Arya Ashrafi

Reinforcement Learning for Decentralized Energy Systems

School of Technology & Innovation
Master's thesis in Smart Energy

Vaasa 2024

Acknowledgments

I would like to thank my gratitude to my supervisor, Prof. Miadreza Shafiekhah, for all of his help and counsel during the writing of this thesis. His guidance not only played a crucial role in molding this study but also served as a motivation and a standard for scholarly integrity and superiority.

Dr. Petri Välisuo, my co-supervisor, has my sincere gratitude as well. His insights have broadened my viewpoint and improved my comprehension of the issue.

I express my sincere appreciation to my family in Iran and here in Vaasa. Their constant support and unshakable faith in my abilities have been the cornerstones of my journey. I owe them the perseverance and tenacity that have allowed me to succeed in my academic career.

Without their continuous support, I would not be here today. For that, I am incredibly grateful.

This work was supported by the Horizon Europe project DiTArtIS (Network of Excellence in Digital Technologies and AI Solutions for Electromechanical and Power Systems Applications), grant agreement number: 101079242.

Arya Ashrafi

UNIVERSITY OF VAASA
School of Technology & Innovation
Author: Arya Ashrafi
Title of the Thesis: Reinforcement Learning for Decentralized Energy Systems
Degree: Master of Science
Program: Smart Energy
Supervisor: Prof. Miadreza Shafiekhah and Dr. Petri Välisuo
Year: 2024 **Number of Pages:** 66

ABSTRACT:
With the rise of electric vehicles (EVs) and smart grid technology, sophisticated energy management systems that guarantee sustainability and efficiency are required. An in-depth examination of reinforcement learning (RL) algorithms in a simulated smart grid system featuring prosumer-generated renewable energy and embedded EV charging stations is presented in this thesis. The study assesses how well the Deep Deterministic Policy Gradient (DDPG), Proximal Policy Optimization (PPO), and Rule-Based Control (RBC) algorithms manage the energy dynamics of 50 prosumer nodes and 40 EVs over a 24-hour period using a Markov decision process framework. The RL algorithms interact with the environment to learn sequential decision-making processes that maximize the overall reward, with a particular focus on balancing energy production, consumption, and vehicle charging demands. The simulation results reveal DDPG's strength in cost-efficient grid energy purchasing and effective state of charge (SOC) management, PPO's potential through exploratory learning, and RBC's advantage in minimizing energy wastage. The findings point towards the necessity of intelligent energy management strategies that not only minimize costs and maximize the use of renewable energy but also enhance the operational efficiency and sustainability of the smart grid and EV ecosystems.

KEYWORDS: Smart Grid, Reinforcement Learning, Electric Vehicles, Energy Management, Demand Response

Contents

Figures

Tables

Abbreviations

ANN	Artificial Neural Network
DDPG	Deep Deterministic Policy Gradient
DRL	Deep Reinforcement Learning
EM-SA	Energy Management System Aggregator
EV	Electric Vehicle
MDP	Markov Decision Process
ML	Machine Learning
MPC	Model Predictive Control
POMDP	Observable Markov Decision Processes
PER	Prioritized Experience Replay
PPO	Proximal Policy Optimization
PV	Photovoltaic
RBC	Rule-Based Control

RES	Renewable Energy Sources
RL	Reinforcement Learning
RNN	Recurrent Neural Network
SAC	Soft-actor Critic
SOC	State of Charge

1 Introduction

The increasing availability of electric cars (EVs) around the world offers tremendous potential as well as significant obstacles, especially when it comes to EV charging and how it affects the electrical grid. The increased use of EVs places more burden on our electrical infrastructure, especially during periods of high demand. This increasing demand has the potential to worsen system congestion, increase the probability of blackouts, reduce the effectiveness of power distribution, and increase consumer electricity costs. These difficulties are exacerbated in places where there is a high concentration of electric vehicles (EVs), as local grid bottlenecks may require major infrastructure modifications to handle the extra load, highlighting the pressing need for creative solutions (Khaki, 2019).

In this context, photovoltaic (PV) systems show up as an economically and sustainably sound addition to conventional energy sources for household use and EV charging. The transition to renewable energy sources, such as solar electricity, is essential for promoting environmental sustainability and preventing climate change. There is a crucial shift to renewable energy sources to mitigate climate change and advance environmental sustainability (Cohen, 2019). PV installations have the potential to significantly lower electricity prices over the long run. The long-term cost savings from producing one's electricity can more than offset the initial high setup costs of PV systems. This increases the attraction and viability of having an electric vehicle by making "fueling" an EV much less expensive for owners. Moreover, EV charging through PV system integration improves resilience and energy independence. A PV system can offer a dependable alternative energy supply during blackouts or times of heavy demand when the grid is stressed, guaranteeing that EVs can still be charged and essential home operations can go on without interruption (Esfandyari, 2019).

While PV systems provide a sustainable energy source for charging EVs in Yao (2024), their intermittent nature poses challenges for consistent EV charging. Solar energy production rises during the day, usually during a time when there is less need for charging, and falls off in the evening when usage usually increases. To ensure that EVs can be

charged with green energy effectively, this mismatch necessitates creative solutions, such as energy storage devices or smart charging techniques, to store extra solar energy generated during the day for use during peak charging hours.

According to Li (2021), technologies like demand response and smart charging offer workable answers to these problems. Smart charging systems optimize the charging process to lower peak demand and more effectively incorporate renewable energy sources by adjusting the charging rate or time of EVs based on grid conditions, PV generation, and EV owner preferences. To further balance the load and maximize the usage of renewable energy, demand response systems might encourage EV owners to charge their vehicles during off-peak hours or when there is excess renewable energy available. To ensure that the grid can support the expanding EV market while maximizing the use of renewable energy sources, advances in grid infrastructure, regulatory frameworks, and consumer engagement techniques are necessary for the widespread application of these technologies in Li (2021).

In this case, Lan (2021) looks into how renewable energy sources can be integrated with demand response and smart charging programs for electric vehicles (EVs), particularly when machine learning (ML) is used. It offers a cutting-edge strategy for improving grid stability and energy consumption optimization. With its ability to forecast energy use, optimize charging schedules, and manage the intermittent nature of renewable energy sources like solar and wind power, machine learning can greatly increase the efficacy and efficiency of these programs.

To optimize energy usage in residential neighborhoods and integrate renewable energy sources for EV charging, this research builds upon previous work to create a powerful reinforcement learning (RL) system specifically designed for 50 residential participants to manage their energy consumption. These residential participants who are considered to be prosumers in this case generate their electricity through photovoltaic (PV) systems installed in their homes, in addition to consuming it. Forty of these homes use electric

vehicles (EVs), which adds a big factor to the energy management equation. This study's main goal is to design a system that will allow these 40 EVs to prioritize the usage of renewable energy by charging from excess energy produced by separate PV systems. When the energy generated by the PV panels is not enough to satisfy the demands of charging, the system will help to obtain the necessary electricity from the grid. Due to the dynamic nature of this dual-source charging approach, energy management becomes more complex, requiring an intelligent system that can make decisions in real-time based on fluctuating energy prices and availability.

To traverse this intricate energy terrain, the study investigates the utilization of two sophisticated reinforcement learning algorithms: Proximal Policy Optimization (PPO) and Deep Deterministic Policy Gradients (DDPG). To provide a baseline for comparison, these algorithms will be assessed against a traditional Rule-Based Control (RBC) system. The comparison analysis will evaluate the effect of renewable energy generation on overall energy management and EV charging efficiency across a range of scenarios, including those without PV systems and those with them. The investigation's key component is its capacity to assess and contrast the efficacy of PPO and DDPG algorithms in handling prosumers' energy demands when using EVs and PV systems. The goal of the study is to determine the best RL approach for minimizing EV charging expenses and optimizing the use of locally produced renewable energy by examining various policy outcomes.

2 Background

2.1 Smart Charging & Demand Response

The management of electric vehicle (EV) charging through rate and timing adjustments in response to grid needs and renewable energy availability is known as smart charging. By doing this, the grid is not as strained during peak hours, and more clean energy generated by PV systems is optimally utilized (Yao et al., 2017). Energy distribution and management take on new dimensions with the integration of photovoltaic (PV) systems and electric vehicles (EVs) into the electric grid. Utilizing these technologies to their fullest potential, maintaining grid stability, and maximizing energy use require smart charging and demand response programs (Radu, 2019). Controlling EV charging rates and schedules to coincide with periods of low electricity demand or high renewable energy generation is made possible by smart charging. This optimizes the utilization of clean energy produced by PV systems and lessens the load on the grid during peak hours.

Programs that encourage consumers to modify their energy use in response to grid conditions further improve this, which are called demand response programs. Demand response (DR) programs include techniques such as price incentives or notifications during periods of high demand, that persuade customers to adjust their energy usage in response to grid conditions (Nezamoddini & Wang, 2016). The purpose of these programs is to maintain equilibrium between the supply and demand of power in the smart grid, hence enhancing stability and mitigating potential hazards. For example, EV owners can be incentivized to charge their cars during peak solar production periods, absorbing extra generation and avoiding cutoff. In the same way, customers can lower their consumption or postpone billing to help balance the grid during periods of high demand and low renewable generation. By encouraging a more robust and effective energy system, these tactics lessen dependency on fossil fuels and accelerate the transition to a sustainable energy future (Ferro, 2018).

The adoption of demand response and smart charging programs for PV systems and EVs adds a degree of efficiency and flexibility that has the potential to completely change conventional power markets. With the help of these initiatives, energy producers, customers, and the grid may interact more dynamically, creating new market opportunities and models.

2.2 Smart Electricity Market

To maximize electricity generation, distribution, and consumption, cutting-edge technology and market mechanisms combine in the smart electricity market, which is a creative and dynamic segment of the energy business. In the context of electric vehicles (EVs), which are contributing a growing amount to the demand for electricity, this market is especially appropriate. Below is the examination of the main features of the smart power market, such as payment structures, types, and market dynamics, with an emphasis on how these factors relate to EV charging.

2.2.1 Electricity Market Dynamics

The smart energy market is a marketplace where utility companies, independent power producers, and consumers—including owners of electric vehicles—trade electricity. These transactions occur when buying orders and selling offers, or matched asks and bids, occur in an auction. The features of these markets are designed to efficiently balance supply and demand. Different auction techniques are used to decide the price of energy, depending on the type of market. Dynamic pricing is crucial for owners of electric vehicles (EVs) who may choose to charge their cars during off-peak hours to save money on electricity (Morstyn, 2018).

2.2.2 Types of Electricity Markets

Day-Ahead Markets: In these markets, agreements are settled one or more days before electricity is actually delivered. Users can therefore plan their output or consumption

based on expected demand. This may mean scheduling EV charging sessions during times when lower expenses are predicted.

Spot Markets: These markets provide a faster reaction time by enabling transactions to be finished up to five minutes before delivery. This makes it possible to respond quickly to sudden changes in supply or demand, such as those resulting from EVs being charged in huge quantities.

Operating Reserve Markets: These markets have the fastest reaction times, ranging from minutes to seconds, and provide backup resources to maintain grid stability. This is especially important for electric vehicles (EVs), as ubiquitous charging can result in considerable increases in the amount of electricity used.

Most power markets operate under the merit order concept, which orders energy sources to be sent in increasing cost order until the demand is met. The price at which supply and demand are equal is known as the clearing price. Since it ensures that the most cost-effective resources are used first, EV owners may benefit from this notion by having their power expenses cut during times of strong renewable energy production (Al-Gabalawy, 2021).

2.3 Machine Learning for Energy Management

A sophisticated method for handling the challenges involved in integrating PV installations and EV charging into the smart grid is machine learning (ML). ML algorithms can forecast demand and optimize energy distribution by examining trends in energy generation, consumption, and a host of other variables, including user behavior and meteorological conditions. For instance, Shin (2020) states that predictive models capable of predicting bursts in solar radiation can be used to strategically schedule EV charging sessions to align with these periods, thus optimizing the usage of renewable energy. To avoid grid overload and boost system efficiency, ML can also dynamically modify charging rates in real time based on grid circumstances and the availability of solar electricity. Moreover, Abdullah, et al. (2021) suggest that reinforcement learning, a subset of machine learning, can be applied to create control techniques that adjust over time, learning from past actions to make more accurate decisions about the distribution and

consumption of energy. As the grid and its users' needs change, energy management solutions must be able to adapt as well. This is because renewable energy sources are naturally unpredictable, and managing their inherent variability and changing consumption patterns is critical. All things considered, the employment of machine learning in this situation not only raises the energy system's operational efficiency but also increases customer comfort and happiness, opening the door for EV and PV system adoption on a larger scale.

Here is a list of why utilizing ML can be beneficial in the context of smart charging and demand response:

Adaptive Smart Charging with Machine Learning: Machine learning algorithms can analyze vast amounts of data from the grid, EVs, and renewable energy sources to predict optimal charging times (Lan, 2021). ML models may dynamically modify EV charging schedules to periods when renewable energy production is high, minimizing dependency on non-renewable energy sources and cutting consumer power bills. This is achieved by identifying patterns in energy consumption and generation.

Demand Response Optimization: To lower total energy demand, demand response programs try to motivate customers to use less energy during off-peak hours or times when renewable energy sources are available. Machine learning enhances demand response technics by predicting times of peak demand and peaks in renewable energy output. ML models can automatically send signals to EVs and other smart appliances. This will help to alter their energy use in real-time and prevent over-load to maintain grid balance.

Integration of Renewable Energy Sources: Renewable energy sources are difficult to integrate into the grid because of their intermittent nature. By forecasting renewable energy generation using machine learning algorithms based on historical data and weather projections, grid operators can more effectively plan for energy storage and delivery. An ML model might, for instance, anticipate a decrease in solar power generation owing to

cloud cover and proactively modify EV charging schedules to maintain grid balance (Shin, 2020).

