Optimizing Hybrid Microgrids in Real-time: A Comparative Analysis of Two Reinforcement Learning Training Methods

***Khawaja Haider Ali[1], Mohammed Alharbi[2], Asif Ali Tahir[3]**

[1] Electrical Engineering department, Sukkur IBA University,65200 Airport Road Sukkur, Pakistan;
[2] Department of Electrical Engineering, College of Engineering, King Saud University, Riyadh 11421, Saudi Arabia
[3] Environment and Sustainability Institute (ESI), University of Exeter, Penryn Campus, Cornwall, TR10 9FE, United Kingdom;
Corresponding author: Khawaja Haider Ali (haiderali@iba-suk.edu.pk); Tel: +923216305772

Received: 25-05-2023, Received in revised form: 08-08-2023, Accepted: 29-08-2023, Published: 31-12-2023

ABSTRACT

Reinforcement learning has been employed in recent research articles to optimize the energy storage system scheduling in microgrids, aiming to reduce overall system costs. However, applying reinforcement learning in real-time scenarios introduces uncertainties and delays due to the extensive training required to develop the optimal policy for the storage system. This work addresses these challenges and explores potential solutions for real-time dispatch control actions of the battery in a grid-tied microgrid. The study considers different approaches for training the agent, distinguishing between online and offline scheduling of the energy storage system. The limitations of these approaches and their implications on real-time performance are also analyzed. By developing a comprehensive microgrid model and comparing two training approaches, this research contributes to novel insights for efficient real-time scheduling of energy storage systems in grid-tied microgrids. The proposed approach presents a promising path towards addressing uncertainties and achieving optimal operation in grid-tied microgrids**.** In terms of average cost per year, the difference between the two approaches is 4% if foresight of the real data is perfect, otherwise the real-time approach is more cost-effective.

Index Terms : Reinforcement Learning, Scheduling, Optimization, Charging, discharging, battery

Introduction

Energy management becomes very important in the last one decade because the consumption of electricity is increasing continuously in every part of the world. Scientists and researchers are trying to get optimal solutions for producing cheap electrical energy. Renewable energy sources are widely used to produce cost-effective electricity. This is the reason microgrids have become very popular everywhere because of their benefits, whether they are small or large.. Nowadays artificial intelligence plays an important role to optimize the microgrid to reduce the tariff rate, which is beneficial for both producers and consumers both. There are different methods developed in previous years through machine learning which are used to cut down the cost of production of electricity in a microgrid. The ultimate goal of optimization of a microgrid is to decrease energy bills while taking into account the energy balance and user comfort.

Microgrids may consists of one or more renwable generating sources, storage system, charge controller and inverter. There can be two modes in which microgrid can operate; offgrid or on grid. In both of these case, if a microgrid is managed through proper planning or scheduling its components, for example, an energy storage system, it will pay a lot in terms of cost saving which is ultimately beneficial for its providers and users. there are different algorithms and strategies suggested to manage the whole microgrid which is either standalone or connected to the main grid, for example selection of renewable energy source, forecasting of the load and demand, sizing of the storage system, scheduling of storage devices, sitting and many more.

Machine learning technique named reinforcement learning (RL) is used to implement this work. There are a lot of applications developed in the past few years using RL. For example; in Atari games, robotics, web system configurations, advertising, and many more. In our work, RL is used to decide the operational mode for the battery: stay idle, charge, or discharge. The RL goal is to minimize the cost of electricity and maximize the self-consumption of locally produced electricity. The planning/scheduling regarding operation had done by sequential decision-making problem using markov decision process (MDP) [1]. The different other algorithms used in optimization techniques are linear programming (LP), mixed integer programming (MIP). The optimization is generally formulated as a mixed integer nonlinear problem (MINLP) for which there is no exact solution method [2]. However, s microgrid modeling needs both continuous and discrete decision variables to specify on/off states of distributed generated units (DGS), loads, or both [3]. It causes the solution space of the consistent optimization problem to be nonconvex. That is why classical mathematical programming techniques are very difficult to be applied directly [4]. Therefore, due to the problem complexity and achieving large economic benefits, this area needs considerable attention to develop better optimization algorithms.

Furthermore, previous studies [5-10] have shown that microgrids can achieve high performance by developing demand response and an optimal framework for utilizing storage devices to compensate for physical imbalances. However, these types of approaches are commonly used in applications which are computationally intensive[11]. Also, they are not suitable for online optimization [11]. Most dynamic programming uses priority list which have heuristic-based techniques [3]. The optimization of microgrids, particularly those with non-linear behavior, is being extensively studied to discover improved solutions, especially for real-time control of the system.This is the reason and motivation to find out more practical solutions to deal with the optimization of the online microgrid problem.

This work aims to optimize the microgrid by controlling the battery commands/actions to provide cost-efficient, reliable, and real-time solution, nevertheless, achieving the following targets:

1. A comfortable and reliable system for the users.
2. Can control the load shedding.
3. Maximise the utilization of renewable energy resources by reducing the dependency on the main grid.

https://doi.org/10.24949/njes.v16i2.757

4. Increase the cost savings, by using the storage system optimally.
5. Sell power to the main grid when the utility tariff is high.
6. The status of charge, discharge, and idle modes are controlled by respecting the different parameters of the battery.

This planning or scheduling mainly consists of model-based and non-model-based approaches. These are further divided into value-based and policy based approaches which are the types of Reinforcement Learning (RL) under the umberella of Artificial Intelligence (AI).Example of value based approach in RL is Q learning, SARSA, value iteration .

In [13], the optimization of hybrid microgrid operation is proposed, focusing on load sharing control and integration into the electricity market. It explores mathematical modeling and introduces decision Tree for microgrid optimization. This study demonstrates improved voltage profiles using different methods and suggests potential for competitive renewable energy integration. In another paper [14], introduces a learning-based control strategy for microgrid energy scheduling, using reinforcement learning algorithms without explicit models. The approach optimizes the use of local renewable energy and reduces operating costs. Simulations with real data demonstrate improved energy utilization efficiency and peak load shaving.

In this work we are using Q learning. On the other hand, the example of Policy based or actor-critic is Reinforce, Crossentropy method. It is not necessary that theoptimizationn problem can be handled by one type of approach every time. This is because all approaches relate to each other in one way or another as shown in below figure 1. It is the fact that in different kinds of Scenarios, optimization should be done by different methods. This is because of the nature of the problem, sometimes it is due to the stochastic behaviour of the system or due to deterministic and non-deterministic characteristics of the environment as well [15].

