IEEE Access
Multidisciplinary ⁝ Rapid Review ⁝ Open Access Journal

Received 10 July 2023, accepted 1 August 2023, date of publication 7 August 2023, date of current version 17 August 2023.

Digital Object Identifier 10.1109/ACCESS.2023.3302895

RESEARCH ARTICLE

Renewable Energy Maximization for Pelagic Islands Network of Microgrids Through Battery Swapping Using Deep Reinforcement Learning

M. ASIM AMIN[1,2], (Graduate Student Member, IEEE), AHMAD SULEMAN[2], MUHAMMAD WASEEM[3], (Member, IEEE), TAOSIF IQBAL[4], SADDAM AZIZ[3], (Senior Member, IEEE), MUHAMMAD TALIB FAIZ[3], LUBAID ZULFIQAR[2], AND AHMED MOHAMMED SALEH[5]

[1]Department of Electrical, Electronic, Telecommunications Engineering and Naval Architecture, University of Genoa, 16145 Genoa, Italy
[2]Rapid Volt (PVT) Ltd., Rajanpur, Punjab 33500, Pakistan
[3]Department of Electronics and Information Engineering, The Hong Kong Polytechnic University, Hong Kong
[4]Electrical Engineering Department, College of Electrical and Mechanical Engineering, National University of Sciences and Technology (NUST), Islamabad 46000, Pakistan
[5]Electrical Engineering Department, University of Aden, Aden, Yemen

Corresponding authors: M. Asim Amin (masim.amin@yahoo.com) and Ahmed Mohammed Saleh (engahmedsaleh14@gmail.com)

ABSTRACT The study proposes an energy management system of pelagic islands network microgrids (PINMGs) based on reinforcement learning (RL) under the effect of environmental factors. Furthermore, the day-ahead standard scheduling proposes an energy-sharing framework across islands by presenting a novel method to optimize the use of renewable energy (RE). Energy sharing across islands is critical for powering isolated islands that need electricity owing to a lack of renewable energy supplies to fulfill local demand. A two-stage cooperative multi-agent deep RL solution based on deep Q-learning (DQN) with central RL and island agents (IA) spread over several islands has been presented to tackle this difficulty. Because of its in-depth learning potential, deep RL-based systems effectively train and optimize their behaviors across several epochs compared to other machine learning or traditional methods. As a result, the centralized RL-based problem of scheduling charge battery sharing from resource-rich islands (SI) to load island networks (LIN) was addressed utilizing dueling DQN. Furthermore, due to its precise tracking, the case study compared the accuracy of various DQN approaches and further scheduling based on the dueling DQN. The need for LIN is also stochastic because of variable demand and charging patterns. Hence, the simulation results, including energy scheduling through the ship, are confirmed by optimizing RE consumption via sharing across several islands, and the effectiveness of the proposed method is validated by state and action perturbation to guarantee robustness.

INDEX TERMS Deep reinforcement learning, pelagic island, microgrids, EMS, renewable energy.

NOMENCLATURE

$r_{i,t,m}$	The energy level of the storage cluster at time *t*.
$EL^{b}_{i,t,m}$	The energy level of each battery at the time *t*.
$N^{CS}_{i,t,s}/N^{BS}_{i,t,s}/N^{Ship}_{i,t}$	Number of batteries at CS/BS/Ship at time. *t*
$N^{CS,C}_{i,t,s}/N^{CS,D}_{i,t,s}$	CS charge/discharge batteries at time *t*.
$N^{BS,C}_{i,t,s}/N^{BS,D}_{i,t,s}$	BS charge/discharge batteries at time. *t*
$N^{C}_{i,t,s}/N^{D}_{i,t,s}$	Ship charge/discharge batteries at time. *t*
$P^{ch,b}_{i,t,m}/P^{dis,b}_{i,t,m}$	Charging/discharging of each battery.
η^{ch}/η^{dis}	Charging/discharging efficiency of each batter.
C^{TOU}_{t}/C_{ship}	Swapping cost of *b* /battery shipping cost of *b*.
C_{de}	Charging cost of degradation cost of *b*.
$G_{i,t,m}/B_{i,t,m}$	Conductivity/admittance of line *m*.

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

$\Re$	Replacement cost of the battery.
$P_{i,t,m}^{DGE}/Q_{i,t,m}^{DGE}$	Active/reactive power of DGE at time t.
$P_{i,t,m}^{\mu T}/Q_{i,t,m}^{\mu T}$	Active/reactive power of MT at time t.
$\upsilon_{i,t,m}^{\mu T}/\upsilon_{i,t,m}^{DGE}$	Cost of fuel or maintenance for MT/DGE at time t.
$P_{i,t,m}^{l}/Q_{i,t,m}^{l}$	Active/reactive power of loads at time t.
$Q_{i,t,m}^{b}$	The battery at time t provides reactive power.
$P_{i,t,m}^{l,sht}/Q_{i,t,m}^{l.sht}$	Active/reactive power of shut loads at time t.
$u_{i,t}^{\mu T}/u_{i,t}^{DGE}$	Control variables of MT/DGE at time t.
$\theta_{i,t,m}$	The phase difference at the bus m.
$P_{i,t}^{PV'}/P_{i,t}^{WT'}$	Predicted power generation of PV/WT at time t.
$V_{i,t,m}$	Voltage magnitude at bus m.
$P_{i,t}^{cur}$	Renewable energy curtailment at time t.
loc	Ship location as (l_x, l_y).
$\varpi^{r}, \varpi^{SH}, \varpi^{l}$	Cost coefficient of storage limit, shipping violation, and load curtailment.
$\omega_{i\rightarrow v,t,s}$	Binary variable for the virtual node for transition between multiple islands i through ship s.
$\omega_{i,t,s}$	Binary variable for the arriving at island i through ship s.
ζ^{l}	Load importance factor.
$s_t^i/a_t^i/\pi$	State space/action space/ policy.
$\rho_{i,t}^{l}$	Load curtailment in the resource-rich island.
α/β	Parameters of connected layers of DNN.
θ/γ	Weight function of prediction/discount factor.
$Q(s, a, \theta)$	State-action weighted function of Q-network.

Notion and Indices:

BS	Battery swapping station.
CS	Charging station.
C	Charge batteries.
D	Discharge batteries.
b	Individual battery parameter.
s	Ship number.
$ship$	Used for the batteries at the ship.
i, j	No. of islands.
m	No. of bus.
t	Time step.
v	Virtual node for ship transition.
l	Connected loads.
$b\,\epsilon N_{i,t}^{CS}/N_{i,t}^{BS}$	Index of No. of batteries.
$i\,\epsilon N_{CS}/N_{BS}$	Index of island number.
$s\epsilon Ship$	Index of ship.
$m,\ n\epsilon\Psi$	Bus nodes of microgrids on islands.
$l\epsilon L$	Number connected loads.

I. INTRODUCTION

Renewable energy offers exciting solutions to the 21st century's fundamental and significant environmental issues. Their incorporation into the current system poses several technological and societal hurdles regarding safe and efficient energy management. Because of the high integration of distributed energy resources (DERs) and renewable generators has changed distributed energy storage facilities into power systems, and the grid has been upgraded from passive to active. Where dispersed resources are deployed close to the consumer load [1].

Researchers have discovered that extensive penetration could weaken the grid's stability and result in blackouts because renewable energy sources are intermittent. For a microgrid (MG), mitigating is viewed as a quick fix. It has several advantages over the primary grid, including durability and dependability, advantages against unsettling cycles of events brought on by nature or humans, and operation as a controllable border [2].

Due to the climate and geographic position, it might be difficult to electrify rural settlements and isolated areas, such as islands, which are often powered by diesel generators. Due to generator aging issues, higher fuel prices, logistical difficulties, and higher emissions of greenhouse gases, few people have access to just power resources. A MILP-based challenge was put up in the literature [3] for the isolated microgrid to implement demand response and renewable energy to reduce generating costs. To reduce prediction error and improve accuracy, which would impact MG operation, a coordinated model predictive control (MPC)-based architecture has been explored [4]. Most of the research has concentrated on many features, such as load curtailment, energy storage, demand response, and forecasts of renewable energy for the electrification of remote areas or islands based on diverse applications. However, there has not been a distributed strategy for PINMGs that considers the load demand and the intermittent nature of RES via network storage system collaboration with nearby islands.

Throughout the next ten years, it will be necessary to cut CO2 emissions, stop using fossil fuels, and improve sustainable power sources due to the rising need for energy. While a few islands have installed RE-like photovoltaic (PV) and wind turbines (WT), they cannot meet the power demand. Some islands based on available resources, such as source Islands (SI), have been enhanced with renewable energy resources and certain natural gas exploration operations set up close to or on the islands. Another load island network (LIN) has a high local energy demand. These islands have a big area and local restoration, but the locally produced energy is insufficient to meet the demand. The grid or neighboring islands must aid in satisfying LIN demand since local generation on LIN is inadequate to supply the local market.

PINMGs are islanded microgrid networks that function together between SI, SLI, and LIN to transfer charged batteries at LIN to meet demand. It also encourages using local diesel generators and microturbines (MT) to meet urgent energy demands. Diesel generators and gas-powered power plants must be reduced to maintain a clean atmosphere and reduce CO2 emissions. The swapping of batteries between islands encourages the usage of RE. It was not considered an economical option, but as time progressed, it gained popularity since it reduced the need to charge for wait periods. The Chinese government supports the battery-swapping business model; 1400 stations must be operational by February 2022 and 26000 by 2025 [5].

Several researchers have worked on multi-microgrids to address the energy management challenge. Several solutions and suggested methods addressed the deficiencies described above [6], [7], [8]. There is not enough common platform to consider the pelagic island's energy management, load, and demand response. A mobile energy storage system has recently gotten much attention, and a joint-disaster network structure for scheduling MGs production and reconfiguration is expected to lower total costs [9]. A resilient routing two-stage architecture explored mobile power dispatching and distribution system operators to test the efficacy of routing and flexibility augmentation in the literature [10]. The decentralized energy management system uses two-stage stochastic programming to offer a steady demand-supply for grid-connected and islanded operations [11].

According to the research method and contribution based on the collaborative structure, participation, and load, management has been discussed in the literature [12], [13], [14], [15], [16], [17], [18], [19] and shown in Table 1. The comparison indicates that several research gaps still need to be filled for the optimal operation of pelagic islands. Most of the work is based on the island microgrid operation in literature, and there is no standard platform for the pelagic island's optimized structure and techniques. It is worth noting that many methods have considered the energy-sharing problem and studied many different aspects regarding the participation and sharing structure given in Table 1 from [12], [13], [14], [15], [16], [17], [18], and [19]. The most relevant literature has considered many different methods to schedule appropriate battery sharing among different islands. The existing methods deliberate only the SI or SLI but do not consider the island network or distribution system without sharing structure between islands. Likewise, the proposed approaches hold drawbacks as they are not cost-effective and exclude the use of power flow parameters.

According to United Nations Goal 7, it is critical to ensure that every user has simple access to clean and cheap energy [20]. According to European Research and Innovation magazine, the islands suffer high power prices while utilizing a large amount of imported fossil fuel. As a result, distant islands begin energy independence and rely on large RE to keep the ecosystem clean and green [21]. In [22], the Indonesian commitment to the Paris Agreement to cut carbon emissions by 29% by 2030 has begun the electrification of 1000 islands to encourage RE. As a result, a robust and secure energy-sharing infrastructure that is cost-effective and maximizes RE consumption on the island level or between islands is required.

The difficulty is to execute adequate scheduling across numerous islands while managing energy locally on each island. There has recently been a tremendous breakthrough in the computationally demanding job of solving challenges like Atari [23], StarCraft [24], and AlphaGo [25] utilizing deep RL-based techniques. As a result, the computationally difficult tasks used by the deep RL multi-agent-based resource management job are considerably simpler and smaller than those used by other traditional systems, making it ideal for more accurate and real-time scheduling of large-scale issues. This paper proposes a multi-agent-based deep reinforcement learning-based distributed approach for multiple islands to share their energy.

Most importantly, the power flow constraints have been considered while formulating the discussed problem. In the following Table 1, the comparison of the related literature and its shortcomings has been discussed. This proposed problem has been considered to minimize the power generation cost DERs, batteries, and DGs and support the load locally on each island. The significance and contribution of this paper:

- A novel energy management framework for pelagic islands is being developed, emphasizing the availability of various intermittent supplies based on environmental conditions.
- Due to environmental constraints affecting energy transfer via ship, a simplified approach is presented to minimize local load and power exchange with neighboring islands using the proposed grid-based deep RL approach.
- The suggested structure illustrates and tracks the day-ahead forecast, matched with hour-based scheduling to satisfy the real-time demand for approaching ships through load curtailment/management, or DEG/MT.
- To overcome the computational expense and minimize training time in RL agents, a ship traveling between islands is justified using the discrete state and action space.

