IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Received 9 November 2022, accepted 4 December 2022, date of publication 12 December 2022, date of current version 21 December 2022.

Digital Object Identifier 10.1109/ACCESS.2022.3228836

RESEARCH ARTICLE

Combined DR Pricing and Voltage Control Using Reinforcement Learning Based Multi-Agents and Load Forecasting

DANYAL AFGAN KHAN[1], AMMAR ARSHAD[1], MATTI LEHTONEN[2], AND KARAR MAHMOUD[2,3], (Senior Member, IEEE)

[1]Ghulam Ishaq Khan Institute of Engineering Sciences and Technology, Topi 23460, Pakistan
[2]Department of Electrical Engineering and Automation, School of Electrical Engineering, Aalto University, 02150 Espoo, Finland
[3]Department of Electrical Engineering, Faculty of Engineering, Aswan University, Aswan 81542, Egypt

Corresponding authors: Karar Mahmoud (karar.mostafa@aalto.fi) and Ammar Arshad (ammar.arshad@giki.edu.pk)

ABSTRACT The demand for energy around the world continues to increase at a very high rate. To sufficiently supply this high demand, it is imperative to employ efficient methods so that the total costs for fulfilling such high demand in energy are minimized. To achieve this ambitious goal, this paper proposes a multi-agent reinforcement learning system for time of use pricing based combined demand response and voltage control. For this purpose, a long short term memory network is employed for day-ahead load forecasting in order to remove future uncertainties. The Q-learning algorithm is used which is a model free algorithm and hence, doesn't require the agent(s) to have prior knowledge of the environment. The role of reinforcement learning in this work is very important since it allows the agent(s) to determine their respective optimal behavior(s) autonomously without explicit training by the end user. To allow effective cooperation among multiple agents, each household is controlled by its own agent, whereas all the household agents are directed by a master agent or service provider. Accordingly, the voltage control agent serves the purpose of checking voltage level violations in the system and removing them through optimal decision making. The proposed system yields very good results, whereby, not only is the overall cost of electricity reduced, but voltage level violations are also removed from the entire system. The implementation of this mechanism reduces the total average aggregated load demand from 5.23 kW to 3.86 kW, while reducing the total aggregated average cost from 94.01 Rs to 60.80 Rs, thanks to the proposed effective multi-agent based system.

INDEX TERMS Reinforcement learning, long short term memory, demand response, multi-agent system, voltage control.

I. INTRODUCTION

The continuous increase in demand for energy has put power systems around the globe under immense stress. The immediate solution to this problem, that comes to mind, is the expansion of power systems, but that in itself comes with a massive con; the huge cost associated with it [1]. An intelligent and viable method, thus, has to be employed to balance the demand and supply of energy without having to invest large amounts for achieving the given purpose.

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

Demand Response (DR) programs are frequently employed to solve the demand-supply imbalance without having to bare the heavy financial constraints, that would otherwise be applicable. DR programs are broadly categorized into two classes, namely, incentive based DR programs and price based DR programs. In incentive based programs, participants get payments for their agreement to curtail load consumption when demand is high. Incentive based schemes are categorized into three types: Direct Load Control (DLC), Interruptible/Curtailable (I/C) and Emergency DR programs. In DLC scheme, the participants get payments for curtailing load consumption under a set limit. This program allows

utilities to remotely power off customers' appliances. In I/C programs, the consumers are required to curtail load consumption in emergency scenarios. Consumers who fail to curtail their respective loads suffer penalties which are agreed upon at the time of the initiation of the scheme. Emergency DR programs are a mix of DLC and I/C programs and are thought of as market based programs [1]. In price based programs, customers are offered time varying electricity prices, which encourages them to shift their respective loads to low priced hours [2]. Price based schemes are divided into two classes, namely, real time pricing (RTP) and time of use (TOU) pricing. In TOU pricing, the electricity prices are high when demand is high (peak hours) while the prices are low when demand is low (off peak hours). The prices for both these sets of times remain constant and are predetermined. In RTP, on the other hand, the electricity prices vary frequently i.e. hourly or minutely and customers are offered price variation in as low as five minutes [3].

The domain of DR is very vast, hence, reasonable amount of study associated with it is available in literature. Reference [1] proposes a reinforcement learning (RL) based single agent system to shift controllable appliances from high demand hours to low demand hours, smoothing the load consumption profile and reducing electricity cost. A real time incentive based DR mechanism for smart grid systems with RL and deep neural network (DNN) is presented in [2]. DNN is used for load and price forecasting, while RL is used to achieve the optimal incentive rates. The author of [3] analyzes the starting of various DR schemes because of slumping technology costs and recognition of consumers' behavior in the electricity market. The author also sheds light on the problems associated with DR implementations across United States of America, China and developed cities of Europe. Reference [4] implements a pricing mechanism that combines long short-term memory (LSTM) models and RL to eradicate the pricing problem of service providers when the consumers' response behavior is not known. In [5], an incentive based DR program with deep learning and RL is proposed, whereas in [6], an hour ahead DR algorithm for home energy management system (HEMS) is implemented. It makes use of artificial neural network (ANN) to predict future prices and a multi-agent RL system for making optimal decisions for various home appliances.

The author of [7] proposes a framework for home energy management (HEM) based on RL for achieving efficient residential DR. In [8], a hybrid price based DR system is proposed which is adaptable to pricing principles, while in [9], a deep RL based DR algorithm for smart facilities energy management is proposed to minimize electricity costs. The author of [10] presents a self scheduling model for HEMS, in which a formulation of linear discomfort index (D1) is proposed, taking into account the preferences of customers in the daily operation of home appliances. An optimization model for residential DR, based on a deep deterministic policy gradient (DDPG) algorithm, is implemented in [11], whereas the author of [12] proposes an intelligent multi-microgrid (MMG) energy management method based on DNN and RL. Reference [13] proposes a dynamic pricing DR algorithm based on RL for energy management in a hierarchical electricity market, whereas the author of [14] proposes a real time DR mechanism for optimal home appliance scheduling using RL. Reference [15] establishes real time pricing models, taking into consideration price based DR measures, and formulates real time pricing sale scheme. Reference [16] proposes a comprehensive pricing based DR for a smart home with different household appliances, while the author of [17] estimates customer elasticity for incentive based DR programs making use of data from surveys on two countries and combined with a comprehensive residential load model. In [18], a real time price based DR scheme is incorporated into the allocated model of distribution generation (DG).

