IEEE Access
Multidisciplinary ⋮ Rapid Review ⋮ Open Access Journal

Received 2 February 2024, accepted 24 February 2024, date of publication 27 February 2024, date of current version 4 March 2024.

Digital Object Identifier 10.1109/ACCESS.2024.3370922

RESEARCH ARTICLE

Optimization of a Cluster-Based Energy Management System Using Deep Reinforcement Learning Without Affecting Prosumer Comfort: V2X Technologies and Peer-to-Peer Energy Trading

METE YAVUZ AND ÖMER CIHAN KIVANÇ, (Senior Member, IEEE)
Department of Electrical and Electronics Engineering, Istanbul Okan University, 34959 İstanbul, Turkey
Corresponding author: Mete Yavuz (mete.yavuzlar@gmail.com)

ABSTRACT The concept of Prosumer has enabled consumers to actively participate in Peer-to-Peer (P2P) energy trading, particularly as Renewable Energy Source (RES)s and Electric Vehicle (EV)s have become more accessible and cost-effective. In addition to the P2P energy trading, prosumers benefit from the relatively high energy capacity of EVs through the integration of Vehicle-to-X (V2X) technologies, such as Vehicle-to-Home (V2H), Vehicle-to-Load (V2L), and Vehicle-to-Grid (V2G). Optimization of an Energy Management System (EMS) is required to allocate the required energy efficiently within the cluster, due to the complex pricing and energy exchange mechanism of P2P energy trading and multiple EVs with V2X technologies. In this paper, Deep Reinforcement Learning (DRL) based EMS optimization method is proposed to optimize the pricing and energy exchanging mechanisms of the P2P energy trading without affecting the comfort of prosumers. The proposed EMS is applied to a small-scale cluster-based environment, including multiple (6) prosumers, P2P energy trading with novel hybrid pricing and energy exchanging mechanisms, and V2X technologies (V2H, V2L, and V2G) to reduce the overall energy costs and increase the Self-Sufficiency Ratio (SSR)s. Multi Double Deep Q-Network (DDQN) agents based DRL algorithm is implemented and the environment is formulated as a Markov Decision Process (MDP) to optimize the decision-making process. Numerical results show that the proposed EMS reduces the overall energy costs by 19.18%, increases the SSRs by 9.39%, and achieves an overall 65.87% SSR. Additionally, numerical results indicates that model-free DRL, such as DDQN agent based Deep Q-Network (DQN) Reinforcement Learning (RL) algorithm, promise to eliminate the energy management complexities with multiple uncertainties.

INDEX TERMS Energy management system, peer-to-peer energy trading, vehicle-to-home, multi-agent reinforcement learning, deep reinforcement learning, smart grids.

I. INTRODUCTION

Optimizing energy production and consumption, resource allocation, and reducing energy costs require the implementation of smart energy management strategies. Advancements

The associate editor coordinating the review of this manuscript and approving it for publication was Alexander Micallef.

in consumer electronic, changes in electronic device usage habits, increased industrialization, and accelerated production of EV have caused excessive energy consumption worldwide [1]. RESs are sustainable and clean energy sources to meet the energy demands, preventing adverse effects of climate change [2], [3]. Applications of Energy Storage System (ESS)s have acquired significant importance on optimizing the utilization of RESs [4]. The combination of multiple residents in a residential area, ESSs, RESs, EVs constitute a localized Microgrid (MG) with on/off-grid and autonomous operation capabilities [5], [6], [7]. The intermittent energy production profiles of RESs and combination of multiple components cause energy management complexities in MGs [8], [9], [10], [11]. Therefore, the implementation of smart energy management architectures are required to ensure efficient utilization of energy in MGs. Consequently, various smart energy management architectures, such as Home Energy Management System (HEMS)s, have been proposed in literature to enhance the operation of MG.

Due to the energy management complexities and intermittent energy production profiles of RESs, HEMSs are of great importance in energy management architectures. Energy cost reduction, flow optimization, enhancement in the overall self-sufficiency, peak shaving, and appliance scheduling are the primary objectives of HEMSs [12], [13]. The conventional HEMSs include shiftable and non-shiftable loads, EVs, RESs, smart meters, Internet of Things (IoT) devices, communication technologies, and bi-directional energy and data exchange capabilities with the electric grid [12], [13], [14]. The HEMS control strategies include long-term benefits based Demand Side Management (DSM) and short-term benefits based Demand Response (DR). DSM strategies improve consumption behaviors of consumers to reduce energy costs in the long term, encouraging consumers to use energy during the off-peak times instead of peak times. Opposite to DSM, DR strategies include short-term and real-term changes in response to price signals and needs of grid reliability. DSM and DR strategies include applications of peak-shaving, appliance scheduling, consumption planning and EV charging planning based on different pricing schemes, including Critical Peak Pricing (CPP), Day-Ahead Pricing (DAP), Time-of-Use (TOU), and Real Time Pricing (RTP) [15], [16], [17], [18], [19], [20], [21], [22], [23], [24], [25].

Several optimization methods, based on different algorithmic approaches and objective functions, have been proposed to improve the process of decision making in HEMSs. Objective functions have been intensely focused on reducing energy costs and optimizing the energy flow. The literature review indicates that proposed HEMSs include 2 fundamental optimization approaches: rule-based and learning based. Rule-based optimization approaches are based on predefined rules, attempting to achieve an optimized solution to a mathematical problem [26]. Rule-based optimization approaches include Action Dependent Heuristic Dynamic Programming (ADHDP) [27], [38], [39], static and dynamic rule-based approaches [28], [29], [30], [31], Particle Swarm Optimization (PSO) [35], [36], [37], and Fuzzy Logic [32], [33]. The primary objectives of rule-based optimization approaches are minimizing the energy costs, reducing energy consumption from the electric grid and emission, optimizing the energy consumption and flow, and maximizing the comforts of consumers [27], [28], [29], [30], [31], [32], [33], [35], [36], [37], [38]. Despite the easy-to-apply feature of rule-based approaches, learning-based optimization approaches are more suitable to solve difficult optimization problems in complex environment with many uncertainties. Therefore, Machine Learning (ML) based optimization methods have been widely investigated in the literature. RL is a sub-branch of ML techniques based on the maximization of an objective function through the interaction with the environment [34]. In addition to the conventional RL algorithm, DRL enhances the advantages of RL algorithms by involving Deep Neural Network (DNN) architectures during the learning. Especially, with the development of the first DQN algorithm using DNN architectures to find the optimal solution [105], DRL algorithms have rapidly gained importance in various application fields, including smart energy management and many others.

DRL algorithms aim to enhance the performance of the conventional RL algorithms by combining the Deep Learning (DL) and RL based optimization methods. DRL algorithms focus on evaluating the Q-function, although conventional RL algorithms evaluate the Q-table instead. Evaluating the Q-table could cause troubles with large-scale data. Therefore, DRL algorithms evaluate the optimal policy through adjusting the weights DNN architecture with training [128]. In general, DRL algorithms perform relatively better than conventional RL algorithms (without DNN) in complex environment with large-scale of information [106]. DRL algorithms differ from conventional RL algorithms in terms of evaluating the optimal policy. The optimization approach of DRL algorithms includes value-based, policy-based, model-based, and hierarchical-based optimization approaches [107]. The most common used algorithms of DRL include DQN and DDQN to find the optimal policy via online and target networks efficiently [108], Deep Deterministic Policy Gradient (DDPG) that uses actor-critic architecture [109], Proximal Policy Optimization (PPO) that evaluate the optimal policy via the combination of policy-based and value-based methods [110] and etc [107]. The application areas of DRL algorithms include task offloading, resource allocation and scheduling precisely in wireless, mobile networks and wireless power transfer [108], [111], [112], [113], financial and advertisement areas such as detecting digital advertising frauds, money phishing attacks, credit card frauds, and decision making in investments [114], [115], [116], [117], supply chain management systems [118], [119], autonomous vehicle technologies for high-speed path following, suspect pursuits, decision making in complex traffic environments,

and lane changing in emergency situation [120], [121], [122], [123], and smart energy management architectures for different types of electric vehicles, grid controls, and P2P energy trading mechanisms [124], [125], [126], [127].

Various RL algorithms have been implemented in HEMSs, including value-based, policy-based, and model-based algorithms. Due to complexity of environmental design in model-based algorithms, value-based and policy-based algorithms are preferred to simulate model-free environments, focusing on the algorithmic evaluation rater than the designing phase. In the literature, the implementation of RL algorithms have been focused on value-based and policy-based DRL algorithms with single-agent environments, using DNN to enhance and accelerate the optimization process [40]. Optimization objectives of RL algorithms include reducing the energy costs, optimizing the energy consumption and flow, optimizing the charging/discharging of ESSs, DSM and DR, and scheduling the charging of EVs [41], [42].

The literature review reveals a significant gap in RL-based energy management systems with multiple prosumers, covering the implementations of P2P energy trading and V2X (V2H, V2L, and V2G) technologies. The majority of the literature are composed of DSM and DR energy management architecture, focusing on the optimization of energy consumption. However, such architectures affect the comfort of prosumers by scheduling the use of appliances and forcing behavioral changes in the daily life activities. To overcome the mentioned issues and fulfill the literature gap, in this paper, optimization of a cluster-based EMS with DDQN agents-based DQN RL algorithm is proposed. Throughout this paper, a cluster-based refers to a relatively small-scale environment with less than 10 prosumers. The proposed cluster-based EMS includes 6 prosumers, 6 ESSs, 3 EVs, and solar panels of each prosumers with equal energy capacity sizes. P2P energy trading and V2X (V2H, V2L, and V2G) technologies are implemented in the proposed EMS to reduce the energy costs and improve the overall SSRs without affecting the comfort of prosumers. To not affect or reduce the comfort of prosumers, conventional DSM and DR energy management architectures are not implemented in the proposed EMS. Instead, energy cost reduction and improvement on the SSRs are achieved via the collaboration of P2P energy trading and V2X technologies with relatively high energy capacities of EVs. In the proposed EMS, prosumers benefit from EVs by meeting the required self energy loads and supplying energy to other prosumers via V2X technologies and P2P energy trading implementations. Sellers profit from the energy transactions with profits and buyers profit from the energy transactions with lower energy prices than the electric grid that has a TOU energy pricing scheme with 3 different time zones and a constant Feed-in-Tariff (FiT) price. The main contributions of this paper are:

1) investigating the effects of the collaboration of P2P energy trading and V2X technologies in a local and cluster-based environment that forms a relatively small-scale network by comparing the proposed EMS with 4 different cases,
2) implementing a cluster-wise scalable EMS to reduce the overall energy costs and improve the SSRs without affecting the comfort of consumers and without applying DSM and DR energy management strategies,
3) proposing a novel hybrid P2P energy trading mechanism that prosumers benefit equally.

The remainder of this paper is organized as follows. The current literature and state-of-the-art on EMS with RL algorithms, V2X technologies, and P2P energy trading concepts are reviewed in Section II. The design of the proposed multi-agent EMS and the MDP formulation are explained in Section III. Multi DDQN agent-based RL algorithm and the training parameters are explained in Section IV. In Section V, the simulations of the proposed system are performed, and the obtained results are analyzed. Finally, the conclusions and future works are discussed in Section VI.

II. THE CURRENT STATE-OF-THE-ART

A. REINFORCEMENT LEARNING BASED ENERGY MANAGEMENT SYSTEMS

RL algorithms have been trending solution approaches to optimization problems with simple to complex environments. Various algorithms, including Q-Learning, DQN, DDPG, PPO, SARSA, Soft Actor-Critic (SAC), Advantage Actor-Critic (A2C), and Twin Delayed Deep Deterministic (TD3), have been implemented in EMSs to meet different requirements. The Q-Learning algorithm is implemented in model-free environments, requiring discrete action space. Sun et al. proposed an implementation of a multi-objective-based Q-learning algorithm to minimize operational costs and pollutant emissions, accelerating the computational process via eligibility trace theory [43]. Dayani et al. implemented a real-time and fuzzy controller based Q-learning algorithm to minimize operational costs and balance energy consumption fluctuations, predicting the future energy consumption [44]. Perera et al. developed a Python program to maximize the utilization of RES and minimize the energy drawn from the grid, forecasting RES energy generation through Artificial Neural Network (ANN) [45]. Chen et al. proposed a preference-based multi-objective RL model to minimize operational costs and optimize appliance usage schedules, managing DR and achieved 8.44% energy cost reduction [46]. Hu et al. approached the energy management problems distinctly, proposing a Q-Learning method to optimize the power distribution scheme and, therefore, maximize social welfare [47]. Contrary to single-agent implementations, Lai et al. proposed a multi-agent Q-Learning algorithm to optimize appliance usage schedules, assigning individual agents to each appliance. Lai et al. compared outcomes of the proposed method with 4 seasons and reduced energy costs by 9.68% in average [48]. Benjamin et al. applied a Q-Learning algorithm to shift 8 appliances usage from peak hours to off-peak hours, thus reducing operational

costs by approximately 38% [49]. Similarly, Rostmnezhad et al. proposed an implementation of a Q-Learning algorithm to optimize the charging/discharging process of Thermal Energy Storage System (TESS) and Battery Energy Storage System (BESS), reducing operational costs by approximately 42% [50]. Moreover, Shouryadhar et al. proposed an enhanced SARSA algorithm to maximize utilization of RES and minimize fossil fuel energy consumption [53]. Xu and et al. proposed a multi-agent Q-Learning HEMS, scheduling appliances and EVs to minimize energy costs approximately by 44.53% for short-term period [85]. An illustration of the implementation of conventional HEMS with shiftable, non-shiftable, and controllable loads can be seen in Fig. 1.

Similar to the Q-Learning algorithm, the DQN algorithm is implemented in model-free environments with discrete action spaces, estimating the optimal Q-values through DNN instead of Q-tables. Hau et al. implemented energy trading concept to a MG using DQN algorithm to minimize operational costs and risks, considering contingency reserves [51]. Xu et al. proposed a system model with DQN algorithm to reduce operational costs by approximately 34%, implementing internal pricing mechanism for residential users [52]. Hong and Lee reduced operational costs around 19% using DQN algorithm and short-horizon forecasts to satisfy energy demands of residential users during energy shortages [54].

DDPG, PPO, SAC, and A2C algorithms are implemented in model-free environments with discrete or continuous action spaces. Wang et al. designed Stackelberg Game theory based DDPG algorithm to maximize energy selling revenue and minimize energy purchasing [55]. Liu et al. approached energy management problems in MG through physical side and proposed DDPG algorithm to optimize AC power flow in stochastic dynamics [56]. Li et al. proposed PPO algorithm to optimize appliance usage schedules, therefore, to minimize operational costs [57]. Weiss et al. proposed of EV implementation into Smart Home (SH) using PPO algorithm to minimize energy costs while satisfying system's constraints [58]. Addition to EV implementations, Tchir et al. implemented V2X technologies with PPO algorithm to minimize operational costs and ensure continuous energy supply [59]. To optimize energy management of multiple residential users, Lee et al. proposed A2C based Federated Reinforcement Learning (FRL) algorithm to be used in multi-agent environment [60]. Moreover, Kahraman et al. studied a comprehensive algorithm comparison using DQN, DDPG and TD3, achieving similar outcomes with different algorithms for short-term (3 days) period [61].

B. VEHICLE-TO-X TECHNOLOGIES

The rapid development of vehicular technologies has facilitated EVs to participate in MG operations. Advancements in power electronic applications and bi-directional EV Charger technologies have enabled the emergence of various V2X technologies, including V2H, V2L, and V2G technologies.

The V2X (V2H, V2L, and V2G) technologies deliver various features to residential users and the electric grid. V2H technologies provide energy during the energy outage of the electric grid and consequently provide uninterrupted energy experiences and support the energy production of RESs. On the other hand, V2L technologies provide energy directly to the interconnected loads through bi-directional chargers, therefore providing uninterrupted energy experiences with sensitive loads requiring continuity. In addition to internal usage, V2G technologies support and balance the electric grid through energy-selling processes [62]. An illustration of the V2X (V2H, V2L, V2G) applications can be seen in Fig. 2.

The application of V2X technologies has several potential advantages to consumers and producers. In V2G technologies, using the battery of EVs as Distributed Energy Resource (DER)s increases the stability of the electric grid by balancing the voltages in the electric grid. Moreover, V2H technologies reduce the energy costs of SHs by providing cheaper energy during peak hours. Implementations of V2H and V2L technologies could protect sensitive electronic devices during energy outages in the electric grid by providing uninterrupted energy [63]. Additionally, V2X technologies reduce pollutant emissions by charging during the daytime and discharging during the evening time and shorten the Return on Investment (ROI) periods of EVs [64]. Besides various advantages, the application of V2X technologies has technical, economic, and social challenges, including accelerating the degradation of EVs' batteries, lack of EV compatibility with V2X technologies, expensive production of compatible chargers, regulation issues, extended charging times and range anxiety of EV owners [64].

Although the application interest of the research includes solely the energy distribution applications (V2G, V2H, and V2L), in addition to energy distribution applications, different types of V2X technologies, such as Vehicle-to-Vehicle (V2V), Vehicle-to-Infrastructure (V2I), and Vehicle-to-Pedestrian (V2P) are investigated intensely in the literature recently. The internet connectivity and wireless communication technologies enable the concept of Internet of Vehicles (IoV) with a wide range of V2X applications. In general, IoV concept includes three virtual layers, namely cloud, infrastructure, and vehicular layers [142]. On contrary, the fundamental components of IoV environment consists of on board units for wireless communication, road side units to connect all the vehicles around the neighborhood, base stations to enhance the overall performance of wireless communication, and trust authority server that secure the connection and functionality of connected vehicles [143]. Besides the energy applications of V2V technologies, EVs could form a vehicular environment (IoV) with connected vehicles through wireless communication methods. In the vehicular environment, vehicles communicate to each other [129] to perform some autonomous tasks, including lateral and longitudinal vehicle following [130], lane changing in noisy environment as tunnel [131], and coordinated brake control for collision avoidance [132]. On the other hand, V2I technologies involve the information exchange

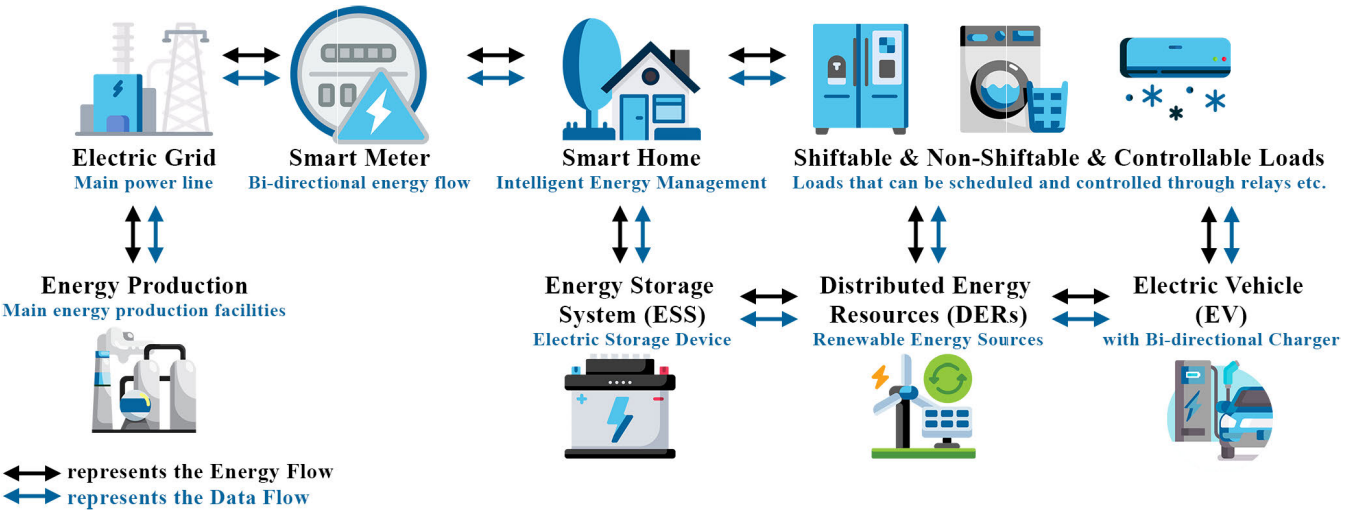

FIGURE 1. The implementation of Reinforcement Learning (RL) algorithms in conventional Home Energy Management Systems (HEMSs) with shiftable, non-shiftable, and controllable loads.

