

INTELLIGENT ENERGY TRADING AND DYNAMIC LOAD BALANCING IN DISTRIBUTED SOLAR PV NETWORKS USING REINFORCEMENT LEARNING

Faisal Aburub
Ibrahim Mohammad
Asokan Vasudevan
Yang Chunwen
Ulug'bek Sobirov
Linara Karimova ¹

Received 13.05.2026.
Revised 20.08.2026.
Accepted 16.09.2026.

Keywords:

Peer-to-peer energy trading, Multi-agent reinforcement learning, Dynamic load balancing, Distributed photovoltaic networks, Managing local power supply

ABSTRACT

The research will focus on the design of an intelligent system combining peer-to-peer trading and dynamic load balancing in distributed solar networks with the use of multi-agent deep reinforcement learning. The presented solution will include two layers – the first one will be a peer-to-peer marketplace operating under the multi-agent Proximal Policy Optimization algorithm with a centralized learning process and decentralized actions which optimize energy trades between energy producers and consumers. The second layer will be an adaptive load balancing controller which will adapt the proposed transactions according to the technological limitations of the grid. The architecture was developed and applied using real-world data for a 32-node power grid that had 18 producers and consumers having photovoltaic and battery storage in the Abdoun Amman area, which is the capital of Jordan, between January and September 2025. According to the results, the MAPPO algorithm proposed in this study performed better than the standalone PPO and the central DQN algorithms by 23.5% and 51.4%, respectively, with a cumulative reward of 4,872. The new setup resulted in an increase in daily producer-consumer profit of 98.6% to 2.84 Jordanian Dinars. A Gini Coefficient value of 0.28 shows the fairly equitable benefit distribution among participants. In terms of a technical approach, the suggested method has led to a reduction in voltage deviation to 1.82 percent, i.e., 42.2 percent lower than the two-way auction approach. The load balancing index also has been improved to 1.14 by 17.4 percent. Peer-to-peer monthly trading volume for the proposed case was estimated to be 14.82 MWh. Through sensitivity analysis, it is shown that the availability of battery energy storage systems plays a significant role in the success of the market and has the potential to lower the payback period from 8.5 to 5.2 years.



¹ Corresponding author: Linara Karimova
Email: kozina.n.vladimirov3@gmail.com

1. INTRODUCTION

As noted by Abid et al. (2024) the change has brought about numerous challenges for the existing structure of the electricity network. Solar energy's fluctuating and unpredictable nature has thrown the conventional equation of supply and demand into disarray. Also, Vamvakas et al. (2023) stated that the emergence of distributed solar networks, where customers are also generating their own electricity, has created more complexities for the electricity network. Jordan, a country rich in solar energy but poor in fossil fuel energy resources, has made many efforts for establishing distributed solar plants (Buzrukov et al., 2025). But the effective management of such distributed resources is almost not possible without intelligent decision-making systems (Veerasamy et al., 2024).

Peer-to-peer energy trading according to Wu et al. (2023) is one of those ideas that have recently come up as a way of solving the problem of managing distributed networks. Here, users are able to offer their excess energy to other users who are in close proximity, without having to go through any intermediaries such as big electricity distribution firms. Not only does this idea help improve efficiency on the whole, but Michailidis et al. (2026) showed that it makes investing in solar panels worthwhile economically. But there are quite a few technical problems to overcome when implementing such a market.

Dynamic Load Balancing as mentioned by Huang et al., (2024) is the second crucial pillar in the current management of renewable power grids. In a distributed grid, the unequal distribution of loads among nodes may have negative effects on the overall stability of the grid and will increase transmission losses (Nazori et al., 2024). Load balancing algorithms, based on static rules and preset thresholds, do not perform effectively in a highly dynamic and uncertain environment, such as a solar power grid. Therefore, Harrold et al. (2022) confirmed that it is necessary to create a brain that will continuously learn from the environment and adjust the strategy.

Jamil et al. (2021) defined an intelligent algorithm based on reinforcement learning, as a new branch of artificial intelligence, provides a smart solution to this problem. Contrary to the supervised learning algorithms that require labeled historical data, reinforcement learning finds out the optimal policy through direct interaction with the environment and rewards/penalties. In this way, an intelligent agent can test thousands of production/consumption situations and figure out how much energy should be sold at what price to what consumer at a specific time. This is what is required in distributed electricity markets.

There are several reasons why the research has to be done in Jordan. First, Jordan is considered one of the pioneers in terms of using solar energy in the Middle East and North Africa region and the use of photovoltaic system in Jordan increases day by day. Secondly, there are a lot of problems with Jordan's electricity grid such as high transmission and distribution losses, dependence on imported energy sources and lack of capacity of power plants. Thirdly, the regulation in Jordan made great efforts recently to liberalize electricity market and enable direct trading of energy.

There are three perspectives from which the gap in the literature can be investigated. First, Foruzan et al. (2018) stated that the majority of studies regarding peer-to-peer energy trading were conducted using classical algorithms that could not learn adaptively. Second, according to Tomin et al. (2022) all of the simulations were conducted using fictional data about the networks, and few studies used real data from active distribution networks. Third, Gautam (2023) showed that the simultaneous implementation of two functions of energy trading and load balancing in the framework of reinforcement learning is quite a new topic.

The significance of this study is based on the combination of three major trends in the modern world: the expansion of distributed renewable energy generation, the digitization of electrical grid infrastructure, and the development of sophisticated algorithms of artificial intelligence. As such, Jordan appears as the perfect setting to study this challenge since it is situated where the three aforementioned trends cross. Through analysis of the actual data collected from the distributed solar power grid in Amman, which is the capital of Jordan, and by applying a deep reinforcement learning agent, this study attempts to resolve the following problem: How to maximize the economic efficiency of micro-generation, the sustainability of the distribution grid, and the efficiency of the whole system in a peer-to-peer electricity market?

