IEEE Access
Multidisciplinary ⋮ Rapid Review ⋮ Open Access Journal

Received 13 October 2022, accepted 14 November 2022, date of publication 24 November 2022, date of current version 12 December 2022.

Digital Object Identifier 10.1109/ACCESS.2022.3224460

RESEARCH ARTICLE

Model-Based Approach on Multi-Agent Deep Reinforcement Learning With Multiple Clusters for Peer-To-Peer Energy Trading

MANASSAKAN SANAYHA, (Member, IEEE), AND PEERAPON VATEEKUL
Department of Computer Engineering, Faculty of Engineering, Chulalongkorn University, Pathumwan, Bangkok 10330, Thailand

Corresponding author: Peerapon Vateekul (peerapon.v@chula.ac.th)

This work was supported by the Second Century Fund (C2F), Chulalongkorn University.

ABSTRACT Peer-to-peer (P2P) energy trading system has the ability to completely revolutionize the current household energy system by sharing energy among residents. As the number of customers employing distributed energy resources (DERs) such as solar rooftops increase, innovation in the double auction market (DA) system is becoming more significant. In this paper, a novel model-based, multi-agent asynchronous advantage actor-centralized-critic with communication (MB-A3C3) approach is carried out. Previous studies are limited since they suffer from unpredictable behavior in renewable energy resources and a large number of prosumers in the peer-to-peer market. As for the model-based strategy, we forecast the trading price and trading quantity in the daily energy trading system in order to overcome unpredictable issues. For the large number of prosumers, the multi-agent and multithreading RL has been chosen as our backbone since the prosumers' behavior can be diverse; time-series clustering is introduced based on their daily trading behavior. With its environmental model and multi-threaded mechanism, MB-A3C3 is seen to be most efficient in carrying out tasks regards time and precision. The model is conducted on a large scale real-world hourly 2012–2013 dataset of 300 households in Sydney having rooftop solar systems installed in New South Wales (NSW), Australia. Results reveal that the MB-A3C3 approach outperforms other reinforcement learning methods (MADDPG and A3C3), producing lower community energy bills for 300 households. When internal trade (trading among houses) increased and external trade (trading to the grid) decreased, our multiple agent RL (MB-A3C3) significantly lowered energy bills by 17%. In closing the gap between the real-world and theoretical problems, the algorithms herein aid in reducing customers' electricity bills.

INDEX TERMS Peer-to-peer energy trading, model-based reinforcement learning, multi-agent reinforcement learning, deep learning approach.

I. INTRODUCTION

The energy sector is constantly innovating. Recently, however, it is continually being disrupted by the "four Ds" of energy: decarbonisation, decentralisation, digitalisation, and democratisation. It is noted that multi-agent structures (MASs) can deal with grid disruptions caused by renewable energy sources, and the system's widely dispersed nature [1]. In the energy economy more effort is needed to establish a comprehensive system for the volatile structure of the market. Digitalization of the energy sector entails greater use of technology, data and advanced systems to better manage energy.

Over the last two decades, there has been a tremendous surge in the development and deployment of various types of machine learning (ML) models for energy systems [2], [3], [4]. Big data has provided a plethora of opportunities as well as problems for making well-informed decisions [5], [6], [7], [8], [9], [10], [11]. In the energy sector, predictive approaches based on ML models have gained in popularity because of their accuracy, effectiveness, and speed [12], [13], [14], [15], [16]. As for electricity price forecasting, an artificial neural network (ANN) has been combined with time-series

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

clustering algorithms [17]. Besides, various types of long short-term memory (LSTM) have been conducted [18], [19], [20], [21]. The application of ML models to traditional energy systems as well as alternative and renewable energy systems has been most beneficial [22], [23].

Peer-to-peer (P2P) energy trading involves a participant submitting bids to a trading system that requires a market operator to manage transactions based on the available data, the quantity required, and the price. Then, employing the double-sided auction approach, all orders are matched where traders can specify the quantity and price at which they want to trade within the boundaries of the price set directly from the grid [24], [25]. Based on previous studies in the P2P market, traders in the double auction (DA) market frequently use a zero intelligence (ZI) trading strategy [26], [27], [28], [29], [30]. The order price that ZI traders determine is the random surplus offset from the value of a particular range, e.g., FiT and ToU. Considering the strategies of all participants along with trading prices, energy supply, and energy consumption, the energy market is seen to be incredibly dynamic. Many previous works have attempted to address the DA market as an optimization problem, using the reinforcement learning (RL) framework [30], [31], [32], [33], [34], [35], [36], [37].

Deep reinforcement learning (DRL) is a subfield of machine learning that combines RL with deep learning. DRL is a fully automated approach that uses a range of inputs from current energy markets to determine maximum profits for intraday market bidding [38], [39], single sided energy markets [40], and power trading competition [41]. In the P2P market, traditional Q-Learning has been applied as a management algorithm to maximize profits through participation in P2P energy trading [42], [43]. The deep Q learning (DQN) algorithm based on the LSTM model has also been used to analyze time-dependent information. The deep deterministic policy gradient (DDPG) has also been put forward to probe strategic bidding in the energy market [44], [45], [46].

To obtain optimum learning for multi-agent decision-making in dynamic and uncertain environments, multi-agent reinforcement learning (MARL) algorithms for collaborative Markov decision processes (MDPs) have been introduced and examined [47]. In energy trading, MARL is capable of optimizing and reducing costs [31], [32], [33]. Thus, multi-agent deep deterministic policy gradients (MADDPG) are enhanced, improving peer-to-peer energy trading in the double auction market [30], [34], [35], [36], [37]. A3C3, which outperforms MADDPG, has been introduced as a distributed asynchronous actor-critic algorithm in a multi-agent setting with differential communication and a centralized critic [48].

Recently, MBRL, a model-based reinforcement learning method has demonstrated promising results in a variety of domains, resulting in a superior bidding strategy. In conjunction with MBRL, Dyna-architecture has been used to improve interaction in a modeled environment through learning and planning based on real-world and simulated experiences [49]. As for the multi-joint dynamics with contact (MuJoCo) benchmark, advanced MBRL algorithms have been able to optimize the reward function [50], [51], [52]. In the energy sector, it is significant that MBRL has been applied in wind energy bidding for a single-agent system, achieving minimized energy costs [53]. As for energy trading tasks, both MBRL and A3C3 have not yet been applied to P2P, but such algorithms may successfully outperform standard benchmarks [54], [55].

In this paper, a novel multi-agent deep reinforcement learning algorithm called "the model-based asynchronous advantage actor-centralized-critic with communication (MB-A3C3)" has been introduced. MB-A3C3 aims to investigate P2P energy trading in solar-installed households, as a multi-agent decision-making model for both competitive and cooperative tasks. Contributions are summarized, as follows:

- MADRL: multi-agent deep reinforcement learning. This technique can effectively handle complex data and many agents because it utilizes a deep learning architecture and a multithreaded framework with communication channels. To reduce training time, it can also be scaled horizontally.
- Agent's daily trading behavior clustering: According to previous research, little consideration has been shown towards prosumers' behavioral traits. This problem is addressed by classifying prosumers into clusters based on their daily trading habits.
- Model-based framework: The model-based concept has been integrated with MADRL viz. "MB-MADRL". As such, MB-MADRL can tackle the problem of a lack of local knowledge by allowing agents to build a functional representation of their environment. MBRL was established on Dyna architecture with multivariate-LSTM to anticipate the whole environmental states for 24 hours ahead, allowing for better policy execution than the model-free reinforcement learning (MFRL). Having a robust forecasting technique, MB-A3C3 represents each cluster as a centralized environmental model. This information enables MB-A3C3 to optimize processes in an accurate manner.

This study sets out to forecast the trading price and trading quantity in the daily energy trading system. Through application of RL algorithms, we are able to use the data to predict both trading price and trading quantity. Little research has been done in this field previously. Nevertheless, the clustering and forecasting methods used in model-based RL shows that our work is new and authentic. Facilitating local power and energy balance, we hope to transform the current household energy system by enabling households to have lower energy bills. Algorithms have been developed, one contribution at a time and applied to actual dataset, revealing their potential to optimize the distribution network.

II. P2P ENERGY TRADING

In the traditional market paradigm, producers and consumers deal with merchants depending on their net consumption. Peer-to-peer trading, however, necessitates the use of new technology and business models having market regulations that govern the P2P archetype [56]. Before trading with a retailer, producers share their production and consumption in local markets at an internal price that is typically set between export and retail prices. Consumers can be thought of as a subset of producers who do not own any local power operations. Producers and consumers confront a complicated quota-decision process because renewable sources of energy like solar photovoltaic (PV) generation are stochastic. Choosing a suitable trading strategy is challenging since all players' strategies are updated in real time.

A. THE DOUBLE AUCTION MARKET MECHANISM

The double auction (DA) market connects many customers and producers who are engaged in the energy market [57], [58]. In the electricity market, the auction term is set at a specific length of time, i.e., an hourly resolution [59]. Procedures are as follows:

1) Traders send their directives to the market whenever an auction period begins. Directives involve a trading price and energy quantity.
2) Purchase orders have to match sale orders. An algorithm is used to match the orders.
3) When two orders are matched, the auctioneer uses the classic mid-pricing approach to determine the market clearing price. The transaction quantity is equal to the minimum quantity between the matched orders.