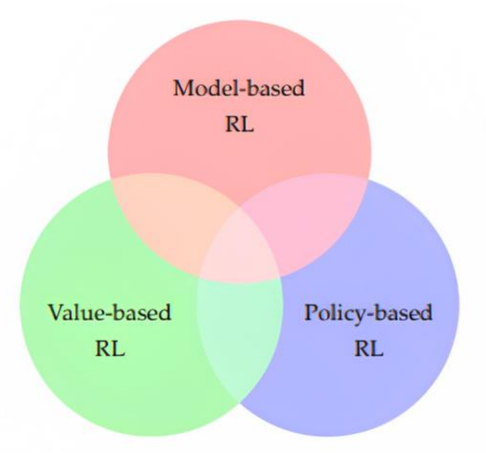


Figure 1. **Venn diagram of different types of RL algorithms[11]**

The extensive literature review of previous similar research by combining different parameters and constraints from different articles to develop a model and algorithm to optimize the battery from RL and dispatch its actions in real time.

This paper proposes an RL to deal with real-time dispatch control actions of energy storage. The above sections describe the Introduction and background.

The rest of this paper is organized as follows. Concept of Q learning used in this paper and then the scope of this work. Section II presents the main model of the microgrid which is used in this article. Section III introduces proposed methods and approaches used to optimize the grid tied microgrid. Methodology behind the implementation of Reinfocement learning in this work is presented in section IV. Finally, the result and conclusion are discussed in section V and VI respectively.

Fundamentals of Q Learning

Q-learning is a value-based method used in RL to find the optimal action-selection policy [16]. This is done through a Q function. The prime goal is to maximize the value function Q [16]. The Q table helps to find the best action for each state [17]. It maximizes the expected reward by selecting the best of all possible actions [17]. Q (state, action) returns the expected future reward of that action in that state. This function can be estimated by using Q-Learning phenomena, which iteratively updates Q(s, a) using the Bellman equation[17]. The bellman equation is given by:

$$Q(s,a) = Q(s,a) + \alpha(Reward + Gamma \times \max\big(Q(s',a')\big) - Q(s,a)) \quad (1)$$

Where:

- Q(s, a) is the current estimate of the action-value function for state s and action a.
- α (alpha) is the learning rate, which determines how much the new information will influence the current estimate. It is a value between 0 and 1.
- γ (gamma) is the discount factor, which determines the importance of future rewards compared to immediate rewards. It is a value between 0 and 1.
- max(Q(s', a')) represents the maximum expected cumulative reward achievable from the next state s onwards, considering all possible actions a in the next state.

This equation allows us to start solving these Morkov's decision processes (MDPs). The Bellman equation is ubiquitous in RL and is necessary to understand how RL algorithms work. In Bellman equations, values of states are expressed as values of other states. By knowing $St + 1$, we can very easily calculate St. This opens up a lot of possibilities for calculating the value for each state iteratively since we can know what the current state is if we know $St + 1$. Bellman equations, can determine optimal policies and train reinforcement-learning agents. [18]. Initially Q learning or RL agent explores (process of exploration) the environment and update the Q-Table [19]. When the Q-Table is ready, the agent will start to exploit the environment and start taking better actions [19].

Initially during exploration of the environment Q-table is updated by taking random action [16]. Once the Q-Table is ready, the agent will start to exploit the environment and start taking better actions [16].

In RL (Q-learning), the environment is explored randomly and after reasonable iterations, the decision is taken according to the statistics of the environment [15]. This is called training of the agent. The field of RL research describes to learn how to act in an environment from previous experiences [20]. This is the beauty and a benchmark of this algorithm over other techniques of optimization as it interacts with the unknown environment and then experiences gathered are used to optimize some objectives such as decrease tariff [19]. This approach, in general, solves sequential decision problems by relying on past findings [21]. The information may be highly dimensional, for example in a certain state, specific action has performed. This is known as a policy. The policy used in this work is defined by :

$\pi(s\times a) = S\times A \rightarrow [1,2,3]$

where $\pi(s,a)$ is denoted by the probability that action a (1,2,3) may be chosen in state s.

During the policy formation or training phase, the time is consumed which can delay the dispatch of a set of actions to the control unit which may not be bearable during online scheduling. But the training is necessary to get optimal solutions [22]. Also, in practice there are many uncertainties that come from a lack of knowledge about the future, for example, load demand varies in real-time scenarios or the change of weather can affect the renewable sources output, which may affect the optimal solution in real time [23]. In this work, this issue had been addressed comprehensively by dividing them into two approaches.

SCOPE OF THIS WORK

In previous research such as [6][9-15][21][26], not all features were simultaneously incorporated, but this paper tries to incorporate all possible scenarios in real-time, which will allow this algorithm to be used in the current grid-tied structure to get the maximum optimization in terms of cost. Also, very helpful in different practical applications. Following are the salient features of the whole network used in this Research Article.

1. The renewable energy which is PV here has a priority to fulfil the demand of load first. If it is not enough then battery or utility grid or combination of all mentioned resources are used to fulfill the demand.
2. Battery can be charged from the PV directly. It can be charged from the utility grid as well.
3. It is also possible that at high tariffs battery can be discharged into the utility grid as a feed-in tariff to earn money. Here, there is one fixed feed-in tariff assumed.

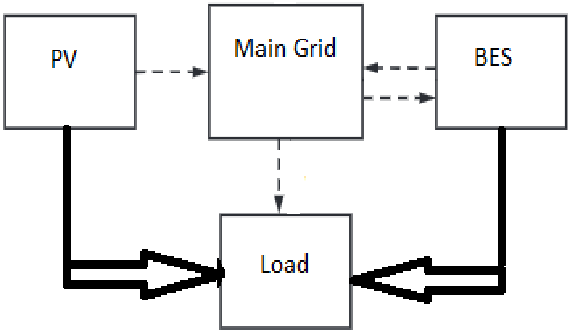


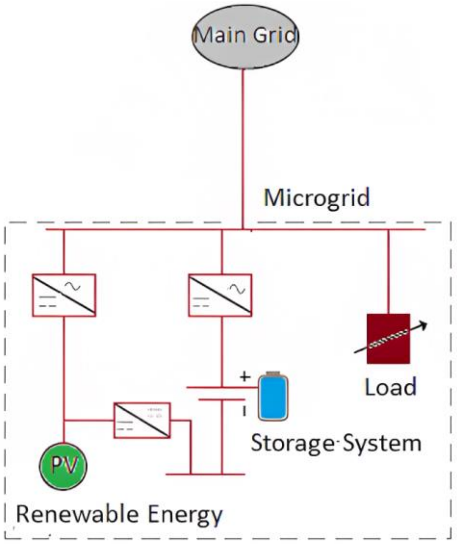
Figure 2. Block diagram of the proposed model.