Enhancing Grid Stability: Smart grids can become more adaptable and resilient to variations in the supply and demand of energy by utilizing machine learning to assess data from several sources, such as energy consumption patterns, EV battery conditions, and projections of renewable energy generation. According to Mololoth et al. (2023), machine learning (ML) can aid in the dynamic management of energy flows, preventing EV charging from adding to grid instability and facilitating the efficient distribution of renewable energy throughout the network.

3 Literature Review

Within the domain of optimizing EV charging in liberalized energy markets, scholars have investigated several approaches with the goal of optimizing cost-saving advantages for EV owners in the face of fluctuating pricing structures and user behaviors. Foster (2013) emphasizes how important it is to incorporate smart grid data to provide EV owners with cost-saving options based on their current level of charge and planned departure time. Better cost optimization and increased charge decision flexibility in the face of price volatility are both made possible by this combination.

Radu (2019) provides a charging management plan for electric vehicles (EVs) to facilitate distributed generation and renewable energy integration. To attain a scheduled aggregated power profile, the technique takes into account the preferences of electric vehicle users. This is done in an effort to lessen the impact of intermittent energy sources like solar and wind power. It is intended to store the excess energy in EV batteries, which will increase electrical network reliability and enable widespread integration of distributed energy resources (DER). This strategy makes it possible for users of flexible electric vehicles to take part in demand response programs, which could be extremely important for enhancing the reliability and effectiveness of future smart grids. Shin (2020) elaborates on this idea of adaptability and suggests a decentralized method of controlling EV charging stations that have energy storage and photovoltaic (PV) systems installed. This methodology makes use of the surplus energy stored in EVs from other stations to enable autonomous control of charging stations through the use of multi-agent deep reinforcement learning. This decentralized approach, which reflects a move toward more independent and effective charging infrastructure management, not only allows for flexibility in decision-making but also reduces total costs through distributed coordination.

Current research emphasizes how important cutting-edge technologies are to improving the energy management of smart grids, especially when electric vehicles (EVs) are included. Examples of these technologies include blockchain and machine learning (ML). The growing interest in machine learning (ML)-driven smart charging techniques for

electric vehicles (EVs) holds great potential for improving demand-side management and cutting operational costs. For instance, López (2018) developed an intelligent EV charging approach that uses deep learning to maximize energy savings and optimize charging times. Their method eliminates the need for projections of future energy prices or vehicle usage by utilizing data on driving habits, the environment, and energy pricing. Their methodology's core is dynamic programming for the examination of historical data, which trains deep neural networks to make economical charging decisions in real-time. According to their research, deep learning can produce charging charges that are almost exactly like the ideal costs that were determined in the past, if not identical.

In the realm of ML, reinforcement learning (RL) has gained vast attention for the energy management of EVs and for providing demand response programs (Abdullah, et al. (2021) and Mololoth, et al. (2023)). In line with that, Vázquez-Canteli (2019) investigated demand response using reinforcement learning (RL) in smart grids with an emphasis on home energy systems. Their thorough analysis of RL techniques for a range of components, such as smart appliances and EVs, highlights how flexible RL is to the dynamics of the environment and how well it can integrate user feedback. Moreover, Li (2022) suggests a workable energy management plan that uses Deep Reinforcement Learning (DRL) to optimize EV charging expenses. The study aims to improve real-time charging management efficiency in the face of fluctuating electricity tariffs and user behaviors by integrating an improved Recurrent Neural Network (RNN) architecture and Deep Deterministic Policy Gradient (DDPG) algorithm.

Arwa (2021) uses a Markov Decision Process formulation to explore energy management and optimal scheduling in the setting of EV charging stations coupled with renewable energy. Through the introduction of an enhanced Q-learning algorithm for managing dynamic stochastic problems and grid tariff models, the study seeks to reduce the cost of purchasing electricity and encourage self-generation via photovoltaic systems. The integration of renewable energy sources with local generation and consumption highlights a significant step towards sustainable and efficient charging infrastructures, despite the

obstacles presented by grid prices and variability in renewable energy supply. The Q-learning technique is used to precisely anticipate loads across a range of scenarios, which helps to optimize plug-in hybrid Electric Vehicle (EV) charging stations This highlights the significance of adaptive and efficient charging strategies in dynamic energy environments.

To optimize operation and reduce expenses, Cai (2023) offers a performance-driven approach to energy management in residential microgrid systems that integrates Reinforcement Learning (RL) and Model Predictive Control (MPC) algorithms. This reflects the increased interest in using RL techniques to optimize the utilization of renewable energy sources as the study presents a viable paradigm for effective energy management in residential microgrid systems by integrating the benefits of both techniques, demonstrating the possibility for collaborative approaches to address local energy concerns. Ye (2020) looks at a bilevel optimization model that may be used to create bidding strategies in deregulated power markets. To deal with non-convexities, the model uses Deep Reinforcement Learning (DRL) to help make cost-effective judgments about charging and unplugging electric vehicles. It emphasizes the role that DRL techniques play in enabling adaptive and responsive decision-making strategies, underscoring the significance of ongoing feedback and strategy refinement in dynamic market situations. Also, Yan (2021) addresses the challenge of optimizing EV charging schedules in the face of fluctuating electricity costs and user behavior by utilizing Deep Reinforcement Learning (DRL) techniques in conjunction with a Markov Decision Process (MDP) framework. Sequential decision-making based on driving behavior is made easier by using off-policy algorithms within the soft-actor critic (SAC) framework. The ultimate goal is to find affordable charging solutions that are in line with user preferences and market dynamics.

Furthermore, Dabbaghjamanesh (2021) illustrates the superior accuracy of Q-learning by utilizing recurrent neural network (RNN) and artificial neural network (ANN) forecasting outcomes. This is especially relevant in scenarios where there are coordinated and disorganized charging behaviors, underscoring the significance of precise load

forecasting for efficient charging optimization. The Energy Management System Aggregator (EM-SA) is part of a demand response program to maximize power sales to the grid while limiting purchases to maximize advantages. The EMSA addresses challenges related to renewable energy integration and variable weather conditions by utilizing a two-level Model Predictive Control (MPC) framework and Reinforcement Learning (RL) algorithms for real-time decision-making. Qiu (2020) promotes the optimization of EVs' continuous charging and discharging levels through the use of deep reinforcement learning algorithms, with a focus on the Deep Deterministic Policy Gradient (DDPG) method. Prioritized experience replay (PER) approaches are used to improve learning efficiency, which leads to more resilient and flexible charging schemes. This study also emphasizes how crucial it is to solve the EV pricing issue that aggregators are facing and how EV flexibility affects average electricity bills and aggregator revenue.

There are even new opportunities for demand management and the safe integration of renewable energy sources thanks to Mololoth, et al.'s (2023) investigation on the synergistic potential of blockchain and machine learning for upcoming smart grids. On the other hand, some researchers and industry stakeholders have turned to advanced control techniques of RL in conjunction with Photovoltaic Systems (PV). The coupling of RL with PVS offers a synergistic approach, leveraging solar energy to charge EVs efficiently while minimizing costs and environmental impact. Huang (2020) uses a hybrid wind-solar storage system in conjunction with a Deep Reinforcement Learning (DRL) control technique to improve prediction quality and optimization while addressing the uncertainties related to the production of renewable energy. This study shows that DRL algorithms are effective in controlling uncertainties related to renewable energy, even when there is a lack of high-dimensional data. This helps power systems utilize renewable energy sources more consistently and profitably. Reinforcement Learning (RL) in solar battery storage to maximize energy use and operational effectiveness in these systems is utilized (Härtel, 2023). The study enhances photovoltaic battery storage systems' performance by utilizing recurrent neural networks (RNNs) and proximal policy optimization (PPO). In addition, there are even new opportunities for demand management and the

safe integration of renewable energy sources thanks to Mololoth, V. K., et al.'s (2023) investigation on the synergistic potential of blockchain and machine learning for upcoming smart grids.

4 Methodology

4.1 Problem Statement

This thesis concerns how electric vehicles are charged at charging stations that have access to renewable energy and can also purchase electricity from the grid. The renewable energy for these stations is provided by solar panels, which can only generate energy during certain hours of the day. A portion of the generated energy is also used for the daily consumption of the station itself, referred to as the load. Therefore, only a part of the generated energy can be sold for charging electric vehicles. Depending on the immediate needs of each station, more electricity can be purchased from the grid if needed for charging. Since purchasing electricity from the grid is costly, the goal of this issue is to charge the vehicles in a way that incurs the lowest cost.

In this scenario, the simulation leverages authentic data on tariffs, the production of renewable energy (specifically from solar panels), and the grid load for each station, sourced from the actual energy consumption and solar power generation records of households. These households are deemed prosumers, reflecting their dual role as both energy consumers and producers within the network. This data also includes the energy usage of electric vehicles associated with these households, based on a study conducted in Portugal (Faia, 2021).

In the problem of charging electric vehicles, the task is to reduce the consumption costs of the stations. While a desired output cannot be specified for each vehicle, for example, it's not possible to state that a vehicle should receive a specific amount of charge at a particular time of day to minimize costs. Solving such problems requires learning a procedure that decides on the charging of each vehicle at every step, considering the variable conditions of the problem. For instance, the moment each vehicle arrives at the station, given the energy level in the device's battery, the instantaneous energy production from the solar panels, and the cost of purchasing energy from the power grid, a decision to charge or not charge a vehicle is made.

Given the characteristics mentioned regarding the electric vehicle charging problem, reinforcement learning is the most suitable method for finding the optimal solution to the problem.

4.2 Reinforcement Learning

Reinforcement learning (RL) is a learning method that is about an action that is supposed to maximize the reward. It differs from supervised learning in that neither the dataset nor the correct decisions are labeled by an external supervisor. Moreover, the purpose of RL is to find the best possible actions that are not defined by learning and exploring based on the environment which the agent is interacting in. It is also not quite similar to unsupervised learning although it is sometimes classified within this category. The reason is that RL is not trying to find a hidden structure within the unlabeled data but it is used for maximizing the reward that the agent is trying to achieve (Sutton & Barto, 2018).

Unlike supervised algorithms, which define a specific desired output for each input, there exists a category of problems where a precise output cannot be established, and the only performance criterion for an agent in these environments is a descriptive quantitative value known as a reward (Co-Reyes et al., 2020). In other words, the reward is the environment's response to the agent's actions. In many environments, a maximum or minimum for the reward cannot be defined, and learning in these problems can be modeled as maximizing the cumulative reward received. Regarding the termination of an episode in a simulated environment, environments can be divided into continuous and episodic. In episodic environments, one or more specific states are considered as the agent's terminal state in the environment, whereas in continuous environments, no terminal state is defined for the agent. A complete cycle of the agent's performance in an episodic environment is called an episode. A cycle in the problem of charging electric vehicles can be simplified to be confined to a single day.

Overall, addressing this problem requires a simulation environment and an agent tasked with optimizing the objectives. In reinforcement learning problems, a simulated environment and a reinforcement learning agent interact sequentially with each other. At each step, considering the current state of the environment and the values of the environment's parameters, the agent selects an appropriate action. The implementation of the mentioned action then changes the current state of the environment and also causes the environment to return a quantitative value indicating the effectiveness of the action performed (Sutton & Barto, 2018). The goal of reinforcement learning is to find actions that create the most significant effect in the environment, in other words, to maximize the cumulative value returned from the environment. In this thesis, many agents are involved in a dynamic and uncertain environment and reinforcement learning is utilized as a trial-and-error method of machine learning algorithm that can interact with this environment and make sequential decisions to learn how to maximize the reward.

In the next step, the simulation environment will be outlined. Then, various reinforcement learning algorithms will be detailed, assessing the strengths of each. The conclusion will present the results derived from each algorithm, showcased through graphs and analyses. To this end, it is necessary to model the reinforcement learning variables in the problem at hand and reduce the charging costs of the vehicles using reinforcement learning algorithms. In the modeled scenario, the function that selects the appropriate action at each step is called a policy, and the best-learned policy is referred to as the optimal policy.

4.3 Modeling & Formulation

Formulation of such problems in RL is done by the Markov decision process (MDP) which is suitable for a random probability pattern that is difficult to predict. In scenarios when decisions are taken randomly or controlled by an agent, MDP can be used which is a mathematical framework to make the decision-making process of the stochastic process. In MDP models, 4 key elements should be defined in a pre-defined environment, namely states, actions, transition probabilities, and reward. This environment is supposed to act

as the artificial world that is a close representation of the real world, and the agent can interact within this environment to achieve its goal over time (Karatzinis, et al., 2022).

4.3.1 RL Environment

The environment is crucial to reinforcement learning (RL) because it interacts with the learning agent by supplying the scenarios that the agent must navigate. under essence, the environment is a model or simulation of the real world, or a particular component of it, that establishes the parameters under which the agent must carry out its duties. This environment can range from virtual simulations for training autonomous vehicles to game settings like chess, or more complex scenarios such as managing energy distribution in a power grid (Narvekar et al., 2020). Environments can be classified into episodic, where interactions are divided into separate episodes with clear endpoints, or continuous, where the interaction goes on indefinitely without predefined ends.

Each RL environment must follow certain rules and characteristics to be acceptable and readable as an environment. Here are the characteristics of a general RL environment in summary (Moussaoui, Akkad, and Benslimane, 2023):

- **State Space (S):** The environment defines the state space, which includes all possible situations the agent might encounter. The complexity of an environment often correlates with the size and diversity of its state space.
- **Action Space (A):** This covers every course of action the agent may pursue. Since the action space's design dictates the extent of the agent's interactions with the surroundings, it can have a substantial impact on the agent's learning and performance.
- **Reward Function (R):** The environment determines the rewards, which are responses to the agent's actions aimed at achieving specific goals. The reward structure is crucial as it guides the learning process by signaling to the agent which actions are beneficial toward achieving its objectives.
- **Transition Function (T):** This refers to how the environment changes in response to the agent's actions. The dynamics can be deterministic (predictable outcomes)

or stochastic (random elements influencing the outcomes), affecting the agent's strategy for learning and decision-making.