At the end of the auction, the auctioneer balances the remaining amount of energy and unmatched orders with the utility company at grid pricing for time-of-use (ToU) and feed-in tariff (FiT) [58], [59]. All traders' pricing schemes are constrained by FiT and ToU to guarantee economic benefits. The prices for bids and asks are always within the grid prices. The buy–sell gap is at the center of the clearing price [60].

B. PROBLEM STATEMENT AND FORMULATION

The above-mentioned double auction market clearing procedures are a model for multi-agent decision-making, defined as a decentralized, partially observable Markov decision process (Dec-POMDP) with discrete time steps [61]. Each agent n selects an action $(a_{n,t})$ based on its policy and private observation $(o_{n,t})$ at time step t. N agents include a set of global states: S, a collection of private observations: O, a collection of action sets: A, a collection of reward functions: R, and a state transition function: T. One auction period (t = 1h) is the time span between two sequential stages. The trading of solar energy, in modified form, can be expressed, as in Eqs. (1), (2), and (3) [30]:

$$s_{n,t} = [P^{inf}_{n,t}, E^{es}_{n,t}, \lambda^{b}_{t}, \lambda^{s}_{t}, q^{da}_{actual_{n,t-1}}, q^{da}_{forecast_{n,t}}, \lambda^{i}_{actual_{n,t-1}}, \lambda^{i}_{forecast_{n,t}}] \tag{1}$$

where $s_{n,t}$ is the state of agent n at time step t. $P^{inf}_{n,t}$ and $E^{es}_{n,t}$ is the inflexible load information and energy storage (ES) battery energy content at time step t. λ^{b}_{t} and λ^{s}_{t} is the grid information for ToU and FiT at time step t. $q^{da}_{actual_{n,t-1}}$ and $\lambda^{i}_{actual_{n,t-1}}$ is the previous trading quantity and price at time step $(t-1)$. $q^{da}_{forecast_{n,1}}$ and $\lambda^{i}_{forecast_{n,t}}$ is the forecast trading quantity and price at time step (t).

$$a_{n,t} = \left(a^{q}_{n,t}, a^{p}_{n,t}\right) \tag{2}$$

where $a_{n,t}$ is the action of agent n at time step t. Both $a^{q}_{n,t}$ and $a^{p}_{n,t}$ represent the energy and price decision submitted to the DA market at time step t.

$$r_{n,t} = -(\lambda_{n,t} q^{da}_{n,t} \Delta t + \lambda^{b}_{t}[q^{grid}_{n,t}]^{+} \Delta t + \lambda^{s}_{t}[q^{grid}_{n,t}]^{-} \Delta t) \tag{3}$$

where $r_{n,t}$ is the immediate reward that the agent n at time step t obtains when the action is executed according to $s_{n,t}$.

At step t, agent n receives its reward $r_{n,t}$ in the form of a negative cost of energy bill, as a result of the DA market clearing procedures. Thus, the agents who are successfully cleared will receive the local price $\lambda_{n,t}$ and the cleared quantity $q^{da}_{n,t}$. Next, each agent n can calculate its corresponding cost in the DA market; the remaining unmatched quantity $q^{grid}_{n,t}$ will be bought or sold through the utility company at ToU λ^{b}_{t} or FiT λ^{s}_{t}. The agents' quantity $q^{grid}_{n,t} = q^{da}_{n,t}$ will be immediately exchanged at ToU λ^{b}_{t} or FiT λ^{s}_{t} if they are unable to be cleared in the DA market.

III. RELATED WORKS

Traders in the double auction market utilize the ZI strategy as a fundamental and popular trading method whereby they can determine their order price as a random surplus offset from its value, based on uniform distribution from a relevant interval viz. ToU and FiT [58], [59]. Because the actual market is quite dynamic in real time, participants are confronted by a complicated process, involving quotation decisions. Choosing an appropriate trading plan in such a complex market situation is difficult.

MARL is a framework for investigating the sequential decision-making problems of agents (producers and consumers) [62], [63]. MARL can also be applied to smart grid applications, i.e., P2P energy trading in the DA market [54].

A. MULTI-AGENT DEEP DETERMINISTIC POLICY GRADIENT (MADDPG) [64]

MADDPG unites the multi-agent actor-critic (MAAC) method with the DDPG algorithm. The algorithm utilizes a multi-agent policy gradient that involves decentralized agents to develop a centralized critic based on all agents' observations and behaviors. Each agent has its own actor and critic network, similar to a single-agent actor-critic architecture. The actor network takes the current state of the agent and suggests an action. However, a critic component differs somewhat from a standard single-agent DDPG. Each agent's critic network can see information concerning all actions and observations of all other agents. A critic network has a better perspective of what is going on whereas an actor network only has access to an agent's observations. A critic network's output is based on an estimated reward having full observation input as well as full action input of all agents. An actor network's output is a suggested action for that particular agent. Only during training time is the critic network active. At execution time, this network will not be available.

B. ASYNCHRONOUS ADVANTAGE ACTOR CENTRALIZED-CRITIC WITH COMMUNICATION (A3C3) [48]

Extended from the asynchronous advantage actor critic (A3C), network updates are carried out by multiple workers using a distributed approach [48]. As depicted in Fig. 1a, A3C3 has distributed worker threads that use actor-critic methods to asynchronously optimize value, policy, and communication networks for agents. By generating periodic local copies of the networks, utilizing them to compute gradients, and applying the gradients on the global networks, multiple workers asynchronously update all networks for each agent. In Fig. 1b, the agent's architecture in A3C3's worker is composed of three networks: 1) a policy (or actor) network, which outputs an action, 2) a communication network, which outputs an outgoing message, and 3) a value (or critic) network, which outputs a value estimation.

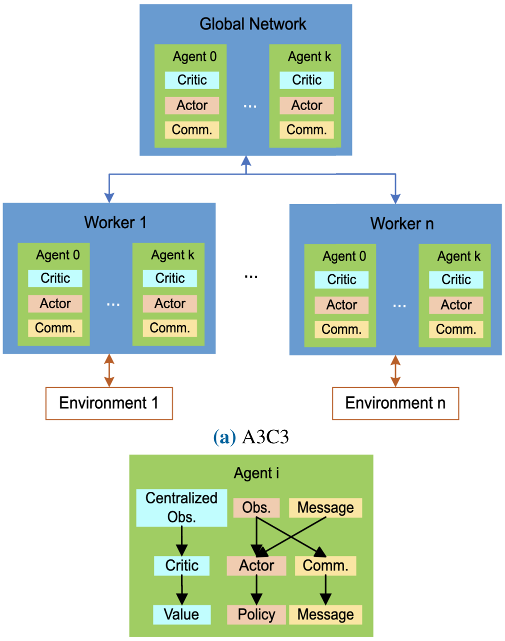
It is acknowledged that A3C3 can learn policies that are very successful and can attain goals in a shorter time than MADDPG [48]. Although A3C3 has never been applied in P2P energy trading, it has been adopted as our core model since A3C3 is seen to outperform MADDPG.

IV. PROPOSED METHOD

In this paper, the model-based deep reinforcement learning algorithm called MB-A3C3 is implemented. In Fig. 2, the schema of MB-A3C3, consisting of three modules, is demonstrated. In Module 1, A3C3 is employed to collect environmental data, including agents' information, actions, and energy bills for the trading period. In Module 2, agents, whether buyers or sellers, are classified according to their daily trading behavior and environmental data from Module 1. After that, agents' trading quantity and price are predicted via a forecasting module (Module 3), using clusters' centralized data obtained from Module 2. For the testing phase, the current state is then utilized to formulate the predicted trading quantity and price. Finally, the model assesses the current state and provides a policy that will result in action in the double auction market: amount of trading quantity and price. The energy bill is determined using all of the variables. Finally, MBRL is utilized to forecast future trading quantities and prices.

A. POLICY MODEL: A3C3-Conv1D WITH DA MECHANISM

The convolutional 1-dimensional A3C3 network was enhanced via application of A3C3, giving A3C3-Conv1D. To develop an optimal trading strategy for each agent, model parameters have been utilized and updated using experience and reward information. Such a strategy has been carried out as a policy model having a customized P2P energy trading environment. The assumptions listed below are attributed to A3C3-Conv1D:



(a) A3C3

(b) Agent

FIGURE 1. **A3C3 and agent's architecture.**

1) The concept of "asynchronous" means that when numerous threads collaborate on the same task and communicate what they have learned, a solution is achieved more efficiently.
2) "Multi-agent actor" and "centralized-critic": the "multi-agent actor" provides values based on their current policy for various acts. A "centralized critic" combines an agent's observations and environmental state information to provide an estimate of the current state and evaluates actions.
3) The term "advantage" describes how much better a certain action is relative to the predicted average value of the situation on which it is based.
4) "Communication" allows agents to share important information explicitly via a communication network based on the performance of other agents.