The following assumption was taken for the case study to simplify the assumptions:

- Only one ship may be linked with each island at a time.
- The renewable forecasting error has not been considered, and the loads are flexible to adjust based on critical and non-critical demand to meet the real-time schedule demand.
- All of the batteries are the same kind and can only be shared until completely charged at the CS, and it continues to discharge until the SOC at the BS hits the lower level.

This paper is organized as follows: Section II outlines the pelagic island structure, Section III explains the energy-sharing structure in detail, and Section IV organizes the energy management system (EMS) and expresses energy management at the individual island level. Section V

TABLE 1. Comparison of a related work in the existing literature.

Ref.	Proposed method	Cost-effective	Power flow	Participation			Sharing structure	
				LIN	SLI	SI	Pelagic	Ship
[11]	Multi-object practical swarm	x	x	x	✓	x	✓	✓
[12]	MILP	x	✓	x	✓	x	x	x
[13]	MPC	✓	x	x	✓	x	x	x
[14]	MILP	✓	x	x	✓	x	x	x
[15]	Neural network	x	x	x	✓	x	x	x
[16]	Nash equilibrium	✓	x	✓	x	x	✓	✓
[17]	MILP	x	x	✓	x	x	✓	✓
[18]	Convex relaxation	✓	✓	x	✓	x	x	x
This work	Multi-agent RL based approach	✓	✓	✓	✓	✓	✓	✓

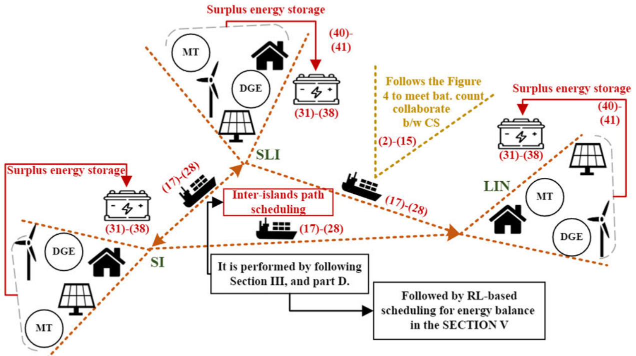


FIGURE 1. Structure of pelagic island microgrids.

additionally discusses the proposed deep RL-based technique, the reward and loss functions, and ship scheduling for accurate scheduling. Furthermore, the suggested issue flowchart has been designed to show the flow of the proposed problem in Section VI. Section VII presents a case study of the suggested issue, then follows Section VIII's conclusion.

II. PELAGIC ISLAND STRUCTURE

The structure of the pelagic islands network microgrids PINMGs has been shown in Figure 1. This networked microgrid has considered three types of pelagic source Islands (SI), load islands (LIN), and a combination of both load and source islands (SLI). Sources islands are enriched with renewable energy resources such as solar and wind. The load island has considered residential and commercial users such as hospitals and critical industries. The islands' energy trading is done with ship swapping to transport the storage batteries from the enriched islands to the load islands.

It is essential to highlight that energy sharing across islands plays a significant role when resources are inadequate to meet local demand or insufficient RE production. As a result, the resources with abundant production distribute excess energy through the charged storage cluster via ships to other distant islands to fulfill local demand and achieve net-zero emission by maximizing renewable energy resources.

A. GOAL OF THE PROPOSED PROBLEM

For the proposed work, this problem is designed to optimally schedule and transfer electricity from resource-rich to load-rich islands. To achieve the objective, the centralized RL-agent-based energy management is proposed to ensure local energy fulfillment in each island and share the surplus energy with the neighboring island by charging at charging stations and sharing the batteries with the neighboring islands through battery swapping using the ship.

The extra energy is stored in the storage cluster batteries using the energy balance equation provided in the EMS level formulation. The storage cluster functions as an energy storage warehouse, transferring energy locally and storing excess power from the surrounding island. The load islands may be motivated to sell electricity by storing it in the storage cluster and making it accessible for exchange with neighboring islands. It is accomplished by lowering aggregated load demand in each island during peak hours to reduce power reliance or interaction during off-schedule times. Its choices on additional charges and discharges are routed via the islanded microgrid operator (IMGO), and loads on each load island control their capacity. Individual decisions are required to reduce energy use during peak hours. It transfers the data to the centralized RL agent in the energy-sharing layer, as shown in Figure 2.

B. OBJECTIVE

The energy management challenge is separated into two stages with the pelagic islands. To begin, the energy-sharing system will address the problem of energy transmission across the various islands through the RL centralized agent. Second, as an IA agent, the IMGO/EMS level structure controls the island's power flow and supply-demand balance. It makes it easier for market clearing, transactive exchange, and analogues to (ISO) independent system operators.

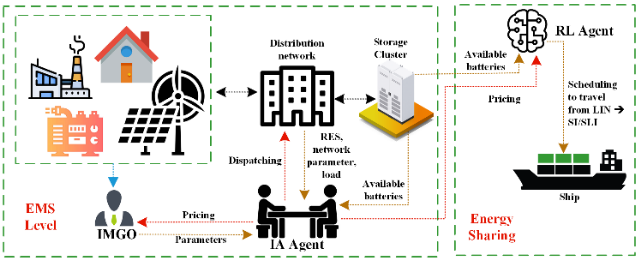


FIGURE 2. Transactive energy-sharing structure.

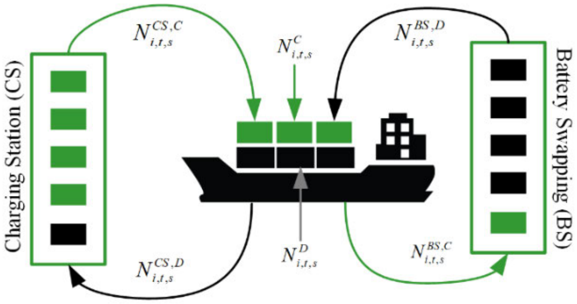


FIGURE 3. Battery swapping platform using a ship.

IMGO also serves as an aggregator for the interconnected system, which includes DER (distributed energy resources) and MG generators such as DEGs (diesel generators). Furthermore, distributed system consumers in the local MG act as agents of local energy management in each MG. Through the IA agent, IMGO, on the other hand, may control the distribution system's local resources and transactive interaction depending on its state and time-of-use (TOU) pricing. The main goal function will assess the most cost-effective method of balancing two layers and use the available resources at the demand end. IMGO and IA agents collaborate to satisfy local generation and demand scheduling through battery swapping with other islands in a day-ahead profile.

III. ENERGY SHARING

Ship swapping between islands allows for energy sharing by exchanging charged batteries via the ship from the source island and from the ship to the load island to fulfill demand and supply balance by raising the storage cluster state-of-charge (SOC).

Furthermore, by providing a better distribution of battery swapping by utilizing the ship on a certain island and joining with adequate batteries for circulation, the problem of capacity and charging delay might be eliminated. Every island has its battery-swapping station, and surplus power produced on the SI/SLI is used to charge the drained batteries. Figure 3 depicts the battery flow. In the case of pelagic islands, electricity transfer between islands through battery swapping is being studied. It is anticipated to fulfill the energy demand in each microgrid, and a mismatch in power and demand balance might result in power flow across islands.

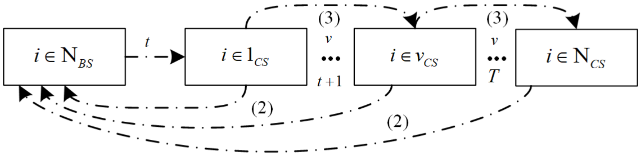


FIGURE 4. Demand fulfillment at BS through several CS collaboration.

A. BATTERY SWAPPING MODEL

Furthermore, the battery count also affects the swapping contribution because it defines its maximum capacity. This approach's main objective is to deal with the optimal use of individual batteries and make the environment more sustainable [26]. The main contribution is to propose the battery swapping (BS) system's infrastructure, which comprises slots for a swapping station and maximum holding capability to yield a total and stable profit. The objective function for shipping profit can be expressed at BS,

$$BS^{pro}_{i,t,BS} = \sum_{t=0}^{T}\left[C^{TOU}_{t}\left[\sum_{j\in N_i, j\neq i} N^{BS,C}_{ij,t,s} \times EL_{j,t}\right] - C_{ship}N^{Ship}_{i,t,s} - C_{de}N^{BS}_{i,t,s}\right] \tag{1}$$

The first term is the total revenue earned by the b batteries swapping with TOU price at BS, the number of available batteries in CS, the battery shipping cost between islands, and the battery degradation cost (38) used at the BS, respectively. Batteries' degradation costs can be minimized by reducing the full battery discharging and overnight charging. In addition, the profit is counted as the batteries sail from CS and reach BS. This swapping process starts with the ship's arrival, enters the CS, exchanges the discharged batteries with the fully charged batteries, leaves the CS, and sails towards the BS station. On the other hand, the batteries at the ship at any time can be calculated $N^{Ship}_{i,t,s} = N^{C}_{i,t,s} + N^{D}_{i,t,s}$.

In addition, the battery swapping service comprises discrete cumulative path methods, including the fulfilled demand and redirected demand, respectively.

Definition 1 (Fulfilled Demand): The fulfilled demand is the total number of batteries fulfilled through the CS at the BS in the time interval T.

$$N^{CS,C}_{i,t,s} = N^{D}_{i,t,s} \tag{2}$$

Definition 2 (Redirected Demand): The redirected request is the total number of batteries not swapped at CS and redirected to another CS to swap the remaining uncharged batteries in the time interval T.

$$N^{D}_{i,t,s} \geq N^{CS,C}_{i,t,s} \tag{3}$$

Therefore, the flow of demand fulfillment through the ship can be visualized in Figure 4.

The grid-based RL scheduling for energy sharing between islands is proposed in *SECTION* V. In addition, *C (inter-island scheduling)* collaborates with the neighboring islands to swap the discharge batteries before leaving for the BS station.

However, the total number of batteries counted at BS can be calculated as follows:

$$N^{BS,D}_{i,t,s} = \sum_{j\in N_i, j\neq i} N^{BS,D}_{ij,t,s}, \quad \forall i\in \mathrm{N}_{BS},\ s\in Ship \tag{4}$$

$$N^{BS,C}_{i,t,s} = \sum_{j\in N_i, j\neq i} N^{BS,C}_{ij,t,s}, \quad \forall i\in \mathrm{N}_{BS},\ s\in Ship \tag{5}$$

$$N^{BS,D}_{i,t,s} + N^{BS,C}_{i,t,s} = N^{BS}_{i,t,s}, \quad \forall i \in \mathrm{N}_{BS},\ s \in Ship \tag{6}$$

$$N^{BS,rem}_{i,t,s} = N^{BS,D}_{i,t,s} - N^{C}_{i,t,s}, \quad \forall i\in \mathrm{N}_{BS},\ s\in Ship \tag{7}$$

$$N^{BS,D}_{i,t,s} \leq N^{C}_{i,t,s}, \quad \forall i \in \mathrm{N}_{CS},\ s \in Ship \tag{8}$$

$$N^{BS,rem}_{i,t,s} \geq 0, N^{BS,rem}_{i,t,s} = N^{BS}_{i,t,s}, \quad \forall i\in \mathrm{N}_{BS},\ s\in Ship \tag{9}$$

where (4)-(5) are the total number of discharge and charge batteries that flow between the ship and storage cluster at BS for the time t. In addition, the total number of batteries at BS is calculated through (6), (7) gives the remaining batteries count at BS, and (8) depicts that the discharge batteries count at the BS must be fulfilled through battery swapping.

B. CHARGING STATION MODEL

The modeling of a battery CS on an island is a discrete model with a limited number of batteries on different islands based on the available capacity with the single row of ships to share the charged batteries with neighboring islands [20]. This swapping process starts with the ship's arrival, enters the CS, exchanges the discharged batteries with the fully charged batteries, leaves the CS, and sails towards the BS station [21].

$$N^{CS,D}_{i,t,s} = \sum_{j\in N_i, j\neq i} N^{CS,D}_{ij,t,s}, \quad \forall i\in \mathrm{N}_{CS},\ s\in Ship \tag{10}$$

$$N^{CS,C}_{i,t,s} = \sum_{j\in N_i, j\neq i} N^{CS,C}_{ij,t,s}, \quad \forall i\in \mathrm{N}_{CS},\ s\in Ship \tag{11}$$

$$N^{CS,D}_{i,t,s} + N^{CS,C}_{i,t,s} = N^{CS}_{i,t,s}, \quad \forall i \in \mathrm{N}_{CS},\ s \in Ship \tag{12}$$

$$N^{CS,rem}_{i,t,s} = N^{CS,C}_{i,t,s} - N^{C}_{i,t,s} + N^{D}_{i,t,s}, \quad \forall i \in \mathrm{N}_{CS},\ s \in Ship \tag{13}$$

$$\sum_{t=1}^{t-(T_c-1)} N^{CS,C}_{i,t,s} \geq \sum_{t=1}^{T} N^{D}_{i,t,s}, \quad \forall i \in \mathrm{N}_{CS},\ s \in Ship \tag{14}$$

$$N^{CS,rem}_{i,t,s} \geq 0, N^{CS,rem}_{i,t,s} = N^{CS}_{i,t,s}, \quad \forall i\in \mathrm{N}_{CS},\ s\in Ship \tag{15}$$

where (10)-(11) are the total number of discharge and charged batteries that flow between the ship and storage cluster at CS in time t. In addition, the total number of batteries at CS is calculated through (12), (13) gives the remaining batteries at CS to be charged, and (16) gives the time for the battery to stay undercharging for at least T_c time slot, and it is calculated as follows:

$$T_c = \left[EL^{b}_{i,t,m} - EL^{b}_{i,t-1,m} \Big/ \eta^{ch} P^{ch,b,\max}_{i,t,m} \right] \tag{16}$$

(15) expresses that the number of charged batteries at the CS must equal the total battery capacity after the swapping [30].

