energies

MDPI

Review

Review and Evaluation of Reinforcement Learning Frameworks on Smart Grid Applications

Dimitrios Vamvakas [1,2,†], Panagiotis Michailidis [1,2,†], Christos Korkas [1,2,*] and Elias Kosmatopoulos [1,2]

[1] Center for Research and Technology Hellas, 57001 Thessaloniki, Greece; dvamvakas@iti.gr (D.V.); panosmih@iti.gr (P.M.); kosmatop@iti.gr (E.K.)
[2] Department of Electrical and Computer Engineering, Democritus University of Thrace, 67100 Xanthi, Greece
* Correspondence: chriskorkas@iti.gr; Tel.: +30-23-1046-4160
† These authors contributed equally to this work.

Abstract: With the rise in electricity, gas and oil prices and the persistently high levels of carbon emissions, there is an increasing demand for effective energy management in energy systems, including electrical grids. Recent literature exhibits large potential for optimizing the behavior of such systems towards energy performance, reducing peak loads and exploiting environmentally friendly ways for energy production. However, the primary challenge relies on the optimization of such systems, which introduces significant complexities since they present quite dynamic behavior. Such cyberphysical frameworks usually integrate multiple interconnected components such as power plants, transmission lines, distribution networks and various types of energy-storage systems, while the behavior of these components is affected by various external factors such as user individual requirements, weather conditions, energy demand and market prices. Consequently, traditional optimal control approaches—such as Rule-Based Control (RBC)—prove inadequate to deal with the diverse dynamics which define the behavior of such complicated frameworks. Moreover, even sophisticated techniques—such as Model Predictive Control (MPC)—showcase model-related limitations that hinder the applicability of an optimal control scheme. To this end, AI model-free techniques such as Reinforcement Learning (RL) offer a fruitful potential for embedding efficient optimal control in cases of energy systems. Recent studies present promising results in various fields of engineering, indicating that RL frameworks may prove the key element for delivering efficient optimal control in smart buildings, electric vehicle charging and smart grid applications. The current paper provides a comprehensive review of RL implementations in energy systems frameworks—such as Renewable Energy Sources (RESs), Building Energy-Management Systems (BEMSs) and Electric Vehicle Charging Stations (EVCSs)—illustrating the benefits and the opportunities of such approaches. The work examines more than 80 highly cited papers focusing on recent RL research applications—between 2015 and 2023—and analyzes the model-free RL potential as regards the energy systems' control optimization in the future.

Keywords: Reinforcement Learning (RL); Hierarchical Reinforcement Learning (HRL); Machine Learning; deep learning; power and energy systems; building energy management; smart buildings; smart grids; electric vehicles

check for updates

Citation: Vamvakas, D.; Michailidis, P.; Korkas, C.; Kosmatopoulos, E. Review and Evaluation of Reinforcement Learning Frameworks on Smart Grid Applications. *Energies* **2023**, *16*, 5326. https://doi.org/10.3390/en16145326

Academic Editor: GM Shafiullah

Received: 12 June 2023
Revised: 30 June 2023
Accepted: 4 July 2023
Published: 12 July 2023

Copyright: © 2023 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (https://creativecommons.org/licenses/by/4.0/).

1. Introduction

As the demand for energy continues to rise, it is imperative to shift towards more efficient control strategies for devices and equipment. Energy efficiency plays a pivotal role in reducing energy consumption, minimizing environmental impact and optimizing resource utilization [1–6]. This trend initiated the urgency for exploiting alternative energy sources to address the rising energy requirements. To this end, the transition to cleaner energy systems is crucial to reduce the impact of fossil fuel consumption on the environment and ensure long-term energy security across the globe [7–9]. Coupled with the growing need to

manage energy consumption and production optimally, the challenge of applying optimal control to energy systems illustrates the primary concern of various studies conducted by researchers and scientists. Some of the most crucial and dynamic energy systems that are thoroughly studied towards control optimization concern Renewable Energy Sources (RESs) frameworks, Building Energy-Management Systems (BEMSs) and Electric Vehicle Charging Stations (EVCSs)—along with the integration of additional equipment as RES and storage technologies in the optimization mix. One of the primary challenges in optimizing the behavior of these energy systems is to aim to balance the trade-off between different objectives such as energy cost, energy efficiency, reliability, user comfort and requirements, grid services and environmental impact. The main challenges in control applications of smart grid nodes, and more specifically for the systems mentioned above, are presented briefly below:

- **Renewable Energy Sources and Storage Solutions for Smart Grid Services and Provisioning:** In RES frameworks, optimal control is empowered by novel methodologies that can be used to optimally exploit the provision of renewable energy, to store excess energy when it is available and discharge it when energy demand is high or energy prices are favorable, to minimize costs, to forecast RES generation, as well as to optimize the operation of the integrated energy-storage systems, such as batteries. Procedures may utilize historical energy production and consumption data, weather forecasts and energy market prices to make optimal energy-storage decisions and RES integration to the grid.
- **Building Energy-Management Systems—BEMSs:** In buildings, optimal control and multiple energy-friendly technologies are being used to reduce energy consumption and carbon emissions. For example, building energy-management systems can use data from sensors and smart meters to optimize heating, ventilation and air-conditioning (HVAC) systems, lighting and other building systems. Solar PV systems and other RES technologies along with storage may be additionally integrated into the optimization mix to provide a reliable and sustainable source of electricity.
- **Electric Vehicle Charging Stations—EVCSs:** Similarly, in charging stations, optimal control and multiple energy-friendly technologies are being used to manage the charging of electric vehicles. Smart charging systems can prioritize the charging of vehicles based on their battery levels and enhance the efficiency of renewable energy sources while minimizing the load on the grid during peak hours.

1.1. Control Strategies for Energy Systems

In order to enable efficient optimal control towards such multiverse and dynamically changing frameworks, conventional Rule-Based Control techniques (RBC) are being traditionally deployed. However, Rule-Based Control (RBC) portrays an oversimplified control scheme that primarily utilizes a set of predetermined rules in order to generate control decisions [10–15]. To this end, even if RBC control schemes may be effective in certain situations, they are not particularly efficient for the optimization of energy systems: (a) RBC practices are not adequate to offer efficient control towards where complex system dynamics and changing environmental conditions emerge. Since RBC relies on a fixed set of rules, it portrays a non-sufficient approach to capture the full range of system behaviors and handle unexpected events. This can result in sub-optimal control decisions and reduced system efficiency [12,16,17]; (b) RBC requires expert knowledge to design the control rules, which can be time-consuming and costly and, of course, prevents optimality due to different manual procedures. As a result, RBC may not be suitable for energy systems with large or complex control requirements [16]. Despite these limitations, RBC methods are still being used in some energy systems, such as HVAC control in buildings, due to their simplicity and ease of implementation. RBC can be useful for basic control tasks, where the system behavior is well-understood and the control objectives are relatively simple.

Another area of algorithms for optimizing the behavior of such systems is Model Predictive Control (MPC) and in general closed-loop feedback control methods such as

genetic algorithms [18–21] that allow the optimization of control inputs based on predictions of the system behavior. However, even if these methods portray a sophisticated approach, they showcase severe limitations that strongly affect their efficiency and applicability in optimizing energy systems such as RES, buildings and EVCS: (a) Implementation of these techniques relies on an accurate mathematical model of the system. Developing and maintaining such a model can be challenging, especially for complex systems with many variables and uncertain parameters. Any inaccuracies in the model can lead to sub-optimal control decisions and reduced system efficiency [22,23]. Taking also into account that such energy systems may be expanded or develop diverse dynamics, the utilized model requires cumbersome recalibration and verification activities in order to integrate the changing system dynamics [24–26]; (b) MPC and model-based approaches usually exhibit high computational requirements. They involve complex optimization-solving problems at each time step, which can be computationally intensive, especially for large-scale energy systems. This can result in significant time delays and may not be suitable for systems with fast dynamics. However, despite such limitations, MPC remains a prevalent choice in various energy systems because of its proficiency/capability in handling constraints and dynamic systems and it can also be integrated with predictive and control algorithms [20,21,27].

During recent years, the rise of Artificial Intelligence has led research to approaches involving novel Machine Learning techniques, able to provide methods that learn how to solve the control problem through trial-and-error approaches, accomplishing very complex and dynamic tasks in various disciplines like robotics, natural language processing and other relevant fields [28–36]. The area of ML that uses these approaches is called Reinforcement Learning (RL): Reinforcement Learning (RL) portrays a promising control approach for energy system optimization owing to its capability to learn from past encounters and adapt to changing system dynamics. RL can handle the complexity and uncertainty of energy systems and develop optimal control policies that maximize performance metrics such as energy efficiency, cost savings and reduced carbon emissions. There are two primary types of RL: model-based and model-free. In model-based RL, the agent has a model of the environment, which is used to simulate the effects of different actions. The agent can then choose actions that maximize a long-term reward function. Model-based RL is similar to Model Predictive Control (MPC) in that both involve a model of the system and use it to predict the effects of different actions and control scenarios [37,38]. It should be noted though that model-based RL is focused on maximizing a long-term reward, while MPC is focused on minimizing a cost function over a finite horizon. Additionally, model-based RL is more flexible and adaptable to system alterations, while MPC assumes a static environment and may require returning if significant changes occur. On the other hand, model-free RL methodologies present a model-independent nature—thus, they do not require the generation of an accurate mathematical model of the system being controlled, which can be difficult to develop and maintain for complex energy systems. Instead, model-free RL algorithms learn from trial-and-error through iterative interactions with the environment, where they receive rewards or penalties for their decisions. In that case, the RL agent revises its control policy according to feedback received from the environment and over time it develops the capacity to choose actions that maximize its cumulative reward over time [38,39]. To this end, these methodologies are adequate for handling complex constraints and adapting to dynamic environments, making such an approach a powerful tool for energy system optimization. RL algorithms can integrate various data sources, such as historical data, real-time sensor measurements and weather forecasts, to learn optimal control policies in real time. Last but not least, it should be noted that model-free or model-based RL methodologies may also be combined with other advanced control techniques, such as Deep Reinforcement Learning (DRL) or even Model Predictive Control (MPC), to further improve the optimization efficiency and performance towards energy systems. For instance, MPC can be employed to generate a high-level control policy, while RL can be used to fine-tune the control parameters and adapt to changing system conditions [40–42].

1.2. Literature Analysis Approach

The approach employed in this comprehensive review article aims to provide an extensive examination and assessment of the scholarly literature pertaining to the subject matter, specifically focusing on Renewable Energy Sources (RESs), Building Energy-Management Systems (BEMSs) and electric vehicle charging systems (EVCSs). The review encompasses the exploration and analysis of various theories and applications discussed within the surveyed papers, including their hypotheses, type, algorithmic methodologies and application testbed. Current work follows a systematic approach, ensuring a rigorous and structured evaluation of the selected literature.

1. **Criteria of Integrated Studies:** First, the criteria of included studies are defined according to the following research topics: renewable energy sources plant control based on Reinforcement Learning methodologies; building energy-management systems control based on Reinforcement Learning methodologies; electric vehicle charging station system control based on Reinforcement Learning methodologies.
2. **Selection of Related Key Words:** Subsequently, we identified and examined the pertinent keywords associated with the subject matter. A research term is a composite of the primary terms: photovoltaics Reinforcement Learning Control, wind power plants Reinforcement Learning Control, building HVAC Reinforcement Learning Control, water heating Reinforcement Learning Control, lighting Reinforcement Learning Control, electric vehicle charging station Reinforcement Learning Control. The selection of keywords was based on the distinctive attributes of RL and the challenges encountered in RES, BEMS and EVCS domains.
3. **Study Selection:** Subsequently, the research articles obtained through the search conducted in online databases, including Google Scholar and Scopus, were subjected to a rigorous assessment by screening the abstracts. To this end, promising and valuable outcomes were selected for further comprehensive examination and analysis.
4. **Data Extraction:** In the survey process, the fourth stage involved the extraction of data, which entailed categorizing the acquired information from each study based on the type of Reinforcement Learning, the utilized algorithmic methodology and the relevant application testbed. Subsequently, an analysis was conducted on the selected studies, addressing various aspects such as the merits, drawbacks, contributions, potential enhancements, practicality and other relevant factors pertaining to the optimal control problem in energy systems. This comprehensive evaluation provided insights and a critical examination of the surveyed literature.
5. **Quality Assessment:** Following the previous steps, the quality of the work is evaluated based on several criteria. These criteria include the number of citations received by the paper, the scientific contribution of the authors and the methodologies employed in the research. These assessments provide insights into the impact, significance and rigor of the work conducted.
6. **Data Analysis:** In the concluding phase, the studies and their outcomes are systematically organized into distinct subcategories and classifications, facilitating a comprehensive comparison of the research findings.

Therefore, this survey paper provides a detailed discussion of the RL-based solutions used in the RES, BEMS and ECSV control problem and explains the challenges of traditional approaches to the problem while guiding scholars to improve the existing methods.

1.3. Previous Literature Work

The literature integrates numerous noteworthy reviews concerning energy systems considering RES, BEMS and Charging Stations (CSs). The current paper's authors have thoroughly examined their contribution to their integrated literature in order to make an example and inspire others to create a distinct review. Deserving of honorable mention, Cao et al. were scientists who offered an extensive examination of the existing body of literature concerning Reinforcement Learning (RL) with regards to fundamental concepts, diverse algorithmic approaches and their practical implementations within power and

energy systems—the paper delves into the complexities encountered and potential future research directions in this field. Additionally, Mosavi et al. (2019) [43] showcased the current advancements in Machine Learning (ML) models employed within energy systems, accompanied by a unique classification system for both models and their respective applications. Employing an innovative approach, ML models are identified, categorized and organized based on ML modeling techniques, energy types and application domains. The recent work of Mason and Grijalva [44] should also be mentioned. This study provided an extensive examination of the academic literature pertaining to the utilization of Reinforcement Learning in the creation of self-governing systems for managing building energy. Moreover, the 2020 work of Wang and Hong [45] provided a thorough examination of prior research that employed RL for building controls. As concerns the RL control in building structures, it is worth mentioning the 2022 work of Shaqour and Hagishima [46] which illustrates recent progress in Deep Reinforcement Learning (DRL)-based Building Energy-Management Systems (BEMSs) across various building categories, following PRISMA guidelines. Last but not least, it is worth mentioning the work of Abdullah et al. [47], where a comprehensive examination was conducted on the current body of literature pertaining to the framework, goals and structure of Reinforcement Learning (RL)-based approaches for coordinating the charging of electric vehicles within power systems.

1.4. Novelty and Contribution

The current work distinguishes itself from previous literature work since it incorporates an examination of three particular energy systems concerning RES, BEMS and CSEV applications. The authors extensively examined numerous studies from a wide range of highly cited papers—almost 400—selecting approximately 80 research papers in order to offer a deep analysis on various aspects of RL implementations, such as the conceptual framework, application areas, outcomes, types of RL used, RL algorithmic methodologies employed and more. To this end, the volume of the current work is significantly wide in comparison to previous works, offering a thorough examination of a wide range of novel RL research implementations in energy systems. Moreover, it provides a comprehensive three-dimensional distribution on multiple aspects that may concern the Reinforcement Learning approach. The work characterizes and evaluates the following: (a) the type of RL implementation (value-based/policy-based, actor–critic); (b) the utilized RL algorithmic methodology; and (c) the application type—concerning the testbed type—of the RL implementation in the three energy systems. To this end, the interested reader may acquire a holistic overview of the recent RL implementation status quo.

It should also be noted that current work focuses on recent RL implementations that concern the time frame 2015–2023 in order to properly identify the emerging trends and directions of the RL approach for the near future. The selected highly cited papers were selected for the number of citations received by the paper, the scientific contribution of the authors and the methodologies employed in the research. Through this meticulous examination, the authors of this work aim to offer valuable insights into the forefront of RL research, shedding light on the latest developments and laying the groundwork for future advancements in the sustainable and efficient Reinforcement Learning (RL) control on energy systems.

1.5. Paper Structure

The remaining sections of the paper are structured as follows: Section 2 illustrates a short and comprehensive review of the main RL terminology and theoretical background. Section 3 presents the main scientific RL implementations in modern power and energy systems such as RES, BEMS and EVCS. Section 4 will present a brief evaluation of the provided RL literature as regards the type of application and Section 5 integrates conclusions as considers RL implementations in energy systems arising from the current work.

2. Background and Overview of Reinforcement Learning

Reinforcement Learning is an approach in Machine Learning that aims to solve dynamic and complex problems, in which autonomous entities, called agents, are trained to take actions that will lead them to an optimal solution [48–50]. The agents learn to make optimal decisions as they exist in an environment and interact with it over time. The learning process is accomplished through a trial-and-error approach, as the agent's actions are taken after receiving feedback from the environment, in the form of a reward or penalty, for their previous action. The basic structure of RL can be seen in Figure 1, where an agent outputs an action and receives the new environment state and its reward, at timestep t. Finding the optimal solution should ultimately lead to the maximization of a cumulative reward over time, commonly referred to as the agent's return.

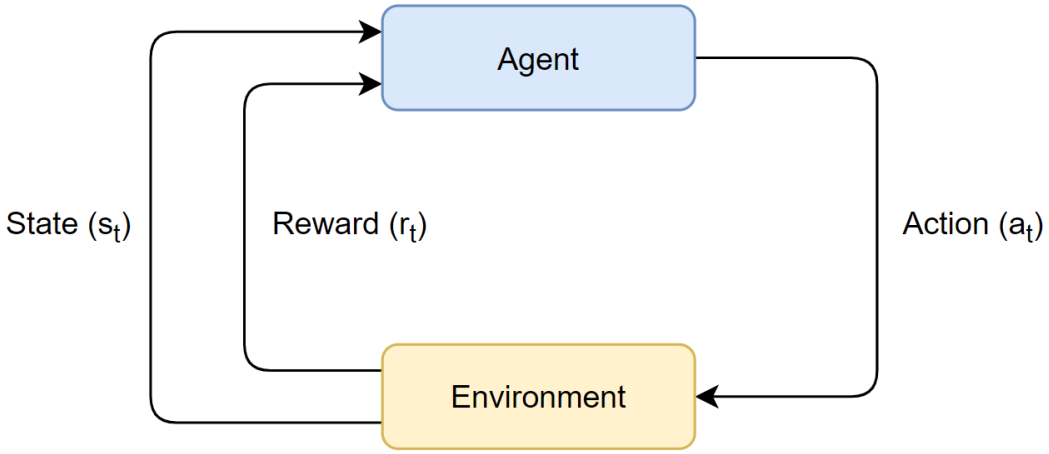

Figure 1. The Reinforcement Learning Framework.