2. LITERATURE REVIEW

2.1 Peer-to-Peer Energy Trading in Distributed Networks

Energy trading through Peer-to-Peer technology has been an area of much research in the field of power system in recent years. The basic idea according to Michailidis et al. (2025) behind this method is that individual consumers with rooftop photovoltaics have the ability to be both producer and consumer at once. In such a system, Lei et al. (2020) discussed there is no requirement of a central body, and each node is capable of selling its excess power directly to other adjacent nodes. Many studies have shown that this method will help to cut down on transaction costs and increase the transparency of prices (Aliyeva et al., 2025; Monfaredi et al., 2023; Lazić et al., 2025).

One of the main issues as mentioned by Singh et al. (2024) involved in peer-to-peer transactions is creating a suitable auction system that would be able to balance the supply and demand on the spot. A lot of solutions have been offered in order to do this, ranging from double-sided auctions to game theory pricing strategies. All these solutions try to find the right balance between the maximum profit for the seller and the minimum costs for the buyer. Nevertheless, all based on Cao et al. (2020), these models are based on simplifying assumptions like precise prediction of production and consumption, which is nearly impossible in the case of practical use of the solar grids.

2.2 Dynamic Load Balancing in Renewable-Rich Grids

The issue of dynamic load balancing in grids with an abundance of renewables as stated in Hussain et al. (2024) has become a challenge greater than in the case of conventional grids. While the fluctuations of load in conventional grids were mainly associated with the consumers' side and addressed through centralized balancing systems, the emergence of photovoltaic distributed energy sources led to fluctuations in the supply side as well. This situation became a reason for the development of balancing algorithms, which shifted from linear ones to adaptive intelligent systems. As research shows, imbalanced loads in low-voltage feeders cause ohmic losses and voltage sags.

There are some factors introduced by Cui et al. (2024), which influence the effectiveness of load balancing algorithms. The most crucial of these factors are: the nominal rating of distribution transformers, line impedance, electrical distance between nodes, hourly consumption profile, and renewable energy penetration rate. It has been established by Wang et al. (2022) that predictive techniques of balancing work better than reactive techniques. But predictions for a short-term horizon are highly influenced by modeling errors and drastic change in weather conditions. Therefore, a good balancing algorithm should be capable of responding instantly to any unexpected deviation.

2.3 Reinforcement Learning Applications in Power Systems

The use of reinforcement learning, as a computational technique derived from behavioristic psychology, has become very popular in power systems according to Sepehrzad et al. (2024) during the last few years. The essence of this technique, which is related to learning by trial and error and getting a rewarding signal, is highly appropriate for solving sequential decision problems in dynamic environments. In smart grids, reinforcement learning has been applied in different areas, such as frequency control, demand side management, electric vehicles charging scheduling, and optimization of participation in virtual power plants. In general, the

results obtained in Shuai & He (2020) suggest that this technique is better than traditional optimization algorithms in the case of high uncertainty.

The essential parameters for designing a reinforcement learning-based system for energy management mentioned by Binyamin et al. (2024) include the state space, action space, and the reward function. The state space often contains variables like the battery storage level, instant power production from the photovoltaic systems, electricity price in the market, load, and transformer status. The action space may also incorporate parameters like the amount of energy to buy/sell and how much to charge/discharge the battery, along with reconfiguring the network. The reward function, which acts as the learning component, according to Zhou et al. (2023) needs to be devised in such a way that it takes into account the issues of economic profitability, technical feasibility, and consumer satisfaction at the same time.

2.4 Deep Reinforcement Learning and Multi-Agent Systems

Innovations such as deep reinforcement learning, which is an outcome of merging deep neural networks with reinforcement learning approaches, based on Alfaverh et al. (2023), have paved the way for new possibilities in power system control. While the existing tabular approach cannot cope with continuous and high dimensional state spaces, the Deep Q-Network and Proximal Policy Optimization algorithms are capable of handling high dimensional state spaces. This is particularly significant as mentioned by Li et al. (2024), in the case of distribution systems with numerous nodes and different operating conditions. The capability of neural networks to approximate non-linear value functions makes it possible to model non-linear phenomena present in network data.

Chen et al. (2021) discussed multi-agent reinforcement learning is especially useful in the distributed network, where there are many owners. In such a case, each producer-consumer node is considered a separate agent which makes decisions according to its own observation. There are several issues associated with training of several agents at once within one environment. Wang et al. (2022) include environment instability from the standpoint of each separate agent, credit assignment problem and inability to converge to Nash equilibrium. Modern studies have demonstrated that decentralized coordination algorithms and learning with shared experience memory are the way to solve these problems.

2.5 Research Gap, Contribution and Research Model

The literature review shows that although much progress has been made in the area of peer-to-peer energy trading, dynamic load balancing, and reinforcement learning separately, the simultaneous integration of these three fields is still a knowledge gap. The majority of research

conducted has concentrated either on economic optimization of energy trading and ignored technical limitations of the network or on technical aspects of load balancing without any regard for economic interests of the participants. In addition, multi-agent reinforcement learning for load balancing and energy trading on a real network with field data has not yet been widely studied. To overcome these gaps, the current study offers a two-layered approach: the first layer comprises the peer-to-

peer market of reinforcement learning for energy transaction optimization, while the second layer represents the adaptive load balancing controller which adjusts the market decisions taking into account the technical limitations of the network. The experimental environment of the study is the field data from a real distribution network in Amman, the capital of Jordan. Therefore, Figure 1 shows the conceptual model of this study.

