A STUDY ON THE MAINTENANCE SCHEDULING OF REFRIGERATION SYSTEMS USING THE REINFORCEMENT LEARNING ALGORITHM

Caio Filipe De Lima Munguba – e-mail**:** caio.munguba@ufpe.br
Federal University of Pernambuco, Mechanical Engineering Department (PPGEM), https://www.ufpe.br/ppgem
Gustavo de novaes pires leite – e-mail**:** gustavonovaes@recife.ifpe.edu.br
Alvaro Antonio Ochoa Villa – e-mail**:** ochoaalvaro@recife.ifpe.edu.br
Kilvio Alessandro Ferraz – e-mail: kilvioferraz@recife.ifpe.edu.br
Federal Institute of Education, Science and Technology of Pernambuco, Campus Recife, DACT/DACS/CACTR/CACSEM, https://portal.ifpe.edu.br/recife/

LN - 20 – Aplicações Industriais e Especiais

Abstract. *Since the outstanding results of Alpha-go™, Reinforcement learning have been attracting attention from most diverse decision-making tasks, from the game playing to engineering domain operations, such as vibration damping or wind-farms management. More recently, researchers have proposed theoretical frameworks for using machine learning, particularly reinforcement learning, to optimize maintenance scheduling. These approaches have seen successful real-world applications in fault detection for industrial structures like pipelines and gas turbines. Refrigeration has also emerged as a promising domain for applying machine learning techniques, whether for control or maintenance optimization, and this is driven by three key factors: the high energy demands and associated emissions of refrigeration systems, the significant costs that refrigeration failures can impose on low-margin retail businesses, and the critical importance of maintaining the cold chain to prevent product spoilage. Minding that, based on related works, we explore the Reinforcement Learning framework for a maintenance schedule to obtain an artificial intelligence actor able to manage the degradations of a refrigeration device over time. In the end, the proposed maintenance scheduler reduced emissions by around 6% and repair costs by around 33% if compared to basic maintenance scheduling methods. This ensures the flexibility of this framework and its suitability for deeper investigation in the search for the autonomous fault detector and maintenance scheduler.*

Keywords: *Reinforcement Learning, Refrigeration, Maintenance, Emissions, Costs.*

1 Introduction

The world cannot live without refrigeration devices. Since the mid-19th century, vapor compression refrigeration (VCR) changed the way society stores live stocks and perishables, improving healthiness and food safety (Cleland, 2020). As the cold chain has become ubiquitous globally, concerns have grown about its environmental impact and operational costs. In response to agreements like the Montreal Protocol and Kyoto Protocol, the refrigeration industry has been transitioning towards Low Global Warming Potential (low-GWP) refrigerants, for instance (Bobbo et al., 2018). And VCR had experienced substantial improvement toward higher Coefficient of Performance (COP). Thus, since the ’90s, despite its continuously growing role, the energy demand by refrigeration equipment has been decreasing over time (Paul et al., 2022).

Like all mechanical devices, refrigeration appliances are also subject to aging. A recent study by Paul *et al* (2022), found that after sixteen years of use, there is an average loss of 27% in performance, leading to higher emissions and low-reliability devices, jeopardizing the cold chain safety. To Loisel *et al* (2021) highlight how Artificial Intelligence (AI) is already aiding in tackling this challenge. Refrigeration has been identified as a promising field for applying machine learning techniques (Dey et al., 2018). The availability of affordable wireless sensors for the Internet of Things (IoT) has greatly facilitated the use of Data Acquisition Systems (DAQs). Consequently, there has been significant growth in research on Fault Detection and Diagnosis (FDD) and Maintenance Scheduling (MS) in recent years, largely leveraging data-driven technologies such as machine learning (Singh et al., 2022).

For instance, Kulkarni *et al* (2018) proposed an automated fault detection system based solely on temperature data, achieving approximately 84% precision by shifting from a corrective/preventive approach to predictive or Condition Based Maintenance (CBM). Zhang *et al* (2020) demonstrated that ensemble-based classifiers can achieve up to 99.58% accuracy in fault detection. Overall, numerous studies have applied supervised or unsupervised learning methods to FDD in Heating, Ventilation, and Air Conditioning (HVAC) systems.

The RL field also popped up as a suitable tool for solving refrigeration tasks, as explored in Barret *et al.* (2015) and Wei *et al.* (2017), who proposed RL based autonomous HVAC control; or prediction, as concluded by Jang *et al* (2021) or Liu *et al.* (2019), who investigated the potential of RL for energy pricing in office HVAC. Meanwhile, Yousefi *et al* (2020) and Zhang *et al* (2020) proposed the theoretical suitability of the RL algorithms for maintenance planning. Later-on, researchers such as Hu *et al.* (2022) proposed a linear programming with RL pipeline to enhance maintenance decisions under uncertainties. The concept, already proven had been explored in theoretical frameworks such as in

Mahmoodzadeh *et al.*, (2020), who proposed RL based CBM for dry gas pipelines, or de Lima Munguba et al. (2023), who proposed a framework for applying RL for CBM in cooling devices.

Reinforcement learning (RL) is particularly suited for autonomous decision-making tasks due to its ability to learn from environmental signals and rewards (Sutton & Barto, 2018). Key attributes include learning by interaction, focusing on long-term returns, being object-oriented, and balancing exploration with exploitation. Considering reinforcement learning's suitability for FDD, this paper aims to contribute by exploring a reinforcement learning-based maintenance scheduler for a refrigeration freezer. The remainder of the paper is structured as follows: Section 2 presents the materials and methods, including the modelling, scheduler design, and experimental setup. Section 3 covers the results and their interpretation. Section 4 provides the overall conclusions of the work.

2 Methods and materials

Freezers, the focus of this work, typically consist of a standard top-opening container to preserve perishables with a compressor, low-GWP refrigerant, static evaporator and condenser, temperature sensor, and control board, operating on a compression cycle. Usually, the internal temperature sensor drives the compressor operation, with the temperature influenced by the load characteristics and mechanical component maintenance state. In this work, the data acquisition setup was managed by an STFMi refrigeration multimeter, a current sensor, a temperature sensor, and lately, data processing, as in Figure 1 (de Lima Munguba et al., 2023).