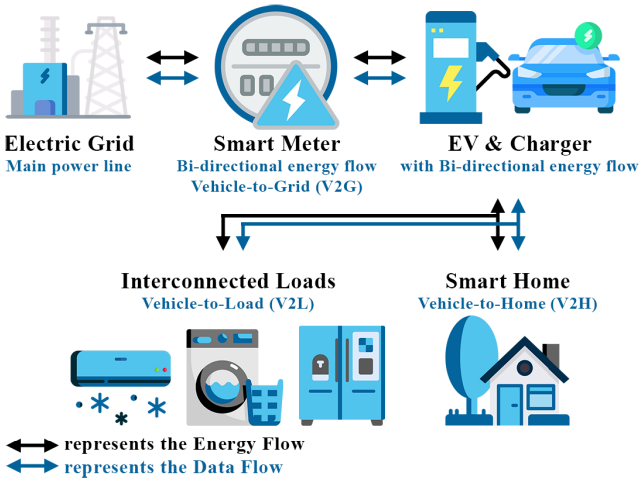

FIGURE 2. The application of Vehicle-to-Grid (V2G), Vehicle-to-Home (V2H), and Vehicle-to-Load (V2L) technologies in Smart Homes (SHs) and Microgrids (MGs).

between vehicles and roadside infrastructure. The roadside infrastructures communicate with vehicles include traffic lights, road signs, and centralized traffic management systems. Vehicle platooning concept have remarkable significance to achieve road safety and high efficiency [133]. The application of V2I technologies include safety and mobility applications, including collision avoidance, collision detection, traffic light detection, toll and fine collection, and smart parking [134], [135], [136], [137]. Similar to V2I, V2P applications are of great importance for efficient traffic management and pedestrian safety. Efficient traffic management could be achieved via V2I applications and waiting times of vehicles and pedestrians could be reduced mutually by implementing cooperative management structures in strategic places as intersections [138], [139]. Pedestrian safety could be achieved via V2I applications by collision avoidance systems that based on path prediction of pedestrians [140], [141].

V2X technologies (V2H, V2L, and V2G) have a limited number of HEMS applications in the current literature. Rule-based energy management approaches of V2X technology integration into HEMSs have constituted the vast majority of the literature. Roche et al. proposed a rule-based energy management approach, including V2H technology, to improve the resilience of HEMSs during energy outages, considering the degradation of batteries in EVs [65]. Rehman et al. approached V2G technologies on the technical side and developed a MATLAB model to simulate V2G technology in terms of load and frequency management in Smart Grid (SG)s [66]. Nowadays, charging costs of EVs have become an important issue. To overcome the stated issue and minimize the charging cost of EVs, Turker et al. proposed a novel V2G algorithm called Optimal Logical Control (OLC) [67]. Hashim et al. implemented a V2G scheduling algorithm in a commercial and residential area to reduce grid load variance and stabilize the electric grid operations by scheduling the charging/discharging process of grid-connected EVs [68]. Hemmati et al. implemented V2H and V2G technologies to improve the utilization and uncertainties of RESs and minimize the daily energy cost of residential users, [69]. In terms of HEMSs cost optimization, Mohamed et al. and Einolander et al. proposed the integration of V2G and V2H technologies into SGs to optimize the energy costs, using different optimization techniques [70], [71].

C. PEER-TO-PEER ENERGY TRADING CONCEPT

Due to the increase in energy consumption, remarkable proliferation in DERs, and facilitation of access to RESs, the P2P energy trading concept has become the prior interest of energy trading applications. The typical P2P

energy trading concept involves the sharing of energy resources among the participants, considered prosumers, in an energy network [72]. P2P energy trading concepts require an energy market to regulate the process, control the dynamics of prosumers, and monitor energy transactions to ensure reliable and efficient energy trading. The design of energy market architectures includes community-based (distributed), centralized, and decentralized energy market architectures [73]. The implementation of P2P energy trading concepts provides several advantages, including peak-shaving, improving the electric grid stability, minimizing operational costs and equipment installations, and increasing the reliability of the electric grid during energy outages [72]. However, besides various advantages, the implementation of P2P energy trading concepts has several challenges, including real-time data acquisition, communication issues, uncertainties of prosumers' dynamics, balancing the selling and purchasing energy amounts, securing transactions, and achieving the optimal pricing mechanism [72], [72], [73]. An illustration of the implementation of P2P energy trading can be seen in Fig. 3.

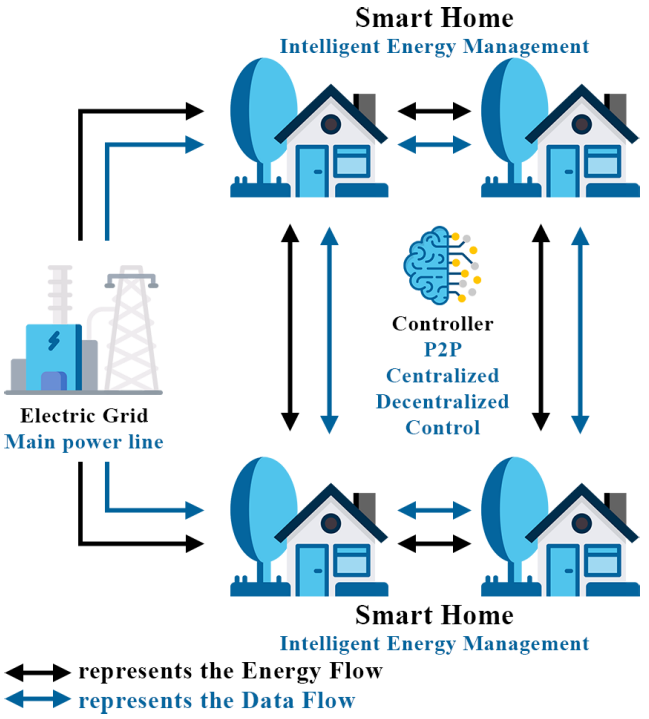

FIGURE 3. The implementation of Peer-to-Peer (P2P) Energy Trading concept with Multiple Prosumers (Consumers & Producers) in a Microgrid (MG) with Smart Homes (SHs).

In the literature, the vast majority of the research on P2P energy trading concepts has focused on developing pricing mechanisms, block-chain applications in energy transactions, and the implementation of P2P energy trading in HEMSs. Developing pricing mechanisms for P2P energy trading concepts has significant importance in achieving optimal energy transactions among prosumers. Anoh et al. proposed a game-theoretic approach based on Stackelberg to achieve an optimal pricing mechanism and maximize consumers' and producers' benefits [74]. Meinke et al. proposed an internal pricing mechanism considering the supply-demand ratio to maximize the economic benefits of prosumers [75]. Trivedi et al. proposed a comprehensive review of optimal pricing mechanisms using cooperative and non-cooperative game-theoretic approaches, including Nash Equilibrium, Nash Bargain, Stackelberg, and Shapley Value [76]. Schiera et al. investigated the different P2P energy trading market strategies and pricing mechanisms to evaluate the different effects of proposed P2P energy trading concepts [77]. To optimize the energy transactions during P2P, ensure data reliability, and maximize prosumers profits, Sarapan et al. proposed a Block-chain-based methodology, considering on-grid and off-grid situations with Islanded MG architecture [78].

Moreover, Zhu et al. implemented a P2P energy trading concept based on a community-based energy market and Lyapunov-based equations with a double auction mechanism to reduce energy costs and maximize the utilization of RESs [79]. Incentives have remarkable importance in developing pricing mechanisms of P2P energy trading concepts. To address the importance of an incentive-based pricing mechanism, Li et al. and Pankiraj et al. proposed an incentive-based pricing mechanism using linear programming-based optimization and rule-based optimization to maximize consumers' and producers' benefits equally [80], [81]. Paudel et al. proposed a novel Hierarchical P2P energy trading concept in MG, including multiple prosumers to improve the efficiency of the energy trading process and reduce energy costs [82]. To realize P2P energy trading with real-time data and equipment, Hayes et al. constructed an open-source laboratory to fulfill the literature gap in hardware and software implementation [83].

The application of RL algorithms in P2P energy trading have relatively insufficient researches in the literature. Various RL algorithms have been applied to P2P energy trading applications to optimize energy trading process and pricing mechanisms. Similar applications with slight differences have been proposed in the literature. Zhou et al. proposed a novel pricing mechanism based on demand and supply ratio to actively encourage prosumers to participate in P2P energy trading. Zhou et al. investigated the proposed method with different case studies, considering the number of smart users in the environment, and reduced energy costs by 25% with 2 smart users involved [86]. Nunna et al. combined multiple (8) MGs as a distribution network, and investigated the effect of P2P energy trading in such network, forecasting the power mismatches across the MGs. Nunna et al. achieved approximately 60$/month profits in maximum [87]. Chen et al. investigated P2P energy trading concept within an environment that include the combination of residential, commercial, and industrial multi-energy MGs with energy conversion among different energy forms.

Chen et al. reduced average hourly costs by 18%, 27%, and 23% in residential, commercial, and industrial microgrids, respectively, using multi-agent TD3 algorithm [88]. Sadeghi and Erol-Kantarci investigated the effects of P2P energy trading with RL algorithm in terms of cost minimization with respect to the number of MGs involved. Sadeghi and Erol-Kantarci concluded that increase in the number of MGs in the environment reduces the average energy costs significantly [89]. Ye et al. built a large-scale model that includes 300 residents, V2X technologies, appliance scheduling, and proposing a novel P2P energy trading mechanism [90]. Similar to the previous paper, Sanayha et al. proposed a model-based multi-agent A3C3 algorithm for a cluster-wise MG with 300 residents. Sanayha et al. clustered the residents with k-means ($k = [2, 10]$) clustering algorithm according to their energy trading behaviors to reduce energy costs by 17% in total [84]. Wu et al. proposed a novel improve mid-market rate P2P energy trading mechanism in a multi-agent environment (multiple MGs) to increase P2P energy trading profits by 25%, implementing DR strategy to schedule appliance usage while maximizing the comforts of prosumers [91]. A similar application example of the proposed multi-agent environment can be seen in Fig. 4.

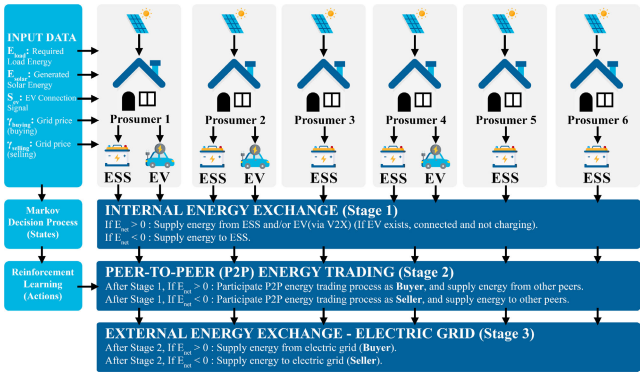

FIGURE 4. Application example of the proposed multi-agent environment with p2p energy trading and V2X implementations, using reinforcement learning algorithm.

III. SYSTEM MODELS AND PROBLEM FORMULATION

Typically, HEMSs include an electric grid to support prosumers and RESs during the limited energy production periods, bi-directional smart meters to monitor energy consumption and production, RESs, battery-based ESSs to store excessive energy production, EVs, controllable, shiftable and non-shiftable loads to apply DR and DSM energy management strategies (see Fig. 1).

The proposed EMS consists of 6 physically and virtually connected homes with energy and information exchange capabilities in decentralized manners within the cluster. Individually, homes, considered as prosumers, in the cluster include battery-based ESS, solar panel based RESs, implementation of V2X (V2H, V2L, and V2G) technologies through EVs with bi-directional chargers, implementation of the P2P energy trading, and system constraints to prevent energy and information exchange violations. Additionally, an electric grid and smart meter are included to improve the reliability of the overall system during energy shortages and monitor the energy consumption and production of prosumers. The illustration and block diagram of the proposed EMS can be seen in Fig. 5. To increase the comfort of prosumers, the proposed EMS does not include DR and DSM energy management strategies. Therefore, the proposed EMS investigates the effects of V2X and P2P implementations on reducing energy costs and improving the overall SSR without affecting the comfort of prosumers.

A. ENERGY MANAGEMENT SYSTEM MODEL

1) ELECTRIC GRID (UTILITY)

To supply energy to prosumers during insufficient energy generation in DERs and insufficient energy transaction in P2P energy trading, an electric grid is implemented. The electric grid is considered as a backup energy source to prevent energy failure situations and maximize satisfaction rates of prosumers. TOU energy prices are adopted as the pricing scheme and the pricing policies can be seen in Fig. 6. Energy buying costs are expressed as follows [94], [95]:

$$C_t^n = E_{grid,t}^n \gamma_t^{buying} \tag{1}$$

where, C_t^n represents the energy cost of the nth prosumer. $E_{grid,t}^n$ and γ_t^{buying} are bought energy from the grid of nth prosumer in kWh and energy buying price with TOU pricing scheme, respectively. Prosumers can sell excessive energy back to the grid. Energy selling profits are expressed as follows [94], [95]:

$$Pr_t^n = E_{selling,t}^n \gamma_t^{selling} \tag{2}$$

where, Pr_t^n represents the profit of the n the prosumer. $E_{selling,t}^n$ and $\gamma_t^{selling}$ are sold energy to the grid in kWh and energy selling price with TOU pricing scheme [84], respectively. The primary objectives of the electric grid are to supply energy to the prosumers during insufficient energy generation in DERs and insufficient energy transactions in P2P energy trading. The second objective of the electric grid is to charge the EVs, if required.

2) ENERGY STORAGE SYSTEM (ESS)

ESSs store the excessive energy generated in RESs and supply energy to the prosumer during energy requirements. The charging and discharging dynamics of ESSs are expressed as follows [46], [61], [94], [95]:

$$E_{ESS,t+1}^n = \begin{cases} E_{ESS,t}^n + \eta_{ESS,charge}^n P_{ESS,t}^n \Delta_t, & P_{ESS,t}^n > 0 \\ E_{ESS,t}^n + \dfrac{P_{ESS,t}^n \Delta_t}{\eta_{ESS,discharge}^n}, & P_{ESS,t}^n \leq 0 \end{cases} \tag{3}$$

where, $E_{ESS,t+1}^n$ and $E_{ESS,t}^n$ represent the current and previous energy in the nth ESS in kWh, respectively. $P_{ESS,t}^n$ is the

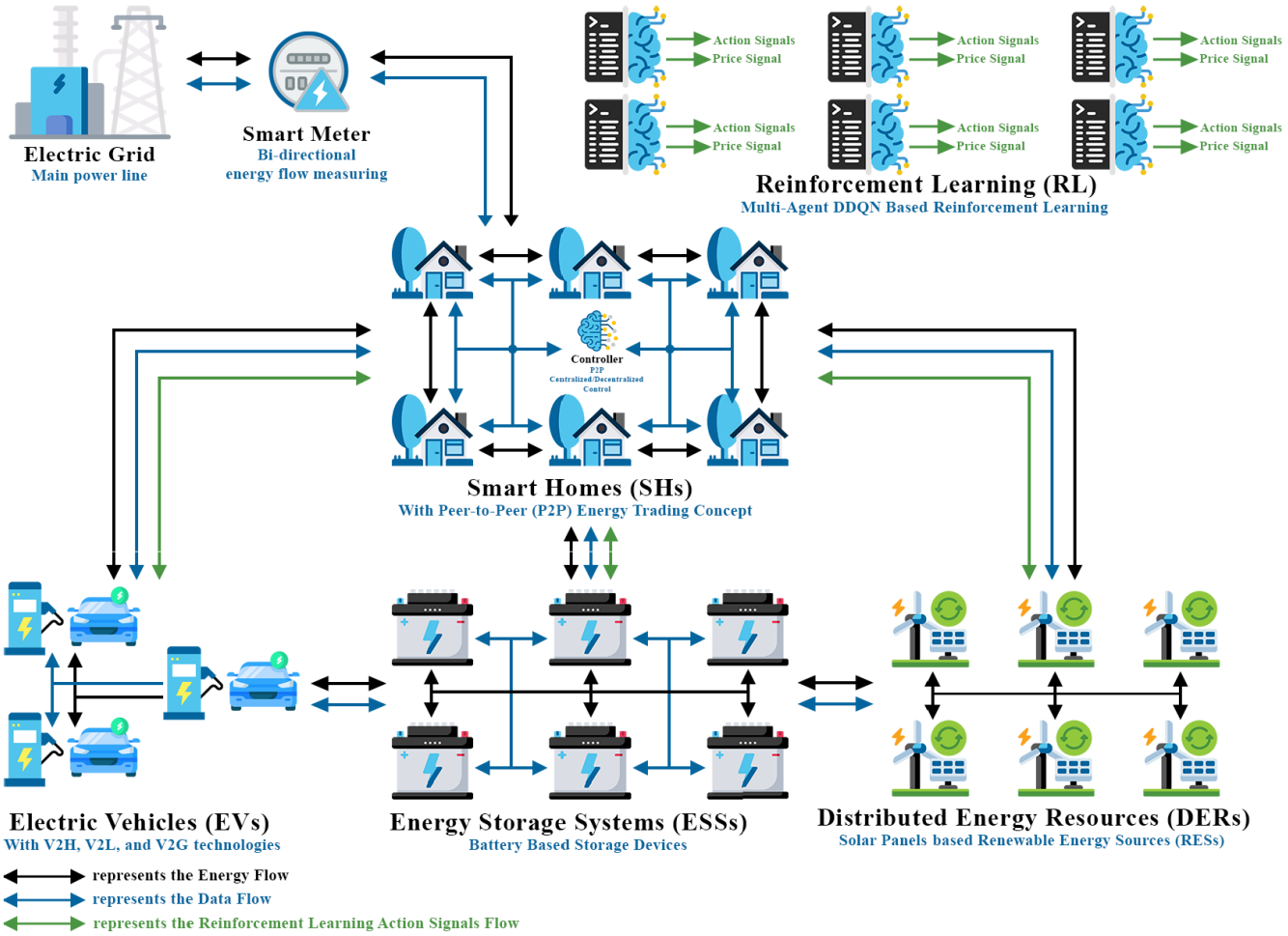

FIGURE 5. The proposed cluster-based Energy Management System (EMS), including multiple (6) prosumers, battery-based ESSs, EVs with V2X technologies, and P2P energy trading mechanism.

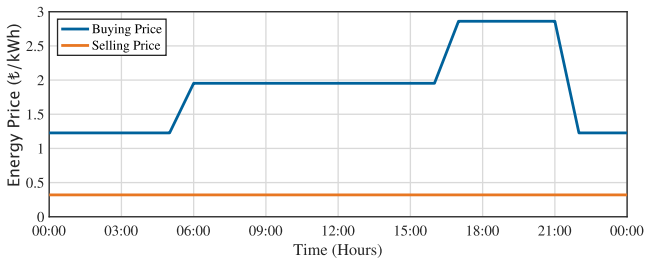

FIGURE 6. Time of Use (TOU) energy pricing scheme, consisting of buying (3 different time zone) and selling prices (a constant feed-in-tariff), respectively.

input power of the nth ESS in kW. Δ_t represents the time step at time t in $T_{simulation}$. $\eta^n_{ESS,charge}$ and $\eta^n_{ESS,discharge}$ represent the charging and discharging efficiency of the nth ESS, respectively. The input power, $P^n_{ESS,t}$, determines the operation mode of ESS: charging mode or discharging mode. The State of Charge (SoC) of each ESS, $SoC^n_{ESS,t+1}$, is expressed as follows [46], [61], [94], [95]:

$$SoC^n_{ESS,t+1} = \frac{E^n_{ESS,t+1}}{E^n_{ESS,capacity}} \tag{4}$$

where $SoC^n_{ESS,t+1}$ represents the current SoC of the nth ESS. $E^n_{ESS,t+1}$ and $E^n_{ESS,capacity}$ are the current energy and total energy capacity of the nth ESS in kWh, respectively. The primary objective of ESSs is to store excessive energy generated in RESs and supply energy to the owner to meet required energy loads. The secondary objective is to supply energy to the other prosumers via P2P energy trading.