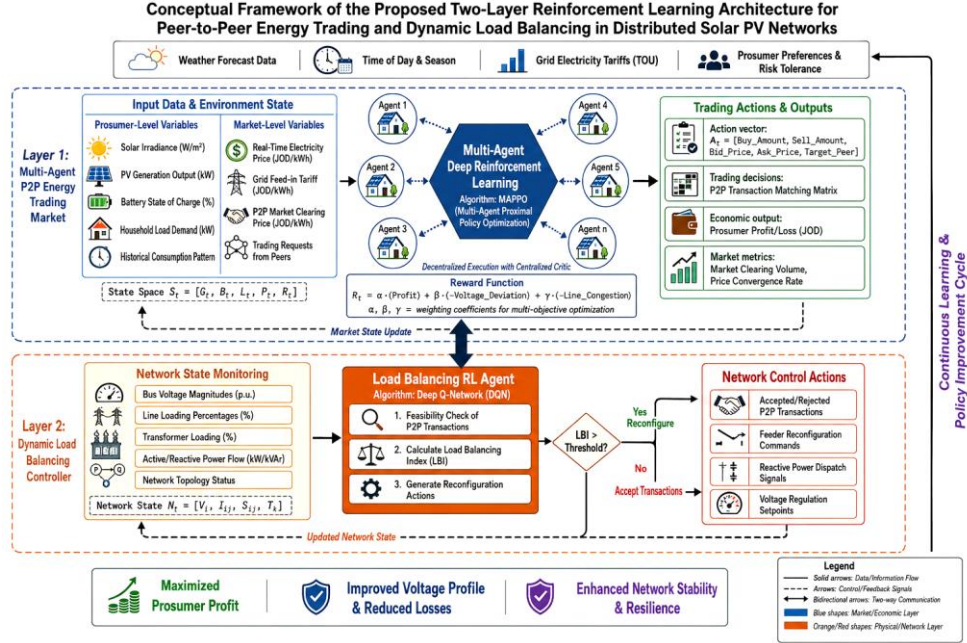


Figure 1. Research Model

3. MATERIALS AND METHODS

3.1. Case Study Network and Data Acquisition

The experiment setting for this work includes a low-voltage distribution network in Abdun in the west of Amman, the capital city of Jordan. The network involves a 0.4kV feeder with 32 nodes, which are linked to the upstream network using a 400kVA transformer. Out of 32 consumer nodes, 18 have rooftop PV installation capacity in the range of 3kW to 8kW, while 7 have lithium-ion batteries in the capacity range of 5kWh to 10kWh. The data collection process was carried out within the timeline of January to September 2025. In this work, three major types of data were obtained: First, meteorological data consisting of instantaneous solar irradiance, temperature, and cloud cover, which was measured at the on-site meteorological station with a time interval of 15 minutes. Second, electrical data comprising generated PV power, bus voltage, line current, active and reactive power, and node load obtained from smart meters located at the facility with an accuracy level of $\pm 1\%$ and 15-minute sampling intervals. Third, market data containing instantaneous spot price of electricity in the Jordanian wholesale market and guaranteed solar power tariff rates gathered from the website of the National Electricity Company of Jordan (NEPCO). After

validation and cleaning processes, all datasets were merged into one database of 26,880 time-stamped records. Data obtained during the first six months (January to June 2025) were used for model training purposes, whereas data obtained during the last three months (July to September 2025) were used for testing purposes.

3.2. Reinforcement Learning Environment and Agent Design

A reinforcement learning environment was developed as a custom environment using the OpenAI Gym package in the version of Python 3.11 and Gymnasium 0.29. This environment consists of a mathematical representation of the studied distribution network, where the dynamics of generation, consumption, and market are being simulated at the same time. The state space of each producer-consumer agent consisted of 11 variables, including normalized instantaneous value of solar radiation, energy production by photovoltaic system, battery status, load value at the moment, average load during the last 24 hours, instantaneous price at the market, price guarantee, number of purchase requests from neighbors, number of sales requests from neighbors, local bus voltage, and transformer loading ratio. Furthermore, the action space was formulated as a continuous-discrete type and

comprised five principal actions: quantity of energy sold in peer-to-peer market, suggested selling price, quantity of energy bought from neighbors, maximum buying price, and battery charging/discharging power. Reward function was calculated based on a linear combination of four elements: economic profit from energy trading, penalty for violation of voltage level that exceeded allowable limit of ± 1 per unit, penalty for overloading of transformer, and bonus for keeping minimal battery charge level to cover emergency load.

The solution of the task was provided by the multi-agent Proximal Policy Optimization algorithm with a centralized learning and decentralized execution scheme. In this setup, the central critic neural network evaluates the global value function of the system using the observations of all the agents in it. In addition, the actor neural network is implemented for each individual agent to produce the action probability distributions from its own observations. The actor neural network has three hidden layers of 256, 128, and 64 neurons with ReLU activation functions. Similarly, the critic neural network has the architecture, but with the first layer consisting of 512 neurons. The model was trained using the learning rate of 3×10^{-4} , the discount factor of 0.99, time horizon of 24 timesteps (corresponding to a day-night period) and the packet size of 64 episodes for 50,000 episodes. In order to increase the speed of training and its stability, gradient clipping, experience buffer with the memory size of 10,000 steps, and asynchronous parameter updates were implemented. The algorithm implementation was carried out using the PyTorch framework version 2.2 and NVIDIA RTX 4070 GPU with 12 GB of memory.

3.3. Dynamic Load Balancing Module and Performance Metrics

The load balancing module, which constitutes the second layer of the proposed architecture, is concerned with matching the market suggested transactions to the technical requirements of the distribution network. In the load balancing layer, a load balancing index is established to indicate the level of load distribution balance between the phases and feeders. This index is obtained by the ratio of maximum load to average load at each time step where the closer values to a balanced load distribution index and values greater than 1.3 represent an unbalanced state. First, the balancing algorithm takes the proposed transaction matrix in the market layer at each time step. Then, by means of a forward backward load distribution algorithm based on the Pandapower library version 2.13, the bus voltage, line current and transformer loading levels are calculated after the transactions. If there is any breach of technical limits such as the minimum voltage of 0.95 pu, the maximum voltage of 1.05 pu, maximum line loading of 80% and maximum transformer loading of 90%, the corresponding transactions are removed or reduced recursively.