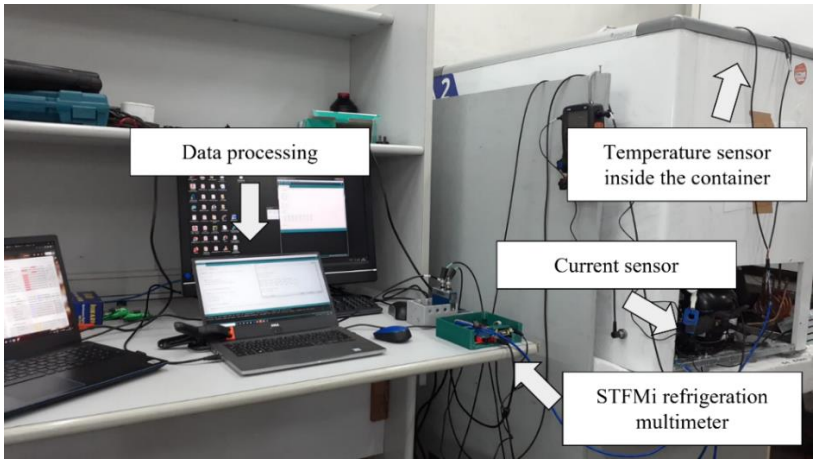

Figure 1. Experimental setup scheme, obtained from (de Lima Munguba et al., 2023).

In principle, the goal of this setup is to gather experimental data to feed the freezer model and create a dynamic state test bench. So, the freezer model, Figure 2, freezer data (1) relies upon the time series constrained by the datasheet, the outcomes then are sent to the costs (2). Note that agent (3) can access both the rewards and the freezer outputs, i.e., temperature (τint_T), Current (I_T), if the door is closed (dr_T) and the compressor delay time (ϑ_T). Then (3) can send back an action (a_T), which will be managed by the degradation model (4) and adjusted by action freezer manager (5) reconfiguring the freezer simulation. In real life, (1) and (4) are the freezer itself, (2) comprises the theoretical costs approximators, (3) is the remote AI algorithm interacting via IoT, and (5) is the action of the maintenance crew. In this work, (4) and (5) are also simulations, and that is why it is named a test bench. The agent must understand the results of the maintenance actions (5) and the costs (2) as the compressor ages.

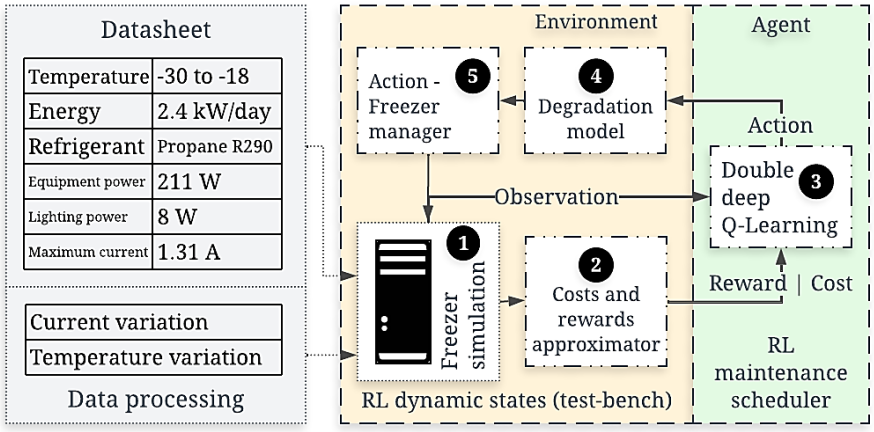

Figure 2. The framework of the studied maintenance scheduler and the test bench.

2.1 Dynamic states

Since RL learns through interaction, a simple time series cannot accomplish the interactivity criterion. As done by Mahmoodzadeh *et al.* (2020), an interactive model was built. The RL does not need to know the physics behind the readings to build the maintenance policy, it just relies on the inputs, understanding the model as a 'black box'. From DAQ, it is possible to infer the Reference Operation (ROP) and Transient Operation (TOP) conditions through the reconciliation method (Dumont et al., 2016). Based on the temperature time series, minding Newton's law of cooling, we established a simplified set of low resolution calibrated readings as the freezer and its operational states. The model is based on a cooling and heating rates, q, in C°/s, observing the heat flow ϕ and the time T. But minding that in the real-world q does not steady, the freezer model manipulates $q_T \in (q, \infty^+)$ when dr_T is open, the load $\varpi_{lT} \in (1{,}1.5)$ for q_T, the degradation of the compressor g, reducing q_T for $\phi < 0$ and the current. Indexing the I_T to the ϕ_T the model can obtain how much energy was expended to reach τmin, the minimum internal temperature. So, the observation O_T can be partially described as the set of equations below, Eq.1:

$$O_T(\tau int_T, I_T) = \begin{cases} \tau int_T = \begin{cases} (q_T * \Delta T) + \tau, \ if\ \tau - \tau' > q\ *5. \\ (q_T * \delta\tau * \Delta T) + \tau, \ if\ \tau - \tau' < q\ *5 \end{cases} \mid q_T = \begin{cases} \frac{q*\varpi_{lT}*\eta v_T + \eta h_T + sf_T}{3} * \eta g_T, \ if\ \phi < 0\ and\ f \neq 0 \\ q = q \forall \phi > 0, \ if\ \phi < 0\ and\ f = 0 \\ q = \varpi_{qT}, \ if\ \phi > 0\ and\ dr_T = True \\ q * a * \left(1 + (1 - \eta i)\right), \ if\ \phi > 0 \end{cases} \\ I_T = \begin{cases} \frac{cg*\varpi_g + cd*\varpi_d}{f}, \ if\ \phi < 0\ and\ f \neq 0 \\ \frac{cd*\varpi_d}{f}, \ if\ \phi > 0 \\ 0, \ if\ f = 0 \end{cases} \end{cases} \quad (1)$$