The author of [19] proposes a voltage management mechanism in unbalanced distribution networks through the implementation of residential DR and on load tap changes (OLTCs). Reference [20] proposes a multi-agent system to obtain flexible price based DR in low voltage distribution networks, while in [21] and [22], a data driven, model free and closed loop control agent, trained using deep RL for voltage control is proposed. The author of [23] proposes a two-time scale voltage regulation scheme for distribution grids. To cover the gap in the literature, this study offers a multi-agent reinforcement learning system for time of use pricing based on combined demand response and voltage control. In order to eliminate future uncertainties, a long short term memory network is used for day-ahead load forecasting. Reinforcement learning agents are used to optimize home appliance scheduling and voltage management. Each home is controlled by its own agent, and all household agents are commanded by a master agent or service provider to allow for successful cooperation among many agents. As a result, the voltage control agent checks for voltage level breaches in the system and eliminates them through optimum decision making. The suggested solution produces excellent results, lowering not just the total cost of power, but also removing voltage level violations from the whole system. Because of the suggested effective multi-agent based system, the deployment of this mechanism decreases the total average aggregated load demand from 5.23 kW to 3.86 kW while lowering the overall aggregated average cost from 94.01 Rs to 60.80 Rs. The main contributions of this paper can be summarized as follows:

- Proposing a multi-agent reinforcement learning system for time of use pricing based on combined demand response and voltage control.
- Precise load forecasting based on LSTM long short-term memory network.
- Proposing effective cooperation among multiple agents where each household is controlled by its own agent; in turn, all the household agents are directed by a master agent or service provider.
- Minimizing the overall cost of electricity, besides removing voltage level violations.

TABLE 1. Shiftable and non-shiftable appliances.

Index No.	Shiftable Appliances		Non Shiftable Appliances	
	Appliance	Rated Power (W)	Appliance	Rated Power (W)
1	AC	1800	UPS	10
2	Electric Heater	1000	Refrigerator	171
3	Iron	1000	Water Pump	1100
4	Washing Machine	255	Water Dispenser	500

II. PROBLEM FORMULATION

This work proposes a multi-agent system for TOU pricing based DR and voltage control, taking into consideration multiple households with varying load consumption profiles, using RL and LSTM, aiming to reduce the overall aggregated cost of electricity for all the households and also to maintain voltage levels over the distribution network within the prescribed limits. In order to cope with future uncertainties, an LSTM network is used to predict the load consumption profile of each house for the next day. RL is then employed for the optimum scheduling of appliances, based on the priority list of each household, which not only reduces the overall cost of electricity but also makes sure that the comfort levels of the residents are not compromised. RL is advantageous in that it is model free. This means that an RL agent, which is the service provider (SP) in this case, does not require prior information about optimal appliance scheduling, instead, the SP discovers it from direct interaction with the customers or households (environment). The appliances are divided into two categories: Shiftable Appliances and Non-Shiftable Appliances. Shiftable Appliances are the type of appliances that can be rescheduled from their normal operating times if the SP requires load to be shifted. For each household, the appliances have different priority settings, which means that the SP has to make sure that each appliance is shifted, keeping in view the priority setting of each household. This is achieved through an RL agent. Non-Shiftable Appliances, on the other hand, are the class of appliances that have to be kept powered on till the need of the particular household from the appliance is satisfied. These appliances can thus, not be rescheduled or shifted to other times and have to be kept on till they satisfy the household's needs. The various shiftable and non-shiftable appliances, relevant to this study, with their respective power ratings, are listed in Table. 1.

The total energy consumed by non-shiftable and shiftable appliances is given by equation 1 and equation 2 respectively, whereas equation 3 represents the total energy consumed by non-shiftable and shiftable appliances combined.

$$E_t^{non} = \sum_{n=1}^{N} e_t^{n,non}.I_t^n \tag{1}$$

$$E_t^{shift} = \sum_{n=1}^{N} e_t^{n,shift}.I_t^n \tag{2}$$

$$E_t^{total} = E_t^{non} + E_t^{shift} \tag{3}$$

III. LSTM AND MULTI-AGENT RL BASED METHODOLOGY

The load consumption data of households is obtained from [24]. The data was collected from 42 households in Lahore, Pakistan, over one minute intervals, for a period of one year. This study considers load consumption data from 9 such households, owing to the fact that most of the households had near similar energy consumption patterns, it was imperative to carefully look through the consumption patterns of each entity and choose households with distinctly varying energy consumption patterns, in order to develop a more generalized mechanism. The following subsections present in detail, the LSTM and multi-agent RL based mechanism.

A. LOAD FORECASTING WITH LSTM

LSTM [25] is basically a recurrent neural network (RNN), which is fundamentally different from traditional feed forward neural networks [26]. RNNs are sequential models, which means that they have the ability to establish correlation between previous and current information. This property of RNNs is particularly useful for time series problems such as load forecasting, where previous sequences of load data are used to predict future value(s) for various households, all having diverse load consumption patterns.

The RNNs, however, have a major limitation of gradient vanishing [27], [28]. Gradient vanishing points towards the fact that the norm of the gradient for long-term components decrease very quickly to zero, restricting the capability of the model to learn long-term temporal correlation, whereas gradient exploding is the opposite scenario. To overcome this limitation, LSTM is frequently employed in forecasting problems. LSTM maintains an internal memory cell throughout its life cycle in order to establish temporal correlations, which makes it an improved version of the conventional RNNs.The basic representation of an LSTM network is depicted in Fig. 1. The compact forms of the equations for the forward pass of an LSTM cell with a forget gate are:

$$f_t = \sigma_g(W_f x_t + U_f h_{t-1} + b_f) \tag{4}$$

$$i_t = \sigma_g(W_i x_t + U_i h_{t-1} + b_i) \tag{5}$$

$$o_t = \sigma_g(W_o x_t + U_o h_{t-1} + b_o) \tag{6}$$

$$\tilde{c_t} = \sigma_c(W_c x_t + U_c h_{t-1} + b_c) \tag{7}$$

$$c_t = f_t \circ c_{t-1} + i_t \circ \tilde{c_t} \tag{8}$$

$$h_t = o_t \circ \sigma_h(c_t) \tag{9}$$

where 'f_t' represents the activation vector of an LSTM's forget gate while 'i_t' is the activation vector of the input gate

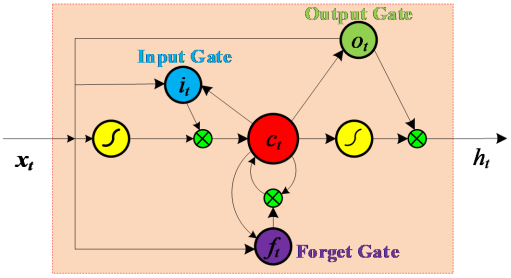


FIGURE 1. The basic representation of an LSTM network explaining how an LSTM network functions.

of an LSTM. Similarly, 'o_t' represents the activation vector of the LSTM's output gate whereas '$\tilde{c_t}$' is the activation vector of the input to an LSTM's cell. Moreover, 'c_t' is the state vector of an LSTM's cell while 'h_t' represents the output vector of an LSTM unit. The '$\circ$' sign in equation 8 and equation 9 represents multiplication.

The LSTM network implemented in this study consists of a sequential layer, hidden layer, LSTM layer and an output layer. The LSTM layer consists of 64 units while the hidden layer consists of 32 units. The LSTM network employed in this paper is depicted in Fig. 2.

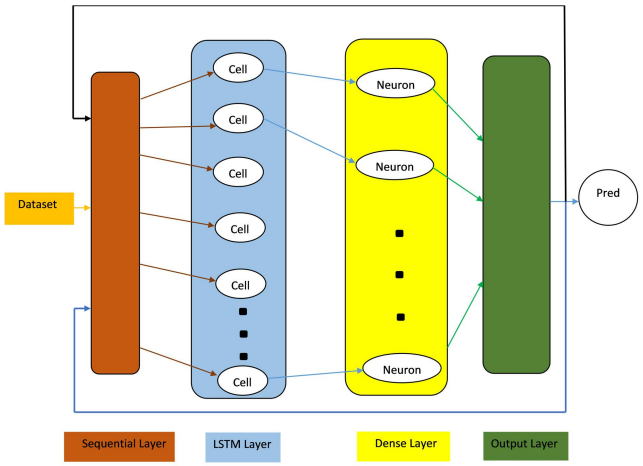


FIGURE 2. LSTM network architecture employed in this study, depicting the different types of layers and their corresponding number, employed in it.