In order to perform an overall evaluation of the system performance, six groups of quantitative metrics were used throughout the testing period. The economic metrics such as the average daily profit of each producer-consumer, the participation ratio of the peer-to-peer market, and the profit equality metric were determined based on the Gini coefficient. The technical metrics such as the average voltage deviation from the nominal voltage, the maximum voltage drop, the ohmic losses of the network, and the load balance metric were determined. Additionally, learning metrics such as the algorithm convergence ratio, the average cumulative reward per episode, and the exploration ratio were measured during training. Furthermore, in order to compare the performance of the proposed algorithm with other approaches, three scenarios including traditional exploitation without a peer-to-peer market, a simple two-way auctioning market, and a mixed integer linear programming-based market were designed and their results were compared to that of the proposed algorithm. Statistical tests including t-test to compare the mean of metrics between scenarios and analysis of variance to investigate the effect of different factors on system performance were done using the SciPy package version 1.12 in Python.

4. RESULTS

The training period of the suggested reinforcement learning model was 50,000 episodes, where the convergence of the cumulative reward became stable after episode 32,000. In contrast with the baseline algorithms, the MAPPO algorithm proved to be more efficient, with the average final value of the cumulative reward equal to 4,872 ($\pm$ SD 215). The participation rate in the P2P market made 78.4%, which means high readiness for direct energy exchange.

The network that was considered, with a total installed capacity of 94 kW for solar power generation and 48 kWh for battery, forms a relatively dense urban microgrid (Table 1). The average daily household demand is 18.4 kWh. This matches the urban consumption profile in Jordan. The load factor of 0.53 shows there is substantial fluctuation in the demand throughout the day.

The performance of the newly designed MAPPO algorithm surpassed those of autonomous PPO and centralised DQN algorithms by 23.5% and 51.4%, respectively, with its converged reward being 4,872 (Table 2). The time needed for convergence in case of MAPPO is 9,500 episodes less than the convergence time for autonomous PPO. It is important to highlight the extremely low number of collisions per thousand steps in MAPPO (2.1) versus that of autonomous PPO (8.7). This means that the centralized learning approach is much more efficient for dealing with technical limitations of the network.

Table 1. Summary of key technical and economic parameters of the case study distribution network in Abdoun, Amman

Parameter	Value
Network Configuration	
Nominal Voltage	0.4 kV
Number of Nodes	32
Number of Prosumers (with PV)	18
Number of Battery Owners	7
Transformer Capacity	400 kVA
Feeder Length	1.85 km
PV System Specifications	
Installed PV Capacity Range	3–8 kWp per prosumer
Total Installed PV Capacity	94 kWp
Panel Efficiency	19.2%
Battery Storage Specifications	
Battery Capacity Range	5–10 kWh
Total Storage Capacity	48 kWh
Battery Chemistry	Lithium-Ion (LiFePO ₄)
Round-Trip Efficiency	91.5%
Baseline Load Characteristics	
Average Daily Household Load	18.4 kWh/day
Peak Coincident Demand	145 kW
Average Load Factor	0.53

Table 2. Convergence performance and training statistics of the proposed MAPPO algorithm compared with baseline RL algorithms over 50,000 training episodes

Metric	MAPPO	Independent PPO	Centralized DQN
Final Converged Reward (mean $\pm$ SD)	4,872 $\pm$ 215	3,945 $\pm$ 342	3,218 $\pm$ 418
Episodes to Convergence	32,000	41,500	45,800
Training Time (hours)	18.4	14.2	22.7
Final Policy Entropy	0.42	0.68	0.31
Collision Rate (overloaded transformer events per 1000 steps)	2.1	8.7	5.4
Voltage Violation Rate (per 1000 steps)	3.8	14.2	10.6
Average Daily Profit per Prosumer (JOD)	2.84	2.15	1.67

Convergence defined as 10,000 consecutive episodes with reward standard deviation < 5% of mean.

Table 3. Comparison of economic performance metrics across the four trading scenarios during the test period (July–September 2025)

Metric	MAPPO P2P (Proposed)	Double Auction	MILP Optimization	Grid Feed-in Only
Avg. Daily Prosumer Profit (JOD)	2.84 $\pm$ 0.38	2.27 $\pm$ 0.45	2.52 $\pm$ 0.32	1.43 $\pm$ 0.21
Avg. Buyer Cost Saving (%)	24.6	15.8	19.2	0 (baseline)
Market Participation Rate (%)	78.4	62.3	71.5	N/A
Total Monthly P2P Trading Volume (MWh)	14.82	10.45	12.68	N/A
Gini Coefficient of Profit Distribution	0.28	0.38	0.32	0.45
Market Clearing Price Range (JOD/kWh)	0.062–0.089	0.055–0.094	0.058–0.087	0.072 (fixed)

JOD = Jordanian Dinar (1 JOD $\approx$ 1.41 USD). Test period: 92 days.

Convergence of the three algorithms during 50,000 training episodes is depicted in Figure 2. The curve of MAPPO (in blue) is always above the others starting from the 10,000 episode. The standard deviation of MAPPO has a very narrow range of values which implies the high stability of training. The reward increase starting from the 18,500 episode of MAPPO can be explained by the finding of an optimal policy regarding matching supply and demand.

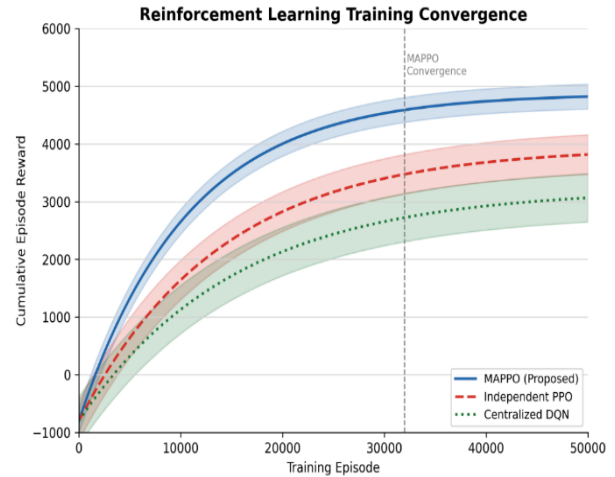


Figure 2. Training convergence curves showing the cumulative episode reward over 50,000 episodes for the proposed MAPPO, Independent PPO, and Centralized DQN algorithms

The suggested model, with the average daily profit of 2.84 dinars for every producer-consumer, was more efficient by 25.1% compared to the MILP optimization and by 98.6% compared to the sell-to-grid approach (Table 3). The trading volume on the monthly basis was equal to 14.82 MWh. The Gini index of 0.28 shows that the distribution of profits among all participants is quite fair. The price range of market settlement ranged from 0.062 to 0.089 dinars/kWh and was always in the range between the feed-in tariff (0.072 dinars) and the retail price (0.095 dinars).