2.1. Markov Decision Processes

The general mathematical formulation of the RL problem is a Markov Decision Process (MDP). An MDP is a mathematical model that describes the interaction between the agent and the environment. It follows the Markov Property, which states that the conditional probability distribution of future states of a stochastic process depends only on the current state and not the previous ones. This means that the history of previous states and their actions' consequences do not define the future states. The Markov Property is the basis for the mathematical framework of Markov models, which are widely used in a variety of fields, including physics, economics and computer science. MDPs describe fully observable RL environments and they consist of a finite number of discrete states. The agent's actions cause the environment to transition from one state to the next and the goal is to discover a policy that is associated with an action to maximize the expected cumulative reward or return, over time.

An MDP can be described by the tuple $(\mathcal{S}, \mathcal{A}, \mathcal{P}, \mathcal{R}, \rho_0, \gamma, \pi)$, where:

- $\mathcal{S}$ denotes the state space, which contains all the possible states an agent can find itself in;
- $\mathcal{A}$ denotes the action space, which contains all the possible actions an agent can take;
- $\mathcal{P}$ denotes the transition probability of the agent, $\mathcal{P}(s_{t+1}|s_t, a_t)$, where $t+1$ is the next time step and t the current one;
- $\mathcal{R}(s_t, a_t)$ denotes the reward function which defines the reward the agent will receive;
- ρ_0 denotes the start-state distribution;
- $\gamma \in [0, 1)$ denotes the discount factor, which determines the influence an action has in future rewards, with values closer to 1 having greater impact;
- $\pi(a|s)$ is the policy which maps the states to the actions taken.

In RL algorithms, the agent interacts with its environment in timesteps $t = 1, 2, 3, \ldots$ In each timestep the agent receives an observation of the environment state, $s_t \in S$ and then

chooses an action, $a_t \in A$. In timestep $t+1$, the agent will receive the arithmetic reward, $r_t \in R \subset \Re$, for the action taken. The policy in an MDP problem gives the probability to transition to state s when executing action a. The main objective in an MDP problem is to find an optimal policy, π^*, which ultimately leads to the maximization of the cumulative reward, given by Equation (1).

$$R_t = \sum_{\kappa=0}^{\infty} \gamma^t r_{t+k}. \quad (1)$$

2.2. Value Functions and Bellman Equations

In the RL framework, value functions have a critical role in estimating the expected cumulative reward that an agent will receive if it follows a particular policy and starts its course of action from a specific state or a state–action pair [48,49]. A fundamental distinction between value functions and reward functions is that the latter provide immediate rewards to the agent, according to its most recent action, whereas the former estimate cumulative rewards in the long term. Thus, the value function estimates the expected sum of rewards that an agent will obtain from a specific state over the entire duration of its interaction with the environment. This makes the value functions difficult to estimate, given that the value of a state or state–action pair must be estimated multiple times throughout an episode from state observations recorded during an agent's lifetime. Bellman Equations are mathematical expressions utilized for the purpose of computing the value functions in an efficient, recursive way. Specifically, the Bellman Equation expresses the value of a state or state–action pair as the sum of the immediate reward received in that state and the expected value of the next state under the policy being followed by the agent [48]. The Bellman Equation takes into account that the optimal value of a state is dependent on the optimal values of its neighboring states and it iteratively estimates and updates the value functions and the policy followed. This iterative process is repeated until the value functions converge to their optimal values and therefore reach the optimal policy. When the agent initiates its actions from a specific state, the value function is referred to as the 'state-value' function, whereas when the agent initiates from a specific state–action pair, it is then called the 'action-value' function. The state value describes the expected performance of an agent, when it follows policy π and has started from state s. The mathematical representation of a typical value function is as seen in Equation (2), while the Bellman Equation for the value function which expresses the dependency of the value of one state with the values of future states is seen in Equation (3).

$$V_\pi(s) = \mathbb{E}[R_t | s_t = s], \quad (2)$$

$$V_\pi(s) = \sum_a \pi(a|s) \sum_{s',r} p(s', r|s, a)[r + \gamma V_\pi(s')], \forall s \in S. \quad (3)$$

Respectively, the action-value describes the expected performance of an agent when it chooses action a, while it finds itself in state s and follows policy π. Then, the mathematical representation of the value function is as in Equation (4) and the Bellman equivalent is as in Equation (5).

$$Q_\pi(s, a) = \mathbb{E}[R_t | s_t = s, a_t = a], \quad (4)$$

$$Q_\pi(s, a) = \sum_{s',r} p(s', r|s, a)[r + \gamma \sum_{a'} \pi(a'|s') Q_\pi(s', a')], \forall s \in S, a \in A. \quad (5)$$

This equation is also known as the Q-function, which calculates the Q-value or 'quality value', of the RL agent. These equations have their optimal equivalents, in which the policies followed are the optimal policies of the agent. The optimal equations, which are given by Equations (6) and (7), are called 'Bellman Optimality Equations' and describe the

optimal action-value function in terms of the optimal value function. These equations in reality describe a system of equations, with each equation corresponding to one state.

$$V^*(s) = \max_{a \in A(s)} \sum_{s',r} p(s', r|s, a)[r + \gamma V^*(s')] \quad (6)$$

$$Q^*(s) = \sum_{s',r} p(s', r|s, a)[r + \gamma \max_{a'} Q^*(s', a')] \quad (7)$$

2.3. Reinforcement Learning Methods and Algorithms

The RL methods can be divided into two categories: model-based and model-free methods [49]. In model-based methods, there is a model which mimics the behavior of an environment and has the ability to observe a state and action and immediately determine the reward of the agent and the future state. The agent utilizes the model in such a way that it can simulate different sequences of actions and predict their outcomes. These methods can achieve faster learning as the model of the system is built beforehand. It can also provide higher sample efficiency, as it leads to the accurate generation of additional training data to improve the agent's performance. In model-free methods, the agent is trained effectively in complex and unknown environments, as it learns directly from the previous experience the agent has gathered. A model which could be inaccurate and lead to potentially unwanted results does not exist; therefore, the agent needs to rely solely on the trial-and-error approach. Model-free and model-based RL methods can, in turn, be divided into two more categories, policy optimization and Q-Learning, also known as policy-based and value-based methods, and learn the model given the model methods. This paper will focus on model-free methods and their subcategories, as the main advantages of RL are modeling and environment agnostic approaches. In conclusion of this section, a comprehensive table is presented, showcasing state-of-the-art RL algorithms. The classification of these algorithms is based on their respective approaches to policy optimization, Q-learning, or a combination of both (see Table 1).

2.3.1. Policy-Optimization and Policy-Based Approaches

Policy-optimization methods directly update the agent's policy and their objective is to discover the optimal policy, which will lead to the maximization of the expected cumulative reward, without explicitly learning a value function. They use stochastic policies, $\pi_\theta(a|s)$, and attempt to optimize the θ parameters of the policy through various techniques. One technique is the gradient ascent technique, which aims to find the maximum of an optimality function $J(\pi_\theta)$, as seen in Equation (8). Another technique is the maximization of local estimates of this function, as seen in Equation (9). Gradient-based policy-optimization algorithms are known as policy gradient methods, as they directly compute the gradient, $\nabla_\theta J(\pi_\theta)$, of the expected cumulative reward in terms of the policy parameters and use this gradient to compute the policy. Policy gradient algorithms are mostly on-policy methods, as in every iteration data gathered from the most recent policy followed is used. They can learn stochastic policies, which can be useful in environments with uncertainty; they can handle continuous action spaces and learn policies that are robust to environments with uncertainty.

$$\theta_{k+1} = \theta_k + \alpha \nabla_\theta J(\pi_\theta)|_{\theta_k} \quad (8)$$

$$\max_\theta J(\theta) = \max_\theta \sum_\tau P(\tau|\theta) R(\tau) \quad (9)$$

Policy-optimization methods also include techniques that introduce a constraint to ensure policy changes are not too large and maintain a certain level of performance. These are called trust region-optimization methods. Additionally, policy gradient algorithms include actor–critic methods, which consist of two components: the actor-network, or the policy function, responsible for the choosing of the agent's action, and the critic network,

or the value function estimator, responsible for the estimation of the value of state–action pairs. The actor receives the critic's estimate and proceeds to update its policy with respect to the policy parameters. Actor–critic methods leverage the benefits of both value-based and policy-based methods. Some of the most well-known policy-optimization methods are Proximal Policy Optimization (PPO) [51], Trust Region Policy Optimization (TRPO) [52], Vanilla Policy Gradient (VPG) [53], Deep Deterministic Policy Gradient (DDPG) [54], Soft Actor–Critic (SAC) [55], Advantage Actor–Critic (A2C) [56] and Asynchronous Advantage Actor–Critic (A3C) [56].

2.3.2. Q-Learning and Value-Based Approaches

Q-Learning methods [57,58] are characterized as off-policy learning methods that do not require a learning model, where the agent learns an action-value function, the Q-function, $Q_\theta(s, a)$, for the optimal action-value function, $Q_*(s, a)$. This function estimates the expected return of selecting action a, in state s, and following the optimal policy from the next state, s'. The Q-function's objective is to maximize quality values of a state, the Q-values, by choosing the appropriate actions for each given state. It simultaneously updates the Q-values of that state. The Q-function is given by Equation (10).

$$Q_\theta(s, a) = \mathbb{E}[R(s, a)] + \gamma \sum P(s'|s, a) \max(Q(s', a')) \tag{10}$$

The best actions of the agent for a state, s, are selected through a greedy method, called ϵ-Greedy. The selection of these actions happen with a probability $1 - \epsilon$, while the selection of a random future action happens with a probability of ϵ, where $\epsilon \in (0, 1)$.

$$a(s) = \arg\max_{\alpha} Q_\theta(s, a) \tag{11}$$

The ϵ-Greedy method is an approach to the exploitation versus exploration problem, where the agent needs to find an equilibrium between exploiting the selection of previously found good actions which yield satisfying rewards and exploring new and potentially better, unknown actions. Q-Learning techniques are characterized by their simplicity and ease of implementation. They are utilized in continuous and discrete state and action spaces and have the ability to learn from experiences that do not have an optimal policy. Some well-known extensions and improved methods that are derived from Q-Learning are Double Q-Learning [59], Deep Q-Networks (DQNs) [60] and Dueling Q-Learning [61]. Double Q-Learning attempts to solve the problem of overestimation of Q-values in certain scenarios by modifying the original Q-Learning algorithm and learning two Q-functions independently as estimators. The Q-function that yields the highest Q-value is the primary estimator whose Q-value will be selected. The other Q-function is the secondary estimator used to estimate the Q-value of the selected action, which helps reduce the overestimation bias that can occur in standard Q-Learning. DQNs make use of deep neural networks which take the current state as input and outputs the estimated Q-value of the Q-function. This is achieved by updating the weights of the network and minimizing the mean squared error between the predicted and actual Q-values. Dueling Q-Learning makes use of a state-value function, which estimates the value of a particular state and an advantage function, which estimates the advantage of taking an action in that state. The Q-value is the sum of the advantage and state-value functions, minus the mean of the advantage of all previous actions.

Table 1. State-of-the-art RL algorithms.

Policy Optimization	Q-Learning	Policy Optimization and Q-Learning-Based
Vanilla Policy Gradient [53]	Double Q-Learning [59]	DDPG [54]
PPO [51]	DQN [60]	SAC [55]
A2C/A3C [56]	Dueling Q-Learning [61]	TD3 [62]
TRPO [52]	C51 [63]	

3. Review of RL-Based Methods for Optimal Control of Energy Systems and Smart Grids

Reinforcement Learning (RL) is particularly useful in situations where the decision-making agent is required to operate in a complex and dynamic environment, such as in energy systems. To this end, RL optimal control has been extensively applied in various fields, including Renewable Energy Source frameworks (RES), Building Energy-Management System (BEMS) control and Electric Vehicle Charging Stations (EVCSs).

This section illustrates numerous essential research works of value-based, policy-based, actor–critic and hybrid control methodologies. By investigating the conceptual scheme of each publication, this section has as its primary aim to briefly analyze the novelty, the experimental procedure and the final outcome of each related work. The literature work concerns the period 2015–2023 and integrates highly cited and valuable research efforts considering RL in RES, buildings and EVCS energy systems as depicted in the literature. Figure 2 portrays the architecture of the current research effort illustrating the applications, RL type and RL algorithmic methodology considered in the examined papers.

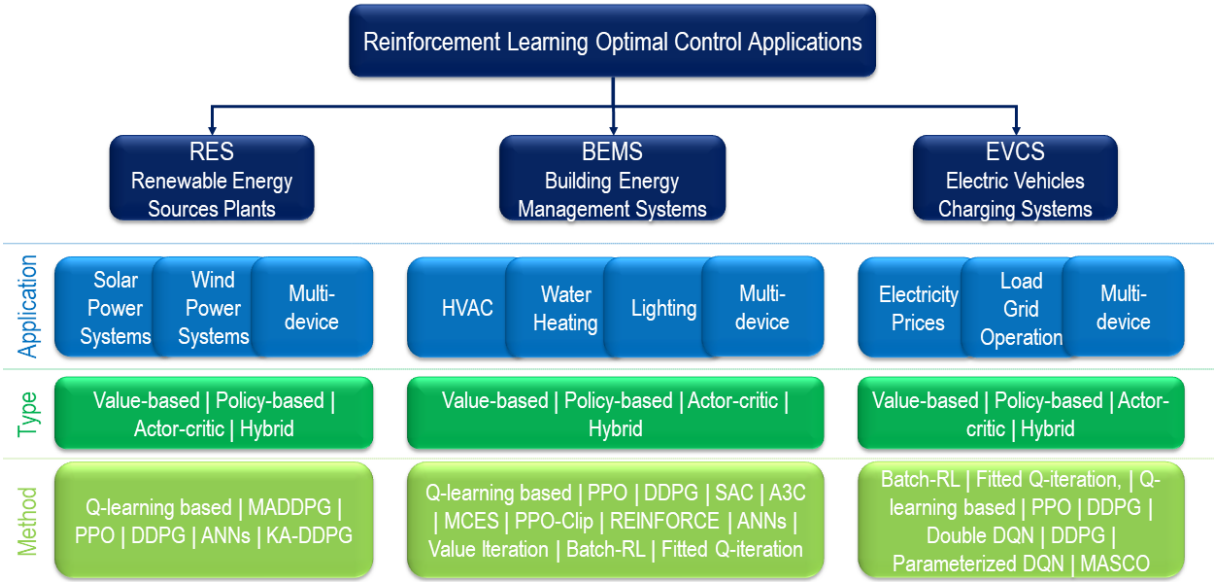

Figure 2. Architecture of the current work, concerning the type of application, the RL type and the RL algorithmic methodology utilized in the examined literature.

3.1. Reinforcement Learning towards Optimal Control in Renewable Energy Source (RES) Plants

Renewable Energy Source (RES) integration with the Energy Grid (EG) is vital for optimally controlling the storage and usage of the electrical energy they produce. By utilizing RESs to charge batteries and feed the grid with additional power in a provisional manner, grid energy demand and production are moderated and energy costs are reduced. The literature integrates most primarily value-based RL such as Q-Learning and Deep Q-Networks (DQNs), which are commonly used in Renewable Energy Sources (RESs) in comparison to policy-based or actor–critic RL control. Hybrid approaches are also present in an effort to exploit the unique benefits of each separate approach. At the end of this

section, Table 2 is introduced as an extensive table which presents a summary of works for Optimal RL Control in RES plants.

3.1.1. Value-Based RL in RES

Kuznetsova et al. (2013) introduced a two-step-ahead RL methodology for a microgrid that involved an end-user, a renewable energy source and an energy-storage system [64]. Their objective was to schedule the battery charging and discharging periods effectively, taking into consideration the generator's stochasticity and random mechanical failures. The agent, in this case, the consumer, wishes to establish optimal and profitable policies on energy use according to the environment's parameters, the battery status and the energy generated by the renewable RES producer. The algorithm presented in that paper was based on the Q-Learning technique and was trained for simulation times of 40 years. It has achieved up to 12% improvement, relative to cases where the consumer does not have explicit goals to follow to select optimal actions.

In their research published in 2015, Wei et al. introduced an intelligent algorithm based on Reinforcement Learning for achieving Maximum Power Point Tracking (MPPT) in variable-speed wind energy-conversion systems [65]. The controller of the wind energy-conversion system utilized a model-free Q-Learning algorithm, enabling it to learn an optimal control action map online. The action values are updated based on received rewards and stored in a *Q*-table. Following an online learning phase, the acquired knowledge of maximum power points is applied to establish an optimal speed–power curve, enabling efficient and rapid Maximum Power Point Tracking (MPPT) control for the wind energy-conversion system. An advantage of the Reinforcement Learning approach was that the wind energy-conversion system learns directly from its interaction with the environment, by eliminating the requirement for precise knowledge of wind turbine parameters or wind speed data; the proposed MPPT control algorithm operates independently. To validate the effectiveness of this algorithm, comprehensive simulation studies are conducted utilizing a 1.5 MW wind energy-conversion system based on a doubly fed induction generator. Additionally, experimental results are obtained by utilizing a 200 W wind energy-conversion system emulator based on a permanent magnet synchronous generator.