C. SHIP-BASED BATTERY SCHEDULING

In a scheduling model for a ship between CS and BS using grid-based scheduling, two nodes are considered: the transit node (enabling battery transfer between CS/BS through the ship) and the parking node (loading and unloading of batteries at CS/BS). The mathematical modeling for the ship is formulated as follows:

$$\sum_{i\in CS/BS} \omega_{i,t,s} = \begin{cases} \sum_{i\in CS/BS} \omega_{i,t,s}, & if\ N^{CS,C}_{i,t,s} = N^{D}_{i,t,s} \\ \sum_{i\in CS/BS} \omega_{i\rightarrow v,t+1,s}, & if\ N^{D}_{i,t,s} \geq N^{CS,C}_{i,t,s} \end{cases} \tag{17}$$

$$\sum_{i\in CS/BS} \omega_{i,1,s} = \omega_{i,0,s}, \quad \forall s \in Ship \tag{18}$$

(17) depicts that ship s is at transit nodes or parking nodes by fulfilling the demand and the flow between CS/BS through the virtual node v with a binary variable $\omega_{i\rightarrow v,t+1,s}$ at $t+1$. In addition, (18) gives the initial position of the ship.

$$\sum_{j\in N_i, j\neq i} N^{D}_{ij,t,s} = \sum_{j\in N_i, j\neq i} N^{BS,D}_{ij,t,s} - \sum_{j\in N_i, j\neq i} N^{CS,D}_{ij,t,s}, \quad \forall s\in Ship,\ t\in T \tag{19}$$

$$\sum_{j\in N_i, j\neq i} N^{C}_{ij,t,s} = \sum_{j\in N_i, j\neq i} N^{CS,C}_{ij,t,s} - \sum_{j\in N_i, j\neq i} N^{BS,C}_{ij,t,s}, \quad \forall s \in Ship,\ t \in T \tag{20}$$

$$\sum_{j\in N_i, j\neq i} N^{C}_{ij,t,s} = N^{C}_{i,t,s}, \sum_{j\in N_i, j\neq i} N^{D}_{ij,t,s} = N^{D}_{i,t,s} \quad \forall i \tag{21}$$

$$N^{D}_{i,t,s} + N^{C}_{i,t,s} = N^{ship}_{i,t,s}, \quad \forall s \in Ship,\ t \in T \tag{22}$$

$$N^{D}_{i,t,s} \geq 0, N^{C}_{i,t,s} \geq 0, \quad \forall s \in Ship,\ t \in T \tag{23}$$

$$N^{D}_{i,T,s} = 0, N^{C}_{i,T,s} = 0, \quad \forall s \in Ship,\ t \in T \tag{24}$$

(19)-(20) provides the charge and discharge batteries carried by the ship s; the total batteries (22), (23) give the stock at the ship should be non-negative, and (21) is the default which gives the sum of batteries either at the CS or BS. In (24), the charge and discharge batteries should be unloaded from the ship at time T.

$$0 \leq N^{BS,D}_{i,t,s} \leq C_s.\omega_{i,t,s}, \quad \forall s \in Ship,\ i \in \mathrm{N}_{BS} \tag{25}$$

$$0 \leq N^{BS,C}_{i,t,s} \leq N^{C}_{i,t-1,s}.\omega_{i,t,s}, \quad \forall s \in Ship,\ i \in \mathrm{N}_{BS} \tag{26}$$

$$0 \leq N^{CS,D}_{i,t,s} \leq N^{D}_{i,t-1,s}.\omega_{i,t,s}, \quad \forall s \in Ship,\ i \in \mathrm{N}_{CS} \tag{27}$$

$$0 \leq N^{CS,C}_{i,t,s} \leq C_s.\omega_{i,t,s}, \quad \forall s \in Ship,\ i \in \mathrm{N}_{CS} \tag{28}$$

Therefore, the (25)-(26) gives the feasible set of charge and discharge batteries swapping between the ship and BS. Similarly, (27)-(28) provide the possible batteries swapping between the ship and CS [32].

Remark 1: It is important to note that battery swapping can only be performed while the ship reaches the parking node, either at BS/CS with $\omega_{i,t,s} = 1$ and $\omega_{i,t,s} = 0$ otherwise.

Remark 2: It is assumed that fully charged/discharged batteries unloading at BS/CS are placed at the start of time t, whereas loading of charge/discharge batteries CS/BS is done at the end of time slot T.

D. INTER-ISLANDS PATH SCHEDULING

For ship sailing and inter-island scheduling, the RL-based methods with a 2D grid-based approach are considered, and ship location as $loc = (l_x, l_y)$ with an agent starting at LIN (BS) to collect the batteries $N^{CS}_{i,t,s}$ after sailing at the ship from SI or SLI (CS) by following the binary variable $\omega_{i,t,s}$ in a minimal number of steps $l_x + l_y$, and return to BS. Therefore, the agent must take action left, right, up, and down by covering grid spaces to collect the charge batteries from CS within a minimum time. The proposed validation is discussed through the proposed RL algorithm and the case study.

IV. EMS LEVEL

The various islands anticipate attaining demand and supply balance, and the discrepancy between demand and supply represents energy transfer between other islands through ship swapping.

A. STORAGE CLUSTER

Because the primary goal of a storage system is to maintain energy balance on a tiny island, and battery life falls exponentially with the depth of discharge and battery efficiency [33], consider extending battery life by limiting the discharge cycle and battery model, which are specified as:

1) CS & BS CONSTRAINTS

To ensure the storage cluster energy level for each period on both BS and CS within the bound and expressed as:

$$r^{\min}_{i,t,m} \leq r_{i,t,m} \leq r^{\max}_{i,t,m} \tag{29}$$

$$r_{i,t,m} = \sum_{j \in N_i, j \neq i} N^{BS}_{ij,t,s} \times EL_{j,t,m} + (N^{CS}_{i,t,s} - N^{Ship}_{i,t,s}) \times EL_{i,t,m}, \ \forall i \tag{30}$$

(29) expresses the storage cluster's SOC boundary limits, and (30) the SOC level of different islands concerning several batteries.

2) ENERGY CONSTRAINTS AT CS

To meet the demand at the CS, the energy consumed at each time instant t should be fully used for the charge batteries offered for the swapping as follows:

$$\sum_{t=1}^{t-T} \left[\eta^{ch} P^{ch,b}_{i,t,m} - P^{dis,b}_{i,t,m} \Big/ \eta^{dis} \right] \Delta t \geq \sum_{t=1}^{\tau} N^{CS,C}_{i,t,s} \times EL_{i,t,m}, \quad \forall t \in T \tag{31}$$

In addition, (32) indicates that by the end of each operation period at time T, most of the energy at CS is utilized to produce fully charged batteries.

$$\sum_{t=1}^{T} \left[\eta^{ch} P^{ch,b}_{i,t,m} - P^{dis,b}_{i,t,m} \Big/ \eta^{dis} \right] \Delta t = \sum_{t=1}^{T} N^{CS,C}_{i,t,s} \times EL_{i,t,m}, \quad \forall t \in T \tag{32}$$

In (33), the maximum energy consumed by the CS for the time t cannot exceed the total discharge batteries available at the CS before swapping to the ship.

$$\sum_{t=1}^{t-T} N^{CS,C}_{i,t,s} \times EL_{i,t,m} \leq \sum_{t=1}^{t-T} N^{D}_{i,t,s} \times EL_{i,t,m}, \quad \forall t \in T \tag{33}$$

The battery charging and discharging with $EL^{b}_{i,t,m}$ power limits are bound between upper and lower limits to maintain battery life [23] with charging/discharging efficiency, and each battery at BS or CS should follow the following constraints:

$$EL^{b,\min}_{i,t,m} \leq EL^{b}_{i,t,m} \leq EL^{b,\max}_{i,t,m} \tag{34}$$

$$P^{ch,b,\min}_{i,t,m} \leq P^{ch,b}_{i,t,m} \leq P^{ch,b,\max}_{i,t,m} \tag{35}$$

$$P^{dis,b,\min}_{i,t,m} \leq P^{dis,b}_{i,t,m} \leq P^{dis,b,\max}_{i,t,m} \tag{36}$$

To avoid malfunction [24], charging and discharging at the same time can be avoided as follows:

$$P^{ch,b}_{i,t,m} \times P^{dis,b}_{i,t,m} = 0 \tag{37}$$

It is important to note that frequent charge/discharge cycles could shorten battery life. Therefore, the degradation cost per depth of discharge (DoD) is as follows:

$$C_{de} = \Re \big/ N_{cyc} \tag{38}$$

It gives the cost per cycle for every DoD. We calculate the whole storage cluster and store the cost every cycle until we get the capital cost and replace the battery and the cycle number given in [25].

B. LOCAL IA FORMULATION

For this problem, the goal is to maximize utilization of RE, energy sharing from the local DGs would not be so high and related to μT (micro-turbine), DGE (diesel generator energy), and consider the individual generator's cost proportional to the power generator by them. To trade-off with these costs, the cost function to be minimized and formulated for the time t can express as,

$$F^{DG}_{i,t,m} = \min \left[\sum_{\mu T=1}^{MT} \upsilon^{\mu T}_{i,t,m} P^{\mu T}_{i,t,m} + \sum_{DGE=1}^{G} \upsilon^{DGE}_{i,t,m} P^{DGE}_{i,t,m} \right], \quad \forall i \in \mathrm{N}_{CS}/\mathrm{N}_{BS}, t \in T \tag{39}$$

Hence, the power balance equations are described as follows (40)-(43):

$$\sum_{PV'} P^{PV'}_{i,t,m} + \sum_{WT'} P^{WT'}_{i,t,m} + \sum_{\mu T} P^{\mu T}_{i,t,m} + \sum_{DGE} P^{DGE}_{i,t,m} - \sum_{l} P^{l}_{i,t,m} + \sum_{b\in N^{CS}_{i,t,s}/N^{BS}_{i,t,s}} \begin{matrix} P^{dis,b}_{i,t,m} \\ -P^{ch,b}_{i,t,m} \end{matrix} + \sum_{l} P^{l,sht}_{i,t,m} - \sum_{i} P^{cur}_{i,t,m} = P_{i,t,m};\ \forall m,n\in\psi,\ \forall i\in \mathrm{N}_{CS}/\mathrm{N}_{BS} \tag{40}$$

$$\sum_{\mu T} Q^{\mu T}_{i,t,m} + \sum_{DGE} Q^{DGE}_{i,t,m} - \sum_{l} Q^{l}_{i,t,m} + \sum_{N^{CS}_{i,t,s}/N^{BS}_{i,t,s}} Q^{b}_{i,t,m} + \sum_{l} Q^{sht}_{i,t,m} = Q_{i,t,m};\ \forall m,n\in\psi,\ \forall i\in \mathrm{N}_{CS}/\mathrm{N}_{BS} \tag{41}$$

$$u^{DGE}_{i,t} P^{DGE,\min}_{i,t,m} \leq P^{DGE}_{i,t,m} \leq u^{DGE}_{i,t} P^{DGE,\max}_{i,t,m} \tag{42}$$

$$u^{\mu T}_{i,t} P^{\mu T,\min}_{i,t,m} \leq P^{\mu T}_{i,t,m} \leq u^{\mu T}_{i,t} P^{\mu T,\max}_{i,t,m} \tag{43}$$

Furthermore, the $P_{i,t,m}/Q_{i,t,m}$ gives the active/reactive power injection at the bus m, respectively. Therefore, the backup provided by DGE and MT is given in (42)-(43). It is important to note that the devices at the bus m should follow the node power balance.