3) RENEWABLE ENERGY SOURCE (RES)

To simulate the dynamics of DERs, solar panel based RESs are implemented in the proposed EMS. Designed DERs are considered as a centralized unit, and it is assumed that prosumers benefit from the generated energy equally. Therefore, the amount of generated solar energy is divided 6 equal parts for each prosumer. The dynamics of the implemented DERs are uncertain, and generated energy is expressed as follows:

$$E_{solar,t} = P_{solar,t}\Delta_t \tag{5a}$$

$$P_{solar,min} \leq P_{solar,t} \leq P_{solar,max} \tag{5b}$$

$$E_{solar,min} \leq E_{solar,t} \leq E_{solar,max} \tag{5c}$$

$$E^n_{solar,t} = \frac{E_{solar,t}}{6} \tag{5d}$$

where, $P_{solar,min}$ and $E_{solar,min}$ are the minimum power and energy generation capacity of DERs. Similarly, $P_{solar,max}$ and $E_{solar,max}$ are the maximum power and energy generation capacity of DERs. $P_{solar,t}$ and $E_{solar,t}$ represent the power and energy generation at time step t. $E^n_{solar,t}$ represents the amount of generated solar energy per prosumer. The primary objective of DERs is to supply energy to the owner to meet the required energy loads. The remaining energy can be stored in ESSs for future energy requirements and supplied to the cluster via P2P energy trading.

4) ELECTRIC VEHICLE (EV) AND VEHICLE-TO-X IMPLEMENTATION

Similar to ESSs, EVs are implemented in the proposed EMS to simulate the dynamics of EVs. Additionally, EVs have V2X (V2H, V2L, and V2G) capabilities to supply energy to home during energy insufficiency (V2H), directly to the connected loads (V2L), and the electric grid (V2G) to improve the operations of the electric grid. The charging and discharging dynamics of EVs are expressed as follows [69], [95]:

$$E^n_{EV,t+1} = \begin{cases} E^n_{EV,t} + \eta^n_{EV,charge} P^n_{EV,t} \Delta_t, & P^n_{EV,t} > 0 \\ E^n_{EV,t} + \dfrac{P^n_{EV,t} \Delta_t}{\eta^n_{EV,discharge}}, & P^n_{EV,t} \leq 0 \end{cases} \tag{6}$$

where $E^n_{EV,t}$ and $E^n_{EV,t+1}$ represent the current and previous energy in the nth EV in kWh, respectively. $P^n_{EV,t}$ is the input power of the nth EV in kW. $\eta^n_{EV,charge}$ and $\eta^n_{EV,discharge}$ represent the charging and discharging efficiency of the nth EV, respectively. The input power $P^n_{EV,t}$ determines the operation mode of EV: charging mode or discharging mode. Discharging mode of EVs are used in V2X operations. The SoC of EVs, $SoC^n_{EV,t+1}$, are expressed as follows [69], [95]:

$$SoC^n_{EV,t+1} = \frac{E^n_{EV,t+1}}{E^n_{EV,capacity}} \tag{7}$$

where, $SoC^n_{EV,t+1}$ represents the current SoC of the nth EV. $E^n_{EV,t+1}$ and $E^n_{EV,capacity}$ are the current energy and total energy capacity of the nth EV in kWh, respectively. In addition to charging operations, EVs supply energy to the owner and other prosumers via V2X and P2P implementations. V2H, V2L, and V2G equations are expressed as follows:

$$E^n_{V2H,t} = E^n_{EVavailable,t} C^n_{V2H,t} \tag{8a}$$

$$E^n_{V2L,t} = E^n_{EVavailable,t} C^n_{V2L,t} \tag{8b}$$

$$E^n_{V2G,t} = E^n_{EVavailable,t} C^n_{V2G,t} \tag{8c}$$

$$E^n_{V2X,t} = E^n_{V2H,t} + E^n_{V2L,t} + E^n_{V2G,t} \tag{8d}$$

where, $E^n_{V2H,t}$, $E^n_{V2L,t}$, and $E^n_{V2G,t}$ represent the computed energies for V2H, V2L, and V2G operations in kWh, respectively. $E^n_{EVavailable,t}$ is the available energy capacity of the nth EV in kWh and is expressed as follows:

$$E^n_{EVavailable,t} = E^n_{EV,capacity,t} SoC^n_{EV,t} \tag{9}$$

$C^n_{V2H,t}$, $C^n_{V2L,t}$, and $C^n_{V2G,t}$ represent the multiplication coefficients of V2H, V2L, and V2G operations that determine the amount of available energy to be used in V2X operations. $E^n_{V2X,t}$ is the total energy amount for V2X operations in kWh at time step t. Subsequently, the ultimate SoCs of EVs are expressed as follows:

$$SoC^n_{EV,t+1} = \frac{E^n_{EV,t+1}}{E^n_{EV,capacity}} - \frac{E^n_{V2X,t}}{E^n_{EV,capacity}} \tag{10}$$

The connection status of EVs to the cluster could be connected (Binary 1) and not connected (Binary 0). EVs consume energy for transportation purposes while not connected to the cluster. When EVs arrive to home and are connected to the cluster, SoCs of EVs are calculated to simulate the dynamics of driving cycles. The calculation of SoC on the arrival is expressed as follows:

$$SoC^n_{EV,t+1} = \frac{E^n_{EV,t-1}}{E^n_{EV,capacity}} - \frac{E^n_{driving,t}}{E^n_{EV,capacity}} \tag{11}$$

where, $E^n_{driving,t}$ represents the energy consumption of the nth EV during the driving cycles. $E^n_{EV,t-1}$ represents the available energy capacity of the nth EV at time $t-1$, when EV leaves home.

5) PEER-TO-PEER (P2P) ENERGY TRADING IMPLEMENTATION

A novel incentive-based hybrid P2P energy trading and pricing mechanism is proposed to be used in the proposed EMS and to simulate the dynamics of P2P energy trading. The proposed P2P energy trading and pricing mechanism allows prosumers to participate in the energy exchange process with equal benefits. The energy trading and pricing flow of the proposed P2P energy trading mechanism is as follows:

1) Required buying energy of the buyers are calculated and reported to the mechanism.
2) Desired selling energy of the sellers are calculated with action signals of RL algorithm and reported to the mechanism.
3) Energy prices of the sellers are calculated with action signals of RL algorithm and reported to the mechanism. The maximum energy price of sellers are limited to the energy price of the electric grid. Therefore, the buyers are always encouraged to buy energy from other peers.
4) All sellers are sorted from lowest to highest energy prices. Seller with the lowest energy price get the priority to sell its desired amount of energy.
5) The amount of energy that seller desire to sell is distributed equally among the all buyers.
6) Energy trading continues until there is no more energy left to sell or buy in the mechanism.

7) After energy trading is completed, energy profits of the sellers and energy costs of the buyers are calculated with energy trading transactions.

Consequently, all buyers benefit equally from the amount of energy sold in the mechanism. Prioritizing the seller rank from lowest to highest energy prices and determining the amount of energy to be sold in the proposed P2P energy trading mechanism require non-cooperative optimization approach inside the cluster to maximize the profits of the sellers and cooperative optimization approach outside the cluster to minimize the energy bought from the electric grid. Algorithm of the proposed P2P energy trading can be seen in the Algorithm 1. Additionally, the simplified process flow of the proposed P2P energy trading mechanism can be seen in Fig. 7.

Algorithm 1 Peer-to-Peer (P2P) Energy Trading

Input: $E^n_{selling,t}$, $E^n_{buying,t}$, $\gamma^{n,P2P}_t$
Output: Energy trading transactions

if $\sum E^n_{selling,t} > 0 and \sum E^n_{buying,t} > 0$ **then**
  Sort the sellers from the lowest $\gamma^{n,P2P}_t$ to the highest $\gamma^{n,P2P}_t$
  for seller **in** sellers **do**
    while $E^{seller}_{selling,t} > 0$ **do**
      $E^{seller}_{selling*,t} = \frac{E^{seller}_{selling,t}}{count_{buyers}}$
      for buyer **in** buyers **do**
        $E^{seller}_{sold,t} = min(E^{seller}_{selling*,t}, E^{buyer}_{buying,t})$
        $E^{buyer}_{bought,t} = E^{seller}_{sold,t}$
        $E^{buyer}_{buying,t} = E^{buyer}_{buying,t} - E^{buyer}_{bought,t}$
        Calculate the energy cost and profit:
        $C^{buyer}_{p2p,t} = E^{buyer}_{bought,t} \gamma^{seller,P2P}_t$
        $Pr^{seller}_{p2p,t} = E^{seller}_{sold,t} \gamma^{seller,P2P}_t$
      end for
      $E^{seller}_{selling,t} = E^{seller}_{selling,t} - E^{seller}_{sold,t}$
    end while
  end for
  Compute the Energy trading transactions
end if

6) CLUSTER MODEL AND SYSTEM CONSTRAINTS

The proposed EMS is composed of 6 integrated homes to build a cluster-based EMS environment. All individual homes are composed of ESS, EV (if applicable) with V2X technologies, DER, and the implementation of P2P energy trading. To eliminate the design complexities, the process flows of each home in the proposed EMS are divided into 3 subsequent steps, namely Before Energy Trading step, P2P Energy Trading step, and After Energy Trading step as seen in Fig. 8.

Initially, in the Before Energy Trading step, the net energy of prosumers are calculated with required energy loads and generated solar energy. Insufficient energies are met through

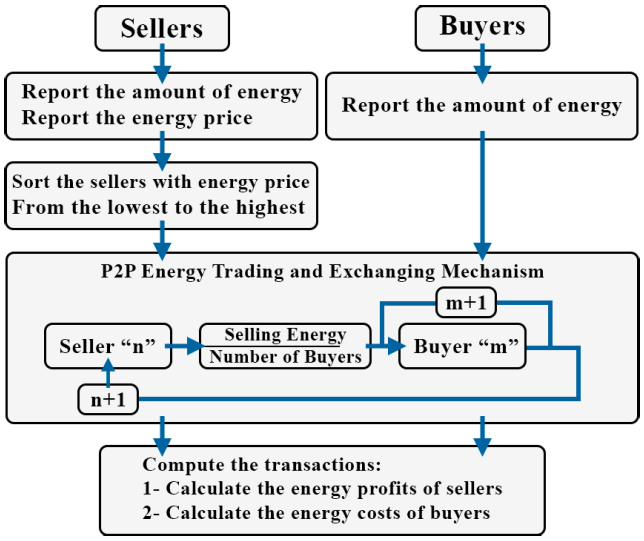

FIGURE 7. The simplified process flow of the proposed hybrid P2P energy trading and exchanging mechanism.

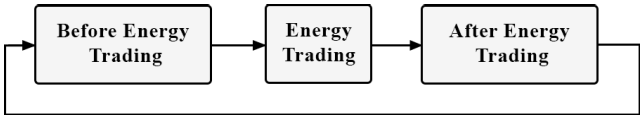

FIGURE 8. The process flow of each home in the proposed EMS: Before Energy Trading (BET), Energy Trading (ET), and After Energy Trading (AET), respectively.

stored energy in ESSs and EVs. It is assumed that EVs are charged, if required, during off-peak hours between 03:00 AM and 06:00 AM. In case of energy surplus, ESSs are charged with the excessive energy.

$$E^n_{net,t} = E^n_{load,t} - E^n_{solar,t} \tag{12}$$

$$E^n_{net,t} = \begin{cases} E^n_{net,t} - (E^n_{ess,t} + E^n_{v2x,t}) & E^n_{net,t} > 0 \\ E^n_{net,t} - E^n_{ess,t}, & E^n_{net,t} \leq 0 \end{cases} \tag{13}$$

where, $E^n_{net,t}$ represents the net energy of the nth prosumer. If the net energy $E^n_{net,t}$ is greater than 0, the net energy is meet from the available energies in ESSs and EVs. In case of insufficient energy supply from ESSs and EVs, remaining energy are met from other peers through P2P energy trading and from the electric grid. Contrarily, if the net energy $E^n_{net,t}$ is less than or equal to 0, ESSs are charged with the net energy. In case of excessive energy after ESS charging process, remaining energy are sold to other peers through P2P energy trading and to the electric grid.

In the P2P Energy Trading step, amounts of energy to be bought and sold are calculated and energy exchanges are computed via proposed P2P energy trading mechanism. The amounts of energy to be sold in P2P energy trading are calculated as follows:

$$E^n_{ess,p2p,t} = E^n_{ess,available,t} C^n_{ess,t} AS^n_{ess,t} \tag{14a}$$

$$E^n_{ev,p2p,t} = E^n_{ev,available,t} C^n_{ev,t} AS^n_{ev,t} \tag{14b}$$

$$E^n_{selling,t} = E^n_{ess,p2p,t} + E^n_{ev,p2p,t} + E^n_{net,t} \tag{14c}$$

where, $E^n_{ess,available,t}$ and $E^n_{ev,available,t}$ are available energy in ESSs and EVs of the nth prosumer, respectively. $C^n_{ess,t}$ and $C^n_{ev,t}$ represent the multiplication coefficients of ESSs and EVs, respectively. $AS^n_{ess,t}$ and $AS^n_{ev,t}$ are binary action signals of ESSs and EVs, respectively. The multiplication coefficients and binary action signals are the optimized action outcomes of the DQN RL algorithm. The energy price of the nth seller is expressed as follows:

$$\gamma^{p2p,selling}_t = \gamma^{buying}_t C^n_{P2P,t} \tag{15}$$

where, $\gamma^{p2p,selling}_t$ represents the calculated P2P energy trading price of the nth seller. γ^{buying}_t is the energy price of the electric grid. $C^n_{P2P,t}$ is the multiplication coefficient of energy price. The multiplication coefficient of energy price is the optimized action outcome of the DQN RL algorithm.

After the P2P energy trading step, in the After Energy Trading step, in case of energy insufficiency, the required energy loads are supplied from the electric grid. In the opposite case, the amounts of excessive energy are supplied to the electric grid. The amount of energy supplied from and to the electric grid are expressed as follows:

$$E^n_{grid,buying,t} = E^n_{net,t} - E^n_{p2p,bought,t} \tag{16a}$$
$$E^n_{grid,selling,t} = E^n_{net,t} - E^n_{p2p,sold,t} \tag{16b}$$

where, $E^n_{p2p,bought,t}$ and $E^n_{p2p,sold,t}$ are the amount of energy bought and sold via P2P energy trading, respectively. $E^n_{grid,buying,t}$ and $E^n_{grid,selling,t}$ represent the amount of energy bought from and sold to the electric grid. Energy costs and profits of the electric grid transactions are expressed as follows.

$$C^n_{grid,t} = E^n_{grid,buying,t}\gamma^{buying}_t \tag{17a}$$
$$Pr^n_{grid,t} = E^n_{grid,selling,t}\gamma^{selling}_t \tag{17b}$$

Therefore, the overall energy costs and profits of each prosumer are expressed as follows:

$$C^n_{total,t} = C^n_{grid,t} + C^n_{p2p,t} + C^n_{ev,charging,t} \tag{18a}$$
$$Pr^n_{total,t} = Pr^n_{grid,t} + Pr^n_{p2p,t} \tag{18b}$$

where, $C^n_{total,t}$, $C^n_{grid,t}$, $C^n_{p2p,t}$, and $C^n_{ev,charging,t}$ are the total energy costs, energy costs of the electric grid, energy costs of P2P energy trading, and energy costs of EV charging, respectively. $Pr^n_{total,t}$, $Pr^n_{grid,t}$, and $Pr^n_{p2p,t}$ are total energy profits, energy profits of the electric grid, and energy profits of P2P energy trading, respectively.

Several system constraints are applied to ensure the stability of the proposed EMS. The ESSs should satisfy the following constraints [46], [61], [94], [95]:

$$0 < E^n_{ess,t} \leq E^{max}_{ess} \tag{19a}$$
$$SoC^{min}_{ess} < SoC^n_{ess,t} \leq SoC^{max}_{ess} \tag{19b}$$
$$P^{min}_{ess} < P^n_{ess,t} \leq P^{max}_{ess} \tag{19c}$$

where, $E^{max}_{ess,t}$ is the maximum energy capacity of ESS of the nth prosumer in kWh. SoC^{min}_{ess} and SoC^{max}_{ess} are minimum and maximum SoC values of the nth prosumer, respectively. P^{min}_{ess} and P^{max}_{ess} minimum and maximum powers of the nth prosumer in kW, respectively. Similar to ESSs, EVs have several constraints to satisfy as follows [69], [95]:

$$0 < E^n_{ev,t} \leq E^{max}_{ev} \tag{20a}$$
$$SoC^{min}_{ev} < SoC^n_{ev,t} \leq SoC^{max}_{ev} \tag{20b}$$
$$0 < P^n_{ev,t} \leq P^{max}_{ev} \tag{20c}$$
$$0 < E^n_{v2x,t} \leq E^{max}_{v2x} \tag{20d}$$

where, E^{max}_{ev} represents the maximum energy capacity of EV of the nth prosumer in kWh. SoC^{min}_{ev} and SoC^{max}_{ev} are minimum and maximum SoC values of the nth prosumer, respectively. P^{max}_{ev} and E^{max}_{v2x} represent the maximum charging power in kW and the maximum V2X discharging energy in kWh, respectively. To prevent energy exchange violations in V2X and P2P energy trading, following constraints should be satisfied:

$$E^n_{buying,t}E^n_{selling,t} = 0 \tag{21a}$$
$$E^n_{ev,charging,t}E^n_{ev,available,t} = 0 \tag{21b}$$

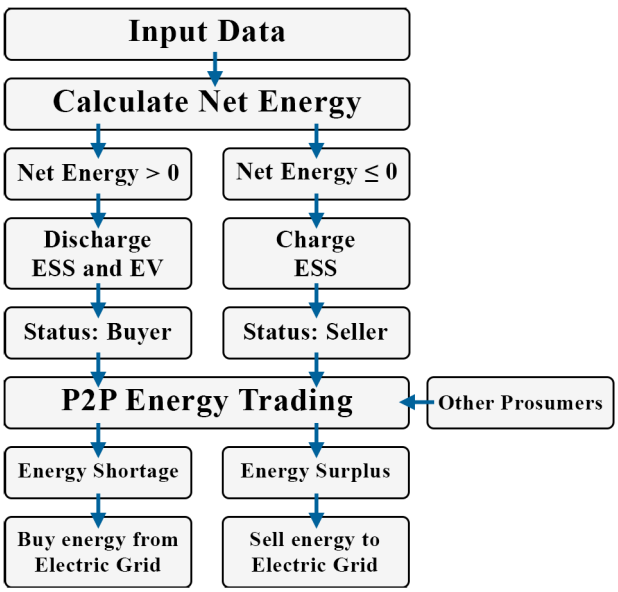

FIGURE 9. The process flow of each home in the proposed EMS and the interaction with other homes.

As seen in the Fig. 9, the functionality of overall environment starts with the net energy calculation. The net energy calculation happens via the produced energy in the solar panels and the required energy by the prosumers. Regarding the amount of net energy, charging/discharging occurs in the ESS devices and EV via V2L and V2H. In this scenario, EVs act in the same way as ESS devices.

Afterwards, prosumers declare the statuses to participate the P2P energy trading as buyer or seller.

1) *Buyer:* Declare the amount of energy to purchase.
2) *Seller:* Declare the amount of energy to sell and the energy price. The amount of energy to sell is the combination of surplus energy ($E^n_{load,t} - E^n_{solar,t} \leq 0$), energy stored in the batteries of ESS and EV.

After the P2P energy trading, prosumers buy energy from the grid, in case of energy shortage. Furthermore, prosumers can sell energy back to the grid, in case of energy surplus.

