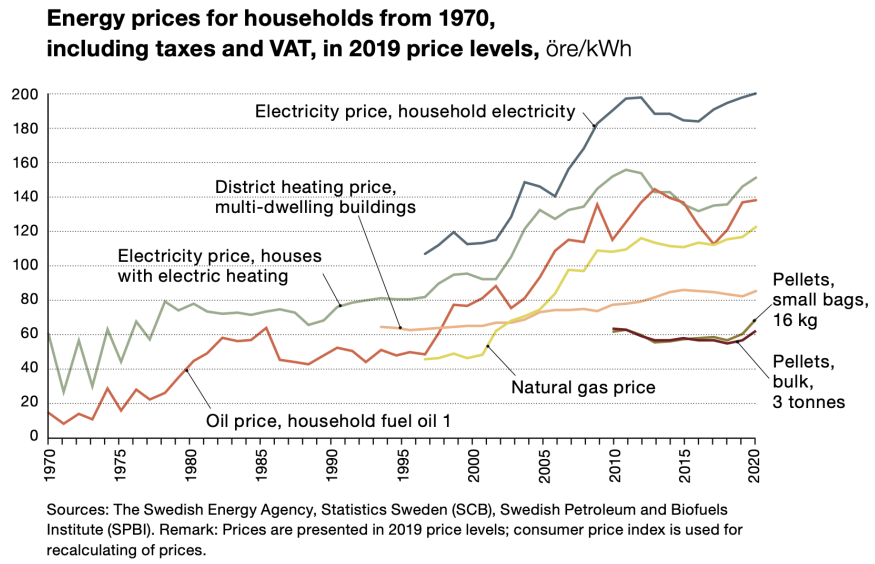

Figure 1. Changing energy prices from different energy sources in Sweden from 1970-2020.

In this paper, deep reinforcement learning models will be built with the objective of optimizing energy costs for an existing residential building. The long-term vision is that by optimizing the cost of electricity in a single household, energy usage can be spread more evenly to reduce the total grid load of the neighborhood, which can lead to better energy conservation overall. With data made available by Internet of Things (IoT) sensors placed in a single-dwelling house located in Lund, Sweden, a deep reinforcement model is applied to understand how the agent can help lower energy costs by turning the heating system on or off at optimized timepoints while maintaining a comfortable indoor temperature. The agent, which represents the model, replaces the notion of the person who would otherwise control the heating system. The use of historical sensor data is more cost-effective compared to the alternative of building a specific thermodynamic model of individual buildings, thus making the proposed method more scalable and quicker to implement.

1.2 Formulating the Energy Cost Optimization Problem

Reinforcement learning (RL) is a machine learning technique that differs in aim from traditional supervised or unsupervised learning. As depicted in Figure 2, RL does not seek to make predictions on future data, yet rather aims to take optimal actions based on the current state of the environment. When it comes to a typical RL problem, an agent interacts with the environment by observing environment states and performs actions that alter the environment with the ultimate objective of maximizing the cumulative reward (Metelli, 2022). Additionally, instead of needing a predefined dataset for algorithm training, RL requires continuous flow of feedback from the environment to the agent based on the results of every action made. In this way, RL uses a Markov decision process that relies only on the resulting current state of the environment in order to choose the next action (Metelli, 2022). Finally, a reward is given based on the result of the action taken by the agent, and the model functions to maximize the total reward earned.

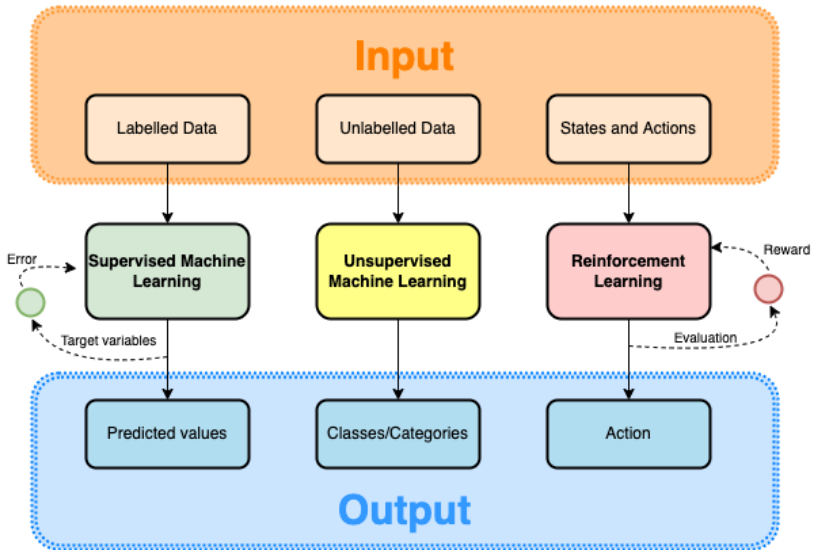

Figure 2. Main differences between supervised, unsupervised, and reinforcement learning.

A major benefit of RL is the fact that no large datasets are required to train the model. This can be quite advantageous to the real-life setting of energy optimization in existing dwellings since it removes the need for a long data collection period for each individual building. Instead, ideally, a reinforcement learning model would be able to read the current temperatures in the building to make a decision about an action, measure the actual resulting temperature after the action has been taken, make a new decision after that, and so forth. However, since a real-life testing environment could not be set up within the length of this study, a prediction of resulting temperatures is made instead.

In the absence of a real-life testing environment, temperature predictions are required for the reinforcement learning models to be trained in a custom reinforcement learning environment. In this thesis, the custom environment defines how the indoor temperature of the house changes based on the decisions made by the RL agent. Specifically, once the model decides whether the heating system should be on or off, the state within the environment is updated with a new indoor temperature. Therefore, to find these new temperatures, different neural networks that predict resulting indoor temperatures were built.

However, it is important to note that these neural networks are not inherently a part of reinforcement learning models, but are only used in this study to substitute for live temperature readings or complex temperature modeling.

Finally, to increase the optimization of RL models, deep reinforcement learning (DRL) is applied. This means that the deep learning process is additionally built into the RL model itself and impacts how the agent decides on the next action. By using a neural network to estimate which action maximizes the reward, the DRL agent is able to determine optimal actions.

The data that will help guide the predictions in the custom environment was collected by sensor strips mounted in a residential home in Lund, Sweden. These sensors were produced by Sensative AB, and the data have been collected and provided by the company as well. The sensor strips are attached to key points both indoors and outdoors, and communicate using either Long Range Wide Area Network (LoRaWAN) or Z-wave technology.

1.3 Original Dataset

The data used in this study are largely gathered by sensors in a single dwelling home. The original dataset was provided by Sensative AB, in which hourly observations were taken between October 20, 2020, to April 30, 2022, for a total of 13,392 observations (a total of 558 days) over the period of 19 months. The variables from the original dataset are listed in Table 1 below.

Variable	**Description**
Date & Time	Date and time of each hourly observation
Outdoor temperature measured by sensor	The hourly outdoor temperature in °C, measured by the outdoor sensor
Outdoor temperature from the weather service	The hourly outdoor temperature in °C, provided by the Swedish weather service (SMHI)
Heating switch status	Proportion of hour that the heating system was turned on
Bathroom, Amy's room, kitchen, Line's room, guest room, living room, entrance temperatures	The hourly indoor temperature per room in the house in °C
Electricity price	The hourly price of electricity in Euro/MWh
Cumulative electricity consumption	The value taken from the consumption meter at that hour, measured in kWh

Table 1. Variables of the original dataset from Sensative AB.

The data have been cleaned and transformed for further analysis. The heating switch status has been changed to the binary variable format from the original decimal value. In addition, the average of all seven indoor temperature sensor values has been calculated to have one overall indoor temperature value

for the building. This condensed form of the dataset was used for training neural networks that were responsible for updating the resulting temperature in the custom environment.

1.4 Research Design and Model Structure

The aim of this study is to explore whether a deep reinforcement model can be built to optimize energy costs while maintaining a comfortable indoor temperature. Thus, the main research question is the following:

RQ: Can a deep reinforcement model lower the energy cost of a household?

The model proposed to address this research question consists of several components that are illustrated in Figure 3. The main aspect of the model is the deep reinforcement learning element depicted in blue, wherein the agent strives to find the optimal actions to take within a given environment to maximize the reward. The neural network for indoor temperature estimation is a separate element that plays a role of a "house" simulator in updating the DRL environment, but is not an inherent part of the DRL model.

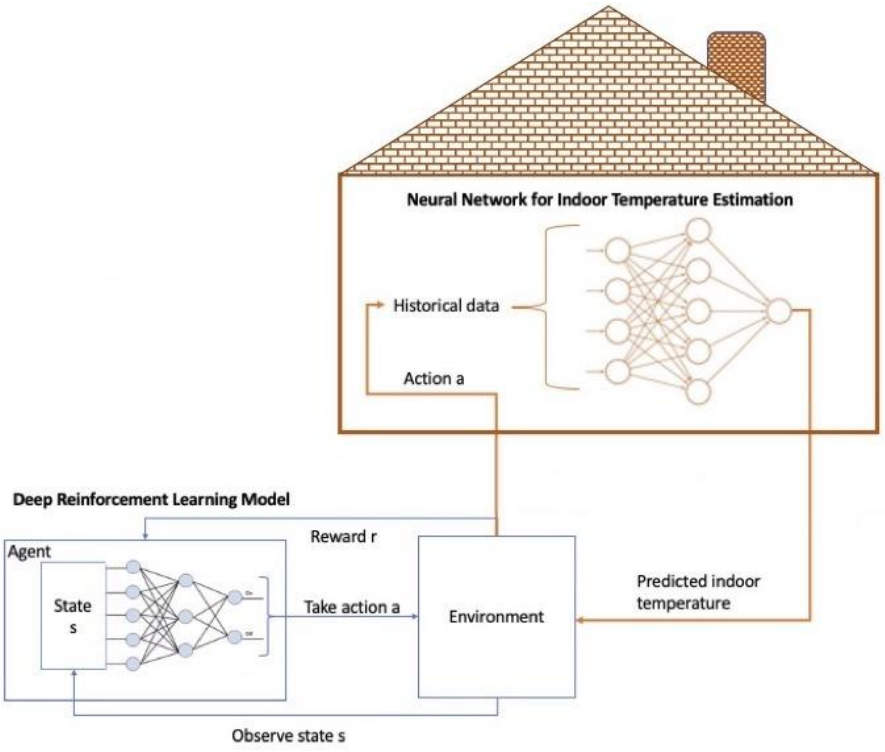

Figure 3. A visualization of how each component in the entire model interacts with each other. Notice that the neural network for temperature estimation exists outside of the reinforcement model, and is not inherently a part of the RL method.

