International Journal of Multiphysics
Volume 18, No. 4, 2024
ISSN: 1750-9548

A Reinforcement Learning Framework for Scalable and Cost-Efficient Energy Management in Smart Grids

Atef Gharbi[1*], Mohamed Ayari[2,3], Akil Elkamel[1], Mahmoud Salaheldin Elsayed[4], Zeineb Klai[4] , Nouha Khedhiri[1]

Department of Information Systems, Faculty of Computing and Information Technology, Northern Border University, Rafha, Saudi Arabia[1]

Department of Information Technology, Faculty of Computing and Information Technology, Northern Border University, Rafha, Saudi Arabia[2]

SYSCOM Laboratory, National Engineering School of Tunis, University of Tunis El-Manar, Tunis 1068, Tunisia[3]

Department of Computer Sciences, Faculty of Computing and Information Technology, Northern Border University, Rafha, Saudi Arabia[4]

*Corresponding author: E-mail address: atef.gharbi@nbu.edu.sa

Abstract

Smart grids integrate renewable energy sources and enable dynamic demand responses to transform energy management. The complexity of managing multiple agent systems with different devices presents challenges in terms of scalability, computational efficiency and real-time adaptability. This paper introduces the new framework MARL-SG (Multi-Agent Reinforcement Learning for Smart Grids), which aims to optimize energy consumption across devices while maintaining grid stability, reducing costs and satisfying users. With MARL-SG, training is centralized, and execution is decentralized to ensure scaling, and advanced technologies such as Heuristic masking ensure the allocation of computational resources to critical tasks. According to the experimental results, the MARL-SG reduces energy costs during peak and off-peak hours, achieves almost perfect grid stability and provides reliable and cost-effective energy distribution. The framework will enable modern smart grids to manage energy more intelligently.

Keywords: Smart grid, demand response, reinforcement learning, Q-learning, Heuristic Masking.

I. INTRODUCTION

As renewable energy sources and intelligent technologies are increasingly integrated into the power grid, energy management has revolutionized, enabling dynamic demand responses, improved efficiency and grid stability. In multi-agent environments with diverse devices, energy management remains challenging due to the need for dynamic user behaviour, variable renewable energy output and real-time optimization. Thus, advanced energy management systems (EMSs) are needed to cope with these complexities. In the case of smart grids, the most promising approach to solving these challenges is reinforcement learning (RL). Researchers have conducted extensive research into RL techniques, such as optimizing energy consumption in smart buildings, managing renewable energy sources in home energy systems, and facilitating peer-to-peer energy trade. Multi-agent deep reinforcement learning (MADRL) enables the control of scalable and distributed residential energy systems and further advances this field [4, 5]. The use of RL-based frameworks for demand response management [6, 7], and intelligent scheduling of energy load [8, 9].

The MARL-SG (Multi-Agent Reinforcement Learning Architecture for Smart Grids) is a new framework for improving energy management in smart grids. To ensure scale and adaptability of large-scale energy systems, MARL-SG combines central training and decentralized execution. It focuses on high priority tasks, such as devices approaching delays, while incorporating heuristic masks to improve computing efficiency. In addition to reducing energy costs, MARL-SG improves grid stability and maintains user satisfaction by responding

dynamically to price signals and grid restrictions. Tests have shown that the framework outperforms traditional methods such as FIFO, which balance cost efficiency, task completion rates, and grid reliability. As a result of prior research in smart grid energy management, this work builds on studies on data-driven optimizations [10], renewable energy forecasts [11], HVAC control strategies [12], and home demand responses [13, 14]. MARL-SG solves the critical challenges of modern energy systems and provides a scalable and effective solution for intelligent energy management.

Despite significant advances in reinforcement learning (RL) and multi-agent systems, the approach to energy management of the existing smart grid remains limited. The RL-based methodology usually involves centralized training and implementation, which limits their scalability in large-scale energy systems involving multiple households and devices. In complex multiagent environments, peer-to-peer energy trading frameworks [3] and distributed control approaches [4, 8] are often difficult to coordinate effectively. Furthermore, computational inefficiency is a major concern because the methods used to respond to demands [6, 7] and schedule loads [8] are often based on comprehensive searches and high-dimensional state action spaces, resulting in significant computational overhead and limiting their real-time applicability. Insufficient integration of dynamic pricing mechanisms and user preferences is another failure. Although the RL framework promised to optimize energy consumption [1, 10], many failed to properly incorporate these aspects, resulting in suboptimal task schedules and a reduction in user satisfaction. In addition, HVAC control and renewable energy management [5, 2, 11] often fail to address critical grid constraints, leading to frequent overloads and instability. Finally, the static task prioritization strategies such as FIFO and energy-based priority [7, 9] lack the flexibility to consider constraints of urgency or delay, resulting in high energy costs and unsatisfactory services. Consequently, new solutions must address scalability, efficiency and adaptability, while maintaining user-centric and grid-oriented energy management.