The same group of scientists, Wei et al. [66], in 2016, proposed an intelligent algorithm based on Artificial Neural Networks (ANNs) and Reinforcement Learning (RL) for Maximum Power Point Tracking (MPPT) in Wind Energy-Conversion Systems (WECSs) with Permanent Magnet Synchronous Generators (PMSGs). The algorithm incorporates Artificial Neural Networks (ANNs) and Q-Learning to acquire the optimal correlation between rotor speed and electrical power output of the Permanent Magnet Synchronous Generator (PMSG). This correlation guides the control of the Wind Energy-Conversion System (WECS) for achieving Maximum Power Point Tracking (MPPT). In cases where the initial correlation becomes obsolete due to factors such as system aging, the online Reinforcement Learning (RL) algorithm adapts and acquires a new optimal relationship. This online learning capability enables continuous performance improvement of the WECS through experiential learning. The proposed MPPT control algorithm's effectiveness is demonstrated through simulation of a 5 MW PMSG-based WECS and experimentation on a 200 W PMSG-based WECS emulator. Moreover, the method is extendable to other types of WECSs, including those based on doubly fed induction generators.

Kofinas et al. [67] introduced a novel RL MPPT control method, utilizing Reinforcement Learning (RL) techniques, which enables tracking and adjustment of the maximum power point of a PV without the requirement for previews of data utilization. A Markov Decision Process (MDP) model was defined for the photovoltaic process involved in Maximum Power Point Tracking (MPPT) and an RL algorithm was introduced and verified across various photovoltaic sources. The primary benefit of the RLMPPT control method was depicted in its versatility since it could sufficiently be applied to different PV sources with minimal setup time. The performance of the RLMPPT algorithmic approach took place using simulations that concerned different operational scenarios under different con-

ditions. Additionally, a comparison with the state-of-the-art perturb and observe method was conducted. The results demonstrated the rapid adaptation and the optimal behavior of the approach, without the requirement for previous operational data.

In 2019, Remani et al. [68] proposed a comprehensive RL approach for the resolution of the load commitment problem (LCP) in residential settings, considering the presence of RES, irrespective of the price type. The LCP involves making decisions on which of the appliances and generation units in a system should be online and which offline, in certain time intervals. The goal of solving the LCP was load scheduling to minimize generation costs and maximize energy adequacy to meet the demands of the system. The novelty of the work by Remani et al. was the introduction of a comprehensive model that considers user comfort, uncertainties concerning RES energy availability and costs for providing an implementable solution. The RL approach presented is Q-Learning-based and takes into account the different environmental variables, such as consumer comfort, stochastic renewable power and energy provider tariffs. Simulation experiments took place in order to assess the effectiveness and adaptability of the concerned methodology. The efficiency of the approach was evaluated using a scenario that involves the occupant user with both programmable and non-programmable household devices, along with PV energy generation. Directions were also provided for selecting the parameters of the load-scheduling model. The simulation time was 90 (24-h) days and the results were separated into two cases. Initially, load scheduling was accomplished without the renewable distributed generator, the PV, while in the second case, the PV was present; the RL Q-Learning-based approach showed slightly better results for scheduled and unscheduled costs and shorter computation times. Overall, this approach offered valuable insights and practical solutions for offering a resolution to the difficulties encountered in programming household loads when RES sources are present, emphasizing occupant's comfort, RES power generation variability and tariff considerations.

In 2019, Diao et al. [69] created an AI agent for voltage control, called Grid Mind, which can learn behaviors via interactions of offline simulations and is adequate for absorbing new changes in the environment, related to both load/generation and topological changes. Grid Mind was a Deep Q-Network-based AI framework, with the main objective of optimizing autonomous grid control in real time, while targeting current and near-future operating conditions. The results show that the agent succeeds in learning optimal policies and receives high numerical rewards after a certain number of episodes. More specifically, there are three cases the authors consider: (i) there are no contingencies in cases of topology changes or malfunctions, (ii) there are contingencies for different changes and malfunctions taking place and (iii) the agent can take only one action towards deciding on the optimal voltage profile for the system.

In the same year, Rocchetta et al. [70] developed a Reinforcement Learning (RL) framework that focuses on optimizing the supervision and sustainment of electricity networks. The power grids were equipped with Prognosis and Health Management (PHM) capabilities, aiming to enhance their Operation and Maintenance (O&M) processes. For the RL framework, the Q-Learning method was utilized in combination with Artificial Neural Networks (ANNs), as ANNs have the ability to effectively replace the traditional planar operation of the state–action value function, enabling Q-Learning to handle complex and continuous environments more efficiently. The test environment was a scaled-down power grid, which incorporates various components, including RES technologies, controllable energy systems, maintenance breaks, forecast and malfunction protection equipment. The RL objective was the maximization of the grid's load balance and the anticipated long-term earnings, while taking into consideration the additional uncertainties of RES in the environment. Two different policies are tested. One takes random actions, while the second takes actions based on the experience gathered. The results indicate that the non-tabular RL framework in combination with ANNs achieved results that are comparable with the Bellman Optimality. They further indicate that the developed algorithm's policy

was a significant improvement when compared to both the random policy and the solution derived from domain specialists to address the optimization challenge.

Another interesting study of 2019, conducted by Zhang et al. [71], proposed a DRL framework that was grounded on the Proximal Policy Optimization (PPO) algorithm. The PPO-based renewable energy-conversion algorithm (PPO-RECA) aimed at reducing the user operational costs in an alternating manner and simultaneously optimizing the energy management of an integrated energy system (IES) with colored RES. This particular framework took into account the uncertainties associated with the user's load demand cases, the variability of spot energy costs and the unpredictable nature of wind power generation. It was tested on a renewable energy IES (RE-IES) environment with a hierarchical structure, where the system operator was responsible for the formulation of the policy followed for energy transition towards heat and electricity demand status and electricity costs provided from an upper-level grid. Furthermore, the trained model was tested on multiple days and the simulation outcomes showcase the algorithm's robustness and adaptability in handling unpredictable scenarios. Through the comparison with algorithms such as Particle Swarm Optimization (PSO) and Deep Q-Networks (DQNs), the algorithm's superiority and its potential to address complex optimization challenges in renewable energy systems are highlighted. It achieves the lowest operating cost for a training period of 30 days, showing an improvement of up to 8.41%.

Taking advantage of the challenges of microgrid operation due to stochastic uncertainties of integrating distributed renewable energy as well as the load demand random variations, Ji et al.—in 2019—introduced an RL methodology for real-time programming of a microgrid considering variations in load demand, RES generation and costs [72]. The challenge was formulated as a Markov Decision Process (MDP) model, aiming to reduce the expenses of the grid. To address such dynamic behavior without an explicit model, a DRL approach was considered. To this end, the MDP was solved using the DQN approach, leveraging a deep feed-forward ANN to determine the value function. The inputs to the network were the current state of the MG and the results provided straightforward actual time generation schemes. By utilizing actual power-grid information—arising from CAISO—the results demonstrated that the considered DRL approach surpassed the state-of-the-art RL practices. Additionally, the approach effectively predicted the trend of such dynamic behavior, without the requirement of an established model.

On the other hand, Phan et al. [73] showcased a maximum power point tracking method grounded on RL to enhance RES generation. The study was dedicated to the advancement of Hybrid Renewable Energy Systems (HRESs) through the integration of storage and hydrogen Fuel Cells (FCs). This work addressed key challenges involved in the system design process, concerning optimal sizing, Maximum Power Point Tracking (MPPT) control and the energy-management system. To determine the best possible design for the HRES, scientists utilized the available information from the Basco area in the Philippines. HOMER software, (https://www.homerenergy.com/, Accessed: 1 May 2023) known for its cost-effectiveness, reliability and environmental friendliness, facilitated the optimization process. According to the examination results, the best potential setup of the energy system concerned PVs (5483 kW), 236 wind turbines (10 kW), storage equipment (20,948 kW–48V DC, integrating 4 modules and 5237 strings), fuel cells (500 kW), a diesel generator (750 W), an electrolyzer (3000 kW), an H-tank (0.5 tons) and a converter (1575 kW). The overall expense was calculated at 0.774/kWh USD, considering a fuel cost of 1 USD/L. Furthermore, the analysis results highlight the integration of a fuel cell framework and storage technology as one of the most favorable potentials for HRES design. Fuel cells serve as a long-term energy-storage solution, while batteries function as short-term energy-storage mediums. The proposed system exhibited cost efficiency and its capability of meeting the required load demand in such frameworks.

Also in 2019, Saenz-Aguirre et al. [74] proposed an original approach to variable speed wind turbine control, utilizing Reinforcement Learning (RL) techniques. The main objective of the yaw control methodology was to determine the optimal wind turbine orientation

adjustment in order to maximize power generation while considering mechanical limitations and loads. The proposed method incorporates the RL Q-Learning algorithm. Initially, a power curve model was established to represent the power generated by an existing onshore wind turbine. Along with a proportional regulator, this helps to create a dataset that captures the behavior of the wind turbine under various yaw control commands. This knowledge was then utilized for learning the optimal control action for every wind turbine state based on the yaw angle, which was stored in a Q(s,a) matrix. To address matrix-management and quantification challenges, an Artificial Neural Network (ANN) with Multilayer Perceptron and Backpropagation (MLP-BP) was employed to model the matrix. Finally, the performance of the synthesized optimal yaw controller was evaluated using wind speed scenarios generated via the TurbSim (1.06.00, NREL, Golden, CO, USA) software application, enabling an analysis of the algorithm's response to different wind speeds. The results of simulations conducted in the Simulink environment provided by MATLAB, (https://www.mathworks.com/products/matlab.html, Accessed: 1 May 2023) as presented in this paper, have demonstrated favorable validation outcomes and indicate that the developed control strategy was well-suited for effectively regulating the yaw angle of wind turbines under varying wind conditions. Furthermore, the advanced RL-based control algorithm devised in this study successfully addressed a significant limitation of conventional control strategies—the requirement for manual tuning of control parameters. In contrast, Liu et al. [75] introduced a novel hybrid model for forecasting wind speeds . The modeling process comprises three main steps. In the first stage, the empirical wavelet transform method was employed to handle the non-stationarity caused by the original wind speed data, by breaking it down into multiple subseries. In the second stage, three different types of deep neural networks (LSTM, DBN and ESN) were utilized to construct individual forecasting models and generate prediction results for each subseries. In the third stage, a Reinforcement Learning method was applied to integrate the outputs of the three deep networks. This integration process combines the forecasting results from each subseries to obtain the final predictions. Comparative analysis of the projected results across multiple kinds of wind speed series showed that the methodology effectively integrates the three deep networks and outperforms classic optimization-based combination methods. Additionally, the ensemble deep Reinforcement Learning model for predicting wind speeds consistently yields accurate results across all cases and exhibits superior accuracy compared to sixteen alternative models and three state-of-the-art models. Finally, in 2021, Jeong et al. [76] introduced a new methodology called 'error compensable forecasting' that shifts the objective of forecasting from error reduction to making errors compensable using a battery system. By utilizing the battery, the strategy aimed to minimize the dispatched error, which represents the deviation between the projected value and the actual dispatched value. The primary challenge of this objective arises from the fact that the stored energy at the current time was influenced by the previous forecasting results. To address this challenge, a Deep RL-based framework named DeepComp was proposed, integrating forecasting within a control loop. In DeepComp, the forecasted value served as a continuous action, necessitating a continuous action space. Proximal Policy Optimization (PPO) was employed as the Reinforcement Learning algorithm, known for its simplicity of implementation and outstanding performance in continuous control tasks. Experiments were conducted using solar and wind power generation data to assess the performance of DeepComp. The results demonstrated that DeepComp outperformed traditional forecasting methods by up to 90% and achieved accurate forecasting with a mean absolute percentage error ranging from 0.58 to 1.22%.

3.1.2. Policy-Based, Actor–Critic and Hybrid RL in RES

In 2020, Cao et al. [77] suggested a novel distributed DRL methodology—abbreviated as MADRL—presented for voltage regulation in multi-agent systems with a significant presence of photovoltaics (PVs) using the distributed DRL framework. The agents are designed to acquire coordinated control strategies by training local policy networks and central critic

networks using historical data. This enables an effective networked distributed control. The comparison review between these alternative methods demonstrated the superior control performance of the examined approach across diverse conditions. The suggested approach exhibited broad applicability and can be readily expanded to encompass diverse integrated systems, including PV, wind and various other types of distributed generation (DG) systems.

Table 2. Summary of works of Optimal Control RL methods for RES.

Reference Year	Field	Methodlogy	Type
[67] 2017	Solar Power Systems	RLMPPT Q-Learning	Value-based
[68] 2019	Solar Power Systems	Q-Learning	Value-based
[73] 2019	Solar Power Systems	hybrid P&O w/Q-Learning	Value-based
[77] 2020	Solar Power Systems	MADDPG	Policy and Value-based, Actor–Critic
[76] 2021	Solar Power Systems	PPO	Policy-based
[65] 2015	Wind Power Systems	MPPT Q-Learning	Value-based
[66] 2016	Wind Power Systems	ANN MPPT Q-Learning	Value-based
[74] 2019	Wind Power Systems	ANNs & Q-Learning	Value-based
[75] 2020	Wind Power Systems	ANNs & Q-Learning	Value-based
[78] 2020	Wind Power Systems	KA-DDPG	Policy and Value-based, Actor–Critic
[64] 2013	Multi	Q-Learning	Value-based
[69] 2019	Multi	DQN	Value-based
[79] 2022	Multi	PPO	Policy-based

The same year, Zhao et al. [78], introduced an innovative RL approach suitable for cooperative wind farm control in scenarios with wind speeds that change over time. To enhance safety and expedite the learning process, the Knowledge-Assisted (KA) RL framework was introduced by integrating an analytical model with RL techniques. Furthermore, a novel algorithm called KA-DDPG was presented. Within the KA learning framework, three methods, namely protection, criteria action and guiding reward methods, are employed. These methods contribute to rejecting unsafe actions, identifying safety areas and facilitating the learning process. Simulation results showcased that the Reinforcement Learning methodology exhibited an average output that was 10% higher than the greedy method and 5% higher in comparison to the PARK model in the simulated environment. Additionally, the KA-DDPG framework ensured safety requirements while achieving faster learning compared to DDPG. According to the final evaluation, the learning process was accelerated directly by the guiding reward method, whereas the criteria action method facilitated faster learning by reaching the safety action area earlier. In order to enable optimal schedule decisions for the Energy-Management System (EMS), Guo et al., in the year 2022 [79], introduced a real-time dynamic Optimal Energy-Management (OEM) solution based on a Deep RL (DRL) algorithm. The proposed DRL algorithm formulated the MG-OEM as a Markov Decision Process (MDP) while considering environmental uncertainties

and employed the Proximal Policy Optimization (PPO) algorithm for solving it. PPO was a policy-based Deep Reinforcement Learning (DRL) algorithm specifically designed to handle continuous state and action spaces. The algorithm consists of two phases: offline training and online operation. During the training process, PPO learns from historical data to capture the uncertain characteristics of renewable energy generation and load consumption. The proposed method was showcased through a case study, demonstrating both its effectiveness and computation efficiency. Overall, the real-time dynamic OEM approach based on DRL offered significant advantages over traditional methods, providing more accurate and efficient decision-making capabilities for the EMS.

3.2. Reinforcement Learning towards Optimal Control in Building Energy-Management Systems (BEMSs)

Building Energy-Management Systems (BEMSs) aim to control the energy usage within a building, taking into account factors such as climate, occupants' behavior and electric device utilization. To this end, the vast majority of research efforts utilize Reinforcement Learning (RL) techniques for the efficient optimization control of various building-related systems: heating, ventilation, air conditioning (HVAC), water heating, lighting systems or even control frameworks that integrate multiple devices of diverse nature. Having as primary concerns to reduce energy wastage, to minimize energy costs and to upgrade indoor comfort and air quality, the relative literature includes several simulation approaches as well as real-life results illustrating the adequacy of RL—and particularly DRL—to act as a straightforward optimization control framework for buildings operations [80]. In summary of this section, Table 3 overviews the works for Optimal RL Control in BEMSs.

3.2.1. Value-Based RL in BEMS

Efficiently controlling heating, ventilation and air-conditioning (HVAC) systems is essential to enhancing the demand side of energy management. However, concerning building thermodynamics, the presence of uncertainties in human activities complicates the task of effective management.Although model-free Reinforcement Learning offers many advantages over existing methods, the existing literature mostly focuses on value-based approaches that struggle to handle complicated HVAC systems. In 2015, Barrett et al. [81] introduced an RL architecture for an intelligent thermostat, which can autonomously control an HVAC system with the goal of optimizing both occupant comfort and energy costs. To enable RL to control the HVAC system successfully, a novel state–action space formalism was proposed. The results show that the RL approach resulted in cost savings of up to 10% in comparison with a programmable thermostat while maintaining high levels of occupant comfort. In 2016, Ruelens et al. [82] employed fitted Q-iteration, a Batch-RL method, to derive a control policy based on a feature representation. In a simulation experiment where an electric water heater with fifty temperature sensors was utilized, the introduced technique achieves good policies at a faster rate, in comparison with using the full state information. In a laboratory experiment utilizing an electric water heater equipped with eight temperature sensors, reducing the state vector did not improve the outcomes of fitted Q-iteration. The 40-day laboratory experiment yielded results which indicate that the approach outperforms a thermostat controller, reducing the electric water heater's total energy-consumption cost by 15%.