B. MULTI-AGENT RL BASED DECISION MAKING

RL is a machine learning algorithm which enables an agent to autonomously work out the perfect behavior in a probabilistic environment, maximizing the cumulative reward as a result. RL algorithm has six parameters, namely, agent, environment, state space, action space, rewards and action-value. At each time step, the agent executes an action, receives the numerical reward for that action and transitions to the next state. The goal of the agent is to maximize the cumulative reward, hence, it has to learn a policy (optimal policy) that allows it to choose the optimal action at each state. A general RL framework is depicted in Fig. 3.

FIGURE 3. General RL framework.

In order to perform the optimal action at each state, the RL problem is modeled as a markov decision process (MDP) framework. The MDP displays the markov property, which states that the transitions in states depend only on the current state and current action performed, and do not depend on any prior environmental states or agent actions.

Q-learning [29], because of its ability to evaluate different actions for different states without needing to have a model of the environment, is used to get the optimal policy ν. The fundamental mechanism of Q-learning is to assign a Q-value i.e. $Q(s, a)$ to each state action pair and then updating this value at each time step for optimising the agent's performance. The optimal Q-value i.e. $Q^*(s, a)$ refers to the maximum discounted future reward $r(s, a)$ while performing action a at state s, and at the same time continuing to follow the optimal policy ν. It is represented by the equation below:

$$Q^*_{\nu}(s_t, a_t) = r(s_t, a_t) + \gamma max_T(a_{t+1}) \times [Q(s_{t+1}, a_{t+1})] \quad (10)$$

When an action is performed, centered on a particular policy ν, a pre-determined reward $r(s, a)$ will be received and the agent will take up a new state s_{t+1}. The action value $Q(s, a)$ is simultaneously updated according to the equation below:

$$Q(s_t, a_t) \leftarrow (1-\alpha)Q(s_t, a_t) + \alpha[r(s_t, a_t) + \gamma max_T(a_{t+1})[Q(s_{t+1}, a_{t+1})]] \quad (11)$$

where α denotes the learning rate which determines how much the old value of $Q(s_t, a_t)$ is affected by the new reward. For instance, $\alpha = 0$ shows that the latest information obtained is not employed in the learning process and thus, the reward obtained has no effect on the Q-value. If $\alpha = 1$, only the new information is taken into account. γ is referred to as the discount rate and depicts the relationship between the future and current rewards. It takes values between [0, 1]. When $\gamma = 0$, the agent takes into account only the current reward, while $\gamma = 1$ refers to the phenomenon where the agent will go for future rewards.

The application of RL in the proposed multi-agent setting enables the master agent (SP) to take the appropriate action

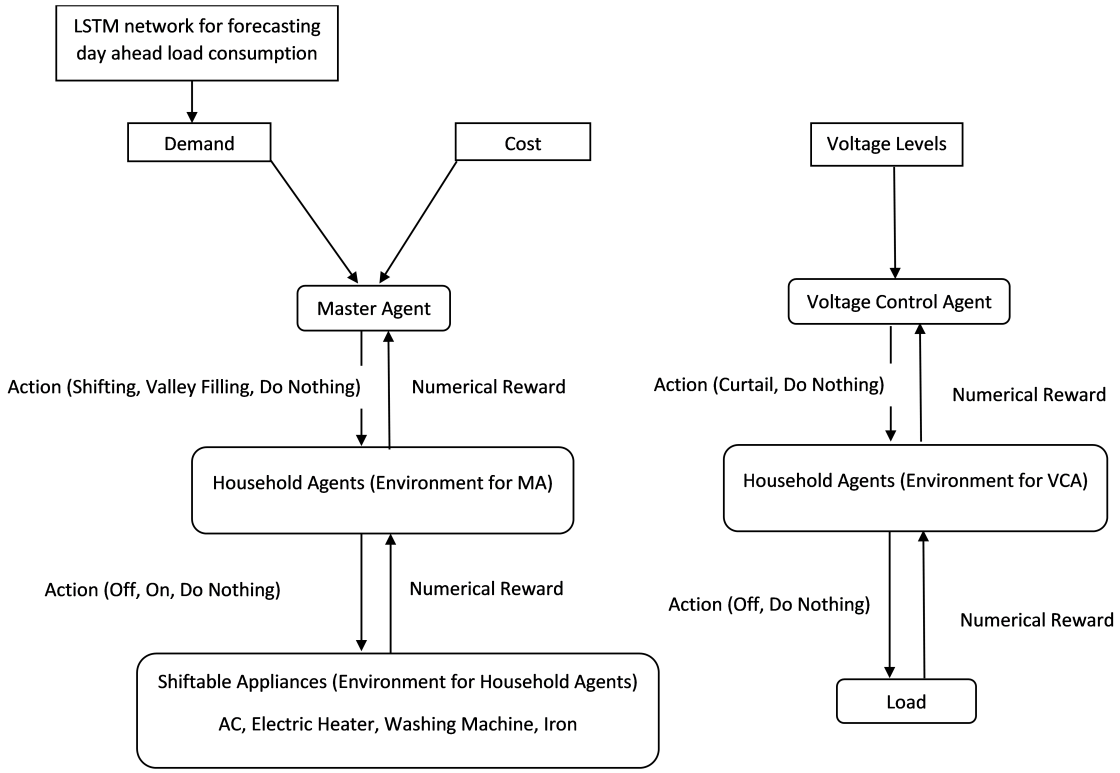


FIGURE 4. Proposed multi-agent system architecture.

(schedule appliances for each household) in each state (combination of aggregated load demand and electricity cost of nine households) and also allows the voltage control agent to monitor the voltage across the distribution network and take the appropriate action to maintain its levels within the prescribed limits. The SP is chosen as the master agent in the appliance scheduling setting because it is responsible for power dispatch and scheduling in a power system. Each agent in the system has its own set of states, actions and the corresponding Q-values and aims to obtain the optimal Q-value, $Q^*(s, a)$. This framework is explained in the following sub-section.

C. LSTM AND MULTI-AGENT RL FRAMEWORK FOR APPLIANCE SCHEDULING AND VOLTAGE CONTROL

Fig. 4 shows the overall framework of the proposed appliance scheduling and voltage control algorithm for TOU pricing based DR using multi-agent RL. An LSTM network forecasts the minutely load of each household for the next day and at each time instant, the master agent (SP) receives the aggregated load and electricity cost of all the households. Based on the combination of both, the master agent takes the appropriate action. The pair of demand and cost constitutes the states of the master agent given as follows:

$$s_t = [E^{total}_{t.index}, C^{total}_{t.index}] \tag{12}$$

The demand is categorized into three levels: high, average and low demand, while the cost is categorized into two types: high and low cost [1], as are given in the following equations.