1) IA OPERATION CONSTRAINTS

Conventionally, the objective function of optimal power flow (OPF) is to minimize the total production cost of active power generation. Therefore, the power flow model is adopted [38] to model the network topology and power flow constraints. This proposed model can simultaneously formulate the nodal voltage and branch power flow and helps to protect the power system's security. Hence, IA adopts the power flow model that does not affect the solvability of the optimization problems with the polar coordinates for voltage is $V_{i,t,m}\angle\theta_{i,t,m}$ for bus m and the branch $P_{i,t,mn}/Q_{i,t,mn}$ active/reactive power flow from the bus m to bus n can be expressed as (44)-(45):

$$P_{i,t,mn} = V_{i,t,m}V_{i,t,n}\left[G_{i,t,mn}\cos\left(\theta_{i,t,m}-\theta_{i,t,n}\right) + B_{i,t,mn}\sin\left(\theta_{i,t,m}-\theta_{i,t,n}\right)\right] \tag{44}$$

$$Q_{i,t,mn} = V_{i,t,m}V_{i,t,n}\left[G_{i,t,mn}\sin\left(\theta_{i,t,m}-\theta_{i,t,n}\right) - B_{i,t,mn}\cos\left(\theta_{i,t,m}-\theta_{i,t,n}\right)\right] \tag{45}$$

where $G_{i,t,mn}$ and $B_{i,t,mn}$ are the real and imaginary parts of the admittance matrix $Y_{i,t,mn} = G_{i,t,mn} + jB_{i,t,mn}$. Therefore, the net injection at each node is equal to the power that leaves each node and expressed as:

$$P_{i,t,m} = V^2_{i,t,m}G_{i,t,mm} + \sum_{m=1,m\neq n} P_{i,t,mn} \tag{46}$$

$$Q_{i,t,m} = -V^2_{i,t,m}B_{i,t,mm} + \sum_{m=1,m\neq n} Q_{i,t,mn} \tag{47}$$

where the first term in (46)-(47) gives the contribution from the nodal shunt element and $f\left(P_{i,t,m}, Q_{i,t,m}, V_{i,t,m}, \theta_{i,t,m}\right) = 0$ for nodal power balance equations (40)-(49).

TABLE 2. Different islands resources capacities.

Type	SI	SLI	LIN
Solar	230 kW	180 kW	0
Wind	200 kW	150 kW	0
Load	15 kW	50 kW	70 kW
Battery/No.	100 kWh/10	100 kWh/10	100 kWh/10

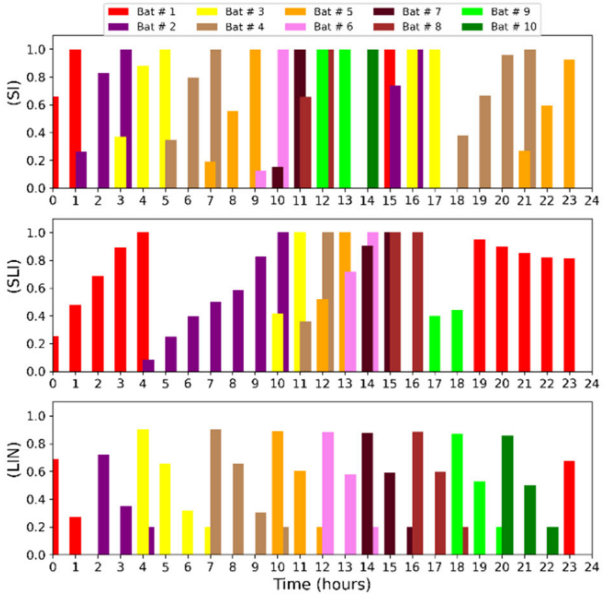


FIGURE 5. Available batteries energy level estimation for energy sharing at different islands.

2) IA SECURITY CONSTRAINTS

The voltage magnitude and branch power flow should flow the security constraints (48)-(49) for the economic operation with the branch line capacity $S_{i,t,\Psi}$.

$$\sqrt{P^2_{i,t,mn} + Q^2_{i,t,mn}} \leq S_{i,t,\psi} \tag{48}$$

$$V^{\min}_{i,t,m} \leq V_{i,t,m} \leq V^{\max}_{i,t,m} \tag{49}$$

3) IA AGENT PARAMETERS

Obtaining data for renewable energy and load profile, which fluctuates dependent on variables such as climate, latitude, and load demand variation such as critical and non-critical loads, is crucial to ensuring proper energy management. Table 2 shows the renewable energy capacity for several islands and data sets obtained from the National Renewable Energy Laboratory (NREL) [27].

4) AVAILABLE BATTERIES ESTIMATION THROUGH IA

The available battery estimation is utilized for energy sharing among different islands. The source islands depict the available charge battery for energy sharing among source islands to load islands from SI →LIN or SLI →LIN. The No. of batteries at SI is available to share at 13 hrs. Moreover, start charging swapped batteries again at 15 hrs.

Available battery estimation is given in Figure 5. From the SOC bar graph, the battery replaces with the next discharge battery until the 10th battery gets charged. In addition, the LIN discharges the battery at 19th hrs., and by adjusting the load profile and considering the shipping time, the swapped batteries from the ship start discharging at 23rd hrs. In the case of SLI, it does not have sufficient batteries because of available local critical load demand and cannot participate in energy sharing. It will raise the flag of energy sharing once the full-charge batteries are available at SLI/SI.

V. REINFORCEMENT LEARNING

In a multi-agent system (MAS), an individual agent acts according to the distributed policy in a typical environment through the central RL agent. Meanwhile, it does not keep the fixed or deterministic approach over time due to its stochastic behavior. It changes its policy over time to maximize the expected reward through the Markov Decision Process (MDP).

A. RL-BASED ENERGY SHARING

In the pelagic island structure, the individual agent will depend on its activities and learn from neighboring island agents' actions to sharpen its policy with each time step. Meanwhile, it does not keep the fixed or deterministic approach over time due to its stochastic behavior.

1) DEEP REINFORCEMENT LEARNING

In this work, (PINMGs) is the environment of distributed IA agents and centralized RL agents to maintain energy balance in each island for the microgrids' network by RES and DGs and provide observation system states to each agent. Based on the PINMGs, this network has a Markov property and represents a tuple $(\mathcal{A}, \mathcal{S}, \mathcal{R}, \gamma)$. For the discrete-time steps, the agent chooses an action space $a_t^i \in \mathcal{A} = \{1, \ldots |\mathcal{A}|\}$ from the state space $s_t^i \in S$ and observes the reward $\mathcal{R}$ with the γ [0, 1] discount factor to trade-off the immediate and future reward $\mathcal{R} = \sum_{t=\tau}^{\infty} R_\tau^i$. Hence, the agent behaves based on the policy π, and the state-action pair (50) and value function (51) are demonstrated as:

$$\mathcal{Q}^{\pi}(\mathcal{S}, \mathcal{A}) = \mathbb{E}\left[\mathcal{R}|\mathrm{s}_{\mathrm{t}}^{\mathrm{i}} = \mathcal{S}, \mathrm{a}_{\mathrm{t}}^{\mathrm{i}} = \mathcal{A}, \pi\right] \tag{50}$$

$$\mathrm{V}^{\pi}(\mathcal{S}) = \mathbb{E}_{\mathcal{A}\sim\pi}\left[\mathcal{Q}^{\pi}(\mathcal{S}, \mathcal{A})\right] \tag{51}$$

Similarly, the advantage function related to value and $\mathcal{Q}$-function to measure the importance of each action (52):

$$\mathrm{A}^{\pi}(\mathcal{S}, \mathcal{A}) = \mathcal{Q}^{\pi}(\mathcal{S}, \mathcal{A}) - \mathrm{V}^{\pi}(\mathcal{S}) \tag{52}$$

2) GRID-BASED PATH SCHEDULING

In the grid-learning environment $E(M, BS, CS)$, the reward depicts the model working efficiently in a map M. In our path scheduling, it focuses on reaching the destination within a minimum time, considered as an obstacle to receiving a negative reward $-re_{i,t,path}$ for wrong decisions and positive $+re_{i,t,path}$ on the right decision with a feasible sequence of actions CS to/from BS as policy $(BS \rightleftharpoons CS)$ [41]. Therefore, the sparse reward for the path schedule is expressed as (53):

$$re_{i,t,path} = \begin{cases} +re_{i,t,path}, & if\ loc = (BS \rightleftharpoons CS) \\ -re_{i,t,path}, & if\ loc \neq (BS \rightleftharpoons CS) \\ 0 & Otherwise \end{cases} \tag{53}$$

In addition, the proposed approach follows the ϵ-greedy policy to balance the exploration.

3) STATE-SPACE

It is a set of all the possible states in an environment and shows the Markovian state for 24 hrs. Furthermore, energy sharing depends on available batteries at CS, BS, ship, and ship locations. From the environment, the current possible states $s_{i,t}$ can be expressed as (54):

$$s_{i,t} = \left[N_{i,t,s}^{CS}, N_{i,t,s}^{BS}, N_{i,t,s}^{Ship}, loc\right] \tag{54}$$

where $N_{i,t,s}^{CS} = N_{i,t,s}^{CS,C} + N_{i,t,s}^{CS,D}$, $N_{i,t,s}^{BS} = N_{i,t,s}^{BS,C} + N_{i,t,s}^{BS,D}$, and $N_{i,t,s}^{Ship} = N_{i,t,s}^{C} + N_{i,t,s}^{D}$, respectively. It is worth noting that the decision variables are related to the batteries being delivered or charged at the charging station. On the other hand, the ship's location is important in deciding the decision of battery swapping at the CS/BS.

4) ACTION-SPACE

Action space $a_{i,t} \in \mathcal{A}$ is the set of all the possible actions (55) that an agent takes to achieve the required states $s_{i,t}$ with the probability transition from one state to another, and it has the dimensions $|\mathcal{S}| \times |\mathcal{S}| \times |\mathcal{A}|$.

$$a_{i,t} = \left[N_{i,t,s}^{CS}, N_{i,t,s}^{Ship}, N_{i,t,s}^{BS}, loc, u_{i,t}^{\mu T}, u_{i,t}^{DGE}\right] \tag{55}$$

As a result, the intended actions are proportional to the number of batteries available at the CS, BS, and ship. Furthermore, the position of the ship influenced the choice of the availability of batteries at the stations. Based on availability, energy sharing controls are actions such as switching MT/DGE or critical and non-critical load management to meet demand at time t.

5) REWARD FUNCTION

The (56) reward function $R_{i,t} \in \mathcal{R}$ is incurred based on the constraint violation to do the excellent action and maximize the profit over the period. $\Delta\rho_{i,t,m}$ constitute storage limit violation based on energy demand (57), not fully exchanging the discharge batteries through the ship, and load curtailment in the resource-rich island $\rho_{i,t,m}^{l} = (1 - \zeta^{N})P_{i,t,m}^{l}$ with a low importance factor.

$$R_{i,t} = \max \sum_{i} \sum_{t} \left[BS_{i,t,BS}^{pro} + re_{i,t,path} - \left[\frac{\Delta\rho_{i,t,m} + \kappa}{\left(P_{i,t,m}^{\mu T} + P_{i,t,m}^{DGE}\right)}\right]\right] \tag{56}$$

$\Delta\rho_{i,t,m}$

$$= \begin{bmatrix} \varpi^{r}\left[r_{i,t,m}^{\max} - N_{i,t,s}^{CS} \times EL_{i,t,m}^{b}\right] + \\ \varpi^{SH}\left[r_{i,t,m}^{\max} - N_{i,t,s}^{Ship} \times EL_{i,t,m}^{b}\right] + \varpi^{l}\rho_{i,t,m}^{l} \end{bmatrix} \tag{57}$$

where ϖ^{r}, ϖ^{SH}, and ϖ^{l} are the cost coefficient updated through the neural network for storage limit, shipping violation, and load curtailment, and κ is the penalty of renewable energy absence supported by DEG/MT.

Furthermore, the physical meaning of the reward is to maximize the profit share through the battery exchange; the second part is also given a positive reward for successfully following the shortest path to the destination; and the third part is associated with storage limit violation and the DEG/MT running time to meet demand by giving a negative reward along with the load curtailment. As a result, since the DEG/MT and load curtailment provide a negative reward, the suggested reward function attempts to maximize the reward by depending less on them.

6) ATTACK MODEL

To show the vulnerability of the proposed work, we verify the effectiveness of the attack model through access to the RL agent action (58) stream $R_{at,a}(\Lambda_t)$, and (59) state-space $R_{at,s}(\Lambda_t)$. The attacker can directly perturb the nominal agent action or state space, and this attack aims to perturb and minimize the long-term discounted reward. Therefore, the attacker greedily designed the perturbation without considering the future concern [42]. Hence, the attacker created the perturbation Λ_t to minimize the future reward at each time, step t:

$$\min_{\Lambda_t} R_{at,s}(\Lambda_t) = R_{i,t}\left(s_{i,t}, \Lambda_t + a_{i,t}\right) + \sum_{k=t+1}^{T} R_{i,t}\left(s_{i,k}, a_{i,k}\right) \tag{58}$$

$$\min_{\Lambda_t} R_{at,a}(\Lambda_t) = R_{t}^{i}\left(s_{i,t} + \Lambda_t, a_{i,t}\right) + \sum_{k=t+1}^{T} R_{i,t}\left(s_{i,k}, a_{i,k}\right) \tag{59}$$

The reward profoundly depends on the evolution of the state trajectory. It is considered a static attack and strictly myopic [44].