Furthermore, policies have constraints (a maximum bound) to make them more realistic:

1) A household's energy storage is determined by offering trading quantity with minimum and maximum energy levels between 2 and 10 kWh [65].
2) In Table 1, trading prices are provided. As highlighted in the grid, ToU is the flexible purchase price for the period; FiT is the set sale price for the entire day. The agent's trading price output, whether buying or selling, is limited to grid prices.
3) When trading in the double auction market, the network capacity threshold is considered peak demand. The algorithm maintains a daily peak demand of 600 kW, which satisfies the capacity of the network [34].

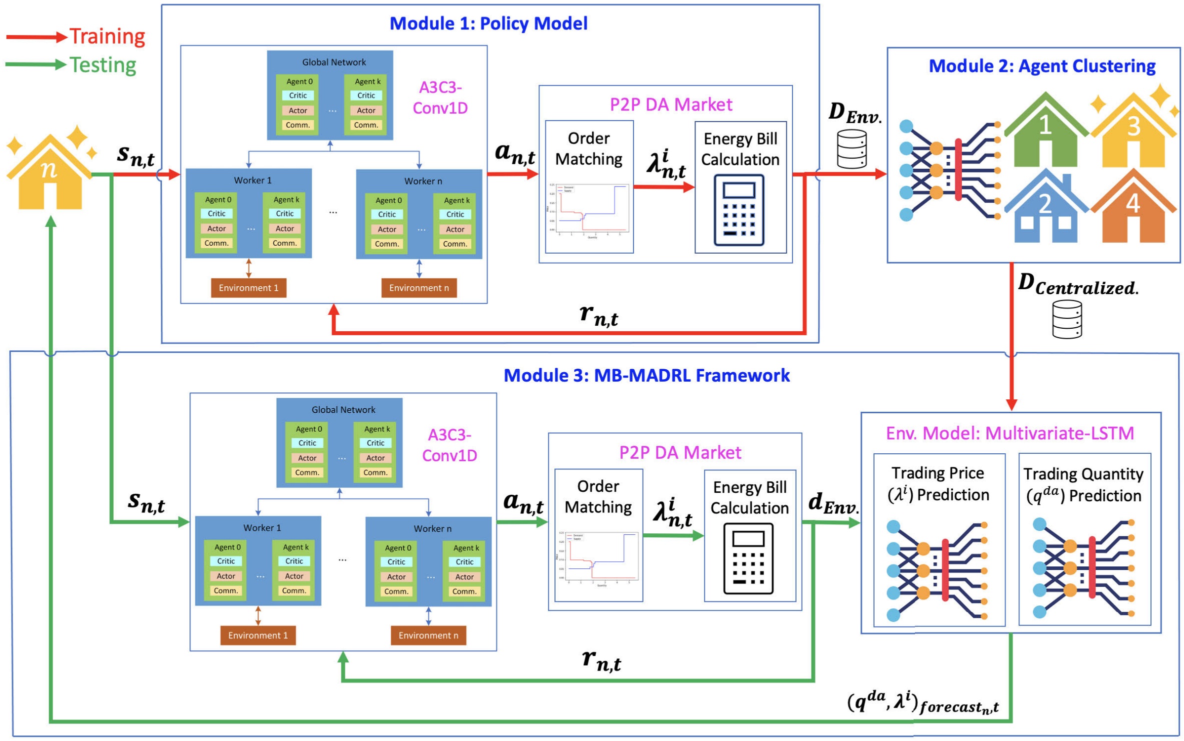


FIGURE 2. Schema of MB-A3C3 modules: (1) Policy model, which outputs agents' trading actions from local observation, (2) Agents' clustering, which categorizes agents into clusters based on their trading behavior from (1), and (3) MB-MADRL framework, which trains each cluster from (2) for trading action prediction.

TABLE 1. Grid pricing by period.

Time	ToU ($/kWh)		FiT (kWh)
	Time	Value	
Shoulder	09:00-16:00	0.13	0.04
Peak	17:00-20:00	0.18	
Off-Peak	21:00-08:00	0.08	

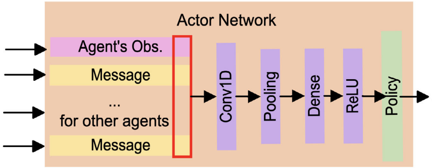


FIGURE 3. Agent's actor network.

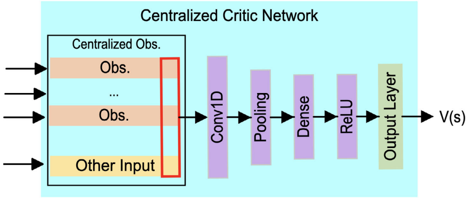


FIGURE 4. Agent's centralized critic network.

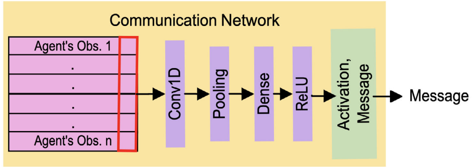


FIGURE 5. Agent's communication network.

A3C3's agent is represented by an actor, a central critic, and an additional communication network, as detailed below:

1) ACTOR NETWORK

As depicted in Fig. 3, local policy is learned by the actor network. For instance, the actor receives all agents' observations and broadcast messages as input. The output layer of the network generates a probability distribution for the agent's actions. The output layer is directly based on the action space of the environment.

2) CENTRALIZED CRITIC NETWORK

In Fig. 4, the agent's centralized network is given, combining all observations of other agents with some additional information from the environment. If the environment allows access to its underlying state, the centralized observations become the entire environmental state s_t. Thus, policy is evaluated by the centralized critic.

3) COMMUNICATION NETWORK

In Fig. 5, the communication network of the agent is depicted. The output layer has a rectifier or ReLU activation function to generate messages. Other output architectures such as continuous valued messages are supported. A communication protocol between agents is learned by the communicator network.

After receiving trading information from all agents, the mechanism of the double auction market matches orders and calculates energy bills, which are defined as a reward for each agent. Unmatched orders trade their energy at the price listed, as in Table 1.

B. AGENT'S DAILY TRADING BEHAVIOR CLUSTERING

For day to day trading, agents are grouped together using dynamic time warping (DTW). Then, each group is assigned a similar trading behavior as a centralized dataset for environmental modeling. DTW is utilized to measure the similarity between an agent's daily trading quantity. Because of its one-to-many determinations, the lowest distance between all points is calculated by DTW, allowing for a one-to-many match. DTW is a more precise way of determining distance than Euclidean distance; data points are moved between each other and focus on the shape rather than the geometry. Two time series do not have to be of identical length, which is a condition of Euclidean distance. Euclidean distance compares two data points with one another [66]. The optimal k for k-means is selected based on the elbow [67] and silhouette method [68].

Due to the large number of agents and their diverse behavior, it is assumed that an agent's daily behavior differs hour by hour. Accordingly, 300 agents are organized into four clusters based on their daily trading behavior. In the literature, DTW is frequently used in conjunction with k-medoids and hierarchical approaches; in some articles, DTW is used in conjunction with k-means [69]. DTW has also been coupled with random-swap and hybrid among non-traditional approaches [70].

C. MODEL-BASED MULTI-AGENT DEEP REINFORCEMENT LEARNING (MB-MADRL) FRAMEWORK

In Fig. 6, the multivariate-LSTM is depicted, and consists of six time-dependent variables:

$$\begin{aligned}(x_1, \ldots, x_6) = [&P_{n,t}^{inf}, E_{n,t}^{es}, \lambda_t^b, \lambda_t^s,\\ &q_{n,t-1}^{da}(or\lambda_{n,t-1}^i), r_{n,t-1}^i]\end{aligned} \tag{4}$$

Herein, the hidden output layer $(h_1, \ldots, h_6)$ is passed from one step of the network to the next. The algorithm LSTM takes into account not just the preceding hour of the input sequence, but also the prior 24h. Such a technique is used to calculate the state's predicted trading quantity $(q_{n,t}^{da})$ and price $(\lambda_{n,t}^i)$. Because of its ability to multiply the output of hidden states by trainable weights, the multivariate-LSTM is used to forecast an agent's trading quantity and price whereas the typical LSTM network simply utilizes the latest hidden state as output [71].

D. THE OVERALL PROCESS OF MB-A3C3

In Algorithm 1, the MB-A3C3 algorithm is demonstrated. The process begins with A3C3-Conv1D having to collect the environmental data. Then, ten random runs of the training process continue until energy bills from the DA market mechanism, calculated in accordance with Eq. (3) using actual data, stabilize. Then, DTW is applied for the time-series clustering to categorize the agents. For the next 24 h trading quantity and price forecasting, each cluster's data is collected to combine the environmental data with the multivariate-LSTM. After that, the MBRL process is applied for the testing phase.