In Al-Jabery et al.'s study [83], the management of electric water heaters in domestic demand-side systems was addressed using Q-Learning and action-dependent heuristic dynamic programming (ADHDP) techniques. Simulation results showed that both Q-Learning and ADHDP techniques can mitigate the energy-consumption cost of DEWHs by approximately 26 and 21%, respectively, while minimizing energy consumption during peak load periods. Customers utilizing ADHDP techniques to manage the operation of their 100-gallon-tank-size DEWHs can save up to USD 466 and 367 annually, respectively, which was better than the cost reduction achieved by advanced control techniques employing equivalent simulation parameters. An interesting study was proposed by Cheng et al.

in 2016 [84], where the RL approach is combined with a human feedback mechanism for regulating the blinds and lights in a solitary office within a building. The primary aim of this work was to optimize both lighting comfort and energy conservation while greatly emphasizing comfort. They discovered that their approach improved luminosity from both the perspective of comfort and energy savings. Specifically, compared to manual and conventional integrated automated controls, their RL technique considerably reduced energy usage. Another interesting study of 2017 [85] examines a fruitful Deep RL methodology for the optimal control of HVAC systems. Using the well-evaluated EnergyPlus simulation framework, the algorithm outperformed the RBC conventional control baseline approach and conventional Q-Learning. Tests using accurate EnergyPlus models and actual weather and pricing data showed that the utilized DRL-based algorithms—which include both a regular DRL algorithm and a heuristic adaptation for multi-zone control—were more efficient at reducing energy costs while still keeping the room temperature at a comfortable level.

In 2018, Chen et al. [86] proposed an RL control strategy based on model-free Q-Learning to optimize the operation of both HVAC and window systems in order to minimize energy usage and thermal discomfort. At every timestep, the control system evaluated the indoor and outdoor environments, including temperature, humidity, solar radiation and wind speed, to generate the optimal control decisions as regards the respective short- and long-term goals. The efficiency of the control approach was assessed through simulations on a thermal model of the building and compared with a rule-based heuristic control strategy. The effectiveness of Reinforcement Learning Control was demonstrated in simulation mode through case studies in hot and humid climates (located in Miami, USA), as well as warm and mild climates (located in Los Angeles, CA, USA). According to the results, the RL control outperformed the heuristic control strategy, resulting in a 13 and 23% reduction in HVAC system energy consumption, 62 and 80% fewer discomfort degree hours and 63 and 77% fewer high-humidity hours in Miami and Los Angeles, respectively. Taking advantage of historical data related to past building–controller interactions, Jia et al. [87] in a research group in 2019 utilized an RL technique and managed to establish a simulation framework that mimics buildings operating dynamically in order to verify and examine different types of Reinforcement Learning algorithms. It was noticeable that the aforementioned framework may further enhance the RL learning process for the HVAC operation by integrating knowledge from human experts as well. Jia et al. tested their approach via a simulation and concluded that their RL methodology worked better than the conventional RL practice control method used in buildings holding significant future potential for further large-scale implementations.

In another significant research effort in 2019, Valladares et al. [88] proposed a DRL AI algorithm targeting balancing optimal thermal comfort and air quality levels while reducing energy consumption from air-conditioning units and ventilation fans. The agent was trained for 10 years of simulated data that concerned a laboratory and classroom environment with up to 60 users. The AI agent managed to successfully achieve a balance between thermal comfort, indoor air quality and energy consumption, resulting in a superior PMV (an index that estimates the thermal comfort sensation of individuals, based on the seven-point scale—+3 hot, +2 warm, +1 slightly warm, 0 neutral, −1 slightly cool, −2 cool, −3 cold) and 10% lower CO2 levels compared to the existing control system, while consuming 4–5% decreased energy consumption. Kazmi et al. [89] in 2019 evaluated the multi-agent modeling and control framework in a large-scale pilot, consisting of over 50 houses, without using any thermostatically controlled load-specific information. Taking into account the storage container, the heating component, human occupant behavior and surrounding conditions, the scientists examined the performance of the RL algorithm on different variations of these factors. The experiment lasted for an entire year and resulted in energy conservation of almost 200 kWh per household, which translates to 20% of the energy required for hot water production. In 2019, Park et al. presented the LightLearn framework [90], which was a system that uses Reinforcement Learning for occupant-centered lighting control in office environments. LightLearn was designed to

adapt its control parameters considering individual occupant behavior and indoor environmental conditions to define customized set points. An 8-week experiment conducted in five offices demonstrated that LightLearn was adequate for learning occupant behaviors and thus improving lighting conditions while balancing occupant comfort and energy consumption. The performance of LightLearn was evaluated against scenarios based on schedules and occupancy for control and a new metric called light–comfort ratio was introduced. Results indicate that only LightLearn successfully balanced occupant comfort and energy usage and highlight that Reinforcement Learning-based occupant-centered control is a promising approach for addressing the discrepancy between occupant comfort and building control goals. Finally, the OCTAPUS framework was proposed by Ding et al. [91]. OCTAPUS integrates a novel DRL framework adequate for finding the optimized control strategies for all building subsystems, incorporating lighting, blinds, window systems and HVAC. A data-driven approach was employed in this framework and the DRL framework includes a unique reward function that explores the balance between energy usage and residents' comfort. The control problem, due to interactions between the four aforementioned building subsystems, was highly dimensional and was also addressed. To manage the data-training requirements of OCTOPUS, scientists contend that calibrated simulations matching the target building operational points are necessary for sufficient training to be acquired for training the DRL framework to find the solution to the control problem. OCTOPUS training utilized 10-year weather data in combination with a building model that was constructed in the building simulator provided by EnergyPlus and calibrated using data from an existing production building. The demonstrated simulations illustrate that OCTOPUS was adequate for achieving energy-consumption reductions that reached 14.26 and 8.1%, when compared to the cutting-edge RBC method in a LEED Gold Certified building and the latest DRL-based method available in the literature, respectively, while ensuring residents' comfort remained within a desired range.

Another fruitful study was carried out by Brandi et al. in 2020 [92]. This study presents the implementation of Deep RL for regulating the temperature setpoint of the supply water in a heating system. The experiment was conducted on an office building using an integrated simulation environment. The efficiency evaluation of various Deep Reinforcement Learning (DRL) control agents, which have been trained with distinct input variable sets and are deployed using different methodologies, were conducted against the reference control (RBC) of supply water temperature to heating system terminal units. The potential of the proposed solution was demonstrated by testing the control agent's adaptability to various occupancy schedules and indoor temperature requirements in different scenarios. The study illustrated the importance of input variable selection in order to achieve energy savings (5–12%) with improved indoor temperature control and dynamic deployment. In the year 2021, Lissa et al. [93] proposed a novel approach for controlling indoor and domestic hot water temperatures using a Deep RL (DRL) algorithm. The primary objective was to optimize the utilization of photovoltaic (PV) energy production and reduce energy consumption. Additionally, a new methodology for defining dynamic indoor temperature setpoints was introduced, which increases flexibility and energy savings. The experimental outcomes demonstrate that the proposed algorithm with a dynamic setpoint leads to an average of 8 to 16% energy reduction when compared to an RBC algorithm, observed during the summer period. The algorithm also ensures user comfort, with less than 1% of time spent outside of the specified temperature setpoints. Another important 2021 study by Jiang et al. [94] focused on reducing energy costs by taking into account changes in electricity prices and demand charges by utilizing a Deep RL technique. Demand charges are fees that power companies charge for using a lot of electricity during peak hours. To save costs, the researchers created a computer program called Deep Q-Network with an action processor that can make decisions based on past data and future predictions. They added a reward function that considers both the energy cost and a penalty for occupants' discomfort. Their approach was also focused on eliminating the problem of sparse rewards caused by demand charges. According to the verification and evaluation procedure on the

simulation environment, their approach managed to decrease the overall energy costs by 6–8% of total costs compared to the conventional RBC policy. Finally, the same year, another study dealing with heating control was carried out by Gupta et al. [95]. The study proposes a DRL heating controller to enhance thermal comfort and reduce energy costs in smart buildings. The controller's performance was assessed through extensive simulation experiments that utilized real-world outdoor temperature data. The obtained results demonstrate that the proposed DRL-based controller outperforms traditional thermostat controllers, providing a 15 to 30% improvement in thermal comfort and a 5 to 12% reduction in energy costs. Current research integrates a secondary set of experiments in order to investigate the performance of a centralized DRL-based controller in comparison with decentralized control, where each heater has its own DRL-based controller. The findings reveal that the decentralized controller performs better in comparison with the centralized one, when there is an increase in the number of buildings and differences in their setpoint temperatures.

3.2.2. Policy-Based, Actor–Critic and Hybrid RL Approaches in BEMS

De Somer et al. [96] in 2017 suggested a scheme to effectively optimize the heating cycles of the Domestic Hot Water (DHW) buffer to maximize the utilization of local photovoltaic (PV) production. To this end, a model-based RL technique was used. The algorithm considers the stochastic behavior of the occupants, predicts PV production and takes the system dynamics into account. The algorithm was tested in a real-life experiment involving six residential buildings, where it was observed that the self-consumption of the PV production was increased significantly as compared to the default thermostat control. According to real-life results that concerned a period of four months, the deployed algorithm significantly increased PV self-consumption compared to the baseline thermostat control approach. Since classical model-based optimal control (MOC) was not sufficient enough for utilizing whole building energy models (BEMs)—due to their high-dimension nature and intensive computational speed—Zhang et al. in 2018 [97] introduced a novel DRL framework for MOC of HVAC systems using BEM. The concerned case study integrated a real-office building in Pennsylvania, USA, and demonstrated the workflow including the building modeling, model calibration and DRL training. The optimal control policy framework resulted in 15% potential reductions in heating energy consumption by controlling the supply water temperature of the heating system. The same year, Gao et al. [98] proposed a novel RL framework—abbreviated as DeepComfort—for optimizing energy consumption and maintaining thermal comfort in smart buildings using Deep RL. They cast the thermal control of the building as a cost-minimization problem that considers both the energy consumption of the HVAC and occupants' thermal comfort. The approach employs an FNN-based deep neural network to predict occupants' thermal comfort and uses Deep Deterministic Policy Gradients (DDPGs) to learn the thermal control policy. The current approach was evaluated using a building thermal control simulation system under various settings. According to the evaluation of Gao et al., the current methodology led to a 14.5% improvement in thermal comfort prediction performance and a 4.31% reduction in HVAC energy consumption, while also improving occupants' thermal comfort by 13.6%.

In 2020, Azuatalam et al. [99] introduced a holistic framework that incorporates an efficient RL controller within a whole building model, which optimizes and controls the HVAC system so as to improve energy efficiency, maintain thermal comfort and achieve the relative demand response objectives. Simulation outcomes demonstrated that RL utilization for normal HVAC operation may lead to up to a 22% weekly energy reduction, compared to a handcrafted baseline controller. Additionally, during demand response periods, by using a demand response-aware RL controller, the average power reductions or increases can reach up to 50% on a weekly basis when compared to the default RL controller, while still maintaining acceptable occupant thermal comfort levels. Another important work was carried out by Du et al. [100] in 2021, proposing a novel approach to optimize multi-zone residential HVAC systems using a DRL methodology. The goal was to minimize energy-consumption costs while maintaining user comfort. The applied methodology

(DDPG) proved to be sufficient for learning through ongoing interaction with a simulated building environment, even in the absence of prior model knowledge. The simulation results indicate that the DDPG-based HVAC control strategy outperformed the state-of-the-art Deep Q-Network (DQN) by reducing energy-consumption costs by 15% and comfort violation by 79%. Moreover, compared to an HVAC RBC strategy, the DDPG-based strategy reduces comfort violation by 98%.

During the same year, Pinto et al. [101] proposed a novel methodology that employs Deep RL and Long Short-Term Memory (LSTM) neural networks for data-driven control of heat pumps and storage systems for four buildings. The simulation environment includes the use of LSTM models that are trained on a synthetic data set from EnergyPlus to assess the dynamics of indoor temperatures. According to the results, the presented algorithm resulted in a reduction of the overall cluster electricity costs, accompanied by a 23% decrease in peak energy demand and a 20% decrease in the Peak Average Ratio (PAR), without compromising indoor temperature control in comparison to a manually optimized RBC. The same authors also proposed [102] a method to increase the energy flexibility of a group of buildings by utilizing cooperation between them to achieve a coordinated approach to energy management. Deep RL was implemented to manage the thermal storage of four buildings equipped with distinct energy systems. The controller's objective was to optimize the energy usage of each building while flattening the cluster load profile. The coordinated energy-management controller was tested against a manually optimized RBC, resulting in an approximately 4% operational cost reduction and an up to 12% decrease in peak demand. Moreover, the control strategy results in a 10% reduction in the average daily peak and a 6% reduction in the average peak-to-average ratio, demonstrating the advantages of a coordinated approach.

Table 3. Summary of works of Optimal Control RL methods for Building Energy-Management Systems.

Reference Year	Field	Methodlogy	Type
[81] 2015	HVAC	ANNs & Q-Learning	Value-based
[85] 2017	HVAC	ANNs & Q-Learning	Value-based
[86] 2017	HVAC	Q-Learning	Value-based
[87] 2019	HVAC	REINFORCE	Policy-based
[98] 2019	HVAC	DDPG	Actor–Critic
[88] 2019	HVAC	Double Q-Learning	Value-based
[99] 2020	HVAC	PPO-Clip	Policy-based
[100] 2021	HVAC	DDPG	Actor–Critic
[94] 2021	HVAC/Multi	DQN	Value-based
[95] 2021	HVAC	DQN	Value-based
[82] 2016	Water Heating	Fitted Q-iteration	Value-based
[83] 2016	Water Heating	Q-Learning	Value-based
[96] 2017	Water Heating	Batch-RL & Fitted Q-iteration	Policy and Value-based

Table 3. *Cont.*

Reference Year	Field	Methodlogy	Type
[97] 2018	Water Heating	A3C	Policy-based
[89] 2019	Water Heating/Multi	Monte-Carlo with Exploring Starts (MCES)	Value-based
[92] 2020	Water Heating	Double Q-Learning with Memory Replay	Value-based
[93] 2021	Water Heating/Multi	ANNs & Q-Learning	Value-based
[101] 2021	Water Heating/Multi	SAC	Actor–Critic
[102] 2021	Water Heating/Multi	SAC	Actor–Critic
[84] 2016	Lighting	Q-Learning	Value-based
[90] 2019	Lighting	Value Iteration	Value-based
[91] 2019	Lighting	Branching Dueling Q-Network	Value-based

3.3. Reinforcement Learning for Optimal Control in Electric Vehicle Charging Stations (EVCSs)

Electric vehicles (EVs) are becoming increasingly popular due to their environmentally friendly nature. To this end, Reinforcement Learning (RL) was used for charging electric vehicles (EVs), as it enables the optimization of charging schedules in real time based on changing conditions, such as fluctuating electricity prices and uncertain arrival and departure times of EVs. RL can help determine the best charging strategy that minimizes the cost of electricity while ensuring that EVs are fully charged when needed. Additionally, RL can be used to optimize charging schedules for multiple EVs, taking into account their individual characteristics and constraints. RL algorithms can adapt and learn from data, allowing for efficient and effective charging of EVs in dynamic and complex environments. This section concludes with Table 4, which presents the works for Optimal RL Control Methods in EVCSs.

3.3.1. Value-Based RL in EVCSs

Beginning in 2015, Vandael et al. [103] addressed the challenge of creating a daily consumption schedule for electric vehicle (EV) fleet recharge. The challenge was grounded on the unpredictable charging tractability of the fleet, which is interrelated with various EV characteristics such as battery size, power curve, power limitations, etc. A heuristic control scheme was used to control the EV charging during operation and the charging operation of the vehicles was determined through Batch-RL methodology. Based on this acquired behavior, an expense-efficient day-ahead consumption scheme was developed. Simulation experiments took place in order to examine the concerned methodology approach for a multistage stochastic programming solution that utilized models from each EV's charging tractability. The experimental outcomes demonstrated that the proposed method was capable of producing efficient consumption schemes of comparable quality to the benchmark solution, without the demand of a thorough day-ahead model per EV's charging tractability.

Chics et al. [104] in 2016 suggested a novel demand response method in order to reduce the long-lasting expenses of charging a separate plug-in electric vehicle (PEV). The method was formulated as a day-by-day control problem for determining the amount of electricity to be charged in the PEV storage within 24 h, modeled as an MDP methodology with unpredictable variations. To address the variations in charging expenses, actual and simulated electricity prices were utilized. Additionally, a Bayesian ANN model was

utilized in order to determine the electricity charges and a Batch-RL approach was also deployed in order to determine the optimal charging scheme by utilizing transition samples and expense-eliminating charging control choices for unknown circumstances. Historical prices were used to construct the RL training information and a linear methodology was established so as to create a set of the best possible charging commands. The proposed method was evaluated through simulations by exploiting actual cost information and the results demonstrated an expense decrease of 10–50% for the PEV user, compared to other relative charging approaches.

In another important work of 2017, Mbuwir et al. [105] proposed batch Reinforcement Learning (RL) for microgrid energy management. The main objective was to develop a control scheme that efficiently schedules a battery device in order to increase the self-consumption of PV generation. To achieve this, a Q-iteration methodology, a common Batch-RL approach, was utilized by an RL agent so as to determine the appropriate control scheme. The approach exploited data and utilized a state–action value function to determine the optimal battery scheduling plan. The algorithm accounts for the storage charge and discharge performances as well as the nonlinearity in the grid for the inverter's performance. To examine the effectiveness of the suggested methodology approach, simulations took place exploiting the available information from Belgian users in a residential setting. By utilizing a model-based technique suitable for benchmarking, the evaluation of this work illustrated a performance difference of about 19%. Overall, this study offered a promising approach to optimizing energy management in microgrids using Batch-RL and its results may have important implications for the development of more efficient and sustainable energy systems.

The same year, Da Silva et al. [106] suggested a novel architecture, named MASCO, that concerned a Multiagent Multi-Objective Reinforcement Learning approach aiming to minimize energy costs and prevent transformer overloads while facilitating EV charging. The architecture considered insignificant presumptions as regards the shared grid and could operate for any potential charge rate while being configured to align with consumer preferences. Using real energy prices, experiments were conducted to evaluate MASCO's effectiveness in balancing energy costs and transformer loads. According to the simulation results, MASCO successfully balances energy costs and prevents transformer overloads, while taking into account consumer preferences. The experiments showed that among the baseline evaluated algorithms, MASCO achieved the best performance for both dynamic and time-of-use tariffs.