1.5 Aim and Contribution to Knowledge

Many studies that focus on energy cost optimization use software-based building simulators to determine methods for energy reduction. However, apart from being costly, these methods require the knowledge of a multitude of parameters in order to create a proper thermodynamic model. In addition, such models often focus on how the building was originally planned to operate and may not reflect real-life usage. Therefore, this study will contribute to the body of knowledge by proposing the use of historical sensor data and neural networks for the simulation of the thermodynamics behavior of the building. The use of sensor data will allow for cost reduction since the data can be collected much easier and will enable a faster setup that would provide the required scalability.

Furthermore, this study will adopt the model-free approach to energy optimization by using deep reinforcement learning algorithms, which typically require no historical data for the agents to be trained. The main contribution in this aspect is the proposed hybrid method where the agent is trained in the custom reinforcement learning environment that is indirectly provided with historical sensor data. Such a hybrid approach helps to overcome the obstacle of not having access to the real-life building or simulation where the changes in the conditions can be observed in real-time.

1.6 Results

For the house studied in this thesis, the heating system is known to consume 12kWh per hour when the system is on. Under this assumption, throughout the length of the studied heating period, which was October 2021 to April 2022, the total cost of electricity was €4,530. Using the final DRL models, the average total cost of electricity was reduced to €3,367. This indicates that the model was able to produce an overall cost saving of approximately 25%.

1.7 Section Outline

The subsequent sections will include the following: Section 2 explores existing literature that is relevant to the theoretical and methodological aspects of this study, Section 3 details the empirical setting and overall research design, Section 4 explains the results from the DRL model, Section 5 discusses these results with the reference to the aforementioned literature, Section 6 provides a summary and conclusion of the thesis, as well as some limitations of this study.

2. Literature Review and Theoretical Framework

Energy usage modeling has been a well-studied field, whether with the intention of reducing costs or consumption. The methodologies have evolved with the technological capabilities of the field, reaching new possibilities with the current state-of-the-art. The research design proposed in this paper is based on historical sensor data, neural networks for temperature predictions, and deep reinforcement learning algorithms - all different technological aspects that have developed to their current form over the years.

2.1 Data-gathering Methods

Historically, the calculation of energy optimization has been conducted using formulas for heating and cooling, usage curves, and cost functions (Löf & Tybout, 1974) with data less granular than what is available today. The lack of ubiquity of sensors meant that data had to be collected more manually. Additionally, processes such as recording hourly indoor and outdoor temperatures would take a lot more man-hours and could potentially contain human errors and irregularities. This meant that many studies were based on simulated data, produced using building specifications.

The usage of simulated data was rife with imperfections, due to input variables being less tailored for real-world usage patterns, and more for engineering, architectural needs, or regulatory certification. Many papers sought to improve the process of creating simulated energy usage models (Eisenhower et al., 2012), including what kind of variables would be necessary to create a more realistic simulation. Today, a large number of studies use EnergyPlus, a software specifically built for whole-building energy modeling (BEM), which uses hundreds of inputs to provide simulated data, and models energy consumption based on physics-based equations.

Though many studies still use simulated data, they are often based on data gathered by sensors in the first place (Lissa et al., 2021). As the cost of sensors decreased, their proliferation became more widespread, and it has become easier to gather realistic data. Research has shown that even changing the placement of sensors could lead to better energy optimization and that automated control of building elements based on human behavior is successful in lowering energy consumption (Sembroiz et al., 2019). Similarly, this thesis uses historical sensor data instead of purely simulated data to model the thermodynamics behavior of the building without complex physical characteristics.

2.2 Model-based vs. Model-free Approaches

The cost of IoT devices has been decreasing, making them more affordable and available for household use, making a regular home "smart". Thus, greater usage of such devices allowed for more accurate and timely data collection of different data points in the household, which in turn, encouraged more research groups to utilize the collected data to address different optimization problems. Better data collection routines alongside the growing domain of knowledge in different machine learning techniques has resulted in increased interest in applying advanced machine learning techniques for energy management, which allowed for better results in terms of energy savings compared to manually controlled methods (Sembroiz et al., 2019).

Overall, the existing studies on energy cost and energy consumption optimization typically fall into one of the two categories: ones that follow a model-based approach and the others that take a model-free approach. As Yu et al. (2020) explain, the model-based approach is created using the information about the thermal dynamics of the building environment, which requires creating a model of the building's thermal dynamics. On the other hand, a model-free approach can be pursued without the complex data required for building a thermodynamic model. Compared to the model-based approach, the model-free approach overcomes the challenge of developing a complex and costly thermal dynamics model of the building since the machine learning algorithm can be constructed without this information (Yu et al., 2020). Thus, with recent advancements in machine learning techniques, further widespread use of model-free approaches can be expected.

Various model-based modeling methods have been implemented with the intention of minimizing cost or energy consumption. Eisenhower et al. (2012) found that using a Support Vector Machine with a Gaussian kernel, a 45% annual energy reduction could be achieved. Another study used Random Forests to save 24.9% in cooling energy used (Bünning et al., 2020). The particle swarm optimization method is also widely used in this domain, with one study achieving 54% in HVAC energy savings (Barber & Krarti, 2022). Suffice to say, there are many advanced model-based techniques that show significant results, but they require the collection of hundreds of data variables in order to accurately model the building behavior.

One example of a model-free approach is reinforcement learning. It can be considered a “trial-and-error” machine learning method since the reinforcement learning agent “learns” the optimal action strategy by trying various actions and then receiving feedback on the results (Zhang & Lam, 2018). As Lissa et al. (2021) further elaborate, reinforcement learning requires no prior knowledge of the environment since the algorithm can learn the optimal policy by interacting with the environment itself and then choosing actions based on past experiences. Markov decision process and its properties are used to create a model of the environment (Mason & Grijalva, 2019). Once the agent learns how to decide on the best action to take for a given state based on the reward function, the optimal policy is constructed. Lissa et al. (2021) further link reinforcement learning and energy management by suggesting that the rewards structure could be based on performing certain actions when the conditions of the environment are favorable, such as when the cost of energy is low. This approach could help reduce costs as much as possible within the set boundary of the comfortable indoor temperatures.

Model-free approaches offer great flexibility when the entire model of the environment is not available, however, reinforcement learning-based methods at times can be unstable (Yu et al., 2020). This issue has been handled by combining reinforcement learning algorithms with deep neural networks, which resulted in greater efficacy of RL methods to an extent that they are now used in tasks involving computer vision and self-driving cars (Mason & Grijalva, 2019). Such a combination of reinforcement learning algorithms with deep neural networks is called deep reinforcement learning (DRL). In DRL, a deep learning model acts as a function approximator for the reinforcement learning agent (Zhang & Lam, 2018). Interestingly, deep reinforcement learning algorithms became more widespread after Mnih et al. (2013) successfully presented their ability to play the Atari video games at the human level (Zhang & Lam, 2018). In addition, due to its flexibility and efficiency over traditional model-based approaches, more research is going to be focused on reinforcement learning and its applications, as is in the case with autonomous building energy

management, where deep reinforcement learning algorithms are expected to keep growing (Mason & Grijalva, 2019). Thus, it can be concluded that deep reinforcement learning algorithms have a lot of potential when it comes to solving energy optimization problems of different kinds.

Furthermore, specific to this thesis's pursuits, the use of historical data was motivated by the research conducted by Natale et al. (2022), where the neural networks were used to create a simulation environment, in which deep reinforcement learning agents were trained to control the temperature of a building zone. The researchers' method depended exclusively on past historical data, which was used to fit physically consistent neural networks that then updated the reinforcement learning model's environment. This allowed them to avoid the complex design stage of physics-based methods while remaining physically consistent with respect to the control inputs (Natale et al., 2022). Similarly, this thesis uses neural networks to simulate the thermodynamics behavior of the dwelling, which is then used in a custom reinforcement learning environment.

2.3 Performance Studies Based on Deep Reinforcement Learning

In the context of autonomous building energy management, Q-learning has become one of the most widespread reinforcement learning algorithms because of its convenient properties of being off-policy and model-free (Mason & Grijalva, 2019). This allows Q-learning algorithms to find the best course of action given the current state independently of the agent's actions since this is an off-policy learning process. As Lissa et al. (2021) remark, in the real-world scenarios with scarce information about the environment, the Q-learning method can be useful because of its capability of making predictions incrementally. In Q-learning, the values of each state-action pair are expressed in a table called a Q-table, yet there is a scalability issue with this approach when the number of states and actions increases (Mason & Grijalva, 2019). In order to address such limitations, a common solution is to replace the Q-table with a function approximator, such as a neural network, as was described in the DRL method. The input to the neural network in this case is the state of the environment and the output is the Q-value for each action. Thus, the use of neural networks allows for scalability and efficiency when handling particularly large state spaces (Mason & Grijalva, 2019).