This paper presents the new framework MARL-SG, which aims to address key constraints in the existing energy management approach. MARL-SG combines centralized training and decentralized execution, allowing efficient scalability of large systems and multi-family devices, as well as ensuring adaptability to dynamic and complex energy environments. This approach improves the applicability of real-time systems by filtering out irrelevant actions or states and focusing on high-priority tasks such as devices close to delayed limitations. As part of its reward function, the framework explicitly incorporates dynamic pricing signals and user satisfaction metrics to ensure that user-centric energy management strategies minimize energy costs and ensure task completion time. By actively monitoring grid capacity, MARL-SG prevents overloads and ensures reliable energy distribution under different demand conditions. MARL-SG dynamically plans tasks based on real-time system conditions such as price signals, device constraints, and grid loads, rather than static priority strategies such as FIFO. As a robust and scalable solution for modern smart grid management, MARL-SG has superior performance in key metrics such as energy savings, grid stability, and project completion rates, as well as traditional methods such as FIFO.

The paper is divided into the following sections: Section II deals with problem formulation, including state, action and reward definitions. Section III describes the MARL-SG algorithm, which integrates reinforcement learning and heuristic-based approaches. Section IV presents experimental results that validate the performance of the framework. Section V summarizes the key findings and future directions.

II. PROBLEM FORMULATION

Multi-Agent Reinforcement Learning (MARL) is used in the framework to optimize energy consumption in smart grid environments. In this model, there is a balance between local decision-making at home and global grid constraints. Centralized training and decentralized execution (CTDE) is a model used by MARL-SG to balance scale and system efficiency. Centralized training phases use grid-level information such as energy consumption and pricing signals to learn policies for all devices. Using reinforcement learning algorithms such as Q-learning, common policies are trained to capture system dynamics. During the decentralized execution phase, each device works independently and makes decisions only based on local conditions and observations. The execution phase does not depend on centralized data, so the system is scalable and efficient in large-scale smart grids.

The Energy Management System (EMS) integrates smart meters and control devices in the household and optimizes operations according to factors such as power costs, delays and user satisfaction. The following sections define the state, actions and rewards.

1. State Space in Smart Grid Systems

The State space S(t) captures information about the devices of each household, grid capacity and real-time constraints.

1.1 Device State in MARL-SG

In MARL-SG, the state of the device is defined by a key parameter that describes its operational status and constraints. There are two types of slots: requested slots (req(t)) and remaining slots (rem(t)). The required slot (req(t)) represents the total number of time slots needed to complete the device's task, and the remaining slot (rem(t)) represents the remaining slot. The Maximum Allowable Delay (MaxD) specifies the maximum tolerable delay when completing the task, and the Current Delay (Delay(t)) records the time when a task is delayed. These parameters provide an overview of the operation status and limitations of each device. When the device is not requested, or a task is requested in the current slot, the device's state and action are set to zero.

Delays increase by one when a task has been delayed previously. If the device is active in the previous slot, the delay value is not changed. The purpose of this parameter is to monitor and manage the progress of the task in relation to permitted delays.

The devices studied are divided into three different groups according to their operational requirements and constraints. Due to their essential functions, Must-Run devices must be activated immediately at request and cannot be delayed under any circumstances. Uninterrupted devices may experience delays if current delays do not exceed the maximum allowed delay (MaxD) but should continue to operate uninterruptedly until the task is completed. There is greater flexibility with interruptible devices, as they can be delayed until the maximum permitted delay (MaxD), deactivated, or stopped, and are not required to complete tasks immediately. This classification makes smart grid device operations priority.

1.2 Grid State: Key Parameters for Grid Operation

The grid state is defined by a key parameter that defines the grid operation status and constraints. Cost Level (Cost(t)) is a dynamic price signal derived from total grid consumption and serves as an incentive to optimize energy use during peak and peak periods. Grid capacity (GridCap) is a measurement that determines the maximum amount of electricity that the grid can handle at any given time and is a constraint on grid stability and overload prevention. These parameters must be considered together to effectively balance energy demand and supply.