- **Policy (π):** A policy is a strategy used by the agent, defined as a mapping from states to actions. The policy determines the action that an agent will take in a given state.
- **Episode:** Many RL issues involve exchanges that can be divided into smaller units called episodes. Every episode has a starting state and a terminal state at the conclusion. Tasks without a clear endpoint in the engagement are known as non-episodic (or ongoing).
- **Discount Factor (γ):** The impact of future benefits is determined by the discount factor, which has a value between 0 and 1. When a factor is near 1, the agent is considered far-sighted because it values benefits that are further in the future, whereas a factor of 0 makes the agent short-sighted since it only considers current rewards.
- **Objective Function:** The objective function in reinforcement learning typically involves maximizing the cumulative reward, which could be the sum of rewards in the case of a finite horizon, or the discounted sum of rewards in the case of an infinite horizon.

To provide an overview of the methodology in this thesis, a detailed flowchart has been made to illustrate the implementation processes of the proposed framework (Figure 1). This flowchart encapsulates the key steps, ranging from the initialization of the environment and calculating main parameters to implementing reinforcement learning (RL) policies and assessing performance.

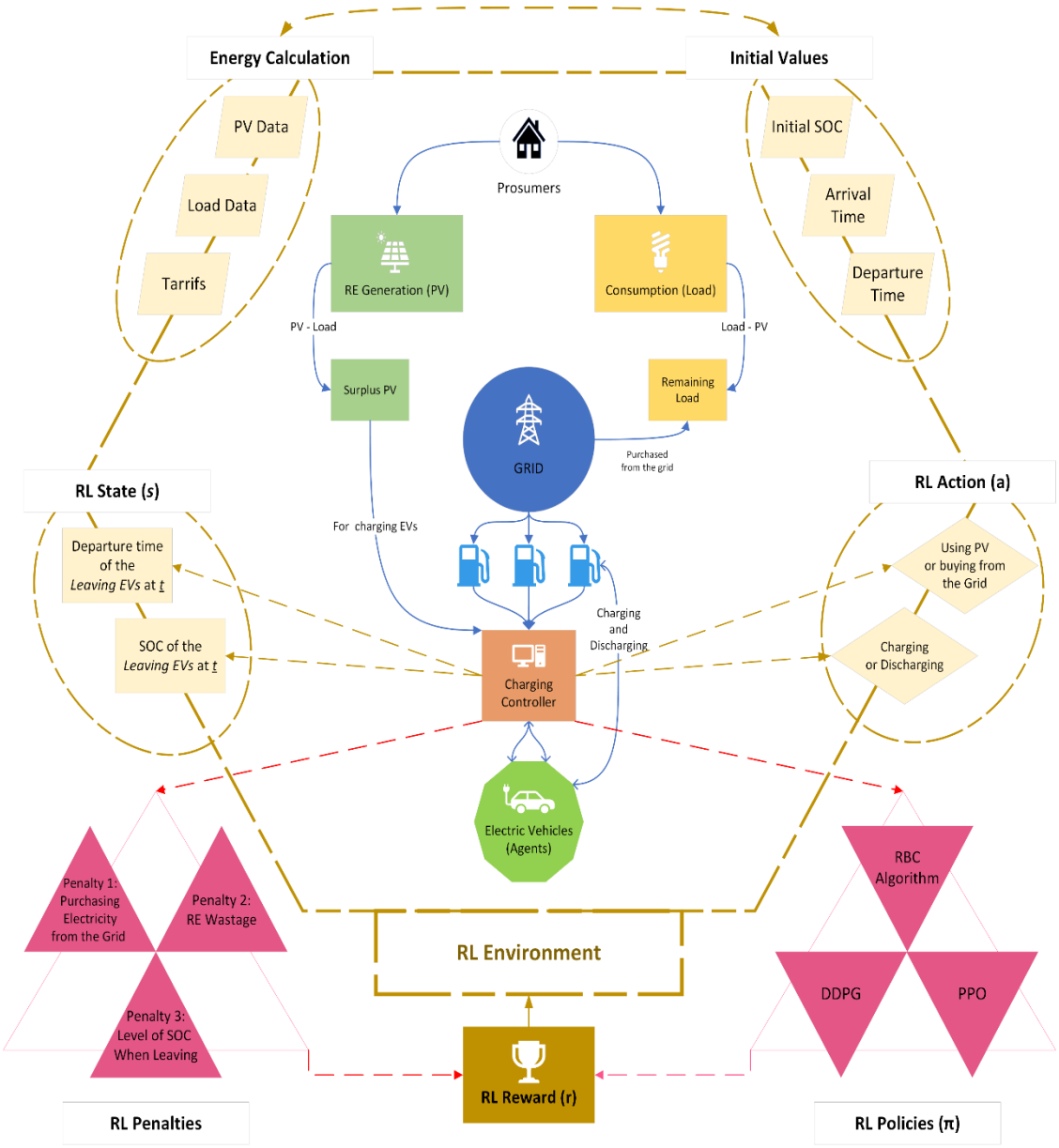

Figure 1. Methodology Workflow: RL-Based EV Charging Framework

4.3.2 Markov decision process (MDP)

As previously stated, a large number of RL issues can be examined using the MDP framework. An environment in an MDP is made of several unique states, and an agent is only ever allowed to be in one of these states at a time. The agent in a Markov Decision Process can also select from a range of possible actions at any given time. Any action performed could cause the agent's status inside the state space to change. Since the MDP

is a stochastic process, a probabilistic transition table governs how it will change states when an action is performed. This model works very well in situations when decisions have to be made sequentially under uncertainty, and each option could have a different effect depending on the system's current state. The construction of strategies that can adapt and function well under a variety of settings is made possible by the stochastic character of MDPs, which represent the randomness and unpredictability of real-world environments (Narvekar et al., 2020).

The mathematical formulas and the principles of MDP in RL are explained accordingly:

Formula (1) represents the transition probability in the context of a MDP, which is a core concept in Reinforcement Learning (RL) (Sutton and Barto, 2018). $P_a(s, \acute{s})$ denotes the probability of transitioning from the state given that action a is taken. In an MDP, this transition probability is crucial because it encapsulates the dynamics of the environment, describing how the environment responds at a future time $t + 1$ based on the agent's current state and action at time t.

$$P_a(s, \acute{s}) = \Pr\left(s_{t+1} = \acute{s} | s_t = s, a_t = a\right) \quad (1)$$

Every action that is taken in a MDP that causes a state to change is referred to as a "step." Every action the agent does in the environment results in a number value being returned to them as a reward. This indicates how effective their actions were. The amount of the reward is determined by the agent's current condition, the action taken, and the agent's subsequent state following the action, which is indicated as $R_a(s.\acute{s})$.

It is essential to remember that the environment alone calculates the reward; the agent is unaware of the process or the rationale behind the value that is rewarded. As a result, the knowledge that the agent has to learn consists of its current state, the action that it chose, and the reward that it received. This configuration emphasizes the reinforcement learning principle of learning from interaction rather than predefined rules by

guaranteeing that the agent's learning is entirely dependent on the input it receives from the environment. By analyzing which behaviors result in greater rewards in particular states over time, the agent optimizes its behavior and adjusts and improves its plan based on actual data rather than theoretical models. Reinforcement learning systems use this ongoing process of interaction and adjustment as the core learning mechanism.

To utilize the MDP framework for modeling our problem, a precise definition of the parameters and variables involved in this process is needed. An MDP can be represented by a tuple (s, a, p, r), where 's' represents the states, 'a' represents the actions, 'p' represents the transition probabilities between states, and 'r' represents the rewards. The state represents the condition of an agent about its environment and can encompass any number of variables. In other words, the state can be considered the agent's perception of its surrounding environment. For example, for an autonomous vehicle, all the input values from the sensors form the vehicle's understanding of its surrounding environment. Thus, the decision-making of an agent to choose an action depends solely on the state it is in. Additionally, the input to a policy is the agent's state. For a policy to improve and be able to choose better actions in each state, it needs to maximize the rewards obtained. Therefore, the improvement of a policy entirely depends on the reward an agent receives for a particular action in a specific state. Hence, having values of state, action, and reward can enable training an agent within an environment.

In scenarios where the environment is not fully known to us, the transition probabilities between states are not available. Such environments are called Partially Observable Markov Decision Processes (POMDPs), and finding the optimal policy in these environments is more challenging than in MDP environments. Our evidence for learning in POMDP environments would be represented by a tuple (s, a, s', r, d), where s' represents the next state and d indicates the end of a complete cycle of the environment's execution. In both MDP and POMDP settings, the key to effective reinforcement learning lies in accurately capturing and utilizing the dynamics of state transitions and reward mechanisms to iteratively refine the decision-making policy. This iterative learning process focuses on

maximizing the cumulative rewards across episodes, adapting the policy based on feedback from the environment to optimize outcomes under given constraints and uncertainties.

Formula (2) represents the optimization of an operator in the environment according to Bellman equations explained in Sutton and Barto (2018):

$$V^{\pi}(s) = \mathbb{E}_{\pi}[\sum_{t=0}^{\infty} \gamma^{t} R(s_t, a_t) | s_0 = s] \quad (2)$$

Here, s represents the state, a represents the action, and R is the function calculating the reward. Additionally, γ is the discount factor, and π represents the policy adopted by the agent. The above formula shows the value of each state according to the chosen policy. Since our environment is a POMDP, instead of calculating V values, formula (3) has to be solved to find the optimal policy (Qiu et al., 2020).

$$Q^{\pi}(s.a) = \mathbb{E}_{\pi}[\sum_{t=0}^{\infty} \gamma^{t} R(s_t.a_t) | s_0 = s.a_0 = a] \quad (3)$$

The $Q - funtcion$, also known as the action-value function, quantifies the expected return of taking an action a in a state s and thereafter following a policy π. Having a function that provides a good approximation of the Q value, at each step, given the state, an action that produces the highest Q value can be selected as the optimal action for that particular state. Qiu (et al., 2020) rewrites the above equation as follows (formula (4)):

$$Q^{\pi}(s_0, a_0) = \mathbb{E}_{\pi}[R_a(s_0, s_1) + \sum_{t=1}^{\infty} \gamma^{t} R(s_t, a_t) | s_1 = s] = \mathbb{E}_{\pi}[R(s_0, a_0) + V^{\pi}(s_1)] \quad (4)$$

Where $\mathbb{E}_{\pi}$ denotes the expected value given that the agent follows policy π after taking action a in state s. $\sum_{t=1}^{\infty} \gamma^{t} R(s_t, a_t) | s_1 = s$ represents the sum of discounted rewards received over the future, starting from state s and action a at time $t = 0$. γ is the discount factor and R is the reward received after executing action a_t in state s_t.

If the above equation is modified as follows (formula (5)), the calculation of the Q value will not depend on the transition probabilities between states. This method is known as Q-learning.

$$Q(s,a) = R(s,a) + \gamma \max_{a} Q(\acute{s},a) \tag{5}$$

Where s' is the next state that the agent occupies as a result of performing action a in state s. To learn the Q function, formula (6) can be used (Arwa & Folly, 2021):

$$Q^{new}(s,a) = (1-\alpha)Q^{old}(s,a) + \alpha\left(R(s,a) + \gamma \max_{a} Q(\acute{s},a)\right) \tag{6}$$

Where α is the learning rate of the algorithm. As shown in the equation, at each step, the Q value for a pair of state and action sums a fraction of its current value and a fraction of the predicted value. Choosing an appropriate learning rate is crucial in effectively learning the agent's behavior.

4.4 RL Policy

Given that the electricity at the station is supplied from two different sources at varying costs, the decision on how much charge each vehicle should receive during each visit to the station can significantly affect the costs incurred. Therefore, selecting an appropriate policy for the amount of charging for each vehicle can be envisioned as an optimization problem aimed at enhancing efficiency and reducing costs. By considering the condition of each vehicle upon entering and exiting a station, as well as the source of the charge for that vehicle, a numerical metric can be defined to evaluate the performance of the station about the specific vehicle.

For solving the defined RL problem, three policies have been used. PPO and DDPG are from the category of reinforcement learning algorithms, and the RBC algorithm is from the category of rule-based algorithms.

4.4.1 PPO Policy

The PPO policy is one of the most commonly used algorithms in reinforcement learning and has become the standard algorithm for RL in many companies. This policy prevents sudden changes in policy and, unlike the DDPG policy, is simple to implement. It also has high stability, and changes in hyperparameters do not significantly affect the algorithm's performance, ensuring it consistently performs well (Schulman et al, 2017). Additionally, in terms of data needed for training, PPO requires fewer data points than other reinforcement learning algorithms, which reduces the learning time.

Specifically, this algorithm tries to prevent sudden changes in policy so that the updated policy deviates only slightly from the current policy, common in off-policy methods. Also, changes in the policy are clipped by a predetermined limit. It is noteworthy that in this algorithm (Cheng et al., 2018), instead of predicting the Q-value, the advantage value is used (formula 16).

$$E^{clip}(\theta) = \hat{\mathbb{E}}_t\left[\min\left(r_t(\theta)\right)\hat{A}_t, clip(r_t(\theta), 1-\varepsilon, 1+\varepsilon)\right] \quad (16)$$

Where θ represents the policy, r_t is the likelihood ratio of events under the new policy relative to the old, and $\hat{A}_t$ represents the advantage.

$$A^{\pi}(s,a) = Q^{\pi}(s,a) - V^{\pi}(s) \quad (17)$$

4.4.2 DDPG Policy

Zhang et al. (2019) present this policy as a combination of Q-learning and policy-learning algorithms. It consists of two separate steps, one for predicting the Q-value and another for deriving the policy from Q-values. This policy falls under the Actor-critic category of algorithms. This category, having an actor, possesses the strengths of policy-based algorithms, and also having a critic, holds the strengths of value-based algorithms. Therefore,

for its implementation, two separate deep neural networks can be used for learning Q-values and policy.

