IEEE Access
Multidisciplinary ⋮ Rapid Review ⋮ Open Access Journal

Received 17 July 2024, accepted 6 August 2024, date of publication 12 August 2024, date of current version 20 August 2024.

Digital Object Identifier 10.1109/ACCESS.2024.3442445

RESEARCH ARTICLE

Optimization of Peer-to-Peer Energy Trading With a Model-Based Deep Reinforcement Learning in a Non-Sharing Information Scenario

NAT UTHAYANSUTHI AND PEERAPON VATEEKUL, (Senior Member, IEEE)
Department of Computer Engineering, Faculty of Engineering, Chulalongkorn University, Pathumwan, Bangkok 10330, Thailand
Corresponding author: Peerapon Vateekul (peerapon.v@chula.ac.th)

ABSTRACT In the realm of sustainable energy distribution, peer-to-peer (P2P) trading within microgrids has emerged as a promising solution, fostering decentralization and efficiency. While previous studies focused on optimizing P2P trading, they often relied on impractical assumptions regarding private information sharing among prosumers. To overcome this limitation, we aim to optimize P2P energy trading within the microgrid based on a realistic assumption (not sharing information), using our proposed model-based multi-agent deep reinforcement learning model. Firstly, our framework integrates long short-term memory (LSTM) for the policy model. Secondly, our model-based framework is based on temporal fusion transformers (TFT) for 24h-ahead net load consumption. Thirdly, the global horizontal index (GHI) is added to the model. Finally, a clustering technique helps to segment a large number of households into small household groups. The experiment was conducted on the Ausgrid dataset, consisting of 300 households in Sydney, Australia. Results demonstrate that our model achieved 4.20% and 3.95% lower microgrid electricity costs than MADDPG and A3C3, the sharing-info-based models. Moreover, it shows 12.48% lower costs than directly trading energy with the utility grid.

INDEX TERMS Model-based deep reinforcement learning, multi-agent deep reinforcement learning, peer-to-peer energy trading, non-sharing information.

I. INTRODUCTION

Emissions from agricultural and industrial sectors have caused global warming [1], [2] requiring society to shift to renewable energy sources such as solar, wind, and biogas. In particular, recent advancements in the energy sector such as solar have notably shifted from centralized systems towards a decentralized paradigm, primarily influenced by the rising penetration of distributed energy resources (DERs) [3]. This transition is facilitating the development of peer-to-peer (P2P) energy trading platforms, where individuals and communities actively engage in direct trading of excess renewable energy, diminishing reliance on traditional power grids and fostering economic benefits by potentially lowering energy costs and enhancing price transparency [4], [5].

Privacy-centric approaches are crucial for building user trust and encouraging the adoption of renewable energy, as they use secure, anonymous transactions that optimize energy distribution and maintain grid stability [6]. By protecting the integrity of energy transactions and empowering prosumers, privacy-preserving mechanisms promote efficient resource allocation and foster trust among stakeholders [7]. This collaborative environment not only reduces dependency on centralized grids but also facilitates the development of innovative, community-centric solutions tailored to local energy needs.

Figure 1 shows the P2P energy trading scenario whereby customers can trade energy among themselves and the utility grid directly. Trading of energy can be inferred as

The associate editor coordinating the review of this manuscript and approving it for publication was Behnam Mohammadi-Ivatloo.

an optimization problem [8]. Herein, deep reinforcement learning (DRL), a subfield of machine learning, is utilized to optimize the effectiveness of energy management [9], [10], [11]. P2P energy trading can be formulated as a multi-agent reinforcement learning (MARL) problem [12]. An agent in an environment has a policy that contains a learning algorithm and a neural network (NN) that can take action to maximize the goal of reducing the cost of electricity in a double auction (DA) market energy trading environment [13].

In Figure 2, the MARL cycle paradigm is demonstrated. Each agent receives observations and takes actions based on their policies in the same environment. Then, the environment provides the next observations and rewards for each agent. The training schemes in MARL are divided into three categories [14]: (i) centralized training centralized execution (CTCE)—all agents have the same policy, (ii) distributed training decentralized execution (DTDE)—all agents have their own policy, and (iii) centralized training decentralized execution (CTDE)—agents have their own policies. However, some of the information can be shared and seen by agents, including public information, such as grid buying and selling energy prices. As for energy trading, decentralized or independent policy algorithms are better suited [12], [15] due to restrictions on observing specific parameters for each agent [16]. Centralized policy algorithms share observation and action parameters, enabling agents to access the private data of other agents despite sharing information.

Most prior works in P2P energy trading have relied on information sharing among prosumers to optimize trading policies [17], [18], [19]. These sharing-based approaches often compromise privacy and security. In contrast, there are few studies that focus on non-sharing scenarios. One notable example is based on the Proximal policy optimization (PPO) algorithm but lacks any enhancements to the policy [20]. This approach has two main limitations: using basic techniques and being tested on a limited data setting. By focusing on the non-sharing aspect, our study addresses these limitations, providing a more robust and privacy-preserving solution for P2P energy trading.

We prioritize customer privacy by ensuring that private data, like energy usage and bid prices, is not shared. Our DTDE model enables customers to trade energy based on their own information only, both during training and testing. We use a policy algorithm, which includes an actor network to choose the best actions and a critic network to evaluate them. Policy algorithms can be policy-based (using only the actor network), value-based (using only the critic network), or actor-critic (using both networks).

There are two main types of policy algorithms: 'on-policy' and 'off-policy'. On-policy algorithms, like PPO [21], learn faster but might not always find the best solution. Off-policy algorithms, such as soft actor-critic (SAC) [22] and Twin-Delayed Deep Deterministic Policy Gradient (TD3) [23], can learn from past experiences and find better solutions, but they are more complicated to set up [24]. We chose PPO because it is simpler to use, more stable, faster to train, and can be easily customized [13], [25]. PPO is also used in tasks like improving ChatGPT by learning from human feedback [26], [27].

Our work is based on the Ausgrid dataset containing energy usage data from 300 households in the Sydney metropolitan area with installed solar cells [28]. Training 300 agents involved applying computational resources and memory. Using clustering techniques to group energy usage behavior helped reduce the number of agents needed for training while maintaining good results. This approach allowed training on regular computers [17]. Augmenting observations with forecasting weather parameters is seen to improve the efficiency of the results. However, the challenge lies in the accuracy of the weather forecast. Inaccuracies in forecasting when compared to actual data can significantly reduce the model's effectiveness, instead of enhancing it [12], [13], [14], [15], [16].

In this paper, we separate our work into three modules: (i) clustering, (ii) forecasting, and (iii) deep reinforcement learning (DRL). To categorize energy consumption, generation behavior and to solve the horizontal scaling problem, we apply a clustering technique [17]. Next, we employ a forecasting model to forecast 24h-ahead of net load energy by comparing long short-term memory (LSTM), DeepAR, and temporal fusion transformers (TFT) such that energy trading performance can be enhanced. Finally, we use DRL to train policy using the PPO algorithm in a decentralized training decentralized execution (DTDE) paradigm for P2P energy trading. This approach ensures that private information is not shared among agents during training and testing. The augmented global horizontal index (GHI) was added as additional public information to the observation parameters, enhancing the DRL performance. Additionally, we customized the policy network and compared the performance of various network architectures, including MLP, LSTM, and attention mechanisms, demonstrating significant improvements. This work is new and authentic in its application of DRL to P2P energy trading. The following is a summary of this paper's contributions:

- Implementation of a non-sharing information policy algorithm for enhancing P2P energy trading performance while maintaining the privacy of customers
- Enhancement of model-based performance to increase the accuracy of the 24h-ahead forecasting net load energy
- Improving the model by applying LSTM and attention to the policy network
- Integrating clustering techniques to optimize computation and memory usage solving the horizontal scaling problem

In this paper, Section II discusses the related work. Section III explains the framework, and methodology of this research. Section IV is the experimental setup of each module including the dataset. Section V shows the experimental

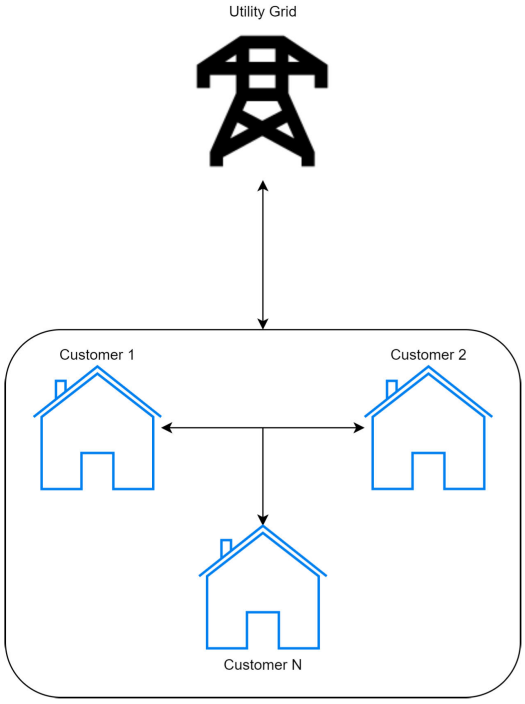

FIGURE 1. P2P energy trading allows customers within the same microgrid to trade with each other and the utility grid directly.

results for each task. Section VI provides an in-depth discussion of the results. Section VII discusses the conclusion and future work.