2. Action Space: Balancing Energy Efficiency and Task Requirements

The Action Space Act (t) of the device defines the possible state of operation that can be taken at any time. These actions include turn on the device (Act(t) = 1) to activate the device to perform its tasks, turn off (Act(t) = 0) to deactivate the device and the sleep mode (Act(t) = -1) to reduce energy consumption. The action space allows for a balance between energy efficiency and task requirements. According to the local state of each device and observations, each agent (household) selects an action. For each device, the following restrictions must be followed to determine the current $Act_{n,d}(t)$ action (e.g. on, off or sleep mode):

- ✓ **No Request**: Devices that have not been requested remain OFF:

$$\text{If } Req_{n,d}(t) = 0 \rightarrow Act_{n,d}(t) = 0$$

- ✓ **Maximum Delay Reached**: When a device has an incomplete task and its current delay equals the maximum allowable delay, it must be activated:

$$\text{If } Req_{n,d}(t)>0 \ \& Delay_{n,d}(t)=MaxD_{n,d} \rightarrow Act_{n,d}(t)=1$$

- ✓ **Non-Interruptible Devices**: If the device is non-interruptible, has a non-zero requested range, and was active in the previous time slot, it will remain ON until the task is completed:

$$\text{If } \text{Req}_{n,d}(t)>0 \ \& \text{Delay}(t)_{n,d}=1 \ \& \text{Act}_{n,d}(t\text{-}1)=1 \rightarrow \text{Act}_{n,d}(t)=1$$

- ✓ **Must-Run Devices**: For must-run devices, the action directly follows the request:

$$\text{If } \text{Req}_{n,d}(t)>0 \ \& \text{Delay}(t)_{n,d}=0 \ \rightarrow \text{Act}_{n,d}(t)=1$$

3. Reward Mechanism: Optimizing Cost, Satisfaction, and Stability

The reward function **R(t)** combines multiple objectives:

$$R_t = \alpha_1 R_{cost}(t) + \alpha_2 R_{satisfaction}(t) + \alpha_3 R_{grid}(t)$$

MARL-SG's reward function balances energy efficiency, user satisfaction and grid stability. Energy savings are rewarded by combining equipment operations with low-cost time slots ($R_{cost}(t)$) and encouraging efficient energy consumption during peak periods.

Satisfaction ($R_{satisfaction}(t)$) guarantees the satisfaction of users by minimizing tasks completed and assigning priority to timely devices completed. Grid stability ($R_{grid}(t)$) prevents overloading by punishing scenarios where the grid capacity is exceeded and maintaining grid reliability. These components together guide the system to optimal energy management while maintaining operational limitations. Adjustment of weights $\alpha_1, \alpha_2, \alpha_3$ can flexibly give priority to user experience, cost and grid efficiency.

4. Objective: Maximizing Cumulative Rewards in MARL-SG

The MARL-SG aims to find the best policy π^* to maximize the expected cumulative rewards over a limited time window T while adhering to operational constraints:

$$\pi^* = \arg\max_{\pi} \mathrm{E}\left[\sum_{t=0}^{T} \gamma^t R_t(S(t), Act(t))\right]$$

Where $\gamma \in [0,1]$ is the discount factor that gives priority to immediate rewards over future benefits.

The operation of the devices in the MARL-SG is subject to specific restrictions to ensure the efficiency and reliability of performance. Maximum delay limits stipulate that the device must be activated immediately when the current delay (Delay(t)) reaches maximum delay (maxD).

According to capacity constraints, total energy consumption ($P_{total}(t)$) cannot exceed the grid capacity (GridCap), adhering to $P_{total}(t) \leq$ GridCap. In addition, priority rules require devices that must run immediately, such as essential medical equipment, to be activated immediately. As a result, these constraints give priority to task completion, ensuring grid stability and ensuring the smooth operation of the devices.