As various research papers suggest, one of the areas of building energy management, in which reinforcement learning has been successfully used is heating, ventilation, and air conditioning (HVAC) control. In their overview of recent literature on energy management focused on HVAC, Mason & Grijalva (2019) outline that the typical environment states for the reinforcement learning algorithms include factors like time of day, outdoor temperature, indoor temperature, weather forecast, and occupancy, and typical reinforcement learning actions include temperature set points, airflow control, and heating or cooling control. In addition, the rewards are typically computed based on energy cost, thermal comfort, or their combination (Mason & Grijalva, 2019).

One example of how Q-learning can be applied to the optimization of HVAC systems can be found in the study conducted by Chen et al. (2018), where a model-free Q-learning approach was utilized to make optimal control decisions for HVAC and window systems in order to minimize both energy consumption and thermal discomfort. After conducting the relevant case studies in Miami and Los Angeles, the authors noted a superior performance of reinforcement learning control. Their model achieved 13% and 23%

lower HVAC system energy consumption, and 62% and 80% lower discomfort degree hours compared to heuristic control, which required knowing the specifications of individual buildings. To further illustrate the implementation of deep reinforcement learning, it is necessary to highlight the 2017 study conducted by Wei et al., in which the authors utilized an artificial neural network to approximate the Q-value that estimates the control actions. Thus, the deep reinforcement learning algorithm applied to the data-driven HVAC control resulted in more effective cost reduction than the conventional Q-learning method.

In addition to HVAC control optimization problems, reinforcement learning can be applied to water heater optimization problems as well since they consume a significant amount of energy as well. As Mason & Grijalva (2019) point out, when it comes to the objective of reducing energy costs by controlling the water heater's usage, the typical state variables include the time of day, current water temperature, and forecasted usage. The action that the reinforcement learning agent makes is normally to turn the heater on or off, and the reward given to the agent is the electricity consumption (Mason & Grijalva, 2019).

In contrast to the previous model-free approaches examined, it is important to discuss the 2018 study done by Kazmi et al., in which a model-based reinforcement learning algorithm was used to optimize the energy efficiency of hot water production. Interestingly, in this particular scenario, a model-based approach was utilized since the model-free controllers would need more data to achieve the same performance level. Overall, the authors concluded that a model-based controller was able to reduce the energy consumption by almost 20% for a set of 32 Dutch houses with no loss of occupant comfort, which, if extrapolated to a year, has the potential to reduce household energy consumption by up to 200 kWh (Kazmi et al., 2018).

In addition to dealing with standalone problems of HVAC or water heater controls, reinforcement learning methods can be used to solve a complex set of problems related to the controls of the home energy management systems. Such systems often include multiple appliances, lighting, photovoltaics (PV), and batteries (Mason & Grijalva, 2019). Of course, this is a much more complex reinforcement learning problem, in which multiple elements need to be considered in order to reduce the overall energy consumption. The states for the reinforcement learning algorithm in this case generally consist of the time of day, temperature information, electricity prices, grid load, and the current usage state of the various appliances meanwhile the actions available are similar to previously discussed problems, which is turning a device or an appliance in the system on or off (Mason & Grijalva, 2019).

The use of deep reinforcement learning can be further illustrated by the study conducted by Lissa et al. (2021), in which a DRL algorithm was used for indoor and domestic hot water temperature control with the aim of reducing energy consumption by optimizing the usage of PV energy production. The results of the study demonstrated that the proposed deep reinforcement learning algorithm combined with the dynamic setpoint achieved on average 8% of energy savings compared to a rule-based algorithm, reaching up to 16% of savings over the summer period, without compromising the comfort temperatures (Lissa et al., 2021). Moreover, the authors of the study pointed out that the renewable energy consumption was 9.5% higher for the deep reinforcement learning model compared to the rule-based, which suggested that less energy was consumed from the grid.

To conclude, there are additional research studies that examine how the application of reinforcement learning within the individual dwelling can be beneficial for energy optimization and the load of the entire grid. Therefore, Mason & Grijalva (2019) suggest that based on the current studies, there is a high potential for utilizing reinforcement learning methods to significantly reduce the electricity costs for the grid.

3. DRL Research Design and Methodology

3.1 Application of DRL on Energy Cost Optimization

The overarching goal of this thesis is to explore how a deep reinforcement learning model can be implemented in order to control a heating system in a way that would optimize energy costs while maintaining comfortable indoor temperatures. Deviating from the more traditional and costly engineering techniques, which required a physical modeling of the building's thermodynamics, this study takes an approach that utilizes sensor-based data instead to supplement the reinforcement learning setup. The aim behind this methodology is to create a model that can be flexible enough to be scaled to other buildings without extensive data requirements and that is less costly to implement. Ideally, a model could be easily fit to existing buildings with simple sensors being the only requirement, compared to simulation-based methods that require the physical aspects of the building to be modeled.

To conclude, the research question for this study is as follows:

RQ: Can a deep reinforcement model lower the energy cost of a single household?

3.2 Research Methodology

3.2.1 Overview

In order to find a solution for the energy cost optimization problem specified and answer the research question, several components were required to build a custom deep reinforcement learning algorithm. First, a neural network was built based on the historical sensor data from the dwelling for the purpose of predicting the resulting indoor temperature based on indoor and outdoor temperatures and the current status of the heating system. Both feedforward and recurrent neural network types were used to build two separate DRL models.

Next, a custom reinforcement learning environment was built using the OpenAI Gym package.Within the custom environment, the states, actions, and rewards of the model were specified. Once the custom environment was built, the reinforcement learning model was ready to be trained.

Lastly, a Deep Q-Network algorithm was employed by combining the resulting Q-Learning reinforcement learning algorithm with a deep neural network, so that the optimal actions could be found. The steps involved in the custom deep reinforcement learning process are depicted by Algorithm 1.

Algorithm 1 Deep Reinforcement Learning with Indoor Temperature Simulation.

Build a neural network for indoor temperature prediction
Create a custom RL environment
Initialize the custom RL environment
    Repeat (for each episode)
    Initialize state s
    Initialize counter c
    repeat
    Choose action a from s using Boltzmann policy
    Take action a
    Predict indoor temperature using action a, state s, historic outside temperature[c]
    Update state (s')
    Calculate reward (r)
    until c is terminal
end

3.2.2 Data Collection and Selection

The dataset consisted of hourly observations taken between 00:00 of October 20, 2020, to 23:00 of April 30, 2022, for a total of 13,392 observations (558 days) over the period of 19 months. The dataset was provided by Sensative AB, a company located in Lund, Sweden. Sensative AB primarily produces sensor strips for consumer and commercial use, and also offers a Digitalization Infrastructure Management System (DiMS) that supports the organization of smart home devices. The sensors were placed in several locations around a specific residential home located in Lund. Since these sensors have been in place for several years, the authors of this study were not involved in the collection of the data. Similarly, the date range of the sensor data was limited to the amount of time the sensors had physically been attached to the building. However, the selection of variables provided was agreed upon at the commencement of the thesis.

To prepare the dataset to be used, several data wrangling tasks were made and some variables were removed. Table 2 details the final variables within the dataset.

Variable	**Description**
Date & Time	Date and time of each hourly observation
Outdoor temperature from the weather service	The hourly outdoor temperature in °C from the house area provided by the SMHI
Heating switch status	Binary variable indicating whether the switch was on or off for the majority of the hour (>=0.5 is on)
Average indoor temperature	The hourly average of all indoor temperatures in °C
Electricity price	The hourly price of electricity in Euro/MWh

Table 2. Variables within the final version of the dataset used.

For the indoor temperatures, sensors were placed in seven locations throughout the home, and have thus gathered various temperatures. Figure 4 shows boxplots of the temperatures in the different rooms. For this study, the average indoor temperature was taken across all rooms each hour to produce a single indoor temperature reading.

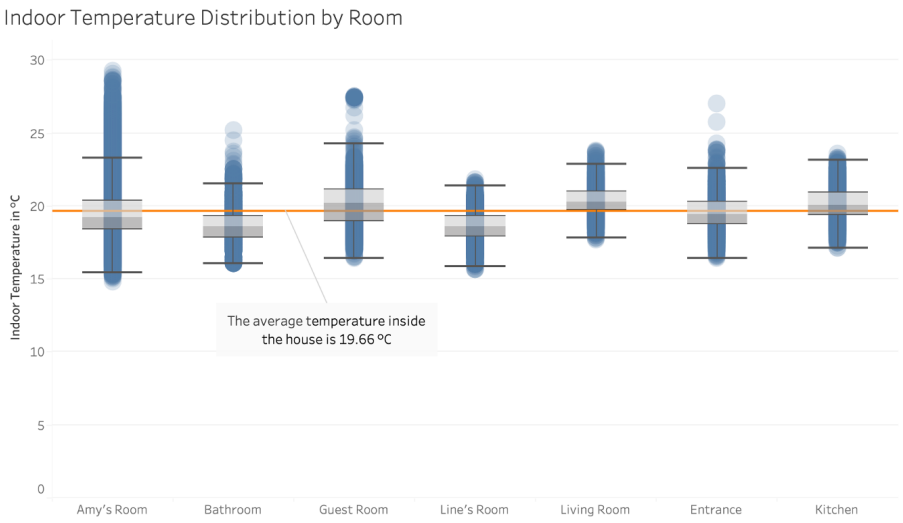

Figure 4. Boxplots of the temperature readings in each location as gathered by sensors.

The outdoor temperatures were gathered by a single sensor placed outside of the dwelling. As can be seen in Figure 5, compared to the outdoor temperatures provided by the SMHI for the same time frames, the sensor data values were, on average, lower by 0.17°C, meaning that the reading is very accurate. However, due to the limited availability of the outdoor sensor data, which started in October 2021 instead of October 2020, the outdoor weather data from the SMHI was used instead and the outdoor sensor data was removed.