The proposed algorithm decreases the net energy cost may be by considering the future prices by using the current system information during the training period. In the training session, the convergence of the Q table will be done by tuning the hyperparameters of the RL such as exploration versus exploitation, learning rate, and discount factor. Once, the algorithm learns pedagogy, the policy has been made which is followed by the data in real-time. But the main problem is that the training which has been made on forecasted data fails or not up to the mark due to changing occurring in real time due to changes of weather conditions or load demand. So, if training is done on forecasted data the forecasting should be appropriate and very near to real-time data, best optimization is not possible. To deal with this problem the 2nd approach (real-time implementation and solution) is proposed in this work which is in section 3 below.

Microgrid Model

The structure as shown in Fig. 3 composed of renewable energy sources (such as PV), the storage energy system, the main grid energy, and the loads. The microgrid proposed in this work consists of a PV system, a battery as the energy storage system, loads, inverters, and a microgrid connected to the main grid.

The inverters convert the Direct Current (DC) to Alternating Current (AC) from the battery and PV. Information on electricity prices is available to microgrid users due to the microgrid's connection to the utility grid [27].



Fgure 3. Proposed microgrid configuration. The microgrid includes PV system, battery storage, loads, inverters, and connects to the main grid.

Loads can be varied at different times of the day. If renewable sources are present within a system or microgrid, its output is dependent on the different conditions, e.g. weather [28]. So, a microgrid may supply power either from a renewable source or from the main grid to loads. The presence of the battery or energy storage system in microgrids has many advantages. One of the advantages can supply power to loads in case of absence or shortage of power from a renewable source to avoid the use of utility power [28]. Also, if the microgrid did not have so much demand at the load side, the energy of the battery can be injected into the main grid as well depending upon the profit gain after selling it to the main grid. For example, if the battery has enough storage and in the remaining days there is not much need for storage to run the load then it is cost-effective to sell the power of the battery to earn or save more cost [29]. The state of charge of the battery should be controlled according to its optimum level. It is also part of the management system of the battery to discharge up to the optimum level when there is a demand from the load side. In other cases, it can deliver or sell its power to the main grid. Battery should take care while supplying its energy to the main grid that after discharging, it should get a low tariff from the main grid or get power from a renewable source to charge again. Otherwise, do not sell energy to the main grid to save its charging for the next day or high-demand load conditions. By RL, the agent decides by knowing load demand, tariff rates, weather conditions, and other necessary riables that what is the appropriate action for the battery [30].

Proposed Method And Approaches

In this research article, sequential decisioking under uncertainties is proposed. In a 24-hour day at every time step (1 hour,15 minutes), stocbehaviorviourof the environment is explored through Reinforcement learning which is done by adopting two different approaches for training the data for a whole year.

In 1st approach, the decision has been made to direct the battery of the microgrid to execute the optimal actions after certain training. This will be done on a daily basis.The training is done on forecasted PV and load profiles. The optimal actionsderived are drivethe d from tforecastedf forcasted data are then used to commad the battery actions, charge, discharge or remain idle for the current year or real data (PV and load) profiles of the year. At the end of the current year it wil,l show the overall yearly cost of the microgrid which is connected to the utility grid. Or daily cost can also be monitored or checked by this approach. The 2nd approach based on an online training mechanism is used to optimise the grid tied microgrid. The training and getting optimal actions simultaneously on a daily basis is the idea behind this approach.there is no need of forcasted one year profile for this 2nd approach because it is implemented on current year. Unlike approach 1 there will be no separate training before dispatching the actions of the battery. The training and dispatching of control actions are done simalteanoulsy at real time. At day 0 the Q table is intialised with state vs actions of the battery (charge, discharge, idle)by getting instant reward. This can be done by usinga learning rate (gamma) equal to zero. Now as day 1 starts, Q table updated it values (actions) vs each state by moving from the 1st hour to 24th hour of the day.The update of Q table does not require intensive training. However, due to instant reward by moving from 1st state to another state it sends the actions command of the battery on run time. It is very important to know that the differehyperparametersers which are used in Q learning are incorporated during this adopted approach. The only difference is that it does not comprises of too many rounds as done in previous 1st approach. For example the actions drived from 1st round are considered to be optimal for that particular 1st day while hyperparameters slightly change itself to show learning. Although this learning is very minor due to only one round or episode but due to time considerations in real time this slight learning is enough for 1st day of the year.The hyperparameter's slight changes on the 1st day reflects to the 2nd day. The 2nd day follows the same procedure as done on 1st day. It also send the control actions to the battery at real time. As, the days progress the hyperparameteres of RL which are responsible for the learning will change more and more until saturation of training happensed.At this time the Q table converges the to give best optimal actions for the battery. These best optimal actions of the battery will be achieved after approximately 50-60 days depending upon the different conditions of the Q learning algorithm such as the sampling period of the time.

The difference between 1st and 2nd approach used in this work is that in 1st approach due to intensive training the dispatch of control actions of the battery cannot be send online or run time of the microgrid due to time consumed for training. But the 2nd approach due to only one round of training per day the control actions per hour need very less time to execute at real time.The 1st approach can give optimal actions or save cost from day 1 of the year depending upon the forecasted PV and load profiles. But the suggested 2nd approach start giving optimal actions after one month or more.So, the 2nd approach is more useful when the cost is saved on annual basis. But to use 1st approach as it is, in online system is quite difficult due to time constraints. Also 1st approach requires very efficient forcasted algorithm for PV and Load profiles.But the work done so far in forcasted algorithm in previous similar researches development does not guarantee the maximum efficiency.

So, It is proposed as a 1st approach in this work that if forecasted profiles of load, and PV are very close to real profiles then training on a daily basis is appropriate. This is because the policy made by using forecasted data can exactly or approximately relate to real-timetime data.

Methodology/Formal Frame Work

The difference between load and PV profile are abbreviated as D (Net demand). The D can either be D>0 or D<0. Before taking action of the battery, the above two cases can be observed. In other words, for every action, there will be two scenarios either load<PV or load>PV. So for every action of the battery charging, discharging and idle, there will be chances of two scenarios each. Therefore total possibilities become six. The details of actions are in below section.

Declare all the states of the system, actions of the battery, Transition probability, Reward and Cost function's described briefly in below section. The hyper parameters are one of the important constraints in RL. These are mainly the discount factor, Exploration vs Exploitation (ε greedy factor) and Learning rate. All constraints for both approaches (training on 1 year forecasted data per day and real time training and optimization for 1 year) used in this work have same features because states are repeated from one day to another till 365 or 366th day of the year depending on regular year or leap year respectively. While others terms such as actions, reward function are also the same. The only difference is regarding their interpretation in both approaches as if training is done on forecasted one-year data.1st approach is trained on a daily basis and stores optimal actions which will be followed by real data.