Moving forward to 2019, Qian et al. [107] illustrated a novel approach to electric vehicle (EV) charging schemes using Deep Reinforcement Learning (DRL). The proposed method extracts feature states from stochastic data such as road velocity, charging costs and holding period at the EVCS plants, using deep-state clustering and representation modeling (DSCRM). The EV charging navigation problem was formulated as an MDP problem with an unknown conversion prospect. To this end, the DRL methodology is adequate for generating the best possible control decisions for the EV charging plan independent of the requirement of previous data. The simulation outcomes showed that the technique is sufficient for determining the optimal decisions, by functioning in an online manner and compromising diverse EV user preferences.

Also in 2019, Sadeghianpourhamami et al. [108] considered a novel approach to regulating a framework of EVCSs by proposing a new MDP concept grounded in RL. The novelty of the paper in contrast with conventional algorithmic approaches was grounded in an RL methodology that manages the entire framework of the vehicles in a coordinated manner. The contribution included a scalable state representation that was detached from the EVCS amount by utilizing a Batch-RL approach—denoted fitted Q-iteration—to determine the optimal charging plan. Simulation experiments are performed considering real-life EV charging data so as to determine the efficiency of the approach. According to the extracted results, the current approach outperformed conventional algorithms in terms of coordination and scalability. Additionally, the considered methodology provided an effective

solution to the problem of jointly controlling EV charging stations, without requiring a priori knowledge of the optimal charging policy.

Finally, Wang et al. [109] presented a new approach using Reinforcement Learning (RL) in order to optimize the charging plan and establish cost-efficiency for a single domestic EVCS. The proposed algorithm relied only on past observations and not on established hypothetical models. To overcome the issue posed by time-dependent continuous state and action spaces in the RL problem, the authors optimized the overall charging rate in order to satisfy the charging requirements of former leaving times. Additionally, they put forward a method that utilizes a linear function approximator based on features for the state-value function, with the objective of improving the algorithm's effectiveness and capacity for generalization. The proposed RL algorithm was evaluated using numerical simulations with actual real-life data and the final outcome exposed that such methodology was adequate to establish 138.5% EVCS earnings improvement in comparison with the baseline methodology, on average.

In 2020, Chang et al. [110] proposed an RL storage charging management approach that aimed to reduce charging expenses for electric storage battery users in cases of markets with alterable energy values. The approach was grounded on the Q-Learning algorithm, the controller was trained for pre-treated energy cost information and the future prices were predicted so as to generate the optimal control framework in a finite Markov decision process (FMDP). One advantage of this method was that it did not require a high level of accuracy in the storage framework model—a phenomenon that was often difficult to be achieve in practice. Moreover, the study trains an ANN (LSTM type) in order to predict energy costs with significant precision, which enhances the control efficiency by extracting principal trends from the potential datasets and utilizing time-series information. The RL-based charging management approach delivered outstanding efficiency for the charging plan and expense minimization, in contrast to the baseline approaches.

Also in 2020, Lee et al. [111] introduced a novel algorithm grounded on the DRL framework, suitable for EV charging and discharging, taking into account energy charges per 60 min. The main goal of the algorithm was to characterize the usage behavior of a designated charging device in a particular location by establishing the probability function—and subsequently to reflect the local characteristics of the DRL agent. To achieve this, the study proposed a parametric free density approximation approach and derived probability function for variables associated with the charger device usage behavior using real-world datasets. Simulation results illustrated that the current approach effectively reduced the total charging expenses and increased the load factor of the target location. The algorithm decreases the charging expenses for the potential customers but also contributes to peak reduction and increases the grid reliability if deployed in multiple chargers. Overall, the proposed algorithm outperformed the conventional approaches as regards expense elimination and peak drop in cases of utilization in a particular charging device in a unique location.

An important work of 2021 was proposed by Tuchnitz et al. [112]. The study illustrated an RL charging planning multi-agent framework, able to generate separate charging plans for multiple electric vehicles using a special Markov Decision Process (MDP). The study was focused on a use-case scenario that included 250 dwellings and 50 passenger EVs, each with a battery capacity of 30 kWh. Simulated electric vehicle user data were applied in order to train and evaluate the RL methodology. Over a period of 300 consecutive days, the proposed system performed adequately by ensuring that each electric vehicle (EV) departed with an adequate charge, preventing the emergence of additional high-demand periods and substantially decreasing the overall load fluctuation by consistently charging EVs during periods of low electrical demand. The charging timetables, spanning a minimum duration of 1 day and formulated upon the arrival of an electric vehicle, enable optimal utilization of the remaining grid capacity. Additionally, the concerned RL methodology was evaluated against other charging strategies in terms of different performance metrics. The findings demonstrated that in contrast to the unmanaged standard charging approach,

the RL-based charging coordination system diminished the load fluctuation by 65% and eliminated the occurrences of surpassing power boundaries from 896 to 0.

3.3.2. Policy-Based, Actor–Critic and Hybrid RL in EVCSs

In an interesting study of 2019, Li et al. [113] described the charging planning framework as a restricted MDP (CMDP) challenge, accounting for the uncertainty of the entry and exit period of the vehicles and the real-time energy costs. In this context, the scientists proposed a model-independent framework grounded on SDRL, which is relieved from the requirement of previous information as regards the randomness or constraints of the process—and thus, it was adequate with respect to achieving optimal control by avoiding the need for manual design or tuning of penalty terms. Instead, the proposed solution exploited a DNN to deliver the optimal scheme of energizing and de-energizing EVs directly throughout the entire process. The simulation outcomes demonstrated the effectiveness of the suggested methodology in reducing charging costs and satisfying charging constraints when compared to baseline solutions.

Also in 2020, Zhang et al. [114] suggested a new technique for the charging control of EVs employing the DLR methodology. The charging procedure was formulated as an MDP and the suggested technique—abbreviated as CDDPG—aimed to decrease the charging procedure expenses while meeting the vehicle owners' needs for storage energy. The approach used an ANN that included an LSTM layer to obtain data from previous electricity costs in order to control the current charging/discharging procedure. To overcome the issue of sparse rewards, two distinct replay buffers were used to store the alterations due to and after the charging period. According to the simulation results, CDDPG outperformed both baseline approaches—DQN and DDPG—in terms of fulfilling the vehicle owner's objectives as concerns storage energy and charging expense elimination. The study presented a promising approach to EV charging control, with the potential for further exploration and refinement using other Reinforcement Learning techniques.

Dorokhova et al. [115] suggested a novel DRL control scheme in 2021 considering EV charging as concerns the integration of RES and more specifically the presence of PVs in the optimization mix. Researchers deployed the RL methodology in order to satisfy the charging control requirement, considering a framework that integrates a utility grid, building load, PV energy production and a particular vehicle. The current group formulated the optimization objective three additional times in the form of Markov Decision Processes (MDPs) with separate action spaces. To solve the proposed MDP formulations, three RL algorithms were deployed, including Double Deep Q-Network learning, Deep Deterministic Policy Gradient (DDPG) and Parametrized Deep Q-Network learning for discrete, continuous and parametrized action spaces, respectively. By utilizing a thorough evaluation procedure on an outstretched test dataset, the Reinforcement Learning Control methodology outperformed both RBC optimization and MPC deterministic and stochastic methodologies. The results also demonstrated that while RLC algorithms show slightly lesser performance on the chosen objectives, they exhibit robustness for generating practical charging-optimization schemes.

In a recent work of 2022, Park et al. [116] suggested a novel framework that utilized a model-independent DRL methodology for satisfying a storage quick-charge challenge while considering safety constraints in the presence of an electrochemical model as the battery simulator. The actor–critic approach—and more specifically the DDPG algorithm—was elected for its capacity to handle continuous state and action spaces. To tackle the issue of state restrictions, a reward function was developed so as to facilitate constraint violation learning. This work was the first of its kind to combine AI and storage management frameworks. The open-source operation of the framework allowed an improved learning performance using advanced RL algorithms, such as Soft Actor–Critic (SAC) and enabled the exploration of other RL perspectives, including inverse RL and hierarchical RL.

Table 4. Summary of works of Optimal Control RL methods of EV Charging Stations.

Reference Year	Field	Methodlogy	Type
[103] 2015	EVCS/Multi	Batch-RL Fitted Q-iteration	Value-based
[104] 2016	EVCS/Multi	Batch-RL Fitted Q-iteration	Value-based
[105] 2017	EVCS/Multi	Batch-RL Fitted Q-iteration	Value-based
[113] 2019	EVCS	Constrained Policy Optimization	Policy-based
[106] 2019	EVCS	MASCO	Value-based
[107] 2019	EVCS	Deep RL & DQN	Value-based
[108] 2019	EVCS	Batch-RL Fitted Q-iteration	Value-based
[109] 2019	EVCS	Q-Learning & DQN	Value-based
[110] 2020	EVCS/Multi	Q-Learning	Value-based
[111] 2020	EVCS	DQN	Value-based
[114] 2020	EVCS/Multi	CDDPG	Actor–Critic
[115] 2021	EVCS	Double DQN, DDPG & Parameterized DQN	Actor–Critic, Policy and Value-based
[112] 2021	EVCS/Multi	Q-Learning & DQN	Value-based
[116] 2022	EVCS/Multi	DDPG	Actor–Critic

4. Evaluation

In this section, valuable insights are drawn regarding the aforementioned literature work. This section justifies the research trends according to different fields of RL application and classifies the integrated literature work. The primary aim of the current section is to illustrate the challenges that each field—or subfield—integrates and the identification of research tendencies regarding each application.

4.1. Evaluation per Type

Application of RL optimal control in such energy systems may vary for the type of RL. To this end, value-based Reinforcement Learning (RL) approaches, such as Q-Learning and Deep Q-Networks (DQNs), are frequently utilized for RES, BEMs and EVCS control due to their integrated simplicity and interoperability. These methods estimate the value function, which represents the expected cumulative rewards for different state–action pairs. They strike a balance between exploration and exploitation, enabling adaptive control policies in complex building environments affected by factors such as occupancy patterns and weather conditions. Additionally, the value-based RL techniques do not rely on explicit system models, allowing direct learning from interactions with the building environment [117]. On the other hand, policy-based methods and actor–critic approaches are less commonly seen in energy systems such as RES, buildings and EVCS control. These approaches face challenges when dealing with high-dimensional action spaces prevalent in building control. Policy-based methods directly learn the policy function, while actor–critic approaches combine policy-based and value-based methods. Moreover, these techniques require more samples and interactions with the environment, which can be difficult and time-consuming to collect in real-world building scenarios [117–119]. Furthermore, interoperability and

safety concerns are key considerations in building control, where stakeholders require transparent and safe control policies. However, even though value-based RL approaches are the prevailing choice due to their simplicity and adaptability to uncertain dynamics, there is certainly a potential for further exploration of policy-based and actor–critic methods in specific building control contexts with tailored modifications and advancements.

Hybrid approaches between value-based, policy-based and actor–critic algorithms are also present in the literature, such as Advantage Actor–Critic (A2C) and Proximal Policy Optimization (PPO), which combine the advantages of the aforementioned methods [118]. These approaches offer improved stability, faster learning and an efficient exploration–exploitation trade-off. By leveraging value estimation for reliable learning signals and policy optimization for effective decision-making, hybrid approaches provide more stable learning dynamics, better sample efficiency and flexibility in RL control tasks. They strike a balance between exploration and exploitation, leading to enhanced performance, convergence and efficiency in RL-based applications [117,118]. The interested reader may find the percentage or share of the examined value-based, policy-based, actor–critic and hybrid Reinforcement Learning approaches in the energy systems RES, BEMS and EVCSs, respectively, in Figure 3.

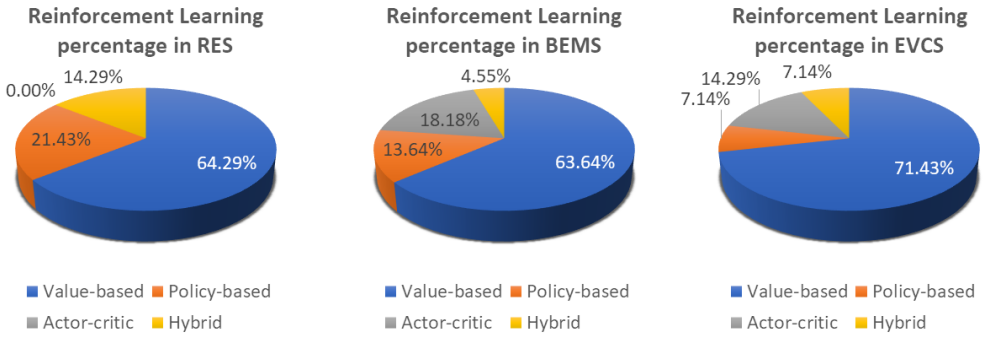

Figure 3. Share of value-based, policy-based, actor–critic and hybrid Reinforcement Learning approaches in the examined energy systems: RES, BEMS and EVCSs, respectively.

4.2. Evaluation per Algorithmic Methodology

Another useful conclusion for the examination of the presented highly cited papers is that, among others, the most common algorithmic approach to enable RL in energy systems—such as RES, BEMS and EVCS—is the value-based Q-Learning algorithmic approach. Q-Learning, compared to other Reinforcement Learning (RL) methods, has distinct advantages that make it well-suited for optimizing control in energy systems [120]. Q-Learning focuses on estimating the action-value function (*Q*-function), enabling it to efficiently learn optimal control policies by iteratively updating *Q*-values based on observed rewards. It scales well to handle large state and action spaces, making it suitable for energy systems with numerous control variables. Another benefit in comparison to other RL algorithmic methodologies is its sample efficiency, as Q-Learning requires fewer interactions with the environment to converge to an optimal policy, which is advantageous in scenarios where real-world data collection may be limited or expensive. Additionally, Q-Learning incorporates an exploration–exploitation trade-off, allowing agents to explore new actions while exploiting the best-known ones, a valuable characteristic for energy systems where finding optimal control policies in uncertain environments is critical [120].

Although Q-Learning has significant strengths, it is essential to consider other RL approaches that concern the aforementioned policy-based or actor–critic methodologies, when specific requirements or constraints of an energy system necessitate their advantages, such as handling high-dimensional action spaces or offering improved convergence properties. According to the evaluated highly cited research works, we mention the following: (1) Proximal Policy Optimization (PPO), which is adequate for ensuring stable learning dynamics, efficient sample utilization and reliable policy updates; (2) Deep De-

terministic Policy Gradient (DDPG), which is able to handle continuous action spaces and is stable for complex policies; (3) Soft Actor–Critic (SAC) approaches, which present efficient exploration–exploitation trade-off through entropy regularization and sample efficiency through off-policy learning [121]. The interested reader may find the share or the percentage of the examined RL algorithmic methodologies in the examined energy systems RES, BEMS and EVCSs, respectively, in Figure 4.

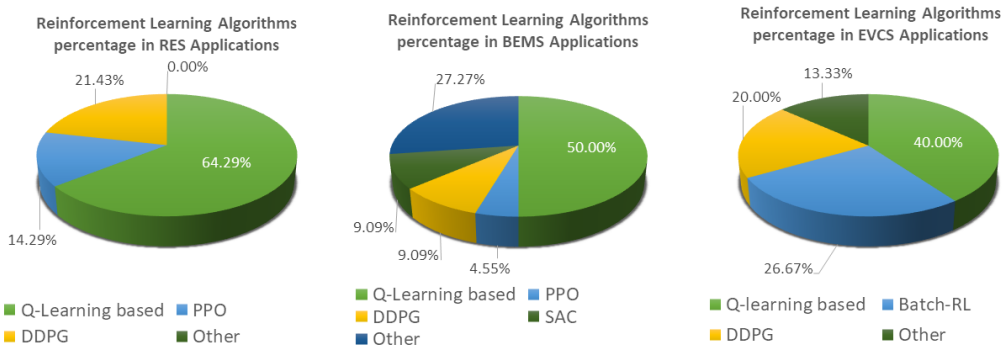

Figure 4. Share of examined RL algorithmic methodologies in the examined Energy Systems: RES, BEMS and EVCSs, respectively.

4.3. Evaluation per Application Field

4.3.1. RL Optimal Control in RES

The goal of RL optimal control in RES is to maximize the efficiency and profitability of renewable energy systems by finding the optimal control strategies for the operation of energy-generation, storage and consumption devices. As the relative literature illustrates:

- RL optimal control may target reducing operational costs, increasing energy production and improving system reliability just as in [68,71–73,122]. In such cases, optimal control in RL works by using a reward-based approach to learn the optimal control policies that can maximize a particular objective function. This objective function can be defined in different ways depending on the specific application of the RES. For example, in a wind or solar farm, the objective may be to maximize the energy output while minimizing the operational costs.
- Additionally, optimal control empowered by RL methodologies can help to overcome the inherent uncertainty and variability of renewable energy sources, energy prices or load demand just as in [70–72,75,79]. RL algorithms are capable of adapting to changing weather conditions, energy demand and other variables in real time, ensuring that the RES operates optimally under different scenarios.
- Last but not least, RL may help to reduce the environmental impact of RES by optimizing energy generation to meet demand while minimizing the use of non-renewable sources of energy just as in [68,71–73]. This tendency may lead to greenhouse gas emission reductions and deliver sustainability in the potential energy system. Figure 5 illustrates the Reinforcement Learning Optimal Control schemes in RES indicating the control targets as denoted in the literature.

RL optimal control frameworks for Renewable Energy Systems (RESs) may also vary based on the type of RES. Different types of renewable energy sources have distinct characteristics, operational requirements, control variables and system dynamics which may influence the design of RL frameworks. This work is primarily focused on solar power systems and wind energy systems as well as hybrid approaches that integrate distributed optimization schemes between them or even the integration of storage, to enable a cooperative control of different frameworks.