Note that $\tau - \tau'$ is conditioned to a q index. This is needed because this model uses a timespan of 300s between the readings. Hence, the temperature in the next timestep τ' is constrained to not generate unreal readings. Therefore, Eq.2, is the power consumption with c_T being the freezer power demand, Em_T, Eq.3, the emissions in g of Co_2, e_T the net emissions (IEA, n.d.), Ta_T, Eq.4, the tariff in €, and p_T, the price of the kWh (Eurostat, n.d.), (de Lima Munguba et al., 2023).

$$Co_T = \frac{\frac{\sum c_T}{1000} * 5 * 1}{60} \quad (2)$$

$$Em_T = e_T \frac{\frac{T * \sum c_T}{1000} * 5 * 1}{60} \quad (3)$$

$$Ta_T = P_T \frac{\frac{T * \sum c_T}{1000} * 5 * 1}{60} \quad (4)$$

2.2 Aging

In this work, only the compressor efficiency is manipulated as $\eta g_T \in (0{,}1)$, so, as TOP. As in Mahmoodzadeh *et al.*(2020) and de Lima Munguba *et al.* (2023), the degradation is described by a Markov decision Process (MDP) with different states $j_T \forall g$. An MDP is essentially the tuple $\left(j_T, a_T, p^a_{jj'}, y_T, \gamma\right)$, being j_T the state, a parametrical abstraction, a_T the action gave by the agent, $p^a_{jj'}$ the transition probability, y_T the cost function, and γ the discount factor. By that, the transition between the states $j \rightarrow j'$, $j_T \in (j, j')$, is driven by both the transition probability and the actions. Thus, in the maintenance frameworks, while j_T can freely transit from j to j' while degrading, return from $j' \rightarrow j$ is impossible without the agent's interference. Since in this work Dg_T is not formally a j_T, this MDP is also a Hidden Markov Model (HMM). The HMM retrieves a data emission subjected to j_T, here ε_T, the Weibull distribution between brackets in Eq. 5, and the speed rate g degrades at each T. So, to simulate the aging with a faulty state, j_T returns a higher $\kappa_j$parameter for j' than to j, increasing ε_T and quickening $Dg_T \mid Dg_T = 1/\eta g_T$. This is easier understood with Eq. 5, where elg is the compressor expected life span.

$$\eta g_T = \frac{\{100 - [\kappa_j x^{\kappa_j - 1} exp\left(-x^{\kappa_j}\right)] * \frac{100}{elg}\}}{100} \ \forall x > 0, \kappa_j > 0 \quad (5)$$

2.3 Action

The agent's maze resides in manage the compressor degradation Dg_T and avoid the costly standing in j' while tracking $O_T(\tau int_T, I_T, dr_T, \vartheta_T)$. It is made by a set of actions a_T, as $a_T \in (0{,}1{,}2)$, and being a_{0T} do nothing, a_{1T} repair and a_{2T} change. The action quantification was inspired by the reduction in failure intensity model proposed by Doyen and Gaudoin (2004). This is made by a discount factor fd as $fd \in (0{,}1)$ applied to Dg_T, so $Dg_{T'} = Dg_T * fd_T$. This detailing is available in Figure 3, based in (de Lima Munguba et al., 2023).

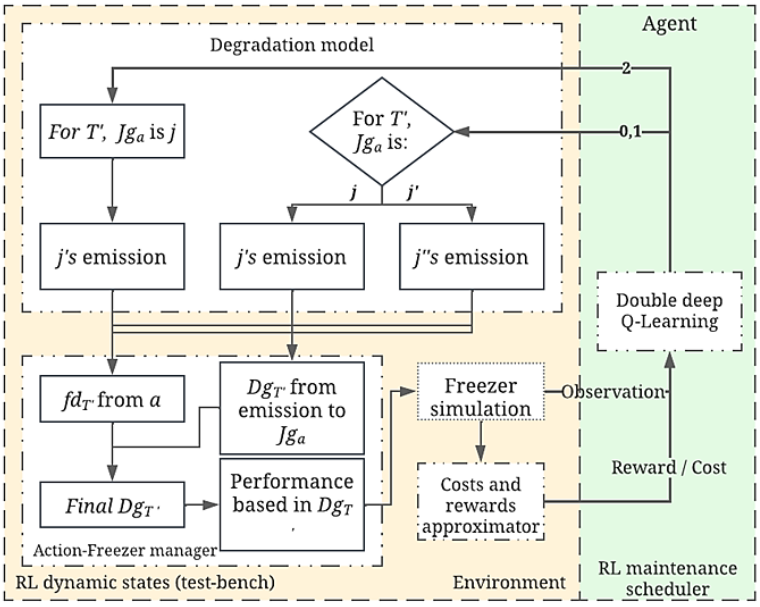

Figure 3. Details of the degradation model and the adjustment of the degradation model to maintenance actions.

2.4 Reinforcement signal

Lastly, there is the reinforcement signal. Designing the reinforcement signal is often considered an art, as the desired agent behavior hinges critically on getting the rewards right (Yousefi et al., 2020). Regarding maintenance, the goals of the reward-cost function $R_T - y_T$ are:

- Avoid complete failures, such as loss of functional capacity and equalization of internal and external temperature.
- Prolong the running time and device availability.
- Reduce maintenance costs.

In Yousefi *et al.* (2020), the R_T is often subdued by y_T if $a_T \in (1,2)$. Each a_T return time I_{aT} and price y_{aT}, but only the prices are reported to the agent. The maintenance prices use the η_{gT} as a guide, so if higher the Dg_T, higher the costs β. Here, $\beta g_{aT} \mid \beta g_{1T} \in \{80, 250\}$, but dissimilarly, $\beta g_{aT} \mid \beta g_{2T} \in \{500, 700\}$ since the cost of a new compressor is independent of the state of the previous one. After all, technician and other costs are given by $\beta r_{aT} \mid \beta r_{(1,2)T} \in \{80, 120\}$. Similarly, I_{aT} comes from $\iota_1 = 0$, $\iota_1 \in \{6{,}18\}$ and $\iota_2 \in \{18, 36\}$. Table 1 summarizes every action-result pair, and then, we have the expected time and cost of each action.