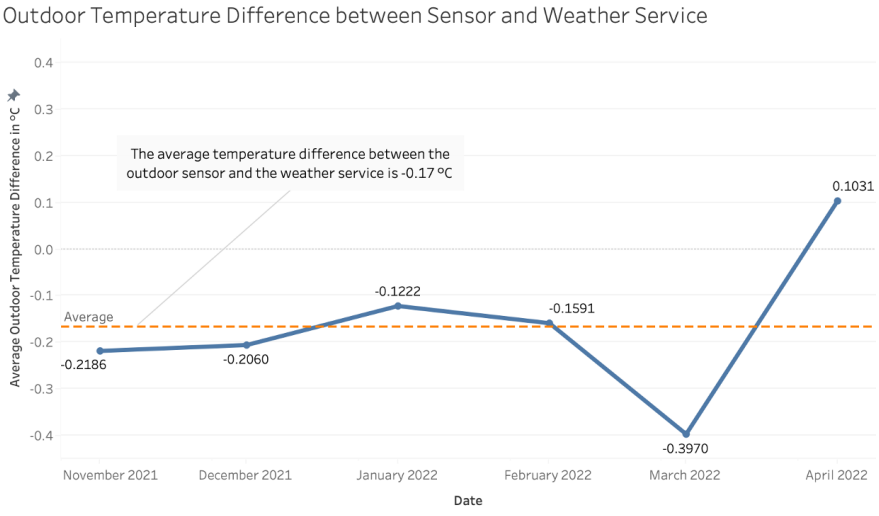

Figure 5. Sensor vs Weather Service Measurements of Outdoor Temperature.

To further understand the outdoor temperature fluctuations, Figure 6 below shows the average temperature per month from October 2020 to April 2022.

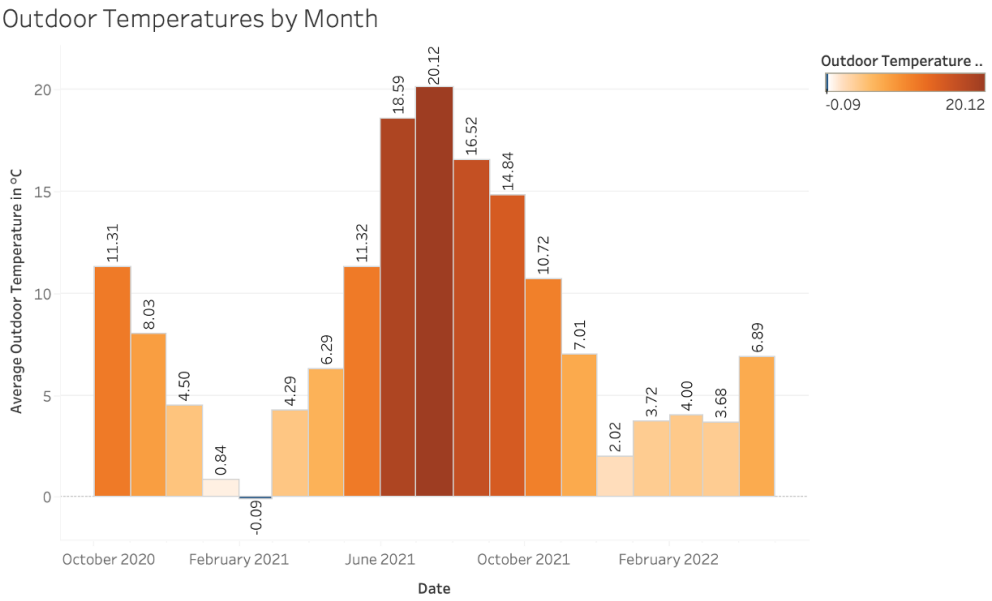

Figure 6. Average outdoor temperature each month from October 2020 to April 2022.

In addition, the electricity prices were gathered in order to be able to compare the actual cost incurred by the heating system throughout the chosen date range with the proposed cost given by the output from the models. As seen in Figure 7, the electricity price has continued its upward trend from Figure 1.

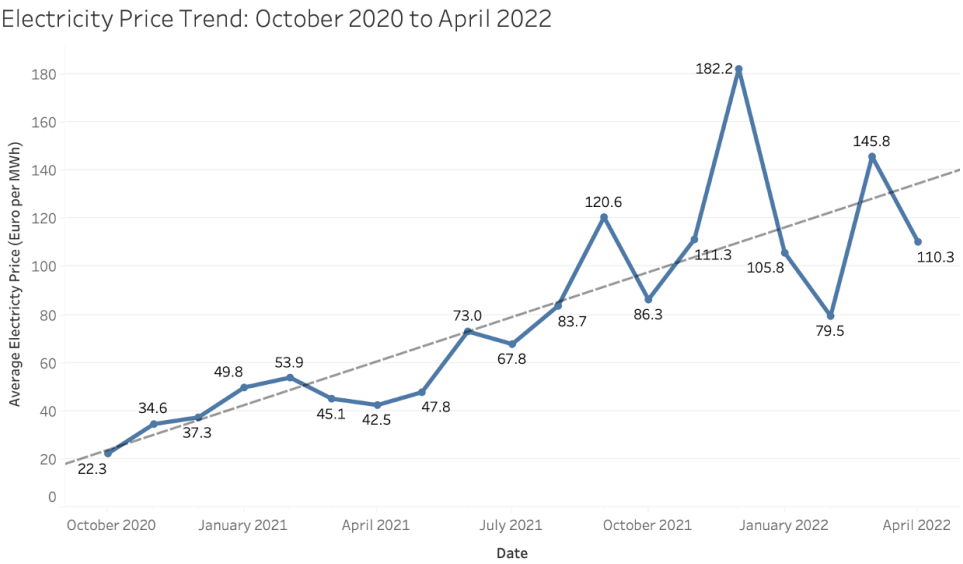

Figure 7. Average price of electricity (in Euro per MWh) per month.

The manual switch controls of the house originally indicated the proportion of the hour wherein the heating system was turned on. To simplify the input from decimal to binary, the values have been transformed to 1 if the heat was on for half an hour or more, and 0 if less than half an hour. The mean action would then indicate the average proportion of time that the heating system is on during a given timeframe. For example, if the mean action at 08:00 is 0.7, it would indicate that the heating system was turned on 70% of the hour, or alternatively that out of all the 08:00 values 70% of them showed the heating system being on. In Figure 8 the mean action across all hours of the day is shown in conjunction with the average electricity price throughout the day. It is clear that the electricity prices are highest when usage is also high, shown by the mean action being over 70% at around 08:00 and 17:00.

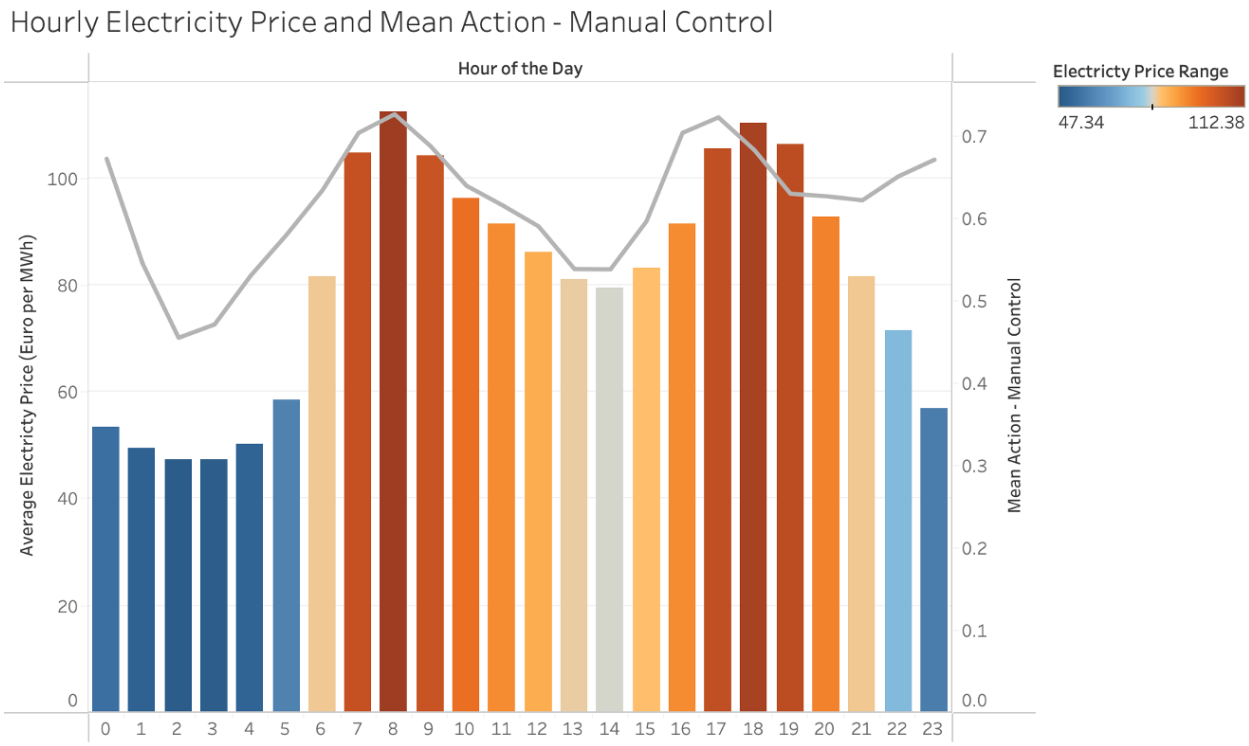

Figure 8. Electricity price and mean action per hour.

From the exploration of the available dataset, it can be concluded that considering a strong trend in the increase of electricity prices and the fact that the bid prices are the highest during the peak hours, there is a need for a method that would help reduce the cost of electricity in the household.