TABLE 2. State indexes.

Demand	Cost	State Index
E_{low}	C_{low}	1
E_{low}	C_{high}	2
$E_{average}$	C_{low}	3
$E_{average}$	C_{high}	4
E_{high}	C_{low}	5
E_{high}	C_{high}	6

$$E^{total}_{t,index} = \begin{cases} E^{low}_{total}, & if E^{total}_t <= 4kW \\ E^{average}_{total}, & if\, 4kW < E^{total}_t <= 6kW \\ E^{high}_{total}, & if E^{total}_t > 6kW \end{cases} \tag{13}$$

$$C^{total}_{t,index} = \begin{cases} C^{low}_{total}, & if C^{total}_t <= 95Rs \\ C^{high}_{total}, & if C^{total}_t > 95Rs \end{cases} \tag{14}$$

The master agent thus has six possible states., the indexes of which are depicted in Table. 2.

There are three actions available to the master agent, i.e. shifting, valley filling and do nothing action given in equation 15.

$$A = [donothing, valleyfilling, shifting] \tag{15}$$

Fuzzy logic is employed as the reward function for determining the action of each agent in a certain state. Fuzzy logic deals with approximate values instead of exact values. For

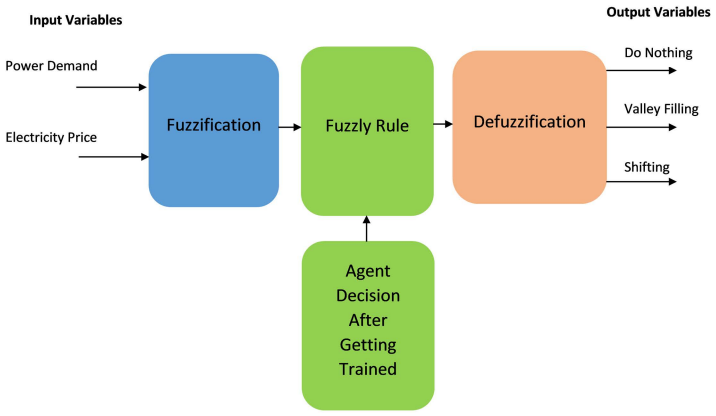


FIGURE 5. Reward function implementation using fuzzy logic.

the Master Agent, there are 3 states and 3 actions available. The actions are graded into *Excellent Action, Satisfactory Action* and *Bad Action*. The reward for *Excellent Action* is kept at $+1.75$, the reward for *Satisfactory Action* is kept at 0, while the reward for *Bad Action* is kept at -1.5. For the Household Agents, again, there are 3 states and 3 actions available. The three actions are *Shifting, Valley Filling* and *Do Nothing* actions. The actions of the Household Agents depend on the state of the Master Agent and Voltage Control Agent. The Household Agents, thus, have only one *Excellent Action* for a particular state, while the remaining action(s) remain Bad, since the value of their action is dependent on whether they follow the commands of the Master Agent or Voltage Control Agent. The reward for the *Excellent Action* is kept at $+2.5$, while that for the *Bad Actions* is kept at -2.25. Finally, for the Voltage Control Agent, there are 2 states and 2 actions available. The actions are simply termed as *Excellent Action* and *Bad Action*, where the reward for an *Excellent Action* is kept at $+1.5$, while the reward for a *Bad Action* is kept at -1.2. Fig. 5 depicts the overall mechanism behind the reward function implementation.

When at a particular time instant, both the aggregated demand and aggregated cost are high, the master agent directs the agents of each household to curtail load consumption (shift appliances based on the priority setting of each). On the contrary, when both the demand and cost are low, the master agent selects the valley filling action, directing each household agent to turn on the shifted appliances. For all the other states, the master agent directs the household agents to remain in their respective present states (do nothing action). Fig. 6 depicts the overall RL framework for the proposed appliance scheduling system.

The household agents constitute the environment of the master agent while the shiftable appliances constitute the environment for the household agents. At each time instant, the master agent takes an action depending on the state. The master agent's actions determine the behavior of the household agents and they then take the appropriate actions depending on the directions of the master agent. After

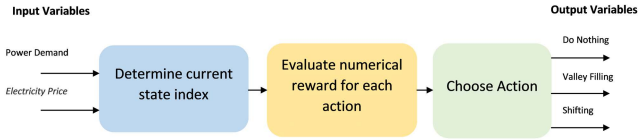


FIGURE 6. Proposed RL framework for appliance scheduling.

this, the voltage control agent monitors the voltage levels across the distribution network at each time instant and whenever the voltage level $<= 0.95$ p.u, it acts to raise the voltage. The functionality of the voltage control mechanism is given by the following equations for sensitivity analysis:

$$\begin{cases} [s_{IP}] = \dfrac{\partial V_i}{\partial I_{PJ}} = -[R] \\ [s_{IQ}] = \dfrac{\partial V_i}{\partial I_{QJ}} = -[X] \end{cases} \tag{16}$$

where S_{IP} and S_{IQ} are the sensitivity matrices with respect to the real and respective part of current, whereas R and X are the real and reactive part of impedances in the impedance matrix $[Z]$ [30]. Fig. 7 depicts the diagram of a 10 bus radial distribution system.

The voltage control agent monitors the voltage at each bus in the network at every time instant and whenever the voltage falls the below the prescribed threshold, it requests the corresponding household to curtail load. It, then, again checks the voltage at the bus and until the voltage level violation is removed, it requests subsequent households to curtail load. Fig. 8 shows the flow chart for the proposed voltage control algorithm.

All the agents in the multi-agent RL setting follow the Epsilon greedy policy to achieve a balance between exploration and exploitation. Exploration is the phenomenon where an agent strives to explore its environment more, sacrificing any immediate reward that might come in its path, for future rewards. Whereas in exploitation, the agent takes the best possible action at the current state, without worrying about future rewards. Following the epsilon greedy policy, the agent

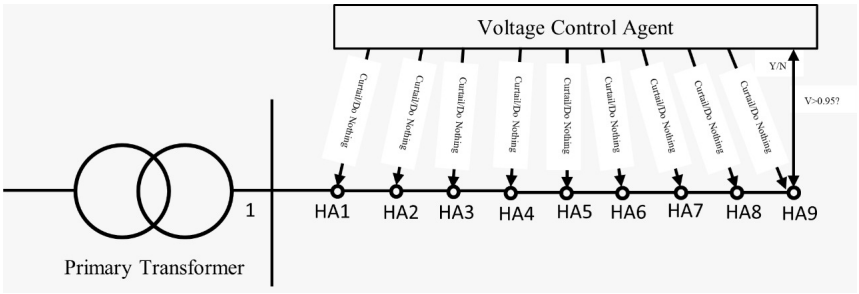


FIGURE 7. Diagram of a 10 bus radial distribution network.