On the other hand, the 2nd approach real time data on a daily basis is not trained as many iterations like the 1st approach, it will train as the day's progresses.

1) System States

In reinforcement learning, it is a very crucial part to define states of the system. As states decide the maximum horizon of the model. Also, careful declaration of states makes the model close to practical. Two types of problems can be dealt with RL methodology, finite horizon or infinite horizon. This work consists of a finite time horizon. We are interested in getting the optimal solution to our problem in a complete day. Day is divided into 24 hours and further it is divided among one hour or 15 minutes time interval. So, there are 24 or 96 states for time respectively.

Equation 2 shows the states of the battery which represent the capacity of the storage system. Here, we considered three states of the battery capacity.

The generalized equation of system states is:

$$\mathbf{S_t \times S_{SE}}$$

where $\mathbf{S_t}$ is the time feature of the state which is divided into 24 or 96 intervals as one day has24 hours. $\mathbf{S_{SE}}$, represents the states of the storage energy in the battery. This is the transition probability of the system.

Total No. of states are$S = 24 \times 3 = 72$ (if sampling time is per hour)

Or

Total No. of states are$S = 96 \times 3 = 72$ (if sampling time is 15 minutes)

2) Transition Probability

The transition probability of the Energy storage system from state, s to state, st+1 when action at is taken can be represented by three states equations, as referenced in [29]

$$\begin{cases} SEmin \le SE < SEmin + \frac{1}{3}capacity \\ SEmin + \frac{1}{3}capacity \ \le SE < SEmin + \frac{2}{3}capacity \\ SEmin + \frac{2}{3}\ capacity \ \le \ SE \ \le \ SEmax \end{cases} \qquad (2)$$

Where:

Capacity= $SEmax - \ SEmin$

At, every time interval t there can be possibility of one SE level out of above 3.

3) Actions

Depending on the actual state out of 72 states, the system chooses between the actions.

A= [1, 2.3], represents the list of actions.

a3=Charging of battery; a2=Idle; battery; a1=Battery discharging.

In section (Methodology/Formal framework) the difference between load and PV were abbreviated as D and for random every action there are two possibilities for having next state (battery) as St+1 and Power utilization from the grid as Pgrid.

Hence, the difference between load and PV makes two cases. Every case has below mentioned equations for SE (t+1) and Pgrid.

Scenerio 1. (D<0)

A=3 solar generation exceeds the load, the excess is used to charge the battery and then the rest is injected into the grid.

A=1 solar generation exceeds the load, all the excess is injected to the grid + the battery?is discharged into the grid.

A =2 solar generation exceeds the load, the excess is injected into the grid

$$\begin{cases} a = 3\ (Charging) \\ SE(next) = \min\big(SE(t) + \min\big(b, -D(t)\big), Cmax\big) \\ Pgrid(t) = \Big(-D(t) - \min\Big(\min\big(b, -D(t)\big), Cmax - SE(t)\Big)\Big) \\ a = 1\ (discharging) \\ SE(next) = SE(t) - \min(b, SE(t) - Cmin) \\ Pgrid(t) = (-D(t) + \min(b, SE(t) - Cmin)) \\ a = 2\ (Idle) \\ SE(next) = \ SE(t); \\ Pgrid(t) = \big(-D(t)\big) \end{cases} \qquad (3)$$

Scenerio 2 (D>0):

a=3 demand deficit, the utility is used to compensate the load and charge the battery.

a=1 demand deficit, the battery together with the utility are feeding the load.

a=2 demand deficit, the utility is compensating the deficit.

$$\begin{cases} a = 3\ (Charging) \\ SE(next) = \min(SE(t) + b, Cmax) \\ Pgrid(t) = \big(D(t) + \min\big(b, Cmax - SE(t)\big)\big) \\ a = 1\ (discharging) \\ SE(next) = SE(t) - \min\big(\min\big(b, D(t)\big), SE(t) - Cmin\big) \\ Pgrid(t) = \big(D(t) - \min\big(\min\big(b, D(t)\big), SE(t) - Cmin\big)\big) \\ a = 2\ (Idle) \\ SE(next) = \ SE(t) \\ Pgrid(t) = \ D(t) \end{cases} \qquad (4)$$

4) States vs Actions

In each state out of 72, we have 1 action which means that at every state depending upon the SE of the battery actions may differ. As at every time step t SE of the battery is checked and depending upon SE level (state) the actions are noted. Here are three states of SE so at every time interval t so, there are three sub-states of SE in which actions are taken. The actions in every SE at time 1 may be different because of the SE different levels. This will be updated in every round depending upon the reward function. As the system will learn with continuous iteration and in last iteration system converges to give action per state which is the best or optimal.

At the end when system converges, we extract the optimal actions (24,96) out of 72,288 respectively depending upon the sampling period of time.

5) Reward and Cost Function

The reward is –ive of cost function (Reward=-cost. So, higher cost make the reward lesser and vice versa. Cost function is described, depending upon the cases D<0 and D>0 respectively by below equations respectively.

Such as:

$$cost(s, a) = -Tariffinj \times Pgrid(t)\ ; D < 0 \qquad (5)$$

$$cost(s, a) = Tariff(t) \quad \times Pgrid(t);\ D > 0 \qquad (6)$$

(Reward (s,a)=-cost (s,a)

where Tariffinj shows the feed in Tariff (microgrid to utility grid) and fixed, while Tariff(t) is the tariff decided by AC grid. It is variable in this work having three different values peak, medium and low.

Exploration vs Exploitation Dilemma

The exploration-exploitation dilemma is a famous tradeoff in RL [8]. In the reinforcement learning setting, no one gives us some batch of data like in supervised learning. We're gathering data as we go, and the actions that we take affects the data that we see, and so sometimes it's worth to take different actions to get new data [30]. The agent starts accumulating information about its environment, it has to make a tradeoff between learning more about its surroundings which is called Exploration and pursuing what seems to be the most promising strategy with the experience gathered so far known as exploitation [30].

There two adopted approaches to handle Exploration VS Exploitation given below:

The agent is expected to perform well without a separate training phase and in that case, an explicit tradeoff between exploration versus exploitation appears so that the agent should explore only when the learning opportunities are valuable enough for the future as compared to what direct exploitation can provide[16]. But it may not be so efficient in terms of optimization of cost.

The second case which is considered in this work, the agent follows a training policy during the first phase of interactions with the environment so as to accumulate training data and hence learn a test policy. The test policy should then be able to maximize a cumulative sum of rewards in a separate phase of interactions. The goal of the training policy is then to ensure the efficient exploration of the state space without constraint directly related to the cumulative reward objective. [3].