Algorithm 1 **MB-A3C3 Algorithm**

```
1. Initial inputs $r_t, \eta_v, \eta_u, \eta_w, T_{max}, t_{max}, \gamma, \beta$, and output $\pi$
2. Initial environments: state, action, and reward
3. Assume global shared parameter vectors $(\theta_\mu, \theta_v, \theta_w)$, global shared counter T = 0, and thread-specific parameter vectors $(\theta'_\mu, \theta'_v, \theta'_w)$
4. Run A3C3 with $D_{train}$ to collect agent's trajectories $D_{env} = (s, a, r)$
5. Time-series clustering with DTW on $D_{env.}$
6. Aggregate each cluster's dataset to $D_{centralized}$
7. Train env. model on $D_{centralized}$ for each cluster.
8. Run A3C3 with env. model from 7. for $D_{test}$
for <episode = 1: $T_{max}$> do
    1. Reset gradients of actor, centralized critic, and communication network $d\theta_\mu \leftarrow 0, d\theta_v \leftarrow 0 \; d\theta_w \leftarrow 0$
    2. Reset $\theta'_\mu = \theta_\mu, \theta'_v = \theta_v, \theta'_w = \theta_w$
    for <agent = 1: N)> do
        for <t = 1: 24> do
            1. Get state $s_t$
            2. Perform $a_t$ according to policy $\pi\left(a_t \mid s_t; \theta'_\mu\right)$ and constraints.
            3. Every actor send its $(s_t, a_t, r_t)$ to env. model according to agent's cluster.
            4. Env. model predicts $a_t$ and send to A3C3 model.
            5. A3C3 action with $r_t = \text{argmax}\, r_t \left\{a_{1_t}, a_{2_t}, \ldots, a_{n_t}\right\}$, receive $r_t$, then transfer to new state
            if <$T/T_{max}$ == 0:> then
                6. $R = \begin{cases} 0, & \text{to terminal } s_t \\ V\left(s_t, \theta'{}_v\right), & \text{to not terminate } s_t \end{cases}$
                for <$i \in \{t, t-1, \ldots, t_{\max}\}$> do
                    7. $R \leftarrow r_i + \gamma R$
                    8. Accumulate gradients with $\theta'_v, \theta'_\mu$, and $\theta'_w$.
            9. Update the gradient of networks: $\theta_\mu$ by $d\theta_\mu$, $\theta_v$ by $d\theta_v$, and $\theta_w$ by $d\theta_w$
            10. Reset gradients of actor, centralized critic, and communication network $d\theta_\mu \leftarrow 0, d\theta_v \leftarrow 0$, and $d\theta_w \leftarrow 0$
            11. Reset $\theta'_\mu = \theta_\mu, \theta'_v = \theta_v$, and $\theta'_w = \theta_w$
```

where r_t is the reward function. η_v, η_u, and η_w are the actor's, centralized critic, and communication network's learning rates. T_{max} is the maximum training episode and t_{max} is the updated time-step. γ is the discount factor. β is the entropy regularization term. π is the policy. D_{train}, $D_{env.}$, $D_{centralized}$, and D_{test} are training, environment, centralized environment, and testing datasets, respectively.

V. EXPERIMENTAL SETUP

The experiment was carried out after assessing the data from Ausgrid's electricity network [72]. The publicly available dataset contained load and rooftop PV generation for 300 residential customers in NSW and the adjacent rural areas. Data was collected over a three-year period; both load and PV generation measurements were taken at 30 min intervals. The algorithm was conducted in a real-time simulation manner [30], [34], [35], [36], [37], [73], [74]. Although the dataset's period took place in 2012 and 2013, the data

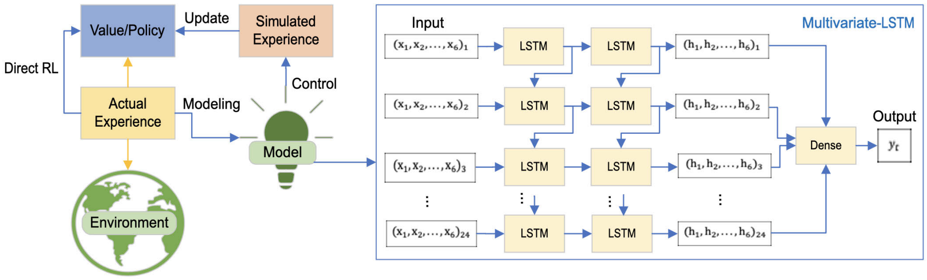

Value/Policy
Update
Simulated Experience
Direct RL
Control
Actual Experience
Modeling
Model
Environment
Input
Multivariate-LSTM
$(x_1, x_2, \dots, x_6)_1$
$(x_1, x_2, \dots, x_6)_2$
$(x_1, x_2, \dots, x_6)_3$
$(x_1, x_2, \dots, x_6)_{24}$
LSTM
$(h_1, h_2, \dots, h_6)_1$
$(h_1, h_2, \dots, h_6)_2$
$(h_1, h_2, \dots, h_6)_3$
$(h_1, h_2, \dots, h_6)_{24}$
Dense
Output
y_t

FIGURE 6. Schema of multivariate-LSTM having six features and a single output.

obtained from the 300 households was integrated and processed in real-time through the sharing economy.

A. EXPERIMENTAL DATA

Between July 1, 2010 and June 30, 2013, data was gathered from the 300 randomly selected solar customers in NSW. Customers had a gross metered solar system installed and were invoiced on a domestic tariff. From June 1, 2012 to May 31, 2013, data was utilized to evaluate performance with the baseline. Various types of data matching the annual statistics of solar home datasets are described [75].

B. HYPERPARAMETERS SETTING AND DETAILS

In Fig. 2, MB-A3C3 hyperparameters are specified [76]. In addition, both TensorFlow and OpenAI Gym were implemented having a Dyna framework in order to evaluate reinforcement learning algorithms; a customized environment using P2P energy trading data was provided.

C. EVALUATION

In Eq. (3), by employing the MB-A3C3 model, the reward function is utilized to reduce the agent's energy bill. It is noted that such action taken by the MB-A3C3 algorithm during training can determine the energy bill, which is the cumulative reward for each episode consisting of 24 steps. Accordingly, over 4,000 episodes, 10 independent runs with 10 random seeds were carried out for random initialization. During training, for every 100 episodes, following the baseline paper, the effectiveness of the households' energy management strategies regarding the test dataset was examined. MB-A3C3 was duly employed to evaluate how well it performed against the policy model under the three modules: 1) clustering: agent trading behavior, 2) forecasting: trading quantity and price, and 3) the MBRL framework.

VI. EXPERIMENTAL RESULTS

A. OVERALL RESULTS

In Table 2, the average community's internal trade, external trade, and net energy bills per day of 8 and 300 households are compared to MARL algorithms. For the multi-agent model, there is one agent per household for the experiment with 8 households (no clustering is applied). For the experiment with 300 households, there is only one agent per cluster. It is assumed that internal trade within communities should increase while external trade directly with the main grid should be reduced by the algorithm. The baseline MADDPG was extended (section III-A) from 8 to 300 households to ensure the validity of the algorithm [30]. Subsequently, of all the 12 algorithms, the MB-A3C3 (LSTM)-DTW algorithm was found to be the winner ($654.95). As a result, when compared with MADDPG ($789.85), household energy bills are seen to have fallen by more than $100. Energy bills turned out to be 17% lower than trading with the grid ($790.51). At the end of the trading day, the community's net energy bills were greatly reduced via the algorithm. Meanwhile, internal trade increased and external trade decreased while peak demand for energy dropped from above 600 to 589.26 kW.

In Figs. 7 (a and b), the training time of the multi-threaded algorithms (A3C3 and MB-A3C3) was compared with the single-threaded (MADDPG). It is acknowledged that despite consuming more training time, the MB-A3C3 (LSTM)-DTW algorithm assessed all relevant data via an agent's clustering and model of the environment. When the number of households increased from 8 to 300, training time reached 1,767.38 min (one model per each agent). However, when assigned to the model-based MB-A3C3, the outcome proved to be 149.36 min. Such an outcome is seen to reduce the time taken for forecasting.

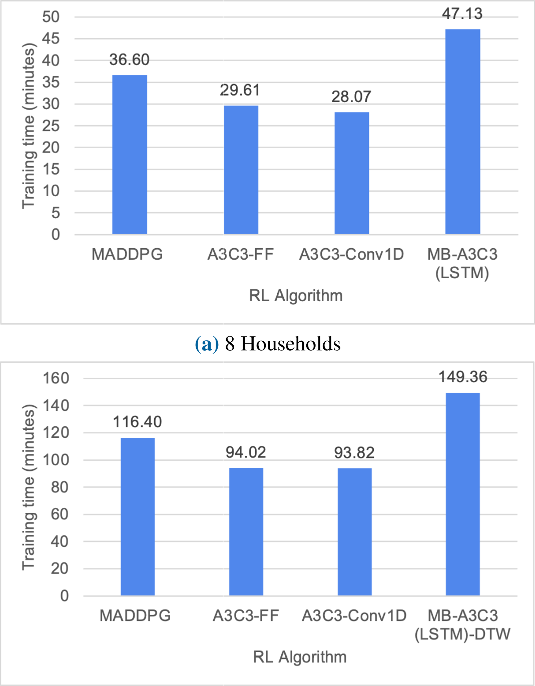
In Fig. 8, the community's average energy bill per day for the training set is presented. The reward of the five RL algorithms' convergence during the training phrase is depicted to illustrate the superior performance of MB-A3C3 (LSTM)-DTW over other algorithms; providing faster convergence and lower energy bills. When trading within a community, the algorithm is optimized under certain constraints and environments. An agent's energy bill is reduced by having a price incentive scheme in the algorithm. It is seen that the reward tends to be lower as agents have no knowledge or experience of how to trade during the first stage. After the training phase, the optimized network parameters, which result from multi-threaded mechanisms, deep learning networks, agents' clustering, and environmental models, efficiently lower the community's energy bills, as shown by the green line in the graph.