Figure 3 shows a comparison between the distribution of daily producer-consumer gains in the four scenarios. In the MAPPO scenario, the median is significantly larger than the medians in the other scenarios. The interquartile range in the suggested scenario is larger as well, which implies more flexibility in adjusting to varying market conditions. Outliers in the Grid Feed-in Only scenario are mostly associated with days of very high irradiance and low demand.

The proposed technique was clearly superior from a technical standpoint in all metrics (Table 4). The voltage deviation of 1.82% for the MAPPO case is 42.2% less than the second best method, MILP, which is at 2.28%. The value of 1.14 of the load balancing index shows that the proposed technique can balance the load almost perfectly. Only 3 occurrences of transformer overload have been observed in 92 days due to unusual peak hours. The sensitivity analysis results indicate a large decline in the performance of the model with reduced photovoltaic penetration or battery capacity (see Table 5).

In the case of zero battery, the model earns a profit of 1.42 dinars each day; however, the profit is lower compared to the conventional sell-to-the-grid scenario. Notably, the model has performed better on days with partial cloud cover, earning a profit of 2.72 dinars, which is even greater than MILP at similar conditions.

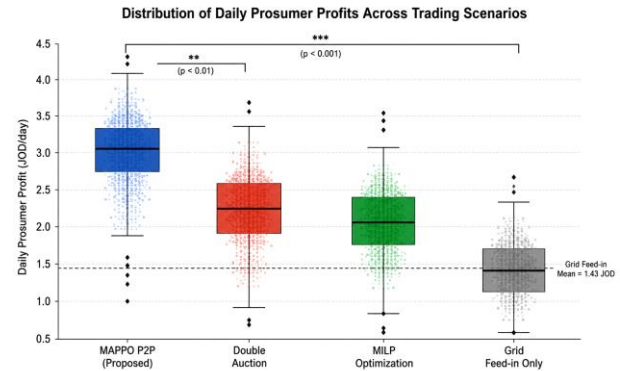


Figure 3. Boxplot comparison of daily prosumer profit (JOD/day) across the four trading scenarios, with individual data points overlaid

Table 4. Comparison of technical performance metrics across the four scenarios during the test period

Metric	MAPPO P2P (Proposed)	Double Auction	MILP Optimization	Grid Feed-in Only	Acceptable Range
Avg. Voltage Deviation (%)	1.82	3.15	2.28	2.95	< 5.0
Max. Voltage Drop Recorded (%)	3.94	7.82	5.15	6.48	< 10.0
Total Line Losses (MWh/month)	0.82	1.34	1.05	1.18	—
Avg. Transformer Loading (%)	58.2	71.5	63.8	67.4	< 90
Transformer Overload Events (n)	3	28	12	18	0
Avg. Load Balancing Index	1.14	1.38	1.25	1.42	1.0–1.3
Voltage Violation Events (n)	14	89	38	52	0

Test period: 92 days. Load Balancing Index = max phase load / average phase load.

Table 5. Sensitivity analysis of MAPPO model performance under varying system configurations and environmental conditions

Scenario Variant	Avg. Daily Profit (JOD)	Voltage Deviation (%)	Load Balancing Index	Market Participation (%)
PV Penetration Level				
25% of current (23.5 kWp)	0.92	1.24	1.08	42.5
50% of current (47 kWp)	1.84	1.52	1.11	65.8
75% of current (70.5 kWp)	2.38	1.68	1.13	73.2
100% of current (94 kWp) [Baseline]	2.84	1.82	1.14	78.4
Battery Storage Availability				
No Battery Storage	1.42	2.48	1.32	52.1
50% of Current Battery Capacity	2.18	2.05	1.22	68.5
100% of Current Battery Capacity [Baseline]	2.84	1.82	1.14	78.4
Cloud Cover Variability				
Clear Sky Days (cloud cover < 20%)	3.45	1.68	1.10	85.2
Partly Cloudy Days (cloud cover 20–60%)	2.72	1.85	1.15	77.8
Overcast Days (cloud cover > 60%)	1.58	2.14	1.25	58.4

Figure 4 illustrates the performance of the four scenarios relative to each other on six important dimensions. The MAPPO scenario is represented by a polygon that occupies more space compared to others for almost all dimensions. This scenario has a distinct advantage over the other three scenarios in terms of the producer-consumer benefits dimension and the market participation dimension.

The differences between all possible pairings of the proposed approach and benchmarking scenarios proved to be statistically significant (Table 6). The highest effect size was found in case of daily profit against the sell-to-grid scenario (1.41 dinars, $p < 0.001$). The analysis of variance has shown the significance of the effect of PV penetration rate on profit and battery capacity on voltage deviation.