Solar Power System Control: The overall objective of RL optimization in solar power systems is to maximize energy generation from solar panels, improve system efficiency and minimize reliance on non-renewable energy sources. To this end, the potential ele-

ments for efficient optimal control via RL methodologies may usually concern the following: load management, where RL algorithms are utilized to control energy-consumption devices within the system, optimizing their operation based on the solar energy availability [68–70,72,79]; battery energy storage, where RL algorithms can control the charging and discharging of energy-storage systems, such as batteries, in solar power systems [73,122]; grid interaction, where RL optimization concerns the interaction between the solar power system and the electrical grid to minimize reliance on the grid and reduce costs [70]; and last but not least, it should be noted that multi-agent RL approaches have also been examined [77] targeting the integration of wide-scale PV implementations or the potential of diverse RES technologies.

Wind Power System Control: The overall objective of RL optimization in wind power systems is to maximize energy generation, increase system efficiency and improve the economic viability of wind farms. According to the reviewed literature, the potential control elements may concern the following: forecasting and uncertainty management, where wind power systems are subject to inherent variability and uncertainty due to fluctuating wind conditions [75,76,78]; turbine or yaw control, where RL algorithms optimize the control parameters of wind turbines, including blade pitch angle, rotor speed and yaw angle in order to maximize energy capture [65,66,74]; and lastly, grid integration—where RL optimization focuses on managing the interaction between the wind power system and the electrical grid—is examined in [71].

Multi-Device Control in RES: A lot of viable research concerns integrating distributed renewable energy along with storage technologies and controllable generators [64,73]. In such cases, the optimization takes place in a coordinated way, demanding an even more sophisticated mathematical algorithmic solution such as multi-agent RL control approaches. In the forthcoming years, the widespread adoption of microgrids is expected due to their crucial role in the integration of distributed renewable energy resources into the primary grid. However, the inherent volatility of renewable energy sources, such as solar and wind energy, poses challenges as they depend highly on weather conditions. This variability, combined with fluctuating demand, can result in unpredictable fluctuations in both power generation and load patterns, thereby complicating the task of optimizing energy management.

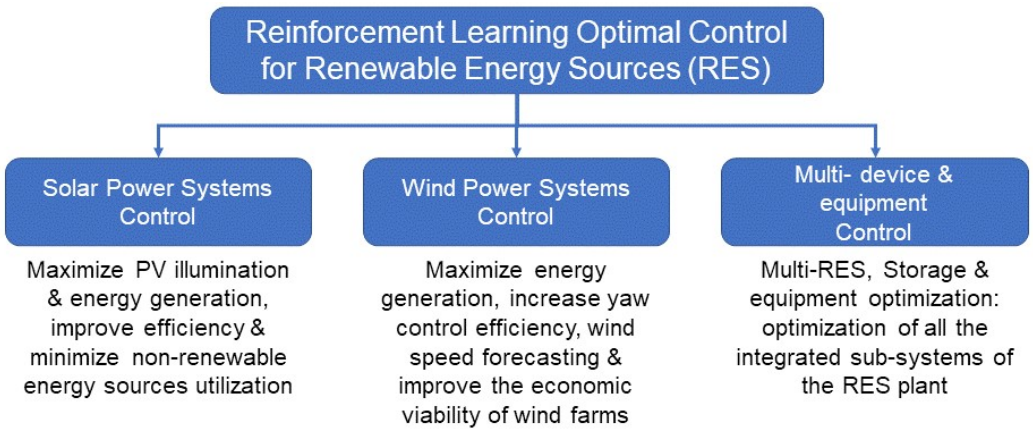

Figure 5. Reinforcement Learning Optimal Control in RES.

4.3.2. RL Optimal Control in BEMS

RL enables the optimization of various performance objectives, including energy consumption, thermal comfort, indoor air quality and peak load reduction, while accounting for uncertainties and varying occupancy patterns. This paper summarizes highly cited papers concerning HVAC, water heating and lighting control, as well as research efforts that concern multi-device optimal control that usually integrates storage or RES equipment in the optimization mix. Figure 6 illustrates the Reinforcement Learning Optimal Control schemes in BEMS, indicating the control targets as denoted in the literature.

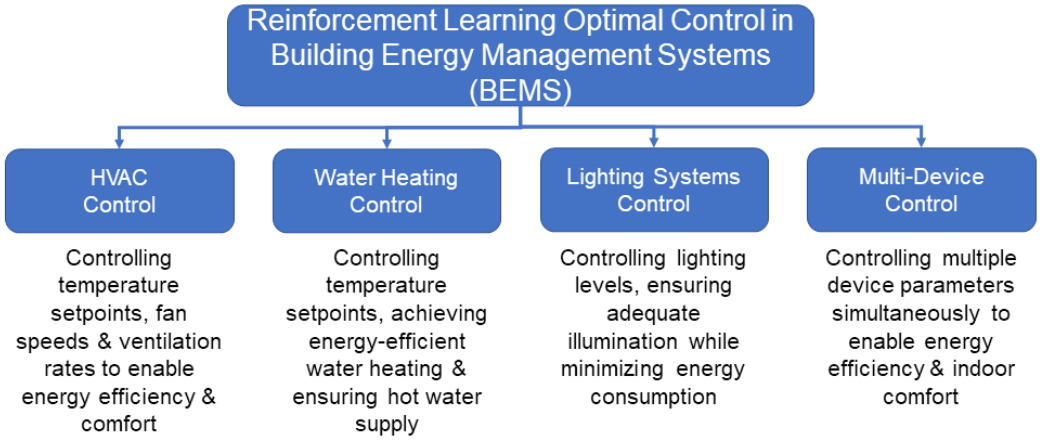

Figure 6. Reinforcement Learning Optimal Control in Buildings.

HVAC Control: By using RL algorithms, HVAC operation can be dynamically adjusted to achieve energy efficiency and occupant comfort. RL can learn the optimal control policies by considering various factors such as outdoor weather conditions, building occupancy, thermal comfort requirements and energy-consumption goals. To this end, the vast majority of research efforts illustrate energy savings and comfort [81,86,88,98–100] as the primary parameters for RL control optimization, while some others integrate cost reduction as well [94,95]. Multi-zone control illustrates a more sophisticated approach to large-scale HVAC control implementation. The approach involves dividing a building into different zones or areas and independently controlling the temperature and airflow in each zone as outlined in [85,100].

Water Heating Control: The capability of electric water heaters to store energy in their water buffer without affecting user comfort makes them an excellent candidate for residential demand response. Nonetheless, due to the stochastic and nonlinear nature of their dynamics, utilizing their flexibility presents a challenge. To this end, RL approaches offer a fruitful control optimization potential with the aim of reducing energy costs without degrading comfort [82,83,89,92,93,96,97,101,102]. Commonly, the RL agent considers various state variables such as the current time of day, water temperature and forecasted usage while deciding whether to turn the heater on or off. It is noticeable that a significant amount of research has focused on demand-side management of domestic electric water heaters (DEWHs) [83,101,102], while others integrate the utilization of RES and especially PVs [93,96], and multi-agent RL control approaches commonly consider water heating in large-scale implementations, such as [89,102].

Lighting System Control: Lighting systems play a crucial role in indoor comfort as they provide the essential illumination required for occupants to carry out tasks, improve visual aesthetics and influence mood and behavior. While lighting systems offer numerous advantages, such as enhancing the overall environment, they also have an impact on energy usage and expenses within buildings. The majority of research applications concern occupant-centered lighting control schemes where the RL approach is combined with a human feedback mechanism for regulating lights [84] along with blinds [90] and windows [91].

Mutli-Device Control in BEMS: Numerous building-related research efforts are focusing on multi-device control, while also integrating Renewable Energy Sources (RESs) and energy storage into comprehensive control systems. RL can control the charging and discharging of energy-storage systems in buildings, such as batteries or thermal storage systems. By learning the optimal scheduling and utilization of stored energy, RL algorithms are adequate with respect to minimizing peak demand, optimizing load balancing and reducing electricity costs. Current work illustrates three storage-related works [89,101,102]—each one dedicated to RL multi-agent algorithmic approaches. Similarly, RL control approaches are adequate with respect to optimizing the integration of RES,

such as solar panels, into building energy systems. This involves learning to maximize the use of locally generated renewable energy while considering energy demand and storage capacity. The present study exemplifies two highly cited research efforts considering the optimal utilization of photovoltaic (PV) energy production and energy-consumption reduction [93,96] in buildings.

4.3.3. RL Optimal Control in EVCS

RL algorithms are also capable of learning from interactions with the charging environment, allowing the generation of intelligent decisions in terms of when and how to charge Electrical Vehicles (EVs). To this end, RL control may consider electricity prices, charging station load and grid constraints, as well as user preferences. As the literature indicates, additional equipment may be integrated into the optimization mix, such as battery usage optimization or RES efficient utilization. By incorporating these variables into the RL algorithm, the charging system can strike a balance between minimizing charging costs and ensuring efficient and reliable charging operations. Figure 7 illustrates the Reinforcement Learning Optimal Control schemes in EVCS, indicating the control targets as outlined in the literature.

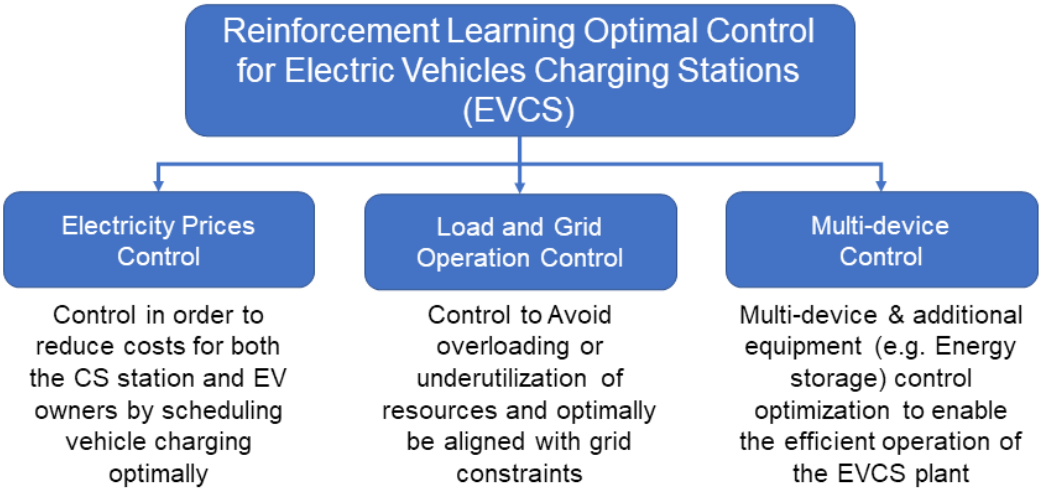

Figure 7. Reinforcement Learning Optimal Control in EVCS.

Electricity Price Control: Overall, RL control enables the charging system to adaptively respond to time-varying electricity prices, optimizing charging schedules to minimize costs while still meeting the charging needs of electric vehicles. By leveraging RL algorithms, charging stations may schedule EV charging optimally, during off-peak hours when electricity prices are lower, reducing costs for both the station operator and EV owners. As outlined in [104,106,110,111,113,114], this approach maximizes the cost-effectiveness of EV charging and encourages efficient utilization of energy resources.

Load and Grid Operation Control: RL algorithms are sufficient with respect to monitoring the charging station load and distributing the available power among connected EVs to maintain a balanced load. By intelligently allocating power resources, RL control ensures that the charging demand is evenly distributed across the station's capacity, avoiding overloading or underutilization of resources, as outlined in [106,111,112,115]. Moreover, RL optimization schemes take into account grid constraints such as maximum power capacity, voltage stability and network congestion. The RL agent optimizes charging schedules to prevent exceeding the capacity limits of the grid infrastructure and mitigate potential issues that could affect the stability and reliability of the electricity grid, as outlined in [105,111,112]. Moreover, RL control considers user preferences, such as desired departure times or minimum charge levels [114]. By incorporating these preferences into the RL optimization scheme, the charging station can prioritize charging based on user requirements while still optimizing load and grid constraints.

Mutli-Device Control in EVCS: By considering battery-specific factors and employing adaptive control strategies, RL control in EV charging stations aims to maximize battery performance, extend battery lifespan and ensure efficient and reliable charging operations. This helps enhance the overall sustainability and user experience of electric vehicles. Taking into account the charging rate optimization, State of Charge management (SOC), battery temperature, battery overall behavior, as well as the smart grid Integration, RL control is adequate with respect to optimizing the charging process and enhancing battery performance and longevity. EVCS and battery-related optimization control can be found in [103–105,110,112,114,116].

5. Research Tendencies and Future Directions

In this section, we delve into the future possibilities and potential avenues for EV energy management grounded in Reinforcement Learning (RL) principles. According to the evaluation, the most commonly utilized RL technique in every energy system is the value-based Q-Learning methodology. RL techniques such as PPO, DDPG and SAC are also utilized, however in a significantly lower percentage. However, their implementation holds significant potential due to their unique attributes, such as adaptability, performance levels, convergence speed, utilization of continuous state and action spaces, the robustness of the proposed solutions and data efficiency, as well as further advancements in exploration strategies and integration with real-time data that unlock even greater potential for optimizing energy systems. Another common practice in all the aforementioned energy system applications is the utilization of hybrid approaches in order to adopt the algorithmic advantages from every perspective. Overall, hybrid algorithmic approaches in RL for energy systems offer the potential to overcome challenges, improve performance and optimize control strategies by combining different algorithms and leveraging their respective strengths. However, addressing these challenges requires careful consideration, domain expertise, thorough evaluation and systematic validation. It is essential to conduct rigorous testing, sensitivity analyses and comparative studies to ensure effectiveness and reliability. Future research will probably leverage hybrid RL on a wider scale, exploiting further the respective algorithmic strengths in handling complexity, uncertainty and specific domain considerations.

A common direction that arises from the evaluation is the adoption of DRL algorithms in the majority of the examined applications in all of the three energy systems. The reason for such a tendency is grounded in the capability of DRL methodologies to learn intricate policies adequately, surpassing those represented by simple neural networks or lookup tables. In the forthcoming years, as the quantity and quality of gathered information increases—due to sensorial equipment expansion—the use of DRL methodologies will become indispensable for developing effective policies in environments with vast state–action spaces. Moreover, advancements in computing power will also enable researchers to train RL agents with increasingly intricate policies. Another potential tendency for future research in energy system RL control involves exploring the potential variants of traditional RL. The literature introduces important research works related to multi-task control schemes. For instance, in RES energy systems, efficiency, costs and load/grid balance make up the primary integrated frameworks in the optimization mix. In BEMS, the usual trade-off considers energy efficiency, comfort and potential user requirements for the multi-task control scheme. Last but not least, the EVCS multi-task control scheme usually integrates costs, time and user requirements as well as harmonization for load and grid balance. It needs to be mentioned that future experimental research works need to investigate further the multi-task optimization control scheme since there is a lack of suitable algorithms as concerns state-of-the-art approaches.

Regarding the application of RL control in large-scale energy systems, it is observed that an increasing number of literary studies implement distributed (multi-agent) RL approaches for the cooperative control of numerous devices, equipment items and assets. Developing control policies for multiple agents offer the advantage of resource pooling,

resulting in greater overall performance—e.g., cost savings overall. Future research efforts may additionally aim to further extend the application of multi-agent RL to even large ecosystems of devices, equipment or even assets.

After illustrating a comprehensive review of the available Reinforcement Learning literature on RES, BEMS and EVCSs, it becomes apparent that existing research on applications is primarily focused on frameworks operating within conventional environmental conditions, while significant alterations to the operational environment are limited. This observation raises an important question regarding the performance of such learned control policies when confronted with significant environmental variations, such as extreme weather conditions, storage exhaustion, PV panel or wind turbine malfunctions, alterations in the number of available equipment items (add/remove equipment), alterations in the number of occupants, etc. It seems the literature is limited in evaluating the robustness of RL methodologies, a fact which may lead to disastrous failure and suboptimal performance. Additionally, in cases in which the RL agent proves robust and capable of adapting to such devastating changes, the expected time frame for proper re-adaption to the new status quo needs to be evaluated as well. To this end, future research efforts need to also provide in-depth investigation, so as to leverage our understanding of the adaptability and robustness of RL agents in the presence of significant status alterations in energy systems.

6. Conclusions and Future Work

This paper provides valuable insights into the application of Reinforcement Learning (RL) in energy systems, specifically renewable energy source plants, building management systems and electric vehicle charging stations. By analyzing 80+ highly cited papers, the current work presents various RL approaches for the control and management of the aforementioned energy systems.

Future work may focus on exploring novel RL approaches that incorporate advanced deep learning techniques, such as Deep Q-Networks or policy gradient methods, to tackle complex decision-making problems in energy systems. Exploring the integration of RL with advanced optimization algorithms, predictive modeling and data analytics may enhance decision-making capabilities and enable more proactive energy-management strategies. By focusing on these areas, future research work needs to contribute to the advancement of energy systems, making them more efficient, sustainable and resilient in meeting the growing demands of our energy future. Additionally, While most of the reviewed works offer solutions related to single-agent RL frameworks, unlocking the potential of multi-agent and hierarchical RL may provide improved data efficiency and accelerated simulation times and also facilitate the use of RL in non-stationary environments, further enabling a more comprehensive understanding of state and action dynamics. Future work is required to additionally deploy distributed control schemes further in order to maximize the performance of energy systems on a wider scale

According to the final evaluation, one may safely assume that RL holds immense potential in energy systems for the future. With its ability to optimize complex environments, RL can adaptively manage numerous energy system operations. RL flexibility enables control strategies for diverse components, considering energy availability, demand patterns, pricing signals and environmental factors. Moreover, RL's adaptability supports the integration of emerging technologies and facilitates the transition to a sustainable energy future. By leveraging RL's optimization capabilities, energy systems can be optimized to improve efficiency, reliability and environmental impact, contributing to a cleaner and more resilient energy ecosystem.