B. DUELING DEEP Q-LEARNING (DQN)

Since the value and $\mathcal{Q}$-function are given in (50)-(52). Therefore, the deep $\mathcal{Q}$-network: $\mathcal{Q}(s, a, \theta)$ estimates the network through the loss function (60) with the target network y_j^{DQN} parameter θ_{j-1}.

$$\begin{aligned} &L_j(\theta_j) \\ &\quad = \mathbb{E}_{s_{i,t},a_{i,t}}\left[\left(y_j^{DQN} - Q(s_{i,t}, a_{i,t};\theta_j)\right)^2\right] \\ &y_j^{DQN} \\ &\quad = \mathbb{E}_{(s_{i,t})'}\left[R_t^i + \gamma \max_{(a^k)'}\left(Q_j\left((s_{i,t})', (a_{i,t})'\right)'; \theta_{j-1}\right) \mid s_{i,t}, a_{i,t}\right] \end{aligned} \tag{60}$$

Furthermore, the network updates the parameters online by freezing the parameter of the target network by gradient descent (61).

$$\begin{aligned} &\nabla_{\theta_j} L_j(\theta_j) \\ &\quad = \mathbb{E}_{s_{i,t},a_{i,t}}\left[\left(y_j^{DQN} - Q\left(s_{i,t}, a_{i,t};\theta_j\right)\right)\nabla_{\theta_i} Q(s_{i,t}, a_{i,t};\theta_j)\right] \end{aligned} \tag{61}$$

DQN uses the deep neural network to approximate the action-value function. It takes input from the state of the environment and gives the Q-value from the output for all the possible actions to be taken. For optimizing the possible actions, it uses an experience replay and target network to stabilize the in-depth learning process [45]. In addition, the improved Double DQN (DDQN) [46] uses the y_j^{DDQN} because y_j^{DQN} use the max operator to select and evaluate an action, and y_j^{DDQN} is defined as (62):

$$\begin{aligned} &y_j^{DDQN} \\ &\quad = \mathbb{E}_{(s_{i,t})'}\left[R_t^i + \left(\gamma \operatorname*{argmax}_{(a_{i,t})'} Q_i\left((s_{i,t})', (a_{i,t})'\right)' \theta_j; \theta_{j-1}\right)\right] \end{aligned} \tag{62}$$

For the dueling DQN network, the stream of one fully connected layer output a scalar $V\left(s_{i,t}, \theta, \beta\right)$, other stream output an $|A|$-dimensional vector $A\left(s_{i,t}, a_{i,t}, \theta, \alpha\right)$ connects to make a stream with θ parameters of convolutional layers, and α, β are parameters of connected layers [47]. Its aggregated module is expressed as (63):

$$\mathcal{Q}(\mathcal{S}, \mathcal{A}, \theta, \alpha, \beta) = V(\mathcal{S}, \theta, \beta) + A(\mathcal{S}, \mathcal{A}, \theta, \alpha) \tag{63}$$

(58) cannot use directly because of notability and force the advantage function to have zero advantage at chosen action.

$$\begin{aligned} \mathcal{Q}(\mathcal{S}, \mathcal{A}, \theta, \alpha, \beta) &= V\left(s_{i,t}, \theta, \beta\right) + A\left(s_{i,t}, a_{i,t}, \theta, \alpha\right) \\ &\quad - \max_{a_{i,t} \in |A|} A\left(s_{i,t}, a_{i,t}{'}, \theta, \alpha\right) \end{aligned} \tag{64}$$

Therefore, the original V and A lose their originality because they are constantly off-target. Still, it also increases the optimization problem's stability, and the advantage only needs to change the mean instead of the optimal action advantage (64). Hence, it provides the estimation for the value function stream, and the alternative module is replaced with the max operator as follows (65):

$$\begin{aligned} &\mathcal{Q}(\mathcal{S}, \mathcal{A}, \theta, \alpha, \beta) \\ &\quad = V\left(s_{i,t}, \theta, \beta\right) + \left(\begin{matrix} A\left(s_{i,t}, a_{i,t}, \theta, \alpha\right) \\ -\frac{1}{|A|}\sum_{a_{i,t}} A_{\left(s_{i,t}, a_{i,t}, \theta, \alpha\right)} \end{matrix}\right) \end{aligned} \tag{65}$$

Dueling DQN is also an extension of DQN, which separates the action values into two different estimates, one estimation used for the state-dependent value function and the other used for the action-dependent advantage function.

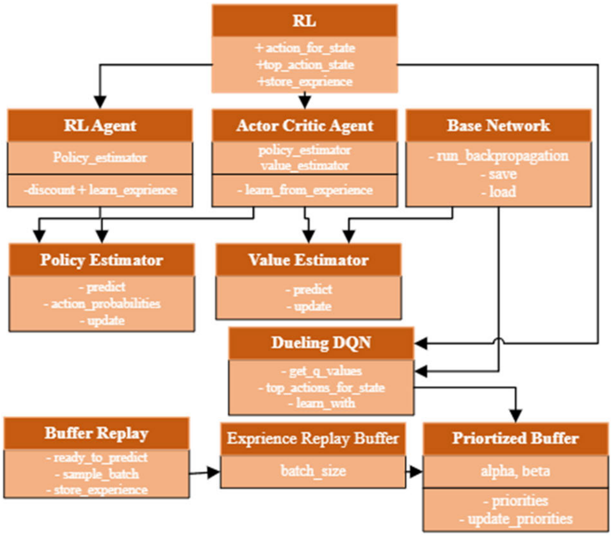


FIGURE 6. Flowchart of proposed Dueling DQN algorithm implementation.

It helps to effectively learn the behaviors of different actions and effectiveness in a different state of the environment [47]. While the TD error is the critical mistake, the reinforce and Actor loss functions, as expected, represent the implementation of the Policy Gradient Theorem and are given as:

$$\nabla_{\theta_i} J(\theta_i) = E_\pi [G_{i,t} \nabla_{\theta_i} \ln \pi_{\theta_i}(a_{i,t}|s_{i,t})] \tag{66}$$

VI. ALGORITHM IMPLEMENTATION

An RL-based algorithm schedules the ship to transport the batteries from resource-rich to load-rich islands in the proposed framework. Simultaneously, the energy demand from the OPF is determined to calculate the available or deficient energy to sustain energy balance. To best execute the energy sharing framework, the CS and BS stations manage to adapt demand depending on demand availability.

The proposed approach's environment, built on several islands, was created using *Python* programming and the *Pandapower* module. The usual TensorFlow and Keras libraries are used for the RL-based scheduling. As a result, Figure 6 depicts the flow chart of the suggested algorithm for scheduling energy sharing across islands. It details the algorithm flow after obtaining the state space from the individual agent's environment and how a centralized RL-based algorithm is used to optimize the scheduling for the optimized route to satisfy demand at the LIN. Furthermore, the Duelling DQN is an extension of DQN that divides the action values into two estimates, one for the state-dependent value function and the other for the action-dependent advantage function. It aids in efficiently learning the behaviors of various acts and their efficacy in various environments [47].

One of DQN's constraints is that the action space must be intrinsically discrete since the value of each action is evaluated using a neuron. An alternate way is to express the output as a probability distribution of the anticipated return, with each action assessed individually. It also works on discrete action spaces when using *SoftMax* distributions. When modeling this way, we shall maximize the value outputs rather than the cost. Modeling in this manner offers advantages and disadvantages. On the one hand, it is simpler since target networks, *replay buffers*, and exploration are unnecessary. On the other hand, bias may be induced due to the substantial variety in reward signals and correlations among states within the same episode.

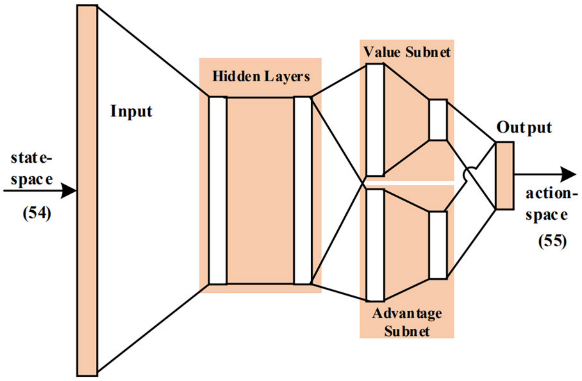


FIGURE 7. Architecture of Dueling DQN algorithm implementation.

An illustration of a Dueling Q-Network's architecture. Keep in mind that value and advantage are modeled differently. Only one neuron is employed to estimate V $(\boldsymbol{s}_{i,t})$, and we require as many neurons to estimate A $(\boldsymbol{s}_{i,t}, \boldsymbol{a}_{i,t})$ and Q $(\mathcal{S}, \mathcal{A}, \boldsymbol{\theta}, \boldsymbol{\alpha}, \boldsymbol{\beta})$ as possible. Hence, the architecture of dueling D-network is given in Figure 7.

A. ASSESSMENT STRATEGY

A comprehensive hyperparameter space search would be inefficient due to the agent's many parameters. According to the protocol, we will test one parameter at a time while freezing the others. For comparison, each combination is tested three times for 50 episodes, and the average of the Return's final moving mean (100-episode window) is chosen. The final model picks the best value calculated for each parameter.

Once we have identified the models' optimum parameters, we compare four distinct seeds and 100 episodes. The moving mean, and 95% confidence interval will be used to compare the agents. For example, a random proposal baseline is presented five times, and the agent is trained for 70K.

B. ALGORITHM FOR THE PROPOSED APPROACH

The RL algorithm for the proposed method is described below, and the parameters used in the algorithm are detailed in Table 3. In contrast, Figure 6 depicts the EMS model implementation, which serves as the foundation for the whole flow of the proposed technique for PINMGs. The organized agent, a Duelling DQN implementation with buffers, an RL implementation with discounted episode rewards as a baseline, and an actor-critic implementation with a value-estimator as the critic will be used to assess the performance of our RL agents.

TABLE 3. Selected parameters for algorithms.

Description	Notion	Value
Episodes	-	70,000
Epsilon decay 1 (50K high exploration)	ϵ	0.000005
Epsilon decay 2 (20K high exploitation)	ϵ	0.0000372
Batch size	-	32
Discount factor	γ	0.98
Learning rate	α	1e-4
Optimizer	-	MSE
Dense layers	-	4
No. of Neuron		32
Hidden layer activation	-	Relu
Output layer activation	-	Linear
Target network update rate	-	10,000
Average reward	-	100 episodes
Loss Monitoring	-	50 episodes
Reward monitoring	-	50 episodes

Algorithm 1 Dueling DQN With Experience Buffer for AI Agent

```
Start with $Q(s_{i,t}, a)$ for all $s_{i,t}, a_{i,t}$.
The state and action spaces follow the (54)-(55).
Get an initial state $s_{i,t}$ and main network with the weight $\theta_i$.
Choose the buffer size, batch size, $\alpha$, $\beta$.
A sequence of main network training and target network
  update was chosen.
Count the steps for iteration as $j$.
Choose the learning rate $\eta$.
Initialize action-value Q with random and target value $\theta_i$.
for every time instant $t$, $s_{i,t}$, $\gamma\epsilon(0, 1)$
    select random action $a_t^i$ by $\epsilon$ probability.
    otherwise, select $a_t^i = argmax_{a_t^i} Q\left(s_{i,t}, a_{i,t}; \theta_i\right)$
    apply $a_t^i$ and calculate the reward.
    $j \leftarrow j+1$
    if network training $= 0$
        store transition $s_{i,t}, a_{i,t}, R_{i,t}, \left(a_{i,t}\right)'$ in memory.
        Sample random mini batch from the reply memory.
        Training and target value construct using (58)-(60).
        perform the gradient descent step (59).
    end
    if target network updated $= 0$
        update the $\theta_{i-1} \rightarrow \theta_i$ at every 10,000 steps.
    end
    store $(s_{i,t}, s_{i,t+1}, a_{i,t}, R_{i,t})$
end
```

VII. FLOW CHART

Figure 8 depicts the flowchart for the proposed PINMGs problem, representing the proposed work's overall architecture. The suggested model expresses the individual component using the following equation number. The one-island model represents the EMS level in detail, while SLI lacks a distribution network for the SI. It follows the flow from the energy-resources→IMGO→IA agent and the dispatch signal sent to the storage cluster through the direct command, and its number ranges from 1 to i. On the other hand, the energy-sharing system is linked with island nature, whether a resource or a load island. It then talks with the RL agent based on the nature.