The epsilon (ε) consider in this paper is as below:

`

$$\varepsilon = \frac{\varepsilon}{\sqrt{M - Mmax}} \qquad (7)$$

6) Learning Rate

The learning rate (α) is how quickly a network abandons old beliefs to new ones. In general, to find a learning rate should low enough that the network converges to something useful, but high enough that it does not have to spend years training it [17].

In this work, α is described by below function equation 8.

$$\alpha = \frac{\alpha}{N - Nmax} \qquad (8)$$

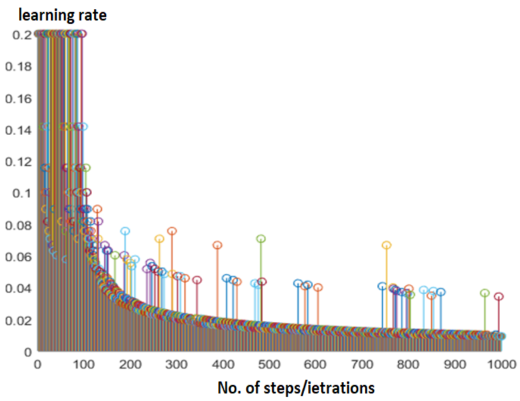


Figure 4. Comparison of Learning Rate and Rewards across Episodes.

Fig. 4 shows the learning rate decay. This change of learning rate happens every day during training for the 1st approach used in this work. While the other novel 2nd approach used in this work show the decay of learning rate after days progress.

7) Discount Factor:

It is the factor, which represents the difference in future and present rewards. RL Important Simulation Parameters used in this work are mentioned in the below Table 1.

TABLE 1
PARAMETERS USED IN THIS WORK

Name	Values
Total Capacity of the battery	12000Wh
Max charging rate of battery	2300W
Minimum charging rate of battery	2300W
Initial SOC of the battery	9000Wh
Min depth of discharge	8400Wh
Learning Rate	0.5
Discount factor	0.9
Epsilon	0.2
Time step Length	1 hour

Results

The aim of this paper was to develop the algorithm using RL , which provides Cost effective optimization and handle the dispatch of control actions of the battery in real-time. This work is simulated using MATLAB 2019(a). In this regard, help is taken from Matlab tool box [31]. However, simulation results were produced through coding.

This section provide the results of used approaches 1 and 2 and compared both of them as well.The results show the convergence timing of the 2nd approach with respect to 1st approach. Below figure 5 represents the profiles used to implement this work.

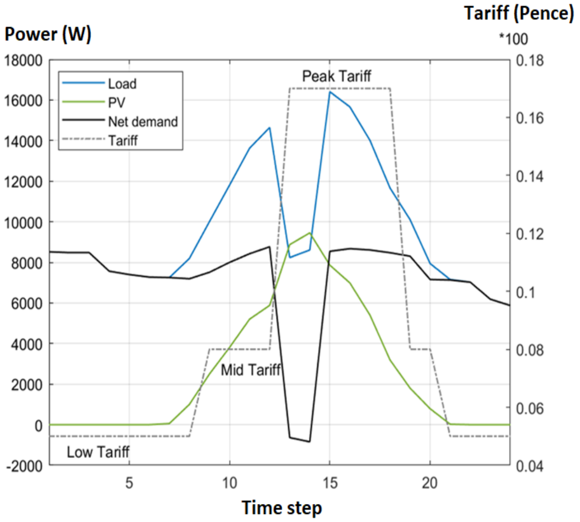


Figure 5. Performance comparison of approaches 1 and 2, Illustrating convergence timing of the 2nd approach.

Below Figure 6 shows the battery commands achieved after training by both approaches 1 and 2. This graph shows charging and discharging of battery on positive and negative y axis respectively. While when battery remains idle figure 6 below show no bar.

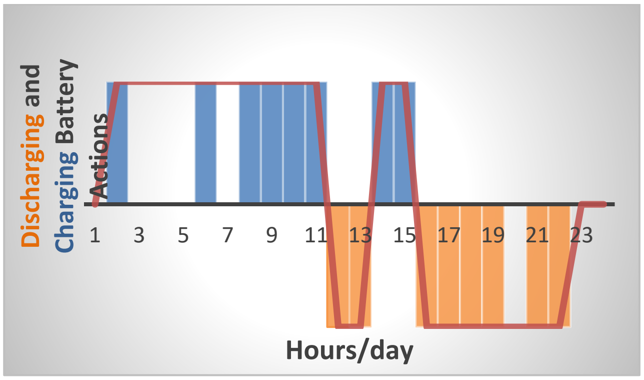


Figure 6. Battery Commands/Actions after training of both approaches 1&2

The Figure 7 below shows the Commulative optimal cost per day for the whole year by comparing both approaches used in this paper.The sampling time for this graph (Fig 7) have one hour per day.

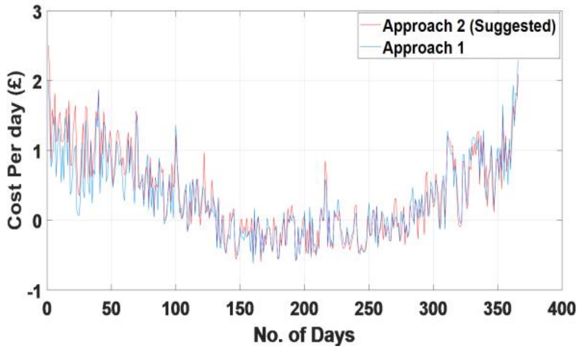


Figure 7. Comparison between both approaches in terms of daily cost using 1 hour Sampling time period

The 1st approach named as conventional training which is on an hourly basis per day shows the average optimal cost for every day of the year. The data used for training is assumed to be forecasted data. It means that the training which has been done by using data profiles for example PV and load, give best Optimal actions of the battery. These actions are then used to command the battery in real time. Though the forecasted data may be different than real-time data but we assume that both data is not so much different. This assumption is to achieve a bench mark for the suggested 2nd approach that how much time it will take to converge with respect to Approach 1. Because the suggested approach is applied to assumed real data. However, it is almost impossible to get real time PV and load profiles exactly by using forcasting. The other pattern of the graph (non training) shows the optimal cost per day by using the 2nd approach which is by following the learning process as the day's progress. In 2nd approach, it has been observed that best cost saving start increasing after 100 days approximately. This is because from day 1 to day 100 agent was learning and then start converging until it reaches in between 100-110 day. This is because now suggested approach which was learning online had updated its Q table with maximum commulative rewards against with each state vs actions.As,in the beginning the suggested approach had not trained so much and does not know the stochastic behavior of the environment completely.So, it take some days to gather information about the unknown environment and then betterly suggest the optimal actions. The sampling time of the suggested approach 2 is also 1 hr per day as in approach 1 in above graph figure 7.