II. RELATED WORKS

In this section, we will discuss the development of P2P energy trading with MARL research. We start with the introduction of the double-side auction (DA) market environment for MARL. We follow by the application of the clustering technique to aggregate groups of energy consumption and generation behavior of customers and the net load forecasting model in order to enhance energy trading performance. Finally, the concern of private information is discussed.

A. INTRODUCTION OF DOUBLE-SIDE AUCTION (DA) MARKET ENVIRONMENT

Qiu et al. [18] introduced the double-side auction (DA) market multi-agent deep reinforcement learning (MARL) environment for P2P energy trading. Moreover, in their work, they used a modified multi-agent deep deterministic policy gradient (MADDPG) algorithm called DA-MADDPG. The purpose of DA-MADDPG was to solve the privacy concerns of prosumers and consumers due to a CTDE scheme of the MADDPG algorithm. As such, the centralized critic approximator collects all the observations and actions of agents, and aggregates the public information from the DA market as input for the centralized critic approximator instead. Thus, we adopted the P2P energy trading environment algorithm based on this research. However, since their experiment was conducted having only three prosumers and three consumers, we extended our dataset to include 300 customers.

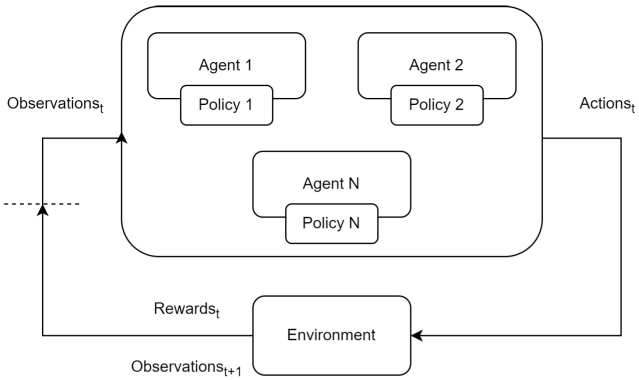

FIGURE 2. The multi-agent reinforcement learning (MARL) cycle paradigm.

B. INTRODUCTION OF MULTI-CLUSTER MB-MARL

Qiu et al. [19] also implemented MADDPG using a parameter sharing (PS) called PS-MADDPG to solve large-scale P2P energy trading between microgrid clusters. The PS framework allows all agents to share actor and critic network parameters, which lead to faster training performance, more stability, and better convergence speed. However, the PS framework is seen to have a security problem such that information regards observations and actions is shared among agents. It is noted that Sanayha and Vateekul [17] scaled the experiment by extending the Ausgrid dataset to 300 customers. Hence, a time-series clustering technique was applied by grouping the customer's energy consumption behavior patterns to reduce computational costs. This research also involved a model-based multi-agent reinforcement learning (MB-MARL) approach by applying a forecasting model to optimize energy-saving performance. However, the main algorithms applied in this research are MADDPG and actor centralized-critic with communication (A3C3), which have centralized critic networks that share all the observations and actions of all agents. While CTDE may perform better than DTDE, an information-sensitive environment like the DA market cannot be overlooked. MADDPG and A3C3 were chosen as benchmarks due to their established effectiveness in P2P energy trading used in prior works as sharing information policy models. Comparing them with our proposed non-sharing policy model highlights our approach's advantages in cost reduction, scalability, and privacy preservation.

C. PRIVACY AND SECURITY

Cao et al. [7] carried out a multi-agent energy trading using blockchain technology to protect the privacy and security of prosumers in microgrids. In essence, they introduced a reputable mechanism to evaluate prosumer's trustworthiness based on their past transactions. Their objective was to promote honest transactions between prosumers and energy trading numbers among each other. Wang et al. [6] applied the DTDE approach in P2P energy trading. They used deep deterministic policy gradient (DDPG) as a base algorithm instead of MADDPG due to the exposure of private information to

the public. They demonstrated a method using the encrypted action-observation memory of other agents provided by a cloud service provider. Besides, they also included weather parameters i.e. indoor and outdoor temperature, in the agents' observations. In our work, we also adopted the DTDE approach and additional weather parameters from the afore-mentioned research. Our work concentrated on the improvement of P2P energy trading without sharing private information. However, we did not focus on the application of the encryption technique.

May and Huang [20] applied P2P energy trading using a dynamic price signal mechanism in the MARL framework with the PPO algorithm. Their framework consisted of two layers: outer and inner RL. Inner RL is a P2P energy trading market where the agents trade energy with each other using a fixed price for an auction period. Outer RL is used for determining the price signal for trading in the next auction period. Their work applied PPO for P2P energy trading and respected data privacy, meaning no private data is shared among prosumers. The difference between their work and ours is the rule of trading price. Their work used a fixed trading price for all agents, while our work allowed agents to submit their trading price between time-of-use (ToU) and feed-in-tariff (FiT) prices.

In our research, we adopt multiple cluster model-based multi-agent deep reinforcement learning for P2P energy trading [17]. We also focus on the aspect of the non-sharing information scenario. Herein, the policy algorithm used for training conforms to the DTDE paradigm, or independent policy. Thus, each policy in a multi-agent environment will not share private information with each other. Moreover, we improve the model by applying state-of-the-art deep neural network (DNN) forecasting models: DeepAR and TFT. Further, we customized the policy network of the PPO learning algorithm in MARL with LSTM and attention, along with the additional weather parameter (GHI). We will compare our improvement with zero intelligence (ZI) [29]; results will be calculated via the test dataset, which assumes trading with the grid as the baseline.

III. METHODOLOGY

This section explains our overall framework. Herein, our aim is to optimize P2P energy trading with no information sharing between agents during training and testing.

A. OVERALL FRAMEWORK

Herein, the overall framework consists of three modules: (i) clustering, (ii) forecasting, and (iii) deep reinforcement learning (DRL). In a mixed-cooperation scenario, sharing information can result in better performance. Sharing information, however, can violate privacy and security. In order to improve the performance of the DTDE paradigm, we set out to improve the version of DRL for P2P energy trading with respect to private information.

In Figure 3, the overall framework of this research is illustrated. In Module 1, the training dataset is used in the clustering task to find the best number of clusters, categorizing the behavior of each cluster. The clustering outputs are the clustered training dataset, aggregating the training dataset by the number of clusters. Customers in the testing dataset are mapped to match the cluster based on the argmax method for the time-series of each customer. Next, in Module 2, the clustered training dataset from the clustering task and the clustered testing dataset are trained in the forecasting model to forecast the net load energy time-series 24h-ahead.

In Module 3, the clustered model-based training dataset, and the clustered model-based testing dataset are further augmented with the weather parameter. In this module, the DA energy trading market environment is used to train and evaluate the agents regards the policy algorithm. The result is the average net daily electricity cost for the 300 customers.

B. MODULE 1: CLUSTERING

We apply clustering to time-series data in order to categorize energy consumption and generation behavior into specific groups. The objective of this module is to reduce the number of trainable agents from hundreds of millions of customers to only a few clusters. This will help with horizontal scaling while training agents on a single computer. The K-means algorithm using dynamic time warping (DTW) is used to cluster time-series. Both elbow method and silhouette scores are used in selecting the best number of clusters.

1) K-MEANS

K-means is an unsupervised machine learning algorithm [30]. K-means clusters data by attempting to divide samples into distinct groups based on their similarity. The goal of this algorithm is to minimize the within-cluster sum of squares (WCSS) between each datapoint within a cluster and the centroid of that cluster. Smaller WCSS values indicate tighter and more compact clusters.

2) EUCLIDEAN DISTANCE AND DYNAMIC TIME WARPING (DTW)

Euclidean distance is the amount of space in Euclidean space that separates two locations in a straight line. It is the length of the shortest path in a Cartesian coordinate system between two points. DTW is a method for comparing two temporal sequences that may have different speeds. DTW is beneficial when comparing data points in sequences with various lengths or temporal abnormalities. The goal of DTW is to warp the time axis non-linearly to find the best alignment between the points of the two sequences. This enables point-to-point comparisons by aligning every point in one sequence with a point in other sequences, regardless of how long or how quickly the sequences change over time. We adopted DTW as our metric due to [17] revealing that DTW has a better performance than Euclidean distance.