3.2.3 Neural Networks for Indoor Temperature Estimation

In this thesis, the ideas for the use of historical sensor data were greatly inspired by the study done by Natale et al. (2022) since it used historical data to train physically consistent neural networks used in simulation environments in order to assess the performance of DRL agents for zone temperature control. Thus, the feed-forward and recurrent neural networks in this study are also trained on the historical data collected using sensors. By learning from patterns in the sensor data, the network simulates the indoor

temperature behavior of the dwelling during different seasons, thus avoiding the manual modeling of the physical properties of the building. This approach was taken to circumvent the need for live implementation of the reinforcement learning model directly in the dwelling in question. Had live temperature readings been available to the RL model, then this neural network component for temperature estimation would not have been necessary.

After the neural network is trained on historical data and has "learned" the typical behavior of the house's temperature changes, it is used in the reinforcement learning environment for the purpose of making a prediction of the indoor temperature based on the environment state (indoor temperature and outdoor temperature metrics) and the action taken by the agent. Thus, the neural network helps to simulate a real-time behavioral reaction of the dwelling based on the environment conditions and the action, which the agent decided to take.

For this study, both a feed-forward network and a recurrent neural network were used for the temperature prediction within the custom RL environment. Specifically, two different RL models were built: one RL model was built with a feed-forward neural network, and the other RL model was built using a recurrent neural network. These two RL models and two neural networks use the same dataset, but their results are independent of each other.

The feed-forward neural network was trained on four inputs: switch status, outdoor temperature, indoor temperature from the past 1 hour, and the indicator variable for the month (e.g. September is 9). The structure of the network was relatively simple as it contains an input layer, one hidden layer, and an output layer. For the hidden layer, the Rectified Linear Unit (ReLU) was chosen as the activation function, followed by the output layer with one output: estimated indoor temperature. The summary of the structure of the network built using TensorFlow and Keras packages and its parameters can be found in the Figure 9 below:

```
Model: "sequential"
_________________________________________________________________
 Layer (type)                Output Shape              Param #
=================================================================
 dense (Dense)               (None, 8)                 40

 dense_1 (Dense)             (None, 12)                108

 dense_2 (Dense)             (None, 1)                 13

=================================================================
Total params: 161
Trainable params: 161
Non-trainable params: 0
_________________________________________________________________
```

Figure 9. Summary of the feed-forward neural network used for indoor temperature estimation.

To account for the heating season and achieve a better physical consistency in predictions, the model was trained on the subset of data covering September 2021 to April 2022 with a total of 5,808 observations. After 200 epochs, the test mean absolute error achieved was 0.16 (measured in °C).

Since the nature of the dataset is sequential, a recurrent neural network was also built to model a more accurate thermodynamics behavior of the building. Since temperature changes often trend in a steady direction as the days warm the spaces and the nights cool, this aspect needs to be considered for predicting the indoor temperature required for the custom environment. The network consisted of six layers presented in Figure 10 below. The model uses Long Short-Term Memory Network (LSTM), which is a variation of a recurrent neural network useful in predicting the long sequences of data.

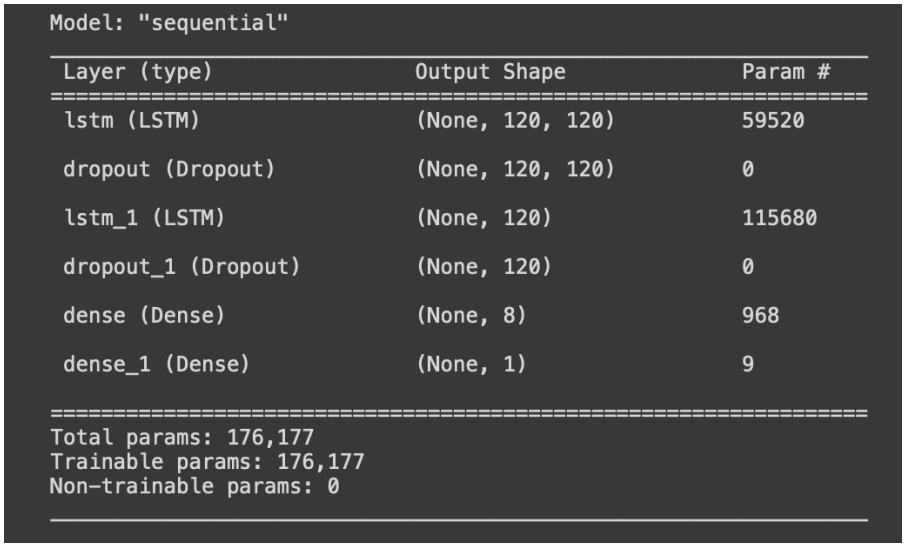

Figure 10. Summary of the recurrent neural network used for indoor temperature estimation.

Since the LSTM was used, the model was trained on the full dataset covering October 20, 2020 to April 30, 2022 with a total of 13,392 observations. The network had a 5-day lookback period (120 observations). After 50 epochs, the test mean absolute error achieved was 0.17 (measured in °C).

Therefore, in order to avoid complex and costly physics-based modeling, such as creating a thermodynamic model of the building, this study will use both feed-forward and recurrent neural networks to estimate the indoor temperature based on historical sensor data collected. Such an approach will not only aid in extending the model to other types of dwellings and buildings, but also will speed up the implementation process if the data collection sensors have been previously installed.

3.2.4 Custom Reinforcement Learning Environments

Environment and Agent

Reinforcement learning setup requires an environment and an agent which interact with one another. The agent, for whom the environment is its "home world", is able to interact with the environment by completing some actions, however, its actions cannot influence the rules or dynamics governing the environment (Metelli, 2022). The agent receives information about the current state of the environment and performs an action. This action makes the environment transition to a new state. In addition, the

environment also sends a reward signal to the agent, which serves as feedback whether the action taken was beneficial or not (Metelli, 2022).

To further exemplify, a widely used environment for developing AI agents for Atari 2600 games such as Pong (see Figure 11), Breakout, SpaceInvaders, Seaquest, and Beam Rider is the Arcade Learning Environment known as ALE (Mnih et al., 2013). For instance, this environment was used by Mnih et al. back in 2013 when they presented the first deep learning model to successfully learn control policies directly from high-dimensional sensory input using reinforcement learning.

With an increased popularity of reinforcement learning methods, there was a need for a toolkit for creating and comparing reinforcement learning algorithms. OpenAI Gym can be considered such a toolkit with its open-source interface for reinforcement learning tasks (OpenAI, 2022). OpenAI Gym not only offers an already predefined suite of environments, like Atari, which has been integrated with the Arcade Learning Environment, but also allows for customization of the environments. This flexibility within an interface is helpful with benchmarking and standardization, and makes it easier for researchers and developers to reproduce results (OpenAI, 2022).

Since there is no predefined suite of environments available for addressing the problem of this thesis, a custom reinforcement learning environment was built using the OpenAI Gym package in Python. In addition, Stable Baselines, a set of improved implementations of Reinforcement Learning algorithms based on OpenAI Baselines, were used to take advantage of common code style and to simplify the code. An agent in this case is single and artificial, representing the model which is being designed and, thus, replacing the concept of the person who would manually control the state of the heating system by turning it on or off.

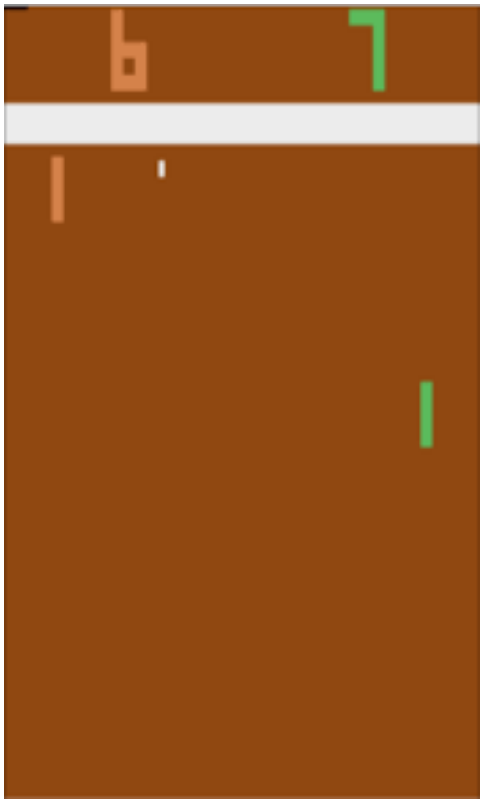

Figure 11. Example of the interface for the Pong game.

States and Actions

Every scenario an agent faces in the environment is a state. For instance, in the research study conducted by Lissa et al. (2021) on deep reinforcement learning for home energy management system control, the state-space *s* consisted of outdoor and indoor temperature (°C), domestic hot water tank temperature (°C), photovoltaic production (kW), and hour of the day. However, in this thesis the state *s* setup is similar to the one described in the research conducted by Natale et al. (2022), where the state *s* observed by the agents at each time step was composed of the input features of the physically consistent neural networks, such as zone temperatures (of the controlled and neighboring room), ambient conditions (temperature and solar irradiation), and time information. Thus, in this environment, the state *s* observed by the agent is comprised of the input features of the neural networks discussed in Section 3.2.3.

Next, it is necessary to specify the actions that can be taken by the agent. These actions are the agent's methods, which allow it to interact and alter the environment, and, therefore, transfer between the states. The decision of which action to take is made by the policy π, which can be viewed as a prescription or a strategy indicating which action a to take in every state s (Metelli, 2022). In this study, the action space is discrete, meaning the agent can either turn on or turn off the heating system.