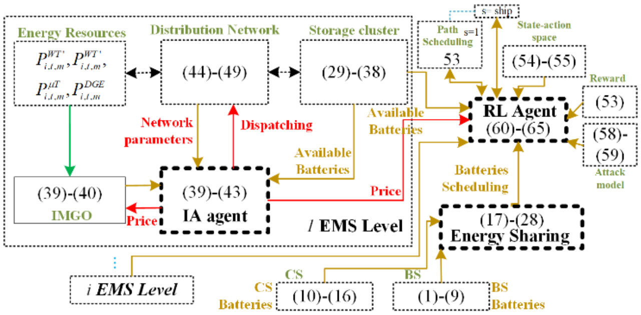


FIGURE 8. Flowchart for the proposed model of PINMGs.

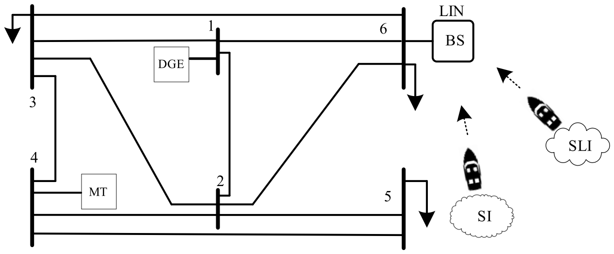


FIGURE 9. Relative position with pelagic island and LIN structure.

VIII. CASE STUDY

In this work, pelagic island network microgrids (PINMGs) are the environment of different agents to maintain energy. Each IA is equipped with standby DEG with a ramp rate of 80kW/h (the peak value of the non-interrupt load is 60kW) with a DEG average fuel cost is 0.6 \$/liter. In addition, the MT with a fixed capacity of 10kW to meet the load demand charge the available batteries to meet the demand. It has a 100kWh capacity, ten batteries, and 100 \$/kWh installment investment with battery DoD cycles [43] and its $Ef = 89$ as a round trip efficiency, and $Life = 1344$ battery lifetime throughput (kWh). In this case, the lead-acid battery has a charge efficiency of 95% or 0.95 [48].

Therefore, the equidistance islands are considered for the case study as 80km with a speed of 20km/h. The ship's sailing time is 4hrs with a 30\$ per battery sailing cost. In addition, the operation cost mainly affects power production through renewable energy and load demand at the load islands. Likewise, energy charging is considered a cost of 0.05\$/kWh for resource-rich islands. The sodium-sulfur batteries are used on the island with a total capacity of 1000kWh. The LIN network structure is given in Figure 9 with a battery-swapping station at bus 6; DEG and MT are connected at buses 1 and 4, respectively.

Hence, the critical loads are connected at bus 6, and the non-critical at buses 3 and 5. The swapping batteries are allowed to work within the 0.2-0.9 SOC with a maximum depth of discharge of 80%. Initially, the batteries are in the LIN, and the ship moves towards the resource-rich islands

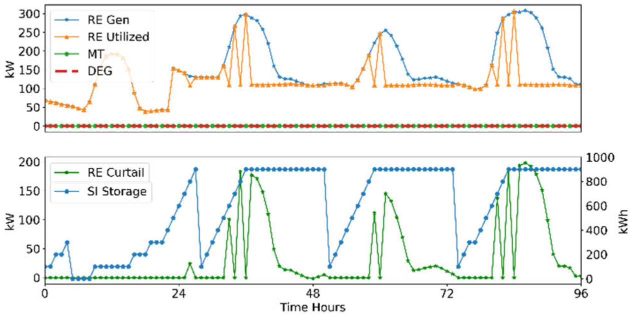
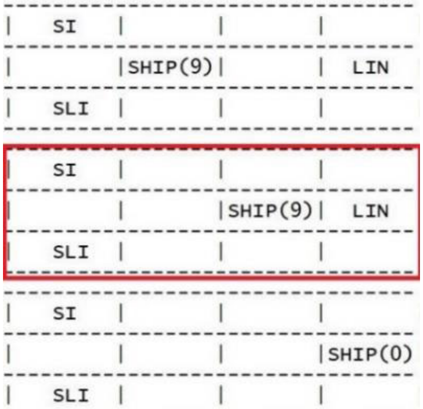


FIGURE 10. Maximizing the usage of RE resources with RE curtailment.



```
-----------------------------------------
|   SI    |         |         |
-----------------------------------------
|         |SHIP(9)|           |   LIN
-----------------------------------------
|   SLI   |         |         |
-----------------------------------------

-----------------------------------------
|   SI    |         |         |
-----------------------------------------
|         |         |SHIP(9)|   LIN
-----------------------------------------
|   SLI   |         |         |
-----------------------------------------

-----------------------------------------
|   SI    |         |         |
-----------------------------------------
|         |         |         |SHIP(0)
-----------------------------------------
|   SLI   |         |         |
```

FIGURE 11. RL grid-base ship movement from LIN to/from SI/SLI.

SI or SLI to utilize the RE generation to meet the demand at LIN and SLI. It is assumed that the ship sailed towards the source-enriched islands for battery swapping and collaborated to meet the demand at LIN. Therefore, the net power flow and voltage on each bus can be depicted through the OPF and provide details about the demand and generation on each island to cooperate with neighboring islands by storage cluster.

The number varies from 1 to ship for the ship, and each ship schedules dispatch from the centralized RL agent. Furthermore, the viability of the RL agent is also verified through the attack model. The impact of the proposed model is discussed in the case study.

The simulation was conducted on an RYZEN 7 (5000 series) with 16 GB RAM and an RTX 3070 GPU, using Python 3.7 with Spyder IDE. Furthermore, the proposed model's validity is justified by the suggested cases, which include renewable energy curtailments of 20% and 50%, to test the proposed algorithm's robustness and recovery to standard operation and energy management by maximizing renewable energy resources. Furthermore, as seen in Figure 10, SI realizes less on non-RE resources while maximizing RE use. Meanwhile, the RE is reduced when the charge batteries at SI are at their maximum. Due to the scarcity of discharge batteries, it is impossible to charge them at CS.

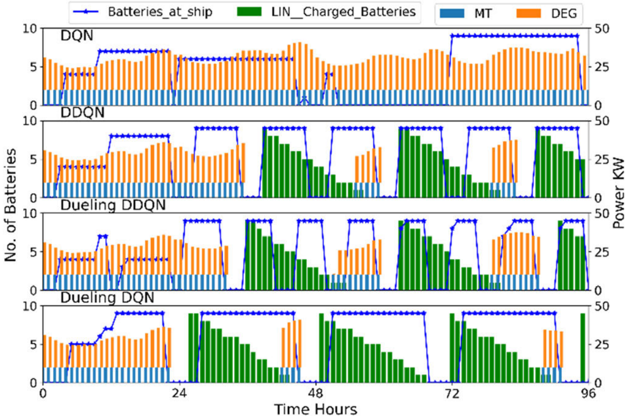


FIGURE 12. Available battery scheduling from SI, SLI→LIN.

A. GRID-BASED SHIP TRAVELING

It can be seen from Figure 11 that the ship is at the (2 × 3) location with a total of 0 to 9 charge batteries on the ship $N_{i,t}^{Ship}$. Moreover, moving towards the LIN by receiving a positive reward. In the proceeding flow, it can be depicted that the ship reached LIN, and charged batteries were swapped at LIN. The swapping batteries $N_{i,t,s}^{BS}$ is exchanged based on the greedy on-policy in the grid-based environment. It can be seen from Figure 11 that the ship is at the (2 × 3) location with a total of 0 to 9 charge batteries on the ship $N_{i,t}^{Ship}$. Furthermore, moving towards the LIN by receiving a positive reward. In the proceeding flow, it can be depicted that the ship reached LIN, and charged batteries were swapped at LIN. The swapping batteries $N_{i,t,s}^{BS}$ is exchanged based on the greedy on-policy in the grid-based environment (58)-(63). Furthermore, the appropriate actions are taken through one week of RL agent training.

B. ANALYSIS OF SCHEDULING RESULTS AT LIN

The effectiveness of the proposed work is verified through the dueling DQN, and its validity in meeting the goal is depicted through the different algorithms such as DQN, DDQN, dueling DDQN, and Dueling DQN are shown in Figure 12. It can be seen that batteries swapped through the resource-rich islands to LIN are gradually improved by reducing non-RE usage from DQN to dueling DQN. The in-depth modeling of the proposed RL methods is discussed in section V, along with their benefits and advantages over the other methods.

As battery swapping depends on the available charge batteries at CS, it is important to note that during the first 24 hours, the LIN starts to rely on non-renewable energy resources. After the sailing period, batteries are available at LIN. Moreover, the sailing time and battery swapping time are considered together. It is important to note that the importance factor prioritizes energy trading through swapping as critical and non-critical loads, and all end-user's nominal demands can be satisfied. It can be compared from Figure 12 that energy demand is to be fulfilled at LIN, and it receives

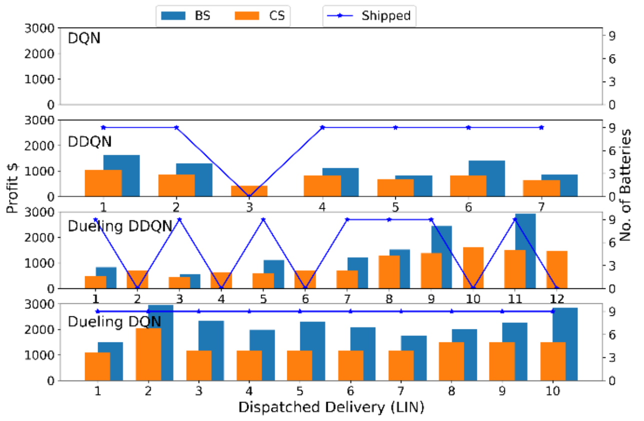


FIGURE 13. Total profit incurred over the successful dispatch delivery (LIN).

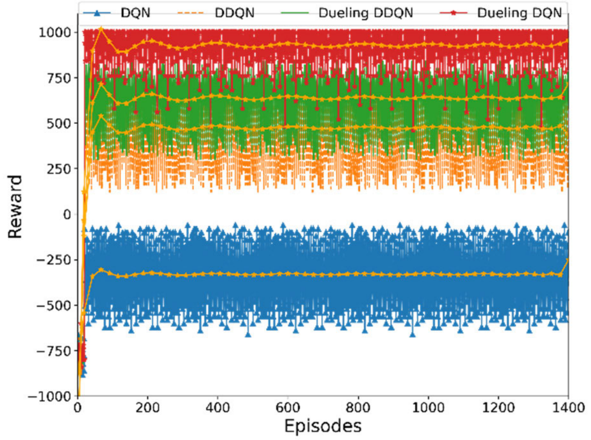


FIGURE 14. Average reward profile for the proposed RL algorithm.

the batteries only once the swapping batteries at LIN are ready to be swapped. Otherwise, it kept waiting until the discharge batteries were ready to be swapped.

Similarly, the comparison with the other methods shows that after the first 24 hrs., the ship keeps sailing and then tries to reach an island for the battery dispatched but fails again to start sailing in the DQN method without considering the available charge batteries. Nevertheless, for the DDQN, the ship started sailing and then dispatched again without considering the charged batteries and whether available batteries were charged completely. However, after 48 hrs., it learns to take the battery and sail towards the load islands but fails because of uncharged batteries' availability [49].

On the other hand, the dueling DDQN-based method fails to start the sailing after getting batteries to charge and dispatch batteries before the charge battery at the LIN. It can be seen that the ship keeps sailing and turns back to the island instead of waiting for the batteries to discharge fully. However, the pattern using the proposed method dueling DQN can be seen that the ship reaches the island once the batteries at the LIN discharge completely, but while battery swapping to meet some critical loads, it can be seen that MT and DEG run to support it.

The ship swapping poses the time-delayed with discrete characteristics. Every round trip is between a resource-rich island and LIN, with a maximum of 10 batteries to be swapped. The energy flow at LIN has a real-time power balance in the islanded microgrid by scheduling and meeting the actual power demand.

In addition, the DQN, DDQN, and dueling DDQN have made many wrong decisions at the BS without considering the discharged batteries at LIN and utilized many non-renewable energy resources. For the DQN-based algorithm, the ship waited long at the LIN but could not dispatch batteries for 96 hrs. and failed to earn a profit because of the penalty received through the RL agent. To meet the demand at LIN, the predicted demand at LIN is 10 times a week to swap at BS.

From Figure 13, the profit incurred over the successful dispatch has been shown by comparing different proposed methods, and its viability can be seen from it. Because the amount is only transferred to the owner once the complete batteries swap with the charge batteries at the load island or the required number of batteries, swap at the BS. By comparing, it can be seen that DQN fails to provide any successful delivery and cannot ensure profit. Similarly, the DDQN and dueling DDQN try to act precisely but fail to provide constant support by dispatching batteries at the CS. At the same time, the proposed dueling DQN method outperformed all the other RL-based methods by showing the most successful deliveries and incurring profit.