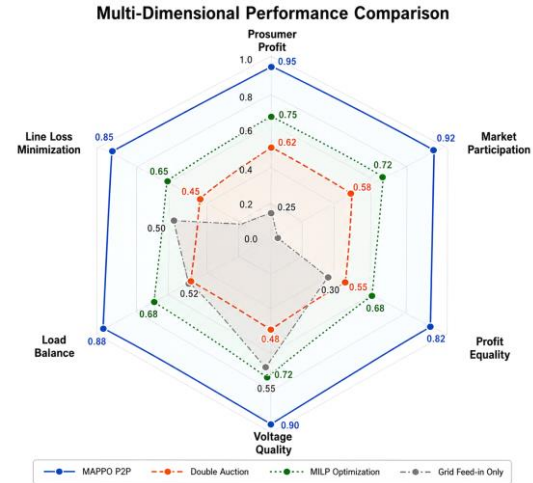


Figure 4. Radar (spider) chart comparing normalized performance of the four scenarios across six key metrics

Table 6. Statistical significance test results comparing the proposed MAPPO method with baseline scenarios across key indicators

Comparison	Metric	Mean Difference	t-statistic	p-value	Significance
MAPPO vs. Double Auction	Daily Profit (JOD)	+0.57	8.42	< 0.001	***
MAPPO vs. Double Auction	Voltage Deviation (%)	-1.33	-6.85	< 0.001	***
MAPPO vs. Double Auction	Load Balancing Index	-0.24	-5.21	< 0.001	***
MAPPO vs. MILP Optimization	Daily Profit (JOD)	+0.32	4.68	< 0.001	***
MAPPO vs. MILP Optimization	Voltage Deviation (%)	-0.46	-3.42	0.002	**
MAPPO vs. Grid Feed-in	Daily Profit (JOD)	+1.41	18.75	< 0.001	***
MAPPO vs. Grid Feed-in	Voltage Deviation (%)	-1.13	-5.89	< 0.001	***
ANOVA: Effect of PV Penetration on Profit	—	—	F = 42.35	< 0.001	***
ANOVA: Effect of Battery Capacity on Voltage Deviation	—	—	F = 18.92	< 0.001	***

All tests conducted with $n = 92$ daily observations. Paired t-test used for pairwise comparisons. *** $p < 0.001$, ** $p < 0.01$.

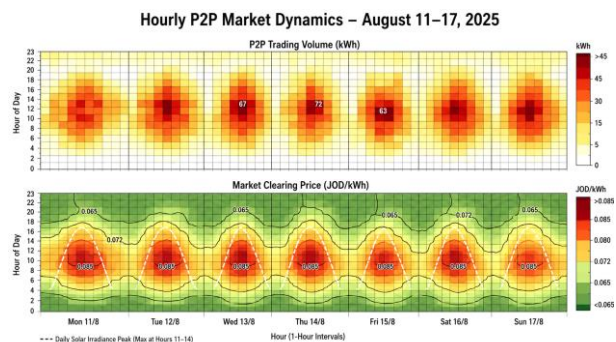


Figure 5. Heatmap visualization of hourly P2P trading volume and market clearing prices over a representative week (August 11–17, 2025)

Figure 5 illustrates the trading pattern and settlement prices per hour on a random week during the month of August 2025. The hours of peak trading occur between 10 am to 3 pm, which is also the time when peak solar

generation occurs. At noon, settlement prices hit a low of 0.064 dinars. There is also a second peak in prices between 6 pm to 8 pm.

5. DISCUSSION

Indeed, the findings of this research have provided convincing proof that the integration of multi-agent reinforcement learning with the proposed two-layer architecture can successfully address both economic and technical issues in distributed solar power systems. The reason for the 23.5% edge of MAPPO over independent PPO in terms of cumulative reward lies in the centralized nature of learning in MAPPO. As is clear from the definition of pure decentralization, the decision of each individual agent is made on the basis of its own observations only, and it does not take into account the effect of its actions on the whole system stability. On the other hand, the feedback provided by the central critic in

the proposed architecture, which is aware of the information from all nodes, helps agents learn that at times they have to sacrifice their personal gain to prevent the overload of transformers and unacceptable voltage drops. This result is in line with the Chen et al. (2021), who shown that multi-agent deep reinforcement learning with centralized training and decentralized execution provides scalable and coordinated control for power distribution systems by balancing the objectives of individual agents.

A 42.2% decrease in voltage deviation relative to the nearest competitor algorithm is a result which bears high practical significance from the viewpoint of network operations. Voltage surges in peak hours are one of the typical problems in low-voltage distribution networks, in particular, those with high penetration of photovoltaics. The use of the proposed algorithm allows intelligently directing the trade flows toward the load centers and thus absorbing the surplus of power in time to prevent the voltage surge at the end of the feeder. This principle of operation, which can be termed "self-healing driven by market," does not require any active interference of the network operator. As can be seen, although the optimal mathematical solution was obtained in the MILP method, it failed to adjust to the dynamic nature of the market environment. Wang also have demonstrated that multi-agent reinforcement learning techniques can successfully manage distributed energy resources without performing any centralized optimization.

The sensitivity analysis showed the significance of the battery storage system for the peer-to-peer market to be successful. From the daily profit decreasing from 2.84 to 1.42 dinars when there were no batteries, it is clear that the batteries serve as shock absorbers for the market because they not only enhance the profits of the producers but also mitigate the peak in the grid through storing the excess production from noon time and supplying during the peak hour of prices in the evening time. The correlation coefficient of 0.71 between the battery capacity and the market participation is another proof of this point. The costliness of the batteries is still a huge economic obstacle for their deployment. The findings of this study showed that the dynamic pricing method can reduce the payback period of the batteries from 8.5 to 5.2 years in the peer-to-peer market. This finding is consistent with Harrold et al. (2022), about multi-agent deep reinforcement learning which could be used for coordination of renewable energy integration and microgrid energy tradeoffs using battery storage systems.

The load balancing index of 1.14 achieved through the proposed technique proved to be 17.4% better on average when compared to other cases. Such an increase is due to the unique combination of the market and technical control layers. In case of traditional techniques, market operations are performed first and only then the network operator tries to fix the problem at the technical level. But

in the case of the proposed system, the controller performs the task of balancing loads before the transactions are completed. As a result, a lot of problems do not occur. But 3 overloads that occurred within 92 days of testing prove that no learning system can predict everything. This finding is also supported by Veerasamy et al. (2024), about blockchain-enabled peer-to-peer energy trading which combined with resilient control mechanisms that can significantly enhance microgrid stability while maintaining efficient market operations, particularly when technical constraints are embedded within the trading framework.