Similar to ESS, when connected to the charger, EVs can supply energy to the owner and other prosumers via V2L and V2H. A balanced approach that integrates both V2X technologies and P2P energy trading offers the most comprehensive optimization solution. Thus, prosumers can leverage the strengths of each component by using the relatively high capacity batteries of EVs in the P2P energy trading via V2X technologies. Additionally, prosumers with EV can leverage the potential profits with TOU pricing scheme, by selling the stored energy to other prosumers via P2P energy trading during the daytime (higher prices) and storing energy during the night-time (cheaper prices). However, amount of energy that a prosumer can sell via V2X technologies during P2P energy trading is strictly limited to prevent the discharge EVs beyond the certain limits and increase the satisfaction of prosumers. Throughout the project, it is assumed that P2P energy trading mechanism ensure the privacy and security during data exchange between vehicles and prosumers. Also, it is assumed that P2P energy trading mechanism ensure the compatibilities, such as communication protocols, standardized interfaces and reliable infrastructure, among all the components in the environments to achieve a seamless integration between V2X technologies and P2P energy trading.

B. MARKOV DECISION PROCESS FORMULATION

The proposed EMS is formulated as MDP tuples to train and simulate DDQN-based multi-agent environment with DRL algorithm. MDP tuples include state spaces (S^n), action spaces (A^n), state transition probability matrices (P^n), reward functions (R^n), and discount factors (γ^n). A typical MDP tuple structure can be seen in Fig. 10.

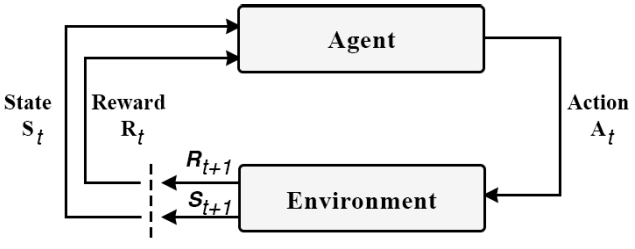

FIGURE 10. Typical Markov Decision Process (MDP) tuple: State, Action, Reward, and Next State.

State spaces (S^n) include the observed information from the environment. Prosumers have identical state space (S^n) design and are as follows:

$$S^n_t = \{E^n_{load,t}, E^n_{solar,t}, SoC^n_{ESS,t-1}, SoC^n_{EV,t-1}, S^n_{EV,t}, \gamma^{buying}_t, \gamma^{selling}_t\} \tag{22}$$

where, $E^n_{load,t} \in [0, \infty)$ and $E^n_{solar,t} \in [0, \infty)$ are required energy loads and generated solar energies, respectively. $SoC^n_{ESS,t-1} \in [0, 1]$ and $SoC^n_{EV,t-1} \in [0, 1]$ represent the previous SoCs of ESSs and EVs at time step $t-1$. $S^n_{EV,t} \subset \{0, 1\}$ are the status signals of EVs, indicating the connection status of EVs with the cluster. γ^{buying}_t and $\gamma^{selling}_t$ are energy buying and selling prices of the grid, respectively.

Action spaces (A^n) include the available set of actions of the observed state (S^n). Prosumers have identical action space (A^n) design and are as follows:

$$A^n_t = \{AS^n_{ESS,t}, C^n_{ESS,t}, AS^n_{EV,t}, C^n_{EV,t}, C^n_{P2P,t}\} \tag{23}$$

where, $AS^n_{ESS,t} \subset \{0, 1\}$ and $AS^n_{EV,t} \subset \{0, 1\}$ are binary indication signals of P2P energy trading participation of ESSs and EVs of each prosumer, respectively. Prosumers decide to supply energy to other peers via P2P with $AS^n_{ESS,t}$ and $AS^n_{EV,t}$ binary signals. $C^n_{ESS,t} \in [0.2, 1]$ and $C^n_{ESS,t} \in [0.2, 1]$ are multiplication coefficients that calculate the amounts of energy to be sold out of the available energy in P2P energy trading. $C^n_{P2P,t} \in [0.6, 0.95]$ represents the price multiplication coefficients that calculate the P2P energy trading prices based on the current electric grid prices. Subsequently, action spaces (A) of each prosumer include 2592 possible discrete action combinations.

State transition probability matrices (P^n) computes the probabilities of the next states ($s^{n'}$) through the current states (s^n) and the executed actions (a^n).

$$P^n_a = Pr(s^n_{t+1} = s' | s^n_t = S, a^n_t = A) \tag{24}$$

Reward functions (R^n) calculate the received immediate rewards through the transition from the state s^n_t to the next state s^n_{t+1} due to the executed action a^n_t. Prosumers have identical reward functions and are expressed as follows:

$$R^n_t = 0.7(\sum_{n=1}^{6} R^n_{cost,t} + R^n_{cost,t}) + 0.3R^n_{grid,t} \tag{25a}$$

$$R^n_{cost,t} = C^{n,normalized}_{total,t} - Pr^{n,normalized}_{total,t} \tag{25b}$$

$$R^n_{grid,t} = \sum_{n=1}^{6} E^{n,normalized}_{grid,t} + E^{n,normalized}_{grid,t} \tag{25c}$$

$$C^{n,normalized}_{total,t} = \frac{C^n_{total,t} - \min C^n_{total,t}}{\max C^n_{total,t} - C^n_{total,t}} \tag{25d}$$

$$Pr^{n,normalized}_{total,t} = \frac{Pr^n_{total,t} - \min Pr^n_{total,t}}{\max Pr^n_{total,t} - Pr^n_{total,t}} \tag{25e}$$

$$E^{n,normalized}_{grid,t} = \frac{E^n_{total,t} - \min E^n_{total,t}}{\max E^n_{total,t} - E^n_{total,t}} \tag{25f}$$

where, $C^{n,normalized}_{total,t}$, $Pr^{n,normalized}_{total,t}$, and $E^{n,normalized}_{total,t}$ are the normalized values of the total energy costs, total energy

profits, and total bought energy from the grid, respectively. Initial part of the reward functions (R^n) computes the penalties of the total energy costs, whereas the last part computes the penalties of the bought energy from the grid. In the reward functions (R^n), 0.7 and 0.3 represent the weight coefficient to emphasize the importance of each penalty part. Consequently, the primary objective of the proposed EMS is to minimize the total energy costs. The secondary objective is to minimize the cumulative bought energy from the grid, encouraging prosumers to participate in P2P energy trading. Discount factors ($\gamma^n = 0.965$) are applied to calculate the expected future reward of each prosumer.

IV. DOUBLE DEEP Q-NETWORK BASED DEEP REINFORCEMENT LEARNING APPROACH

DDQN agent-based DRL algorithm is a model-free, online, and off-policy optimization method. DDQN agents are value-based RL agent that train the critic networks to estimate the discounted expected cumulative long-term rewards. The fundamental objective of RL algorithms is to maximize the expected cumulative long-term reward. Besides maximizing the expected cumulative long-term reward, careful design considerations are required to achieve a stable training process. Stability in the training process refers to a point, where relatively low fluctuations occur in the produced value function estimations [101]. During the training phase, stability measures could be monitored by tracking the mean squared error between the critic and target critic, tracking the behavior of the value function estimates by analyzing the learning curves or terminating the training phase with predefined rules (e.g. episode count and reward threshold). One criteria that assists to achieve stability and boost performance is to balancing exploration and exploitation [102]. Gathering information about the environment through exploration and maximizing the reward with the best action possible are crucial to help DRL algorithms to converge within the desired training ranges. Various RL algorithms employ different strategies to balance exploration and exploitation, such as epsilon-greedy [96], softmax exploration [97], Upper Confidence Bound (UCB) and Thompson sampling [98]. Such strategies aim to systematically trade off between exploration and exploitation based on the agent's current knowledge and the uncertainty in its estimates of action values. To overcome the exploration and exploitation dilemma and balance the learning process, Epsilon-Greedy method is implemented in the learning phase. Epsilon-Greedy method balance the learning process with a probability constant ϵ, by increasing the exploitation probability and decreasing the exploration probability over the episodes [96]. Therefore, Epsilon-Greedy method aims to increase the convergence rate and speed [103]. However, it should be noted that several potential problems could occur during Epsilon-Greedy exploration and exploitation method, such as limited exploration, exploitation-exploration trade-off, greedy selection bias and sensitivity to hyper-parameters. To overcome such potential problems, adjusting the proper ϵ value is of great importance to ensure a stable training. A typical DDQN agent-based DRL, including DNN, can be seen in Fig. 11.

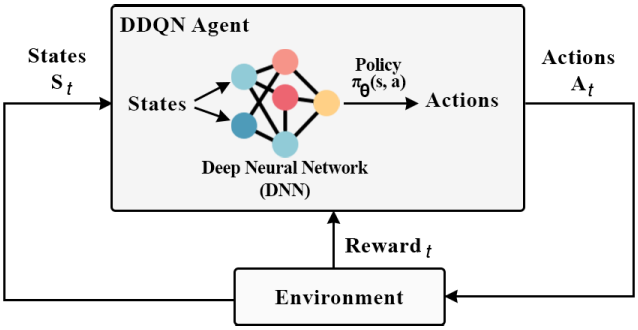

FIGURE 11. DDQN agent-based Deep Reinforcement Learning algorithm flow chart, using deep neural network architecture to estimate the optimal policy.

The expected cumulative long-term rewards of each agent with discount factors applied are as follows [92], [93]:

$$\mathbb{E}^n = \sum_{t=1}^{T_{simulation}} \gamma^n r(s_t^n, a_t^n) \tag{26}$$

Therefore, state-value functions (V^n), based on the expected cumulative long-term rewards by following a policy (π), of each agent are expressed as follows [92], [93]:

$$V^{\pi,n}(s_t^n) = \mathbb{E}^n \left[\sum_{t=1}^{T_{simulation}} \gamma^n r_t^n | s_t^n = S^n, \pi^n \right] \tag{27}$$

The optimum expected state-value functions ($V^{*,n}$) of each agent are expressed as follows [92], [93]:

$$V^{*,n}(s_t^n) = \max V^{\pi,n}(s_t^n) \tag{28}$$

Similar to state-value functions (V^n), action-value functions (Q^n) of each agent are calculated with $\mathbb{E}^n$, considering the executed actions a_t^n and following policies π^n, and are expressed as follows [92], [93]:

$$Q^{\pi,n}(s_t^n, a_t^n) = \mathbb{E}^n \left[\sum_{t=1}^{T_{simulation}} \gamma^n r_t^n | s_t^n = S^n, a_t^n = A^n, \pi^n \right] \tag{29}$$

The optimum action-value functions ($Q^{*,n}$), in the given states s_t^n, actions a_t^n, and policies π^n, of each agent are expressed as follows [92], [93]:

$$Q^{*,n}(s_t^n, a_t^n) = \max Q^{\pi,n}(s_t^n, a_t^n) \tag{30}$$

DDQN agents include two critic networks with two separate action-value functions (Q^n): Critics $Q^n(S^n, A^n; \phi^n)$ and Target Critics $Q_t^n(S^n, A^n; \phi_t^n)$. Critics $Q^n(S^n, A^n; \phi^n)$ store the estimates of the expected cumulative rewards of each

agent. Conversely, Target Critics $Q_t^n(S^n, A^n; \phi_t^n)$ are updated periodically to enhance the overall performance and stability of the optimization process. Update processes of Target Critics $Q_t^n(S^n, A^n; \phi_t^n)$ through Critics $Q^n(S^n, A^n; \phi^n)$ are expressed as follows [92], [93]:

$$Q^{*,n}(s_t^n, a_t^n) = r_t^n + \gamma^n Q_t^n(s_{t+1}^n, \max Q^n(S^n, A^n; \phi^n); \phi_t^n) \tag{31a}$$

$$\phi_t^n = \tau^n \phi^n + (1 - \tau^n)\phi_t^n \tag{31b}$$

where, ϕ^n and ϕ_t^n are the network parameters of critics and target critics, respectively. τ^n represents the smoothing factor of each network. The network losses of each network are calculated as follows [92], [93]:

$$L^n = \frac{1}{2M} \sum_{t=1}^{M} (r_t^n - Q^n(S^n, A^n; \phi^n))^2 \tag{32}$$

where, L^n and M represent the network losses of each network and the number of sampled experiences from the stored experiences, respectively. Loss functions are of great importance to improve the overall performance and stability of the learning phase. Adam optimizer is one of the most commonly used optimization method during the training of DNNs [104]. Therefore, Adam optimizer is used to train the critic approximator that updates the parameters of online and target networks. The convergence analysis of DRL algorithms refer to finding of an optimal policy over time and directly related to losses of the online and target networks. The convergence criterion of DRL algorithm is expressed as follows:

$$\lim_{t \to \inf} Q^{\pi,n}(s_t^n, a_t^n) = Q^{*,n}(s_t^n, a_t^n) \tag{33}$$

where, the Q value of target network equals to the optimum Q value that maximizes the reward function when following the policy π. Environment experiences are stored in Prioritized Experience Replay (PER) buffers to improve and accelerate the learning process. Epsilon-greedy methods and update processes, using decay rates, of each agent are expressed as follows [92], [93]:

$$\epsilon_t^n = \epsilon_t^n (1 - \alpha_{decay,t}^n) \tag{34}$$

where, ϵ_t^n and $\alpha_{decay,t}^n$ represent the probability thresholds of epsilon-greedy method and decay rates of each agent, respectively. Multi DDQN agent based DRL algorithm is explained in the Algorithm 2.

In the proposed EMS environment, all agents share identical agent and training options and are described in Table 1.

Estimation of the critic networks, $Q_t^n(S^n, A^n; \phi_t^n)$, are achieved through DNN. Architectural design of each DNN has similar network parameters and are summarized in Table 2.

Input layer and output layer of the DNN architecture represent the number of states and possible discrete action combinations of each agent, respectively. Four different hidden layers are combined and implemented in the DNN architecture with different numbers of nodes and Rectified Linear Unit (ReLu) activation functions. The illustration of the DNN architecture can be seen in Fig. 12.

Multi-layer perceptron structure is the keystone foundation of the overall structure of a DDQN based DRL optimization algorithm which includes multiple neural networks. Therefore, the computational complexity of a DRL algorithm relies on the action selection through optimum policies with neural networks. Generally, complexity analysis in multi-agent systems is performed as agent-wise and network-wise [99]. Although complexity analysis of agent-wise and network-wise are almost exactly the same, the only difference occurs

Algorithm 2 Multi DDQN Agent Based DRL Optimization Algorithm

Initialize: Q^n, Q_t^n, ϕ_t^n, τ^n, PER buffer D^n

for each agent (n) **do**
  for each episode (t) **do**
    for each environment step **do**
      Observe state s_t^n and choose $a_t^n \sim \phi^n(s_t^n, a_t^n)$ using epsilon-greedy with probability ϵ_t^n
      Execute a_t^n and observe the next states s_{t+1}^n and rewards r_t^n
      Store $(s_t^n, a_t^n, r_t^n, s_{t+1}^n)$ in the experience buffer D^n
    end for
    for each update step **do**
      Sample $e_t^n = (s_t^n, a_t^n, r_t^n, s_{t+1}^n) \sim D^n$
      Compute target critic (Q_t^n) value: $Q^{*,n}(s_t^n, a_t^n) = r_t^n + \gamma^n Q_t^n(s_{t+1}^n, \max Q^n(S^n, A^n; \phi^n); \phi_t^n)$
      Compute network losses (L^n): $L^n = \frac{1}{2M} \sum_{t=1}^{M} (r_t^n - Q^n(S^n, A^n; \phi^n))^2$
      Update target critic (Q_t^n) parameters: $\phi_t^n = \tau^n \phi^n + (1 - \tau^n)\phi_t^n$
      Update Epsilon-Greedy threshold (ϵ^n): $\epsilon_t^n = \epsilon_t^n (1 - \alpha_{decay,t}^n)$
    end for
  end for
end for

TABLE 1. Agent and training options of each DDQN agent. All agents have identical agent and training options.

Symbol	Description	Value
ϵ_{max}^n	Maximum value of Epsilon-Greedy thresholds	1.0
ϵ_{min}^n	Minimum value of Epsilon-Greedy thresholds	0.1
α_{decay}^n	Epsilon decay rates	4.5e-4
T_s^n	Simulation sample times (s)	1.0
τ^n	Target smoothing factors	0.01
γ^n	Discount factors of the expected long-term rewards	0.965
D_{length}^n	PER buffer lengths	10e5
$D_{minibatch}^n$	Mini-batch sizes	16
$T_{episode}$	Total episode number	5000
T_{step}	Total step numbers per episode	240

TABLE 2. Summary table of DNN architecture and network parameters. Due to the identical action/state space designs, each agent have identical DNN architecture and network parameters.

#	Layer Type	Node Size	Activation Function	Output Size	Trainable Parameters
1	Input Layer	7	-	7	-
2	Hidden Layer	3200	ReLu	3200	25600
3	Hidden Layer	1600	ReLu	1600	5121600
4	Hidden Layer	400	ReLu	400	640400
5	Hidden Layer	200	ReLu	200	80200
6	Output Layer	2592	-	2592	520992

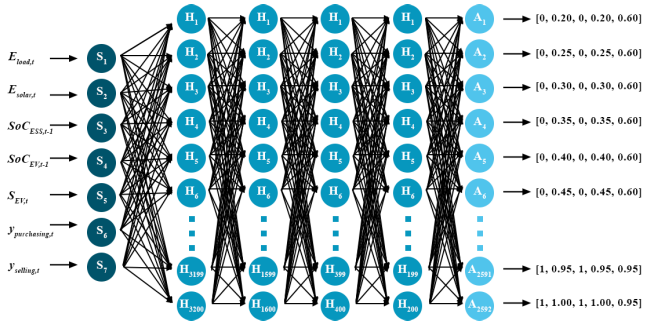

FIGURE 12. The architecture design of DNNs, including 7 input layers (states), 4 hidden layers (3200, 1600, 400, and 200 nodes), 2592 output layers (possible discrete action combinations).

on the number of agents involved. In the simplest form, the complexity analysis of DDQN-based DRL algorithm is composed of the complexity analysis of feed-forward pass in multi-layer perceptron structure mostly during the exploitation (with trained network) that forms the online and target networks [100].

The complexity analysis of agent-wise feed-forward ANN consists of input layer (with 7 nodes), 4 hidden layer with rectified linear unit activation function and different node size (3200, 1600, 400, and 200 nodes respectively) and output layer (with 2592 nodes). Therefore, the complexity of the input layer, hidden layers, and output layer are expressed as follows [99], [100]:

$$C_{input}^{agent} = O(S \times U_i) \tag{35a}$$
$$C_{hidden}^{agent} = O(n \times U_i \times U_j) \tag{35b}$$
$$C_{output}^{agent} = O(A \times U_j) \tag{35c}$$

where, S and A represent the number of states and actions, respectively. In the ANN structure, n, U_i, and U_j represent the number of hidden layers, neurons (node size) at the inputs and outputs, respectively. Consequently, the overall complexity of the DRL algorithm is expressed as follows [99], [100]:

$$C_{DRL}^{agent} = C_{input}^{agent} + C_{hidden}^{agent} + C_{output}^{agent} \tag{36}$$

The complexity analysis of network-wise feed-forward ANN shares the similar equations with agent-wise, except the number of agents involved in the environment. Therefore, the complexity of the input layer, hidden layers, and output layer are expressed as follows [99]:

$$C_{input}^{network} = O(I \times S \times U_i) \tag{37a}$$
$$C_{hidden}^{network} = O(I \times n \times U_i \times U_j) \tag{37b}$$
$$C_{output}^{network} = O(I \times A \times U_j) \tag{37c}$$

where, I represents the number of agents involved in the environment. Consequently, the overall complexity of the DRL algorithm is expressed as follows [99]:

$$C_{DRL}^{network} = C_{input}^{network} + C_{hidden}^{network} + C_{output}^{network} \tag{38}$$

It should be noted that each agent in the environment have the same ANN architecture for online and target networks.

TABLE 3. Time of Use (TOU) energy pricing scheme, consisting of 3 time zones, namely mid-peak, on-peak, and off-peak. The electric grid buys energy with Feed-in-Tariff (FiT) prices.