Figure 14 shows that the ship has started improving the scheduling through the DDQN to duel the DQN RL algorithm and earn profit based on the energy collaboration between SI and SLI at BS. The training process in the deep learning model is considered at every time step by following the greedy policy with the probability of $1 - \epsilon$ and random action chosen on ϵ decreasing value. Furthermore, the average reward profile comparison is compared, and the DQN-based method fails to maximize constant reward. From the profile comparison, the dueling DQN based outperforms the other methods and accumulates the constant and maximum average reward compared with other algorithms.

It is very important to note that a penalty is incurred for every deviation to maintain the energy balance during decision-making. In addition, the improving discount factor stabilizes and advances the reward function for every discrete action. Hence, the reward profile with decreasing ϵ is given in Figure 15. In addition, the decreasing learning rate through the given epochs with the rising discount factor provides decreasing training loss, as shown in Figure 15. Moreover, the increasing discount factor improves further reward function through 70,000 iterations. Furthermore, the deep RL-based methods analyze and compare the naive policy and parameters used for the DNN-based algorithms listed in Table 3.

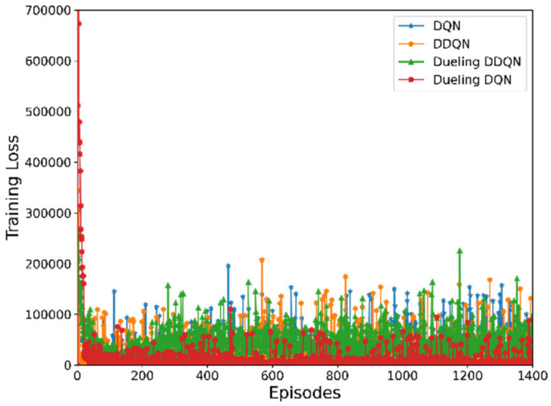


FIGURE 15. Training loss of PINMGs under Q-learning-based algorithms.

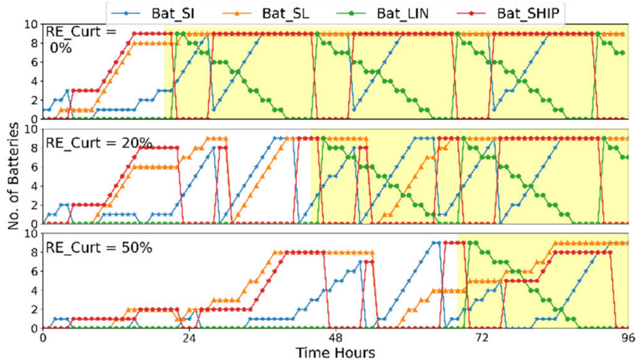


FIGURE 16. The cooperative energy comparison during energy curtailment.

C. EFFECT OF RENEWABLE ENERGY CURTAILMENT

It is a common phenomenon of curtailment with the rising penetration of renewable energy to meet the power balance issues, as PV production is higher during the afternoon. Wind produces more power at night and pushes DEG and MT to meet the energy demand during peak hours to meet the downward reserves unable to provide renewable [30].

Therefore, the cooperative energy management between islands is justified through renewable energy curtailment to verify the effectiveness of the proposed approach. The comparison has been drawn based on the 0%, 20%, and 50% curtailment. With 0% curtailment, the battery was swapped on the LIN after the first 24 hours. While curtailment reaches 20%, it takes 24 hours to readjust the resources, collaborate between resource-rich islands, and provide downward reserves.

Our proposed algorithm enables islands and resource collaboration while minimizing the usage of non-renewable resources [31]. In addition, with 50% renewable energy curtailment, our RL agent spends more time learning to reduce the use of DEG and MT and maximize the renewable usage to meet the demand at LIN. It can be seen from the resource adjustments in Figure 16. On the other hand, with the curtailment of 0%, the agent learns very fast and starts to adjust their resources in the first 24 hrs. Likewise, for the curtailment of 20%, the agent learns in the first two days and starts performing the best possible ways to adjust the resources to meet the demand by scheduling between battery swapping stations and battery charging stations [50].

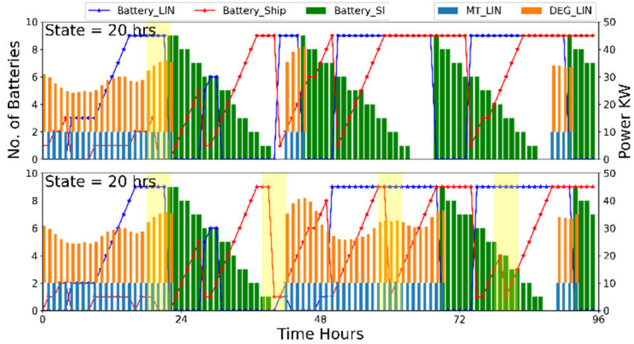


FIGURE 17. State attacks with single/multiple time step every 20 hrs.

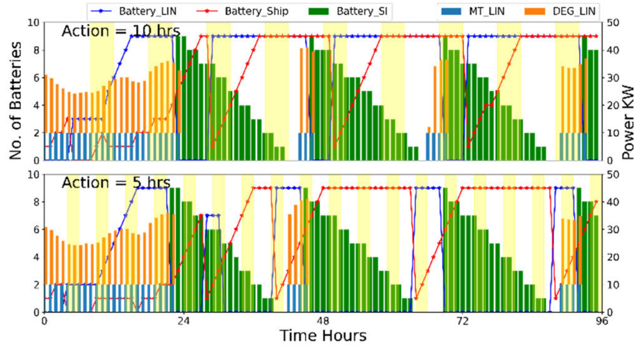


FIGURE 18. Multiple actions attack every 10 and 20 hrs.

D. DISCUSSION ON THE ATTACK MODEL

The robustness of the RL agent is verified through the series of attacks over the action and state space to degrade the performance of the trained RL agent (58)-(59). Therefore, the perturbation in the state attack is applied at the time step 20th hrs. for the charged batteries at the resource-rich island and trained RL agents to act under the nominal conditions through the gradient computation.

It can be observed from Figure 17 that trained RL agents mitigate the single attack and allow the ship to dispatch the charge batteries at the LIN without relying much on non-RE. In addition, multiple state attacks were applied at every 20th-hour time step to verify the RL viability further. Under the worst attack situation, the RL agent relies upon the non-RE but manages the cooperation between SI and SLI to dispatch batteries at LIN to meet the load demand and minimize the non-RE [32]. At first, the attack did not impact the ship's movement much. During the second attack, it had much impact and again started using the non-RE resources. After learning for a day, the agent returned to normal and started working normally.

Hence, the further validity of the proposed approach is verified through the action attack by perturbing the final decision of battery dispatches from source islands to LIN in Figure 17. The action attack applied to the learned RL agent and the $s_{i,t}$ and $a_{i,t}$ are not independent during the transition but depend on the evolutions of state trajectories. It can be observed from Figure 18 that the single-action attack is more stable because the agent learns the correct actions very quickly.

Therefore, multiple attacks at every 5th and 10th hour are applied to perturb the battery dispatch decisions. The ship followed the wrong dispatch decision with action perturbation, but after a few steps, the ship learned the appropriate policy to dispatch batteries at the right time. Therefore, the viability of the proposed algorithm is justified to manage the energy resources and ship to LIN to maximize the RE usage and minimize the non-RE at LIN. The RL algorithm was trained for a week and performed well under state and action perturbations attacks during standard operating conditions.

IX. CONCLUSION

This study presents a two-stage energy management strategy for PINMGs based on ship swapping across resource-rich islands to load islands (to fulfill demand) with the help of a centralized RL agent and IA agents. As a result, by optimizing the usage of RE and overcoming RE intermittency using RL-based scheduling and energy storage during RE availability, the simulation maximizes profit through energy sharing across various islands. This technique is effective compared to other typical RL algorithms for day-ahead scheduling. Consequently, the simulation results demonstrate the practicality and efficiency of the proposed work. Finally, the robustness of the trained RL agent is validated by perturbing the single/multiple states and actions to undertake energy trading across distinct pelagic islands.

Furthermore, as a future study, this research gives a beginning motivation for the PINMGs-based challenge to be extended on a broad scale for the island's electrification.

REFERENCES

[1] Z. Wang, B. Chen, J. Wang, M. M. Begovic, and C. Chen, ''Coordinated energy management of networked microgrids in distribution systems,'' *IEEE Trans. Smart Grid*, vol. 6, no. 1, pp. 45–53, Jan. 2015.

[2] A. Suleman, M. A. Amin, M. Fatima, B. Asad, M. Menghwar, and M. A. Hashmi, ''Smart scheduling of EVs through intelligent home energy management using deep reinforcement learning,'' in *Proc. 17th Int. Conf. Emerg. Technol. (ICET)*, Swabi, Pakistan, Nov. 2022, pp. 18–24.

[3] H. Li, Z. Wan, and H. He, ''Constrained EV charging scheduling based on safe deep reinforcement learning,'' *IEEE Trans. Smart Grid*, vol. 11, no. 3, pp. 2427–2439, May 2020.

[4] D. Michaelson, H. Mahmood, and J. Jiang, ''A predictive energy management system using pre-emptive load shedding for islanded photovoltaic microgrids,'' *IEEE Trans. Ind. Electron.*, vol. 64, no. 7, pp. 5440–5448, Jul. 2017.

[5] Z. Yang. (Mar. 21, 2022). EV battery swapping was left for dead. Now, it's being revived in China. Protocol. [Online]. Available: https://www.protocol.com/climate/electric-vehicle-battery-swap-china

[6] C. Battistelli, Y. P. Agalgaonkar, and B. C. Pal, ''Probabilistic dispatch of remote hybrid microgrids including battery storage and load management,'' *IEEE Trans. Smart Grid*, vol. 8, no. 3, pp. 1305–1317, May 2017.

[7] H. Xie, S. Zheng, and M. Ni, ''Microgrid development in China: A method for renewable energy and energy storage capacity configuration in a megawatt-level isolated microgrid,'' *IEEE Electrific. Mag.*, vol. 5, no. 2, pp. 28–35, Jun. 2017.

[8] S. Yao, P. Wang, and T. Zhao, ''Transportable energy storage for more resilient distribution systems with multiple microgrids,'' *IEEE Trans. Smart Grid*, vol. 10, no. 3, pp. 3331–3341, May 2019.

[9] S. Lei, C. Chen, H. Zhou, and Y. Hou, ''Routing and scheduling of mobile power sources for distribution system resilience enhancement,'' *IEEE Trans. Smart Grid*, vol. 10, no. 5, pp. 5650–5662, Sep. 2019.

[10] J. Arkhangelski, M. Abdou-Tankari, and G. Lefebvre, ''Day-ahead optimal power flow for efficient energy management of urban microgrid,'' *IEEE Trans. Ind. Appl.*, vol. 57, no. 2, pp. 1285–1293, Mar. 2021.

[11] X. Wei, M. A. Amin, Y. Xu, T. Jing, Z. Yi, X. Wang, Y. Xie, D. Li, S. Wang, and Y. Zhai, ''Two-stage cooperative intelligent home energy management system for optimal scheduling,'' *IEEE Trans. Ind. Appl.*, vol. 58, no. 4, pp. 5423–5437, Jul. 2022.

[12] D. Romero-Quete and C. A. Cañizares, ''An affine arithmetic-based energy management system for isolated microgrids,'' *IEEE Trans. Smart Grid*, vol. 10, no. 3, pp. 2989–2998, May 2019.

[13] H. Morais, P. Kádár, P. Faria, Z. A. Vale, and H. M. Khodr, ''Optimal scheduling of a renewable micro-grid in an isolated load area using mixed-integer linear programming,'' *Renew. Energy*, vol. 35, no. 1, pp. 151–156, Jan. 2010.

[14] B. V. Solanki, A. Raghurajan, K. Bhattacharya, and C. A. Cañizares, ''Including smart loads for optimal demand response in integrated energy management systems for isolated microgrids,'' *IEEE Trans. Smart Grid*, vol. 8, no. 4, pp. 1739–1748, Jul. 2017.

[15] M. Hu, Y.-W. Wang, J.-W. Xiao, and X. Lin, ''Multi-energy management with hierarchical distributed multi-scale strategy for pelagic islanded microgrid clusters,'' *Energy*, vol. 185, pp. 910–921, Oct. 2019.

[16] Q. Sui, F. Wei, C. Wu, X. Lin, and Z. Li, ''Day-ahead energy management for pelagic island microgrid groups considering non-integer-hour energy transmission,'' *IEEE Trans. Smart Grid*, vol. 11, no. 6, pp. 5249–5259, Nov. 2020.

[17] M. Dabbaghjamanesh, A. Kavousi-Fard, and Z. Y. Dong, ''A novel distributed cloud-fog based framework for energy management of networked microgrids,'' *IEEE Trans. Power Syst.*, vol. 35, no. 4, pp. 2847–2862, Jul. 2020.

[18] R. Xiong, J. Cao, and Q. Yu, ''Reinforcement learning-based real-time power management for hybrid energy storage system in the plug-in hybrid electric vehicle,'' *Appl. Energy*, vol. 211, pp. 538–548, Feb. 2018.