The difference between the net average cost per year of both approaches is approximately 4%.The training on daily basis in approach 1 is more cost effective in one year analysis then the 2nd approach. The reason is quite obvious as approach one save cost from the day one while the 2nd approach become efficient in terms of cost after 100 days approximately as shown in figure 7. But please noted that if the percentage age of error between forecasted and real data become high then the approach 1 become less cost efficient then the proposed new algorithm (Approach 2).

It is also tested in this work that if sampling time used in RL become shorter than the convergence time of the 2nd approach become less.For example as in figure 8 below the sampling time is 15 minutes rather than 1 hr. It means that after each 15 minutes of the day the RL training is performed and test both approaches 1 and 2.The approach 2 now have more trainings in a day which can lead shortening of its convergence time.So, it will show cost saving earlier compared to Figure 7.

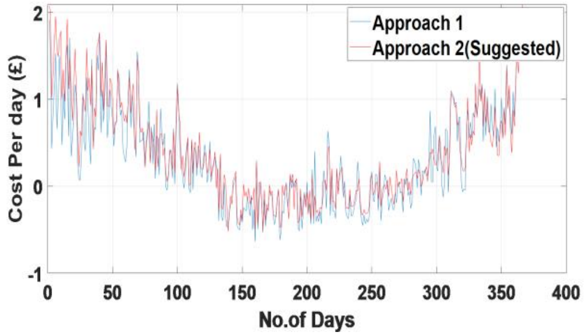


Figure 8. Comparison between both approaches in terms of daily cost using 15 Minutes Sampling time period

Figure 8 above shows that the suggested approach start converging approximately between 90 to 100 days. However, a comparison between both approaches in above figure 8 suggests that approach 2 (suggested) gives best optimal solution after 90 days.

It is also noticed that the convergence achieved in figure 8 is better in terms of cost than figure 7. Therfore, by decreasing the sampling period in terms of time the results bocme better. So, in this case the optimal average cost per year in both approaches are approximately same. But if the forcasted data used in approach 1 become more different than the real data then approach 2, suggested in this work perform better.

Power Import From Main Grid

The figure 9 below shows the power imported per day from the main grid**.**

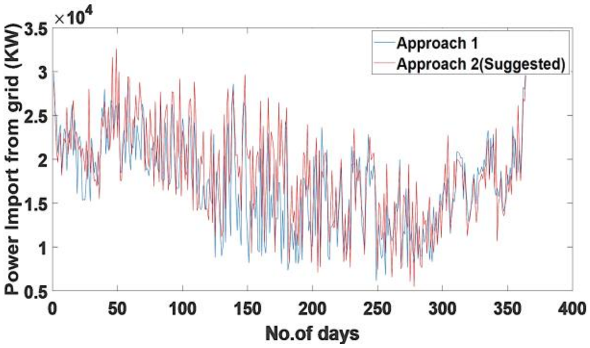


Figure 9. Comparison between both approaches in terms Power imported,daily from Main grid

The power from the utility grid which is utilized during the shortfall of microgrid power especially at peak hours are shown in the above graph in comparison to both approaches used in this work. Figure 9 above shows that the total average poer imported from the main grid may show an increasing and decreasing trend after convergence for approach 2. This can be an argument that the Power imported from main grid should be decreased after the convergence of Approach 2. But the justification to show a bit different behavithanhen the therotical concept is due to adopted model approach used in this work. It is not necessary that the power from main grid is imported during the shortfall of microgrid power. It can be imported even when the battery has enough power to fulfill the demand of the load. The reason is that the battery power may be saved for the time when main grid power has high tariff rate and there are not enough PV sources power avialble to satify the user demand.Hence, the main objective for this work which is to decrease or optimize the overall average cost of the utility grid per year by controlling the battery actions (charge,discharge,Idle) have been achieved.When used 15 minutes sampling period for time the total average power imported from the main grid become same in both approaches 1 and 2 depending the difference between forcasted and real data profiles.If forcasted and real data are more different then the total average import from the main grid is also decreased in approach 2 as compared with approach 1.

Storage Energy Of The Battery

This work takes care of all the parameters which can affect the efficiency of battery in terms of its life and state of charge.It is also suggested in this work that the battery should not charge or discharge at once upto a certain level. Table 1 shows all the used parameters for the battery in this work.Figure 10 below is the energy storage of the battery taken when the sampling period is 15 minutes.

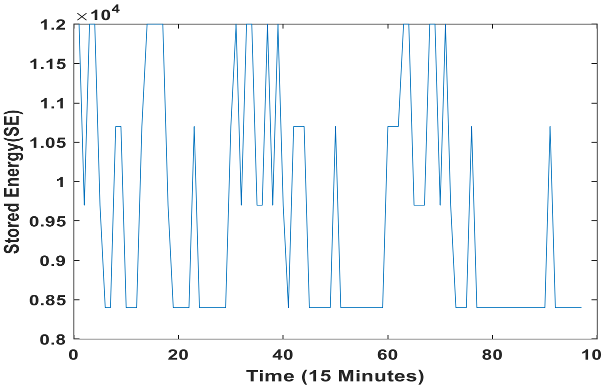


Figure 10. Battery energy storage profile at 15 minute sampling period.

Validity Of Suggested Approach

The training approach adopted in this work named as 2nd approach is checked by using different data sets of PV and load profiles.To,observe the behavior of suggested approach 2 on different data we consider below function to add or subtract noise in load and PV profiles.

$$Load = Load \pm c \times (rand(1, X) \qquad (09)$$
$$Pv = Pv \pm c \times (rand(1, X) \qquad (10)$$

where X is the sampling period; for example if the sampling time is 1hr then the X =1*24*365=8760.If sampling time is 15 minutes then X=4*24*365=35040.

C is the constant that is chosen to add or subtract the amount of noise in the data (Pv, load).

CASE 1:

If the load is increased or decreased while keeping the PV constant. The behavior of the optimal actions should be changes. For example if load is increased while PV remain constant the overall average cost per year is increased and vice versa.