Time Zone	Starting Time	Ending Time	Selling Price	Buying Price
Mid-Peak	06:00	17:00	1.95	0.32
On-Peak	17:00	22:00	2.86	0.32
Off-Peak	22:00	06:00	1.22	0.32

V. SIMULATION RESULTS AND ANALYSIS

In total, 8760 data are collected through online-available datasets with sampling time $T_s^n = 60$ minutes. Consequently, the resolution of the proposed EMS is $T_{resolution=1}$ hour. 240 data (10 days) are selected randomly from the collected dataset ($T_{simulation}$). Collected data include required energy loads (E_{load}^n in kWh), solar energy generation (E_{solar}^n in kWh), status signals of EVs ($S_{EV}^n \subset \{0, 1\}$), and energy consumption of EVs during the driving cycles ($E_{driving}^n$ in kWh). Arrival and departure times of EVs are assigned randomly. However, due to training complexities, it is assumed that EVs are charged during the off-peak hours, between 03:00 AM and 06:00 AM, if necessary. Adapted TOU pricing scheme consists of 3 time zones, namely mid-peak, on-peak, and off-peak. FiT pricing scheme is adapted for energy selling transactions from prosumers to the grid. Adapted FiT consists of a constant value throughout the day. The electric price of the electric grid can be seen in Table 3. ESS of each prosumer include 7 parameters and prosumers have identical simulation parameters. Total energy capacities of ESSs (E_{ESS}^n) are designed as 28.8 kWh/cluster and 4.8 kWh/prosumer. Maximum ($SoC_{ESS,max}^n$) and minimum ($SoC_{ESS,min}^n$) SoCs are set to 1.0 and 0.1, respectively. Charging ($\eta_{ESS,charge}^n$) and discharging ($\eta_{ESS,discharge}^n$) efficiencies are set to 0.95 and 0.95, respectively. Maximum charging powers ($P_{ESS,max}^n$) are 30 kW/cluster and 5 kW/prosumer. Similarly, maximum discharging powers ($P_{ESS,min}^n$) are 12 kW/cluster and 2 kW/prosumer. The simulation parameters of ESSs can be seen in Table 4.

TABLE 4. Simulation parameters of Energy Storage Systems (ESSs). Each prosumer have a same size of ESS and identical parameters.

Symbol	Parameter	Cluster	Prosumer
E^n_{ESS}	Total energy capacity (kWh)	28.80	4.80
$SoC^n_{ESS,max}$	Maximum SoC	1.00	1.00
$SoC^n_{ESS,min}$	Minimum SoC	1.00	0.10
$\eta^n_{ESS,charge}$	Charging Efficiency	0.95	0.95
$\eta^n_{ESS,discharge}$	Discharging Efficiency	0.95	0.95
$P^n_{ESS,max}$	Maximum charging power (kW)	30	5
$P^n_{ESS,min}$	Maximum discharging power (kW)	12	2

Similar to ESSs, EV of each prosumer include 7 parameters. The difference is that EVs have distinct parameters. The proposed EMS includes 3 EVs, owned by Prosumer 1, Prosumer 2, and Prosumer 4. Prosumers 3, 5, and 6 do not have EVs. Total energy capacities (E^n_{EV}) are designed as 28 kWh, 26.3 kWh, and 37.3 kWh, respectively. Maximum SoCs ($SoC^n_{EV,max}$) are set to 0.7, 0.65, and 0.68, respectively. Minimum SoCs ($SoC^n_{EV,min}$) are set to 0.25, 0.25, and 0.30, respectively. Charging ($\eta^n_{EV,charge}$) and discharging ($\eta^n_{EV,discharge}$) efficiencies are 0.95 and 0.92, respectively. Maximum discharging powers ($P^n_{EV,max}$) are 3 kW, 2.3 kW, and 2 kW, respectively. Level 2 AC charging infrastructure is adopted and charging powers ($P^n_{EV,charge}$) are 11 kW, 7.4 kW, and 7.4 kW, respectively. The simulation parameters of EVs can be seen in Table 5.

TABLE 5. Simulation parameters of Electric Vehicles (EVs). EVs are owned by Prosumers 1, 2, and 4.

Symbol	Parameter	P1	P2	P4
E^n_{EV}	Energy capacity (kWh)	28.0	26.3	37.3
$SoC^n_{EV,max}$	Maximum SoC	0.70	0.65	0.68
$SoC^n_{EV,min}$	Minimum SoC	0.25	0.25	0.30
$\eta^n_{EV,charge}$	Charging Efficiency	0.95	0.95	0.95
$\eta^n_{EV,discharge}$	Discharging Efficiency	0.92	0.9	0.92
$P^n_{EV,max}$	Maximum discharging power (kW)	3.0	2.3	2.0
$P^n_{EV,charge}$	Charging power (kW)	11	7.4	7.4

A. EVALUATION PHASE

In total, prosumers require 809.04 kWh/cluster (10 days) of energy loads. On average, prosumers require 134.84 kWh/prosumer (10 days) of energy loads. Daily average energy requirements are 13.48 kWh/prosumer. However, prosumers have distinct load requirement profiles (see in Fig. 13) to further improve the efficiency of learning phase. The descriptive statistical values of the energy load requirements for each prosumer can be seen in Table 6.

It is assumed that solar panels are designed in a centralized architecture and prosumers benefit from the generated solar energy equally (see in Fig. 14). In total, solar panels generate 559.44 kWh/cluster (10 days) of energy. Per prosumer, solar panels generate 93.24 kWh/prosumer (10 days). Daily solar energy generations per prosumer are 9.32 kWh/prosumer.

TABLE 6. Energy load requirements of prosumers, explaining the total, daily, maximum, minimum, average, and standard deviation of the required energy.

Prosumer	Evaluation (10 days)					
	Total (kWh)	Daily (kWh)	Max (kWh)	Min (kWh)	Mean (kWh)	Std (kWh)
Prosumer 1	142.02	14.20	0.88	0.29	0.58	0.08
Prosumer 2	139.23	13.92	1.06	0.36	0.57	0.13
Prosumer 3	112.34	11.23	0.84	0.14	0.46	0.11
Prosumer 4	91.29	9.12	0.82	0.08	0.38	0.18
Prosumer 5	144.48	14.44	0.84	0.38	0.59	0.08
Prosumer 6	179.68	17.96	1.35	0.36	0.75	0.26

Therefore, the generated solar energy can solely meet the required energy by 65.65%, 66.96%, 83%, 102.13%, 64.54%, and 51.89% of each prosumer, respectively. However, due to the intermittent energy generation profile of solar panels, excessive energies are stored in ESSs for future energy requirements and P2P energy trading. The descriptive statistical values of the generated solar energy can be seen in Table 7.

TABLE 7. Solar energy generations, explaining the total, daily, maximum, minimum, average, and standard deviation of the generated energy.

Prosumer	Evaluation (10 days)					
	Total (kWh)	Daily (kWh)	Max (kWh)	Min (kWh)	Mean (kWh)	Std (kWh)
Cluster	559.44	55.94	10.54	0	2.32	3.41
Prosumer	93.24	9.32	1.76	0	0.39	0.57

The proposed EMS includes 2 different input data for EVs. The first input data includes the binary signals that represent the connection status of each EV to the cluster at time step t. Binary 1 represents that EVs are connected to the cluster and 0 represents that EVs are not connected. When connected to the cluster, EVs supply energy to the owners and other peers via V2X and P2P energy trading when required. Typically, EVs are connected to the cluster between 06:00 AM and 08:00 PM on the weekdays. On the weekends, EVs are generally connected to the cluster throughout the day. During the simulation (240 hours), the total connection duration (see in Fig. 15) to the cluster of each prosumers' EV are 176, 188, and 192 hours. The second input data includes daily energy consumption signals that represent the energy consumption of EVs for transportation purposes during the daily driving cycles (see in Fig. 16). After each driving cycle, SoCs of each EV are recalculated with the energy consumption signals. In total, EVs consume 149 kWh/cluster during the driving cycles. The total energy consumption of each prosumer, during the simulation (10 days), are 53.37 kWh, 45.04 kWh, and 28.18 kWh, respectively. The total connection duration and energy consumption of each prosumer's EV can be seen in Table 8.

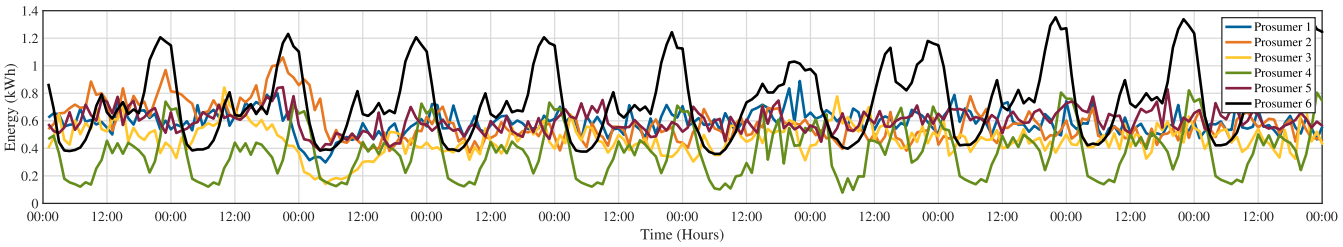

FIGURE 13. Energy load requirements of prosumers. Prosumers have different load profiles to improve the efficiency of the proposed EMS with different input data.

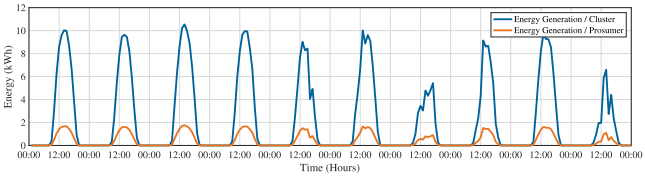

FIGURE 14. Solar energy generations per cluster and prosumer. Solar panels are designed in a centralized architecture and generated energies are distributed equally to prosumers.

TABLE 8. Connection duration (hours) to the cluster and energy consumption (kWh) of EVs.

Prosumer	Evaluation (10 days)	
	Total Connection	Total Energy
Prosumer 1	176 Hours	53.37 kWh
Prosumer 2	188 Hours	45.04 kWh
Prosumer 4	192 Hours	28.18 kWh

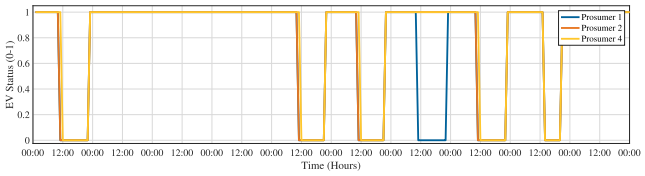

FIGURE 15. Connection status of EVs to the cluster. Status are represented as binary signals, where 1 represent that EVs are connected to the cluster and 0 represent that EVs are not connected to the cluster.

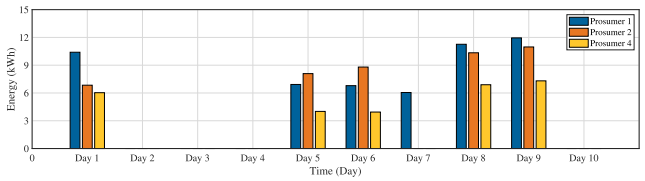

FIGURE 16. Daily energy consumption (for driving) of EVs that represent the required energy for transportation purposes.

The training phase is carried out in 5000 episodes, including 240 steps (10 days) per episode. Agents converge to the optimal solutions (Q-Values) after approximately 1500 episodes with learning rates $\alpha^n_{learning} = 0.001$. Training of 6 DDQN agents takes approximately 18 hours to complete. After the training phase, the computational process of the multi DDQN agent based DRL algorithm takes 17.18 milliseconds per step in MATLAB. Agent 4 has the highest average reward with 186.11. Contrarily, Agent 2 has the lowest average reward with 112.77. Average rewards of each agent, ($R^n_{average}$), are 156.06, 112.77, 155.43, 186.11, 137.83, and 133.06, respectively. Estimated average Q-Values are 156.40, 119.86, 164.10, 181.33, 143.63, and 142.90, respectively. Between episode 1500 and 5000, where steady outcomes are achieved, Mean Absolute Percentage Errors (MAPEs) between average rewards and estimated average Q-Values are 4.49%, 5.57%, 7.90%, 5.88%, 5.69%, and 5.28 %, respectively. Therefore, estimated average Q-values approach to the average rewards and verify the accuracy (MAPEs < 10%) of the DNN architectures. The outcomes of the learning phase for each agent can be seen in Fig. 17.

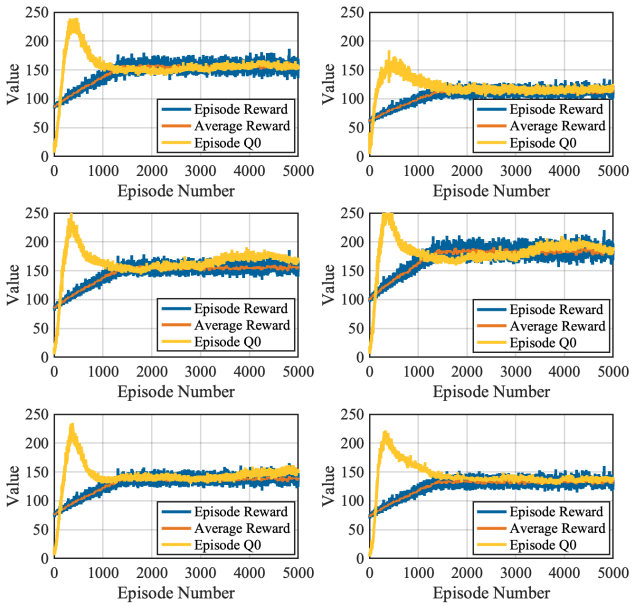

FIGURE 17. Agent learning curves of the proposed HEMS, composed of episode rewards, average rewards, and estimated average Q-Values.

B. CASE STUDIES

To compare the outcomes of the proposed EMS, 4 different case studies are examined. The case studies are differentiated

depending on the existence of V2X technologies and P2P energy trading implementations within the cluster. The content of the examined case studies are:

1) Case Study 1: represents the proposed EMS and includes the collaboration of V2X technologies and P2P energy trading implementations.
2) Case Study 2: includes the implementation of V2X technologies without P2P energy trading.
3) Case Study 3: includes the implementation of P2P energy trading without V2X technologies
4) Case Study 4: represents a rule-based conventional HEMS without the implementations of V2X technologies and P2P energy trading.

TABLE 9. Comparisons of V2X and P2P energy trading implementations in the case studies.

Prosumer	Case 1		Case 2		Case 3		Case 4	
	V2X	P2P	V2X	P2P	V2X	P2P	V2X	P2P
Prosumer 1	True	True	True	False	False	True	False	False
Prosumer 2	True	True	True	False	False	True	False	False
Prosumer 3	False	True	False	False	False	True	False	False
Prosumer 4	True	True	True	False	False	True	False	False
Prosumer 5	False	True	False	False	False	True	False	False
Prosumer 6	False	True	False	False	False	True	False	False

In case studies 1 and 3, the prosumers can supply energy to the other prosumers via stored energy in ESSs and surplus energy generations in RESs through P2P energy trading. Additionally, in case study 1, Prosumers 1, 2, and 4 can supply energy to the other prosumers via V2X technologies through P2P energy trading. In case study 4, all prosumers meet the self required energy loads only via stored energy in ESS, generated solar energy, and from the grid if necessary. In addition to stored energy in ESS and generated solar energy, Prosumers 1, 2, and 4 can meet the self required energy loads via V2X technologies in case study 2.

TABLE 10. The implementation of DRL algorithm in each case study.

Implementation	Case 1	Case 2	Case 3	Case 4
DRL Algorithm	Yes	No	Yes	No
P2P Energy Trading	Yes	No	Yes	No
V2X Technologies (in P2P trading)	Yes	No	No	No
V2X Technologies (self consumption)	Yes	Yes	No	No

On the other hand, the implementation of DRL algorithm relies on the existence of P2P energy trading in the environment. The reason for this is that the outcomes of DRL algorithm specifies the amounts of energy to be sold in P2P energy trading and the energy prices of each seller. Therefore, according to Table 9, DDQN-based DRL algorithm is implemented in various cases based on different requirements. The implementation status of DRL algorithm in each case can be seen in Table 10.

C. NUMERICAL RESULTS AND DISCUSSIONS

Compared to case studies 2, 3, and 4, energy costs are the lowest in case study 1, where V2X technologies and P2P energy trading are implemented with the proposed EMS. In the cluster, energy costs in case study 1 are 13.19%, 16.81%, and 19.18% lower than in case studies 2, 3, and 4, respectively. Therefore, the collaboration of V2X technologies and P2P energy trading have significant impact (19.18%) on reducing the cumulative energy costs without affecting the comfort of the prosumers. The energy cost comparison of each prosumer can be seen visually in Fig. 18 and numerically in Table 11.

TABLE 11. Total energy cost comparisons in (TRY).

Prosumers	Case 1	Case 2	Case 3	Case 4
Prosumer 1	134.28	140.29	154.48	158.46
Prosumer 2	101.94	122.86	142.57	143.76
Prosumer 3	36.74	42.40	42.30	42.40
Prosumer 4	42.10	50.36	56.71	59.11
Prosumer 5	73.61	83.68	82.71	83.68
Prosumer 6	129.10	154.43	142.65	154.43
Totals	**517.76**	**594.02**	**621.41**	**641.84**

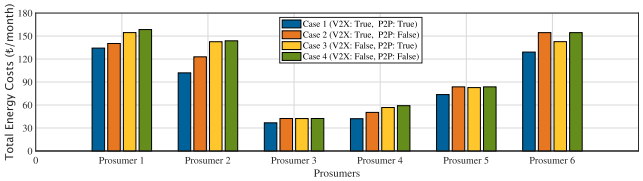

FIGURE 18. Case by case total energy cost comparisons of each prosumer in (TRY).

The case studies 1 and 3 have dynamic energy pricing schemes due to the P2P energy pricing based on the RL action outcomes. The average energy prices of each case study are 0.89 TRY/kWh, 1.71 TRY/kWh, 1.48 TRY/kWh, and 1.73 TRY/kWh. As seen on the average prices of case studies 1 and 3, P2P energy trading has a significant impact on reducing the average energy prices. Compared to case study 4, the collaboration of V2X technologies and P2P energy trading (case study 1) could reduce the average energy prices by 47.92%. P2P energy trading implementation without V2X technologies (case study 3) could reduce the average energy prices 14.16%. On contrary, V2X technologies without P2P energy trading implementation (case study 2) could reduce the average energy prices 0.46%, thus having an insignificant impact. The average energy prices of each case can be seen in the Fig. 19.

The SSRs are expressed as follows:

$$SSR^n = 1 - \sum_{t=1}^{T_{simulation}} \frac{E^n_{grid,buying}}{E^n_{load}} \tag{39}$$

where, $E^n_{grid,buying}$ and E^n_{load} are energy bought from the grid and the required energy loads, respectively. The average

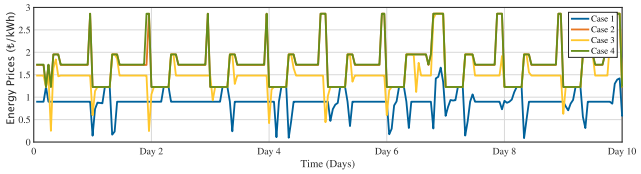

FIGURE 19. Average energy prices (TRY/kWh) of each case, respectively (missing values, where energy requirements are supplied from self-resources, are filled with mean values).

SSRs (as seen in Fig. 20), SSR_{avg}, of each case are 65.87%, 60.01%, 56.36%, and 56.48%, respectively. Consequently, compared to case study 4, case studies 1 and 2 indicates that collaboration of V2X technologies and P2P energy trading could increase SSRs 9.39% and V2X technologies solely could increase 3.53%. However, P2P energy trading implementation without V2X technologies (case study 3) could decrease SSRs (-0.12%), due to the required energy transactions and insufficient self-resources. Therefore, optimization issues occur between reducing energy costs and increasing SSR with P2P energy trading implementation without V2X technologies.