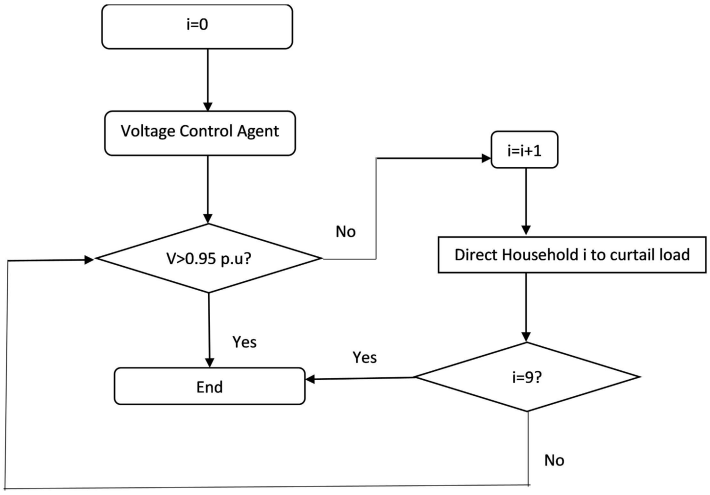


FIGURE 8. Flow chart for voltage control mechanism.

either selects a random action with probability ϵ, or selects a greedy action (best possible action at the current state with reference to the Q-table), with probability $1 - \epsilon$. The agent, as a result, explores its action space with an element of randomness, but does not become completely random. After executing an action at a given state, the agent receives a numerical reward $r(s_t, a_t)$ and transitions to the next state s_{t+1}. This procedure is repeated till the end of the day. Algo 1 explains the complete mechanism of the proposed system.

IV. RESULTS AND PERFORMANCE ANALYSIS

The performances of the LSTM model for load forecasting of all the households and multi-agent RL algorithm for TOU pricing based demand response (optimal appliance scheduling) and voltage control are presented in this section.

A. LOAD FORECASTING MODEL

An LSTM model was employed to predict the minutely variation in load consumption of all households for one day. The historical load consumption data was obtained from households based in Lahore, Pakistan. Since real time demand is not implementable in Pakistan, there was a need to forecast future load demand. The load consumption data set for a period of one year i.e. 1 June 2018 to 31 May 2019 was available, where 65% of the data set was used to train the LSTM model while the rest of the data set was used for testing the model. Load consumption of individual households was predicted by

Algorithm 1 Appliance Scheduling and Voltage Control With Multi-Agent RL System

```
Do day ahead load forecasting with LSTM
Set γ, ϵ and α parameters and define rewards r(s_t, a_t) for each agent
Initialize Q(s_t, a_t) to zero
for each iteration do
    for each time step do
        for each agent do
            Chose a random state s_t
            Select a random action a_t from all possible actions for the chosen state
            Execute the chosen action a_t,
            receive a numerical reward r(s_t, a_t)
            and observe the next state s_{t+1}
            Determine the maximum Q-value for the next state in the Q-matrix
            Update Q(s_t, a_t) using equation 2
            Set the next state as current state
        end for
    end for
end for
```

TABLE 3. LSTM performance on each household's data.

House No	MAE	MAPE
1	0.11	6%
2	0.06	11%
3	0.02	8%
4	0.02	5%
5	0.04	15%
6	0.03	5%
7	0.18	11%
8	0.01	3%
9	0.01	14%

TABLE 4. Experimental error in value function of master agent during training.

No. of Iterations	MAE
1000	0.9562
2000	0.5634
3000	0.5414
4000	0.4808
5000	0.3693

the LSTM model and the forecasted consumption for each household was summed and fed to the master agent as one its state parameters. Python's colab environment was employed to run the forecasting simulations and the process was a smooth one.

Fig. 9 and Fig. 10 depict the comparisons of actual load vs forecasted load with the LSTM model for household 4 and household 7, the households for which the LSTM gave the highest and lowest performance respectively. It can be seen that the LSTM model has accurately approximated the variations in load for both the households over time. The mean absolute error (MAE) and mean absolute percentage error (MAPE), represented by equation 17 and equation 18 respectively, are the performance metrics used to evaluate the LSTM model. Table. 3 compares the performance of the LSTM model on each household's load consumption data, while Table. 4 depicts the experimental error in the value function for the master agent over the course of training.

$$MAE = \frac{1}{n}\sum_{i=1}^{n}|y_i - y_i'| \tag{17}$$

$$MAPE = \frac{100}{n}\sum_{i=1}^{n}\frac{|y_i - y_i'|}{y_i} \tag{18}$$

B. TOU PRICING BASED DR ALGORITHM

A multi-agent system was employed for optimal scheduling of household appliances and voltage control at each bus of the distribution network. Each household was controlled by its own agent, separately trained, while the master agent (SP) controlled all the households agents. The household agents operated on the directions of the master agent to control the various shiftable household appliances, whereas the voltage control agent monitored the voltage at each bus and maintained it within a prescribed limit i.e V > 0.95 p.u.

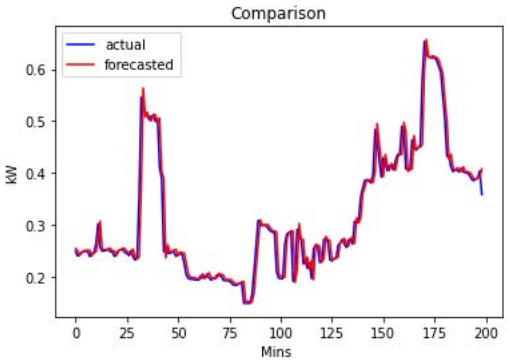


FIGURE 9. LSTM performance for house 4.

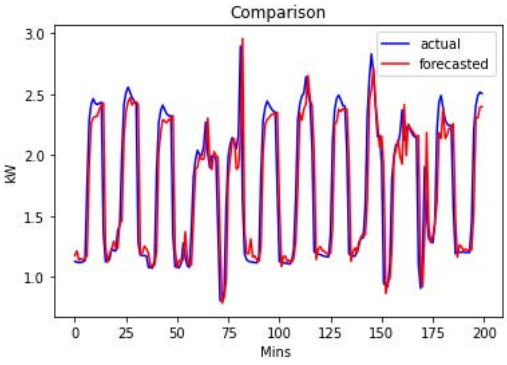


FIGURE 10. LSTM performance for house 7.

It is very natural for each household to have different priority setting for each appliance, thus, each household agent was trained to shift or turn on appliances in accordance with the priority setting of its corresponding household.

The hyperparameters of the Q-learning algorithm, i.e α, γ and ϵ, were all set to 0.9 to achieve a balance between the agent striving for future rewards and at the same time, giving importance to current rewards. The selection of these values for the hyperparameters also maintained a balance with regards to how much the latest reward received affected the Q-value.

Each agent was trained for a considerable amount of time, which allowed the Q-values of each agent to converge to their respective maximums. The agents were then enabled to choose the optimal actions for appliance scheduling and voltage control in a given state. The effectiveness of the overall multi-agent system for decreasing the aggregated load consumption can be seen in Fig. 11. It can be seen that the implementation of the DR algorithm reduced the overall load consumption significantly as compared to the scenario where DR was not employed. The average load consumption with DR was reduced to 3.86 kW from 5.23 kW without DR. There is a small window of time where valley filling was done i.e. the appliances shifted from peak hours were turned back on.

Fig. 12 depicts the total cost of electricity at each time instant without and with the multi-agent DR algorithm implementation. It can be seen that the cost of electricity was markedly reduced with the TOU pricing based DR algorithm

TABLE 5. Various scenarios.