Table 1. Estimation of the cost of the intervention for each of the actions.

Action	Time	Price
0	$\mathrm{I}_{aT} = 0$	$y_{0T} = 0$
1	$\mathrm{I}_{aT} = \frac{maxDc_T}{100} * (max\iota_1 - min\iota_1) + min\iota_1$	$y_{1T} = \frac{Dc_T}{100} * (max\beta g_1 - min\beta g_1) + min\beta g_1 + \beta r_1$
2	$\mathrm{I}_{aT} = \iota_2$	$y_{2T} = \beta g_2 + \beta r_2$

The sizing of y_{aT}, despite inspired in real-world values, aims to guide the agent as well, avoiding unwanted actions and penalizing bad policies while ensuring the value of good decisions. Here, the best decisions are those that lead to both minimal y_{aT} and Ta_T in the long term and are unknown till the agent builds its action policy over R_T, so it must be built wisely since the R_T must give to the agent a 'feeling' about the environmental status. This feeling means how close the agent is to its target, so closer, higher rewards (Knowles et al., 2011; Kongkijpipat et al., 2022; Koprinkova-Hristova, 2014; Valet et al., 2022). Here, it was done by six indicators related to the freezer's working status.

First, spi_T is the reward for reaching the set-point temperature. spi_T is purposed shaped like a linear equation between the outside temperature τout_T and the set-point temperature τmin to assure the agent is rewarded by chasing reach τmin overtime. Therefore, when there is no energy in the system, $spi = 0$. Since the temperature readings can variate largely without any sign of malfunction, the shaping of spi_T also has the purpose of stabilizing the rewards. This is done by summing spi_T and the temperature difference indicator, dti_T. dti_T measures the $\Delta\tau \; \forall \; T \rightarrow T'$, and on this basis, rewards the agent when the system goes away from τout_T.

Ii_T reads the current and penalizes the agent when the readings are out of the ROP, Ip while ϱ is a coefficient defined to say how much Ii_T is relevant to R_T. This paper adopted $\varrho = 3.23$. The freezer in this study does not have variable speed technology. So, τin_T cycles around τmax and τmin as the compressor is activated and deactivated. This is cycling is accounted by coi_T. Most of the time, $coi_T = 0$, but if $\xi_T > 4$, a cascading algorithm is called. First, ξ_T counts the compressor time during the cooling cycle, and then, ϑ_T returns an indicator to the agent. Since the cycling can be misunderstood by the agent, ϑ_T is only accounted when the compressor time buffers over a threshold. However, coi_T does not obliterate the agent's decision-making, and therefore does not drive it, just ensures dti_T.

Another indicator simply boosts the agent when there is a transition $j' \rightarrow j$ during the training. This is a 'trick' for accelerating the learning. Another 'trick' is to penalize the agent when there is a complete failure. Lastly, since the agent must function as little as needed, an action penalty was also set to prevent biases and induced malfunction.

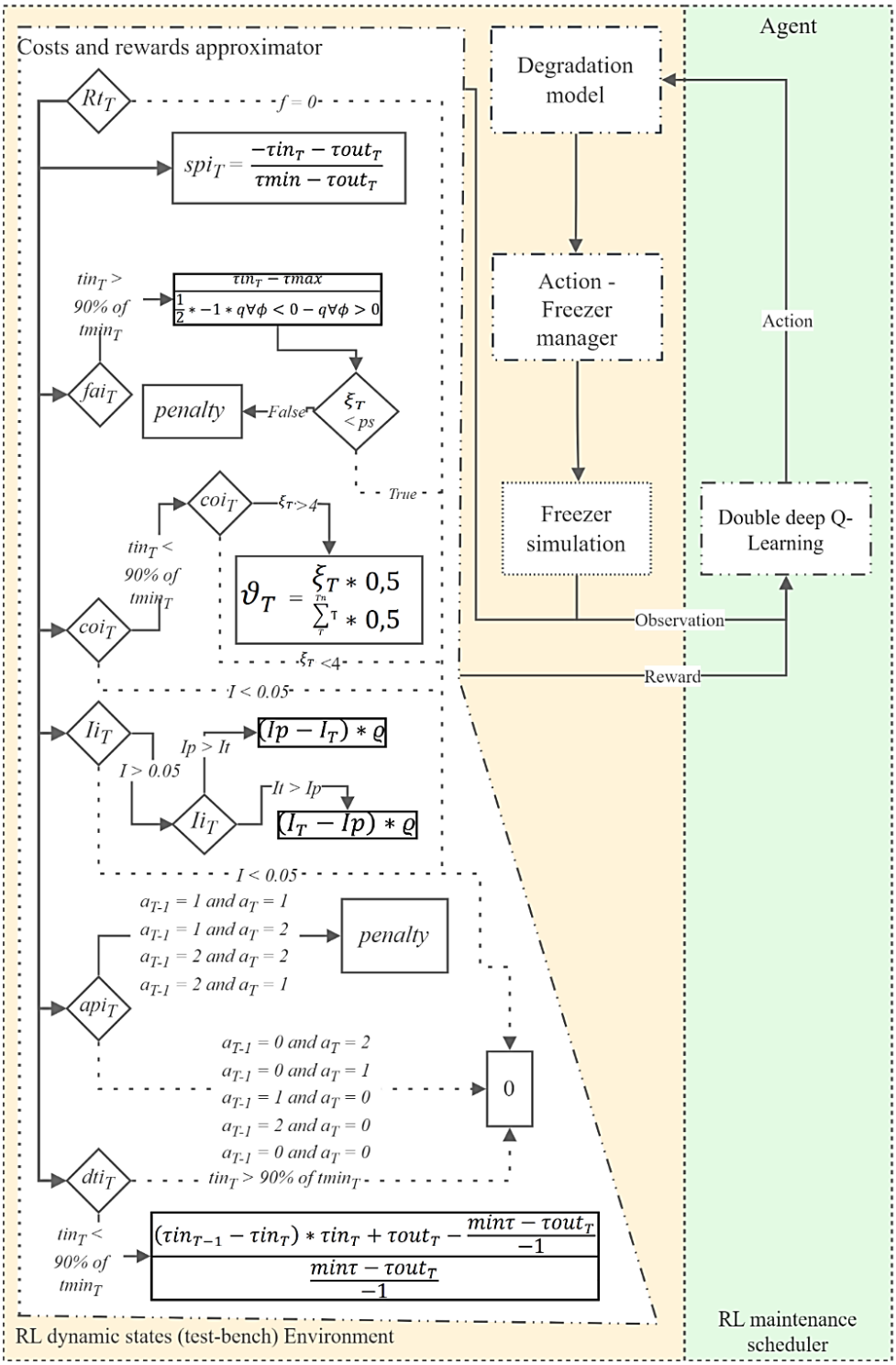

Figure 4. Details of the costs and rewards approximator.