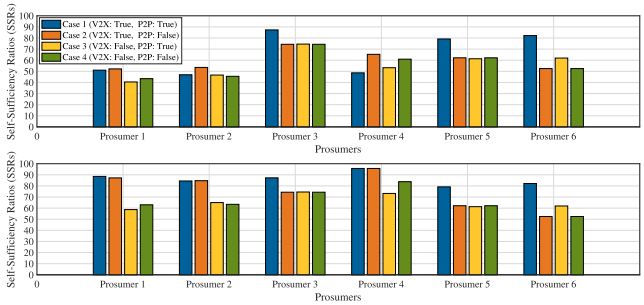

FIGURE 20. The Self-Sufficiency Ratios (SSR) of each prosumer and case, with and without EV charging energies are included.

As seen on the SSRs of Prosumers 1, 2, and 4, collaboration of V2X technologies and P2P energy trading reduces the prosumers' SSR due to the energy transactions of EVs through V2X technologies. However, the average SSR of all prosumers are reduced and therefore have significant impact in cluster-wise. The numerical comparison of SSRs of each prosumer and case study can be seen in Table 12. Furthermore, the cumulative energy costs of all prosumers are reduced with the proposed EMS (collaboration of V2X technologies and P2P energy trading). Therefore, it is concluded that all prosumers benefit from the proposed EMS in terms of energy costs. Prosumers with EVs benefit from the profits of P2P energy trading via V2X technologies. Prosumers without EVs benefit from the cheaper energy prices that supplied from other peers through P2P energy trading and V2X technologies. The purchased energy from the grid can be seen in Fig. 21. Additionally, Fig. 22 and Table 13 shows the utilization of different energy resources for each case studies.

TABLE 12. The Self-Sufficiency Ratio (SSR) comparisons of prosumers in (%). EV charging energies are included in the comparisons.

Prosumers	Case 1	Case 2	Case 3	Case 4
Prosumer 1	51.06	52.19	40.44	43.41
Prosumer 2	46.93	53.53	46.65	45.53
Prosumer 3	87.28	74.35	74.53	74.35
Prosumer 4	48.64	65.32	53.26	60.95
Prosumer 5	79.15	62.18	61.33	62.18
Prosumer 6	82.19	52.48	61.95	52.48
Averages	**65.87**	**60.01**	**56.36**	**56.48**

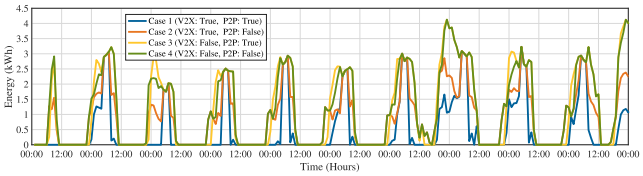

FIGURE 21. Energy bought from the electric grid in each case, respectively. Compared to other case studies, the proposed EMS can reduce the peak energy demands by approximately 4.06%.

Among the all case studies, the total required energy loads and the total generated solar energies are same and are 809.04 kWh and 559.44 kWh, respectively. The amount of charging and discharging energy of ESSs are similar in case studies. However, the amount of charging and discharging energy of EVs are different due to the V2X technologies and P2P energy trading implementations. In case studies 1 and 2, EVs require 304.04 kWh and 218.99 of energy for charging, respectively. On the other hand, in case studies 3 and 4, EVs require 153.10 kWh of energy for charging. Similar to amount of charging energy, the amount of discharging energy of EVs are different. In case studies 1 and 2, EVs discharge 87.65 kWh and 74.90 kWh to meet required self energy loads. Contrarily, since case studies 3 and 4 do not have V2X implementations, EVs discharge 0.00 kWh of energy.

In total, to meet the required energy loads of each case, 118.21 kWh, 212.17 kWh, 284.67 kWh, and 287.06 kWh of energy are bought from the grid, respectively. With energy required to charge EVs are added, 422.25 kWh, 431.15 kWh, 437.77 kWh, and 440.16 kWh of energy are bought from the grid to charge EVs and meet the required energy loads. Cumulatively, compared to case study 4, energy dependencies on the electric grid of case studies 1, 2, and 3 are reduced 53.73 kWh/month, 27.03 kWh/month, and 7.17 kWh/month, respectively.

During the P2P energy trading process, 100.75 kWh and 23.47 kWh of energy are exchanged between prosumers in case studies 1 and 3, respectively. In case study 3, all exchanged energy are supplied from stored energy in ESSs. However, in case study 1, 21.81 kWh and 78.93 kWh of energy are supplied from ESSs and EVs to other prosumers. Therefore, V2X technologies have significant impact on reducing the cumulative energy costs via P2P energy trading implementation.

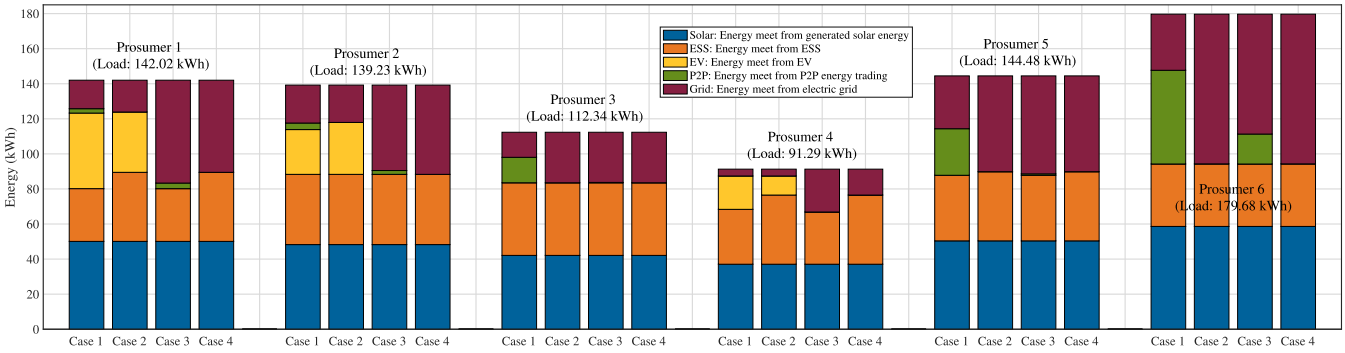

FIGURE 22. Utilization of different energy resources to supply the required energy loads of each case and prosumer.

TABLE 13. Summarizing the energy resource utilization (kWh), energy costs (TRY), and self-sufficiency ratios (%) in each case (cumulative results).

Case	Load	Solar	ESS		EV		P2P		Electric Grid			Cost	SSR
			Charge	Discharge	Charge	Discharge	ESS	EV	Total	Bought	Sold		
Case 1	809.04	559.44	214.66	215.88	304.04	87.65	21.81	78.93	100.75	422.25	58.23	517.76	65.87
Case 2	809.04	559.44	212.39	235.43	218.99	74.90	0.00	0.00	0.00	431.15	60.50	594.02	60.01
Case 3	809.04	559.44	214.78	214.35	153.10	0.00	23.47	0.00	23.47	437.77	58.11	621.41	56.36
Case 4	809.04	559.44	212.39	235.43	153.10	0.00	0.00	0.00	0.00	440.16	60.50	641.84	56.48

Therefore, the proposed EMS reduces the energy costs (by 19.18%) with the collaboration of V2X technologies and P2P energy trading implementations without affecting the comfort of the prosumers. Also, the proposed EMS improve the SSR (9.39%) to further increase the grid independence (53.73 kWh/month) in a small-scale environment, including 6 prosumers.

VI. CONCLUSION AND FUTURE WORKS

In this paper, a smart EMS, including 6 prosumers and DDQN multi-agent environment, is proposed and optimized with DQN algorithm to minimize the cumulative energy costs, improve the SSR and grid independence, and optimize the utilization of RESs and EVs through V2X (V2H, V2L, and V2G) technologies and P2P energy trading implementations. The proposed EMS is formulated as a MDP with discrete state and actions spaces, determining the participation status of each prosumer in the P2P energy trading, energy prices, and amount of energy to be sold in P2P energy trading. The multi-agent environment is trained for 5000 episodes, including 240 steps (10 days) per episode. Average reward and episode Q-values are calculated as 146.87 and 151.37, respectively. Therefore, average MAPEs (5.80%) value shows a well-designed DNN architecture to estimate future rewards. In the evaluation phase, the outcomes of proposed EMS are compared with 4 cases, differentiated on the existence of the V2X technologies and P2P energy trading implementations. Numerical results show that, compared to conventional rule-based EMS, the proposed EMS reduces energy costs by 19.18% and average energy prices (TRY/kWh) by 47.92% without affecting the comfort of prosumers. Additionally, the proposed EMS increases SSRs by 9.39%, thus increasing the grid independence by 53.73 kWh/month.

In future works, it is planned to implement a large-scale environment (with more than 1000 prosumers), including V2X technologies, P2P energy trading with cluster-based different pricing mechanisms, DSM and DR energy management architectures with optimized comforts of prosumers, and categorized ESSs, EVs, and RESs with different energy capacity size. Future works will further investigate the effects of energy capacity sizing, different pricing mechanisms used in P2P energy trading, and appliance scheduling on reducing energy costs and optimizing the comfort of prosumers. It is planned to integrate a machine learning based forecasting model to forecast the dynamic behavior of prosumers, solar energy generations, and required energy loads to further improve the SSRs and the utilization of different energy resources.

REFERENCES

[1] M. A. Usova and V. I. Velkin, "Possibility to use renewable energy sources for increasing the reliability of the responsible energy consumers on the enterprise," in *Proc. 17th Int. Ural Conf. AC Electr. Drives (ACED)*, Ekaterinburg, Russia, Mar. 2018, pp. 1–4, doi: 10.1109/ACED.2018.8341682.

[2] P. Pinson, L. Mitridati, C. Ordoudis, and J. Østergaard, "Towards fully renewable energy systems: Experience and trends in Denmark," *CSEE J. Power Energy Syst.*, vol. 3, no. 1, pp. 26–35, Mar. 2017, doi: 10.17775/CSEEJPES.2017.0005.

[3] Y. Sun, Z. Zhao, M. Yang, D. Jia, W. Pei, and B. Xu, "Overview of energy storage in renewable energy power fluctuation mitigation," *CSEE J. Power Energy Syst.*, vol. 6, no. 1, pp. 160–173, Mar. 2020, doi: 10.17775/CSEEJPES.2019.01950.

[4] L. Gigoni, A. Betti, E. Crisostomi, A. Franco, M. Tucci, F. Bizzarri, and D. Mucci, "Day-ahead hourly forecasting of power generation from photovoltaic plants," *IEEE Trans. Sustain. Energy*, vol. 9, no. 2, pp. 831–842, Apr. 2018, doi: 10.1109/TSTE.2017.2762435.

[5] H. Zhao, Z. Shen, Z. Wang, J. Guo, E. Hu, Y. Wang, Z. Shen, X. Song, and Y. Xing, "Research on the influence of microgrid on power supply mode of medium-voltage distribution network considering reliability and economy," in *Proc. 2nd IEEE Conf. Energy Internet Energy Syst. Integr. (EI2)*, Beijing, China, Oct. 2018, pp. 1–6, doi: 10.1109/EI2.2018.8582308.

[6] G. Mejía-Ruiz, M. R. A. Paternina, J. R. Rodríguez, A. Zamora, G. Bolivar-Ortiz, and C. Toledo-Santos, "A bidirectional isolated charger for electric vehicles in V2G systems with the capacity to provide ancillary services," in *Proc. 52nd North Amer. Power Symp. (NAPS)*, Tempe, AZ, USA, Apr. 2021, pp. 1–6, doi: 10.1109/NAPS50074.2021.9449674.

[7] E. J. Palacios-Garcia, E. Rodriguez-Diaz, A. Anvari-Moghaddam, M. Savaghebi, J. C. Vasquez, J. M. Guerrero, and A. Moreno-Munoz, "Using smart meters data for energy management operations and power quality monitoring in a microgrid," in *Proc. IEEE 26th Int. Symp. Ind. Electron. (ISIE)*, Edinburgh, U.K., Jun. 2017, pp. 1725–1731, doi: 10.1109/ISIE.2017.8001508.

[8] L. E. Zubieta, "Power management and optimization concept for DC microgrids," in *Proc. IEEE 1st Int. Conf. DC Microgrids (ICDCM)*, Atlanta, GA, USA, Jun. 2015, pp. 81–85, doi: 10.1109/ICDCM.2015.7152014.

[9] M. Zhai, Y. Liu, T. Zhang, and Y. Zhang, "Robust model predictive control for energy management of isolated microgrids," in *Proc. IEEE Int. Conf. Ind. Eng. Eng. Manage. (IEEM)*, Singapore, Dec. 2017, pp. 2049–2053, doi: 10.1109/IEEM.2017.8290252.

[10] E. Rodriguez-Diaz, E. J. Palacios-Garcia, A. Anvari-Moghaddam, J. C. Vasquez, and J. M. Guerrero, "Real-time energy management system for a hybrid AC/DC residential microgrid," in *Proc. IEEE 2nd Int. Conf. DC Microgrids (ICDCM)*, Nuremburg, Germany, Jun. 2017, pp. 256–261, doi: 10.1109/ICDCM.2017.8001053.

[11] A. T. Eseye, D. Zheng, J. Zhang, and D. Wei, "Optimal energy management strategy for an isolated industrial microgrid using a modified particle swarm optimization," in *Proc. IEEE Int. Conf. Power Renew. Energy (ICPRE)*, Shanghai, China, Oct. 2016, pp. 494–498, doi: 10.1109/ICPRE.2016.7871126.

[12] K. Patel and A. Khosla, "Home energy management systems in future smart grid networks: A systematic review," in *Proc. 1st Int. Conf. Next Gener. Comput. Technol. (NGCT)*, Dehradun, India, Sep. 2015, pp. 479–483, doi: 10.1109/NGCT.2015.7375165.

[13] B. Han, Y. Zahraoui, M. Mubin, S. Mekhilef, M. Seyedmahmoudian, and A. Stojcevski, "Home energy management systems: A review of the concept, architecture, and scheduling strategies," *IEEE Access*, vol. 11, pp. 19999–20025, 2023, doi: 10.1109/ACCESS.2023.3248502.

[14] U. Zafar, S. Bayhan, and A. Sanfilippo, "Home energy management system concepts, configurations, and technologies for the smart grid," *IEEE Access*, vol. 8, pp. 119271–119286, 2020, doi: 10.1109/ACCESS.2020.3005244.

[15] J. Liu, S. Zhang, and J. Wang, "Demand-side management strategy for local energy system supporting renewable energy sources integration," in *Proc. 6th Int. Conf. Energy, Electr. Power Eng. (CEEPE)*, Guangzhou, China, May 2023, pp. 1332–1337, doi: 10.1109/CEEPE58418.2023.10167321.

[16] G. A. Raiker, S. Reddy B., L. Umanand, S. Agrawal, A. S. Thakur, K. Ashwin, J. P. Barton, and M. Thomson, "Internet of Things based demand side energy management system using non-intrusive load monitoring," in *Proc. IEEE Int. Conf. Power Electron., Smart Grid Renew. Energy (PESGRE)*, Cochin, India, Jan. 2020, pp. 1–5, doi: 10.1109/PESGRE45664.2020.9070739.

[17] M. Anzar, R. Iqra, A. Kousar, S. Ejaz, M. S. Alvarez-Alvarado, and A. K. Zafar, "Optimization of home energy management system in smart grid for effective demand side management," in *Proc. Int. Renew. Sustain. Energy Conf. (IRSEC)*, Tangier, Morocco, Dec. 2017, pp. 1–6, doi: 10.1109/IRSEC.2017.8477255.

[18] Y. Zhang, Y. Ma, S. Zhang, L. Chen, and H. Liu, "Building-level demand-side energy management based on game theory," in *Proc. IEEE Int. Conf. Mechatronics Autom. (ICMA)*, Guangxi, China, Aug. 2022, pp. 65–69, doi: 10.1109/ICMA54519.2022.9856118.

[19] P. Mundra, A. Arya, S. Gawre, and S. Mehroliya, "Independent demand side management system based on energy consumption scheduling by NSGA-II for futuristic smart grid," in *Proc. IEEE-HYDCON*, Hyderabad, India, Sep. 2020, pp. 1–6, doi: 10.1109/HYDCON48903.2020.9242816.

[20] Q. Luo and S. Zhang, "Two-stage optimal scheduling of virtual power plant considering demand response and forecast errors," in *Proc. 4th Int. Conf. Electr. Eng. Control Technol. (CEECT)*, Shanghai, China, Dec. 2022, pp. 852–856, doi: 10.1109/CEECT55960.2022.10030666.

[21] Q. Liu, B. Wang, J. Cao, X. Peng, and K. Huang, "Automatic demand response evaluation method of regional power grid," in *Proc. IEEE Sustain. Power Energy Conf. (iSPEC)*, Nanjing, China, Dec. 2021, pp. 2493–2497, doi: 10.1109/iSPEC53008.2021.9735801.

[22] C. Arun, R. Aswinraj, M. T. Bijoy, M. Nidheesh, and R. Rohikaa Micky, "Day ahead demand response using load shifting technique in presence of increased renewable penetration," in *Proc. IEEE 7th Int. Conf. Converg. Technol. (I2CT)*, Mumbai, India, Apr. 2022, pp. 1–6, doi: 10.1109/I2CT54291.2022.9825258.

[23] Y. Nan, C. Chenggang, L. Xiaotong, C. Jing, and X. Peifeng, "Research on demand response strategy of HVAC based on deep reinforcement learning," in *Proc. 5th Int. Conf. Power Renew. Energy (ICPRE)*, Shanghai, China, Sep. 2020, pp. 456–460, doi: 10.1109/ICPRE51194.2020.9233115.

[24] G. Xue, C. Wu, W. Niu, X. Dou, S. Wang, and Y. Fu, "Optimal scheduling of integrated power and gas energy systems considering demand response," in *Proc. IEEE 6th Conf. Energy Internet Energy Syst. Integr. (EI2)*, Chengdu, China, Nov. 2022, pp. 2569–2573, doi: 10.1109/EI256261.2022.10116545.

[25] S. Patil and S. R. Deshmukh, "Development of control strategy to demonstrate load priority system for demand response program," in *Proc. IEEE Int. WIE Conf. Electr. Comput. Eng. (WIECON-ECE)*, Bangalore, India, Nov. 2019, pp. 1–6, doi: 10.1109/WIECON-ECE48653.2019.9019950.

[26] L.-X. Wang and J. M. Mendel, "Generating fuzzy rules by learning from examples," *IEEE Trans. Syst., Man, Cybern.*, vol. 22, no. 6, pp. 1414–1427, Nov. 1992, doi: 10.1109/21.199466.

[27] G. Sterling and B. Tyler, "Renewable energy management using action dependent heuristic dynamic programming," in *Proc. IEEE Int. Smart Cities Conf. (ISC2)*, Kansas City, MO, USA, Sep. 2018, pp. 1–5, doi: 10.1109/ISC2.2018.8656942.

[28] M. J. Sanjari, H. Karami, and H. B. Gooi, "Analytical rule-based approach to online optimal control of smart residential energy system," *IEEE Trans. Ind. Informat.*, vol. 13, no. 4, pp. 1586–1597, Aug. 2017, doi: 10.1109/TII.2017.2651879.

[29] S.-Y. Chen and C.-H. Chang, "Optimal power flows control for home energy management with renewable energy and energy storage systems," *IEEE Trans. Energy Convers.*, vol. 38, no. 1, pp. 218–229, Mar. 2023, doi: 10.1109/TEC.2022.3198883.

[30] D. M. Minhas, J. Meiers, and G. Frey, "A rule-based expert system for home power management incorporating real-life data sets," in *Proc. 3rd Int. Conf. Smart Grid Renew. Energy (SGRE)*, Doha, Qatar, Mar. 2022, pp. 1–6, doi: 10.1109/SGRE53517.2022.9774212.