Bus No.	Without DR					With DR and VC Applied on All Buses					With DR and VC Applied on 5 Buses				
	V_{max}	V_{min}	V_{dev}	$Load_{avg}$	$Cost_{avg}$	V_{max}	V_{min}	V_{dev}	$Load_{avg}$	$Cost_{avg}$	V_{max}	V_{min}	V_{dev}	$Load_{avg}$	$Cost_{avg}$
2	0.995	0.963	0.005	1.78	32.47	0.996	0.983	0.002	1.33	23.27	0.996	0.98	0.004	1.65	28.52
3	0.992	0.934	0.009	0.34	5.68	0.994	0.973	0.003	0.32	5.12	0.99	0.96	0.006	0.29	5.33
4	0.989	0.906	0.013	0.25	4.69	0.989	0.966	0.004	0.23	4.54	0.99	0.95	0.007	0.245	4.58
5	0.986	0.882	0.017	0.21	3.71	0.985	0.963	0.004	0.18	3.52	0.98	0.94	0.008	0.205	3.57
6	0.984	0.857	0.020	0.32	5.79	0.982	0.959	0.005	0.29	5.25	0.98	0.94	0.009	0.29	5.25
7	0.982	0.837	0.023	0.38	7.25	0.978	0.955	0.006	0.33	6.16	0.98	0.93	0.009	0.33	6.16
8	0.979	0.820	0.026	1.28	22.77	0.975	0.952	0.006	0.74	12.78	0.97	0.93	0.010	0.74	12.78
9	0.979	0.815	0.027	0.30	5.42	0.973	0.951	0.007	0.23	4.06	0.97	0.92	0.010	0.23	4.06
10	0.978	0.811	0.027	0.35	6.24	0.972	0.951	0.007	0.23	4.12	0.97	0.92	0.011	0.23	4.12

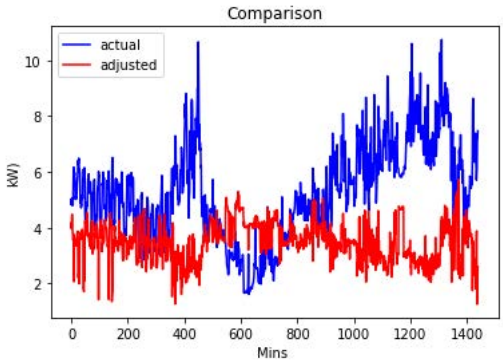


FIGURE 11. Comparison of load consumption without and with DR.

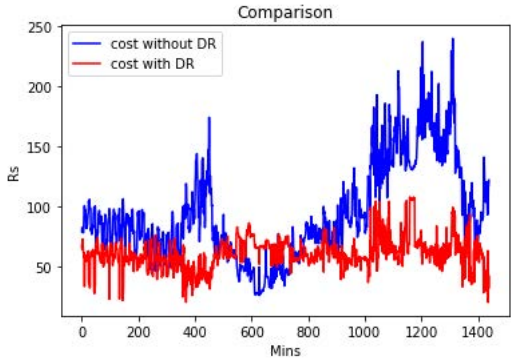


FIGURE 12. Comparison of electricity cost without and with DR.

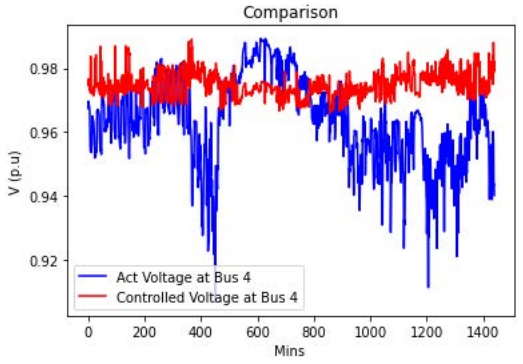


FIGURE 13. Comparison of voltage levels at bus 4.

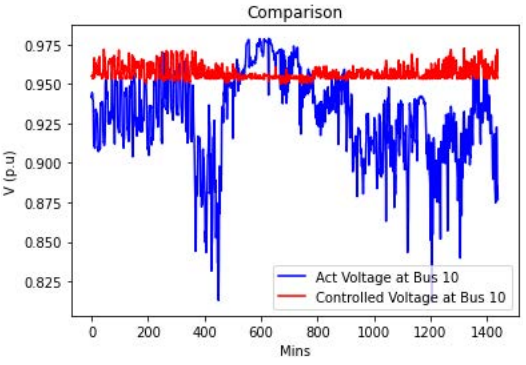


FIGURE 14. Comparison of voltage levels at bus 10.

as compared to the situation where the DR algorithm was not employed. The average cost without the implementation of the proposed DR strategy was 94.01 Rs as compared to 60.80 Rs with its implementation.

Fig. 13 and Fig. 14 depict the effect of the multi-agent system on the voltage levels at bus 4 and bus 10 respectively. It can be seen that the voltage levels are very poor without the implementation of the mechanism proposed in this study. The DR algorithm along with the voltage control agent, maintained the voltage levels at each bus, within the prescribed limits i.e. $V > 0.95$ p.u. This, thus, added to the effectiveness of the proposed system, where not only the total cost of electricity was reduced, but the voltage levels across the distribution network were also kept within an acceptable range.

Table. 5 shows the maximum voltage, minimum voltage, standard deviation in the voltage, average load and average cost at each bus in the distribution network for three different scenarios i.e. without DR, with DR and voltage control applied on all buses and with DR but voltage control applied on only five buses farthest from the source bus. It can be seen that the standard deviation in the voltage is markedly reduced after the implementation of the proposed algorithm as compared to the case without DR and voltage control. Moreover, employing voltage control only on the five farthest buses from the source bus also yields very good results, whereby the voltage level at each bus remains at acceptable levels i.e $V > 0.9$ p.u and hence the SP will have to pay less amount for load curtailment to the remaining four households, adding to its profitability.

Latency in communication forms a core part of the proposed voltage control strategy. Fig. 15 shows the number of

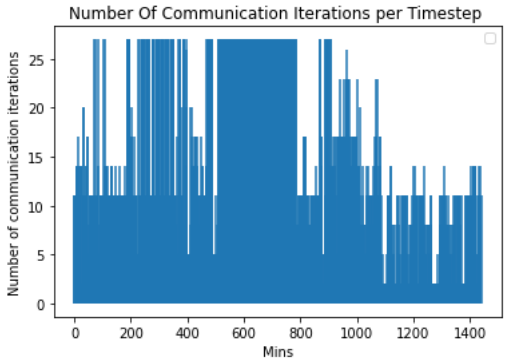


FIGURE 15. Number of communication iterations at each time step for voltage control.

communication iterations between the agents in order to keep the voltage levels within the prescribed limit. It takes 20 ms on average for machine-to-machine interaction, based on a research conducted on LTE network communication [31].