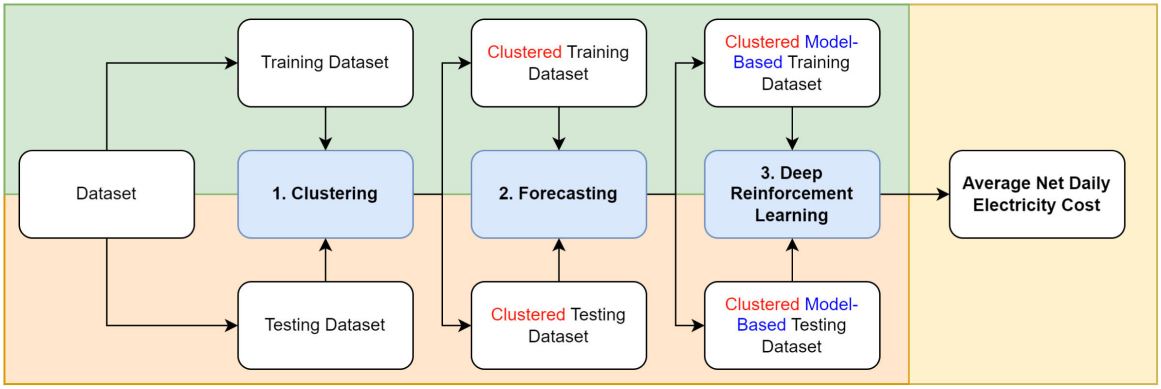

FIGURE 3. The overall framework of the research consists of three modules: starting with (i) clustering time-series data to aggregate energy consumption and generation behavior, (ii) forecasting to predict 24h-ahead net load, and (iii) deep reinforcement learning to optimize P2P energy trading, resulting in a lower average daily net energy cost.

3) ELBOW METHOD

The elbow method is a technique used in cluster analysis to determine the optimal number of clusters. In this method, the average distance is plotted on the y-axis and the number of clusters on the x-axis. As the number of clusters increases, the average distance tends to decrease because the points are closer to their respective centroids. However, at some point, the decrease in the average distance will slow down significantly, creating an "elbow" in the graph. The number of clusters at which this change occurs is considered the optimal number for clustering. The approach is based on computing the within-cluster-sum of squared errors (WCSS) for various cluster densities (k).

4) SILHOUETTE SCORE

A silhouette score is a measure used in clustering analysis to determine the degree of separation between clusters. It ranges from -1 to 1 where a high value indicates that the object is well matched to its own cluster and poorly matched to neighboring clusters. If the average silhouette score of all points in a cluster is close to $+1$, the cluster is dense and well-separated from other clusters. A value around 0 indicates overlapping clusters, and a value close to -1 means that points have been assigned to the wrong cluster.

C. MODULE 2: FORECASTING

Time-series forecasting is a method for predicting future values based on past observations of a time-dependent sequence of data points. Time-series data frequently show seasonal fluctuations, patterns, and trends that can be observed and used for forecasting. Recurrent neural networks (RNNs) from deep learning have been used to forecast time-series, and they beat statistical techniques like ARIMA (auto-regressive integrated moving average) [31].

1) LSTM

Long short-term memory (LSTM) [32] is a recurrent neural network (RNN). LSTM was created to solve the vanishing gradient problem and effectively capture long-term dependencies in sequential data. Both gates and a cell state are two essential parts of the LSTM network that aid learning, and manage information over lengthy sequences.

2) DEEPAR

DeepAR is an autoregressive recurrent neural network (RNN) based on LSTM cells by Amazon [33]. DeepAR excels in datasets with multiple interrelated time-series, offering not just point forecasts but also probabilistic predictions to gauge uncertainty effectively. This model is in stark contrast to traditional time-series models that typically focus on single series in isolation. Furthermore, DeepAR supports the incorporation of additional covariates and categorical variables, allowing the model to account for external influences and time-dependent features, such as holidays, promotions, or price changes, thereby contextualizing forecasts. DeepAR uses RNN to create a likelihood model where the network predicts the mean and standard deviation that are best suited for the probability distribution of the target data for the next time step.

3) TEMPORAL FUSION TRANSFORMERS (TFT)

The temporal fusion transformers (TFT) [34] is a transformer-based deep learning model designed specifically for multi-horizon forecasting time-series data. TFT is expert at capturing complex temporal relationships and accommodating diverse data types. TFT excels in handling multivariate inputs, distinguishing between static (time-invariant) and dynamic (time-variant) features, selectively leveraging relevant information through its attention mechanism. Thus, TFT enables highly accurate and interpretable forecasts. The model has the ability to discern and utilize temporal patterns. Combined with its interpretability features like variable selection, TFT is particularly valuable for applications across various domains requiring robust and insightful time-series analysis.

In Figure 4, the architecture of TFT is shown. Key components of the transformer architecture within TFT include multi-head self-attention mechanism, positional encodings, layer normalization, and feed-forward networks. The multi-head self-attention mechanism allows the model to focus

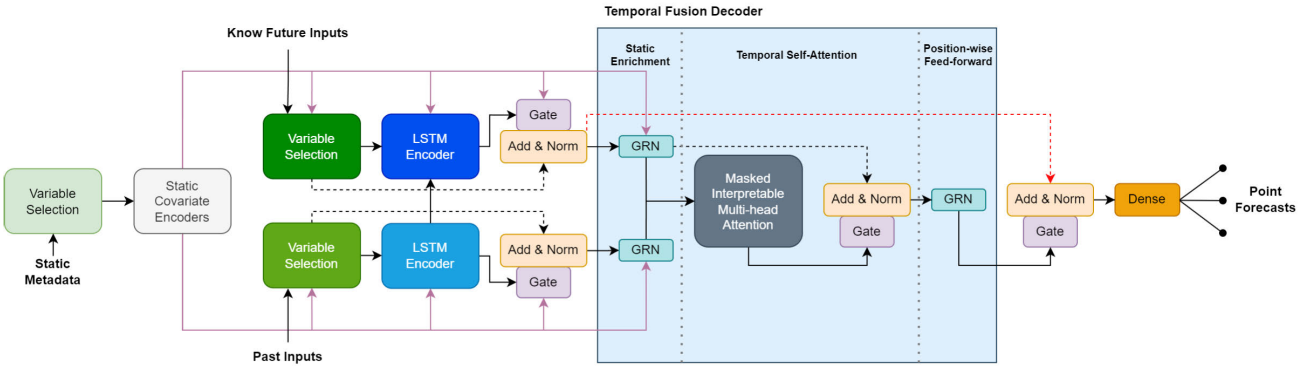

FIGURE 4. TFT architecture [34]. The model incorporates static metadata, time-varying inputs, and future time-varying inputs that are known beforehand. Variable Selection is deployed for choosing the most important features from the inputs. Gated Residual Network blocks facilitate effective flow of information through skip connections and gating layers. TFT combines LSTMs, which handle local temporal processing, with multi-head attention mechanisms that allow the model to consider information from any time points.

on different time steps and different features, understanding the underlying patterns and how they interact with each other over time. Positional encodings are added to the input embeddings to provide the model with information about the position of data points in the time-series. Both layer normalization and feed-forward networks help in stabilizing the learning process and allow for each layer to learn the function of the residuals of the previous layers. Hence, the model is able to learn more complex functions. In the TFT paper, it was compared with many algorithms, e.g., ARIMA and DeepAR. The results showed that TFT significantly surpassed ARIMA and slightly outperformed DeepAR in terms of forecasting accuracy. Therefore, apart from LSTM, we decided to include TFT and DeepAR, but did not use ARIMA in our experiment.

D. MODULE 3: DEEP REINFORCEMENT LEARNING (DRL)

This research uses the DRL approach for P2P energy trading, focusing on a non-sharing information multi-agent scheme. First, we describe the importance of non-sharing information for P2P energy trading. Then, we explain the advantages and concept of the state-of-the-art algorithm (PPO), which is used as the learning algorithm for agents. The last section explains the functions of the model-based P2P energy trading environment, including observation, action, and reward. The algorithm of the whole framework of this research, called the clustered model-based P2P energy trading algorithm, is also explained.

1) NON-SHARING INFORMATION MULTI-AGENTS

In Figure 5, the difference between non-sharing and sharing information MARL is demonstrated. In a non-sharing or independent MARL approach, each agent operates independently. In essence, each agent makes decisions based on their own observations and rewards without coordinating with other agents. When privacy is a concern in P2P energy trading, agents (consumers or prosumers) might not want to share proprietary information about their consumption, generation, or trading strategies: non-sharing MARL is more suitable. Non-sharing MARL has more scalability and simplicity than sharing MARL. Agents can operate independently without the need for extensive communication or coordination with others. Therefore, there will be no communication overheads. Non-sharing MARL is also modeled and trained independently, meaning that the complexities involved in multi-agent coordination, negotiation, or communication protocols are inherently absent, simplifying the design and implementation and making it more useful in a real-world application. The only downside of non-sharing MARL is that the result might be suboptimal due to the lack of communication compared to sharing MARL. However, this problem can be solved since non-sharing MARL is easier to implement for optimization, leading to the performance better than sharing MARL.

2) PROXIMAL POLICY OPTIMIZATION (PPO)

PPO [21] is a popular 'on'-policy reinforcement learning algorithm used in DRL for training agents to make decisions in an environment. The highlight of PPO is a balance between sample efficiency, stability, and ease of implementation. PPO belongs to the class of policy gradient algorithms, which aim to optimize the policy directly. The main concept of PPO is to make small updates to the policy while ensuring that the new policy does not deviate significantly from the old one. This helps in stabilizing the learning process and prevents drastic policy changes that might lead to instability. The proximal part in PPO refers to the use of a clipped objective function that constrains the policy update to a certain range. By placing a limit on how much the updated policy can differ from the old one, PPO prevents large policy changes that could be detrimental to learning. The algorithm generally works by collecting data from interactions with the environment, then it uses this data to compute advantages (how good or bad certain actions are in comparison to the expected outcome) and update the policy accordingly. The objective is to maximize the expected reward.