[31] A. Abbasi, H. A. Khalid, H. Rehman, and A. U. Khan, "A novel dynamic load scheduling and peak shaving control scheme in community home energy management system based microgrids," *IEEE Access*, vol. 11, pp. 32508–32522, 2023, doi: 10.1109/ACCESS.2023.3255542.

[32] L. Yao, W. H. Lim, and C.-C. Lai, "Self-learning fuzzy controller-based energy management for smart home," in *Proc. IEEE Int. Conf. Internet Things (iThings) IEEE Green Comput. Commun. (GreenCom) IEEE Cyber, Phys. Social Comput. (CPSCom) IEEE Smart Data (SmartData)*, Chengdu, China, Dec. 2016, pp. 87–93, doi: 10.1109/iThings-GreenCom-CPSCom-SmartData.2016.41.

[33] A. Soetedjo, Y. I. Nakhoda, and C. Saleh, "Simulation of fuzzy logic based energy management for the home with grid connected PV-battery system," in *Proc. 2nd Int. Conf. Smart Grid Smart Cities (ICSGSC)*, Kuala Lumpur, Malaysia, Aug. 2018, pp. 122–126, doi: 10.1109/ICSGSC.2018.8541271.

[34] R. S. Sutton and A. G. Barto, "Reinforcement learning: An introduction," *IEEE Trans. Neural Netw.*, vol. 9, no. 5, p. 1054, Sep. 1998, doi: 10.1109/TNN.1998.712192.

[35] A. J. Litchy and M. H. Nehrir, "Real-time energy management of an islanded microgrid using multi-objective particle swarm optimization," in *Proc. IEEE PES Gen. Meeting | Conf. Expo.*, National Harbor, MD, USA, Jul. 2014, pp. 1–5, doi: 10.1109/PESGM.2014.6938997.

[36] A. Salem, A. El-Shenawy, and M. S. Hamad, "Energy management strategy for grid connected DC hybrid micro grid using particle swarm optimization technique," in *Proc. 32nd Int. Conf. Microelectron. (ICM)*, Aqaba, Jordan, Dec. 2020, pp. 1–5, doi: 10.1109/ICM50269.2020.9331823.

[37] A. Ignat, E. Lazar, and D. Petreus, "Energy management for an islanded microgrid based on particle swarm optimization," in *Proc. IEEE 24th Int. Symp. for Design Technol. Electron. Packaging? (SIITME)*, Iasi, Romania, Oct. 2018, pp. 213–216, doi: 10.1109/SIITME.2018.8599272.

[38] J. Xue, C. Yan, L. Liu, J. Wang, X. Zhao, and X. Wang, "ADHDP-based housing energy management for two housing units with mobile storage," in *Proc. IEEE 9th Data Driven Control Learn. Syst. Conf. (DDCLS)*, Liuzhou, China, Nov. 2020, pp. 69–73, doi: 10.1109/DDCLS49620.2020.9275280.

[39] D. Liu, Y. Xu, Q. Wei, and X. Liu, "Residential energy scheduling for variable weather solar energy based on adaptive dynamic programming," *IEEE/CAA J. Autom. Sinica*, vol. 5, no. 1, pp. 36–46, Jan. 2018, doi: 10.1109/JAS.2017.7510739.

[40] H. Zhang, D. Wu, and B. Boulet, "A review of recent advances on reinforcement learning for smart home energy management," in *Proc. IEEE Electr. Power Energy Conf. (EPEC)*, Edmonton, AB, Canada, Nov. 2020, pp. 1–6, doi: 10.1109/EPEC48502.2020.9320042.

[41] H. Zhang, S. Seal, D. Wu, F. Bouffard, and B. Boulet, "Building energy management with reinforcement learning and model predictive control: A survey," *IEEE Access*, vol. 10, pp. 27853–27862, 2022, doi: 10.1109/ACCESS.2022.3156581.

[42] X. Zhou, J. Wang, X. Wang, and S. Chen, "Deep reinforcement learning for microgrid operation optimization: A review," in *Proc. 8th Asia Conf. Power Electr. Eng. (ACPEE)*, Tianjin, China, Apr. 2023, pp. 2059–2065, doi: 10.1109/acpee56931.2023.10135713.

[43] Q. Sun, D. Wang, D. Ma, and B. Huang, "Multi-objective energy management for we-energy in energy internet using reinforcement learning," in *Proc. IEEE Symp. Ser. Comput. Intell. (SSCI)*, Honolulu, HI, USA, Nov. 2017, pp. 1–6, doi: 10.1109/SSCI.2017.8285243.

[44] A. B. Dayani, H. Fazlollahtabar, R. Ahmadiahangar, A. Rosin, M. S. Naderi, and M. Bagheri, "Applying reinforcement learning method for real-time energy management," in *Proc. IEEE Int. Conf. Environ. Electr. Eng. IEEE Ind. Commercial Power Syst. Eur.*, Genova, Italy, Jun. 2019, pp. 1–5, doi: 10.1109/EEEIC.2019.8783766.

[45] M. K. Perera, K. T. M. U. Hemapala, and W. D. A. S. Wijayapala, "Developing a reinforcement learning model for energy management of microgrids in Python," in *Proc. Int. Conf. Comput. Intell. Knowl. Economy (ICCIKE)*, Dubai, United Arab Emirates, Mar. 2021, pp. 68–73, doi: 10.1109/ICCIKE51210.2021.9410754.

[46] S.-J. Chen, W.-Y. Chiu, and W.-J. Liu, "User preference-based demand response for smart home energy management using multiobjective reinforcement learning," *IEEE Access*, vol. 9, pp. 161627–161637, 2021, doi: 10.1109/ACCESS.2021.3132962.

[47] C. Hu, Y. Zhou, and Y. Lei, "Optimal energy management in microgrids based on reinforcement learning," in *Proc. 1st Int. Conf. Cyber-Energy Syst. Intell. Energy (ICCSIE)*, Shenyang, China, Jan. 2023, pp. 1–6, doi: 10.1109/ICCSIE55183.2023.10175284.

[48] B.-C. Lai, W.-Y. Chiu, and Y.-P. Tsai, "Multiagent reinforcement learning for community energy management to mitigate peak rebounds under renewable energy uncertainty," *IEEE Trans. Emerg. Topics Comput. Intell.*, vol. 6, no. 3, pp. 568–579, Jun. 2022, doi: 10.1109/TETCI.2022.3157026.

[49] A. Benjamin and A. Q. H. Badar, "Reinforcement learning based cost-effective smart home energy management," in *Proc. IEEE 3rd Int. Conf. Sustain. Energy Future Electric Transp. (SEFET)*, Bhubaneswar, India, Aug. 2023, pp. 1–5, doi: 10.1109/sefet57834.2023.10245183.

[50] Z. Rostmnezhad and L. Dessaint, "Power management in smart buildings using reinforcement learning," in *Proc. IEEE Power Energy Soc. Innov. Smart Grid Technol. Conf. (ISGT)*, Washington, DC, USA, Jan. 2023, pp. 1–5, doi: 10.1109/ISGT51731.2023.10066398.

[51] C. Hau, K. K. Radhakrishnan, J. Siu, and S. K. Panda, "Reinforcement learning based energy management algorithm for energy trading and contingency reserve application in a microgrid," in *Proc. IEEE PES Innov. Smart Grid Technol. Eur. (ISGT-Europe)*, The Hague, Netherlands, Oct. 2020, pp. 1005–1009, doi: 10.1109/ISGT-Europe47291.2020.9248752.

[52] T. Xu, T. Chen, and C. Gao, "Intelligent energy management strategy with internal pricing sensitivity for residential customers based on reinforcement learning and Internet-of-Things," in *Proc. IEEE 6th Conf. Energy Internet Energy Syst. Integr. (EI2)*, Chengdu, China, Nov. 2022, pp. 1760–1765, doi: 10.1109/EI256261.2022.10116749.

[53] K. Shouryadhar, C.-N. Chen, and Y.-C. Chen, "An enhanced reinforcement learning based approach of energy management optimization for microgrids," in *Proc. IET Int. Conf. Eng. Technol. Appl.*, Changhua, Taiwan, Oct. 2022, pp. 1–2, doi: 10.1109/IET-ICETA56553.2022.9971705.

[54] S.-H. Hong and H.-S. Lee, "Robust energy management system with safe reinforcement learning using short-horizon forecasts," *IEEE Trans. Smart Grid*, vol. 14, no. 3, pp. 2485–2488, May 2023, doi: 10.1109/TSG.2023.3240588.

[55] Y. Wang, Z. Yang, L. Dong, S. Huang, and W. Zhou, "Energy management of integrated energy system based on Stackelberg game and deep reinforcement learning," in *Proc. IEEE 4th Conf. Energy Internet Energy Syst. Integr. (EI2)*, Wuhan, China, Oct. 2020, pp. 2645–2651, doi: 10.1109/EI250167.2020.9346692.

[56] L. Liu, J. Zhu, J. Chen, and H. Ye, "Deep reinforcement learning for stochastic dynamic microgrid energy management," in *Proc. IEEE 4th Int. Electr. Energy Conf. (CIEEC)*, Wuhan, China, May 2021, pp. 1–6, doi: 10.1109/CIEEC50170.2021.9511049.

[57] H. Li, Z. Wan, and H. He, "A deep reinforcement learning based approach for home energy management system," in *Proc. IEEE Power Energy Soc. Innov. Smart Grid Technol. Conf. (ISGT)*, Washington, DC, USA, Feb. 2020, pp. 1–5, doi: 10.1109/ISGT45199.2020.9087647.

[58] X. Weiss, Q. Xu, and L. Nordström, "Energy management of smart homes with electric vehicles using deep reinforcement learning," in *Proc. 24th Eur. Conf. Power Electron. Appl.*, Hanover, Germany, Sep. 2022, pp. 1–9.

[59] Z. Tchir, M. Z. Reformat, and P. Musilek, "Home energy management with V2X capability using reinforcement learning," in *Proc. IEEE Conf. Artif. Intell. (CAI)*, Santa Clara, CA, USA, Jun. 2023, pp. 89–91, doi: 10.1109/cai54212.2023.00046.

[60] S. Lee and D.-H. Choi, "Federated reinforcement learning for energy management of multiple smart homes with distributed energy resources," *IEEE Trans. Ind. Informat.*, vol. 18, no. 1, pp. 488–497, Jan. 2022, doi: 10.1109/TII.2020.3035451.

[61] A. Kahraman and G. Yang, "Home energy management system based on deep reinforcement learning algorithms," in *Proc. IEEE PES Innov. Smart Grid Technol. Conf. Eur. (ISGT-Europe)*, Novi Sad, Serbia, Oct. 2022, pp. 1–5, doi: 10.1109/ISGT-Europe54678.2022.9960575.

[62] C. Corchero and M. Sanmarti, "Vehicle-to-everything (V2X): Benefits and barriers," in *Proc. 15th Int. Conf. Eur. Energy Market (EEM)*, Lodz, Poland, Jun. 2018, pp. 1–4, doi: 10.1109/EEM.2018.8469875.

[63] A. Alsharif, A. A. Ahmed, M. M. Khaleel, A. S. D. Alarga, Omer. S. M. Jomah, and I. Imbayah, "Comprehensive state-of-the-art of vehicle-to-grid technology," in *Proc. IEEE 3rd Int. Maghreb Meeting Conf. Sci. Techn. Autom. Control Comput. Eng.*, Benghazi, Libya, May 2023, pp. 530–534, doi: 10.1109/MI-STA57575.2023.10169116.

[64] R. Khezri, D. Steen, and L. A. Tuan, "A review on implementation of vehicle to everything (V2X): Benefits, barriers and measures," in *Proc. IEEE PES Innov. Smart Grid Technol. Conf. Eur. (ISGT-Europe)*, Novi Sad, Serbia, Oct. 2022, pp. 1–6, doi: 10.1109/ISGT-Europe54678.2022.9960673.

[65] R. Roche, F. Berthold, F. Gao, F. Wang, A. Ravey, and S. Williamson, "A model and strategy to improve smart home energy resilience during outages using vehicle-to-home," in *Proc. IEEE Int. Electr. Vehicle Conf. (IEVC)*, Florence, Italy, Dec. 2014, pp. 1–6, doi: 10.1109/IEVC.2014.7056106.

[66] U. ur Rehman and M. Riaz, "Vehicle to grid system for load and frequency management in smart grid," in *Proc. Int. Conf. Open Source Syst. Technol. (ICOSST)*, Lahore, Pakistan, Dec. 2017, pp. 73–78, doi: 10.1109/ICOSST.2017.8279008.

[67] H. Turker and S. Bacha, "Optimal minimization of plug-in electric vehicle charging cost with vehicle-to-home and vehicle-to-grid concepts," *IEEE Trans. Veh. Technol.*, vol. 67, no. 11, pp. 10281–10292, Nov. 2018, doi: 10.1109/TVT.2018.2867428.

[68] M. S. Hashim, J. Y. Yong, V. K. Ramachandaramurthy, K. M. Tan, and M. Tariq, "Coordinated vehicle-to-grid scheduling to minimize grid load variance," in *Proc. Int. Conf. Electr., Electron. Comput. Eng. (UPCON)*, Aligarh, India, Nov. 2019, pp. 1–6, doi: 10.1109/UPCON47278.2019.8980281.

[69] R. Hemmati, H. Mehrjerdi, N. A. Al-Emadi, and E. Rakhshani, "Mutual vehicle-to-home and vehicle-to-grid operation considering solar-load uncertainty," in *Proc. 2nd Int. Conf. Smart Grid Renew. Energy (SGRE)*, Doha, Qatar, Nov. 2019, pp. 1–4, doi: 10.1109/SGRE46976.2019.9020685.

[70] A. A. S. Mohamed, A. M. Zulkefli, S. M. V. Sagar, U. M. Gajjarapu, A. Thomas, and V. Bhavaraju, "Smart vehicle-to-home (V2H) platform enabled home energy management system (HEMS) for backup supply," in *Proc. IEEE Green Technol. Conf. (GreenTech)*, Denver, CO, USA, Apr. 2023, pp. 73–77, doi: 10.1109/GreenTech56823.2023.10173790.

[71] J. Einolander, A. Kiviaho, and R. Lahdelma, "Household electricity cost optimization with vehicle-to-home technology and mixed-integer linear programming," in *Proc. Int. Conf. Future Energy Solutions (FES)*, Vaasa, Finland, Jun. 2023, pp. 1–5, doi: 10.1109/fes57669.2023.10182713.

[72] W. Tushar, T. K. Saha, C. Yuen, D. Smith, and H. V. Poor, "Peer-to-peer trading in electricity networks: An overview," *IEEE Trans. Smart Grid*, vol. 11, no. 4, pp. 3185–3200, Jul. 2020, doi: 10.1109/TSG.2020.2969657.

[73] A. Takkabutra, C. Chupong, and B. Plangklang, "Peer-to-peer energy trading market: A review on current trends, challenges and opportunities for Thailand," in *Proc. 18th Int. Conf. Electr. Engineering/Electronics, Comput., Telecommun. Inf. Technol. (ECTI-CON)*, Chiang Mai, Thailand, May 2021, pp. 1076–1079, doi: 10.1109/ECTI-CON51831.2021.9454828.

[74] K. Anoh, S. Maharjan, A. Ikpehai, Y. Zhang, and B. Adebisi, "Energy peer-to-peer trading in virtual microgrids in smart grids: A game-theoretic approach," *IEEE Trans. Smart Grid*, vol. 11, no. 2, pp. 1264–1275, Mar. 2020, doi: 10.1109/TSG.2019.2934830.

[75] R.-J. Meinke, H. Sun, and J. Jiang, "Optimising demand and bid matching in a peer-to-peer energy trading model," in *Proc. IEEE Int. Conf. Commun. (ICC)*, Dublin, Ireland, Jun. 2020, pp. 1–6, doi: 10.1109/ICC40277.2020.9148652.

[76] R. R. Trivedi, C. P. Barala, P. Mathuria, R. Bhakar, and S. Sharma, "Peer-to-peer energy trading: Energy pricing using game theory models," in *Proc. IEEE IAS Global Conf. Renew. Energy Hydrogen Technol.*, Male, Maldives, Mar. 2023, pp. 1–6, doi: 10.1109/GlobConHT56829.2023.10087444.

[77] D. S. Schiera, C. De Vizia, A. Zarri, R. Borchiellini, A. Lanzini, E. Patti, and L. Bottaccioli, "Modelling and techno-economic analysis of peer-to-peer electricity trading systems in the context of energy communities," in *Proc. IEEE Int. Conf. Environ. Electr. Eng. IEEE Ind. Commercial Power Syst. Eur.*, Prague, Czech Republic, Jun. 2022, pp. 1–6, doi: 10.1109/EEEIC/ICPSEurope54979.2022.9854537.

[78] W. Sarapan, N. Boonrakchat, A. Paudel, T. Booraksa, P. Boonraksa, and B. Marungsri, "Optimal peer-to-peer energy trading by applying blockchain to islanded microgrid considering V2G," in *Proc. 19th Int. Conf. Electr. Engineering/Electronics, Comput., Telecommun. Inf. Technol. (ECTI-CON)*, Prachuap Khiri Khan, Thailand, May 2022, pp. 1–4, doi: 10.1109/ECTI-CON54298.2022.9795559.

[79] H. Zhu, K. Ouahada, and A. M. Abu-Mahfouz, "Peer-to-peer energy trading in smart energy communities: A Lyapunov-based energy control and trading system," *IEEE Access*, vol. 10, pp. 42916–42932, 2022, doi: 10.1109/ACCESS.2022.3167828.

[80] J. Li, Y. Ye, D. Papadaskalopoulos, and G. Strbac, "Computationally efficient pricing and benefit distribution mechanisms for incentivizing stable peer-to-peer energy trading," *IEEE Internet Things J.*, vol. 8, no. 2, pp. 734–749, Jan. 2021, doi: 10.1109/JIOT.2020.3007196.

[81] J. Pankiraj, A. Yassine, and S. Choudhury, "Incentive-based peer-to-peer distributed energy trading in smart grid systems," in *Proc. Int. Symp. Netw., Comput. Commun. (ISNCC)*, Montreal, QC, Canada, Oct. 2020, pp. 1–6, doi: 10.1109/ISNCC49221.2020.9297278.

[82] A. Paudel and G. H. Beng, "A hierarchical peer-to-peer energy trading in community microgrid distribution systems," in *Proc. IEEE Power Energy Soc. Gen. Meeting (PESGM)*, Portland, OR, USA, Aug. 2018, pp. 1–5, doi: 10.1109/PESGM.2018.8586168.

[83] B. P. Hayes, S. Thakur, and E. Barrett, "Design of an open-source laboratory demonstrator for peer-to-peer trading in local energy markets," in *Proc. 12th Medit. Conf. Power Gener., Transmiss., Distrib. Energy Convers. (MEDPOWER)*, Nov. 2020, pp. 342–347, doi: 10.1049/icp.2021.1226.

[84] M. Sanayha and P. Vateekul, "Model-based approach on multi-agent deep reinforcement learning with multiple clusters for peer-to-peer energy trading," *IEEE Access*, vol. 10, pp. 127882–127893, 2022, doi: 10.1109/ACCESS.2022.3224460.

[85] X. Xu, Y. Jia, Y. Xu, Z. Xu, S. Chai, and C. S. Lai, "A multi-agent reinforcement learning-based data-driven method for home energy management," *IEEE Trans. Smart Grid*, vol. 11, no. 4, pp. 3201–3211, Jul. 2020, doi: 10.1109/TSG.2020.2971427.

[86] S. Zhou, Z. Hu, W. Gu, M. Jiang, and X.-P. Zhang, "Artificial intelligence based smart energy community management: A reinforcement learning approach," *CSEE J. Power Energy Syst.*, vol. 5, no. 1, pp. 1–10, Mar. 2019, doi: 10.17775/CSEEJPES.2018.00840.

[87] H. S. V. S. K. Nunna, A. Sesetti, A. K. Rathore, and S. Doolla, "Multiagent-based energy trading platform for energy storage systems in distribution systems with interconnected microgrids," *IEEE Trans. Ind. Appl.*, vol. 56, no. 3, pp. 3207–3217, May 2020, doi: 10.1109/TIA.2020.2979782.