Author Contributions: Conceptualization, D.V. and C.K., methodology, D.V., P.M., C.K. and E.K., software, D.V. and P.M., validation, D.V., P.M. and C.K., formal analysis, D.V. and P.M., investigation, D.V., P.M., C.K. and E.K., resources, D.V., P.M., C.K. and E.K., data curation, D.V., P.M. and C.K., writing—original draft preparation, D.V. and P.M., writing—review and editing, D.V., P.M. and C.K., visualization, D.V. and P.M., supervision, C.K. and E.K., project administration, E.K., funding acquisition, C.K. and E.K. All authors have read and agreed to the published version of the manuscript.

Funding: The research leading to these results was partially funded by the European Commission H2020-EU.2.1.5.2.—LC-EEB-07-2020—Smart Operation of Proactive Residential Buildings (IA) (Grant agreement ID: 958284) PRECEPT https://www.precept-project.eu/ (accessed on 1 June 2023) and CL5-2021-D4-02-02—Efficient, sustainable and inclusive energy use (Grant agreement ID: 101079951) REHOUSE https://rehouse-project.eu// (accessed on 1 June 2023).

Data Availability Statement: No new data have been presented in this paper.

Conflicts of Interest: The authors declare no conflict of interest.

Abbreviations

The following abbreviations are used in this manuscript:

A2C	Advantage Actor–Critic
A3C	Asynchronous Advantage Actor–Critic
ADHDP	Action-Dependent Heuristic Dynamic Programming
ANN	Artificial Neural Network
BEM	Building Energy Model
BEMS	Building Energy-Management Systems
DBN	Deep Belief Network
DC	Direct Current
DDPG	Deep Deterministic Policy Gradient
DEWH	Domestic Electric Water Heating Systems
DG	Distributed Generation
DHW	Domestic Hot Water
DNN	Deep Neural Network
DQN	Deep Q-Network
DRL	Deep Reinforcement Learning
DSCRM	Deep-State Clustering and Representation Modeling
EG	Energy Grid
ESN	Echo State Network
EVCS	Electric Vehicle Charging Stations
FC	Fuel Cell
FMDP	Finite Markov Decision Process
FNN	Feedforward Neural Network
HRES	Hybrid Renewable Energy Systems
HRL	Hierarchical Reinforcement Learning
HVAC	Heating Ventilation Air Conditioning
IES	Integrated Energy System
KA	Knowledge-Assisted
LCP	Load Commitment Problem
LSTM	Long Short-Term Memory
MDP	Markov Decision Process
MLP-BP	Multilayer Perceptron and Backpropagation
MOC	Model-based Optimal Control
MPC	Model Predictive Control
MPPT	Maximum Power Point Tracking
PEV	Plug-in Electric Vehicle
PHM	Prognosis and Health Management
PMSG	Permanent Magnet Synchronous Generators
PMV	Predicted Mean Vote Index
PPO	Proximal Policy Optimization
PV	Photovoltaic
RBC	Rule-based Control
RES	Renewable Energy Sources
RL	Reinforcement Learning
RLMPPT	Reinforcement Learning Maximum Power Point Tracking
SAC	Soft Actor–Critic
SDRL	Symbolic Deep Reinforcement Learning

TD3	Twin-Delayed Deep Deterministic Policy Gradient
TRPO	Trust Region Policy Optimization
VPG	Vanilla Policy Gradient
WECS	Wind Energy-Conversion System

References

1. Mikayilov, J.I.; Mukhtarov, S.; Dinçer, H.; Yüksel, S.; Aydın, R. Elasticity analysis of fossil energy sources for sustainable economies: A case of gasoline consumption in Turkey. *Energies* **2020**, *13*, 731. [CrossRef]
2. Martins, F.; Felgueiras, C.; Smitkova, M.; Caetano, N. Analysis of fossil fuel energy consumption and environmental impacts in European countries. *Energies* **2019**, *12*, 964. [CrossRef]
3. Zahraoui, Y.; Basir Khan, M.R.; AlHamrouni, I.; Mekhilef, S.; Ahmed, M. Current status, scenario and prospective of renewable energy in Algeria: A review. *Energies* **2021**, *14*, 2354. [CrossRef]
4. Abas, N.; Kalair, A.; Khan, N. Review of fossil fuels and future energy technologies. *Futures* **2015**, *69*, 31–49. [CrossRef]
5. Holechek, J.L.; Geli, H.M.; Sawalhah, M.N.; Valdez, R. A global assessment: Can renewable energy replace fossil fuels by 2050? *Sustainability* **2022**, *14*, 4792. [CrossRef]
6. Shafiee, S.; Topal, E. When will fossil fuel reserves be diminished? *Energy Policy* **2009**, *37*, 181–189. [CrossRef]
7. Halkos, G.E.; Gkampoura, E.C. Reviewing usage, potentials and limitations of renewable energy sources. *Energies* **2020**, *13*, 2906. [CrossRef]
8. Yan, J.; Chou, S.K.; Desideri, U.; Lee, D.J. Transition of clean energy systems and technologies towards a sustainable future. Fifteenth International Conference on Atmospheric Electricity (ICAE 2014), Norman, Oklahoma, USA, 15–20 June 2014. *Appl. Energy* **2015**, *160*, 619–1006. [CrossRef]
9. Dominković, D.F.; Bačeković, I.; Pedersen, A.S.; Krajačić, G. The future of transportation in sustainable energy systems: Opportunities and barriers in a clean energy transition. *Renew. Sustain. Energy Rev.* **2018**, *82*, 1823–1838. [CrossRef]
10. Michailidis, P.; Pelitaris, P.; Korkas, C.; Michailidis, I.; Baldi, S.; Kosmatopoulos, E. Enabling optimal energy management with minimal IoT requirements: A legacy A/C case study. *Energies* **2021**, *14*, 7910. [CrossRef]
11. Michailidis, I.T.; Sangi, R.; Michailidis, P.; Schild, T.; Fuetterer, J.; Mueller, D.; Kosmatopoulos, E.B. Balancing energy efficiency with indoor comfort using smart control agents: A simulative case study. *Energies* **2020**, *13*, 6228. [CrossRef]
12. Michailidis, I.T.; Schild, T.; Sangi, R.; Michailidis, P.; Korkas, C.; Fütterer, J.; Müller, D.; Kosmatopoulos, E.B. Energy-efficient HVAC management using cooperative, self-trained, control agents: A real-life German building case study. *Appl. Energy* **2018**, *211*, 113–125. [CrossRef]
13. Tamani, N.; Ahvar, S.; Santos, G.; Istasse, B.; Praca, I.; Brun, P.E.; Ghamri, Y.; Crespi, N.; Becue, A. Rule-based model for smart building supervision and management. In Proceedings of the 2018 IEEE International Conference on Services Computing, San Francisco, CA, USA, 2–7 July 2018; pp. 9–16.
14. De Hoog, J.; Abdulla, K.; Kolluri, R.R.; Karki, P. Scheduling fast local rule-based controllers for optimal operation of energy storage. In Proceedings of the Ninth International Conference on Future Energy Systems, Karlsruhe, Germany, 12–15 June 2018; pp. 168–172.
15. Kermadi, M.; Salam, Z.; Berkouk, E.M. A rule-based power management controller using stateflow for grid-connected PV-battery energy system supplying household load. In Proceedings of the 2018 9th IEEE International Symposium on Power Electronics for Distributed Generation Systems (PEDG), Charlotte, NC, USA, 25–28 June 2018; pp. 1–6.
16. Schreiber, T.; Netsch, C.; Baranski, M.; Mueller, D. Monitoring data-driven Reinforcement Learning Controller training: A comparative study of different training strategies for a real-world energy system. *Energy Build.* **2021**, *239*, 110856. [CrossRef]
17. Fu, Y.; Xu, S.; Zhu, Q.; O'Neill, Z.; Adetola, V. How good are learning-based control vs model-based control for load shifting? Investigations on a single zone building energy system. *Energy* **2023**, *273*, 127073. [CrossRef]
18. Jahedi, G.; Ardehali, M. Genetic algorithm-based fuzzy-PID control methodologies for enhancement of energy efficiency of a dynamic energy system. *Energy Convers. Manag.* **2011**, *52*, 725–732. [CrossRef]
19. Ooka, R.; Komamura, K. Optimal design method for building energy systems using genetic algorithms. *Build. Environ.* **2009**, *44*, 1538–1544. [CrossRef]
20. Parisio, A.; Wiezorek, C.; Kyntäjä, T.; Elo, J.; Strunz, K.; Johansson, K.H. Cooperative MPC-based energy management for networked microgrids. *IEEE Trans. Smart Grid* **2017**, *8*, 3066–3074. [CrossRef]
21. Mariano-Hernández, D.; Hernández-Callejo, L.; Zorita-Lamadrid, A.; Duque-Pérez, O.; García, F.S. A review of strategies for building energy management system: Model predictive control, demand side management, optimization and fault detect & diagnosis. *J. Build. Eng.* **2021**, *33*, 101692.
22. Michailidis, I.T.; Kapoutsis, A.C.; Korkas, C.D.; Michailidis, P.T.; Alexandridou, K.A.; Ravanis, C.; Kosmatopoulos, E.B. Embedding autonomy in large-scale IoT ecosystems using CAO and L4G-CAO. *Discov. Internet Things* **2021**, *1*, 1–22. [CrossRef]
23. Jin, X.; Wu, Q.; Jia, H.; Hatziargyriou, N.D. Optimal integration of building heating loads in integrated heating/electricity community energy systems: A bi-level MPC approach. *IEEE Trans. Sustain. Energy* **2021**, *12*, 1741–1754. [CrossRef]
24. Artiges, N.; Nassiopoulos, A.; Vial, F.; Delinchant, B. Calibrating models for MPC of energy systems in buildings using an adjoint-based sensitivity method. *Energy Build.* **2020**, *208*, 109647. [CrossRef]
25. Forgione, M.; Piga, D.; Bemporad, A. Efficient calibration of embedded MPC. *IFAC-PapersOnLine* **2020**, *53*, 5189–5194. [CrossRef]

26. Storek, T.; Esmailzadeh, A.; Mehrfeld, P.; Schumacher, M.; Baranski, M.; Müller, D. Applying Machine Learning to Automate Calibration for Model Predictive Control of Building Energy Systems. In Proceedings of the Building Simulation 2019, Rome, Italy, 2–4 September 2019; Volume 16, pp. 900–907.
27. Saad, A.A.; Youssef, T.A.; Elsayed, A.T.; Amin, A.M.A.; Abdalla, O.H.; Mohammed, O.A. Data-Centric Hierarchical Distributed Model Predictive Control for Smart Grid Energy Management. *IEEE Trans. Ind. Inform.* **2019**, *15*, 4086–4098. [CrossRef]
28. Nian, R.; Liu, J.; Huang, B. A review on Reinforcement Learning: Introduction and applications in industrial process control. *Comput. Chem. Eng.* **2020**, *139*, 106886. [CrossRef]
29. Coronato, A.; Naeem, M.; de Pietro, G.; Paragliola, G. Reinforcement Learning for intelligent healthcare applications: A survey. *Artif. Intell. Med.* **2020**, *109*, 101964. [CrossRef] [PubMed]
30. Polydoros, A.S.; Nalpantidis, L. Survey of model-based Reinforcement Learning: Applications on robotics. *J. Intell. Robot. Syst.* **2017**, *86*, 153–173. [CrossRef]
31. Khan, M.A.M.; Khan, M.R.J.; Tooshil, A.; Sikder, N.; Mahmud, M.P.; Kouzani, A.Z.; Nahid, A.A. A systematic review on Reinforcement Learning-based robotics within the last decade. *IEEE Access* **2020**, *8*, 176598–176623. [CrossRef]
32. Michailidis, I.T.; Michailidis, P.; Alexandridou, K.; Brewick, P.T.; Masri, S.F.; Kosmatopoulos, E.B.; Chassiakos, A. Seismic Active Control under Uncertain Ground Excitation: An Efficient Cognitive Adaptive Optimization Approach. In Proceedings of the 2018 5th International Conference on Control, Decision and Information Technologies (CoDIT), Thessaloniki, Greece, 10–13 April 2018; pp. 847–852.
33. Karatzinis, G.D.; Michailidis, P.; Michailidis, I.T.; Kapoutsis, A.C.; Kosmatopoulos, E.B.; Boutalis, Y.S. Coordinating heterogeneous mobile sensing platforms for effectively monitoring a dispersed gas plume. *Integr.-Comput.-Aided Eng.* **2022**, *29*, 411–429. [CrossRef]
34. Salavasidis, G.; Kapoutsis, A.C.; Chatzichristofis, S.A.; Michailidis, P.; Kosmatopoulos, E.B. Autonomous trajectory design system for mapping of unknown sea-floors using a team of AUVs. In Proceedings of the 2018 European Control Conference (ECC), Limassol, Cyprus, 12–15 June 2018; pp. 1080–1087.
35. Keroglou, C.; Kansizoglou, I.; Michailidis, P.; Oikonomou, K.M.; Papapetros, I.T.; Dragkola, P.; Michailidis, I.T.; Gasteratos, A.; Kosmatopoulos, E.B.; Sirakoulis, G.C. A Survey on Technical Challenges of Assistive Robotics for Elder People in Domestic Environments: The ASPiDA Concept. *IEEE Trans. Med. Robot. Bionics* **2023**, *5*, 196–205. [CrossRef]
36. Michailidis, I.T.; Manolis, D.; Michailidis, P.; Diakaki, C.; Kosmatopoulos, E.B. Autonomous self-regulating intersections in large-scale urban traffic networks: A Chania city case study. In Proceedings of the 2018 5th International Conference on Control, Decision and Information Technologies (CoDIT), Thessaloniki, Greece, 10–13 April 2018; pp. 853–858.
37. Moerland, T.M.; Broekens, J.; Plaat, A.; Jonker, C.M. Model-based Reinforcement Learning: A survey. *Found. Trends® Mach. Learn.* **2023**, *16*, 1–118. [CrossRef]
38. Pong, V.; Gu, S.; Dalal, M.; Levine, S. Temporal difference models: Model-free Deep RL for model-based control. *arXiv* **2018**, arXiv:1802.09081.
39. Sun, W.; Jiang, N.; Krishnamurthy, A.; Agarwal, A.; Langford, J. Model-based rl in contextual decision processes: Pac bounds and exponential improvements over model-free approaches. In Proceedings of the Conference on Learning Theory, Phoenix, AZ, USA, 25–28 June 2019; pp. 2898–2933.
40. Lu, R.; Hong, S.H.; Zhang, X. A dynamic pricing demand response algorithm for smart grid: Reinforcement Learning approach. *Appl. Energy* **2018**, *220*, 220–230. [CrossRef]
41. Aktas, A.; Erhan, K.; Özdemir, S.; Özdemir, E. Dynamic energy management for photovoltaic power system including hybrid energy storage in smart grid applications. *Energy* **2018**, *162*, 72–82. [CrossRef]
42. Korkas, C.D.; Baldi, S.; Michailidis, P.; Kosmatopoulos, E.B. A cognitive stochastic approximation approach to optimal charging schedule in electric vehicle stations. In Proceedings of the 2017 25th Mediterranean Conference on Control and Automation (MED), Valletta, Malta, 3–6 July 2017; pp. 484–489.
43. Mosavi, A.; Salimi, M.; Faizollahzadeh Ardabili, S.; Rabczuk, T.; Shamshirband, S.; Varkonyi-Koczy, A.R. State of the art of Machine Learning models in energy systems, a systematic review. *Energies* **2019**, *12*, 1301. [CrossRef]
44. Mason, K.; Grijalva, S. A review of Reinforcement Learning for autonomous building energy management. *Comput. Electr. Eng.* **2019**, *78*, 300–312. [CrossRef]
45. Wang, Z.; Hong, T. Reinforcement Learning for building controls: The opportunities and challenges. *Appl. Energy* **2020**, *269*, 115036. [CrossRef]
46. Shaqour, A.; Hagishima, A. Systematic Review on Deep Reinforcement Learning-Based Energy Management for Different Building Types. *Energies* **2022**, *15*, 8663. [CrossRef]
47. Abdullah, H.M.; Gastli, A.; Ben-Brahim, L. Reinforcement Learning based EV charging management systems–a review. *IEEE Access* **2021**, *9*, 41506–41531. [CrossRef]
48. Sutton, R.S.; Barto, A.G. *Reinforcement Learning: An Introduction*, 2nd ed.; MIT Press: Cambridge, MA, USA, 2018.
49. Wiering, M.; Otterlo, M.v. *Reinforcement Learning: State-of-the-Art*; Springer: Berlin/Heidelberg, Germany, 2012.
50. Arulkumaran, K.; Deisenroth, M.P.; Brundage, M.; Bharath, A.A. Deep Reinforcement Learning: A brief survey. *IEEE Signal Process. Mag.* **2017**, *34*, 26–38. [CrossRef]
51. Schulman, J.; Wolski, F.; Dhariwal, P.; Radford, A.; Klimov, O. Proximal Policy Optimization Algorithms. *arXiv* **2017**, arXiv:1707.06347.