Thus, R_T, to be maximized by the agent is defined by Eq.6.

$$R_T = spi_T + coi_T + dti_T - li_T - fai_T - ps_T - ap_T - \mathcal{Y}_{aT} \quad (6)$$

3 Double deep Q-learn maintenance scheduler

The Q -learning, presented in 1989 by Watkins (Watkins, 1989), is an immensely popular RL algorithm. In Q-Learn, each state-action pair, called a function Q, is stored in a table. The action chosen for each state, the 'policy', is the highest value in the table for the state-action pair. The basic function of the Q - learn is described by Eq.7 (van Hasselt et al., 2015):

$$Q_\pi(j,a) = E[r1 + \gamma r2 + \cdots \mid J_0 = j, A_0 = a, \pi] \quad (7)$$

where γ is the discount function, E means expected value, r is the reward, j is the state, and a is the action. Thus, the optimal value is $Q_\pi(j,a) = max_\pi Q_\pi(j,a)$. There are problems when the j is not directly observable, such as in this work. To them, the optimal value becomes $Q(j, a.\Theta_t)$, and thus the table update Q for an action a_t in the state j_t, receiving and resulting r_T, in the state j_t, becomes Eq.8 (van Hasselt et al., 2015), where α is the learning rate and ∇ the gradient operator.

$$\Theta_T, = \Theta_T + \alpha\left(Y_T^Q - Q(j_T, a_T.\Theta_T)\right)\nabla_{\Theta T} Q(j_T, a_T.\Theta_T) \quad (8)$$

Deep Q-learning is about replacing the Q-table with a multilayer neural network that takes care of returning a vector of actions for a given state $Q(j.\Xi)$, which Ξ is the neural network parameters. For Hasselt *et al.* (2015), the Deep Q-Network (DQN) has two relevant attributes compared to traditional Q-learning: the so-called target network and the experience replay. The target network corresponds to Eq.9, differing in that the parameter Θ becomes the Ξ, and is updated each time T from the network, and the experience replay is a memory defined to store the tuple (j, a, r, j') for j non-terminal during the phase of neural network training.

$$Y_T^Q = r_T, + \gamma^{max}_a Q(j_T,, a.\Xi'_T) \quad (9)$$

However, approaches like DQN and Q-learning can be overly optimistic due to limitations in the value function approximator's flexibility (van Hasselt et al., 2015). One solution, which can positively help in maintenance tasks, is to change the target network function to Eq.10.

$$Y_T^{doubleQ} = r_T, + \gamma Q(J_T,, ^{max}_a Q(J_T,, a.\Xi_T).\Xi'_T) \quad (10)$$

The main difference is in the application of a second approximator ${\Xi'}_T$, the output target network of the neural network. Then, the first Ξ_T become the traditional DQN configurator, and ${\Xi'}_T$ become the actions evaluator. The higher stability of Double Deep-Q Network (DDQN) face DQN and tabular Q-learn drove the choice of its application as the maintenance scheduler in this work.

4 Results and discussion

This section presents the results of both training and evaluation of the proposed pipeline. First, the training will be discussed. Then, the agent's results are compared to preventive and corrective programs.

4.1 Training and agent birth

The training and testing were performed under the hyperparameters available in Table 2. Over several tests, those returned the most stable and successful results.

Table 2. Hyperparameters and architecture.

Learning rate	$1x10^{-4}$
Network architecture	Dense network, layers: Input: 10, no activation Hidden: 48,32,32,24. ReLu activation Output: sixteen, no activation
Type of algorithm	Double DQN
Training episodes	750
Batch size	100

Discount fator	0.99

The training process for this HMM exhibits some typical characteristics for RL agents, as illustrated in Figure 5. First, the agent prioritizes exploring the environment and learning how to avoid penalties. |After, the next step is to evaluate the optimal action policy. The left roughness is often known as the exploration x exploitation dilemma because the agent begins towards the borders of its actions and then converges to a best strategy. When that happens, in this HMM, the penalties vanish and the episode reward becomes stable, as can be seen. This means the training has been successful and the evaluation can be performed.

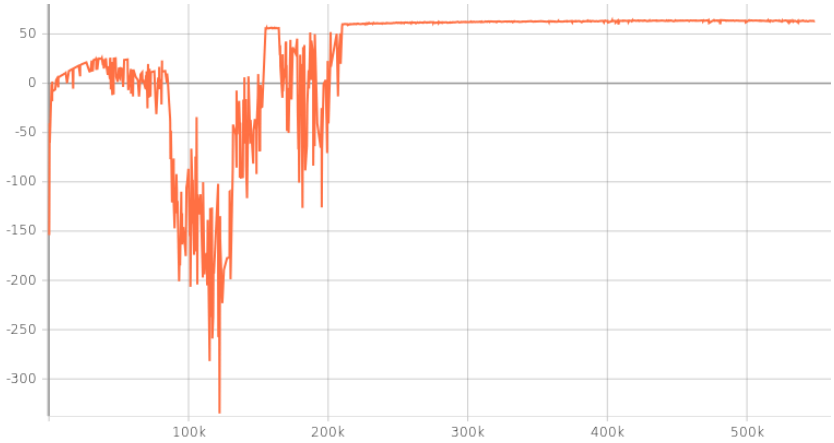

Figure 5. Agent update return per episode overtime.