Episodes and Rewards

An episode comprises all states between the first and the last state of the environment, and the agent's goal is to maximize the total reward it obtains during an episode. Furthermore, each episode can be considered as a separate "round of the game" for the agent "to play". The selected episode duration in various research studies depends on the simulation setup and the heating system. For example, in the research study conducted by Lissa et al. (2021), the episode duration is 8 months (from May to December), after which the episode is finished and the environment is restarted (2020). However, in their research setup of the radiant heating system, Zhang & Lam (2018) made the duration of one training episode three months (from January to March) in order to account for the heating season. Thus, based on the availability of historical data and to account for the heating season, in this study the selected episode length is from October 1st, 2021 to April 30th, 2022.

The final step in configuring this custom environment is to create an appropriate reward function. A reward is a numerical value received by the agent from the environment as a direct response to the agent's actions. As was mentioned earlier, the maximization of the total reward is the ultimate goal of the agent during the episode. The rewards can be designed to suit different purposes. For example, a reward can be positive if the agent's action has been helpful in achieving the goal, negative when a certain situation needs to be deterred, or zero, possibly indicating a status-quo (Fuchida, Aung & Sakuragi, 2010). The reward function for this study has been inspired by research conducted by Natale et al. (2022) and Wei et al. (2017), since both took into account the total penalty of temperature violations. In addition, the reward function designed by Wei et al. (2017) considers the energy cost of the control action. Both studies use negative rewards since the deep reinforcement learning algorithm will try to maximize the total reward.

Therefore, as shown in the reward equation below, the deep reinforcement learning algorithm in this study will attempt to obtain as high of a reward as possible (closest to 0), thus balancing the objective of minimizing energy cost with the goal of maintaining the temperature within the desired comfort range (18°C to 22°C). Since the electricity price is given in Euro/MWh and the hourly usage is measured in kWh, the price is scaled down by 1000.

$$Reward = -0.001(electricity\ cost) - max\{T_L - T, 0\} - max\{T - T_U, 0\}$$

$$T_L = Lower\ temperature\ bound$$

$$T_U = Upper\ temperature\ bound$$

3.2.5 Deep Reinforcement Learning Setup

The final aspect of this custom deep reinforcement learning model is to "upgrade" from the ordinary reinforcement learning process to deep reinforcement learning process. As was previously mentioned, it was Mnih et al. from DeepMind Technologies who first introduced the use of deep neural networks in Q-learning back in 2013.

Q-learning is a model-free and an off-policy reinforcement learning method, which creates a matrix called Q-table with Q-values, which are an estimation of how beneficial it is to take a particular action at a particular state (Mason & Grijalva, 2019). An agent can "refer to" this Q-table (see Figure 12) to maximize its reward in the long run. This approach is only practical for smaller environments with action and states spaces both being discrete, and can easily become infeasible as the number of states and actions increases (Lissa et al., 2021).

State \ Action	0	1	2	3	4
1	0	0	-1	0	
2	0	0	100	-1	
3	-1	0	-1	100	
4	0	100	-1	0	

Figure 12. Sample Q-table containing Q-values, in which the maximum expected future reward for each action at each state is displayed.

Combining reinforcement learning with an artificial neural network to replace the Q-table is known as Deep Reinforcement Learning or Deep Q Network (DQN) (Mason & Grijalva, 2019). Since artificial neural networks are universal function approximators, one of the possible solutions to the dimensionality curse is to replace the Q-table by using an artificial neural network to estimate the Q-values (Lissa et al., 2021). As can be seen in Figure 13, this approach can manage larger state-action spaces, thus bringing more possibilities when working with a larger number of variables and observations.

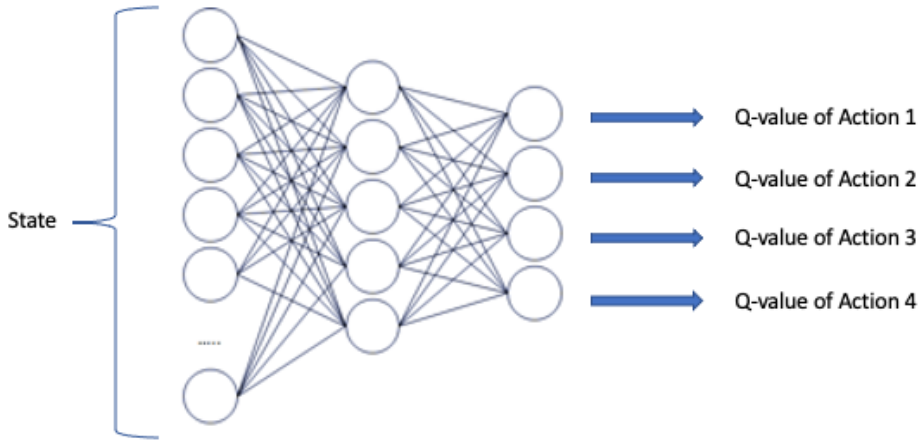

Figure 13. Deep Q-Learning Approach

In this study, the action space used is discrete and consists of only two actions (on or off), however, the state space is continuous. Thus, these conditions require the use of the Deep Q Network (DQN) approach.

The two deep reinforcement learning models - one using feed-forward neural network and the other using recurrent neural network for indoor temperature estimation - are created using Keras RL package. However, in order to be able to compare both methods, the structure of the neural network used in the DQN is the same and is displayed in Figure 14 below. The activation function is Rectified Linear Unit (ReLU), and the output layer contains two outputs representing all possible actions.

```
Model: "sequential_1"
_________________________________________________________________
 Layer (type)                Output Shape              Param #
=================================================================
 flatten (Flatten)           (None, 1)                 0

 dense_3 (Dense)             (None, 8)                 16

 dense_4 (Dense)             (None, 16)                144

 dense_5 (Dense)             (None, 2)                 34

=================================================================
Total params: 194
Trainable params: 194
Non-trainable params: 0
_________________________________________________________________
```

Figure 14. Deep Q Network Used.

Next, using Keras RL package guidelines, a DQN agent was created. Instead of using a ϵ-*greedy* policy as was suggested by Lissa et al. (2021) to address the exploration and exploitation dilemma, the Boltzmann policy was used. Reinforcement learning process is plagued by the exploration and exploitation dilemma because the agent needs to balance between exploring the environment further in order to find new ways to gain rewards and exploiting previous knowledge by adhering to paths that have historically led to rewards (Beyer et al., 2019). Compared to the popular ϵ-*greedy* approach, which combines a greedy method with occasional random decisions with probability ϵ, the Boltzmann policy was used because it

involves selecting an action with weighted probabilities based on expected rewards (Masadeh, Wang & Kamal, 2018).

In addition, the sequential memory, which is used by the DQN agent to store information about the environment's states, actions, and rewards, has been specified to store the agent's experiences for the length of 3 episodes. Both DRL models are trained on the subset of the data representing the heating season, which is October 1st, 2021 to April 30th, 2022. Overall, each model was trained for 6 episodes.

3.3 Source Critical Consideration

The data collected for this study was provided by Sensative AB, a private limited company located in Sweden. Specifically, the indoor temperatures and heating switch actions were measured by sensors built by Sensative, the outdoor temperatures were collected from the Swedish Meteorological and Hydrological Institute (SMHI), and the electricity prices were gathered from Nord Pool AS. The latter sources can be considered highly trustworthy and accurate. As the sensor data were not recorded with academic stringency in mind, there may be erroneous values in recording, reporting, or collating. Lastly, the authors declare that the results of this study have not been influenced by the collaboration with Sensative AB.

4. Empirical Results from DRL Models

The aim of this study was to address to main research question:

RQ: Can a deep reinforcement model lower the energy cost of a single household?

In order to investigate this, two separate DRL models were built with the aim of reducing energy costs. The first model used a feed-forward neural network (FFNN) to make the temperature predictions used in the custom environment, and the second RL model used a recurrent neural network (RNN) to make the same prediction. In order to specifically isolate the ability of the RL agent to affect indoor temperatures through heating (i.e. not through cooling, since the system in the house is not capable of that), the DRL models were run on an episode only from September 2021 to April 2022. Overall, each model was trained for 6 episodes.

Additionally, to account for the recurrent neural network needing a designated period of time to be trained, leading to actions starting mid-month, the final actions and energy cost outputs considered were taken from October 2021 to April 2022 across all three scenarios: the historical information about manual control of the heating, DRL using feed-forward neural network for indoor temperature predictions, and DRL using recurrent neural network for indoor temperature predictions.

The results will be evaluated using the following four aspects:

1. Total cumulative reward achieved during each episode
2. Mean action of the DRL agent during each episode
3. Indoor temperature control
4. Overall energy cost-savings

Overall, it was shown that the DRL agent was able to lower not only energy costs but also energy consumption, while maintaining the appropriate indoor temperatures.

4.1 Reward Analysis

Since the agent has an objective to maximize the total reward accumulated per episode, considering the specified reward function which accounts for both electricity price and comfort temperature violations, it would be preferred to have a total episode reward as close to 0 as possible. The closer the reward is to 0, the lower would be the overall energy cost of the episode.

$$Reward = -0.001(electricity\ cost) - max\{T_L - T, 0\} - max\{T - T_U, 0\}$$

$$T_L = Lower\ temperature\ bound$$

$$T_U = Upper\ temperature\ bound$$

As can be seen in the Table 3 below, each FNN- and RNN-based DRL model had 6 training episodes with 5,807 and 5,688 steps respectively to account for the heating season. It can be seen that both algorithms managed to achieve a quite similar reward. However, the DRL algorithm which used RNN performed relatively better by achieving an average reward of -619.39.

Episode Summary Table

	Network Type					
	Epsiode Steps		Episode Reward		Mean Action	
Episode	FFNN	RNN	FFNN	RNN	FFNN	RNN
1	5,807	5,688	-635.61	-619.36	38.40%	37.60%
2	5,807	5,688	-632.16	-619.55	50.00%	51.40%
3	5,807	5,688	-634.80	-619.36	51.50%	52.20%
4	5,807	5,688	-632.16	-619.36	50.40%	51.20%
5	5,807	5,688	-632.16	-619.36	49.60%	50.80%
6	5,807	5,688	-632.26	-619.36	51.00%	51.70%
Average	**5,807**	**5,688**	**-633.19**	**-619.39**	**48.48%**	**49.15%**

Table 3. Summary of steps, rewards, and mean action for every episode for both DRL models with FFNN and RNN temperature prediction models.