3) MODEL-BASED P2P DOUBLE AUCTION ENERGY TRADING ENVIRONMENT

The agents that are trading in the P2P DA market energy trading environment [18] as shown in Figure 6 with the

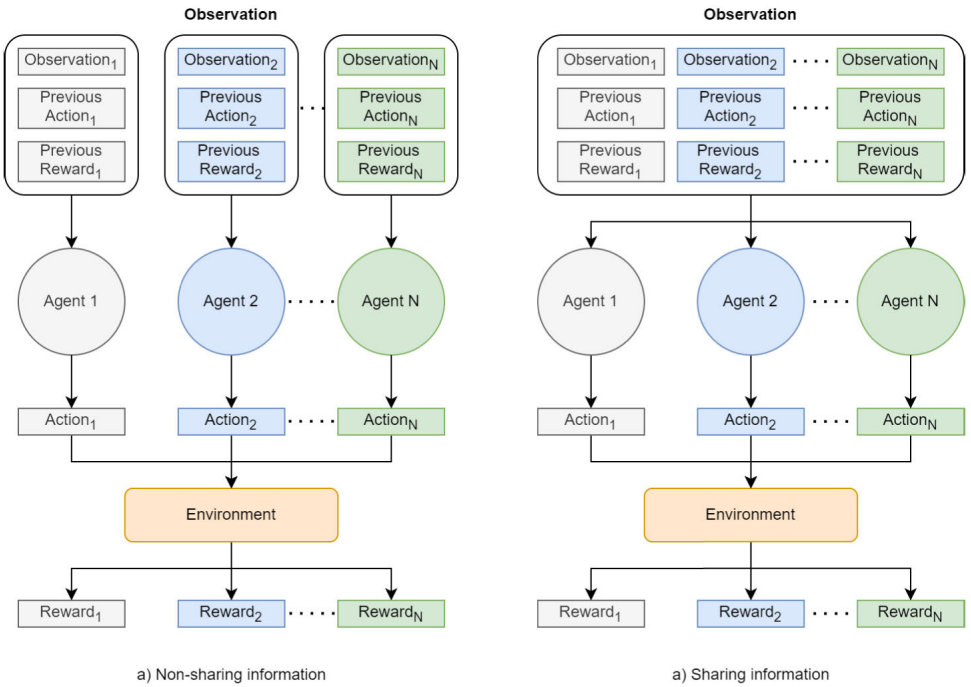

FIGURE 5. Comparison between non-sharing (left) and sharing (right) information MARL for one cycle. Non-sharing or independent agents operate independently based on their own observations and rewards. No private information is shared in this scenario, which is more suitable for privacy-related tasks like P2P energy trading. Sharing information MARL has a benefit from communication between agents. Private information of other agents, including observation (consumption, generation, and energy storage), action (trading price and quantity), or even reward (energy cost), can be seen by every agent, compromising privacy and security.

observation $o_{n,t}$ at time t can be expressed as:

$$o_{n,t} = \left(L_{n,t},\ L_{n,t+1},\ E_{n,t},\ \lambda_t^s,\ \lambda_t^b,\ W_t\right) \tag{1}$$

where $L_{n,t}$ is the net load energy at time t, $L_{n,t+1}$ is the predicted net load energy at time $t+1$ from the forecasting model, E_t is the energy storage at time t, λ_t^s is the grid selling price (FiT) at time t, λ_t^b is the grid buying price (ToU) at time t, and W_t is the weather parameter at time t.

In Figure 7, after agents receive their observations, the policy mapping function maps out policies to match the agent's clusters based on their energy consumption and generation behavior. Then, the agents pursue their actions based on their policy $\pi(a_t|o_t)$. The action $a_{n,t}$ of each agent can be expressed as:

$$a_{n,t} = \left(P_{n,t}, Q_{n,t}\right) \tag{2}$$

where $P_{n,t}$ is the trading price magnitude submitted to the DA market ranges (from 0 to 1): 0 means keeping the energy and 1 means selling energy at the maximum with the cap of 2 kWh per auction period. $Q_{n,t}$ is the quantity of energy magnitude submitted to the DA market, which ranges from -1 to 1: a negative value means discharging or selling, and a positive value means charging or buying.

After all agents submit their actions, the DA market starts the auction with the clearing algorithm and publishes the public order book outcome, including trading prices and quantities. The cleared orders count as internal trading between agents, and the remaining orders are traded with the grid. The reward $r_{n,t}$ is the negative of the electricity cost and can be calculated as follows:

$$r_{n,t} = -\left(\sum_{i=0}^{k} \lambda_{n,t}^i q_{n,t}^i + \lambda_t^b q_{n,t}^b - \lambda_t^s q_{n,t}^s\right) \tag{3}$$

where $\sum_{i=0}^{k} \lambda_{n,t}^i q_{n,t}^i$ represents the sum of internal trading from cleared order i to k where $\lambda_{n,t}^i$ and $q_{n,t}^i$ denote the trading price and trading quantity of order i. Additionally, $\lambda_t^s q_{n,t}^s$ and $\lambda_t^b q_{n,t}^b$ signify the external trading cost where $q_{n,t}^s$ is the quantity sold to the grid, and $q_{n,t}^b$ is the quantity bought from the grid.

In Algorithm 1, we explained the whole process of this research. Starting with clustering time-series training data, aggregated k clusters use the argmax method to simplify the model to assign a cluster number for each customer. As such, there is no significant difference in performance between argmax and predicted clusters from forecasting time-series for this dataset, making the framework simpler and more straightforward. Then, we trained two separate forecasting models for the aggregated time-series and individual time-series for training and testing, respectively, to forecast 24h-ahead net load energy. For training in the P2P DA energy trading environment, clustered agents trade their energy for 24h consecutively per episode by random starting datetime index with 1h increments; policies are updated until $episodes = T_{max}$. For testing, each customer trades

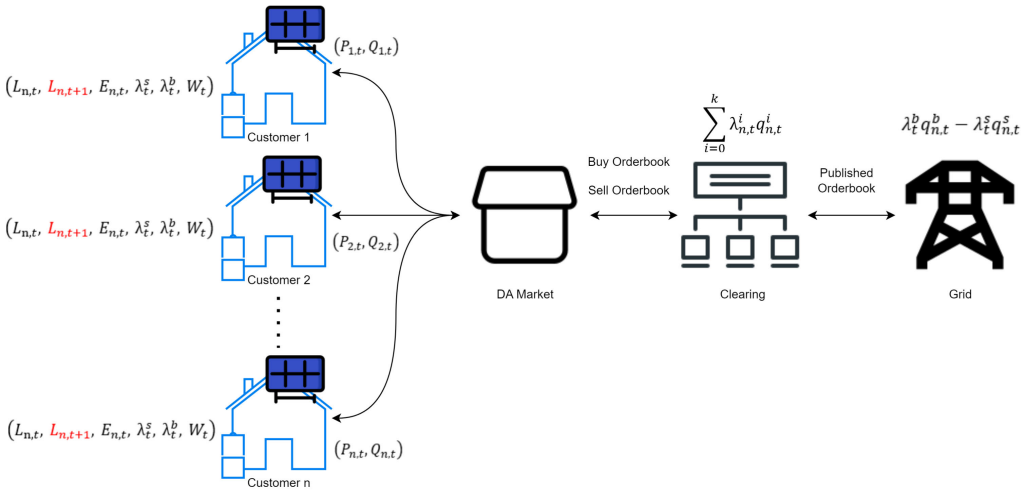

FIGURE 6. P2P Double Auction Energy Trading Environment. Observation parameter, $L_{n,t+1}$ (red), is the predicted net load from the forecasting model.

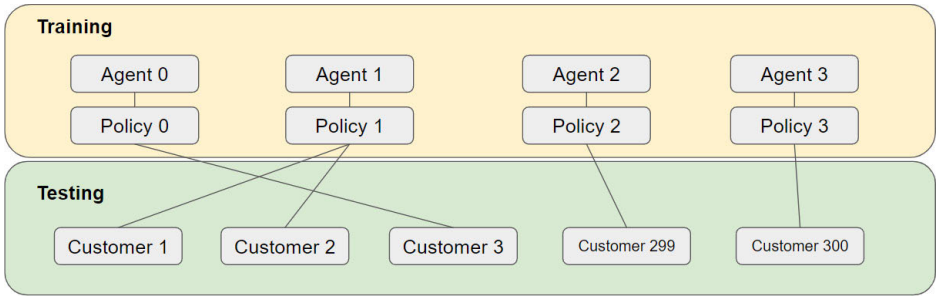

FIGURE 7. During the training phase, all customers are aggregated into 4 clusters (agents 1-4) based on their energy consumption and generation behavior. During the testing phase, customers decide their actions based on their mapped policies.

their energy according to the assigned cluster number, which indicates the policy to be used for submitting trading energy quantity and price. During this phase, all policies have stopped updating. The environment is simulated as real-world energy trading. All customers trade continuously for the whole testing dataset timesteps.