V. CONCLUSION

In this study, a multi-agent RL system was proposed for TOU pricing based DR and voltage control, the aim being to reduce the total cost of electricity and remove voltage level violations from the system. An LSTM network was employed for day ahead load forecasting to remove uncertainties from the system. The RL system in combination with the LSTM network was used to make the optimal decisions with regards to appliance scheduling of each household and voltage control. The effectiveness of the proposed scheme is depicted in the simulation results, which prove that not only was the overall cost of electricity reduced, but voltage levels were also maintained within the prescribed limits. The work done in this paper is summarized as follows:

1. This paper implemented a decentralized, multi-agent DR system, where each household's load was controlled by its respective agent, and each household agent was controlled by a master agent (SP).

2. The household agents did not need to communicate with each other, reducing the overall complexity of the system, and also making sure that the privacy of the customers was not compromised.

3. This study also took into account the diversity of the households or customers with respect to their load consumption patterns, making sure that the appliance scheduling for each household was done according to its respective preference or priority.

4. This paper employed RL for the optimum scheduling of appliances for each household. RL is adaptive and model free, allowing the SP to independently determine the optimum appliance schedule for each household, without needing to have prior knowledge about the system.

5. This paper accomplished real time performance by predicting the load consumption data of each household through the use of LSTM networks, thus, mitigating future uncertainties.

6. Apart from the optimum scheduling of appliances, this work also implemented a mechanism to maintain the voltage levels across the distribution network of all the 9 households within prescribed limits, using a separate RL agent for voltage control.

7. The electricity cost with and without DR were compared.

In the future, the aim is to extend this work to incentive based DR, which too, forms a core part of the DR field. Furthermore, if available, Real Time Pricing will also be employed in future work.

REFERENCES

[1] F. Alfaverh, M. Denai, and Y. Sun, "Demand response strategy based on reinforcement learning and fuzzy reasoning for home energy management," *IEEE Access*, vol. 8, pp. 39310–39321, 2020.

[2] R. Lu and S. H. Hong, "Incentive-based demand response for smart grid with reinforcement learning and deep neural network," *Appl. Energy*, vol. 236, pp. 937–949, Feb. 2019.

[3] M. F. Tahir, C. Haoyong, I. Ibn, N. Ali, and S. Ullah, "Demand response programs significance, challenges and worldwide scope in maintaining power system stability," *Int. J. Adv. Comput. Sci. Appl.*, vol. 9, no. 6, pp. 1–11, 2018.

[4] X. Kong, D. Kong, J. Yao, L. Bai, and J. Xiao, "Online pricing of demand response based on long short-term memory and reinforcement learning," *Appl. Energy*, vol. 271, Aug. 2020, Art. no. 114945.

[5] L. Wen, K. Zhou, J. Li, and S. Wang, "Modified deep learning and reinforcement learning for an incentive-based demand response model," *Energy*, vol. 205, Aug. 2020, Art. no. 118019.

[6] R. Lu, S. H. Hong, and M. Yu, "Demand response for home energy management using reinforcement learning and artificial neural network," *IEEE Trans. Smart Grid*, vol. 10, no. 6, pp. 6629–6639, Nov. 2019.

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

[8] H. J. Monfared, A. Ghasemi, A. Loni, and M. Marzband, "A hybrid price-based demand response program for the residential micro-grid," *Energy*, vol. 185, pp. 274–285, Oct. 2019.

[9] R. Lu, R. Bai, Z. Luo, J. Jiang, M. Sun, and H.-T. Zhang, "Deep reinforcement learning-based demand response for smart facilities energy management," *IEEE Trans. Ind. Electron.*, vol. 69, no. 8, pp. 8554–8565, Aug. 2022.

[10] M. S. Javadi, A. E. Nezhad, P. H. J. Nardelli, M. Gough, M. Lotfi, S. Santos, and J. P. S. Catalão, "Self-scheduling model for home energy management systems considering the end-users discomfort index within price-based demand response programs," *Sustain. Cities Soc.*, vol. 68, May 2021, Art. no. 102792.

[11] C. Deng and K. Wu, "Residential demand response strategy based on deep deterministic policy gradient," *Processes*, vol. 9, no. 4, p. 660, Apr. 2021.

[12] Y. Du and F. Li, "Intelligent multi-microgrid energy management based on deep neural network and model-free reinforcement learning," *IEEE Trans. Smart Grid*, vol. 11, no. 2, pp. 1066–1076, Mar. 2020.

[13] R. Lu, S. H. Hong, and X. Zhang, "A dynamic pricing demand response algorithm for smart grid: Reinforcement learning approach," *Appl. Energy*, vol. 220, pp. 220–230, Jun. 2018.

[14] H. Li, Z. Wan, and H. He, "Real-time residential demand response," *IEEE Trans. Smart Grid*, vol. 11, no. 5, pp. 4144–4154, Sep. 2020.

[15] P. Zhang, X. Dou, W. Zhao, M. Hu, and X. Zhang, "Analysis of power sales strategies considering price-based demand response," *Energy Proc.*, vol. 158, pp. 6701–6706, Feb. 2019.

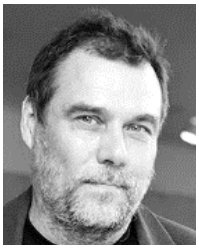
[16] Y. Liu, L. Xiao, G. Yao, and S. Bu, "Pricing-based demand response for a smart home with various types of household appliances considering customer satisfaction," *IEEE Access*, vol. 7, pp. 86463–86472, 2019.

[17] A. Asadinejad, A. Rahimpour, K. Tomsovic, H. Qi, and C.-F. Chen, "Evaluation of residential customer elasticity for incentive based demand response programs," *Electric Power Syst. Res.*, vol. 158, pp. 26–36, May 2018.