In addition, the DRL algorithm which used FFNN had a larger variability of cumulative rewards per episode. Its standard deviation over 6 episodes was 1.58 whereas the standard deviation of the reward when RNN was used was 0.08, indicating that the DRL models using RNN were able to find an optimal combination of actions faster and more reliably.

4.2 Action Analysis

The action space in the current problem setting is discrete, with 0 indicating that the heating should be turned off and 1 indicating that the heating is turned on. Looking at the sum of all actions gives the total hours that the heating system is on over the given episode, and in this way indicates what the total energy consumption would be. Table 4 shows the total hours of heat consumption for both models across all episodes. On average, the DRL agents were able to decrease the total energy consumption by 19.43%.

Total Hours of Heating Per Episode

Model Type	Episode	Total Hours of Heating	Percentage Decrease from Manual Control
Manual	Manual Control	3,225	
FFNN	FFNN Episode 1	2,230	30.85%
	FFNN Episode 2	2,524	21.73%
	FFNN Episode 3	2,650	17.83%
	FFNN Episode 4	2,594	19.56%
	FFNN Episode 5	2,513	22.08%
	FFNN Episode 6	2,593	19.59%
RNN	RNN Episode 1	2,139	33.67%
	RNN Episode 2	2,607	19.16%
	RNN Episode 3	2,677	16.99%
	RNN Episode 4	2,612	19.01%
	RNN Episode 5	2,581	19.97%
	RNN Episode 6	2,632	18.39%

Table 4. Total hours of heating used across each episode and the percent decrease from the manually controlled hours.

By taking an average of the actions during an entire episode, the amount of time the heating was on during an episode can be seen, which shows the proportion of energy used. For instance, if the mean action of an episode was 0.5, it would indicate that the heating was on for exactly half the time across the entire episode. Although the reward function for the DRL was not tuned for optimizing the mean action (i.e. reducing usage as much as possible), it is still interesting to see that the models were able to reduce energy consumption overall compared to the manual control.

For readability, the values have been transformed to percentage values indicating the percentage of time that the heating system was on. A summary of the mean action for the last episode and an overall average are presented in Table 5. Overall, it can be observed that the DRL model using RNN had a higher overall mean action compared to the model using FFNN. Both algorithms start with a relatively small value of mean action (around 38%) and then reach the value of around 51% in the last training episode. Since both algorithms have similar settings that balance exploration and exploitation, the changes in mean actions are quite similar.

In order to study the research question from different aspects, it is interesting to look at the difference between actions taken by a person versus the DRL algorithms. For the purpose of comparison, only the actions of the last training episode are taken.

Mean Action Monthly Comparison

Month of Date & Time	Mean Action - Manual	Mean Action - FFNN - Episode 6	Mean Action - RNN - Episode 6
October 2021	58.71%	53.09%	52.02%
November 2021	55.67%	50.83%	53.89%
December 2021	73.47%	51.61%	50.27%
January 2022	70.85%	49.73%	51.61%
February 2022	58.97%	52.08%	51.34%
March 2022	62.53%	48.92%	52.28%
April 2022	62.78%	50.56%	50.69%
Average	**63.38%**	**50.96%**	**51.73%**

Table 5. Mean action per month for episode 6.

As can be seen in Table 5, when it comes to the DRL algorithm using FFNN, the percentage of time per month that the heating system was on was an average of 50.96% instead of 63.38% when it was controlled manually. Similarly, the DRL method using RNN had an average of 51.73%. In a simulated environment of the dwelling across all heating months both DRL methods were able to keep the switch time relatively stable around the same value compared to the human action.

4.3 Indoor Temperature Analysis

Keeping the indoor temperature within a band of 18°C to 22°C was one of the main goals of the DRL agent. The results show that the final DRL model was able to remain in this temperature range throughout every episode. Across the six episodes, the average indoor temperature for the DRL models with FFNN was 19.98°C, and the average temperature for the DRL models with RNN was 19.61°C. Figure 15 shows

a box plot of indoor temperatures resulting from the manual control and the different DRL agents for one of the episodes. It is apparent that the range of temperatures is much smaller for the DRL models.

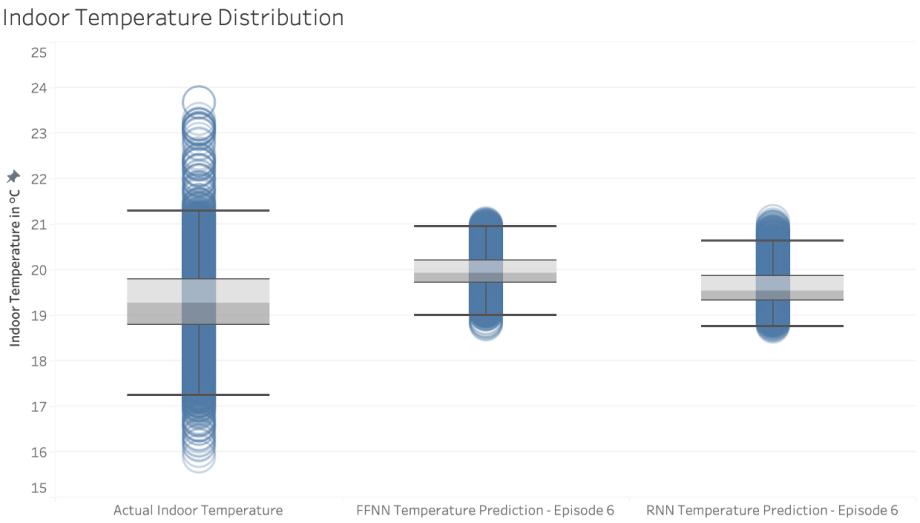

Figure 15. Box plot showing the indoor temperature distribution of the manual control, FFNN DRL model, and RNN DRL model.

The indoor temperatures of the DRL models' 6th episode can also be seen compared to the manual controlled temperatures in Figure 16 and Figure 17. In each graph, the temperature trend can be seen to decrease slightly until January before it rises again towards May, reflecting the seasonal changes of winter and spring. Using the DRL agent, the indoor temperature is maintained at a steadier rate compared to the fluctuations from manually controlling the heating system. This could be interpreted to mean that the DRL model was more successful in maintaining an even temperature compared to manually turning the heat on and off. Contrastingly, this could also be due to the neural networks' predictions of indoor temperature not being as erratic as the actual temperature measurements. Nonetheless, it is clear that the DRL model is successful in keeping the temperatures within the desired temperature band.

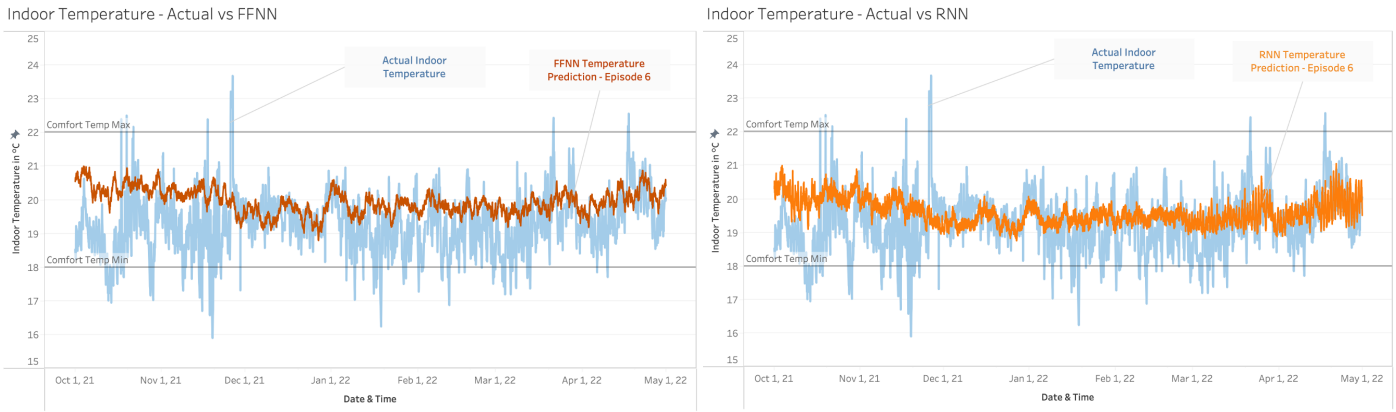

Figure 16 and 17. Model temperatures for Episode 6 plotted against the actual temperatures.

4.4 Cost Analysis

The main aim of this study was to see whether the DRL agent could decrease electricity costs. Across all episodes, it was found that the agents were able to achieve an average cost savings of 27.58% with the FFNN DRL model, and 26.60% with the DRL model using RNN compared to the manual actions of the homeowner. For every month, a reduction in cost was successfully achieved, which means that the DRL agent was able to maintain the indoor heat while minimizing cost. Table 6 shows the total costs per month in episode 6, as well as the percent decrease that was achieved between the DRL models and the manual control. Across the models, the lowest monthly percentage decrease was 2.83% and the highest achieved monthly decrease was 32.00%. The DLR model using FFNN was able to lower costs more than the RNN DRL model. This could be due to the differences in how well the NN were able to accurately update resulting temperatures. Nonetheless, the cost savings were apparent across both models for every episode.