IV. EXPERIMENTAL SETUP

In this section, we explain the main dataset and the augmented dataset used in our research, as well as the implementation details of each module from the previous section. All of the important hyperparameters are displayed in Appendix B.

A. DATASET

The main dataset used in this research is the dataset from Ausgrid [28], sourced from 300 customers in Ausgrid's electricity network. The period is one year, from July 2012 to June 2013, with a 70:30 train-test ratio. The data contains electrical energy consumed and generated every half an hour (aggregated to 1h). The dataset also contains generator capacity and postcodes. In Figure 8, the visualization of the Ausgrid dataset is observed. The yellow dome in the middle (6:00 to 18:00) displays energy generation. The light blue line indicates the energy consumption, which shows the shoulder between 6:00 and 16:00 and the peak between 17:00 and

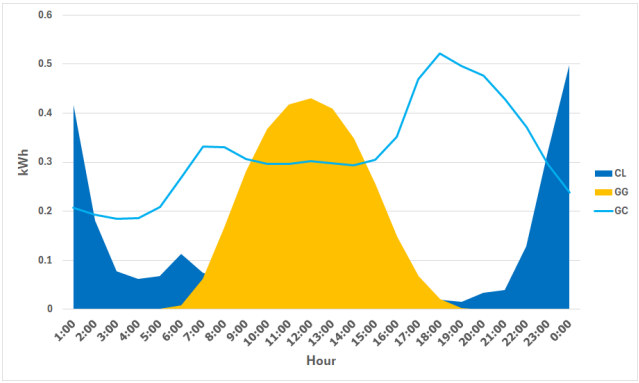

FIGURE 8. The visualization of the Ausgrid dataset during July 2012 to June 2013 showing the characteristics of energy consumption (GC), control load consumption (CL), and energy generation (GG).

22:00. The blue area displays the control load, which refers to the usage of high-load appliances, such as heaters and hot water systems. The peak of the control load occurs from 21:00 to 3:00.

The global horizontal irradiance (GHI) is a term used in the fields of solar energy and meteorology to describe the total amount of solar radiation received at the Earth's surface on a horizontal plane. It represents the total solar energy that reaches a specific location over a specific time period,

Algorithm 1 Clustered Model-Based P2P Energy Trading Algorithm

1: Time-series clustering on training dataset with DTW method from n customers to k clusters
2: Aggregate training time-series dataset to k clusters
3: Argmax each customer's time-series training dataset to map the policy based on the cluster number
4: Train forecasting models for individual n customers and N clusters
5: Initialize policy configuration and maximum training episode $= T_{max}$
6: Initialize P2P DA energy trading environment for training
7: **for** *episode* $\leftarrow$ 1 to T_{max} **do**
8: Random datetime index
9: **for** $t \leftarrow$ 1 to 24 **do**
10: **for** *agent* $\leftarrow$ 1 to N **do**
11: Submit a_t based on policy $\pi(a_t|o_t)$
12: **end for**
13: Allocate buy and sell order book
14: Execute clearing algorithm and published cleared order book
15: **for** *agent* $\leftarrow$ 1 to N **do**
16: Receive r_t from cleared order book
17: **end for**
18: **if** $t = 24$ **then**
19: Reset the environment
20: **end if**
21: **end for**
22: Update the policies
23: **end for**
24: For testing, repeat 6 to 23, remove the random datetime index, environment reset, and policy update

measured in watts per square meter (W/m^2) [35]. GHI is used as additional weather information in this research, obtained from the National Solar Radiation Database (NSRDB) [36].

B. MODULE 1: CLUSTERING

A time-series machine learning analysis tool called TSlearn [37] has been applied to cluster and map the energy consumption behavior of the customers. The clustering algorithm used in this research is K-means. To find and visualize the best number of clusters, we compare the results with the elbow method and the silhouette scores using dynamic time warping (DTW) as the metric, with clusters ranging from 2 to 10.

C. MODULE 2: FORECASTING

We compare the forecasting models between LSTM, DeepAR, and TFT. Pytorch-Forecasting is the main framework for constructing the forecasting models to forecast 24h-ahead net load energy. The variates for all models, including hour, day of the week, and month, are obtained using the cyclical index transformation. The time-series for each group is normalized by the group normalizer. The learning rate is 0.001 with the seven-day (168h) lookback. The optimizer is Adam, training the models with 20 epochs using the last day as validation data. The last day in the training dataset is used for the forecasting task as validation. The accuracy of the models is evaluated by the root mean square error (RMSE) method.

D. MODULE 3: DEEP REINFORCEMENT LEARNING (DRL)

We train and evaluate the experiment for DRL under the Ray RLlib [25] framework. The discharging and charging efficiency of the energy storage battery is 95%, with the minimum storage at 2 kWh and the maximum at 10 kWh. The energy storage of all agents at $t = 0$ is 6 kWh. For each training episode, consisting of 24 steps, the initial setting of the date and time index at step 0 is randomized. With each subsequent step, the time index is incremented by one until it reaches 24 steps. Upon completion of these steps, the environment resets, and the date and time index are randomized again. This cycle repeats until the reward from the model stabilizes and converges. A limit of 4,000 episodes has been set for this training process. The experiment assessed the performance of the proposed method against several baselines, including direct trading with the grid and the zero intelligence (ZI) approach. All agents were required to execute actions entirely at random, without consideration of the environmental state or any accumulated knowledge from past experiences. This methodology was also applied in the evaluation of the MADDPG and A3C3 algorithms. We also extended the experiment by changing the policy network from multi-layer perceptron (MLP) to LSTM and attention [38], [39]. PPO relies on current information and doesn't effectively utilize past observation data. LSTM and attention mechanisms are useful when agents need to base their decisions on long-term dependencies or when they only have partial observations. This improvement is expected to enhance the PPO algorithm. Performance was evaluated by the average daily electricity cost of 300 customers with an average of three runs.

V. EXPERIMENTAL RESULTS

In this section, we will discuss the experimental results: the clustering module for selecting the best K clusters, the comparison of the accuracy of the forecasting models in the forecasting module, and the optimization of MB-MARL.

A. MODULE 1: CLUSTERING

We used four clusters for the training dataset [17]. In Figure 9, illustrates the results of two cluster validation methods: the silhouette score on the left and the elbow method on the right. The silhouette score assesses the cohesion and separation of clusters, with the highest value at $k = 4$, suggesting optimal cluster definition at this number. Concurrently, the elbow method, which evaluates the within-cluster sum of squares, demonstrates a pronounced bend at $k = 4$, indicating

TABLE 1. Root Mean Square Error (RMSE) comparison of forecasting models between LSTM, DeepAR, and TFT for 300 customers. Boldface refers to the winner. % difference column is the difference compared with LSTM.

Model	kWh	% Diff.
LSTM	0.5438	0%
DeepAR	0.4999	8%
TFT	**0.4296**	**21%**

that increasing the number of clusters beyond this point yields minimal improvement in clustering quality. Both methods independently corroborate $k = 4$ as the most effective number of clusters. Figure 10 shows four separate clusters, each depicting variations in energy consumption and generation over time. The horizontal axis in each panel represents time, while the vertical axis measures energy values—positive values for consumption and negative values for generation. The bold red line in each graph likely represents the average or median trend across the data points. Cluster 1 shows minimal fluctuations, primarily indicating steady energy consumption. Cluster 2 exhibits a slight upward trend, suggesting increasing energy consumption. Cluster 3 has the most pronounced variation, showing significant dips into negative values, which implies substantial energy generation during daytime, surpassing consumption. Cluster 4 shows considerable volatility with both peaks and troughs, indicating periods of high energy consumption as well as significant generation. In Figure 11, an overview of energy consumption and generation in the dataset is given. Cluster 2 is seen to have the highest number reaching 34,502, followed by Cluster 1 (21,597), Cluster 4 (11,536), and Cluster 3 (9,165).

B. MODULE 2: FORECASTING

The RMSE between the clustered training dataset and the individual 300 customers training dataset is shown in Table 1. TFT performed best being 21% better than LSTM at 0.4296. Likewise, DeepAR performed 8% better than LSTM at 0.4999. Figure 12 illustrates the validation results for 24h-ahead predictions using LSTM, DeepAR, and TFT models across two scenarios: 4 clusters and 300 customers, employing a 168h lookback period. All models performed well on the clustered training dataset due to the aggregation of data, resulting in an easier pattern for prediction. Among the models, TFT and DeepAR outperformed LSTM, largely due to their superior handling of multiple time-series data, capturing complex temporal dependencies more effectively. The TFT model, in particular, demonstrated the highest proficiency, benefiting from its advanced architectural features like variable selection mechanisms and adaptive gating layers, which optimize responsiveness to changing data patterns.

C. MODULE 3: DEEP REINFORCEMENT LEARNING (DRL)

1) BASELINES EVALUATION

In Table 2, the average daily net electricity cost is shown. When compared with baselines (MADDPG and A3C3), our model achieved the most substantial reduction in average daily net electricity costs. Specifically, it outperformed MADDPG by 4.20% and A3C3 by 3.95%, with our costs amounting to \$460 versus \$479.88 and \$478.70 for MADDPG and A3C3, respectively. Our best model is seen to save electricity costs by 12.48%, 15.29%, and 7.14% when compared with grid trading (\$526.20), random ZI (\$543.63), and non-MB PPO (\$495.91), respectively.