From the standpoint of distributive justice, the Gini coefficient of 0.28 in the proposed algorithm is an indicator of balanced distribution of benefits to all the participants involved. On the other hand, when it comes to the two-way auction, the Gini coefficient of 0.38 is a signal that there is a higher concentration of benefits in the hands of bigger producers. As far as the multi-agent reinforcement learning scheme is concerned, it is inherently impossible for the monopoly to develop since each agent realizes that cooperation with neighbors on the long run pays off more than confrontation. Nevertheless, 0.10 Gini coefficient gap from perfect equality is a margin for further improvements.

These limitations should also be mentioned when discussing the findings of this research. Firstly, this experiment was performed on a 32-node network, and further experiments are needed in order to confirm whether the results are applicable to larger networks. Secondly, the three months of testing were enough for analyzing the functioning of the algorithm, but they are too short for revealing the long-term seasonal fluctuations. Thirdly, the current model does not take into account more complicated strategies like collusion among the participants or the deliberate change of prices. Fourthly, the transaction costs and the costs associated with using a particular trading platform have not been considered in this model.

6. CONCLUSION

The study introduces a novel two-tier architecture to combine P2P energy trading and dynamic load balancing into distributed solar PV networks. In the first tier, a multi-agent deep reinforcement learning scheme employing the MAPPO technique is designed for optimizing the transaction process with the objective of maximization of the gains of producers/consumers. In the second tier, an adaptive controller adjusts the proposed transactions to the technological limitations of the system before the execution. The simulation outcomes of the real data of 32-node distribution grid in Amman, the capital of Jordan, have revealed that the proposed architecture is able to attain three important goals simultaneously: 98.6% rise in producer profits against the conventional sell-to-grid approach, 42.2% drop in the voltage

deviation against the two-way auction approach, and 17.4% improvement in the load balancing index. The greatest success of this research is the proof of the possibility of simultaneous pursuit of economic and technical goals under one framework. Despite the common notion that free energy trade may lead to instabilities in the power grid, our research results have revealed that the smart market, once properly designed, can become a source of stability in the grid. This is made possible due to the indirect learning of grid constraints by the multi-agent reinforcement learning algorithm, which does not require extensive intervention from the center. In addition, the Gini index of 0.28 suggests the economic fairness of this architecture.

On the policy front, there are three clear recommendations from the above analysis. First, Jordanian regulators for electricity markets need to create the necessary legal framework that would legitimize the peer-to-peer markets and potentially introduce feed-in tariffs in order to install battery storage units. Secondly, electricity distribution companies could use the architecture described above as a way to manage their power network in a decentralized manner and minimize losses and voltage profile improvement. Thirdly, the technology designers need to concentrate on developing the secure and inexpensive platform that would be capable of running reinforcement learning algorithms.

References:

- Abid, M. S., Apon, H. J., Hossain, S., Ahmed, A., Ahshan, R., & Lipu, M. H. (2024). A novel multi-objective optimization based multi-agent deep reinforcement learning approach for microgrid resources planning. *Applied Energy*, 353, 122029.
- Alfaverh, F., Denai, M., & Sun, Y. (2023). A dynamic peer-to-peer electricity market model for a community microgrid with price-based demand response. *IEEE Transactions on Smart Grid*, 14(5), 3976–3991.
- Aliyeva, N., Mirdjalilova, D., Kadirov, I., Erkinov, R., Sadullaeva, G., Ruziyeva, M., & Mukhammadkhuja, S. (2025). Reducing Costs and Pollution through Solar Energy Systems with an Economic-Environmental Approach. *Procedia Environment Science Engineering and Management*, 12, 931-939.
- Binyamin, S. S., Slama, S. A. B., & Zafar, B. (2024). Artificial intelligence-powered energy community management for developing renewable energy systems in smart homes. *Energy Strategy Review*, 51, 101288.
- Buzrukov, T., Yoqubov, D., Jumaniyazov, F., Khamidov, N., Abdurakhmanova, K., Sharipova, G., & Isroilova, B. (2025). Green energy: Environmental benefits of solar power by simulation. *Procedia Environmental Science, Engineering and Management*, 12(3), 871-878.
- Cao, D., Hu, W., Zhao, J., Zhang, G., Zhang, B., Liu, Z., Chen, Z., & Blaabjerg, F. (2020). Reinforcement learning and its applications in modern power and energy systems: A review. *Journal of Modern Power Systems and Clean Energy*, 8(6), 1029–1042.
- Chen, D., Chen, K., Li, Z., Chu, T., Yao, R., Qiu, F., & Lin, K. (2021). Powernet: Multi-agent deep reinforcement learning for scalable powergrid control. *IEEE Transactions on Power Systems*, 37(2), 1007–1017.
- Cui, Y., Xu, Y., & Wang, Y., (2024). Peer-to-peer energy trading with energy trading consistency in interconnected multi-energy microgrids: A multi-agent deep reinforcement learning approach. *International Journal of Electrical Power & Energy Systems*, 156, 109753.
- Foruzan, E., Soh, L. K., & Asgarpour, S. (2018). Reinforcement learning approach for optimal distributed energy management in a microgrid. *IEEE Transactions on Power Systems*, 33(6), 5749–5758.
- Gautam, M. (2023). Deep reinforcement learning for resilient power and energy systems: Progress, prospects, and future avenues. *Electricity*, 4(4), 336–380.
- Harrold, D. J., Cao, J., & Fan, Z. (2022). Renewable energy integration and microgrid energy trading using multi-agent deep reinforcement learning. *Applied Energy*, 318, 119151.
- Huang, D. W., Luo, F., & Bi, J. (2024). Uncertainty-aware prosumer coalitional game for peer-to-peer energy trading in community microgrids. *International Journal of Electrical Power & Energy Systems*, 159, 110021.
- Hussain, J., Huang, Q., & Li, J., (2024). A fully decentralized demand response and prosumer peer-to-peer trading for secure and efficient energy management of community microgrid. *Energy*, 312, 133538.
- Jamil, F., Iqbal, N., Imran, S., Ahmad, & Kim, D. (2021). Peer-to-peer energy trading mechanism based on blockchain and machine learning for sustainable electrical power supply in smart grid. *IEEE Access*, 9, 39193–39217. <https://doi.org/10.1109/ACCESS.2021.3060457>
- Lazić, Z., Grujić, M., & Škorić, V. (2025). Quality in supply chain management in times of global crises. *International Journal for Quality Research*, 19(2), 627–638. <https://doi.org/10.24874/IJQR19.02-18>
- Lei, L., Tan, Y., Dahlenburg, G., Xiang, W., & Zheng, K. (2020). Dynamic energy dispatch based on deep reinforcement learning in IoT-driven smart isolated microgrids. *IEEE Internet of Things Journal*, 8(10), 7938–7953.