DDPG is especially suitable for environments where the dimension of state or action is high. As an off-policy learning algorithm, it requires fewer data points for learning. Moreover, off-policy learning ensures that the learning and action selection are from two separate networks, leading to more stability for the actor-network (Zhang et al., 2019).

4.4.3 Rule-Based Controller

Rule-based algorithms are typically used in environments where sufficient knowledge of the environment exists, and a set of predefined rules can be written for them (Karmaker et al., 2023). The implementation of the mentioned rules is very straightforward and usually stems from complete human knowledge of an environment and its governing policies. In this method, learning does not occur, and the vehicle charging is determined by the time it detaches from the charger. If the vehicle leaves the station in less than three hours, it is charged at full capacity; otherwise, depending on the availability of renewable energy, the station proceeds to charge the vehicle.

$$action_t^i = \begin{cases} 1 & Tleave_t^i \leq 3 \\ \frac{(G_t + G_{t+1})}{2} & Tleave_t^i \leq 3 \end{cases} \quad (18)$$

Where G_t is the first element of the states, representing the battery energy level.
Table 1 highlights the key steps taken in this thesis for implementing the DDPG, PPO, and RBC policies.

Table 1. The Pseudocode of the RL Policies and Final Evaluation

Algorithm 1 RL Policies and Final Evaluation
1: Initialize DDPG/PPO/RBC Environment
2: import necessary libraries (**`gym`, `stable_baselines3`, `RBC`, etc**.)
3: Create directories for saving models and logs

DDPG & PPO Policies
4: Create and Train Model
5: - Initialize custom environment using `**gym.make**`
6: - Optionally set up action noise for DDPG
7: - Create model (`DDPG` or `PPO`) with `**MlpPolicy**` and custom environment
8: - Train and save model periodically

9: Evaluate and Use Model
10: - Evaluate the model using `**evaluate_policy**`
11: - Test trained agent by predicting and stepping through the environment

RBC Algorithm
1: RBC Class
2: - Create `**select_action**` function
3: - Determine departure time and solar radiation
4: - Set action based on departure time and radiation
5: - Return selected action

6: Run RBC Algorithm
7: - Parse arguments for custom environment
8: - Initialize custom environment and reset state
9: - Loop until done:
10: - Select action using RBC
11: - Step through environment and update state
12: - Track rewards

Evaluation and Comparison
1: Evaluate Models
2: - Loop through evaluation episodes
3: - **for** each episode:
4: - Evaluate DDPG model and record rewards
5: - Evaluate PPO model and record rewards
6: - Evaluate RBC algorithm and record rewards

7: **Plot and Show Results**
8: - Calculate mean rewards for DDPG, PPO, and RBC by '**np.mean(final_reward)**'
9: - '**pl.plot(final_reward)**' Plot and display reward comparison graph

5 Numerical Studies

5.1 Daily Energy Dynamics

In this environment, there are 40 electric vehicles and 50 prosumers. Each household is equipped with solar panels that produce a specific amount of electricity during the day depending on weather conditions. Each station is also connected to the power grid and can purchase electricity from it if there is a shortage in production and use it for charging vehicles. The load imposed on the network for each of these nodes and the amount of electricity produced by each of these prosumers is pre-defined and available in Faia (et al., 2021) – specified in the PV.csv and Load.csv files. The time step defined in this environment is 15 minutes, and the total simulation time is 24 hours, meaning that 96 steps are carried out in this simulation. The amount of electricity consumed and produced by a prosumer helps to better understand the environment and the performance of each agent.

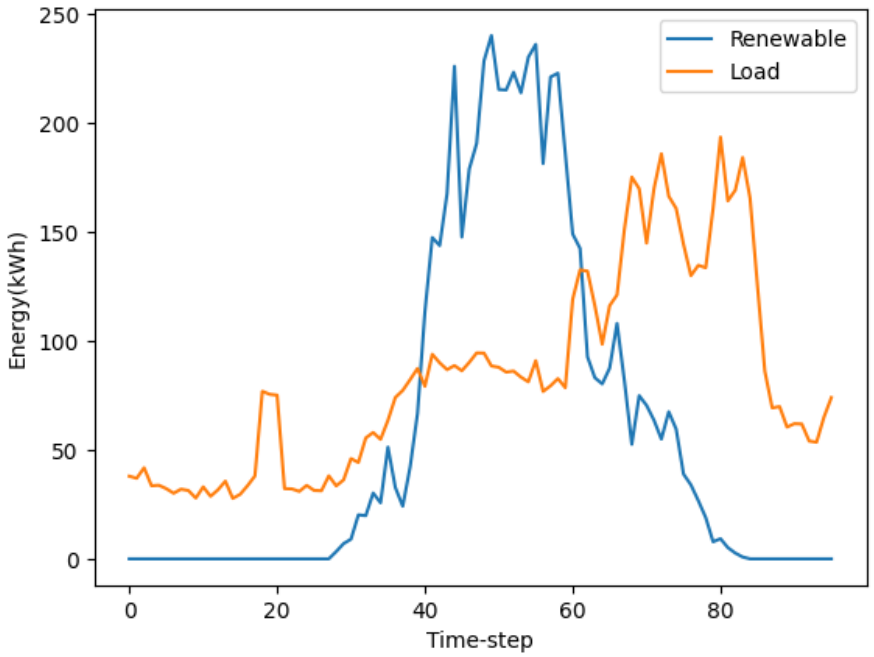

Figure 2. Consumption (Load) and Generated Electricity by RE (Renewable)

Figure 2 illustrates the daily profile of energy consumption and production of the mentioned 50 prosumers. The vertical axis measures the energy in kilowatt-hours (kWh), depicting the household's energy dynamics. Upon examination of Figure 2, it is evident that renewable energy production—denoted by the blue line—experiences a significant increase during the central daylight hours. This pattern is indicative of solar power generation, which aligns with the peak insolation periods typically observed in middle latitudes. The load curve, depicted by the orange line, illustrates daily fluctuations in energy consumption. It reveals a rise in energy use during the early morning as people prepare for work, followed by a more significant increase in the evening, which lasts until about 9 o'clock. This evening peak likely results from household members returning home and using various appliances simultaneously.

A critical observation from the graph is the intersection points of the load and renewable lines, which divide the transition between net consumption and net production phases. During the early morning and late evening hours (00:00-05:00 and 20:00-23:45), the household's energy consumption surpasses its production, thereby necessitating the procurement of additional electricity from the grid.

Conversely, the middle of the day (07:30-17:30) is characterized by an excess of renewable energy production over consumption. This surplus provides an opportunity for the household to act as an energy provider, potentially channeling excess electricity for storage or ancillary uses such as charging electric vehicles. The profile also reveals potential for optimization. Energy storage solutions could be employed during the production surplus to alleviate the demand during deficit periods. This would enhance energy independence and contribute to a more balanced and self-sustaining household energy system.

5.2 Solar Energy Surplus

The strength of the decision-making policy is when the renewable energy is more than the energy consumed by the prosumers. Now, to calculate the amount of energy a

prosumer can use to charge electric vehicles, the remaining renewable energy at each moment must be calculated.

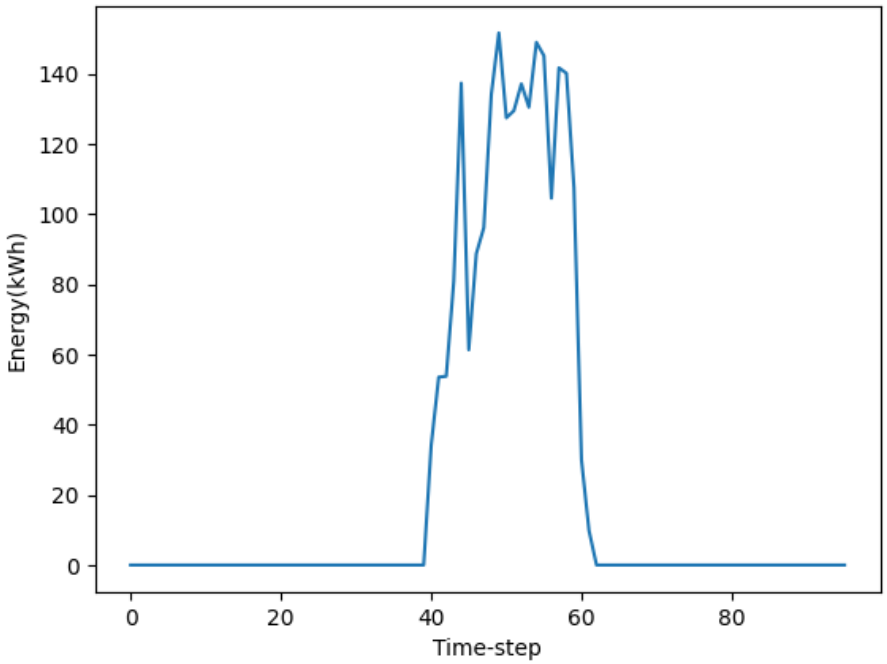

Figure 3. The Remaining RE after consumption

Figure 3 provides a visual representation of the residual renewable energy available at a prosumer household, postulated after the subtraction of consumed load from the gross renewable production over 24 hours. The peak suggests that the highest availability of surplus renewable energy occurs in the central daylight hours, providing the most opportune time for energy-intensive activities such as charging electric vehicles (EVs) in this case.

The surplus initiates at time-step 05:00, peaks around time-step 12:30, and wanes by time-step 17:30. The data presented in the chart implies that there exists a window between these intervals where the generated solar energy exceeds the household's consumption demands. Given that the total number of electric vehicles is 40, it can be inferred that the household's energy management system must prioritize vehicle charging schedules to align with the surplus energy availability. The variance in vehicle activity

over the day affects the charging demand. Hence, strategic scheduling is paramount to ensure that the vehicles are adequately charged when most active while optimizing the use of surplus renewable energy.

5.3 Tariffs

According to Faia (et al., 2021), Table 2 demonstrates the price of electricity with six different tariffs at different times of the day, with values stored in the Tariff.csv file as well. To streamline the simulation, a consistent tariff, “FIXED T1”, which is a fixed tariff is applied for the cost of electricity procured from the grid.

Table 2. The price of electricity at different times with six kinds of Tariffs

TIME	PERIODS	FIXED T1	HOURLY T1	FIXED T2	HOURLY T2	TRI-HOURLY T2
00:00:00	Off-Peak	0.1456	0.0923	0.1548	0.0968	0.0942
00:15:00	Off-Peak	0.1456	0.0923	0.1548	0.0968	0.0942
00:30:00	Off-Peak	0.1456	0.0923	0.1548	0.0968	0.0942
00:45:00	Off-Peak	0.1456	0.0923	0.1548	0.0968	0.0942
01:00:00	Off-Peak	0.1456	0.0923	0.1548	0.0968	0.0942
01:15:00	Off-Peak	0.1456	0.0923	0.1548	0.0968	0.0942
01:30:00	Off-Peak	0.1456	0.0923	0.1548	0.0968	0.0942
01:45:00	Off-Peak	0.1456	0.0923	0.1548	0.0968	0.0942
02:00:00	Off-Peak	0.1456	0.0923	0.1548	0.0968	0.0942
02:15:00	Off-Peak	0.1456	0.0923	0.1548	0.0968	0.0942
02:30:00	Off-Peak	0.1456	0.0923	0.1548	0.0968	0.0942
02:45:00	Off-Peak	0.1456	0.0923	0.1548	0.0968	0.0942
03:00:00	Off-Peak	0.1456	0.0923	0.1548	0.0968	0.0942
03:15:00	Off-Peak	0.1456	0.0923	0.1548	0.0968	0.0942
03:30:00	Off-Peak	0.1456	0.0923	0.1548	0.0968	0.0942
03:45:00	Off-Peak	0.1456	0.0923	0.1548	0.0968	0.0942
04:00:00	Off-Peak	0.1456	0.0923	0.1548	0.0968	0.0942
04:15:00	Off-Peak	0.1456	0.0923	0.1548	0.0968	0.0942
04:30:00	Off-Peak	0.1456	0.0923	0.1548	0.0968	0.0942

04:45:00	Off-Peak	0.1456	0.0923	0.1548	0.0968	0.0942
05:00:00	Off-Peak	0.1456	0.0923	0.1548	0.0968	0.0942
05:15:00	Off-Peak	0.1456	0.0923	0.1548	0.0968	0.0942
05:30:00	Off-Peak	0.1456	0.0923	0.1548	0.0968	0.0942
05:45:00	Off-Peak	0.1456	0.0923	0.1548	0.0968	0.0942
06:00:00	Off-Peak	0.1456	0.0923	0.1548	0.0968	0.0942
06:15:00	Off-Peak	0.1456	0.0923	0.1548	0.0968	0.0942
06:30:00	Off-Peak	0.1456	0.0923	0.1548	0.0968	0.0942
06:45:00	Off-Peak	0.1456	0.0923	0.1548	0.0968	0.0942
07:00:00	Off-Peak	0.1456	0.0923	0.1548	0.0968	0.0942
07:15:00	Off-Peak	0.1456	0.0923	0.1548	0.0968	0.0942
07:30:00	Off-Peak	0.1456	0.0923	0.1548	0.0968	0.0942
07:45:00	Off-Peak	0.1456	0.0923	0.1548	0.0968	0.0942
08:00:00	Off-Peak	0.1456	0.0923	0.1548	0.0968	0.0942
08:15:00	Peak	0.1456	0.1833	0.1548	0.2027	0.1715
08:30:00	Peak	0.1456	0.1833	0.1548	0.2027	0.1715
08:45:00	Peak	0.1456	0.1833	0.1548	0.2027	0.1715
09:00:00	Peak	0.1456	0.1833	0.1548	0.2027	0.1715
09:15:00	On-Peak	0.1456	0.1833	0.1548	0.2027	0.2942
09:30:00	On-Peak	0.1456	0.1833	0.1548	0.2027	0.2942
09:45:00	On-Peak	0.1456	0.1833	0.1548	0.2027	0.2942
10:00:00	On-Peak	0.1456	0.1833	0.1548	0.2027	0.2942
10:15:00	On-Peak	0.1456	0.1833	0.1548	0.2027	0.2942
10:30:00	On-Peak	0.1456	0.1833	0.1548	0.2027	0.2942
10:45:00	Peak	0.1456	0.1833	0.1548	0.2027	0.1715
11:00:00	Peak	0.1456	0.1833	0.1548	0.2027	0.1715
11:15:00	Peak	0.1456	0.1833	0.1548	0.2027	0.1715
11:30:00	Peak	0.1456	0.1833	0.1548	0.2027	0.1715
11:45:00	Peak	0.1456	0.1833	0.1548	0.2027	0.1715
12:00:00	Peak	0.1456	0.1833	0.1548	0.2027	0.1715
12:15:00	Peak	0.1456	0.1833	0.1548	0.2027	0.1715
12:30:00	Peak	0.1456	0.1833	0.1548	0.2027	0.1715
12:45:00	Peak	0.1456	0.1833	0.1548	0.2027	0.1715