III. Algorithm and Heuristic for MARL-SG

The following description of the training and execution algorithms of MARL-SG incorporates advanced scaling and efficiency techniques. For clarity, the heuristic approach is described separately. MARL-SG is engaged in two phases to address the challenges of multi-agent energy management in smart grids. During the training phase (centralized), global policies are trained using shared parameters and global models of learning are integrated to improve sample efficiency. To ensure the stability of the grid, this phase focuses on managing the behavior of different devices and optimizing energy consumption. The execution phase (decentralized) is a stage in which each device independently executes the learned policy, applying heuristic masking to focus on the relevant states and actions. By ensuring the efficiency and adaptability of computing, algorithms can scale and function in dynamic energy environments.

Algorithm 1: MARL-SG Training and Execution

Input:

- State space S, Action space A, Reward function R(S,A)
- Device constraints Req, Rem, MaxD and grid capacity GridCap

Output:

- Optimized policy π^*

1. Initialize Parameters:

- Initialize the policy grid π_θ and the value function Q_ϕ.
- Set up shared parameters across agents shared_params.
- Initialize an empty replay buffer B.
- Initialize the world model M.

2. While Not Converged

a) Episode Simulation:

- For each episode:

1. Initialize the environment and device states (Req, Rem, MaxD, GridCap) to obtain the initial state s_0.
2. Simulate Time Steps: For t=0 to T−1:
   - Sample an action $a_t \sim \pi_\theta(s_t)$.
   - Execute a_t in the environment to obtain the next state s_{t+1} and reward r_t.
   - Store the transition (s_t, a_t, r_t, s_{t+1}) in the replay buffer B.

b) Train World Model:

- Train the world model M using transitions stored in B.
- Generate synthetic transitions (s_t, a_t, r_t, s_{t+1}) using M and augment B.

c) Policy and Value Function Optimization:

For each update step k = 1 to N:

1. Sample a mini-batch from B (including synthetic transitions).
2. Update the policy π_θ and the value function Q_ϕ using a reinforcement learning algorithm.
3. Update shared parameters shared_params across agents for stability and scalability.

d) Performance Evaluation:

- Evaluate the performance metrics (e.g., cumulative reward, energy cost, grid stability).

- If the performance metrics meet the desired threshold, set converged ←True.

3. Return Optimal Policy:

- Return the optimized policy π_θ as π^*.

To achieve robust and scalable energy management, MARL-SG combines centralized training with decentralized implementation. Data from all devices are aggregated in the centralized training phase to ensure the policy is robust and generalized. Using a learned global model, policy can handle edge cases and invisible scenarios more effectively. In the decentralized execution phase, policies apply independently to each device to ensure the scalability of large grids. Further improving performance is the heuristic approach, which dynamically focuses on devices that are approaching deadlines or have a significant impact on user satisfaction and cost. It aligns actions with grid constraints and prevents overload by monitoring the total load of the grid. The focus of critical devices reduces computational requirements and enables scalability and efficiency in real time. Using the heuristic algorithm, computational resources focus on the most important components through dynamic filtering states and actions. This guarantees real-time scalability and avoids grid instability.

Algorithm 2: Heuristic Masking for Efficient Decision-Making

Input:

- Current state S_t
- Device parameters Req, Rem, MaxD, GridCap

Output:

- Filtered state S_t'

Steps:

1. Compute Device Urgency:
   - For each device d, calculate urgency: $U_d = \frac{Rem_d}{Req_d} + \frac{Delay_d}{MaxD}$
2. Select Critical Devices:
   - Rank devices by U_d.
   - Select the top k devices with the highest scores.
3. Evaluate Grid Stability:
   - Compute current grid load: $Load_t = \sum_{d \in \text{active}} Power_d$
   - If $Load_t$ > GridCap, deactivate least urgent devices until $Load_t \leq$ GridCap.
4. Filter State:
   - Construct reduced state S_t' with selected devices.

IV. Experimental

Based on a grid configuration with a 50 kW capacity, the simulation included five time slots from 18h00 to 22:59, with dynamic prices rising during peak times (6 units) and declining during peak times (4 units). In the simulation, each household contains a random power consumption of 10 devices (1–4 kW), task requirements and maximum delays. We compared two methods: MARL-SG and the first-in-first-out strategy FIFO. The simulation evaluation criteria included the energy costs during peak and off-peak hours, which measure total energy expenditures across time frames at different prices; the completion rate of tasks, which represents the percentage of completed tasks within the limits of reasonable delays; and the stability of the grid, which represents the percentage of time that the grid load remains within the allowable capacity. To make a robust statistically significant comparison, 100 episodes have been simulated per method.