Monthly Cost Comparison

Month of Date & Time	Cost of Manual Control	Cost of DRL with FFNN	Cost of DRL with RNN	Percent Difference between FFNN and Manual	Percent Difference between RNN and Manual
October 2021	€ 400.03	€ 293.63	€ 292.38	-26.60%	-26.91%
November 2021	€ 306.71	€ 287.61	€ 298.04	-6.23%	-2.83%
December 2021	€ 1,174.15	€ 840.29	€ 840.00	-28.43%	-28.46%
January 2022	€ 706.36	€ 480.32	€ 492.53	-32.00%	-30.27%
February 2022	€ 439.25	€ 325.85	€ 335.46	-25.82%	-23.63%
March 2022	€ 874.09	€ 636.88	€ 686.75	-27.14%	-21.43%
April 2022	€ 629.02	€ 499.04	€ 481.47	-20.66%	-23.46%
Grand Total	**€ 4,529.61**	**€ 3,363.62**	**€ 3,426.63**	**-25.74%**	**-24.35%**

Table 6. The monthly heating costs for each model in episode 6.

5. Discussion and Critical Reflection

In order to critically assess the results achieved in this thesis, first, it is important to benchmark them against the other research projects mentioned in the Literature Review section. This thesis was able to show that deep reinforcement learning, combined with sensor data used for indoor temperature simulation, is a promising method of achieving energy cost savings through heating system control. The overall average cost savings of 25% and energy savings of 19% is in line with many previous studies mentioned, such as Kazmi et al. (2018) with 20% energy savings, Lissa et al. (2021) with a high of 16% savings, and Chen et al. (2018) with between 13% and 23% savings. Similar to Natale et al. (2022) who also used historical data paired with a neural network in order to conduct the custom environment, the shared positive results are a boon to this specific methodology.

Moreover, it is necessary to highlight the consequences of employing feed-forward and recurrent neural networks for indoor temperature estimation. It is essential to note that compared to Natale et al. (2022) who used a specifically tuned physically-consistent neural network for their indoor temperature predictions, the FFNN and RNN used in this study are not as accurate, and could have potentially caused the results to be more on the optimistic side. Nevertheless, the results from the DRL agent still indicate that deep reinforcement learning is a viable method for energy cost and consumption optimization. Additionally, the usage of historical sensor data introduces a new level of flexibility when it comes to the model-free methodology.

Of course, the study conducted in this thesis could be taken further to the real test environment of the actual dwelling in Lund, Sweden. The real-world use of a trained DRL agent such as the one in this study could be altered with minimum modifications to accept a live reading of the resulting temperature at a set time interval and make a new decision based on that. This approach, in turn, would avoid the necessity of predicting the resulting temperature and allow the DRL agent to directly interact with the real environment.

Furthermore, the results of this study need to be considered in conjunction with the reward function that was constructed. In this particular scenario with a single dwelling, the reward function was designed specifically to minimize the cost of energy while respecting the desired indoor comfort temperatures. However, the objective of the reinforcement learning algorithm can be potentially made more complex by adding other parameters like energy consumption, the penalty for energy use, or additional scaling. The reward function also adds flexibility in that it would be easy to tweak parameters such as target temperature to suit the needs of each individual household.

Similarly, the states of the environment in this study were motivated by the availability of variables in the dataset provided by Sensative AB, and were narrowed down to outdoor temperature, indoor temperature, and the status of the heating system. However, the model can be enriched by adding supplementary parameters like heating system readings; forecasted energy usage; other weather parameters, like precipitation and wind speed; other sensor readings, like open window indicator or dwelling occupancy; and grid load. Additionally, the control actions that were used in the models described in this thesis were changed from decimal values to binary to simplify the model structure and ease the computational load,

which led to the action space being discrete. However, changing the action space to continuous, by making the control action represent the fraction of the hour that the system has to be on, could potentially lead to better accuracy of the results and additional cost and energy savings.

Finally, the last item to remark on is the chosen model-free approach. In this thesis, the models were built using the Deep Q-Network method since the action space was discrete and the state space was continuous. However, other model-free reinforcement learning algorithms can be considered, such as Deep Deterministic Policy Gradient (DDPG), Asynchronous Advantage Actor-Critic (A3C), or Proximal Policy Optimization (PPO). To conclude, as in any machine learning problem, there are multiple model aspects to be considered and various parameters to be tuned.

6. Conclusion

Reinforcement learning has historically often been used in gaming and other virtual environments. Here, we have shown that it can also be used on real-world data, and have successfully answered the question as to whether a DRL agent can learn to lower energy costs. The departure from regular machine learning in needing large amounts of data is provided by the flexibility of the custom environment in reinforcement learning. This provides a rigorous base for the RL model to test out decisions and find the best actions.

On the other hand, the dependence of reinforcement learning models on the custom environment introduces some challenges. The custom environments built in this study rely on neural networks to update the current state of the environment. Since this neural network contains only a few variables, there is a possibility that the environment is not as robust as it could be. In further research, improving the performance of this neural network could lead to a better RL model. Alternatively, the neural network could be forgone by allowing the custom environment to be updated directly by live sensor readings. In this way, the RL model would have immediate access to the results of its actions, making for a more robust model.

This study has explored the possibilities for reinforcement learning to be used in conjunction with sensor data in a way that would ideally be more scalable and widely applicable to existing buildings. By adopting the model, not only would the household energy costs go down, but the overall energy consumption could also be reduced. If a similar result could be replicated across the board, a cost reduction of 25% and usage reduction of 19% would make for significant savings on a macro level. Although lofty, our research shows that the methodology holds promise for energy optimization problems.

References

Barber, K. A. & Krarti, M. (2022). A Review of Optimization Based Tools for Design and Control of Building Energy Systems, *Renewable and Sustainable Energy Reviews*, vol. 160

Beyer, L., Vincent, D., Teboul, O., Gelly, S., Geist, M., & Pietquin, O. (2019). MULEX: Disentangling Exploitation from Exploration in Deep RL, ArXiv e-prints

Bünning, F., Huber, B., Heer, P., Aboudonia, A., Lygeros, J. (2020). Experimental Demonstration of Data Predictive Control for Energy Optimization and Thermal Comfort in Buildings, *Energy and Buildings*, vol. 211

Chen,Y., Norford, L.K., Samuelson, H.W., & Malkawi, A. (2018). Optimal Control of HVAC and Window Systems for Natural Ventilation through Reinforcement Learning, *Energy and Buildings*, vol. 169, pp. 195-205

Eisenhower, B., O'Neill, Z., Narayanan, S., Fonoberov, V. A., & Mezić, I. (2012). A Methodology for Meta-model Based Optimization in Building Energy Models, *Energy and Buildings*, vol. 47, pp. 292-301

Energimyndigheten. (2021). Energy in Sweden 2021 - An Overview, Available online: https://energimyndigheten.a-w2m.se/Home.mvc?ResourceId=198022 [Accessed 11 April 2022]

Fuchida, T., Aung, K.T. & Sakuragi, A. (2010). A Study of Q-learning Considering Negative Rewards, *Artificial Life and Robotics,* vol. 15, pp. 351–354

Kazmi, H., Mehmood, F., Lodeweyckx, S., & Driesen, J. (2018). Gigawatt-hour Scale Savings on a Budget of Zero: Deep Reinforcement Learning Based Optimal Control of Hot Water Systems, *Energy*, vol. 144, pp. 159-168

Lissa, P., Deane, C., Schukat, M., Seri F., Keane, M., & Barrett, E. (2021). Deep Reinforcement Learning for Home Energy Management System Control, *Energy and AI*, vol. 3

Löf, G.O.G. & Tybout, R.A. (1974). The Design and Cost of Optimized Systems for Residential Heating and Cooling by Solar Energy, *Solar Energy*, Volume 16, Issue 1

Masadeh, A., Wang, Z., & Kamal, A. (2018). Convergence-Based Exploration Algorithm for Reinforcement Learning, *Electrical and Computer Engineering Technical Reports and White Papers,* vol. 1, Available online: https://core.ac.uk/download/pdf/212815835.pdf [Accessed 21 April 2022]

Mason, K. & Grijalva, S. (2019). A Review of Reinforcement Learning for Autonomous Building Energy Management, *Computers & Electrical Engineering*, vol. 78, pp. 300-312

Metelli, A.M. (2022). Configurable Environments in Reinforcement Learning: An Overview. In: Piroddi, L. (eds) Special Topics in Information Technology, SpringerBriefs in Applied Sciences and Technology, Springer, Cham

Mnih, V., Kavukcuoglu, K., Silver, D., Graves, A., Antonoglou, I., Wierstra, D., & Riedmiller, M. (2013). Playing Atari with Deep Reinforcement Learning, ArXiv e-prints

Natale, L., Svetozarevic, B., Heer, P., & Jones, C. (2022). Near-optimal Deep Reinforcement Learning Policies from Data for Zone Temperature Control, ArXiv e-prints

Sembroiz, D., Careglio, D., Ricciardi, S., Fiore, U. (2019). Planning and Operational Energy Optimization Solutions for Smart Buildings, *Information Sciences*, vol. 476, pp. 439-452

Yu, L., Xie, W., Xie, D., Zou, Y., Zhang, D., Sun, Z., Zhang, L., Zhang, Y., & Jiang, T. (2020). Deep Reinforcement Learning for Smart Home Energy Management, *IEEE Internet of Things Journal*, vol. 7, no.4, pp. 2751–2762

Wei, T., Wang,Y., & Zhu, Q. (2017). Deep Reinforcement Learning for Building HVAC Control, In Proceedings of the 54th Annual Design Automation Conference 2017 (DAC '17), Association for Computing Machinery, New York, NY, USA, Article 22, pp.1–6

Zhang, Z. & Lam, K.P. (2018). Practical Implementation and Evaluation of Deep Reinforcement Learning Control for a Radiant Heating System, In Proceedings of the 5th Conference on Systems for Built Environments (BuildSys '18), Association for Computing Machinery, New York, NY, USA, pp. 148–157