B. EFFECT OF MULTITHREADED AND DEEP LEARNING IN POLICY MODEL

In this section, it is seen that A3C3-FF can outperform the baseline MADDPG. Performance is further improved by

TABLE 2. The average community's internal trade, external trade, and net energy bills per day: 8 to 300 households are compared to MARL algorithms. Boldface refers to the winner method.

Algorithm	Internal (kWh) ↑		External (kWh) ↓		Net bills ($) ↓	
	8*	300	8*	300	8*	300
1: Deep Learning based A3C3						
A3C3-FF (baseline)	68.64	241.95	311.86	6,012.49	32.91	738.10
A3C3-Conv	66.32	264.17	310.58	5,956.16	31.18	732.61
A3C3-LSTM	65.58	242.05	313.95	6,114.99	34.59	742.79
2: Model based RL + 3: Agent clustering (one model per cluster)						
MB-A3C3 (LSTM)-Randomly	65.58	272.22	309.07	5,705.82	30.89	730.69
MB-A3C3 (LSTM)-Location-based	66.77	315.89	310.36	5,601.28	31.55	675.18
MB-A3C3 (LSTM)-DTW	**72.81**	**326.51**	**306.51**	**5,590.06**	**29.17**	**654.95**
MB-A3C3 (GRU)-Randomly	67.89	314.44	312.59	5,649.13	33.03	674.10
MB-A3C3 (GRU)-Location-based	66.37	317.43	311.15	5,654.46	32.21	672.61
MB-A3C3 (GRU)-DTW	69.37	321.77	310.71	5,597.87	32.54	673.26
MB-A3C3 (Transformer)-Randomly	69.22	315.68	316.66	5,697.35	33.10	738.86
MB-A3C3 (Transformer)-Location-based	69.41	314.78	317.11	5,738.70	32.20	719.60
MB-A3C3 (Transformer)-DTW	70.86	317.80	318.67	5,601.17	32.80	723.73



(a) 8 Households

(b) 300 Households

FIGURE 7. Training time (min) of each RL algorithms by number of households.

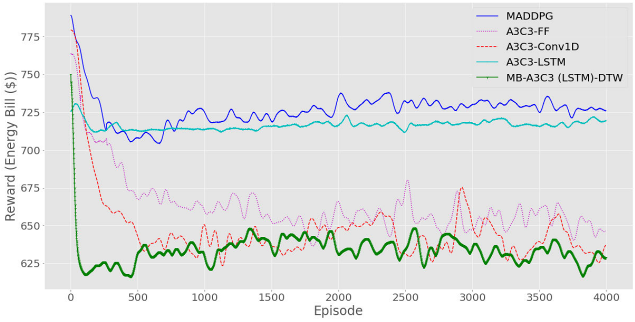


FIGURE 8. The community's energy bill per day over 4,000 episodes of training: faster convergence and lower bills are preferred.

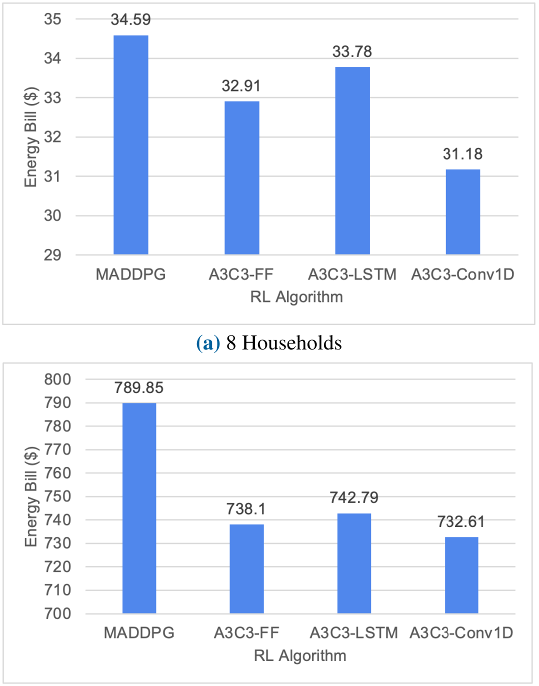
applying deep learning techniques, e.g., Conv1D rather than the feed-forward architecture: FF. In Table 2, it is noted that performance of the A3C3-Conv1D model is found to be superior to that of the single-threaded MADDPG, attaining a reduction in energy bills of 9.86% (from 34.59 to 31.18) and 7.25% (from 789.85 to 732.61) for both 8 and 300 households, respectively.

In Fig. 9a, by comparing results with the different network architectures, the A3C3-Conv1D algorithm outperformed A3C3-FF and A3C3-LSTM, revealing much lower energy bills in both 8 (Fig. 9) and extended 300 households (Fig. 9b). When a policy model considers the correlation between observations in a short timestep to take proper action, CNN is seen to perform better than LSTM because LSTM is usually applied for processing sequences of data. CNN is designed to exploit "spatial correlation" in data.

C. EFFECT OF AGENT'S TRADING BEHAVIOR TIME SERIES CLUSTERING

In this experiment, 300 agents (households) are clustered based on their trading behaviors. In our comparison, there are three strategies: (1) eight households: randomly selected, (2) location: based on postcode, and (3) time-series: k-means. In Fig. 10, the optimal number of k results are shown: k=4 was chosen.

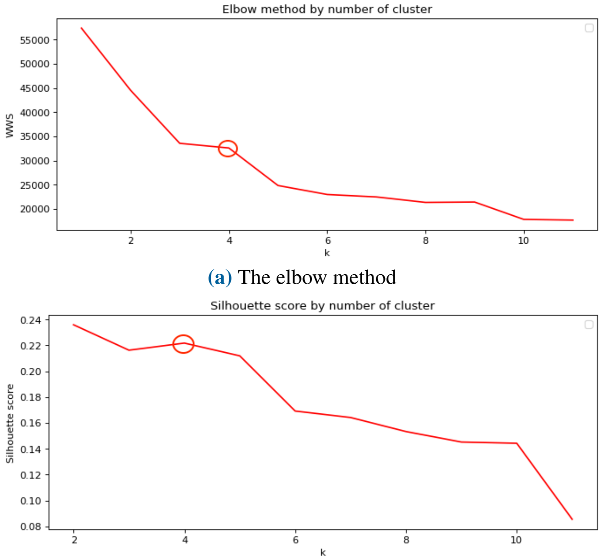
For the k-means method, two clustering algorithms were compared: namely, DTW and Euclidean. DTW was chosen as our clustering algorithm since its silhouette scores, i.e., clustering performance measures proved to be higher than the Euclidean scores: 0.23 and 0.17, respectively. In our paper, the time series k-means is also called "DTW". Due to DTW's calculations, each household is classified into a



(a) 8 Households

(b) 300 Households

FIGURE 9. Community's net energy bills of MARL algorithms by number of households.



(a) The elbow method

(b) The silhouette method

FIGURE 10. The number of clusters chosen from the elbow and silhouette method should therefore be 4.

cluster. Fig. 11 depicts the diverse trading quantities among four clusters, which vary in time.

In Fig. 12, a comparison is made of the three clustering methods applied to our winner from the previous experiment viz. MB-A3C3 (LSTM), as seen in Table 2. Results demonstrate that DTW proved to be the winner, revealing the cheapest energy bill ($654.95) for the 300 households.

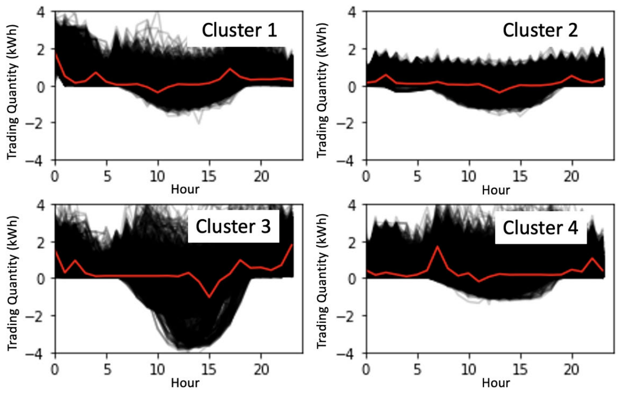


FIGURE 11. Four clusters of the 300 households. The red line shows the centroid of each cluster.