4.2 Maintenance scheduling

Beforehand, the RL algorithm successfully strived for the highest reward in the long term. That pursuit also led to lower emissions, consumption, repair costs, and downtime. In addition, none of the experiments reported a break in the cold chain or loss of perishables, i.e., complete failure. Aftermath, the results are in Table 3.

Table 3. Maintenance scheduling results.

-	Program	Rewards	Emissions g of CO2	Consumption kWh	Repair cost	Downtime repair in hours
Median	Corrective	49877652.7	3588950.9	14723.1	0	0
	Preventive	846181.6	3415093.4	14027.8	1077.48	1.18
	DDQN Predictive	899039.03	3407202.6	13996.4	774.49	1,09
Coefficient of variation	Corrective	-0.53	0.017	0.016	1.87	1.86
	Preventive	-2.84	0.020	0.019	0.46	0.35
	DDQN Predictive	0.018	0.002	0.002	0.75	0.28

From a ML perspective, the shaping of the $R_T - y_T$ and O_T signals provided the agent with sufficient information to develop an effective policy. Given that the problem was a MDP nested within an HMM, this planning approach demonstrates the potential of deep reinforcement learning algorithms, such as DDQN, as maintenance schedulers. Maintenance schemes for refrigeration systems are often limited and complex due to multiple nonlinearities and parameters. However, as with other ML approaches, RL has shown remarkable results that outperform more extensive modeling, potentially easing wider adoption. These findings align with similar research in related domains.

From a refrigeration standpoint, the RL-based approach aligns with CBM and its associated benefits. Stability is a crucial feature, and the RL agent's CBM strategy resulted in lower coefficient of variation (CV) over the long term. This means the action policy maintained the system closer to the ROP, translating to reduced life-cycle costs.

From the environmental standpoint, this approach helps reduce emissions and improve food safety. For the low-margin retail sector, the ability of the RL agent to optimize costs is a significant advantage, as discussed previously.

This is better assessed in Figure 6 (a), which displays the approach of the agent. Here, 60% of maintenance costs by agent actions are below the minimum threshold of preventive, squared, and 100% of the power costs are below the preventive median. Another question might be raised about the 40% over the minimum threshold. This can be explained by the cost approximator itself. Its disclaimed, however that since they are computed as part of the $R_T - y_T$, whenever the

agent calls the costs are accounted. Sometimes, the agent repeatedly calls for action to avoid the penalties, when that happens, whether the repair is executed once as in real world, but the costs were computed twice in the test bench, as a penalty.

Availability, as presented earlier, also was a concern of the agent. Since there was a right reflection in the TOP status, the repairs were grouped into a narrower performance range, lowering downtime. This predictability is desirable because acting before the failure also gives time to manage unplanned maintenance activities. With no tracking, the freezer often reaches high degrees of degradation, pressuring local maintenance crew and often breaking the cold chain. That never happened under the agent, as in Figure 6 (b).

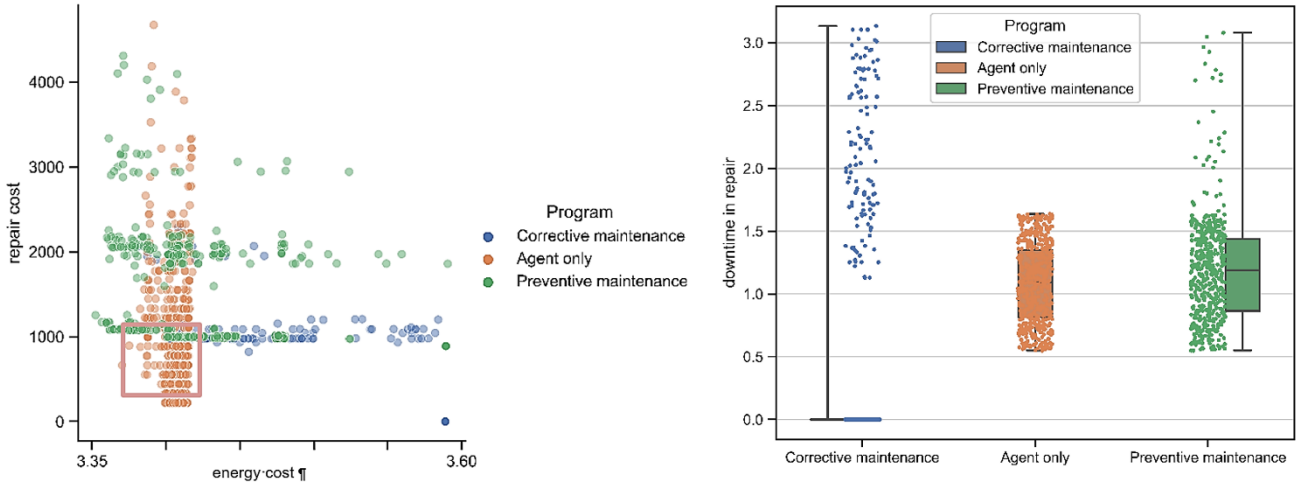

Figure 6. Repair cost x energy cost in the lifecycle (a), and average theoretical downtime in repair (b).

Lately, emissions are also a major concern. In this viewpoint, the built policy is also managed more efficiently than the traditional scheduling. In Figure 7, there is a huge spike in the KDE around $3.4x10^6$, while the corrective program regularly drives the freezer to the highest possible consumption over time and the preventive program generates a much-scattered range of possible emissions. This once more proves the potential of CBM tracking with RL, which can become a powerful tool in the pursuit of the global climate-changing defeating goals.