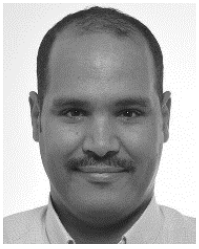
[18] S. He, H. Gao, H. Tian, L. Wang, Y. Liu, and J. Liu, "A two-stage robust optimal allocation model of distributed generation considering capacity curve and real-time price based demand response," *J. Modern Power Syst. Clean Energy*, vol. 9, no. 1, pp. 114–127, 2021.
[19] M. M. Rahman, A. Arefi, G. M. Shafiullah, and S. Hettiwatte, "A new approach to voltage management in unbalanced low voltage networks using demand response and OLTC considering consumer preference," *Int. J. Electr. Power Energy Syst.*, vol. 99, pp. 11–27, Jul. 2018.
[20] S. Davarzani, R. Granell, G. A. Taylor, and I. Pisica, "Implementation of a novel multi-agent system for demand response management in low-voltage distribution networks," *Appl. Energy*, vol. 253, Nov. 2019, Art. no. 113516.
[21] J. Duan, D. Shi, R. Diao, H. Li, Z. Wang, B. Zhang, D. Bian, and Z. Yi, "Deep-reinforcement-learning-based autonomous voltage control for power grid operations," *IEEE Trans. Power Syst.*, vol. 35, no. 1, pp. 814–817, Jan. 2020.
[22] S. Wang, J. Duan, D. Shi, C. Xu, H. Li, R. Diao, and Z. Wang, "A data-driven multi-agent autonomous voltage control framework using deep reinforcement learning," *IEEE Trans. Power Syst.*, vol. 35, no. 6, pp. 4644–4654, Nov. 2020.
[23] Q. Yang, G. Wang, A. Sadeghi, G. B. Giannakis, and J. Sun, "Two-timescale voltage control in distribution grids using deep reinforcement learning," *IEEE Trans. Smart Grid*, vol. 11, no. 3, pp. 2313–2323, Nov. 2019.
[24] A. Nadeem and N. Arshad, "PRECON: Pakistan residential electricity consumption dataset," in *Proc. 10th ACM Int. Conf. Future Energy Syst.*, Jun. 2019, pp. 52–57.
[25] S. Hochreiter and J. Schmidhuber, "Long short-term memory," *Neural Comput.*, vol. 9, no. 8, pp. 1735–1780, 1997.
[26] W. Kong, Z. Y. Dong, Y. Jia, D. J. Hill, Y. Xu, and Y. Zhang, "Short-term residential load forecasting based on LSTM recurrent neural network," *IEEE Trans. Smart Grid*, vol. 10, no. 1, pp. 841–851, Jan. 2019.
[27] Y. Bengio, P. Simard, and P. Frasconi, "Learning long-term dependencies with gradient descent is difficult," *IEEE Trans. Neural Netw.*, vol. 5, no. 2, pp. 157–166, Mar. 1994.
[28] S. Hochreiter, Y. Bengio, P. Frasconi, J. Schmidhuber, "Gradient flow in recurrent nets: The difficulty of learning long-term dependencies," IEEE, Germany, 2001.
[29] C. J. C. Watkins and P. Dayan, "Q-learning," *Mach. Learn.*, vol. 8, nos. 3–4, pp. 279–292, 1992.
[30] A. Arshad, J. Ekström, and M. Lehtonen, "Multi-agent based distributed voltage regulation scheme with grid-tied inverters in active distribution networks," *Electr. Power Syst. Res.*, vol. 160, pp. 180–190, Jul. 2018.
[31] N. Maskey, S. Horsmanheimo, and L. Tuomimaki, "Latency analysis of LTE network for M2M applications," in *Proc. 13th Int. Conf. Telecommun. (ConTEL)*, Jul. 2015, pp. 1–7.
[32] J. Yang, M. Xi, J. Wen, Y. Li, and H. H. Song, "A digital twins enabled underwater intelligent internet vehicle path planning system via reinforcement learning and edge computing," *Digit. Commun. Netw.*, May 2022, doi: 10.1016/j.dcan.2022.05.005.
[33] M. Xi, J. Yang, J. Wen, H. Liu, Y. Li, and H. H. Song, "Comprehensive ocean information-enabled AUV path planning via reinforcement learning," *IEEE Internet Things J.*, vol. 9, no. 18, pp. 17440–17451, Sep. 2022.
[34] D. A. Khan, A. Arshad, and Z. Ali, "Performance analysis of machine learning techniques for load forecasting," in *Proc. 16th Int. Conf. Emerg. Technol. (ICET)*, Dec. 2021, pp. 1–6.
[35] 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.
[36] B. Wang, Y. Li, W. Ming, and S. Wang, "Deep reinforcement learning method for demand response management of interruptible load," *IEEE Trans. Smart Grid*, vol. 11, no. 4, pp. 3146–3155, Jul. 2020.
[37] A. Lesage-Landry and D. S. Callaway, "Batch reinforcement learning for network-safe demand response in unknown electric grids," *Electric Power Syst. Res.*, vol. 212, Nov. 2022, Art. no. 108375.
[38] Z. Li, Z. Sun, Q. Meng, Y. Wang, and Y. Li, "Reinforcement learning of room temperature set-point of thermal storage air-conditioning system with demand response," *Energy Buildings*, vol. 259, Mar. 2022, Art. no. 111903.
[39] A. Amer, K. Shaban, and A. Massoud, "DRL-HEMS: Deep reinforcement learning agent for demand response in home energy management systems considering customers and operators perspectives," *IEEE Trans. Smart Grid*, early access, Aug. 15, 2022, doi: 10.1109/TSG.2022.3198401.

DANYAL AFGAN KHAN received the B.Sc. degree in electrical engineering from the University of Engineering and Technology, Peshawar, Pakistan. He is currently pursuing the M.Sc. degree with the Ghulam Ishaq Khan Institute of Engineering Sciences and Technology, Topi, Pakistan. His research interest includes artificial intelligence techniques employed in power systems for load forecasting and demand response.

AMMAR ARSHAD received the B.Sc. degree in electrical engineering from the University of Engineering and Technology, Lahore, Pakistan, in 2013, and the M.Sc. and D.Sc. degrees from Aalto University, Finland, in 2016 and 2019, respectively. He has been with the Ghulam Ishaq Khan Institute of Engineering Sciences and Technology, since 2020. His main research interests include PV integration in distribution networks, distributed voltage control, enhancement of PV hosting capacity, and utilization of smart grid technologies.



MATTI LEHTONEN received the B.Sc. and M.Sc. degrees in electrical engineering from Aswan University, Aswan, Egypt, in 2008 and 2012, respectively, and the Ph.D. degree from the Electric Power and Energy System Laboratory (EPESL), Graduate School of Engineering, Hiroshima University, Hiroshima, Japan, in 2016. Since 2010, he has been with Aswan University, where he is currently an Associate Professor with the Department of Electrical Engineering, Faculty of Engineering. Since 2019, he has been a Postdoctoral Researcher with the Prof. M. Lehtonen's Power Systems and High Voltage Engineering Group, School of Electrical Engineering, Aalto University, Espoo, Finland. His research interests include power systems, renewable energies, smart grids, distributed generation, optimization, applied machine learning, the IoT, industry 4.0, electric vehicle, and high voltage. Since 2021, he has been a Topic Editor of *Sensors* and *Energies* (MDPI) journals. He has also become a Guest Editor for three Special Issues in *Energies*, *Catalysts*, and *Forecasting* (MDPI) journals. Further, he is a Guest Editor for Special Issues in *Catalysts* and *Forecasting* (MDPI) journals. In 2022, he becomes a Guest Editor Special Issue in *Frontiers in Energy Research* journal on the topic of Smart Grids.



KARAR MAHMOUD (Senior Member, IEEE) received the B.Sc. and M.Sc. degrees in electrical engineering from Aswan University, Aswan, Egypt, in 2008 and 2012, respectively, and the Ph.D. degree from the Electric Power and Energy System Laboratory (EPESL), Graduate School of Engineering, Hiroshima University, Hiroshima, Japan, in 2016. Since 2010, he has been with Aswan University, where he is currently an Associate Professor with the Department of Electrical Engineering, Faculty of Engineering. Since 2019, he has been a Postdoctoral Researcher with the Prof. M. Lehtonen's Power Systems and High Voltage Engineering Group, School of Electrical Engineering, Aalto University, Espoo, Finland. His research interests include power systems, renewable energies, smart grids, distributed generation, applied machine learning, and electric vehicles. Since 2021, he has been a Topic Editor of *Sensors* and *Energies* (MDPI) journals. He has also become a Guest Editor of three special issues in *Energies*, *Catalysts*, and *Forecasting* (MDPI) journals. Further, he is a Guest Editor for Special Issues in *Catalysts* and *Forecasting* (MDPI) journals. In 2022, he becomes a Guest Editor Special Issue in *Frontiers in Energy Research* journal on the topic of Smart Grids.