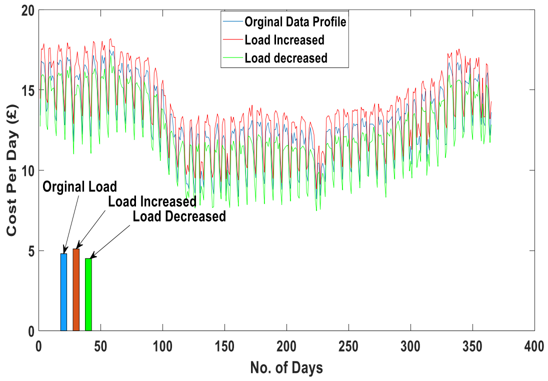


Figure 11. Effect of Load Variation on Optimal Actions and Yearly Average Cost.

The above figure 11 comapare the changings done in load profile with respect to orginal laod.It is validating the results as in theory that increase or decrease of load can increase or decrease the daily cost respectively.

CASE 2:

The suggested approach is checked that on different PV profiles it is working or not. By using equation 10 different PV profiles are generated to see the behavior of approach 2.

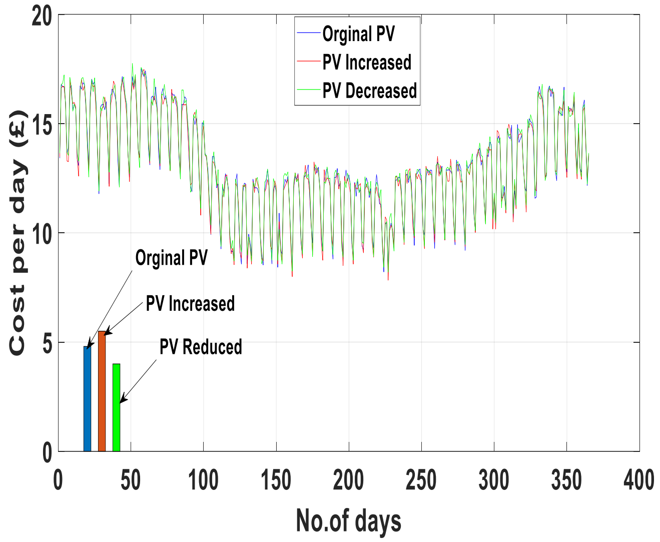


Figure 12. Evaluating approach 2 on diverse PV profiles.

Above figure 12 suggested that increase or decrease of PV will decrease or increase the daily cost respectively.

Conclusion:

This research proposes two approaches for optimizing the control actions of a battery in a Microgrid. The first approach relies on forecasted PV and load data, requiring intensive daily training and potentially being less adaptable to sudden weather changes. In contrast, the second approach introduces an online training mechanism using RL, allowing real-time control actions and continuous learning from day one to the end of the year. This approach avoids the need for forecasting, making it more resilient to abrupt weather changes. Furthermore, the second approach achieves cost savings by considering the average performance over the entire year. Finally, the second approach presents a promising alternative for efficient and adaptive battery control in Microgrid systems.

In future, this Aagorithm can be validated by any other optimal technique such as by linear programing or by solving the problem by other MDP techniques. Comparison of both solutions give more accurate idea about the validity of Reinforcement learning in different optimization problems.
This technique can be tried on real system on run time. The Load, PV and Tariff profiles are also used accurately according to the region where Microgrid is installed.

REFERENCES

1. Aljohani, T. M., Ebrahim, A., & Mohammed, O. (2021). Real-Time metadata-driven routing optimization for electric vehicle energy consumption minimization using deep reinforcement learning and Markov chain model. Electric Power Systems Research, 192, 106962.
2. Kadowaki, T., & Ambai, M. (2022). Lossy compression of matrices by black box optimisation of mixed integer nonlinear programming. Scientific Reports, 12(1), 15482.
3. De Mel, I., Klymenko, O. V., & Short, M. (2022). Balancing accuracy and complexity in optimisation models of distributed energy systems and microgrids with optimal power flow: A review. Sustainable Energy Technologies and Assessments, 52, 102066.
4. A. Das and Z. Ni, "A Computationally Efficient Optimization Approach for Battery Systems in Islanded Microgrid," in IEEE Transactions on Smart Grid, vol. 9, no. 6, pp. 6489-6499, Nov. 2018.
5. F. Ruelens, B. J. Claessens, S. Vandael, B. De Schutter, R. Babuška and R. Belmans, "Residential Demand Response of Thermostatically Controlled Loads Using Batch Reinforcement Learning," in IEEE Transactions on Smart Grid, vol. 8, no. 5, pp. 2149-2159, Sept. 2017.
6. E. Chaimaa, M. R. El-Fenni, and H. Dahmouni. "Cost-Effective Energy Usage in a Microgrid Using a Learning Algorithm." , Hindawi Wireless Communications and Mobile Computing Volume 2018, Article ID 9106430, 11 pages.
7. S.Bahramirad “Reliability-constained optimal sizing of energy storage system (ESS) in a microgrid”, in IEEE transactions on smart grid, volume.3 no.4, December 2012.
8. Xu, M. Bishop, D. G. Oikarinen and C. Hao, "Application and modeling of battery energy storage in power systems," in CSEE Journal of Power and Energy Systems, vol. 2, no. 3, pp. 82-90, Sept. 2016.
9. U.Mukthar, “Energy Management Strategy For Renewable Sources Using Battery Scheduling Process”, Volume 3, Special Issue 3, March 2014, 2014 International Conference on Innovations in Engineering and Technology (ICIET’14) On 21st&22ndMarch Organized by K.L.N. College of Engineering, Madurai, Tamil Nadu, India.
10. Kofinas, Panagiotis, G. Vouros, and A. I. Dounis. "Energy management in solar microgrid via reinforcement learning using fuzzy reward." Advances in Building Energy Research 12.1 (2018): 97-115.
11. M. Castronovo “ Offline Policy search in Bayesian Reinforcement learning” , PhD dissertation, Advisors:

Damien Ernst University of Liège, Faculty of Applied Sciences, Department of Electrical Engineering & Computer Science, 2016.