2) EFFECT OF THE ACCURACY OF FORECASTING MODELS IN MB-MARL

As observed in Table 1, TFT had the lowest RMSE value, followed by DeepAR and LSTM. In Table 2, all cases using TFT resulted in better average daily net electricity cost by 0.90% and 2.64%, on average, compared with DeepAR and LSTM, respectively.

3) EFFECT OF INCORPORATING GHI

In Table 2, we have presented the electricity costs with and without the inclusion of global horizontal irradiance (GHI). The PPO with MLP as the policy network and LSTM model as the forecasting model exhibits a decrease in costs from \$487.55 to \$484.27 (0.67% reduction). A shift in the forecasting model from LSTM to DeepAR and TFT results in reductions from 481.60 *to*478.49 for DeepAR (0.65% reduction), and from \$475.03 to \$473.16 for TFT (0.39% decrease). These figures underscore the subtle yet consistent impact of integrating GHI on reducing electricity costs across various models.

4) EFFECT OF THE POLICY NETWORK MODELS

Changing the policy network from MLP to LSTM further reduced the average daily net electricity cost by 2.63%. LSTM leverages, between external and internal trading best, resulted in the lowest average daily net electricity cost. In contrast, using attention increased the average daily net electricity cost by 0.74%. However, attention helped to increase internal trading. When using TFT as the forecasting model, internal trading value reached 123.13 kWh.

VI. DISCUSSION

Our work addresses important aspects of P2P energy trading, focusing on making it more scalable and efficient while maintaining private information. We applied clustering technique to group customers into smaller sets, which helps in handling large numbers more effectively during training. The TFT model achieved better accuracy in predicting the 24h-ahead net load while GHI provides PV generation information. These two parameters effectively inform the agents to make better trading decisions. Furthermore, we applied LSTM to enhance the decision-making process of our model, leading to further reductions in electricity costs.

A. IMPORTANCE OF COST REDUCTION IN P2P ENERGY TRADING

The proposed model achieved a 4.20% and 3.95% reduction in microgrid electricity costs relative to MADDPG and

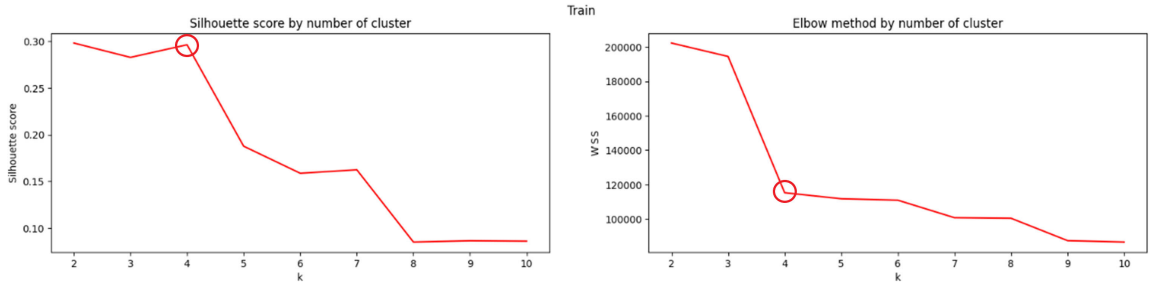

FIGURE 9. Comparison of results between silhouette score (left) and elbow method (right). Both reveal similar results, with $k = 4$ as the best number of the cluster groups.

TABLE 2. Comparative analysis of average daily net electricity costs ($), external and internal trading quantity (kWh) and value ($), calculated over three separate runs. for various combinations of policy networks, forecasting models, and weather parameter (GHI). Boldface refers to the winner.

Policy	Policy Network	Forecast. Model	Weather Param.	Quantity (kWh)			Value ($)			Electricity Cost ($) (↓)
				External (↓)		Internal (↑)	External (↓)		Internal (↑)	
				Buy	Sell		Buy	Sell		
Grid	-	-	-	4,512.23	620.79	0	493.87	26.14	0	526.20
ZI	-	-	-	4,993.95	789.58	**1,374.75**	575.21	31.58	113.37	543.63
PPO	MLP	-	-	4,510.41	492.97	547.28	515.63	19.72	34.92	495.91
PPO	MLP	LSTM	-	4,459.13	411.66	604.42	504.02	16.47	48.03	487.55
PPO	MLP	DeepAR	-	4,362.88	378.91	298.92	496.76	15.16	21.99	481.60
PPO	MLP	TFT	-	4,259.44	315.57	351.90	487.66	12.62	32.64	475.03
PPO	MLP	LSTM	GHI	4,310.49	367.12	305.56	499.00	14.68	19.47	484.32
PPO	MLP	DeepAR	GHI	4,292.39	306.14	486.64	490.64	12.25	44.20	478.39
PPO	MLP	TFT	GHI	**4,216.97**	292.36	311.80	484.91	11.69	32.65	473.22
PPO	LSTM	LSTM	GHI	4,302.16	378.71	280.62	489.85	15.15	27.38	474.71
PPO	LSTM	DeepAR	GHI	4,245.31	**293.03**	525.93	474.74	**11.72**	60.07	463.02
PPO	LSTM	TFT	GHI	4,316.13	350.91	450.93	**474.55**	14.04	49.78	**460.52**
PPO	Attention	LSTM	GHI	4,423.11	452.83	494.85	507.23	18.11	54.19	489.12
PPO	Attention	DeepAR	GHI	4,333.64	369.80	540.23	496.10	14.79	101.26	481.31
PPO	Attention	TFT	GHI	4,288.25	304.97	675.29	488.35	12.20	**123.13**	476.15
MADDPG	MLP	TFT	GHI	4,282.35	379.33	241.45	495.05	15.17	19.25	479.88
A3C3	MLP	TFT	GHI	4,269.41	382.60	162.81	494.00	15.30	11.71	478.70

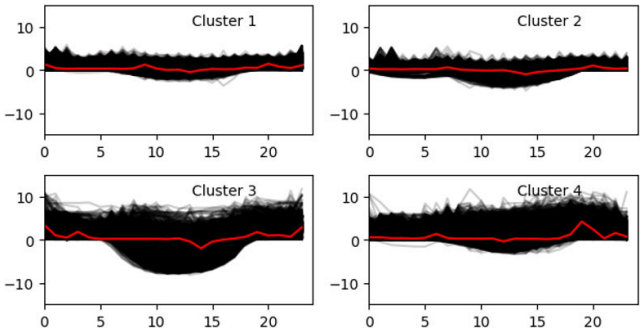

FIGURE 10. Visualization of 4 clusters showing energy consumption (positive) and generation (negative).

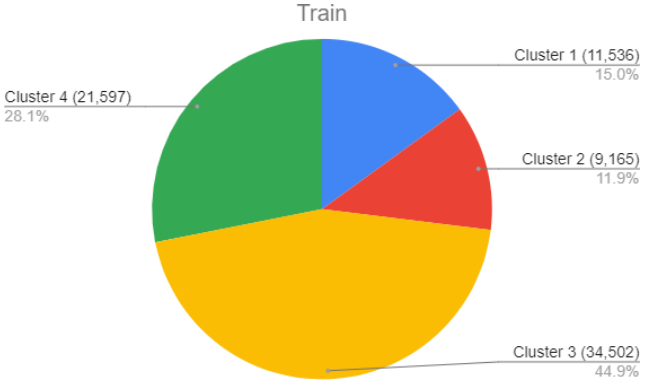

FIGURE 11. Population of each cluster in the training dataset.

A3C3, respectively, and a substantial 12.48% reduction when compared to direct utility grid trading. These cost reductions are noteworthy as they significantly lower the energy expenses for end-users while maintaining the privacy of user information, thereby enhancing the economic viability of peer-to-peer (P2P) energy trading as an alternative to traditional energy procurement methods.