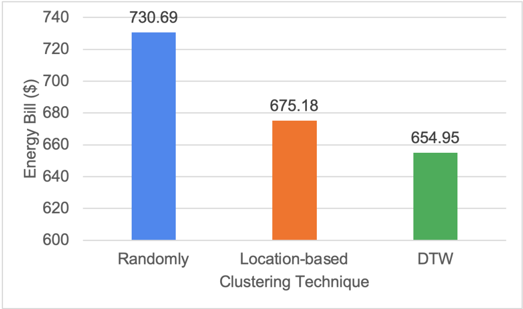


FIGURE 12. The community's energy bill for 300 households, applying clustering techniques for comparison.

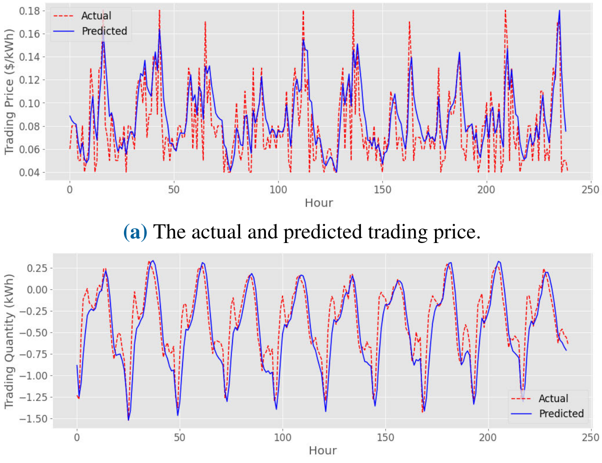
TABLE 3. Evaluation aspects: (1) predicted trading price and (2) predicted trading quantity, as determined via root mean squared error (RMSE) and mean absolute percentage error (MAPE). Boldface refers to the winner.

Method	Predicted price ($\lambda^{i}_{n,t}$)		Predicted quantity ($q^{da}_{n,t}$)	
	RMSE	MAPE	RMSE	MAPE
Mutivariate-LSTM	**0.0344**	**15.82%**	**0.0263**	**10.39%**
GRU	0.0582	17.17%	0.0379	12.87%
Transformer	0.0412	16.75%	0.0290	11.94%

D. EFFECT OF FORECASTING MODELS IN MB-MADRL FRAMEWORK

In Table 3, it is noted that the multivariate-LSTM excelled in terms of both RMSE and MAPE on the testing set over GRU and the transformer. The winner, the multivariate-LSTM, shows a marginal error of only 0.0344 dollars per kWh (15.82%) and 0.0263 kWh (10.39%) for the trading price and the trading quantity, respectively. Such an outcome reveals that the multivariate-LSTM algorithm proved to be the best since it provided less error than others in forecasting. In forecasting both trading prices and trading energy, our research has broken new ground.

In Fig. 13, the predicted trading price and quantity forecasting results, as determined by the winner (multivariate-LSTM) for one randomly selected household, is depicted. In Fig. 13a, it is seen that both predicted trading price and trading quantity differ quite dramatically due to fluctuation in householder's decisions. Rather than the trading quantity, which results from the agents' consumption-generation activity, the trading price volatility makes it more difficult



(a) The actual and predicted trading price.

(b) The actual and predicted trading quantity.

FIGURE 13. Results for (a) actual and predicted trading price and (b) quantity using the winner's forecasting model (multivariate-LSTM) for the first 240 timesteps of the testing dataset from an example household (randomly selected).

to capture patterns. In Fig. 13b, the forecasting result of the trading quantity is very promising since there is a pattern in energy usage (trading quantity), signifying its stable trend. According to the accurate forecast, the policy model can learn to act and minimize energy bills more efficiently. As shown in Table 2, MB-A3C3 (LSTM)-DTW outperformed other algorithms by providing higher internal trade, lower external trade, and reduced community energy bills for both 8 and 300 households.

VII. DISCUSSION

A. MBRL WITH FORECASTING MODEL

As investigated in Section IV, the MBRL framework begins by collecting environmental data and training the model to forecast. It is a requirement for MBRL that the forecasting model be accurate to ensure precise information for agents. The algorithm must be able to utilize the productive information to optimize the reward for the community's energy bill.

B. NUMBER OF K IN CLUSTERING METHOD

The clustering method was introduced to reduce the number of forecasting models (one model per cluster), assuming that homes in the same cluster behave similarly. Since it is quite costly to develop a forecasting model separately for each household (a total of 300 households), three clustering techniques were tested to determine the winner: random matching, location-based clustering, and k-means (DTW) clustering.

The results of clustering depend on the number of clusters (k); bias-variance trade-off determines the cluster number. If overfitting is taken into consideration, a large cluster will produce a tiny bias while a small number of clusters will produce a minor variation (sometimes favorable for generalization or interpretation) and is typically great for prediction. In Fig. 14, the community's energy bill and peak demand for 300 households diverge between k = 4 and 7; between k = 8 and 10, a tight race begins. It is projected that if k is increased to 300, the result will remain the same while requiring a significant amount of computational resources. It is significant that the winner of the selected number of clusters (k = 4) exhibits the lowest energy bill and peak demand.

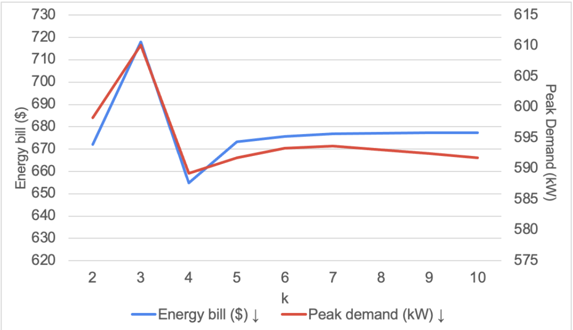


FIGURE 14. The inspection of energy bill and peak demand from k = 2 to 10 using the winner's clustering method (k-means (DTW)).

VIII. CONCLUSION

In this paper, a model-based multi-agent deep reinforcement learning algorithm called MB-A3C3 is presented. Firstly, the baseline A3C3 was enhanced by using the 1D convolutional network. Secondly, RL can support a large number of households (agents) by clustering those houses based on their trading behaviors using dynamic time warping (DTW). Thirdly, the environment was forecasted using multivariate LSTM; this is called model-based RL. Besides, both the multivariate-LSTM and CNN network are seen to improve multi-agent deep reinforcement learning. For large-scale households, the time-series clustering strategy based on trading behavior was utilized as an agent-based model. The experiment was conducted on the Ausgrid data set based on 300 households in NSW, Australia. Results demonstrate that our MB-A3C3, being less time-consuming and less complex, proved to be superior to other RL algorithms, producing costs 17% lower than traditional grid trading. It is significant that MB-A3C3 leveraged internal trading between households, thereby decreasing external trading under the grid's price incentives and constraints. Herein, the algorithms are seen to potentially aid in reducing customers' electricity bills. Further research must investigate various regulations to embrace more real-world scenarios of electricity consumers, producers, and power system operators to create more opportunities for P2P energy trading. Moreover, by adding other related factors, e.g., weather and system information, we can further improve the approach to make it more accurate. Training agents with more factors can provide more optimized policies.

DECLARATION OF COMPETING INTEREST

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this article.

DATA AVAILABILITY

Datasets related to this article can be found at an open-source online data repository hosted at Data to share (see: https://www.ausgrid.com.au/Industry/Our-Research/Data-to-share/Solar-home-electricity-data).

REFERENCES

[1] A. R. Khare and B. Y. Kumar, "Multiagent structures in hybrid renewable power system: A review," *J. Renew. Sustain. Energy*, vol. 7, no. 6, Nov. 2015, Art. no. 063101.

[2] M. Torabi, S. Hashemi, M. R. Saybani, S. Shamshirband, and A. Mosavi, "A hybrid clustering and classification technique for forecasting short-term energy consumption," *Environ. Prog. Sustain. Energy*, vol. 38, no. 1, pp. 66–76, Jan. 2019.

[3] B. Najafi, S. Faizollahzadeh Ardabili, A. Mosavi, S. Shamshirband, and T. Rabczuk, "An intelligent artificial neural network-response surface methodology for accessing the optimum biodiesel and diesel fuel blending conditions in a diesel engine from the viewpoint of exergy and energy analysis," *Energies*, vol. 11, no. 4, p. 860, Apr. 2018.

[4] M. Hosseini Imani, S. Zalzar, A. Mosavi, and S. Shamshirband, "Strategic behavior of retailers for risk reduction and profit increment via distributed generators and demand response programs," *Energies*, vol. 11, no. 6, p. 1602, Jun. 2018.

[5] A. Dineva, A. Mosavi, S. Faizollahzadeh Ardabili, I. Vajda, S. Shamshirband, T. Rabczuk, and K.-W. Chau, "Review of soft computing models in design and control of rotating electrical machines," *Energies*, vol. 12, no. 6, p. 1049, Mar. 2019.