52. Schulman, J.; Levine, S.; Abbeel, P.; Jordan, M.; Moritz, P. Trust Region Policy Optimization. In Proceedings of the 32nd International Conference on Machine Learning, Lille, France, 7–9 July 2015; Volume 37, pp. 1889–1897.
53. Sutton, R.S.; McAllester, D.; Singh, S.; Mansour, Y. Policy Gradient Methods for Reinforcement Learning with Function Approximation. In *Advances in Neural Information Processing Systems*; Solla, S., Leen, T., Müller, K., Eds.; MIT Press: Cambridge, MA, USA, 1999; Volume 12.
54. Lillicrap, T.P.; Hunt, J.J.; Pritzel, A.; Heess, N.M.O.; Erez, T.; Tassa, Y.; Silver, D.; Wierstra, D. Continuous control with deep Reinforcement Learning. *arXiv* **2015**, arXiv:1509.02971.
55. Haarnoja, T.; Zhou, A.; Abbeel, P.; Levine, S. Soft Actor–Critic: Off-Policy Maximum Entropy Deep Reinforcement Learning with a Stochastic Actor. In Proceedings of the 35th International Conference on Machine Learning, Stockholm, Sweden, 10–15 July 2018; Volume 80, pp. 1861–1870.
56. Mnih, V.; Badia, A.P.; Mirza, M.; Graves, A.; Lillicrap, T.; Harley, T.; Silver, D.; Kavukcuoglu, K. Asynchronous Methods for Deep Reinforcement Learning. In Proceedings of the 33rd International Conference on Machine Learning, New York, NY, USA, 20–22 June 2016; Volume 48, pp. 1928–1937.
57. Watkins, C. Learning from Delayed Rewards. Ph.D. Thesis, King's College, London, UK, 1989.
58. Watkins, C.; Dayan, P. Q-Learning. *Mach. Learn.* **1992**, *8*, 279–292. [CrossRef]
59. Hasselt, H. Double Q-Learning. In *Advances in Neural Information Processing Systems*; Lafferty, J., Williams, C., Shawe-Taylor, J., Zemel, R., Culotta, A., Eds.; Curran Associates, Inc.: Red Hook, NY, USA, 2010; Volume 23.
60. Mnih, V.; Kavukcuoglu, K.; Silver, D.; Rusu, A.A.; Veness, J.; Bellemare, M.G.; Graves, A.; Riedmiller, M.A.; Fidjeland, A.K.; Ostrovski, G.; et al. Human-level control through deep Reinforcement Learning. *Nature* **2015**, *518*, 529–533. [CrossRef]
61. Wang, Z.; Schaul, T.; Hessel, M.; Hasselt, H.; Lanctot, M.; Freitas, N. Dueling Network Architectures for Deep Reinforcement Learning. In Proceedings of the 33rd International Conference on Machine Learning, New York, NY, USA, 20–22 June 2016; Volume 48, pp. 1995–2003.
62. Fujimoto, S.; van Hoof, H.; Meger, D. Addressing Function Approximation Error in Actor–Critic Methods. *arXiv* **2018**, arXiv:1802.09477.
63. Bellemare, M.G.; Dabney, W.; Munos, R. A Distributional Perspective on Reinforcement Learning. *arXiv* **2017**, arXiv:1707.06887.
64. Kuznetsova, E.; Li, Y.F.; Ruiz, C.; Zio, E.; Ault, G.; Bell, K. Reinforcement Learning for microgrid energy management. *Energy* **2013**, *59*, 133–146. [CrossRef]
65. Wei, C.; Zhang, Z.; Qiao, W.; Qu, L. Reinforcement-learning-based intelligent maximum power point tracking control for wind energy-conversion systems. *IEEE Trans. Ind. Electron.* **2015**, *62*, 6360–6370. [CrossRef]
66. Wei, C.; Zhang, Z.; Qiao, W.; Qu, L. An adaptive network-based Reinforcement Learning method for MPPT control of PMSG wind energy-conversion systems. *IEEE Trans. Power Electron.* **2016**, *31*, 7837–7848. [CrossRef]
67. Kofinas, P.; Doltsinis, S.; Dounis, A.; Vouros, G. A Reinforcement Learning approach for MPPT control method of photovoltaic sources. *Renew. Energy* **2017**, *108*, 461–473. [CrossRef]
68. Remani, T.; Jasmin, E.; Ahamed, T.I. Residential Load Scheduling With Renewable Generation in the Smart Grid: A Reinforcement Learning Approach. *IEEE Syst. J.* **2019**, *13*, 3283–3294. [CrossRef]
69. Diao, R.; Wang, Z.; Shi, D.; Chang, Q.; Duan, J.; Zhang, X. Autonomous Voltage Control for Grid Operation Using Deep Reinforcement Learning. In Proceedings of the 2019 IEEE Power & Energy Society General Meeting (PESGM), Atlanta, GA, USA, 4–8 August 2019; pp. 1–5. [CrossRef]
70. Rocchetta, R.; Bellani, L.; Compare, M.; Zio, E.; Patelli, E. A Reinforcement Learning framework for optimal operation and maintenance of power grids. *Appl. Energy* **2019**, *241*, 291–301. [CrossRef]
71. Zhang, B.; Hu, W.; Cao, D.; Huang, Q.; Chen, Z.; Blaabjerg, F. Deep Reinforcement Learning–based approach for optimizing energy conversion in integrated electrical and heating system with renewable energy. *Energy Convers. Manag.* **2019**, *202*, 112199. [CrossRef]
72. Ji, Y.; Wang, J.; Xu, J.; Fang, X.; Zhang, H. Real-time energy management of a microgrid using deep reinforcement learning. *Energies* **2019**, *12*, 2291. [CrossRef]
73. Phan, B.C.; Lai, Y.C. Control strategy of a hybrid renewable energy system based on Reinforcement Learning approach for an isolated microgrid. *Appl. Sci.* **2019**, *9*, 4001. [CrossRef]
74. Saenz-Aguirre, A.; Zulueta, E.; Fernandez-Gamiz, U.; Lozano, J.; Lopez-Guede, J.M. Artificial neural network based Reinforcement Learning for wind turbine yaw control. *Energies* **2019**, *12*, 436. [CrossRef]
75. Liu, H.; Yu, C.; Wu, H.; Duan, Z.; Yan, G. A new hybrid ensemble deep Reinforcement Learning model for wind speed short term forecasting. *Energy* **2020**, *202*, 117794. [CrossRef]
76. Jeong, J.; Kim, H. DeepComp: Deep Reinforcement Learning based renewable energy error compensable forecasting. *Appl. Energy* **2021**, *294*, 116970. [CrossRef]
77. Cao, D.; Hu, W.; Zhao, J.; Huang, Q.; Chen, Z.; Blaabjerg, F. A multi-agent deep Reinforcement Learning based voltage regulation using coordinated PV inverters. *IEEE Trans. Power Syst.* **2020**, *35*, 4120–4123. [CrossRef]
78. Zhao, H.; Zhao, J.; Qiu, J.; Liang, G.; Dong, Z.Y. Cooperative wind farm control with deep Reinforcement Learning and knowledge-assisted learning. *IEEE Trans. Ind. Inform.* **2020**, *16*, 6912–6921. [CrossRef]
79. Guo, C.; Wang, X.; Zheng, Y.; Zhang, F. Real-time optimal energy management of microgrid with uncertainties based on deep Reinforcement Learning. *Energy* **2022**, *238*, 121873. [CrossRef]

80. Sierla, S.; Ihasalo, H.; Vyatkin, V. A Review of Reinforcement Learning Applications to Control of Heating, Ventilation and Air Conditioning Systems. *Energies* **2022**, *15*, 3526. [CrossRef]
81. Barrett, E.; Linder, S. Autonomous HVAC control, A Reinforcement Learning approach. In Proceedings of the Machine Learning and Knowledge Discovery in Databases: European Conference, ECML PKDD 2015, Porto, Portugal, 7–11 September 2015; pp. 3–19.
82. Ruelens, F.; Claessens, B.J.; Quaiyum, S.; De Schutter, B.; Babuška, R.; Belmans, R. Reinforcement Learning applied to an electric water heater: From theory to practice. *IEEE Trans. Smart Grid* **2016**, *9*, 3792–3800. [CrossRef]
83. Al-Jabery, K.; Xu, Z.; Yu, W.; Wunsch, D.C.; Xiong, J.; Shi, Y. Demand-side management of domestic electric water heaters using approximate dynamic programming. *IEEE Trans. Comput.-Aided Des. Integr. Circuits Syst.* **2016**, *36*, 775–788. [CrossRef]
84. Cheng, Z.; Zhao, Q.; Wang, F.; Jiang, Y.; Xia, L.; Ding, J. Satisfaction based Q-Learning for integrated lighting and blind control. *Energy Build.* **2016**, *127*, 43–55. [CrossRef]
85. Wei, T.; Wang, Y.; Zhu, Q. Deep Reinforcement Learning for building HVAC control. In Proceedings of the 54th Annual Design Automation Conference 2017, Austin, TX, USA, 18–22 June 2017; pp. 1–6.
86. Chen, Y.; Norford, L.K.; Samuelson, H.W.; Malkawi, A. Optimal control of HVAC and window systems for natural ventilation through Reinforcement Learning. *Energy Build.* **2018**, *169*, 195–205. [CrossRef]
87. Jia, R.; Jin, M.; Sun, K.; Hong, T.; Spanos, C. Advanced building control via deep Reinforcement Learning. *Energy Procedia* **2019**, *158*, 6158–6163. [CrossRef]
88. Valladares, W.; Galindo, M.; Gutiérrez, J.; Wu, W.C.; Liao, K.K.; Liao, J.C.; Lu, K.C.; Wang, C.C. Energy optimization associated with thermal comfort and indoor air control via a deep Reinforcement Learning algorithm. *Build. Environ.* **2019**, *155*, 105–117. [CrossRef]
89. Kazmi, H.; Suykens, J.; Balint, A.; Driesen, J. Multi-agent Reinforcement Learning for modeling and control of thermostatically controlled loads. *Appl. Energy* **2019**, *238*, 1022–1035. [CrossRef]
90. Park, J.Y.; Dougherty, T.; Fritz, H.; Nagy, Z. LightLearn: An adaptive and occupant centered controller for lighting based on Reinforcement Learning. *Build. Environ.* **2019**, *147*, 397–414. [CrossRef]
91. Ding, X.; Du, W.; Cerpa, A. Octopus: Deep Reinforcement Learning for holistic smart building control. In Proceedings of the 6th ACM International Conference on Systems for Energy-Efficient Buildings, Cities and Transportation, New York, NY, USA, 13–14 November 2019; pp. 326–335.
92. Brandi, S.; Piscitelli, M.S.; Martellacci, M.; Capozzoli, A. Deep Reinforcement Learning to optimise indoor temperature control and heating energy consumption in buildings. *Energy Build.* **2020**, *224*, 110225. [CrossRef]
93. Lissa, P.; Deane, C.; Schukat, M.; Seri, F.; Keane, M.; Barrett, E. Deep Reinforcement Learning for home energy management system control. *Energy AI* **2021**, *3*, 100043. [CrossRef]
94. Jiang, Z.; Risbeck, M.J.; Ramamurti, V.; Murugesan, S.; Amores, J.; Zhang, C.; Lee, Y.M.; Drees, K.H. Building HVAC control with Reinforcement Learning for reduction of energy cost and demand charge. *Energy Build.* **2021**, *239*, 110833. [CrossRef]
95. Gupta, A.; Badr, Y.; Negahban, A.; Qiu, R.G. Energy-efficient heating control for smart buildings with deep Reinforcement Learning. *J. Build. Eng.* **2021**, *34*, 101739. [CrossRef]
96. De Somer, O.; Soares, A.; Vanthournout, K.; Spiessens, F.; Kuijpers, T.; Vossen, K. Using Reinforcement Learning for demand response of domestic hot water buffers: A real-life demonstration. In Proceedings of the 2017 IEEE PES Innovative Smart Grid Technologies Conference Europe (ISGT-Europe), Turin, Italy, 26–29 September 2017; pp. 1–7.
97. Zhang, Z.; Chong, A.; Pan, Y.; Zhang, C.; Lu, S.; Lam, K.P. A deep Reinforcement Learning approach to using whole building energy model for hvac optimal control. In Proceedings of the 2018 Building Performance Analysis Conference and SimBuild, Chicago, IL, USA, 26–28 September 2018; Volume 3, pp. 22–23.
98. Gao, G.; Li, J.; Wen, Y. Energy-efficient thermal comfort control in smart buildings via deep Reinforcement Learning. *arXiv* **2019**, arXiv:1901.04693.
99. Azuatalam, D.; Lee, W.L.; de Nijs, F.; Liebman, A. Reinforcement Learning for whole-building HVAC control and demand response. *Energy AI* **2020**, *2*, 100020. [CrossRef]
100. Du, Y.; Zandi, H.; Kotevska, O.; Kurte, K.; Munk, J.; Amasyali, K.; Mckee, E.; Li, F. Intelligent multi-zone residential HVAC control strategy based on deep Reinforcement Learning. *Appl. Energy* **2021**, *281*, 116117. [CrossRef]
101. Pinto, G.; Deltetto, D.; Capozzoli, A. Data-driven district energy management with surrogate models and deep Reinforcement Learning. *Appl. Energy* **2021**, *304*, 117642. [CrossRef]
102. Pinto, G.; Piscitelli, M.S.; Vázquez-Canteli, J.R.; Nagy, Z.; Capozzoli, A. Coordinated energy management for a cluster of buildings through deep Reinforcement Learning. *Energy* **2021**, *229*, 120725. [CrossRef]
103. Vandael, S.; Claessens, B.; Ernst, D.; Holvoet, T.; Deconinck, G. Reinforcement Learning of heuristic EV fleet charging in a day-ahead electricity market. *IEEE Trans. Smart Grid* **2015**, *6*, 1795–1805. [CrossRef]
104. Chiş, A.; Lundén, J.; Koivunen, V. Reinforcement Learning-based plug-in electric vehicle charging with forecasted price. *IEEE Trans. Veh. Technol.* **2016**, *66*, 3674–3684.
105. Mbuwir, B.V.; Ruelens, F.; Spiessens, F.; Deconinck, G. Battery energy management in a microgrid using batch reinforcement learning. *Energies* **2017**, *10*, 1846. [CrossRef]
106. Da Silva, F.L.; Nishida, C.E.; Roijers, D.M.; Costa, A.H.R. Coordination of electric vehicle charging through multiagent Reinforcement Learning. *IEEE Trans. Smart Grid* **2019**, *11*, 2347–2356. [CrossRef]

107. Qian, T.; Shao, C.; Wang, X.; Shahidehpour, M. Deep Reinforcement Learning for EV charging navigation by coordinating smart grid and intelligent transportation system. *IEEE Trans. Smart Grid* **2019**, *11*, 1714–1723. [CrossRef]
108. Sadeghianpourhamami, N.; Deleu, J.; Develder, C. Definition and evaluation of model-free coordination of electrical vehicle charging with Reinforcement Learning. *IEEE Trans. Smart Grid* **2019**, *11*, 203–214. [CrossRef]
109. Wang, S.; Bi, S.; Zhang, Y.A. Reinforcement Learning for real-time pricing and scheduling control in EV charging stations. *IEEE Trans. Ind. Inform.* **2019**, *17*, 849–859. [CrossRef]
110. Chang, F.; Chen, T.; Su, W.; Alsafasfeh, Q. Control of battery charging based on Reinforcement Learning and long short-term memory networks. *Comput. Electr. Eng.* **2020**, *85*, 106670. [CrossRef]
111. Lee, J.; Lee, E.; Kim, J. Electric vehicle charging and discharging algorithm based on Reinforcement Learning with data-driven approach in dynamic pricing scheme. *Energies* **2020**, *13*, 1950. [CrossRef]
112. Tuchnitz, F.; Ebell, N.; Schlund, J.; Pruckner, M. Development and evaluation of a smart charging strategy for an electric vehicle fleet based on Reinforcement Learning. *Appl. Energy* **2021**, *285*, 116382. [CrossRef]
113. Li, H.; Wan, Z.; He, H. Constrained EV charging scheduling based on safe deep reinforcement learning. *IEEE Trans. Smart Grid* **2019**, *11*, 2427–2439. [CrossRef]
114. Zhang, F.; Yang, Q.; An, D. CDDPG: A deep-reinforcement-learning-based approach for electric vehicle charging control. *IEEE Internet Things J.* **2020**, *8*, 3075–3087. [CrossRef]
115. Dorokhova, M.; Martinson, Y.; Ballif, C.; Wyrsch, N. Deep Reinforcement Learning Control of electric vehicle charging in the presence of photovoltaic generation. *Appl. Energy* **2021**, *301*, 117504. [CrossRef]
116. Park, S.; Pozzi, A.; Whitmeyer, M.; Perez, H.; Kandel, A.; Kim, G.; Choi, Y.; Joe, W.T.; Raimondo, D.M.; Moura, S. A deep Reinforcement Learning framework for fast charging of li-ion batteries. *IEEE Trans. Transp. Electrif.* **2022**, *8*, 2770–2784. [CrossRef]
117. Belousov, B.; Abdulsamad, H.; Klink, P.; Parisi, S.; Peters, J. *Reinforcement Learning Algorithms: Analysis and Applications*; Springer: Berlin/Heidelberg, Germany, 2021.
118. Kabanda, G.; Kannan, H. A Systematic Literature Review of Reinforcement Algorithms in Machine Learning. In *Handbook of Research on AI and Knowledge Engineering for Real-Time Business Intelligence*; IGI Global: Hershey, PA, USA, 2023; pp. 17–33.
119. Mosavi, A.; Faghan, Y.; Ghamisi, P.; Duan, P.; Ardabili, S.F.; Salwana, E.; Band, S.S. Comprehensive review of deep Reinforcement Learning methods and applications in economics. *Mathematics* **2020**, *8*, 1640. [CrossRef]
120. Glorennec, P.Y. Reinforcement Learning: An overview. In Proceedings of the European Symposium on Intelligent Techniques (ESIT-00), Aachen, Germany, 14–15 September 2000; pp. 14–15.
121. Cao, D.; Hu, W.; Zhao, J.; Zhang, G.; Zhang, B.; Liu, Z.; Chen, Z.; Blaabjerg, F. Reinforcement Learning and its applications in modern power and energy systems: A review. *J. Mod. Power Syst. Clean Energy* **2020**, *8*, 1029–1042. [CrossRef]
122. Muriithi, G.; Chowdhury, S. Optimal energy management of a grid-tied solar PV-battery microgrid: A Reinforcement Learning approach. *Energies* **2021**, *14*, 2700. [CrossRef]

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.