[19] Z. Li, M. A. Amin, C. Wang, J. Ding, B. Peng, and L. Du, ''Efficient reactive power control using reinforcement learning under inaccurate power network model,'' in *Proc. 6th Int. Conf. Power Renew. Energy (ICPRE)*, Shanghai, China, Sep. 2021, pp. 350–355.

[20] United Nations. (2022). *Goal 7 | Department of Economic and Social Affairs*. [Online]. Available: https://sdgs.un.org/goals/goal7

[21] *Europe's Islands are Leading the Charge in the Clean Energy Transition | Research and Innovation*. Accessed: Mar. 10, 2023. [Online]. Available: https://ec.europa.eu/research-and-innovation/en/horizon-magazine/europes-islands-are-leading-charge-clean-energy-transition

[22] GIZ. *1,000 Islands—Renewable Energy for Electrification Programme*. Accessed: Mar. 10, 2023. [Online]. Available: https://www.giz.de/en/worldwide/63533.html

[23] M. Waseem, I. A. Sajjad, S. S. Haroon, S. Amin, H. Farooq, L. Martirano, and R. Napoli, ''Electrical demand and its flexibility in different energy sectors,'' *Electr. Power Compon. Syst.*, vol. 48, nos. 12–13, pp. 1339–1361, 2020.

[24] T. Rashid, M. Samvelyan, C. Schroeder, G. Farquhar, J. Foerster, and S. Whiteson, ''QMIX: Monotonic value function factorization for deep multi-agent reinforcement learning,'' in *Proc. Int. Conf. Mach. Learn.*, 2018, pp. 4295–4304.

[25] D. Silver, T. Hubert, J. Schrittwieser, I. Antonoglou, M. Lai, A. Guez, M. Lanctot, L. Sifre, D. Kumaran, T. Graepel, T. Lillicrap, K. Simonyan, and D. Hassabis, ''A general reinforcement learning algorithm that masters chess, shogi, and go through self-play,'' *Science*, vol. 362, no. 6419, pp. 1140–1144, Dec. 2018.

[26] R. Zhang, G. Li, S. Bu, S. Aziz, and R. Qureshi, ''Data-driven cooperative trading framework for a risk-constrained wind integrated power system considering market uncertainties,'' *Int. J. Electr. Power Energy Syst.*, vol. 144, Jan. 2023, Art. no. 108566, doi: 10.1016/j.ijepes.2022.108566.

[27] S. Aziz, M. Irshad, S. A. Haider, J. Wu, D. N. Deng, and S. Ahmad, ''Protection of a smart grid with the detection of cyber-malware attacks using efficient and novel machine learning models,'' *Frontiers Energy Res.*, vol. 10, p. 1102, Aug. 2022.

[28] H. Wang, R. Cai, B. Zhou, S. Aziz, B. Qin, N. Voropai, L. Gan, and E. Barakhtenko, ''Solar irradiance forecasting based on direct explainable neural network,'' *Energy Convers. Manage.*, vol. 226, Dec. 2020, Art. no. 113487, doi: 10.1016/j.enconman.2020.113487.

[29] 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.

[30] B. Li, K. Xie, W. Zhong, X. Huang, Y. Wu, and S. Xie, ''Operation management of electric vehicle battery swapping and charging systems: A bilevel optimization approach,'' *IEEE Trans. Intell. Transp. Syst.*, vol. 24, no. 1, pp. 528–540, Jan. 2023, doi: 10.1109/TITS.2022.3211883.

[31] Y. Liang, Z. Ding, T. Zhao, and W.-J. Lee, ''Real-time operation management for battery swapping-charging system via multi-agent deep reinforcement learning,'' *IEEE Trans. Smart Grid*, vol. 14, no. 1, pp. 559–571, Jan. 2023, doi: 10.1109/TSG.2022.3186931.

[32] H. Pandžic and V. Bobanac, ''An accurate charging model of battery energy storage,'' *IEEE Trans. Power Syst.*, vol. 34, no. 2, pp. 1416–1426, Mar. 2019, doi: 10.1109/TPWRS.2018.2876466.

[33] S. Liu, C. Chen, Y. Jiang, Z. Lin, H. Wang, M. Waseem, and F. Wen, ''Bi-level coordinated power system restoration model considering the support of multiple flexible resources,'' *IEEE Trans. Power Syst.*, vol. 38, no. 2, pp. 1583–1595, Mar. 2023, doi: 10.1109/TPWRS.2022.3171201.

[34] S. Lee and D.-H. Choi, ''Federated reinforcement learning for energy management of multiple smart homes with distributed energy resources,'' *IEEE Trans. Ind. Informat.*, vol. 18, no. 1, pp. 488–497, Jan. 2022.

[35] M. Waseem, Z. Lin, S. Liu, I. A. Sajjad, and T. Aziz, ''Optimal GWCSO-based home appliances scheduling for demand response considering end-users comfort,'' *Electr. Power Syst. Res.*, vol. 187, Oct. 2020, Art. no. 106477.

[36] M. A. Amin, A. Suleman, T. Korõtko, S. Aziz, M. U. Naseer, and N. Ahmad, ''Profit maximization by integrating demand response in multiple VPPs optimal scheduling,'' in *Proc. Int. Conf. Electr. Eng. Sustain. Technol. (ICEEST)*, Lahore, Pakistan, Dec. 2022, pp. 1–6.

[37] M. Hu, Y.-W. Wang, X. Lin, and Y. Shi, ''A decentralized periodic energy trading framework for pelagic islanded microgrids,'' *IEEE Trans. Ind. Electron.*, vol. 67, no. 9, pp. 7595–7605, Sep. 2020.

[38] J. Li, M. A. Amin, J. Shi, L. Cheng, F. Lu, B. Geng, A. Liu, and S. Zhou, ''Energy trading of multiple virtual power plants using deep reinforcement learning,'' in *Proc. Int. Conf. Power Syst. Technol. (POWERCON)*, Dec. 2021, pp. 892–897.

[39] C. P. Mediwaththe, E. R. Stephens, D. B. Smith, and A. Mahanti, ''Competitive energy trading framework for demand-side management in neighborhood area networks,'' *IEEE Trans. Smart Grid*, vol. 9, no. 5, pp. 4313–4322, Sep. 2018.

[40] *Solar Integration Data and Tools*. Accessed: Apr. 2022. [Online]. Available: https://www.nrel.gov/grid/solar-integrationdata.html

[41] L. Lv, S. Zhang, D. Ding, and Y. Wang, ''Path planning via an improved DQN-based learning policy,'' *IEEE Access*, vol. 7, pp. 67319–67330, 2019.

[42] W. Infante, J. Ma, X. Han, and A. Liebman, ''Optimal recourse strategy for battery swapping stations considering electric vehicle uncertainty,'' *IEEE Trans. Intell. Transp. Syst.*, vol. 21, no. 4, pp. 1369–1379, Apr. 2020.

[43] K. Thirugnanam, M. S. E. Moursi, V. Khadkikar, H. H. Zeineldin, and M. Al Hosani, ''Energy management of grid interconnected multimicrogrids based on P2P energy exchange: A data driven approach,'' *IEEE Trans. Power Syst.*, vol. 36, no. 2, pp. 1546–1562, Mar. 2021, doi: 10.1109/TPWRS.2020.3025113.

[44] M. Waseem, Z. Lin, Y. Ding, F. Wen, S. Liu, and I. Palu, ''Technologies and practical implementations of air-conditioner based demand response,'' *J. Modern Power Syst. Clean Energy*, vol. 9, no. 6, pp. 1395–1413, Nov. 2021, doi: 10.35833/MPCE.2019.000449.

[45] X. Y. Lee, S. Ghadai, K. L. Tan, C. Hegde, and S. Sarkar, ''Spatiotemporally constrained action space attacks on deep reinforcement learning agents,'' in *Proc. AAAI Conf. Artif. Intell.*, 2020, vol. 34, no. 4, pp. 4577–4584.

[46] Y. Wang, M. Mao, L. Chang, and N. D. Hatziargyriou, ''Intelligent voltage control method in active distribution networks based on averaged weighted double deep Q-network algorithm,'' *J. Mod. Power Syst. Clean Energy*, vol. 11, no. 1, pp. 132–143, Jan. 2023, doi: 10.35833/MPCE.2022.000146.

[47] R. Zhang, K. Xiong, Y. Lu, B. Gao, P. Fan, and K. B. Letaief, ''Joint coordinated beamforming and power splitting ratio optimization in MU-MISO SWIPT-enabled HetNets: A multi-agent DDQN-based approach,'' *IEEE J. Sel. Areas Commun.*, vol. 40, no. 2, pp. 677–693, Feb. 2022, doi: 10.1109/JSAC.2021.3118397.

[48] M. Sewak, ''Deep Q network (DQN), double DQN, and dueling DQN: A step towards general artificial intelligence,'' in *Deep Reinforcement Learning: Frontiers of Artificial Intelligence*. Springer, 2019, pp. 95–108. [Online]. Available: https://link.springer.com/chapter/10.1007/978-981-13-8285-7_8

[49] ScienceDirect. *Practical Handbook of Photovoltaics*. Accessed: Mar. 11, 2023. [Online]. Available: https://www.sciencedirect.com/book/9780123859341/practical-handbook-of-photovoltaics

[50] M. Waseem, Z. Lin, S. Liu, Z. Zhang, T. Aziz, and D. Khan, ''Fuzzy compromised solution-based novel home appliances scheduling and demand response with optimal dispatch of distributed energy resources,'' *Appl. Energy*, vol. 290, May 2021, Art. no. 116761.

M. ASIM AMIN (Graduate Student Member, IEEE) received the M.Eng. degree in electrical engineering from Xi'an Jiaotong University, Xi'an, China, in 2017. He is currently pursuing the Ph.D. degree with the University of Genoa, Italy. He was the Solution Manager and the Product Manager of CYG SUNRI and Growatt New Energy, Shenzhen, China, from 2017 to 2020. He is a Marie Skłodowska Curie Research Fellow and the CTO of Rapid Volt (PVT) Ltd., Pakistan. His research interests include integrated energy management systems (iEMS), transactive energy, the IoT, virtual power plant (VPP), P2P, and demand response.

AHMAD SULEMAN received the M.S. degree from the University of Punjab, Lahore, Pakistan. He is currently the Research and Development Director with Rapid Volt (PVT) Ltd., Pakistan. His research interests include autonomous AI, decision support, energy management systems, demand-response, robotics, reinforcement learning, autonomous AI, and decision support systems.

MUHAMMAD WASEEM (Member, IEEE) received the M.Sc. degree in electrical engineering from the University of Engineering and Technology Taxila, Pakistan, in 2017, and the Ph.D. degree in electrical engineering from Zhejiang University, China, in 2022. He is currently a Postdoctoral Fellow with the Centre for Advances in Reliability and Safety, The Hong Kong Polytechnic University, Hong Kong. His research interests include power system analysis, demand-side management, integrated energy management systems, and smart grids.

TAOSIF IQBAL received the M.S. degree in control systems from the Ivanovo State University of Chemical Technology, in 2003, and the Ph.D. degree in power engineering from Xi'an Jiaotong University, in 2019. He has been an Assistant Professor with the Department of Electrical Engineering, College of E&ME, NUST, since 2006. His research interests include modeling and controlling power converters and their applications in P.V. systems, adjustable speed drives, and artificial intelligence.

SADDAM AZIZ (Senior Member, IEEE) received the M.Eng. degree in electrical engineering from Chongqing University, Chongqing, China, in 2015, and the Ph.D. degree in optoelectronics engineering from Shenzhen University, Shenzhen, China, in 2019. He was a Research Associate Professor, from 2019 to 2021. He is a Postdoctoral Fellow with the Centre for Advances in Reliability and Safety, The Hong Kong Polytechnic University, Hong Kong.

MUHAMMAD TALIB FAIZ received the Ph.D. degree from the Department of Electrical Engineering, Shanghai Jiaotong University, Shanghai, China, in 2021. He is a Postdoctoral Fellow with The Hong Kong Polytechnic University, Hong Kong, SAR, China. His interests include digital control in power electronics converters, electric vehicle onboard chargers, microgrids, and renewable energy generation systems.

LUBAID ZULFIQAR received the M.S. degree in electrical engineering from Xi'an Jiaotong University, Xi'an, China, in 2018. He is currently the CEO of Rapid Volt (PVT) Ltd., Pakistan. His research interests include optimization, renewable energy generation systems, demand response, artificial intelligence, robotics, and low-voltage circuit breakers.

AHMED MOHAMMED SALEH was born in Yemen. He received the bachelor's degree (Hons.) in electrical engineering from the University of Aden, Yemen, in 2015. He is currently pursuing the Ph.D. degree. He was a Laboratory Engineer with the University of Aden, until 2019. His research interests include smart grids, hybridizing renewable energy sources, and optimization techniques.