1. Energy Cost Comparison

Figure 1 shows that MARL-SG costs less during peak and off- peak hours than FIFO, while FIFO costs more at both times. Using intelligent simulated decision-making to delay non-critical tasks during peak times, MARL-SG can reduce energy costs by dispersing energy consumption during peak times when costs are lower. On the contrary, FIFO incurs higher costs because the tasks are executed in the order requested, without considering cost variations, resulting in excessive energy consumption during peak hours.

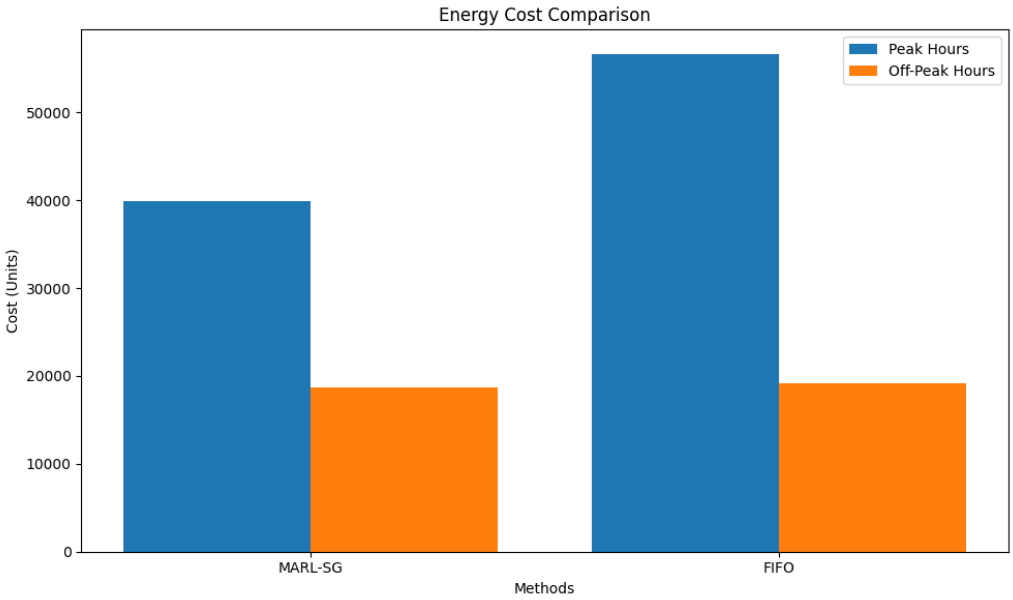

Figure 1. Comparison of Energy Costs During Peak and Off-Peak Hours for MARL-SG and FIFO Methods

2. Grid Stability

Figure 2 shows that MARL-SG had a nearly perfect grid stability, nearly 100%, keeping the grid load well within the capacity, whereas FIFO had a significantly lower stability, about 60%, and more cases of over-the-grid capacity. MARL-SG maintains almost perfect stability by actively monitoring grid capacity and dynamically adjusting device activation, thereby avoiding grid overload. However, the lower stability of FIFO is due to the lack of capacity management because it processes requests without considering the current load of the grid and often exceeds its maximum capacity.

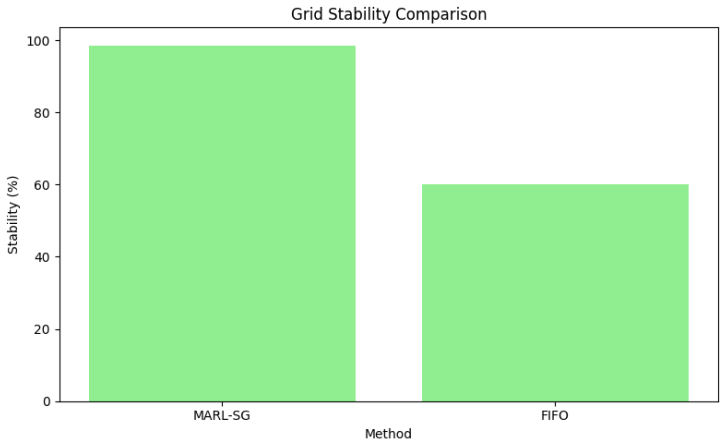

Figure 2. Grid Stability Comparison Between MARL-SG and FIFO Methods

V. Conclusion

The paper introduces MARL-SG, a framework for strengthening learning to optimize energy management in smart grids. MARL-SG combines centralized training and decentralized execution to achieve scalability and adaptability. Advanced technologies, such as heuristic masking and the generation of synthetic data, ensure computational efficiency and dynamic response to real-time pricing and grid constraints. Experimental studies have shown that MARL-SG reduces energy costs during peak and off-peak hours and maintains near-perfect grid stability. Consequently, it is ideal for dynamic prices and tight grid capacity scenarios where cost efficiency and grid reliability are essential. MARL-SG is a significant advance in smart grid management and offers an efficient and flexible approach to balancing energy costs, grid stability, and user satisfaction. To improve the performance and applicability of MARL-SG, future research will improve the RL model, incorporate advanced neural architectures and test it in real energy systems.