13:00:00	Peak	0.1456	0.1833	0.1548	0.2027	0.1715
13:15:00	Peak	0.1456	0.1833	0.1548	0.2027	0.1715
13:30:00	Peak	0.1456	0.1833	0.1548	0.2027	0.1715
13:45:00	Peak	0.1456	0.1833	0.1548	0.2027	0.1715
14:00:00	Peak	0.1456	0.1833	0.1548	0.2027	0.1715
14:15:00	Peak	0.1456	0.1833	0.1548	0.2027	0.1715
14:30:00	Peak	0.1456	0.1833	0.1548	0.2027	0.1715
14:45:00	Peak	0.1456	0.1833	0.1548	0.2027	0.1715
15:00:00	Peak	0.1456	0.1833	0.1548	0.2027	0.1715
15:15:00	Peak	0.1456	0.1833	0.1548	0.2027	0.1715
15:30:00	Peak	0.1456	0.1833	0.1548	0.2027	0.1715
15:45:00	Peak	0.1456	0.1833	0.1548	0.2027	0.1715
16:00:00	Peak	0.1456	0.1833	0.1548	0.2027	0.1715
16:15:00	On-Peak	0.1456	0.1833	0.1548	0.2027	0.2942
16:30:00	On-Peak	0.1456	0.1833	0.1548	0.2027	0.2942
16:45:00	On-Peak	0.1456	0.1833	0.1548	0.2027	0.2942
17:00:00	On-Peak	0.1456	0.1833	0.1548	0.2027	0.2942
17:15:00	On-Peak	0.1456	0.1833	0.1548	0.2027	0.2942
17:30:00	On-Peak	0.1456	0.1833	0.1548	0.2027	0.2942
17:45:00	On-Peak	0.1456	0.1833	0.1548	0.2027	0.2942
18:00:00	On-Peak	0.1456	0.1833	0.1548	0.2027	0.2942
18:15:00	On-Peak	0.1456	0.1833	0.1548	0.2027	0.2942
18:30:00	On-Peak	0.1456	0.1833	0.1548	0.2027	0.2942
18:45:00	On-Peak	0.1456	0.1833	0.1548	0.2027	0.2942
19:00:00	On-Peak	0.1456	0.1833	0.1548	0.2027	0.2942
19:15:00	On-Peak	0.1456	0.1833	0.1548	0.2027	0.2942
19:30:00	On-Peak	0.1456	0.1833	0.1548	0.2027	0.2942
19:45:00	On-Peak	0.1456	0.1833	0.1548	0.2027	0.2942
20:00:00	On-Peak	0.1456	0.1833	0.1548	0.2027	0.2942
20:15:00	On-Peak	0.1456	0.1833	0.1548	0.2027	0.2942
20:30:00	On-Peak	0.1456	0.1833	0.1548	0.2027	0.2942
20:45:00	Peak	0.1456	0.1833	0.1548	0.2027	0.1715
21:00:00	Peak	0.1456	0.1833	0.1548	0.2027	0.1715

21:15:00	Peak	0.1456	0.1833	0.1548	0.2027	0.1715
21:30:00	Peak	0.1456	0.1833	0.1548	0.2027	0.1715
21:45:00	Peak	0.1456	0.1833	0.1548	0.2027	0.1715
22:00:00	Peak	0.1456	0.1833	0.1548	0.2027	0.1715
22:15:00	Off-Peak	0.1456	0.0923	0.1548	0.0968	0.1715
22:30:00	Off-Peak	0.1456	0.0923	0.1548	0.0968	0.0942
22:45:00	Off-Peak	0.1456	0.0923	0.1548	0.0968	0.0942
23:00:00	Off-Peak	0.1456	0.0923	0.1548	0.0968	0.0942
23:15:00	Off-Peak	0.1456	0.0923	0.1548	0.0968	0.0942
23:30:00	Off-Peak	0.1456	0.0923	0.1548	0.0968	0.0942
23:45:00	Off-Peak	0.1456	0.0923	0.1548	0.0968	0.0942

In this thesis, it's assumed that all electric vehicles share identical specifications regarding battery capacity (45 kWh), the period required for a full charge, the charging and discharging rate (8 kWh), and the efficiency of the charging and discharging process (90%). Moreover, for each vehicle, it must have access to the amount of charge stored in the battery in the first departure, which is saved in the SOC.csv file.

The variables in this problem are the battery charge levels of each vehicle. At any given time, a vehicle may stop at one of the charging stations. Therefore, in the simulation environment, it is needed to provide the time of arrival and departure from the station for each vehicle, which are saved in the Evolution.csv file and can be observed in Figure 4.

5.4 EVs Presence during the day

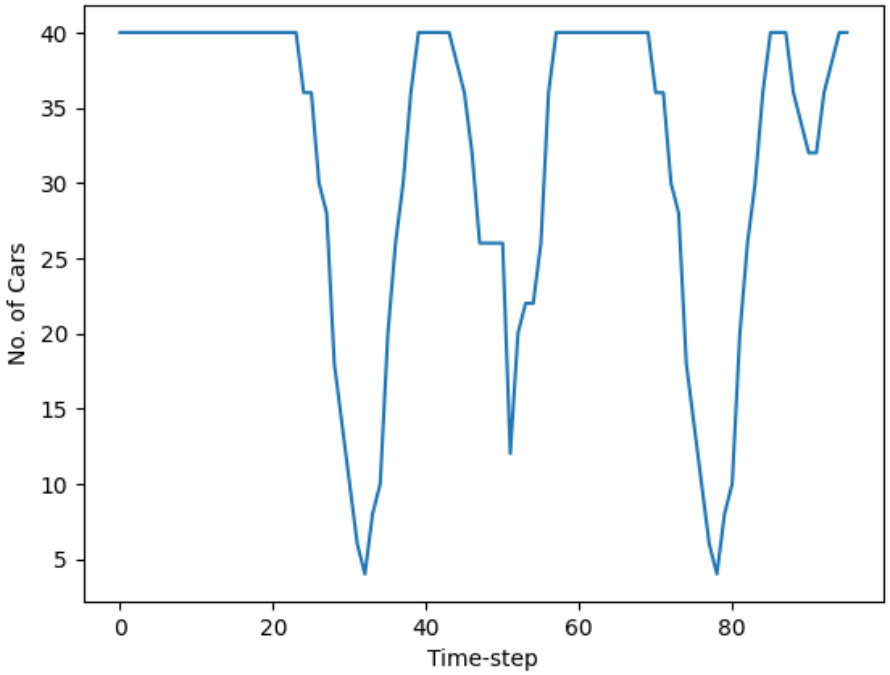

Figure 4. Present cars over a 24-hour period

Figure 4 depicts the fluctuation in the quantity of electric vehicles (EVs) present within the simulated environment over the specified day. The graph shows sharp variations in the number of vehicles, ranging from as few as 10 to the maximum fleet size of 40. Notably, there are pronounced dips in the vehicle count, suggesting times when a significant portion of the fleet is not present—presumably when the vehicles are out for usage.

Tracking the presence of each car at specific times, indicated by t, is crucial to precisely calculating penalties in the algorithm and to better understand how EVs move. Creating $state$ part of the Reinforcement Learning (RL) environment depends on this tracking. Based on this presence data, the EVs that depart within the next 15 minutes are extracted for the $action$ phase. Also, the SOC of an EV that is expected to depart in the next fifteen minutes will be extracted as well for additional analysis and reward calculations.

Each EV is an agent that enters a station with a specific amount of charge and either increases or decreases its battery charge. At the start of each simulation, when the environment is reinitialized, the values of load, amount of produced energy, and electricity prices at each step of the algorithm are read from the simulation files. Table 3 shows the core functions of the environment, such as initialization, step function, resetting, generating observations, and setting the seed.

Table 3. The Pseudocode of the RL Environment

Algorithm 2 Reinforcement Learning Environment
Define Environment
Create an environment class using `**gym.Env**`
Initialize EV parameters and observation/action spaces
EV Charging Environment (Gym Environment)
Initialization '(__init__)':
Initialize key variables (`**number_of_cars**`, `**number_of_days**`, etc.)
Set EV parameters (e.g., `**EV_capacity**`, `**charging_rate**`)
Define action and observation spaces using `**spaces.Box**`
Initialize random seed and other variables
Step Function (`**step**`):
Simulate actions using `**Simulate_Actions3.RL_Actions**`
Update and record relevant metrics (`**reward_History`, `Grid_trade_Evol**`, etc.)
Check if the episode is done and save results using `**savemat**`
Reset Function (`**reset**`):
Reset environment state and load initial values using `**loadmat**`
Initialize necessary variables **(`timestep`, `done`, etc.**)
Get Observations (`**get_obs**`):
Calculate the state of the environment using `**Simulate_Station3.RL_States**`
Return the current state as a numpy array
Seed Function (`seed`):
Set random seed for reproducibility using `**seeding.np_random**`
Close **(`close`):**

Given the problem definition, reinforcement learning algorithms can be used to optimize decision-making about the manner and amount of charging for the vehicles. For this purpose, a clear definition of state, reward, and action is needed:

• **State:** The state of the environment at any moment is a vector twice the length of the number of vehicles presented in the simulation, which includes the amount of charge stored in each vehicle's battery and the movement time of each vehicle.

• **Reward:** The defined reward consists of three parts that determine the total reward at each stage. The first part is the amount of electricity purchased from the power grid. For this purpose, at each step, the total energy required for electric vehicles is calculated and the total remaining renewable energies from it is subtracted to determine the amount of electricity needed for purchase. Then, by multiplying this amount by the cost of electricity at that time, the cost of purchasing energy from the power grid is determined. It should be noted that this amount is considered a penalty term and needs to be minimized for optimization. To calculate the purchased energy, first, a computation of the total remaining renewable energies according to formula (7) must be done.

$$r^t = ReLU(\sum_{i \in stations}[{E_{renewable}}^t - Load^t]) \quad (7)$$

The formula (7) represents the total remaining renewable energy generated by prosumers, where ${E_{renewable}}^t$ indicates the renewable energy generated at each household and $Load^t$ indicates the consumption load of that prosumer. The ReLU function is defined as formula (8):

$$ReLU(x) = \begin{cases} x & x \geq 0 \\ 0 & otherwise \end{cases} \quad (8)$$

Now, having the amount of remaining renewable energy at any moment, the formula (9) calculates the cost of purchasing the deficit load from the power grid.

$$E_{charge} = \min(AvR, (1 - SOC) \times Capacity) \quad (9)$$

E_{charge} is the energy required to charge the battery. AvR is the available rate of charging (8 kWh) according to the average rate of charging extracted from Faia (et al., 2021).

$Capacity$ is the total energy capacity of the battery, which is considered to be 45 kWh for all the EVs here according to the average capacity of EVs in Faia (et al., 2021).

In the defined environment, there is no possibility of selling electricity from the stations to the power grid. That is why the second term of the reward is a penalty on the amount of energy that is produced by the prosumers but not consumed by the vehicles. Formula 10 is also considered a penalty term presented for the environment to maintain the explore and exploit process of the RL environment.

$$P_{RE}{}^{t} = ReLU\left(\sum_{i \in Cars} r^{t} - E_i{}^{t}\right).\frac{Price^{t}}{2} \tag{10}$$

$P_{RE}{}^{t}$ is the power associated with renewable energy at time, $E_i{}^{t}$ represents the total energy injected into or taken from the vehicles by the stations, and r^{t} denotes the total remaining renewable energies at moment t.

In the defined environment, in an optimal state, a vehicle that visits a charging station should not leave the station with a low charge and should benefit from the services as much as possible. Therefore, a penalty is considered in this thesis for leaving the station with a low charge. The mentioned penalty is calculated according to the following formula 11:

$$P_{SOC} = \begin{cases} [(1 - SOC_i) \times 2]^2 \\ (1 - SOC_i) \times 2 \\ (1 - SOC_i) \times 3 \\ (1 - SOC_i) \times 5 \end{cases} \tag{11}$$

Where SOC represents the remaining charge of each vehicle, and $SOC_i \in [0,1]$. This function imposes a penalty based on how much the vehicle's battery is below full charge (100%) when it leaves the charging station. The closer SOC_i is to 1 (or 100%), the smaller the penalty because it would be a smaller number. The exact penalties are determined by multiplying the deficit from the full charge by different factors (2 squared, 2, 3, or 5).