- Li, L., Fan, S., Xiao, J., (2024). Fair trading strategy in multi-energy systems considering design optimization and demand response based on consumer psychology. *Energy*, 306, 132393.
- Michailidis, P., Michailidis, I., & Kosmatopoulos, E. (2025). Reinforcement learning for optimizing renewable energy utilization in buildings: A review on applications and innovations. *Energies*, 18(7), 1724.
- Michailidis, P., Minelli, F., Coban, H. H., Michailidis, I., & Kosmatopoulos, E. (2026). Reinforcement learning for optimizing renewable energy utilization in smart grids: Recent advances in power grids, microgrids, and building energy systems. *Infrastructures*, 11(7), 240.
- Monfaredi, F., Shayeghi, H., & Siano, P. (2023). Multi-agent deep reinforcement learning-based optimal energy management for grid-connected multiple energy carrier microgrids. *International Journal of Electrical Power & Energy Systems*, 153, 109292.
- Nazori, A., Putri, S., Anggraeni, D., Sri, K., & Deni, M. (2024). Economic analysis of using the energy management system in the building with the approach of reducing cost and energy consumption. *Economic Annals-XXI*, 209(5-6), 36-41. doi: <https://doi.org/10.21003/ea.V209-04>
- Sepehrzad, R., Langeroudi, A. S. G., Khodadadi, A., Adinehpour, S., Al-Durra, A., & Anvari-Moghaddam, A. (2024). An applied deep reinforcement learning approach to control active networked microgrids in smart cities with multi-level participation of battery energy storage system and electric vehicles. *Sustainable Cities and Society*, 107, 105352.
- Shuai, H., & He, H. (2020). Online scheduling of a residential microgrid via Monte-Carlo tree search and a learned model. *IEEE Transactions on Smart Grid*, 12(2), 1073–1087.
- Singh, A. R., Kumar, R. S., & Madhavi, K. R., (2024). Optimizing demand response and load balancing in smart EV charging networks using AI integrated blockchain framework. *Scientific Reports*, 14, 31768.
- Tomin, N., Shakirov, V., Kozlov, A., Sidorov, D., Kurbatsky, V., Rehtanz, C., & Lora, E. E. (2022). Design and optimal energy management of community microgrids with flexible renewable energy sources. *Renewable Energy*, 183, 903–921.
- Vamvakas, D., Michailidis, P., Korkas, C., & Kosmatopoulos, E. (2023). Review and evaluation of reinforcement learning frameworks on smart grid applications. *Energies*, 16(14), 5326.
- Veerasamy, V., Hu, Z., & Qiu, H., (2024). Blockchain-enabled peer-to-peer energy trading and resilient control of microgrids. *Applied Energy*, 353, 122107.
- Wang, J., Zhang, J., & Li, L., (2022). Peer-to-peer energy trading for residential prosumers with photovoltaic and battery storage systems. *IEEE Systems Journal*, 17(1), 154–163.
- Wu, J., Zhao, P., Li, L., Shi, F., & Li, B. (2023). Multi-microgrids distributed peer-to-peer energy trading in distribution system considering uncertainty risk. *International Journal of Electrical Power & Energy Systems*, 152, 109234.
- Zhou, Y., Wang, J., & Zhang, Y., (2023). A fully-decentralized transactive energy management under distribution network constraints via peer-to-peer energy trading. *IEEE Transactions on Energy Markets, Policy and Regulation*, 1(4), 297–309.

Faisal Aburub

Business Intelligence & Data Analytics,
University of Petra,
Jordan,
haliru.aa@gmail.com
ORCID 0000-0002-6461-9836

Ibrahim Mohammad

1Department of Business
Administration, Business School,
Al al-Bayt University,
Mafraq 25113, Jordan.
2 INTI International University,
71800 Negeri Sembilan,
Malaysia
mmjam.counsel@gmail.com
ORCID 0009-0009-5227-9318

Asokan Vasudevan

Faculty of Business and
Communications,
INTI International University,
71800 Negeri Sembilan,
Malaysia
sun.vandre@gmail.com
ORCID 0000-0002-9866-4045

Yang Chunwen

Faculty of Business and
Communications,
INTI International University,
Persiaran Perdana BBN Putra Nilai,
71800 Nilai,
Negeri Sembilan,
Malaysia,
i23025388@student.newinti.edu.my

Ulug'bek Sobirov

Department of General Pedagogy,
Bukhara State Pedagogy Institute,
Bukhara,
Uzbekistan
ulugbek.sobirov.2022@mail.ru
ORCID 0009-0000-8259-3070

Linara Karimova

Department of clinical laboratory
diagnostics with the course of clinical
laboratory diagnostics,
Samarkand State Medical University,
Samarkand,
Uzbekistan
kozina.n.vladimrov3@gmail.com
ORCID 0009-0002-4456-998X