ACKNOWLEDGMENT

The authors extend their appreciation to the Deanship of Scientific Research at Northern Border University, Arar, KSA for funding this research work through the project number “NBU-FFR-2024-2441-05“.

Bibliography

[1] Yu, L., Qin, S., Zhang, M., Shen, C., Jiang, T., & Guan, X. (2021). A review of deep reinforcement learning for smart building energy management. *IEEE Internet of Things Journal*, *8*(15), 12046-12063.

[2] Ali, A. O., Elmarghany, M. R., Abdelsalam, M. M., Sabry, M. N., & Hamed, A. M. (2022). Closed-loop home energy management system with renewable energy sources in a smart grid: A comprehensive review. *Journal of Energy Storage*, *50*, 104609.

[3] Qiu, D., Ye, Y., Papadaskalopoulos, D., & Strbac, G. (2021). Scalable coordinated management of peer-to-peer energy trading: A multi-cluster deep reinforcement learning approach. *Applied energy*, *292*, 116940.

[4] Charbonnier, F., Morstyn, T., & McCulloch, M. D. (2022). Scalable multi-agent reinforcement learning for distributed control of residential energy flexibility. *Applied Energy*, *314*, 118825.

[5] Yu, L., Sun, Y., Xu, Z., Shen, C., Yue, D., Jiang, T., & Guan, X. (2020). Multi-agent deep reinforcement learning for HVAC control in commercial buildings. *IEEE Transactions on Smart Grid*, *12*(1), 407-419.

[6] Kumari, A., & Tanwar, S. (2021). A reinforcement-learning-based secure demand response scheme for smart grid system. *IEEE Internet of Things Journal*, *9*(3), 2180-2191.

[7] Wan, Y., Qin, J., Yu, X., Yang, T., & Kang, Y. (2021). Price-based residential demand response management in smart grids: A reinforcement learning-based approach. *IEEE/CAA Journal of Automatica Sinica*, *9*(1), 123-134.

[8] Chung, H. M., Maharjan, S., Zhang, Y., & Eliassen, F. (2020). Distributed deep reinforcement learning for intelligent load scheduling in residential smart grids. *IEEE Transactions on Industrial Informatics*, *17*(4), 2752-2763.

[9] Lu, T., Chen, X., McElroy, M. B., Nielsen, C. P., Wu, Q., & Ai, Q. (2020). A reinforcement learning-based decision system for electricity pricing plan selection by smart grid end users. *IEEE Transactions on Smart Grid*, *12*(3), 2176-2187.

[10] Ahsan, F., Dana, N. H., Sarker, S. K., Li, L., Muyeen, S. M., Ali, M. F., ... & Das, P. (2023). Data-driven next-generation smart grid towards sustainable energy evolution: techniques and technology review. *Protection and Control of Modern Power Systems*, *8*(3), 1-42.

[11] Mostafa, N., Ramadan, H. S. M., & Elfarouk, O. (2022). Renewable energy management in smart grids by using big data analytics and machine learning. *Machine Learning with Applications*, *9*, 100363.

[12] Liu, X., Ren, M., Yang, Z., Yan, G., Guo, Y., Cheng, L., & Wu, C. (2022). A multi-step predictive deep reinforcement learning algorithm for HVAC control systems in smart buildings. *Energy*, *259*, 124857.

[13] Kotsiopoulos, T., Sarigiannidis, P., Ioannidis, D., & Tzovaras, D. (2021). Machine learning and deep learning in smart manufacturing: The smart grid paradigm. *Computer Science Review*, *40*, 100341.

[14] Pallonetto, F., De Rosa, M., Milano, F., & Finn, D. P. (2019). Demand response algorithms for smart-grid ready residential buildings using machine learning models. *Applied energy*, *239*, 1265-1282.