From an efficiency perspective, these cost savings underscore the model's capacity to optimize energy distribution within the microgrid, effectively balancing supply and demand while reducing dependency on the utility grid. By reducing transaction costs and improving energy management, the overall efficiency of the microgrid is enhanced. This efficiency translates to fewer energy losses and more stable grid operations, which are critical for the feasibility and scalability of P2P trading systems. Moreover, these cost reductions foster the adoption of renewable energy sources by making decentralized energy trading more attractive and

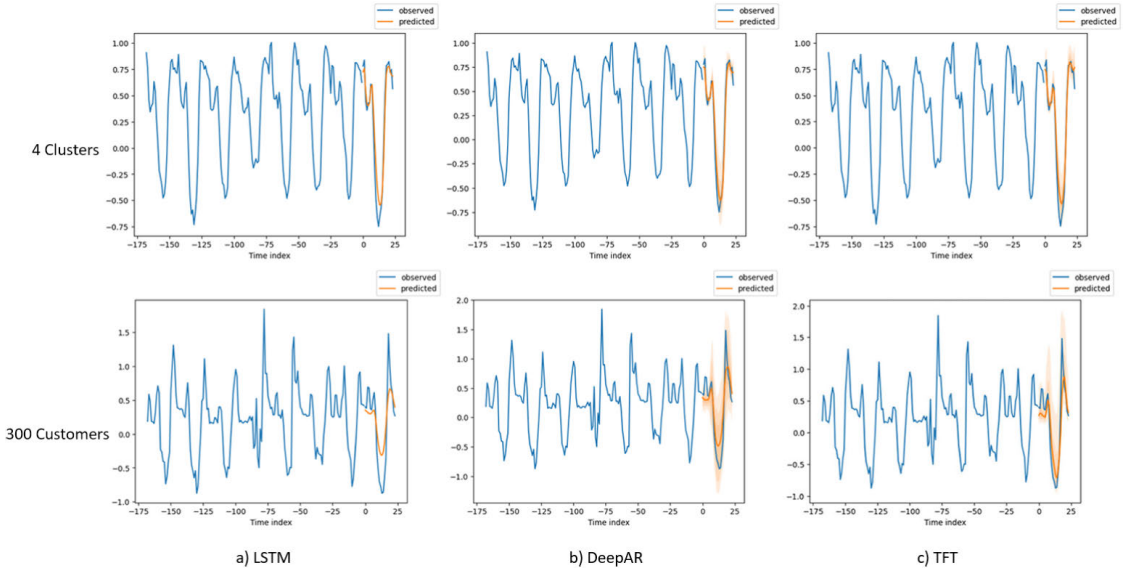

FIGURE 12. Validation results for the 4 clusters and 300 customers: each row reveals 24h-ahead prediction (orange) with 168h lookback period (blue) following LSTM, DeepAR, and TFT models.

financially sustainable for prosumers, thereby promoting a more resilient and sustainable energy ecosystem.

B. SCALABILITY OF THE PROPOSED MODEL

We applied a clustering technique to categorize energy consumption and generation patterns into four distinct clusters. This approach significantly reduces computational resource requirements. Increasing the number of clusters does not yield a notable improvement in performance. Consequently, our proposed model demonstrates the capability to maintain its performance without necessitating re-training, even as the population size varies.

C. PRACTICAL IMPLEMENTATION IN REAL-WORLD ENERGY TRADING SYSTEMS

While our work demonstrates the concept of scalability and cost reduction performance in real-world energy trading systems, one of the main challenges is the integration of the model with existing energy management systems and infrastructure. Ensuring compatibility and seamless data transfer between the model and the hardware/software used in energy trading systems is crucial. Moreover, real-world deployment will require rigorous testing to ensure the reliability and security of the model in live environments. Solutions to potential challenges could include developing standardized protocols for transactions, enhancing the model's robustness against cyber threats, and creating user-friendly interfaces for stakeholders to interact with the system.

D. FUTURE RESEARCH DIRECTIONS

Future research in P2P energy trading with MARL can focus on enhancing data privacy and security, particularly with blockchain integration, to prevent cyber-attacks. Improving the computational efficiency and scalability of MARL algorithms is essential for handling larger networks and real-time decision-making. MB-MARL can enhance the predictive accuracy of energy supply and demand, promoting better trading performance. Exploring the socio-economic impacts on different community scales and regulatory environments will provide insights for policy development. Finally, developing frameworks for interoperability among various renewable energy sources and storage systems can maximize renewable utilization and reduce grid dependency.

VII. CONCLUSION

In this paper, we developed a method based on multi-agent deep reinforcement learning for P2P energy trading within the microgrid. Our approach is seen to enhance the MB-MARL algorithm in the non-sharing information scenario providing the same or better performance compared with sharing information. Results show that increasing the policy network using LSTM and adding the weather parameter (GHI) boosts performance. The accuracy of the forecasting model also has a significant impact on the outcomes. The study employed the clustering technique to solve the horizontal scaling problem on the Ausgrid data set, which included 300 Australian households. Applying PPO as the learning policy algorithm, LSTM as the policy network, TFT as the forecasting model, and GHI as the weather parameter, the average daily net electricity cost is reduced to \$460.52, 4.20% lower than MADDPG, 3.95% lower than A3C3, and 12.48% lower than trading directly with the utility grid.

For future work, we aim to extend our model to larger datasets and enhance P2P energy trading from within to between microgrids. We plan to integrate other renewable sources such as wind, and hydroelectricity. Additionally, we intend to utilize blockchain technology for secure, transparent transactions and assess the environmental impacts

of P2P energy trading, focusing on renewable adoption rates and carbon emissions.

APPENDIX A NOMENCLATURE

Abbreviations

A3C3	Actor centralized-critic with communication.
CTCE	Centralized training centralized execution.
CTDE	Centralized training decentralized execution.
DDPG	Deep deterministic policy gradient.
DRL	Deep reinforcement learning.
DTDE	Distributed training decentralized execution.
DTW	Dynamic time warping.
FiT	Feed-in-tariff.
GHI	Global horizontal index.
LSTM	Long short-term memory.
MADDPG	Multi-agent deep deterministic policy gradient.
MARL	Multi-agent reinforcement learning.
MB-MARL	Model-based multi-agent reinforcement learning.
P2P	Peer-to-peer.
PPO	Proximal policy optimization.
RNNs	Recurrent neural networks.
SAC	Soft actor-critic.
TD3	Twin-Delayed Deep Deterministic Policy Gradient.
TFT	Temporal fusion transformers.
ToU	Time-of-use.
WCSS	Within-cluster sum of squares.
ZI	Zero intelligence.

Symbols

λ_t^b	Grid buying price (ToU) at time t.
λ_t^s	Grid selling price (FiT) at time t.
λ_t^s	Internal trading price of the agent n at time t of order i.
$\lambda_{n,t}^i$	Internal trading price of the agent n at time t of order i.
$a_{n,t}$	Action of the agent n at time t.
$E_{n,t}$	Energy storage of the agent n at time t.
$L_{n,t+1}$	Predicted net load energy of the agent n at time $t+1$.
$L_{n,t}$	Net load energy of the agent n at time t.
$o_{n,t}$	Observation of the agent n at time t.
$P_{n,t}$	Trading price magnitude of the agent n at time t.
$Q_{n,t}$	Energy quantity magnitude of the agent n at time t.
$q_{n,t}^i$	Internal trading quantity of the agent n at time t of order i.
$q_{n,t}^s$	Grid selling quantity of the agent n at time t.
$r_{n,t}$	Reward of the agent n at time t.
W_t	Weather parameter (GHI) at time t.

APPENDIX B HYPERPARAMETERS

In Table 3, we list the hyperparameters for environment configuration, PPO algorithm, and model.

TABLE 3. Hyperparameters overview.

Hyperparameter	Value
Algorithm	PPO
Episode length	24
Training episodes	4,000
Batch size	2400
Training steps	96,000
Learning rate	0.0001
Number of workers	4
SGD minibatch size	128
Discount factor	0.99
KL coefficient	0.2
KL target	0.01
GAE Lambda	1
MLP layers	256, 256
Entropy reg. coefficient	0
Gradient clipping	10
Value function loss coefficient	1
Training agents	4
Testing agents	300
Energy storage max (kWh)	10
Energy storage min (kWh)	2
Energy storage at t=0 (kWh)	6
Max charge/discharge per period (kWh)	2
Charge/discharge efficiency	0.95
Grid selling price ($)	0.04
Grid buying price between 09:00-16:00 ($)	0.13
Grid buying price between 17:00-20:00 ($)	0.18
Grid buying price between 21:00-08:00 ($)	0.08
Sequential length	24
LSTM cell size	256
Attention head	1
Attention head dimension	32
Attention dimension	64
Attention transformer unit	1
Attention init GRU gate bias	2
Attention position-wise MLP dimension	32
Attention memory	48

REFERENCES

[1] E. Elahi, M. Zhu, Z. Khalid, and K. Wei, "An empirical analysis of carbon emission efficiency in food production across the Yangtze river basin: Towards sustainable agricultural development and carbon neutrality," *Agricult. Syst.*, vol. 218, Jun. 2024, Art. no. 103994.

[2] C. Isaboke, Y. Chen, and A. Bagonza, "Carbon disclosures and industry environment sensitivity on firm performance," *Green Low-Carbon Economy*, vol. 1, pp. 1–27, Aug. 2023.

[3] M. I. A. Shah, A. Wahid, E. Barrett, and K. Mason, "Multi-agent systems in peer-to-peer energy trading: A comprehensive survey," *Eng. Appl. Artif. Intell.*, vol. 132, Jun. 2024, Art. no. 107847.

[4] M. Mahmoud and S. B. Slama, "Peer-to-peer energy trading case study using an AI-powered community energy management system," *Appl. Sci.*, vol. 13, no. 13, p. 7838, Jul. 2023.

[5] L. Pu, S. Wang, X. Huang, X. Liu, Y. Shi, and H. Wang, "Peer-to-peer trading for energy-saving based on reinforcement learning," *Energies*, vol. 15, no. 24, p. 9633, Dec. 2022.

[6] J. Wang, L. Li, and J. Zhang, "Deep reinforcement learning for energy trading and load scheduling in residential peer-to-peer energy trading market," *Int. J. Electr. Power Energy Syst.*, vol. 147, May 2023, Art. no. 108885.