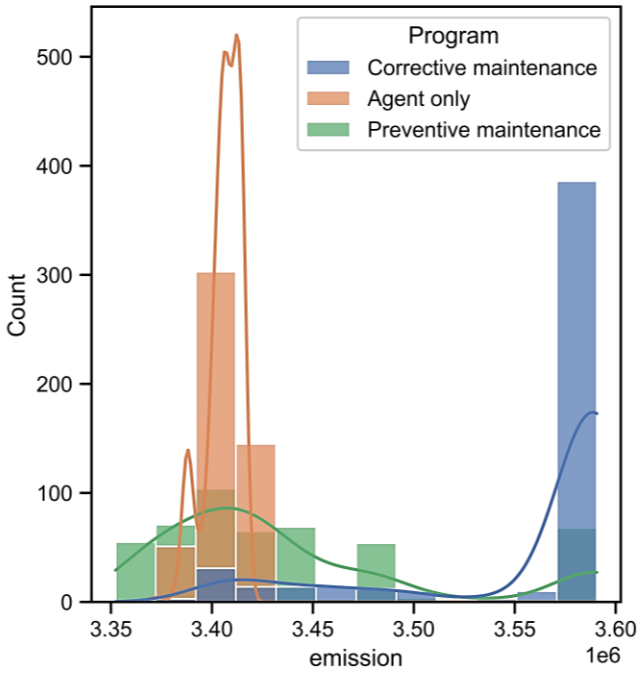

Figure 7. Freezer emissions in the life cycle for each maintenance program.

5 Conclusions and future work

In this paper, refrigeration cycle simulation data were learned using the RL method to generate maintenance scheduling based on aging and fault. It was demonstrated that the performance can be estimated by the agent based on the O_T and $R_T - y_T$ signals quite accurately, and thus performing its desired task as designed.

The results showed the proposed method identified faults and avoided complete failures. For the quantitative parameters, the agent managed to pursue the higher rewards, which also reflected in 6% fewer emissions compared to corrective maintenance and 28% lower repair costs, if compared to preventive, but major increases were found while improving system availability by reducing the number and time of maintenance stops. There was no complete failure under agent management. All these benefits are reflected in better performance and cost predictability, relevant to perishable loads and grocery stores.

Although its promising results, the proposed method is still limited by the freezer model used as input data. Thus, in continuation of this research, the authors are advancing the RL agent to overcome the current results to then, developing a trustable and efficient RL solution that requires minimal data inputs when applied to a real freezer.

6 Acknowledgments

The first author thanks the CAPEs for the scholarship of the master's degree and the PPGEM/UFPE. The second and the third authors thank the IFPE for its financial support throughout the Call 10/2019/Propesq. The third author thanks the CNPq for the scholarship of Productivity nº 309154/2019-7.

7 References

Barrett, E., & Linder, S. (2015). Autonomous hvac control, a reinforcement learning approach. *Lecture Notes in Computer Science (Including Subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, *9286*, 3–19. https://doi.org/10.1007/978-3-319-23461-8_1

Bobbo, S., Nicola, G. di, Zilio, C., Brown, J. S., & Fedele, L. (2018). Low GWP halocarbon refrigerants: A review of thermophysical properties. *International Journal of Refrigeration*, *90*, 181–201. https://doi.org/10.1016/J.IJREFRIG.2018.03.027

de Lima Munguba, C. F., de Novaes Pires Leite, G., Ochoa, A. A. V., & Lopez Droguett, E. (2023). Condition-based maintenance with reinforcement learning for refrigeration systems with selected monitored features. *Engineering Applications of Artificial Intelligence*, *122*. https://doi.org/10.1016/j.engappai.2023.106067

Dey, M., Rana, S. P., & Dudley, S. (2018). Semi-supervised learning techniques for automated fault detection and diagnosis of HVAC systems. *Proceedings - International Conference on Tools with Artificial Intelligence, ICTAI*, *2018-November*, 872–877. https://doi.org/10.1109/ICTAI.2018.00136

Doyen, L., & Gaudoin, O. (2004). Classes of imperfect repair models based on reduction of failure intensity or virtual age. *Reliability Engineering & System Safety*, *84*(1), 45–56. https://doi.org/10.1016/S0951-8320(03)00173-X

Dumont, O., Quoilin, S., & Lemort, V. (2016). Importance of the reconciliation method to handle experimental data in refrigeration and power cycle: application to a reversible heat pump/organic Rankine cycle unit integrated in a positive energy building. *International Journal of Energy and Environmental Engineering*, *7*(2), 137–143. https://doi.org/10.1007/S40095-016-0206-4

Eurostat. (n.d.). *Development of electricity prices for household consumers, EU, 2008-2021 (EUR per kWh) v2.png - Statistics Explained*. Retrieved August 22, 2022, from https://ec.europa.eu/eurostat/statistics-explained/index.php?title=File:Development_of_electricity_prices_for_household_consumers,_EU,_2008-2021_(EUR_per_kWh)_v2.png

Hu, J., Wang, Y., Pang, Y., & Liu, Y. (2022). Optimal maintenance scheduling under uncertainties using Linear Programming-enhanced Reinforcement Learning. *Engineering Applications of Artificial Intelligence*, *109*, 104655. https://doi.org/10.1016/J.ENGAPPAI.2021.104655

IEA. (n.d.). *Average CO2 emissions intensity of hourly electricity supply in the European Union, 2018 and 2040 by scenario and average electricity demand in 2018 – Charts – Data & Statistics - IEA*. Retrieved August 22, 2022, from https://www.iea.org/data-and-statistics/charts/average-co2-emissions-intensity-of-hourly-electricity-supply-in-the-european-union-2018-and-2040-by-scenario-and-average-electricity-demand-in-2018

Jang, D., Spangher, L., Srivistava, T., Khattar, M., Agwan, U., Nadarajah, S., & Spanos, C. (2021). Offline-online reinforcement learning for energy pricing in office demand response: Lowering energy and data costs. *BuildSys 2021 - Proceedings of the 2021 ACM International Conference on Systems for Energy-Efficient Built Environments*, 131–139. https://doi.org/10.1145/3486611.3486668

Knowles, M., Baglee, D., & Wermter, S. (2011). Reinforcement learning for scheduling of maintenance. *Res. and Dev. in Intelligent Syst. XXVII: Incorporating Applications and Innovations in Intel. Sys. XVIII - AI 2010, 30th SGAI Int. Conf. on Innovative Techniques and Applications of Artificial Intel.*, 409–422. https://doi.org/10.1007/978-0-85729-130-1_31