12. https://joshgreaves.com/reinforcement-learning/understanding-rl-the-bellman-equations,retreived on 15June 2022.
13. Bilbao, J., Bravo, E., García, O., Rebollar, C., & Varela, C. (2022). Optimising energy management in hybrid microgrids. Mathematics, 10(2), 214.
14. Bi, W., Shu, Y., Dong, W., & Yang, Q. (2020, October). Real-time energy management of microgrid using reinforcement learning. In 2020 19th International Symposium on Distributed Computing and Applications for Business Engineering and Science (DCABES) (pp. 38-41). IEEE.
15. S. Kim & H. Lim, 2018."Reinforcement Learning Based Energy Management Algorithm for Smart Energy Buildings," Energies, MDPI, Open Access Journal, vol. 11(8), pages 1-19, August.
16. K. H. Ali, M. Sigalo, S. Das, E. Anderlini, A. A. Tahir, and M. Abusara, “Reinforcement Learning for Energy-Storage Systems in Grid-Connected Microgrids: An Investigation of Online vs. Offline Implementation,” *Energies*, vol. 14, no. 18, p. 5688, Sep. 2021, doi: 10.3390/en14185688.
17. https://towardsdatascience.com/reinforcement-learning-demystified-exploration-vs-exploitation-in-multi-armed-bandit-setting-be950d2ee9f6,retrived on 12 January 2023.
18. Kofinas, Panagiotis, George Vouros, and Anastasios I. Dounis. "Energy management in solar microgrid via reinforcement learning." Proceedings of the 9th Hellenic Conference on Artificial Intelligence, pages1-7,May, 2016.
19. K. H. Ali, M. Abusara, A. A. Tahir, and S. Das, “Dual-Layer Q-Learning Strategy for Energy Management of Battery Storage in Grid-Connected Microgrids,” *Energies*, vol. 16, no. 3, p. 1334, Jan. 2023, doi: 10.3390/en16031334.
20. T. A. Nguyen and M. L. Crow, "Stochastic Optimization of Renewable-Based Microgrid Operation Incorporating Battery Operating Cost," in IEEE Transactions on Power Systems, vol. 31, no. 3, pp. 2289-2296, May 2016.
21. Essayeh, M. Raiss El-Fenni and H. Dahmouni, "Towards an intelligent home energy management system for smart Microgrid applications," 2016 International Wireless Communications and Mobile Computing Conference (IWCMC), Paphos, 2016, pp. 1051-1056.
22. L. I. Minchala-Avila, L. Garza-Castañon, Y. Zhang and H. J. A. Ferrer, "Optimal Energy Management for Stable Operation of an Islanded Microgrid," in IEEE Transactions on Industrial Informatics, vol. 12, no. 4, pp. 1361-1370, Aug. 2016.
23. A. Das and Z. Ni, "A Computationally Efficient Optimization Approach for Battery Systems in Islanded Microgrid," in IEEE Transactions on Smart Grid, vol. 9, no. 6, pp. 6489-6499, Nov. 2018.
24. D. E. Olivares, C. A. Cañizares and M. Kazerani, "A centralized optimal energy management system for microgrids," 2011 IEEE Power and Energy Society General Meeting, Detroit, MI, USA, 2011, pp. 1-6.
25. X. Wu, X. Wang and Z. Bie, "Optimal generation scheduling of a microgrid," 2012 3rd IEEE PES Innovative Smart Grid Technologies Europe (ISGT Europe), Berlin, 2012, pp. 1-7.
26. C. Essayeh, M. R. El-Fenni and H. Dahmouni, "Optimal Energy Exchange in Micro-Grid Networks: Cooperative Game Approach," 2018 Renewable Energies, Power Systems & Green Inclusive Economy (REPS-GIE), Casablanca, 2018, pp. 1-6.
27. Chaouachi, Aymen & Kamel, Rashad & Andoulsi, Ridha & Nagasaka, Ken. (2013). Multiobjective Intelligent Energy Management for a Microgrid. Industrial Electronics, IEEE Transactions on. 60. 1688-1699. 10.1109/TIE.2012.2188873.
28. Mbuwir, B.V.; Ruelens, F.; Spiessens, F.; Deconinck, G. Battery Energy Management in a Microgrid Using Batch Reinforcement Learning. Energies 2017, 10, 1846.
29. Wu, Di & Kintner-Meyer, Michael & Yang, Tao & Balducci, Patrick. (2016). Economic Analysis and Optimal Sizing for behind-the-meter Battery Storage. 10.1109/PESGM.2016.7741210.
30. Z. Zhang, J. Wang and X. Cao, "Economic dispatch of microgrid considering optimal management of lithium batteries," 2014 International Conference on Power System Technology, Chengdu, 2014, pp. 3194-3199.
31. K. H. Ali, H. Hyder, and M. A. Khan, “Sensitivity Analysis of Reinforcement Learning to Schedule the Battery in Grid-tied Microgrid,” vol. 6, no. 3, 2022.
32. Sutton, R.S.; Barto, A.G. *Reinforcement Learning: An Introduction*, 2nd ed.; MIT Press: Cambridge, UK, 2018.
33. Mathworks. The Help documents of Optimization Toolbox included in MATLAB 2017b. www.mathworks.com. [Online] retrieved on 30,Dec 2019.

Author Biographies

Khawaja Haider Ali received the B.Eng in Electronics Engineering from Bahauddin Zakariyia University (BZU), Multan Pakistan in 2009. He Completed his MSc in Electrical Engg and Info technology from Otto Von Gurieke University Magdeburg, Germany, in 2012, and PhD degree in Renewable energy from University of Exeter UK in 2022. He had implemented the algorithm to optimise the grid-tied microgrid using AI (Reinforcement learning) by scheduling the storage system. During his PhD he had published 2 Journal impact factor Papers related to his PhD topic. Teaching and research have been part of his career for more than a decade. Currently he is serving as an Assistant Professor in Electrical Engineering department, Sukkur IBA University, Pakistan.

Mohammed Alharbi received the B.S. degree in electrical engineering from King Saud University, Riyadh, Saudi Arabia, in 2010, and the M.S. degree in electrical engineering from the Missouri University of Science and Technology, Rolla, MO, USA, in 2014, the Ph.D. degree in electrical engineering from the North Carolina State University, Raleigh, NC, USA, in 2020. He was a Teaching Assistant with King Saud University from September 2010 till May 2011. He was a project engineer at the Freedom Systems Center in the North Carolina State University, Raleigh, NC, USA from January 2016 till December 2019, where he was involved in designing and constructing a modular multi-level converter for control validations. In August 2020, he joined the Department of Electrical Engineering, King Saud University, Riyadh, Saudi Arabia, where he is currently an Assistant Professor. His research interests include medium voltage and high-power converters, modular multi-level converter (MMC) controls, multi-terminal HVdc systems, and grid integration of renewable energy systems.

Prof. Asif Tahir received his PhD degree in Inorganic Chemistry from Quaid-I-Azam University Islamabad Pakistan in 2009.Currently he is Associate Professor and Director of Research in Renewable Energy at Department of Engineering, University of Exeter (UoE). He has secured research funding (>10m) as PI and CoI for various research projects. He is specialized in the fabrication of nanomaterials using state-of-the-art techniques for solar energy conversion and building energy efficiency. His research focus on energy material design, green hydrogen production, electrochemical energy storage, thermal energy storage, device characterisation and optimisation of device for high performance. He has published 115 peer-reviewed research papers in high impact journals and one book chapter. His publication has received an overall citation of 5630 and his h-index is 38.