These factors correspond to different tiers or conditions for the penalties. This is a policy to ensure that EVs on the road maintain a higher average charge level to keep the system more reliable for EV transportation or the stability of the grid in a vehicle-to-grid (V2G) system.

Given that all three parts considered for the reward are in the form of a penalty term, the reward value of the environment is defined as formula 12:

$$reward = -1 \times (P_{EV} + P_{RE} + P_{SOC}) \quad (12)$$

Action (action): The action for each vehicle is a number in the range $a_i \in [-1, +1]$, where positive numbers indicate the injection of electricity from the station to the vehicle, and negative numbers indicate the injection of electricity from the vehicle to the station. Furthermore, the numbers are proportional. The number "+1" indicates a full charge of an empty battery, and -1 indicates a complete discharge of a fully charged battery. To calculate how each vehicle is charged by the station, formula (9) is used. Based on that, the lower the charge level of a vehicle, the higher the rate it receives from the station, although this rate should not exceed the average rate of 8 kWh.

To calculate the discharge of each vehicle, the following formula is used (formula 13):

$$E_{charge} = \min(AvR, SOC \times Capacity) \quad (13)$$

Then, to apply the calculated energies above in the charge level of each vehicle, the formula 15 is presented:

$$SOC^{new} = SOC^{old} + \frac{a \times E_{charge}}{Capacity} \quad (14)$$

Where SOC^{old} is the previous charge level of the vehicle, SOC^{new} is the new charge level of the vehicle, and a is the action applied. It should be noted that in the above formula, the battery energy level must remain within the range of -1 to +1.

A detailed presentation of this section can be found in Table 4, which focuses on supporting functions and data processing utilities, such as energy calculations, initial value settings, action simulations, and state calculations.

Table 4. The Pseudocode of EV charging utilities

Algorithm 3 EV Charging Utilities

```
1: Energy_Calculation
2:    load data from CSV files:
3:       - tariffs_data = pd.read_csv('Tariffs.csv')
4:       - pv_data = pd.read_csv('PV.csv')
5:       - load_data = pd.read_csv('Load.csv')
6:    define function Energy_Data(self):
7:       set experiment parameters:
8:          - days_of_experiment = 1
9:          - number_of_prosumers = 50
10:         - price_flag = self.price_flag       # which tariff is set for the simulation
11:         - pv_flag = self.pv_flag             # whether it would be used or not
12:      initialize arrays:
13:         - Renewable = np.zeros([number_of_prosumers, 96])
14:         - Load = np.zeros([number_of_prosumers, 96])
15:      for each prosumer do
16:         for each timestep do
17:            calculate the remaining RE after using it for consumption
18:            calculate the remaining load after using RE
19:      determine price structure based on price flag
20:      create Price array based on chosen tariff
21:      return Load, Renewable, Price
22:   end procedure

23: Initial_Values
24:   define function EVs_Calculations():
25:      load data from CSV file (Moves)
```

```
    initialize data structures:
      - arrival = {}
      - departure = {}
      - SOC = np.zeros([number_of_cars, 96])
    for each car do
      for each timestep do
        identify changes in movement state
        assign SOC based on departure duration
    handle last timestep departure
    ensure all cars are fully charged at midnight
    calculate presence and evolution of cars
    return SOC, ArrivalT, DepartureT, present_cars, evolution_of_cars
  end procedure

Action_Simulation
  define function RL_Actions(self, actions):
    set current hour based on timestep
    initialize EV charging demands and SOC
    for each car do
      determine charging or discharging action
      calculate energy demand or supply
      update SOC for next timestep
    compute rewards based on penalties:
      - Penalty_EV
      - Penalty_RE
      - Penalty_SOC
    calculate final reward
    return reward, Grid_trade, RE_surplus, Penalty_SOC, SOC
  end procedure

Station_State
  define function RL_States(self):
    initialize key variables:
      - SOC = self.SOC
      - Arrival = self.Invalues['ArrivalT']
      - Departure = self.Invalues['DepartureT']
      - present_cars = self.Invalues['present_cars']
    identify cars departing soon
    compute hours until departure
```

```
63:     calculate SOC for each car
64:     return leave, Departure_hour, Battery
65: end procedure
```

6 Results (Simulation Implementation)

6.1 An Evaluation of DDPG, PPO, and RBC

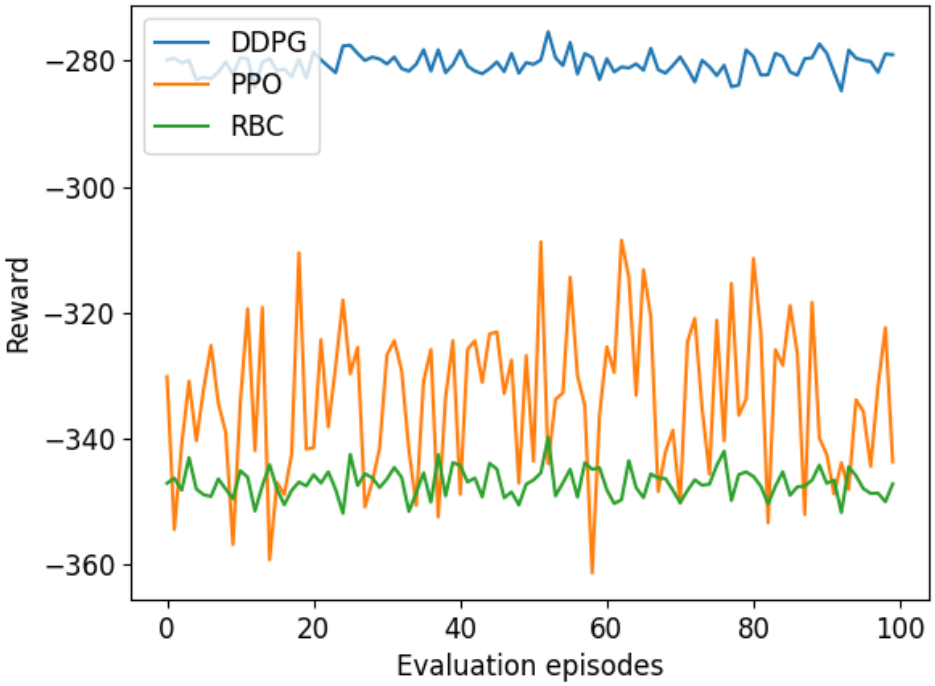

Figure 5. The Reward obtained by each agent in different policies

Figure 5 offers a comparative assessment of three distinct policies, Deep Deterministic Policy Gradient (DDPG), Proximal Policy Optimization (PPO), and Rule-Based Control (RBC) in terms of their reward outcomes over a sequence of 100 evaluation episodes. The 'Reward' axis represents the numerical score accrued by each algorithm, signifying its effectiveness within the simulated environment, post-learning phase. The 'Evaluation episodes' axis provides a temporal sequence of tests, which are indicative of the policies' performance over time.

A cursory inspection of Figure 5 reveals that the DDPG algorithm, depicted by the blue line, consistently achieves the highest rewards throughout the evaluation episodes, indicating its superior performance in this specific environment. It maintains a notable degree of stability with minimal fluctuation in reward, suggesting that its policy is both

effective and reliable over successive episodes. The advantage of DDPG most likely stems from its ability to gracefully handle continuous action spaces and deterministic policy approach in decision making. It is crucial for controlling time and charging rates in a setting where EV charging stations are present. Moreover, by effectively balancing exploration and exploitation, the Ornstein-Uhlenbeck action noise was incorporated into the algorithm, allowing it to sustain high performance through efficient search of the action space around similar states.

The PPO algorithm, represented by the orange line, while outperforming the RBC algorithm, exhibits significant volatility in its performance. This on-policy algorithm may have trouble adjusting its policy in a complicated reward landscape like the EV charging simulation. The high degree of reward variability in PPO may be made worse by policy optimization's usage of a multi-epoch stochastic gradient ascent, which may not always converge optimally in the specified environmental configuration.

In contrast, the green line corresponding to the RBC algorithm shows the least variability among the three, though it consistently yields the lowest rewards. Rule-based systems exhibits this patter as they follow predetermined rules instead of adjusting to feedback from the environment. Because it lacks learning and adaptive processes, RBC's performance is consistent, which indicates how robust it is. Nonetheless, the steadily decreasing rewards show that the RBC's static rules fall short of the adaptive algorithms in terms of capturing complexities and optimizing decision-making.

When examining stability during the evaluation phase, it is evident that the DDPG and RBC algorithms demonstrate a degree of dependability, with the former also achieving superior performance. On the other hand, the variability of the PPO algorithm’s rewards indicates a potential for improvement in policy consistency.

To further scrutinize the comparative efficiency of the DDPG and PPO algorithms, an analysis beyond the scope of Figure 5 would be required, focusing on the reward

trajectories during the learning phase. Such an analysis would elucidate the learning progression and the eventual stabilization of policy performance.

6.2 Learning Trajectories of DDPG and PPO

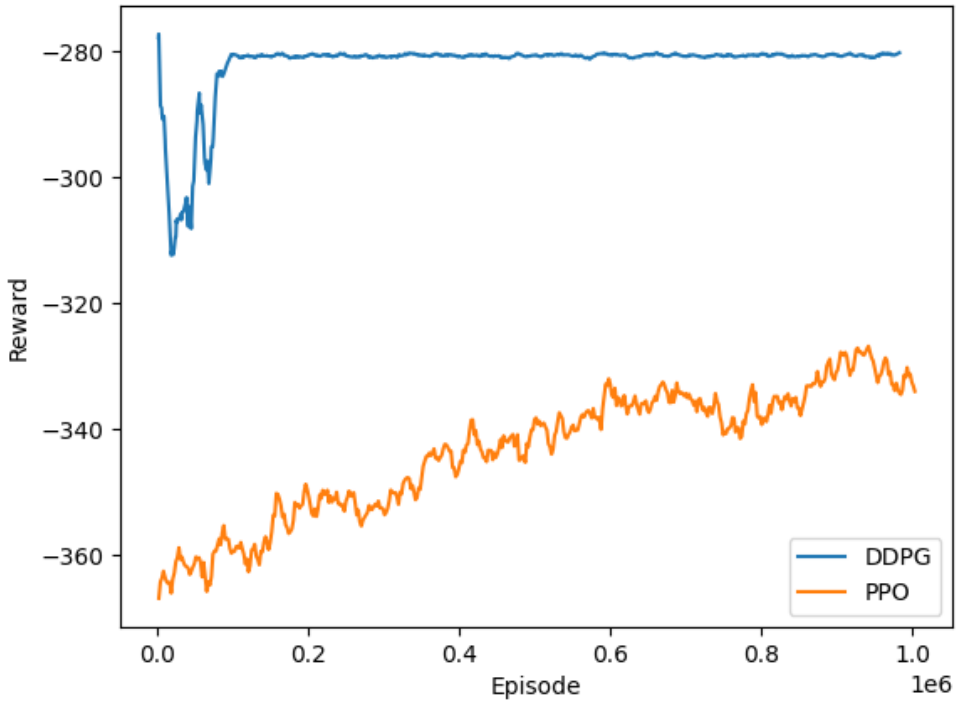

Figure 6. DDPG and PPO Policy Reward Training

Figure 6 provides a comparison of the learning trajectories of the DDPG and PPO algorithms throughout training episodes, with the number of episodes reaching one million. The vertical axis tracks the cumulative reward gained during training.

The DDPG algorithm, indicated by the blue line, demonstrates rapid convergence towards a local optimum early in the training process. The architecture of DDPG incorporates noise processes. However, since it is deterministic, it is naturally focused on taking advantage of the knowledge it gains immediately to maximize rewards. This quick stabilization of reward suggests that DDPG efficiently bootstraps knowledge from its experiences, leveraging this to achieve a policy that yields consistent results. The algorithm's

exploration tends to diminish as it converges to what it considers as an optimal policy, based on its experience. This may cause the algorithm to prematurely plateau and thus fail to sufficiently examine possible better policies outside of the present policy's near area.

Conversely, the PPO algorithm, illustrated by the orange maintains exploration throughout its training process due to being on-policy. That is why it shows a more gradual increase in rewards over the training episodes. Every update is based on new data, and the clipping method aids in a deeper exploration of the policy space, avoiding the policy from prematurely convergent to local optima that are not optimal. This trend suggests that PPO continues to explore the policy space throughout the training process, potentially leading to a better overall understanding of the environment. Such persistent exploration could be advantageous in environments where the initial apparent local optima are not the best possible policies.

Exploration (looking for new information) and exploitation (using existing knowledge to maximize reward) must be balanced at the heart of reinforcement learning. The DDPG method favors quick exploitation that works well when the ideal course of action doesn't alter all that much. On the other hand, PPO's approach of persistently examining the policy space enhances its flexibility. However, it can lead to a delayed convergence in situations when a policy is readily apparent.

6.3 The Cost-Effectiveness of Policies

To analyze the economic impact of each policy, Figure 7 is proposed to involving the integration of time-variable electricity pricing into the cost calculation, thereby multiplying the quantity of electricity purchased from the grid at each time step by the electricity price.