[6] L. W. Chong, Y. W. Wong, R. K. Rajkumar, R. K. Rajkumar, and D. Isa, "Hybrid energy storage systems and control strategies for stand-alone renewable energy power systems," *Renew. Sustain. Energy Rev.*, vol. 66, pp. 174–189, Dec. 2016.

[7] N. Curry and P. Pillay, "Biogas prediction and design of a food waste to energy system for the urban environment," *Renew. Energy*, vol. 41, pp. 200–209, May 2012.

[8] K. Amarasinghe, D. L. Marino, and M. Manic, "Deep neural networks for energy load forecasting," in *Proc. IEEE 26th Int. Symp. Ind. Electron. (ISIE)*, Jun. 2017, pp. 1483–1488.

[9] V. H. Quej, J. Almorox, J. A. Arnaldo, and L. Saito, "ANFIS, SVM and ANN soft-computing techniques to estimate daily global solar radiation in a warm sub-humid environment," *J. Atmos. Solar-Terrestrial Phys.*, vol. 155, pp. 62–70, Mar. 2017.

[10] M. A. Mat Daut, M. Y. Hassan, H. Abdullah, H. A. Rahman, M. P. Abdullah, and F. Hussin, "Building electrical energy consumption forecasting analysis using conventional and artificial intelligence methods: A review," *Renew. Sustain. Energy Rev.*, vol. 70, pp. 1108–1118, Apr. 2017.

[11] B. Yildiz, J. I. Bilbao, and A. B. Sproul, "A review and analysis of regression and machine learning models on commercial building electricity load forecasting," *Renew. Sustain. Energy Rev.*, vol. 73, pp. 1104–1122, Jun. 2017.

[12] S. A. Kalogirou, "Applications of artificial neural-networks for energy systems," *Appl. Energy*, vol. 67. nos. 1–2, pp. 17–35, Sep. 2000.

[13] S. Faizollahzadeh Ardabili, B. Najafi, M. Alizamir, A. Mosavi, S. Shamshirband, and T. Rabczuk, "Using SVM-RSM and ELM-RSM approaches for optimizing the production process of methyl and ethyl esters," *Energies*, vol. 11, no. 11, p. 2889, Oct. 2018.

[14] O. Y. Al-Jarrah, P. D. Yoo, S. Muhaidat, G. K. Karagiannidis, and K. Taha, "Efficient machine learning for big data: A review," *Big Data Res.*, vol. 2, no. 3, pp. 87–93, 2015.

[15] K. Amasyali and N. M. El-Gohary, "A review of data-driven building energy consumption prediction studies," *Renew. Sustain. Energy Rev.*, vol. 81, pp. 1192–1205, Jan. 2018.

[16] Y. Peng, A. Rysanek, Z. Nagy, and A. Schlüter, "Using machine learning techniques for occupancy-prediction-based cooling control in office buildings," *Appl. Energy*, vol. 211, pp. 1343–1358, Feb. 2018.

[17] I. P. Panapakidis and A. S. Dagoumas, "Day-ahead electricity price forecasting via the application of artificial neural network based models," *Appl. Energy*, vol. 172, pp. 132–151, Jun. 2016.

[18] J. K. Kolberg and K. Waage, "Artificial intelligence and Nord pool's intraday electricity market Elbas: A demonstration and pragmatic evaluation of employing deep learning for price prediction: Using extensive market data and spatio–temporal weather forecasts," M.S. thesis, Econ. Bus. Admin., Norwegian School Econ., Bergen, Norway, 2018. [Online]. Available: https://openaccess.nhh.no/nhh-xmlui/handle/11250/2560898

[19] K. Yan, H. Shen, L. Wang, H. Zhou, M. Xu, and Y. Mo, "Short-term solar irradiance forecasting based on a hybrid deep learning methodology," *Information*, vol. 11, no. 1, p. 32, Jan. 2020.

[20] Z. Zhao, C. Xia, L. Chi, X. Chang, W. Li, T. Yang, and A. Y. Zomaya, "Short-term load forecasting based on the transformer model," *Information*, vol. 12, no. 12, p. 516, Dec. 2021.

[21] J. Zhang, H. Zhang, S. Ding, and X. Zhang, "Power consumption predicting and anomaly detection based on transformer and K-means," *Frontiers Energy Res.*, vol. 9, Oct. 2021, Art. no. 779587.

[22] C. Voyant, G. Notton, S. Kalogirou, M.-L. Nivet, C. Paoli, F. Motte, and A. Fouilloy, "Machine learning methods for solar radiation forecasting: A review," *Renew. Energy*, vol. 105, pp. 569–582, May 2017.

[23] A. Mosavi, A. Lopez, and A. Varkonyi-Koczy, "Industrial applications of big data: State of the art survey," in *Proc. Int. Conf. Global Res. Educ.*, Sep. 2017, pp. 225–232.

[24] E. A. Soto, L. B. Bosman, E. Wollega, and W. D. Leon-Salas, "Peer-to-peer energy trading: A review of the literature," *Appl. Energy*, vol. 283, Feb. 2021, Art. no. 116268.

[25] H. Javed, M. Irfan, M. Shehzad, H. Abdul Muqeet, J. Akhter, V. Dagar, and J. M. Guerrero, "Recent trends, challenges, and future aspects of P2P energy trading platforms in electrical-based networks considering blockchain technology: A roadmap toward environmental sustainability," *Frontiers Energy Res.*, vol. 10, Mar. 2022, Art. no. 810395.

[26] H. Muhsen, A. Allahham, A. Al-Halhouli, M. Al-Mahmodi, A. Alkhraibat, and M. Hamdan, "Business model of peer-to-peer energy trading: A review of literature," *Sustainability*, vol. 14, no. 3, p. 1616, Jan. 2022.

[27] J. Guerrero, A. Chapman, and G. Verbic, "Trading arrangements and cost allocation in P2P energy markets on low-voltage networks," in *Proc. IEEE Power Energy Society General Meeting (PESGM)*, Aug. 2019, pp. 1–5.

[28] X. Yan, J. Lin, Z. Hu, and Y. Song, "P2P trading strategies in an industrial park distribution network market under regulated electricity tariff," in *Proc. IEEE Conf. Energy Internet Energy Syst. Integr. (EI)*, Nov. 2017, pp. 1–5.

[29] K. Chen, J. Lin, and Y. Song, "Trading strategy optimization for a prosumer in continuous double auction-based peer-to-peer market: A prediction-integration model," *Appl. Energy*, vol. 242, pp. 1121–1133, May 2019.

[30] 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. IJCAI*, Aug. 2021, pp. 2913–2920.

[31] G. Gao, Y. Wen, and D. Tao, "Distributed energy trading and scheduling among microgrids via multiagent reinforcement learning," *IEEE Trans. Neural Netw. Learn. Syst.*, early access, May 12, 2022, doi: 10.1109/TNNLS.2022.3170070.

[32] A. Ghasemi, A. Shojaeighadikolaei, K. Jones, M. Hashemi, A. Bardas, and R. Ahmadi, "A multi-agent deep reinforcement learning approach for a distributed energy marketplace in smart grids," in *Proc. IEEE Int. Conf. Commun., Control, Comput. Technol. Smart Grids (SmartGridComm)*, pp. 1–6, Nov. 2020.

[33] A. Liu and Z. Zhao, "Multi-agent learning in repeated double-side auctions for peer-to-peer energy trading," in *Proc. Annu. Hawaii Int. Conf. Syst. Sci.*, Jan. 2021, pp. 1–10.

[34] 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.

[35] C. Samende, J. Cao, and Z. Fan, "Multi-agent deep deterministic policy gradient algorithm for peer-to-peer energy trading considering distribution network constraints," *Appl. Energy*, vol. 317, Jul. 2022, Art. no. 119123.

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

[37] J. Wang, L. Li, and J. Zhang, "Deep reinforcement learning for energy trading and load scheduling in residential peer-to-peer energy trading market," *SSRN Electron. J.*, 2022. [Online]. Available: https://papers.ssrn.com/sol3/papers.cfm?abstract_id=4102931, doi: 10.2139/ssrn.4102931.

[38] I. Boukas, D. Ernst, T. Théate, A. Bolland, A. Huynen, M. Buchwald, C. Wynants, and B. Cornélusse, "A deep reinforcement learning framework for continuous intraday market bidding," *Mach. Learn.*, vol. 110, no. 9, pp. 2335–2387, Sep. 2021.

[39] F. Verdaasdonk, S. Demir, and N. G. Paterakis, "Intra-day bidding strategies for storage devices using deep reinforcement learning," in *Proc. Int. Conf. Smart Energy Syst. Technol. (SEST)*, Lódz, Sep. 2022, pp. 1–9.

[40] E. Subramanian, Y. Bichpuriya, A. Achar, S. Bhat, A. Singh, V. Sarangan, and A. Natarajan, ''Learn: A reinforcement learning based bidding strategy for generators in single sided energy markets,'' in *Proc. ACM Int. Conf. Future Energy Syst.*, Jun. 2019, pp. 121–127.