[88] T. Chen, S. Bu, X. Liu, J. Kang, F. R. Yu, and Z. Han, "Peer-to-peer energy trading and energy conversion in interconnected multi-energy microgrids using multi-agent deep reinforcement learning," *IEEE Trans. Smart Grid*, vol. 13, no. 1, pp. 715–727, Jan. 2022, doi: 10.1109/TSG.2021.3124465.

[89] M. Sadeghi and M. Erol-Kantarci, "Deep reinforcement learning based coalition formation for energy trading in smart grid," in *Proc. IEEE 4th 5G World Forum (5GWF)*, Montreal, QC, Canada, Oct. 2021, pp. 200–205, doi: 10.1109/5GWF52925.2021.00042.

[90] Y. Ye, Y. Tang, H. Wang, X.-P. Zhang, and G. Strbac, "A scalable privacy-preserving multi-agent deep reinforcement learning approach for large-scale peer-to-peer transactive energy trading," *IEEE Trans. Smart Grid*, vol. 12, no. 6, pp. 5185–5200, Nov. 2021, doi: 10.1109/TSG.2021.3103917.

[91] Y. Wu, T. Zhao, H. Yan, M. Liu, and N. Liu, "Hierarchical hybrid multi-agent deep reinforcement learning for peer-to-peer energy trading among multiple heterogeneous microgrids," *IEEE Trans. Smart Grid*, vol. 14, no. 6, pp. 4649–4665, Nov. 2023, doi: 10.1109/TSG.2023.3250321.

[92] Y. Liu, D. Zhang, and H. B. Gooi, "Optimization strategy based on deep reinforcement learning for home energy management," *CSEE J. Power Energy Syst.*, vol. 6, no. 3, pp. 572–582, Sep. 2020, doi: 10.17775/CSEEJPES.2019.02890.

[93] H. H. Goh, Y. Huang, C. S. Lim, D. Zhang, H. Liu, W. Dai, T. A. Kurniawan, and S. Rahman, "An assessment of multistage reward function design for deep reinforcement learning-based microgrid energy management," *IEEE Trans. Smart Grid*, vol. 13, no. 6, pp. 4300–4311, Nov. 2022, doi: 10.1109/TSG.2022.3179567.

[94] N. Kodama, T. Harada, and K. Miyazaki, "Home energy management algorithm based on deep reinforcement learning using multistep prediction," *IEEE Access*, vol. 9, pp. 153108–153115, 2021, doi: 10.1109/ACCESS.2021.3126365.

[95] G. Wei, M. Chi, Z.-W. Liu, M. Ge, C. Li, and X. Liu, "Deep reinforcement learning for real-time energy management in smart home," *IEEE Syst. J.*, vol. 17, no. 2, pp. 2489–2499, Jun. 2023, doi: 10.1109/JSYST.2023.3247592.

[96] A. Masadeh, Z. Wang, and A. E. Kamal, "Reinforcement learning exploration algorithms for energy harvesting communications systems," in *Proc. IEEE Int. Conf. Commun. (ICC)*, Kansas City, MO, USA, May 2018, pp. 1–6, doi: 10.1109/ICC.2018.8422710.

[97] B. Chen, W. Deng, and J. Du, "Noisy softmax: Improving the generalization ability of DCNN via postponing the early softmax saturation," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Honolulu, HI, USA, Jul. 2017, pp. 4021–4030, doi: 10.1109/CVPR.2017.428.

[98] S. Thadikamalla and P. Joshi, "Exploration strategies in adaptive traffic signal control: A comparative analysis of Epsilon–Greedy, UCB, softmax, and Thomson sampling," in *Proc. 7th Int. Symp. Innov. Approaches Smart Technol. (ISAS)*, Nov. 2023, pp. 1–8, doi: 10.1109/isas60782.2023.10391701.

[99] F. Rasheed, K. A. Yau, R. Md. Noor, C. Wu, and Y.-C. Low, "Deep reinforcement learning for traffic signal control: A review," *IEEE Access*, vol. 8, pp. 208016–208044, 2020, doi: 10.1109/ACCESS.2020.3034141.

[100] M. Elsayed, A. Badawy, A. E. Shafie, A. Mohamed, and T. Khattab, "A deep reinforcement learning framework for data compression in uplink NOMA-SWIPT systems," *IEEE Internet Things J.*, vol. 9, no. 14, pp. 11656–11674, Jul. 2022, doi: 10.1109/JIOT.2021.3131524.

[101] W. Jia, J. Li, and Y. Zhao, "DQN algorithm based on target value network parameter dynamic update," in *Proc. IEEE 4th Int. Conf. Comput. Commun. Eng. Technol. (CCET)*, Beijing, China, Aug. 2021, pp. 285–289, doi: 10.1109/CCET52649.2021.9544323.

[102] I. J. Sledge and J. C. Príncipe, "Balancing exploration and exploitation in reinforcement learning using a value of information criterion," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, New Orleans, LA, USA, Mar. 2017, pp. 2816–2820, doi: 10.1109/ICASSP.2017.7952670.

[103] Y. Li, Y. Xu, G. Li, Y. Gong, X. Liu, H. Wang, and W. Li, "Dynamic spectrum anti-jamming access with fast convergence: A labeled deep reinforcement learning approach," *IEEE Trans. Inf. Forensics Security*, vol. 18, pp. 5447–5458, 2023, doi: 10.1109/tifs.2023.3307950.

[104] S. Bock and M. Weiß, "A proof of local convergence for the Adam optimizer," in *Proc. Int. Joint Conf. Neural Netw. (IJCNN)*, Budapest, Hungary, Jul. 2019, pp. 1–8, doi: 10.1109/IJCNN.2019.8852239.

[105] V. Mnih, K. Kavukcuoglu, D. Silver, A. A. Rusu, J. Veness, M. G. Bellemare, A. Graves, M. Riedmiller, A. K. Fidjeland, G. Ostrovski, S. Petersen, C. Beattie, A. Sadik, I. Antonoglou, H. King, D. Kumaran, D. Wierstra, S. Legg, and D. Hassabis, "Human-level control through deep reinforcement learning," *Nature*, vol. 518, no. 7540, pp. 529–533, Feb. 2015, doi: 10.1038/nature14236.

[106] K. Arshad, R. F. Ali, A. Muneer, I. A. Aziz, S. Naseer, N. S. Khan, and S. M. Taib, "Deep reinforcement learning for anomaly detection: A systematic review," *IEEE Access*, vol. 10, pp. 124017–124035, 2022, doi: 10.1109/ACCESS.2022.3224023.

[107] K. Yu, K. Jin, and X. Deng, "Review of deep reinforcement learning," in *Proc. IEEE 5th Adv. Inf. Manage., Communicates, Electron. Autom. Control Conf. (IMCEC)*, vol. 5. Chongqing, China, Dec. 2022, pp. 41–48, doi: 10.1109/IMCEC55388.2022.10020015.

[108] R. Zhang, C. Pan, Y. Wang, Y. Yao, and X. Li, "Federated deep reinforcement learning for multimedia task offloading and resource allocation in MEC networks," *IEICE Trans. Commun.*, early access, Jan. 30, 2024, doi: 10.23919/TRANSCOM.2023EBP3116.

[109] S. Li, J. Chen, S. Liu, C. Zhu, G. Tian, and Y. Liu, "MCMC: Multi-constrained model compression via one-stage envelope reinforcement learning," *IEEE Trans. Neural Netw. Learn. Syst.*, early access, Jan. 30, 2024, doi: 10.1109/TNNLS.2024.3353763.

[110] J. Kang, J. Chen, M. Xu, Z. Xiong, Y. Jiao, L. Han, D. Niyato, Y. Tong, and S. Xie, "UAV-assisted dynamic avatar task migration for vehicular metaverse services: A multi-agent deep reinforcement learning approach," *IEEE/CAA J. Autom. Sinica*, vol. 11, no. 2, pp. 430–445, Feb. 2024, doi: 10.1109/jas.2023.123993.

[111] S. Liu et al., "Fight against intelligent reactive jammer in MEC networks: A hierarchical reinforcement learning based hybrid hidden strategy," *IEEE Wireless Commun. Lett.*, early access, Jan. 30, 2024, doi: 10.1109/LWC.2024.3360235.

[112] X. Wang, X. Wan, H. Ji, H. Hu, Z. Chen, and Y. Xiao, "Joint UAV deployment and user scheduling for wireless powered wearable networks," *IEEE Internet Things J.*, early access, Jan. 30, 2024, doi: 10.1109/JIOT.2024.3360078.

[113] C. Wang, D. Deng, L. Xu, and W. Wang, "Resource scheduling based on deep reinforcement learning in UAV assisted emergency communication networks," *IEEE Trans. Commun.*, vol. 70, no. 6, pp. 3834–3848, Jun. 2022, doi: 10.1109/TCOMM.2022.3170458.

[114] S. Baik, "Detecting digital ad fraud by active reinforcement learning," in *Proc. 24st Asia–Pacific Netw. Oper. Manag. Symp.*, Sejong, South Korea, 2023, pp. 338–340.

[115] A. N. Njoya, V. L. T. Ngongag, F. Tchakounté, M. Atemkeng, and C. Fachkha, "Characterizing mobile money phishing using reinforcement learning," *IEEE Access*, vol. 11, pp. 103839–103862, 2023, doi: 10.1109/ACCESS.2023.3317692.

[116] K. Zhu, N. Zhang, W. Ding, and C. Jiang, "An adaptive heterogeneous credit card fraud detection model based on deep reinforcement training subset selection," *IEEE Trans. Artif. Intell.*, early access, Jan. 30, 2024, doi: 10.1109/TAI.2024.3359568.

[117] Z. Yi, X. Cao, Z. Chen, and S. Li, "Artificial intelligence in accounting and finance: Challenges and opportunities," *IEEE Access*, vol. 11, pp. 129100–129123, 2023, doi: 10.1109/ACCESS.2023.3333389.

[118] S. Dangi, K. Ghanshala, and S. Sharma, "Live memory forensics integrated reinforcement learning model for optimized demand forecasting in autonomous supply chain management system," in *Proc. 14th Int. Conf. Comput. Commun. Netw. Technol. (ICCCNT)*, Delhi, India, Jul. 2023, pp. 1–6, doi: 10.1109/icccnt56998.2023.10307291.

[119] R. Mishra, M. Kaif, A. Raj, and R. Deep, "Block chain enabled farmer centric supply chain management using reinforcement learning," in *Proc. 14th Int. Conf. Comput. Commun. Netw. Technol. (ICCCNT)*, Delhi, India, Jul. 2023, pp. 1–7, doi: 10.1109/icccnt56998.2023.10307187.

[120] J. Liu et al., "Reinforcement learning-based high-speed path following control for autonomous vehicles," *IEEE Trans. Veh. Technol.*, early access, Jan. 11, 2024, doi: 10.1109/TVT.2024.3352543.

[121] X. Li et al., "Progression cognition reinforcement learning with prioritized experience for multi-vehicle pursuit," *IEEE Trans. Intell. Transp. Syst.*, early access, Jan. 26, 2024, doi: 10.1109/TITS.2024.3354196.

[122] T. Liu, Y. Yang, W. Xiao, X. Tang, and M. Yin, "A comparative analysis of deep reinforcement learning-enabled freeway decision-making for automated vehicles," *IEEE Access*, vol. 12, pp. 24090–24103, 2024, doi: 10.1109/ACCESS.2024.3358424.

[123] C. Ding, I. W.-H. Ho, E. Chung, and T. Fan, "V2X and deep reinforcement learning-aided mobility-aware lane changing for emergency vehicle preemption in connected autonomous transport systems," *IEEE Trans. Intell. Transp. Syst.*, early access, Jan. 17, 2024, doi: 10.1109/TITS.2024.3350334.

[124] J. Wu, Z. Huang, and C. Lv, "Transformer-based traffic-aware predictive energy management of a fuel cell electric vehicle," *IEEE Trans. Veh. Technol.*, early access, Jan. 19, 2024, doi: 10.1109/TVT.2024.3355895.

[125] Y. Zhang, M. Yue, J. Wang, and S. Yoo, "Multi-agent graph-attention deep reinforcement learning for post-contingency grid emergency voltage control," *IEEE Trans. Neural Netw. Learn. Syst.*, early access, Jan. 25, 2024, doi: 10.1109/TNNLS.2023.3341334.

[126] Y. Han, J. Wu, H. Chen, F. Si, Z. Cao, and Q. Zhao, "Enhancing grid-interactive buildings demand response: Sequential update-based multi-agent deep reinforcement learning approach," *IEEE Internet Things J.*, early access, Jan. 22, 2024, doi: 10.1109/JIOT.2024.3357109.

[127] J. Zheng, Z.-T. Liang, Y. Li, Z. Li, and Q.-H. Wu, "Multi-agent reinforcement learning with privacy preservation for continuous double auction-based P2P energy trading," *IEEE Trans. Ind. Informat.*, early access, Jan. 17, 2024, doi: 10.1109/TII.2023.3348823.

[128] Z. Wen, W. Zhang, and M. Qian, "A comprehensive review of deep reinforcement learning for object detection," in *Proc. Int. Symp. Artif. Intell. its Appl. Media (ISAIAM)*, Xi'an, China, May 2021, pp. 146–150, doi: 10.1109/ISAIAM53259.2021.00038.

[129] P. Kumar and K. B. Ali, "Intelligent traffic system using vehicle to vehicle (V2V) & vehicle to infrastructure (V2I) communication based on wireless access in vehicular environments (WAVE) std," in *Proc. 10th Int. Conf. Rel., INFOCOM Technol. Optim.*, Noida, India, Oct. 2022, pp. 1–5, doi: 10.1109/ICRITO56286.2022.9964590.

[130] J. Liu, Z. Wang, and L. Zhang, "Event-triggered vehicle-following control for connected and automated vehicles under nonideal vehicle-to-vehicle communications," in *Proc. IEEE Intell. Vehicles Symp. (IV)*, Nagoya, Japan, Jul. 2021, pp. 342–347, doi: 10.1109/IV48863.2021.9575727.

[131] H. Zhao, H. Wu, N. Lu, X. Zhan, E. Xu, and Q. Yuan, "Lane changing in a vehicle-to-everything environment: Research on a vehicle lane-changing model in the tunnel area by considering the influence of brightness and noise under a vehicle-to-everything environment," *IEEE Intell. Transp. Syst. Mag.*, vol. 15, no. 2, pp. 225–237, Mar. 2023, doi: 10.1109/MITS.2022.3197462.

[132] M. Hu, J. Li, Y. Bian, J. Wang, B. Xu, and Y. Zhu, "Distributed coordinated brake control for longitudinal collision avoidance of multiple connected automated vehicles," *IEEE Trans. Intell. Vehicles*, vol. 8, no. 1, pp. 745–755, Jan. 2023, doi: 10.1109/TIV.2022.3197951.

[133] P. Zhang, D. Tian, J. Zhou, X. Duan, Z. Sheng, D. Zhao, and D. Cao, "Joint optimization of platoon control and resource scheduling in cooperative vehicle-infrastructure system," *IEEE Trans. Intell. Vehicles*, vol. 8, no. 6, pp. 3629–3646, Jun. 2023, doi: 10.1109/TIV.2023.3265866.

[134] R. Q. Malik, Khairun. N. Ramli, Z. H. Kareem, M. I. Habelalmatee, and H. Abbas, "A review on vehicle-to-infrastructure communication system: Requirement and applications," in *Proc. 3rd Int. Conf. Eng. Technol. Appl. (IICETA)*, Najaf, Iraq, Sep. 2020, pp. 159–163, doi: 10.1109/IICETA50496.2020.9318825.

[135] M. Shawki and M. S. Darweesh, "Collision probability computation for road intersections based on vehicle to infrastructure communication," in *Proc. 32nd Int. Conf. Microelectron. (ICM)*, Aqaba, Jordan, Dec. 2020, pp. 1–4, doi: 10.1109/ICM50269.2020.9331802.

[136] C. Teng, G. Ligang, W. Zexu, S. Qin, and N. Hao, "Car following model based on driving risk field for vehicle infrastructure cooperation," in *Proc. 6th CAA Int. Conf. Veh. Control Intell. (CVCI)*, Nanjing, China, Oct. 2022, pp. 1–6, doi: 10.1109/CVCI56766.2022.9964837.

[137] O. Dokur and S. Katkoori, "Vehicle-to-Infrastructure based algorithms for traffic light detection, red light violation, and wrong-way entry applications," in *Proc. IEEE Int. Symp. Smart Electron. Syst. (iSES)*, Warangal, India, Dec. 2022, pp. 25–30, doi: 10.1109/iSES54909.2022.00158.

[138] R. Barnett, C. M. Hume, and A. Taylor, "Comparison of connected automated vehicle to pedestrian interaction systems to reduce vehicle waiting times," in *Proc. Syst. Inf. Eng. Design Symp. (SIEDS)*, Charlottesville, VA, USA, Apr. 2021, pp. 1–5, doi: 10.1109/SIEDS52267.2021.9483718.

[139] P. Cai, J. He, and Y. Li, "Hybrid cooperative intersection management for connected automated vehicles and pedestrians," *J. Intell. Connected Vehicles*, vol. 6, no. 2, pp. 91–101, Jun. 2023, doi: 10.26599/jicv.2023.9210008.

[140] Y. Zhang, X. Wang, K. Zhuo, W. Jiao, and W. Yang, "Research on pedestrian vehicle collision warning based on path prediction," in *Proc. 7th Int. Conf. Transp. Inf. Saf. (ICTIS)*, Xi'an, China, Aug. 2023, pp. 1267–1272, doi: 10.1109/ictis60134.2023.10243767.

[141] M. Golchoubian, M. Ghafurian, K. Dautenhahn, and N. L. Azad, "Pedestrian trajectory prediction in pedestrian-vehicle mixed environments: A systematic review," *IEEE Trans. Intell. Transp. Syst.*, vol. 24, no. 11, pp. 11544–11567, Nov. 2023, doi: 10.1109/tits.2023.3291196.

[142] L. Gao, C. Wu, T. Yoshinaga, X. Chen, and Y. Ji, "Multi-channel blockchain scheme for Internet of Vehicles," *IEEE Open J. Comput. Soc.*, vol. 2, pp. 192–203, 2021, doi: 10.1109/OJCS.2021.3070714.

[143] X. Wang, Z. Ning, X. Hu, L. Wang, B. Hu, J. Cheng, and V. C. M. Leung, "Optimizing content dissemination for real-time traffic management in large-scale Internet of Vehicle systems," *IEEE Trans. Veh. Technol.*, vol. 68, no. 2, pp. 1093–1105, Feb. 2019, doi: 10.1109/TVT.2018.2886010.

METE YAVUZ received the B.Sc. and M.Sc. degrees in mechatronics engineering from Istanbul Okan University, İstanbul, Turkey, in 2021 and 2023, respectively.

His current research interests include energy management systems, reinforcement learning applications in energy management, optimization of energy flow and appliance scheduling in smart grids and smart homes, peer-to-peer energy trading implementations, integration of vehicle-to-x technologies, smart energy management in electric vehicles, and machine learning applications in autonomous driving.

ÖMER CIHAN KIVANÇ (Senior Member, IEEE) received the B.Sc. degree in electrical engineering from Istanbul Technical University, İstanbul, Turkey, in 2011, and the Ph.D. degree in electrical engineering from Istanbul Okan University, İstanbul, in 2016.

In 2011, he joined Power Electronics and Energy Conversion (PEEC) Laboratory, a research group, Istanbul Okan University, where he was a Graduate Research and Teaching Assistant with the Department of Electrical and Electronics Engineering. Between 2016 and 2021, he was an Assistant Professor with the Department of Electrical and Electronics Engineering. Since 2021, he has been an Associate Professor with the Department of Electrical and Electronics Engineering, Istanbul Okan University. He is also an Electrical Engineer with Mekatro R&D Company. His research interests include sensorless motor control strategies, electric drive systems, electric vehicles, robotics, and energy management systems.