Kongkijpipat, P., Sandee, C., Vachirapaneegul, S., Sumetpipat, K., & Vatiwutipong, P. (2022). Wet Gas Pipeline Maintenance Process Using Reinforcement Learning. *2022 19th International Joint Conference on Computer Science and Software Engineering, JCSSE 2022*. https://doi.org/10.1109/JCSSE54890.2022.9836258

Koprinkova-Hristova, P. (2014). Reinforcement Learning for Predictive Maintenance of Industrial Plants. *Information Technologies and Control*, *11*(1), 21–28. https://doi.org/10.2478/itc-2013-0004

Kulkarni, K., Devi, U., Sirighee, A., Hazra, J., & Rao, P. (2018). Predictive Maintenance for Supermarket Refrigeration Systems Using only Case Temperature Data. *Proceedings of the American Control Conference*, *2018-June*, 4640–4645. https://doi.org/10.23919/ACC.2018.8431901

Liu, T., Xu, C., Guo, Y., & Chen, H. (2019). A novel deep reinforcement learning based methodology for short-term HVAC system energy consumption prediction. *International Journal of Refrigeration*, *107*, 39–51. https://doi.org/10.1016/J.IJREFRIG.2019.07.018

Loisel, J., Duret, S., Cornuéjols, A., Cagnon, D., Tardet, M., Derens-Bertheau, E., & Laguerre, O. (2021). Cold chain break detection and analysis: Can machine learning help? *Trends in Food Science & Technology*, *112*, 391–399. https://doi.org/10.1016/J.TIFS.2021.03.052

Mahmoodzadeh, Z., Wu, K. Y., Droguett, E. L., & Mosleh, A. (2020). Condition-based maintenance with reinforcement learning for dry gas pipeline subject to internal corrosion. *Sensors (Switzerland)*, *20*(19), 1–26. https://doi.org/10.3390/s20195708

Paul, A., Baumhögger, E., Elsner, A., Reineke, M., Hueppe, C., Stamminger, R., Hoelscher, H., Wagner, H., Gries, U., Becker, W., & Vrabec, J. (2022). Impact of aging on the energy efficiency of household refrigerating appliances. *Applied Thermal Engineering*, *205*. https://doi.org/10.1016/J.APPLTHERMALENG.2021.117992

Singh, V., Mathur, J., & Bhatia, A. (2022). A comprehensive review: Fault detection, diagnostics, prognostics, and fault modeling in HVAC systems. *International Journal of Refrigeration*, *144*, 283–295. https://doi.org/10.1016/J.IJREFRIG.2022.08.017

Sutton, R. S., & Barto, A. G. (2018). *Reinforcement Learning: An Introduction Second edition, in progress*.

Valet, A., Altenmüller, T., Waschneck, B., May, M. C., Kuhnle, A., & Lanza, G. (2022). Opportunistic maintenance scheduling with deep reinforcement learning. *Journal of Manufacturing Systems*, *64*, 518–534. https://doi.org/10.1016/J.JMSY.2022.07.016

van Hasselt, H., Guez, A., & Silver, D. (2015). Deep Reinforcement Learning with Double Q-learning. *30th AAAI Conference on Artificial Intelligence, AAAI 2016*, 2094–2100. https://doi.org/10.48550/arxiv.1509.06461

Watkins, C. J. C. H. (1989). *Learning From Delayed Rewards* [King's College]. https://www.cs.rhul.ac.uk/~chrisw/new_thesis.pdf

Wei, T., Wang, Y., & Zhu, Q. (2017). Deep Reinforcement Learning for Building HVAC Control. *Proceedings - Design Automation Conference*, *Part 128280*. https://doi.org/10.1145/3061639.3062224

Yousefi, N., Tsianikas, S., & Coit, D. W. (2020). Reinforcement learning for dynamic condition-based maintenance of a system with individually repairable components. *Quality Engineering*, *32*(3), 388–408. https://doi.org/10.1080/08982112.2020.1766692

Zhang, N., & Si, W. (2020). Deep reinforcement learning for condition-based maintenance planning of multi-component systems under dependent competing risks. *Reliability Engineering and System Safety*, *203*. https://doi.org/10.1016/j.ress.2020.107094

Zhang, Z., Han, H., Cui, X., & Fan, Y. (2020). Novel application of multi-model ensemble learning for fault diagnosis in refrigeration systems. *Applied Thermal Engineering*, *164*. https://doi.org/10.1016/j.applthermaleng.2019.114516

UM ESTUDO SOBRE PROGRAMAÇÃO DE MANUTENÇÃO DE SISTEMAS DE REFRIGERAÇÃO USANDO APRENDIZADO POR REFORÇO

Resumo. *Desde os expressivos resultados do AlphaGo™, o aprendizado por reforço vem atraindo atenção como otimizador para tarefas de tomada de decisão, abrangendo desde jogos até operações no domínio da engenharia, como amortecimento de vibrações ou gerenciamento de parques eólicos. Mais recentemente, pesquisadores propuseram arcabouços teóricos para aplicação de APRENDIZADO POR REFORÇO como programador de manutenção. E essas propostas têm-se mostrado bem-sucedidas quando aplicadas a estruturas industriais, como oleodutos e turbinas a gás. A refrigeração também se apresenta como um campo propício a aplicação de técnicas de Aprendizado de Máquina principalmente por três fatores: a alta demanda energética e as emissões associadas aos sistemas de refrigeração, os significativos custos que as falhas de refrigeração podem impor a empresas varejistas de baixa margem, e a importância crítica de manter a cadeia de frio para evitar a deterioração de produtos. Com base em trabalhos relacionados, exploramos a seara do Aprendizado por Reforço como programador de manutenção, para obter um agente de inteligência artificial capaz de gerenciar a degradação de um dispositivo de refrigeração ao longo do tempo. Ao final, o programador de manutenção proposto reduziu as emissões em cerca de 6% e os custos de reparo em aproximadamente 33%, em comparação com os métodos corretivo e preventivo, demonstrando sua viabilidade e a oportunidade tecnológica do desenvolvimento de agentes capazes de detectar falhas e programar manutenção de forma autônoma.*

Palavras-chave: Aprendizado por Reforço, Refrigeração, Manutenção, Emissões, Custos.