[41] S. Ghosh, E. Subramanian, S. Bhat, S. Gujar, and P. Paruchuri, ''Vidyutvanika: A reinforcement learning based broker agent for a power trading competition,'' in *Proc. AAAI Conf. Artif. Intell.*, vol. 33, pp. 914–921, Jul. 2019.

[42] H. Zang and J. Kim, ''Reinforcement learning based peer-to-peer energy trade management using community energy storage in local energy market,'' *Energies*, vol. 14, no. 14, p. 4131, Jul. 2021.

[43] J.-G. Kim and B. Lee, ''Automatic P2P energy trading model based on reinforcement learning using long short-term delayed reward,'' *Energies*, vol. 13, no. 20, p. 5359, Oct. 2020.

[44] Y. Ye, D. Qiu, M. Sun, D. Papadaskalopoulos, and G. Strbac, ''Deep reinforcement learning for strategic bidding in electricity markets,'' *IEEE Trans. Smart Grid*, vol. 11, no. 2, pp. 1343–1355, Aug. 2020.

[45] H. Xu, H. Sun, D. Nikovski, S. Kitamura, K. Mori, and H. Hashimoto, ''Deep reinforcement learning for joint bidding and pricing of load serving entity,'' *IEEE Trans. Smart Grid*, vol. 10, no. 6, pp. 6366–6375, Nov. 2019.

[46] M. Christensen, C. Ernewein, and P. Pinson, ''Demand response through price-setting multi-agent reinforcement learning,'' in *Proc. 1st Int. Workshop Reinforcement Learn. Energy Manag. Buildings Cities*, Nov. 2020, pp. 1–5.

[47] S. Kar, J. Moura, and H. V. Poor, ''Distributed reinforcement learning in multi-agent networks,'' in *Proc. 5th IEEE Int. Workshop Comput. Adv. Multi-Sensor Adapt. Process. (CAMSAP)*, Dec. 2013, pp. 296–299.

[48] D. Simoes, N. Lau, and L. Paulo Reis, ''Multi-agent actor centralized-critic with communication,'' *Neurocomputing*, vol. 390, pp. 40–56, May 2020.

[49] R. S. Sutton, ''Dyna, an integrated architecture for learning, planning, and reacting,'' *ACM SIGART Bull.*, vol. 2, no. 4, pp. 160–163, Jul. 1991.

[50] V. Pong, S. Gu, M. Dalal, and S. Levine, ''Temporal difference models: Model-free deep RL for model-based control,'' 2018, *arXiv:1802.09081*.

[51] H. Xu, Y. Li, Y. Tian, T. Darrell, and T. Ma, ''Algorithmic framework for model-based reinforcement learning with theoretical guarantees,'' 2018, *arXiv:1807.03858*.

[52] 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.*, Jun. 2019, pp. 1–12.

[53] M. Sanayha and P. Vateekul, ''Model-based deep reinforcement learning for wind energy bidding,'' *Int. J. Electr. Power Energy Syst.*, vol. 136, Mar. 2022, Art. no. 107625.

[54] D. Zhang, X. Han, and C. Deng, ''Review on the research and practice of deep learning and reinforcement learning in smart grids,'' *CSEE J. Power Energy Syst.*, vol. 4, no. 3, pp. 362–370, Sep. 2018.

[55] K. Chandrasekaran, P. Kandasamy, and S. Ramanathan, ''Deep learning and reinforcement learning approach on microgrid,'' *Int. Trans. Electr. Energy Syst.*, vol. 30, no. 10, Oct. 2020, Art. no. e12531.

[56] M. R. Alam, M. St-Hilaire, and T. Kunz, ''Peer-to-peer energy trading among smart Homes,'' *Appl. Energy*, vol. 238, pp. 1434–1443, Mar. 2019.

[57] N. Z. Aitzhan and D. Svetinovic, ''Security and privacy in decentralized energy trading through multi-signatures, blockchain and anonymous messaging streams,'' *IEEE Trans. Depend. Sec. Comput.*, vol. 15, no. 5, pp. 840–852, Sep. 2018.

[58] J. Guerrero, D. Gebbran, S. Mhanna, A. Chapman, and G. Verbic, ''Towards a transactive energy system for integration of distributed energy resources: Home energy management, distributed optimal power flow, and peer-to-peer energy trading,'' *Renew. Sustain. Energy Rev.*, vol. 132, p. 27, Jun. 2020.

[59] D. Friedman, *The Double Auction Market: Institutions, Theories, and Evidence*, 1st ed. Routledge, 1993, doi: 10.4324/9780429492532.

[60] A. M. Alabdullatif, E. H. Gerding, and A. Perez-Diaz, ''Market design and trading strategies for community energy markets with storage and renewable supply,'' *Energies*, vol. 13, no. 4, p. 972, Feb. 2020.

[61] Y. Shoham and K. Leyton-Brown, *Multiagent Systems: Algorithmic, Game-Theoretic, and Logical Foundations*, 1st ed. Cambridge Univ. Press, Jan. 2009.

[62] M. Wooldridge, *An Introduction to Multiagent Systems*, 2nd ed. Wiley, Mar. 2022.

[63] L. P. Kaelbling, M. L. Littman, and A. W. Moore, ''Reinforcement learning: A survey,'' *J. Artif. Intell. Res.*, vol. 4, no. 1, pp. 237–285, Jan. 1996.

[64] R. Lowe, Y. Wu, A. Tamar, J. Harb, P. Abbeel, and I. Mordatch, ''Multi-agent actor-critic for mixed cooperative-competitive environments,'' in *Proc. Adv. Neural Inf. Process. Syst.*, Jun. 2017, pp. 1–12.

[65] D. Papadaskalopoulos and G. Strbac, ''Nonlinear and randomized pricing for distributed management of flexible loads,'' *IEEE Trans. Smart Grid*, vol. 7, no. 2, pp. 1137–1146, Mar. 2016.

[66] A. Javed, B. S. Lee, and D. M. Rizzo, ''A benchmark study on time series clustering,'' *Mach. Learn. Appl.*, vol. 1, Sep. 2020, Art. no. 100001.

[67] R. L. Thorndike, ''Who belongs in the family?'' *Psychometrika*, vol. 18, pp. 267–276, Dec. 1953.

[68] P. J. Rousseeuw, ''Silhouettes: A graphical aid to the interpretation and validation of cluster analysis,'' *J. Comput. Appl. Math.*, vol. 20, no. 1, pp. 53–65, 1987.

[69] R. Alcaraz, F. Hornero, and J. Rieta, ''Dynamic time warping applied to estimate atrial fibrillation temporal organization from the surface electrocardiogram,'' *Med. Eng. Phys.*, vol. 35, 2013, doi: 10.1016/j.medengphy.2013.03.004.

[70] S. Aghabozorgi, A. Seyed Shirkhorshidi, and T. Ying Wah, ''Time-series clustering—A decade review,'' *Inf. Syst.*, vol. 53, pp. 16–38, Oct. 2015.

[71] S. Hochreiter and J. J. Schmidhuber, ''Long short-term memory,'' *Neural Comput.*, vol. 9, no. 8, pp. 1735–1780, 1997.

[72] E. L. Ratnam, S. R. Weller, C. M. Kellett, and A. T. Murray, ''Residential load and rooftop PV generation: An Australian distribution network dataset,'' *Int. J. Sustain. Energy*, vol. 36, no. 8, pp. 787–806, Sep. 2017.

[73] X. Wang et al., ''Rolling horizon optimization for real-time operation of prosumers with peer-to-peer energy trading,'' *Energy Rep.*, vol. 9, pp. 321–328, 2023.

[74] J. Wang, J. Zhang, L. Li, and Y. Lin, ''Peer-to-peer energy trading for residential prosumers with photovoltaic and battery storage systems,'' *IEEE Syst. J.*, early access, Jul. 26, 2022, doi: 10.1109/JSYST.2022.3190976.

[75] *Data Description and Statistic*. Accessed: Apr. 20, 2022. [Online]. Available: https://drive.google.com/file/d/1qc-R0jYng7axay3j447uu5XFgktVYHzH/view?usp=sharing

[76] *Our Hyperparameter Tuning*. Accessed: Apr. 23, 2022. [Online]. Available: https://drive.google.com/file/d/1yMvKZSQf1EmwVUy-1Ns46lEp_Vi9UByP/view?usp=sharing

MANASSAKAN SANAYHA (Member, IEEE) was born in Bangkok, Thailand, in 1989. She received the B.E. degree in computer engineering from the King Mongkut's University of Technology North Bangkok, Bangkok, in 2012, and the M.Sc. degree in computer science from Chulalongkorn University, Bangkok, in 2017, where she is currently pursuing the Ph.D. degree with the Department of Computer Engineering. Her main research interests include machine learning, reinforcement learning, and advanced approaches in power utilities and the energy sector, mainly in energy forecasting and trading.

PEERAPON VATEEKUL 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 machine learning, data mining, deep learning, text mining, big data analytics, hierarchical multi-label classification, natural language processing, data quality management, and applied deep learning and reinforcement learning techniques in various domains, such as, healthcare, geoinformatics, hydrometeorology, and energy trading.