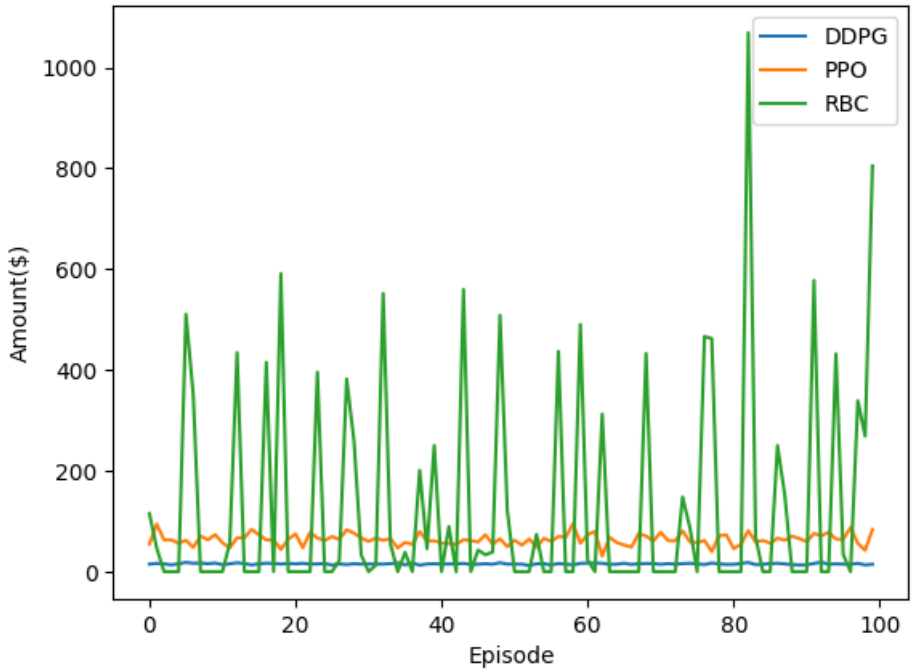

Figure 7. Amount of electricity needed to be purchased from the grid

According to The DDPG algorithm, represented by the blue line, shows the most stable and lowest costs throughout, indicating its efficiency in utilizing available renewable energy and minimizing reliance on the grid. This suggests a well-optimized policy that has learned to predict and match the household's energy consumption with the production of renewable energy. With DDPG, EVs are charged when solar energy is plentiful and nearly free, reducing the need to use the grid to generate electricity during pricey peak demand hours. Because DDPG is able to learn, it can adjust its charging approach to best save money by basing it on the average daily solar energy curve.

In contrast, the exploratory tendency of PPO can result in less consistency in how well it matches charging with peak solar times as compared to DDPG. Even though it is equally capable of responding to solar availability. The results may fluctuate, indicating a higher level of performance variability as a result. PPO may investigate methods that infrequently move charging to less-than-ideal times to see whether they could produce better long-term results.

On the other hand, RBC does not optimally reduce costs, as its static nature doesn't adapt to real-time changes in solar output. It may cause energy to be taken from the grid at high-cost times or underuse abundant solar energy during low-demand times if the regulations are overly strict or out of sync with real consumption patterns.

Table 5. The cost of electricity purchased from the power grid using PV

	PPO	**DDPG**	**RBC**
Total Cost (€)	6,378	1,533	13,086

The data in Table 5 quantifies the total costs for each policy, providing a comparison of their economic impact. With DDPG incurring the lowest total cost of 1,533€, it drastically undercuts the expenses of PPO and RBC, which are 6,378€ and 13,086€ respectively. The cost savings are substantial when considering the hypothetical scenario without solar energy, where costs would escalate to 14,439€ in DDPG, 16,648€ in PPO, and 18,849€ in RBC according to Table 6, demonstrating the significant financial advantage of integrating renewable energy sources into the energy management system.

Table 6. The cost of electricity purchased from the power grid without PV

	PPO	**DDPG**	**RBC**
Total Cost (€)	16,648	14,439	18,849

The analyses of both Figure 7 and Table 5 highlight the superior performance of the DDPG algorithm in reducing electricity purchase costs, evidencing the potential of reinforcement learning algorithms to offer notable economic benefits. However, there is an observed inefficiency across all policies concerning the distribution of surplus energy. The simulation constraints, specifically the inability to sell excess electricity back to the grid, lead to a scenario where surplus renewable energy is not fully utilized, resulting in wastage. In an optimized real-world application, policies would ideally be designed to incorporate mechanisms such as energy storage systems or grid feedback to capture and

redistribute surplus energy, thereby maximizing the utilization of renewable energy production and further enhancing cost savings.

6.4 Energy Wastage

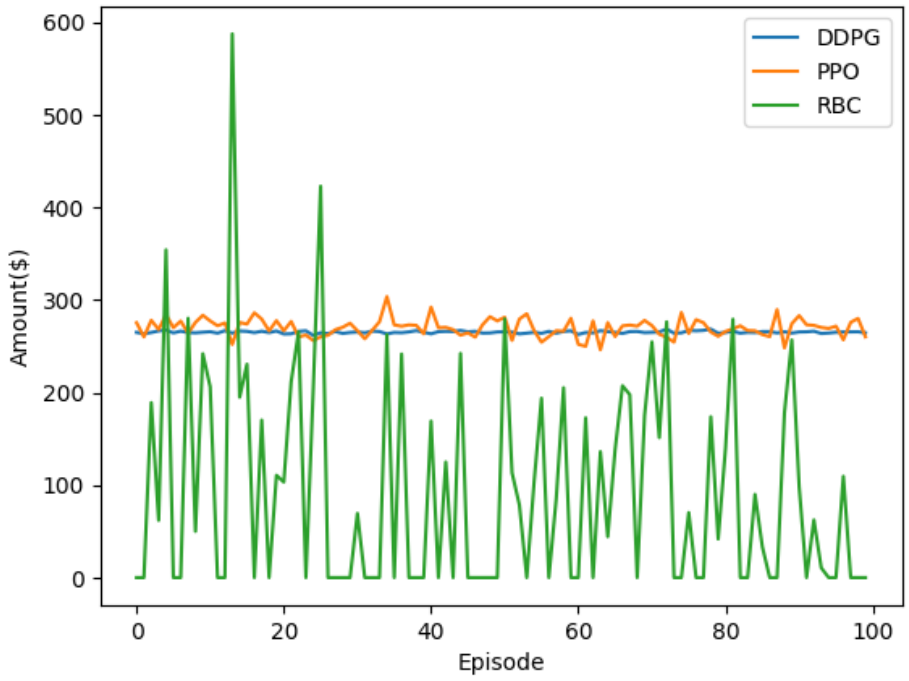

Figure 8. Wasted RE in three policies

Figure 8 illustrates the wastage costs associated with the unused renewable energy (RE) for the DDPG, PPO, and RBC energy management policies for 100 episodes. The chart delineates the amount of surplus renewable energy that was not utilized for any beneficial purpose, such as charging electric vehicles (EVs) or selling back to the grid, which is considered 'wasted' in this context.

The RBC policy, depicted by the green line, exhibits significantly lower peaks of wastage costs compared to the other two algorithms although it is a less flexible and adaptive method than the machine learning-based policies. This indicates that the RBC's rule-based approach has been pre-configured with more efficient usage of the generated

renewable energy under the constraints given in the simulation. Using this method enables RBC to be precisely designed to use renewable energy sources before using grid electricity. These guidelines can result in a more efficient use of the generated electricity if they are designed to maximize the immediate use of renewable energy as it is created (for example, by arranging charging during hours of peak solar production). However, the performance of the RBC method varies between episodes because its fixed rules do not match the fluctuating conditions of renewable energy production (because of meteorological and seasonal variations reasons) and consumption patterns.

Table 7. Wastage of Electricity Generated by RE

	PPO	**DDPG**	**RBC**
Total Cost (€)	26994	26519	9349

In Table 7, the total wastage costs quantified over the evaluation period further emphasize the findings from Figure 8. The RBC algorithm has resulted in the least total wastage (9,349€), significantly outperforming the machine learning-based DDPG (26,519€) and PPO (26,994€) algorithms in terms of reducing energy wastage. This could imply that while the machine learning algorithms are more adept at optimizing for cost through purchasing decisions, they are not as efficient in handling the distribution of surplus energy within the simulation's parameters. For their policies to be optimized, DDPG and PPO both use environmental learning. These algorithms may not effectively match the supply of renewable energy with the demand for electricity during early episodes.

While the comparison of rewards is a common metric for evaluating the performance of reinforcement learning algorithms, Figure 8 and Table 7 underscore the importance of considering additional operational variables, such as energy wastage when assessing overall system efficiency. This approach recognizes that the optimal functioning of an energy management system must balance multiple objectives, including cost minimization and sustainable energy utilization.

6.5 SOC Management

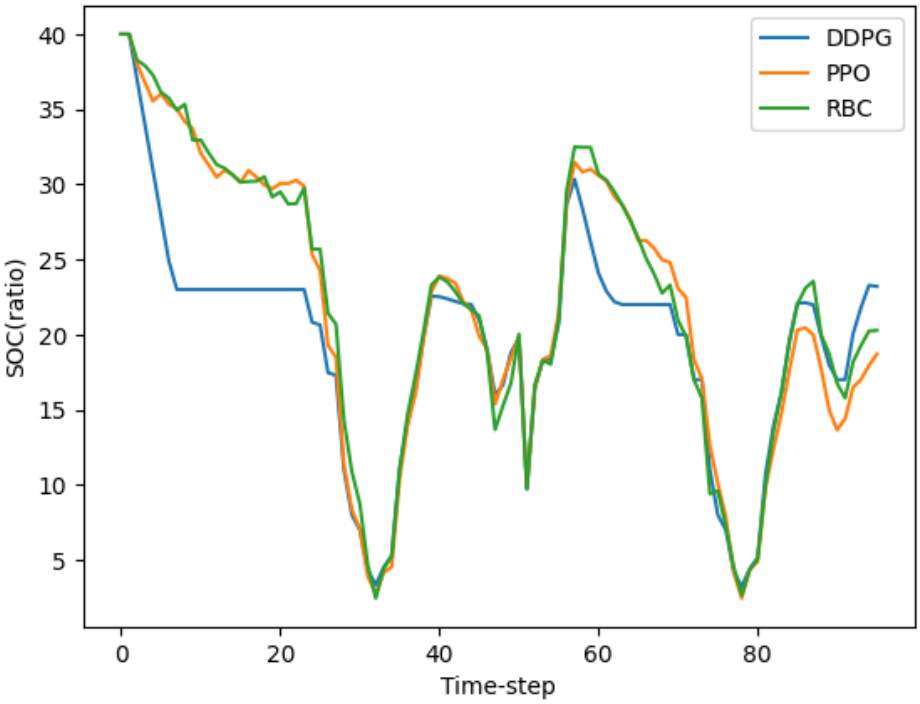

Figure 9. Cumulative SOC in three policies

Figure 9 presents a temporal visualization of the cumulative state of charge (SOC) ratios for a fleet of electric vehicles (EVs) managed under the three mentioned policies. Each SOC ratio ranges from -1 to +1, with the aggregate at each 15-minute time step providing a snapshot of the fleet's overall charge status.

The graph portrays the DDPG algorithm as maintaining a relatively balanced SOC, avoiding extremes in undercharging or overcharging the fleet. This suggests an optimal energy distribution strategy that aligns vehicle charging needs with the availability of renewable energy, thus ensuring that vehicles are adequately charged for use without incurring unnecessary energy wastage. The SOC trajectories for the PPO and RBC algorithms show greater fluctuations. Both algorithms experience moments where the cumulative SOC dips or peaks more sharply, implying periods of potential overcharging or under-utilization of the fleet's battery capacity. However, it is worth noting that neither PPO nor RBC

consistently underperforms compared to DDPG, as they all occasionally cross over one another's paths.

The data on the presence of cars (Figure 4) shows significant variations in the number of cars that need to be charged at various time intervals. In response, the best charging approach would modify the SOC levels as necessary: lower energy output during peak hours to save energy or divert it to other uses, and raise SOC during off-peak hours to meet increased demand without going overboard or wasting energy. This behavior suggests the existence of an advanced energy management system covered by policies like the DDPG, which appears to adjust to fluctuating vehicle counts while maximizing energy use and making sure cars are charged appropriately without wasting or consuming excessive amounts of energy.

7 Discussion

The exploration of reinforcement learning in the operation of smart grid energy systems, as observed in this study, is a testament to the transformative potential of machine learning in the energy sector. With each algorithm revealing distinct strengths, the study paints a comprehensive picture of the current capabilities and areas ripe for improvement.

Currently, any excess energy not used for charging electric vehicles (EVs) is considered wasted, representing an opportunity loss both economically and in energy resource management. Future studies could focus on two main strategies to address this issue:

- **Grid Feed-In Systems:** Incorporating the option to sell surplus electricity back to the grid would not only prevent wastage but could also provide a financial return to the prosumers and contribute to the overall stability of the grid.
- **Battery Energy Storage Systems (BESS):** The installation of battery storage systems offers another viable solution for capturing and retaining surplus energy. By storing excess production, energy can be utilized during periods of high demand or low generation, thereby enhancing the reliability and resilience of the smart grid system.

Building on the existing research, further algorithmic refinement could enable a more nuanced balance between immediate reward maximization and long-term strategic planning. Adaptive algorithms that can respond to real-time pricing and engage with BESS and grid feed-in options would significantly advance smart grid management capabilities. Expanding the scope to include variable renewable energy sources (like wind energy), larger-scale simulations, and behavioral models of human energy consumption will enhance the realism and applicability of the research. Furthermore, the integration of these technologies must be considered within the context of the evolving policy and regulatory environment, which will undoubtedly shape the operational framework of future smart grids.

8 Conclusion

In this thesis, a dynamic and uncertain environment populated by 50 prosumer nodes and 40 electric vehicles (EVs) is examined, each interacting within a smart grid framework over a 24-hour simulation period. Reinforcement learning (RL), characterized by its trial-and-error approach, interacts with this environment to sequentially make decisions that aim to maximize the reward. By framing the problem within the structure of a MDP, we engage with a system where outcomes are probabilistically determined and inherently difficult to predict. Each household in the simulation is equipped with solar panels that generate variable amounts of electricity, contingent upon weather conditions. Additionally, each station is connected to the power grid, allowing for the purchase of electricity to supplement any shortfall in production, which can then be used to charge the EVs. The vehicles enter the station with varying charge levels, and the challenge lies in either increasing or decreasing the battery charge efficiently. The simulation is structured into 96 time steps of 15 minutes each, ensuring a detailed resolution of energy dynamics.