[7] M. Cao, Z. Yin, Y. Wang, L. Yu, P. Shi, and Z. Cai, "A reliable energy trading strategy in intelligent microgrids using deep reinforcement learning," *Comput. Electr. Eng.*, vol. 110, Sep. 2023, Art. no. 108796.

[8] A. Timilsina and S. Silvestri, "P2P energy trading through prospect theory, differential evolution, and reinforcement learning," *ACM Trans. Evol. Learn. Optim.*, vol. 3, no. 3, pp. 1–22, Sep. 2023.
[9] V. Franois-Lavet, D. Taralla, D. Ernst, and R. Fonteneau, "Deep reinforcement learning solutions for energy microgrids management," in *Proc. Eur. Workshop Reinforcement Learn.*, 2016, pp. 1–24.
[10] J. R. Vázquez-Canteli and Z. Nagy, "Reinforcement learning for demand response: A review of algorithms and modeling techniques," *Appl. Energy*, vol. 235, pp. 1072–1089, Feb. 2019.
[11] T. Yang, L. Zhao, W. Li, and A. Y. Zomaya, "Reinforcement learning in sustainable energy and electric systems: A survey," *Annu. Rev. Control*, vol. 49, pp. 145–163, Jun. 2020.
[12] D. J. B. Harrold, J. Cao, and Z. Fan, "Renewable energy integration and microgrid energy trading using multi-agent deep reinforcement learning," *Appl. Energy*, vol. 318, Jul. 2022, Art. no. 119151.
[13] A. K. Shakya, G. Pillai, and S. Chakrabarty, "Reinforcement learning algorithms: A brief survey," *Expert Syst. Appl.*, vol. 231, Nov. 2023, Art. no. 120495.
[14] R. Lowe, Y. I. Wu, A. Tamar, J. Harb, O. P. Abbeel, and I. Mordatch, "Multi-agent actor-critic for mixed cooperative-competitive environments," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 30, 2017, pp. 1–11.
[15] M. Janner, J. Fu, M. Zhang, and S. Levine, "When to trust your model: Model-based policy optimization," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 32, 2019, pp. 1–10.
[16] N. Lambert, A. Wilcox, H. Zhang, K. S. J. Pister, and R. Calandra, "Learning accurate long-term dynamics for model-based reinforcement learning," in *Proc. 60th IEEE Conf. Decis. Control (CDC)*, Dec. 2021, pp. 2880–2887.
[17] 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.
[18] D. Qiu, J. Wang, J. Wang, and G. Strbac, "Multi-agent reinforcement learning for automated peer-to-peer energy trading in double-side auction market," in *Proc. 13th Int. Joint Conf. Artif. Intell.*, Aug. 2021, pp. 2913–2920.
[19] D. Qiu, Y. Ye, D. Papadaskalopoulos, and G. Strbac, "Scalable coordinated management of peer-to-peer energy trading: A multi-cluster deep reinforcement learning approach," *Appl. Energy*, vol. 292, Jun. 2021, Art. no. 116940.
[20] R. May and P. Huang, "A multi-agent reinforcement learning approach for investigating and optimising peer-to-peer prosumer energy markets," *Appl. Energy*, vol. 334, Mar. 2023, Art. no. 120705.
[21] J. Schulman, F. Wolski, P. Dhariwal, A. Radford, and O. Klimov, "Proximal policy optimization algorithms," 2017, *arXiv:1707.06347*.
[22] T. Haarnoja, A. Zhou, K. Hartikainen, G. Tucker, S. Ha, J. Tan, V. Kumar, H. Zhu, A. Gupta, P. Abbeel, and S. Levine, "Soft actor-critic algorithms and applications," 2018, *arXiv:1812.05905*.
[23] S. Dankwa and W. Zheng, "Twin-delayed DDPG: A deep reinforcement learning technique to model a continuous movement of an intelligent robot agent," in *Proc. 3rd Int. Conf. Vis., Image Signal Process.*, Aug. 2019, pp. 1–5.
[24] J. W. Mock and S. S. Muknahallipatna, "A comparison of PPO, TD3 and SAC reinforcement algorithms for quadruped walking gait generation," *J. Intell. Learn. Syst. Appl.*, vol. 15, no. 1, pp. 36–56, 2023.
[25] E. Liang, R. Liaw, P. Moritz, R. Nishihara, R. Fox, K. Goldberg, J. E. Gonzalez, M. I. Jordan, and I. Stoica, "RLlib: Abstractions for distributed reinforcement learning," 2017, *arXiv:1712.09381*.
[26] N. Stiennon, L. Ouyang, J. Wu, D. Ziegler, R. Lowe, C. Voss, A. Radford, D. Amodei, and P. F. Christiano, "Learning to summarize with human feedback," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 33, 2020, pp. 3008–3021.
[27] L. Ouyang, J. Wu, X. Jiang, D. Almeida, C. L. Wainwright, P. Mishkin, C. Zhang, S. Agarwal, K. Slama, A. Ray, J. Schulman, J. Hilton, F. Kelton, L. Miller, M. Simens, A. Askell, P. Welinder, P. Christiano, J. Leike, and R. Lowe, "Training language models to follow instructions with human feedback," in *Proc. Adv. Neural Inf. Process. Syst. (NIPS)*, 2022, pp. 27730–27744.
[28] S. H. E. Data and N. Sydney. *Solar Home Electricity Data*. Accessed: Nov. 2, 2023. [Online]. Available: https://www.ausgrid.com.au/Industry/Our-Research/Data-to-share/Solarhome-electricity-data
[29] D. Friedman, *The Double Auction Market: Institutions, Theories, and Evidence*. Evanston, IL, USA: Routledge, 2018.
[30] M. Ahmed, R. Seraj, and S. M. S. Islam, "The k-means algorithm: A comprehensive survey and performance evaluation," *Electronics*, vol. 9, no. 8, p. 1295, Aug. 2020.
[31] C. Ying, W. Wang, J. Yu, Q. Li, D. Yu, and J. Liu, "Deep learning for renewable energy forecasting: A taxonomy, and systematic literature review," *J. Cleaner Prod.*, vol. 384, Jan. 2023, Art. no. 135414.
[32] Y. Yu, X. Si, C. Hu, and J. Zhang, "A review of recurrent neural networks: LSTM cells and network architectures," *Neural Comput.*, vol. 31, no. 7, pp. 1235–1270, Jul. 2019.
[33] D. Salinas, V. Flunkert, J. Gasthaus, and T. Januschowski, "DeepAR: Probabilistic forecasting with autoregressive recurrent networks," *Int. J. Forecasting*, vol. 36, no. 3, pp. 1181–1191, Jul. 2020.
[34] B. Lim, S. Ö. Arık, N. Loeff, and T. Pfister, "Temporal fusion transformers for interpretable multi-horizon time series forecasting," *Int. J. Forecasting*, vol. 37, no. 4, pp. 1748–1764, Oct. 2021.
[35] J. S. Stein, C. W. Hansen, and M. J. Reno, *Global Horizontal Irradiance Clear Sky Models: Implementation and Analysis*. Livermore, CA, USA: Sandia National Laboratories, 2012.
[36] M. Sengupta, A. Weekley, A. Habte, A. Lopez, and C. Molling. (2005). *Validation of the National Solar Radiation Database (NSRDB)(2005–2012)*. [Online]. Available: https://nsrdb.nrel.gov/data-viewer
[37] R. Tavenard, J. Faouzi, G. Vandewiele, F. Divo, G. Androz, C. Holtz, M. Payne, R. Yurchak, M. Rußwurm, K. Kolar, and E. Woods, "Tslearn, a machine learning toolkit for time-series data," *J. Machine Learn. Res.*, vol. 21, no. 1, pp. 4686–4691, 2020.
[38] E. Parisotto, F. Song, J. Rae, R. Pascanu, C. Gulcehre, S. Jayakumar, M. Jaderberg, R. L. Kaufman, A. Clark, S. Noury, M. Botvinick, N. Heess, and R. Hadsell, "Stabilizing transformers for reinforcement learning," in *Proc. Int. Conf. Mach. Learn.*, 2020, pp. 7487–7498.
[39] L. Liu, J. Liu, and J. Han, "Multi-head or single-head? An empirical comparison for transformer training," 2021, *arXiv:2106.09650*.

NAT UTHAYANSUTHI received the B.Sc. degree in metallurgical engineering from the Department of Metallurgical Engineering, Faculty of Engineering, Chulalongkorn University, Thailand, in 2016, where he is currently pursuing the M.Sc. degree in computer science with the Department of Computer Engineering, Faculty of Engineering. His main research interests include deep reinforcement learning, time-series forecasting, and energy trading.

PEERAPON VATEEKUL (Senior Member, IEEE) received the Ph.D. degree from the Department of Electrical and Computer Engineering, University of Miami (UM), Coral Gables, FL, USA, in 2012.

Currently, he is an Associate Professor with the Department of Computer Engineering, Faculty of Engineering, Chulalongkorn University, Thailand. His research interests include the domains of machine learning, data mining, deep learning, text mining, and big data analytics. To be more specific, his works include variants of classification (hierarchical multi-label classification), natural language processing, data quality management, and applied deep learning techniques in various domains, such as medical images and videos, satellite images, meteorological data, and text.