Given this complex backdrop, the study applies RL algorithms to navigate the intricate decisions surrounding the charging strategies for EVs. These decisions hinge on a well-defined set of states, rewards, and actions that determine the algorithms' efficiency in optimizing for both the individual vehicle requirements and the broader energy demands of the grid. Three various RL algorithms (Policies) are used in this framework, namely DDPG, PPO, and RBC. The comparative evaluation of the Deep Deterministic Policy Gradient (DDPG), Proximal Policy Optimization (PPO), and Rule-Based Control (RBC) algorithms demonstrated that each possesses distinct advantages and challenges in managing the delicate balance between energy consumption, production, and storage.

DDPG emerged as a robust approach for cost-effective energy management, adeptly minimizing the need for grid electricity purchases by effectively aligning energy consumption with renewable production. Its rapid convergence suggests a strong capability for bootstrapping and immediate application of learned policies, though this may come

at the cost of reduced exploration in the policy space. While PPO, with its continuous exploration, has the potential for better adaptability in dynamic environments, its performance was marked by greater volatility in cost-effectiveness and energy utilization. This characteristic could potentially yield improvements over time as the algorithm further explores the environment. On the other hand, the RBC algorithm's performance highlighted the benefits of predefined rules in reducing energy wastage, emphasizing the value of predictable and stable energy management strategies in certain contexts. While not as economically efficient in reducing costs as its machine learning counterparts, its effectiveness in minimizing wastage presents a compelling case for its inclusion in a hybrid strategy.

When considering the overall sustainability and economic viability of smart grid systems, it is clear that a multifaceted approach is necessary. The integration of renewable energy not only provides an avenue for cost savings but also presents challenges in terms of energy distribution and storage. The inability to return surplus energy to the grid in this simulation underscored the importance of effective energy storage solutions or alternative strategies to fully capitalize on the generated renewable energy. Through the lens of EV fleet management, the study accentuated the importance of managing the state of charge (SOC) levels to prevent overcharging and ensure the vehicles are charged by renewable energy availability, thereby enhancing the operational efficiency of the fleet.

Overall, the analyses suggest that while reinforcement learning algorithms such as DDPG and PPO show promise in optimizing for economic gains and adaptability, the predictability and stability of rule-based systems like RBC offer valuable benefits. An integrated approach that combines the predictive and learning capabilities of machine learning with the consistency of rule-based algorithms may yield the most effective smart grid energy management system, one that maximizes renewable energy usage, minimizes costs, and promotes sustainability in the evolving landscape of energy distribution and consumption.

References

Al-Gabalawy, M. (2021). Reinforcement learning for the optimization of electric vehicle virtual power plants. *International Transactions on Electrical Energy Systems, 38*(8). doi:https://doi.org/10.1002/2050-7038.12951

Arwa, E., & Folly, K. (2021, June). Improved Q-learning for Energy Management in a Grid-tied PV Microgrid. *SAIEE Africa Research Journal, 112*(2), 77-88. doi:10.23919/SAIEE.2021.9432896

Cai, W., Kordabad, A., & Gros, S. (2023, March). Energy management in residential microgrid using model predictive control-based reinforcement learning and Shapley value. *Engineering Applications of Artificial Intelligence, 119*, 0952-1976. doi:10.1016/j.engappai.2022.105793

Chen, G., Peng, Y., & Zhang, M. (2018). An Adaptive Clipping Approach for Proximal Policy Optimization. *ArXiv*.

Chiş, A., Lundén , J., & Koivunen, V. (2017, May). Reinforcement Learning-Based Plug-in Electric Vehicle Charging With Forecasted Price. *IEEE Transactions on Vehicular Technology, 66*(5), 3674-3684. doi:10.1109/TVT.2016.2603536

Cohen, J., Azarova, V., Kollmann, A., & Reichl, J. (2019). Q-complementarity in household adoption of photovoltaics and electricity-intensive goods: The case of electric vehicles. *Energy Economics*, 567-577. doi:https://doi.org/10.1016/J.ENECO.2019.08.004

Co-Reyes, J., Sanjeev, S., Berseth, G., Gupta, A., & Levine, S. (2020). Ecological Reinforcement Learning. *ArXiv*.

Dabbaghjamanesh, M., Moeini, A., & Kavousi-Fard, A. (2021, June). Reinforcement Learning-Based Load Forecasting of Electric Vehicle Charging Station Using Q-Learning Technique. *IEEE Transactions on Industrial Informatics, 17*(6), 4229-4237. doi:10.1109/TII.2020.2990397

Esfandyari, A., Norton, B., Conlon, M., & McCormack, S. (2019). Performance of a campus photovoltaic electric vehicle charging station in a temperate climate. *Solar Energy*, 762-771. doi:https://doi.org/10.1016/J.SOLENER.2018.12.005

Faia, R., Soares, J., Fotouhi Ghazvini, M., Franco, J., & Vale, Z. (2021). Local Electricity Markets for Electric Vehicles: An Application Study Using a Decentralized Iterative Approach. *Frontiers in Energy Research*, 9. doi:10.3389/fenrg.2021.705066

Ferro, G., Laureri, F., Miniciardi, R., & Robba, M. (n.d.). An optimization model for electrical vehicles scheduling in a smart grid. *Sustainable Energy, Grids and Networks*, 14, 62-70. doi:https://doi.org/10.1016/j.segan.2018.04.002

Foster, J., & Caramanis, M. (2013, Aug.). Optimal Power Market Participation of Plug-In Electric Vehicles Pooled by Distribution Feeder. *IEEE Transactions on Power Systems, 28*(3), 2065-2076. doi:10.1109/TPWRS.2012.2232682

Härtel, F., & Bocklisch, T. (2023). Minimizing Energy Cost in PV Battery Storage Systems Using Reinforcement Learning. *IEEE Access, 11*, 39855-39865. doi:10.1109/ACCESS.2023.3267978

Huang, S., Yang, M., Zhang, C., Yun, J., Gao, Y., & Li, P. (2020). A Control Strategy Based on Deep Reinforcement Learning Under the Combined Wind-Solar Storage System. *2020 IEEE 3rd Student Conference on Electrical Machines and Systems (SCEMS)*, 819-824. doi:10.1109/SCEMS48876.2020.9352436

Karatzinis, G., Korkas, C., Terzopoulos, M., Tsaknakis, C., Stefanopoulou, A., Michailidis, I., & Kosmatopoulos, E. (2022). Chargym: An EV Charging Station Model for Controller Benchmarking. In *Artificial Intelligence Applications and Innovations* (pp. 241-252).

Karmaker, A., Hossain, M., Pota, H., Onen, A., & Jung, J. (2023). Energy Management System for Hybrid Renewable Energy-Based Electric Vehicle Charging Station. *IEEE Access*, 11, 27793-27805. doi:https://doi.org/10.1109/ACCESS.2023.3259232

Khaki, B., Chung, Y., Chu, C., & Gadh, R. (2019). Probabilistic Electric Vehicle Load Management in Distribution Grids. *2019 IEEE Transportation Electrification Conference and Expo (ITEC)*, 1-6. doi:https://doi.org/10.1109/ITEC.2019.8790535

Lan, T., Jermsittiparsert, K., Alrashood, S., Rezaei, M., Al-Ghussain, L., & Mohamed, M. (2021). An Advanced Machine Learning Based Energy Management of

Renewable Microgrids Considering Hybrid Electric Vehicles' Charging Demand. *Energies*, 14(3), 569. doi:https://doi.org/10.3390/EN14030569

Li , S., & et al. (2022, May). Electric Vehicle Charging Management Based on Deep Reinforcement Learning. *Journal of Modern Power Systems and Clean Energy, 10*(3), 719-730. doi:10.35833/MPCE.2020.000460

Li, Y., Han, M., Yang, Z., & Li, G. (2021). oordinating Flexible Demand Response and Renewable Uncertainties for Scheduling of Community Integrated Energy Systems With an Electric Vehicle Charging Station: A Bi-Level Approach. *IEEE Transactions on Sustainable Energy*, 12(4), 2321-2331. doi:https://doi.org/10.1109/TSTE.2021.3090463

López, K., Gagné, C., & Gardner, M. (2018). Demand-Side Management Using Deep Learning for Smart Charging of Electric Vehicles. *IEEE Transactions on Smart Grid*, 1-1. doi:10.1109/TSG.2018.2808247

Mololoth, V., Saguna, S., & Åhlund, C. (2023). Blockchain and Machine Learning for Future Smart Grids: A Review. *Energies 2023*, 16, 528. doi: https://doi.org/10.3390/en16010528

Morstyn, T., Teytelboym, A., & Mcculloch, M. (2018). Matching Markets with Contracts for Electric Vehicle Smart Charging. *IEEE Power & Energy Society General Meeting (PESGM)*, 1-5. doi:https://doi.org/10.1109/PESGM.2018.8586361

Moussaoui, H., Akkad, N., & Benslimane, M. (2023). Reinforcement Learning: A review. *International Journal of Computing and Digital Systems*. doi:https://doi.org/10.12785/ijcds/1301118

Najafi, S., Shafie-khah, M., Siano, P., Wei, W., & Catalão, P. (2019, December). Reinforcement learning method for plug-in electric vehicle bidding. *IET Smart Grid, 2*(4), 529-536. doi:10.1049/iet-stg.2018.0297

Narvekar, S., Peng, B., Leonetti, M., Sinapov, J., Taylor, M., & Stone, P. (2020). Curriculum Learning for Reinforcement Learning Domains: A Framework and Survey. *ArXiv*. Retrieved from abs/2003.04960

Nezamoddini, N., & Wang, Y. (2016). Risk management and participation planning of electric vehicles in smart grids for demand response. *Energy*, 116, 836-850. doi:https://doi.org/10.1016/J.ENERGY.2016.10.002

Ojand , K., & Dagdougui, H. (2022, Jan.). Q-Learning-Based Model Predictive Control for Energy Management in Residential Aggregator. *19*(1), 70-81. doi:10.1109/TASE.2021.3091334

Qiu, D., Ye, Y., Papadaskalopoulos, D., & Strbac, G. (2020, Sept.-Oct.). A Deep Reinforcement Learning Method for Pricing Electric Vehicles With Discrete Charging Levels. *IEEE Transactions on Industry Applications, 56*(5), 5901-5912. doi:10.1109/TIA.2020.2984614

Radu, A., Eremia, M., & Toma, L. (n.d.). Optimal charging coordination of electric vehicles considering distributed energy resources. *2019 IEEE Milan PowerTech*, 1-6. doi:http://dx.doi.org/10.1109/PTC.2019.8810756

Schulman, J., Wolski, F., Dhariwal, P., Radford, A., & Klimov, O. (2017). Proximal Policy Optimization Algorithms. Retrieved from abs/1707.06347

Shin, M., Choi, D.-H., & Kim, J. (2020, May). Cooperative Management for PV/ESS-Enabled Electric Vehicle Charging Stations: A Multiagent Deep Reinforcement Learning Approach. *IEEE Transactions on Industrial Informatics, 16*(5), 3493-3503. doi:10.1109/TII.2019.2944183

Sutton, R., & Barto, A. (2018). *Reinforcement learning: an introduction.* Westchester Publishing Services.

Vandael, S., Claessens, B., Ernst, D., Holvoet , T., & Deconinck, G. (2015, July). Reinforcement Learning of Heuristic EV Fleet Charging in a Day-Ahead Electricity Market. *IEEE Transactions on Smart Grid, 6*(4), 1795-1805. doi:10.1109/TSG.2015.2393059

Vázquez-Canteli, P., José, R., & Nagy, Z. (2019). Reinforcement learning for demand response: A review of algorithms and modeling techniques. *Applied Energy, Elsevier, 235(C)*, 1072-1089. doi:10.1016/j.apenergy.2018.11.002

Wan, Z., Li, H., He, H., & Prokhorov, D. (2019, Sept.). Model-Free Real-Time EV Charging Scheduling Based on Deep Reinforcement Learning. *IEEE Transactions on Smart Grid, 10*(5), 5246-5257. doi:10.1109/TSG.2018.2879572

Yan, L., Chen, X., Zhou, J., Chen, Y., & Wen, J. (2021, Nov.). Deep Reinforcement Learning for Continuous Electric Vehicles Charging Control With Dynamic User Behaviors. *IEEE Transactions on Smart Grid, 12*(6), 5124-5134. doi:10.1109/TSG.2021.3098298

Yao, L., Lim, W., & Tsai, T. (2017). A Real-Time Charging Scheme for Demand Response in Electric Vehicle Parking Station. *IEEE Transactions on Smart Grid*, 8, 52-62. doi:https://doi.org/10.1109/TSG.2016.2582749

Yao, M., Da, D., Lu, X., & Wang, Y. (2024). A Review of Capacity Allocation and Control Strategies for Electric Vehicle Charging Stations with Integrated Photovoltaic and Energy Storage Systems. *World Electric Vehicle Journal*, 15(3):101. doi:https://doi.org/10.3390/wevj15030101

Ye, X., Ji, T., Li, M., & Wu, Q. (2018). Optimal Control Strategy for Plug-in Electric Vehicles Based on Reinforcement Learning in Distribution Networks. *2018 International Conference on Power System Technology (POWERCON)*, 1706-1711. doi:10.1109/POWERCON.2018.8602101

Ye, Y., Qiu, D., Sun, M., Papadaskalopoulos, D., & Strbac, G. (2020, March). Deep Reinforcement Learning for Strategic Bidding in Electricity Markets. *IEEE Transactions on Smart Grid, 11*(2), 1343-1355. doi:10.1109/TSG.2019.2936142

Zhang, Z., Chen, J., Chen, Z., & Li, W. (2019). Asynchronous Episodic Deep Deterministic Policy Gradient: Toward Continuous Control in Computationally Complex Environments. *IEEE Transactions on Cybernetics*, (99):1-10. doi:10.1109/TCYB.2019.2939174