



The University of Tehran Press

Development of a Deep Reinforcement Learning Framework for Microgrid Energy Management under High Renewable Energy Penetration Scenarios

Mona Mirrazavi¹ | Hossein Yousefi^{2*} | Ahmad Hajinezhad³

1. M.Sc. Student in Energy Systems Engineering, Faculty of Energy Engineering and Sustainable Resources, College of Interdisciplinary Science and Technologies, University of Tehran, Tehran, Iran. Email: mona.mirrazavi@ut.ac.ir

2. Corresponding Author, Full Professor of Energy Systems Engineering, Faculty of Energy Engineering and Sustainable Resources, College of Interdisciplinary Science and Technologies, University of Tehran, Tehran, Iran. Email: hosseinyousefi@ut.ac.ir

3. Associate Professor of Energy Systems Engineering, Faculty of Energy Engineering and Sustainable Resources, College of Interdisciplinary Science and Technologies, University of Tehran, Tehran, Iran. Email: hajinezhad@ut.ac.ir

ARTICLE INFO

Article type:
Research Paper

Article History:
Received: 09 February 2026
Revised: 16 March 2026
Accepted: 12 May 2026
Published Online: 23 September 2026

Keywords:
Deep Reinforcement Learning,
Energy Management,
Microgrid,
Renewable Energy Sources.

ABSTRACT

With the increasing penetration of renewable energy resources in microgrids, intelligent energy management has become increasingly important due to the intermittent nature of solar and wind generation. This study presents a deep reinforcement learning-based energy management framework for a grid-connected microgrid consisting of photovoltaic panels, a wind turbine, a battery energy storage system, a diesel generator, and the utility grid. The main novelty of this work lies in combining Double Deep Q-Network (Double DQN) with Prioritized Experience Replay (PER) and designing a multi-objective reward function that simultaneously considers operating cost, carbon emissions, supply reliability, and battery state-of-charge management before peak-price hours. This structure enables the agent to learn not only an economic operating policy but also decisions that adapt to the temporal patterns of renewable generation and electricity prices. The proposed model is evaluated over a 24-hour scheduling horizon after 2,000 training episodes. The results show satisfactory convergence, reducing the daily operating cost from approximately \$10 to \$3.44, corresponding to a 66% cost reduction. In addition, total carbon emissions decrease from nearly 41 kg CO₂ to 16.6 kg CO₂, representing about a 60% reduction. Moreover, load shedding is completely eliminated, and the demand is fully supplied throughout the day. Battery behavior analysis indicates that the agent charges the battery during periods of high renewable generation and discharges it during peak-price hours. The results confirm the effectiveness and scalability of the proposed method for energy management of renewable-rich microgrids.

Cite this article: Mirrazavi, M.; Yousefi, H. & Hajinezhad, A. (2026). Development of a Deep Reinforcement Learning Framework for Microgrid Energy Management under High Renewable Energy Penetration Scenarios. *Journal of Sustainable Energy Systems*, 5 (4), 763-786. DOI: <http://doi.org/10.22059/ses.2026.413067.1228>



© Authors retain the copyright and full publishing rights.
DOI: <http://doi.org/10.22059/ses.2026.413067.1228>

Publisher: University of Tehran Press.

Introduction

The increasing penetration of renewable energy resources in modern microgrids has created new challenges for energy management due to the intermittent nature of photovoltaic (PV) and wind generation. Efficient coordination of distributed energy resources is therefore essential to reduce operating cost, limit carbon emissions, and maintain reliable power supply. Conventional optimization methods often face difficulties when dealing with nonlinear operational constraints and time-varying uncertainties. In this study, a reinforcement learning (RL) based energy management strategy is developed for a grid-connected microgrid consisting of PV panels, a wind turbine, a battery energy

storage system (BESS), a diesel generator, and the utility grid. The main objective is to determine an intelligent operational policy that minimizes daily operating cost and environmental impact while ensuring uninterrupted load supply.

Materials and methods

The proposed microgrid is modeled over a 24-hour scheduling horizon with one-hour time steps. The system state includes load demand, renewable generation, electricity price, and battery state of charge (SOC). The control action is defined as a discrete set of battery charging, discharging, or idle decisions. A deterministic transition model updates the system state according to physical constraints such as battery capacity, charging/discharging limits, diesel generation limits, and grid import capability. A Double Deep Q-Network (Double DQN) is employed as the learning algorithm to avoid overestimation of action values and improve policy stability. Prioritized Experience Replay (PER) is incorporated to emphasize critical experiences such as peak-price hours, renewable shortages, and low SOC conditions. The reward function is designed to guide the agent toward economical and environmentally favorable behavior. It includes grid electricity cost, diesel operating cost, carbon-emission penalty, load shedding penalty, and an incentive term that encourages maintaining adequate SOC before high-price periods. The agent is trained for 2,000 episodes. During training, an epsilon-greedy exploration policy is used, where epsilon gradually decreases from 1.0 to 0.01. All input variables are normalized before entering the neural network.

Results

The learning process shows clear convergence. In the early episodes, cumulative rewards are strongly negative because of inefficient battery usage and excessive dependence on grid power. As training progresses, rewards stabilize at positive values, indicating successful policy learning. Daily operating cost decreases significantly from approximately \$10 in the initial episodes to about \$3.44 after convergence, corresponding to nearly 66% cost reduction. The learned policy also reduces total carbon emissions from about 41 kg CO₂ to approximately 16.6 kg CO₂, representing nearly 60% improvement. Most importantly, load shedding is completely eliminated. In the daily operational test, the battery begins the day with around 50% SOC, charges during midday when renewable generation is high, and discharges during evening peak-price hours. Grid import occurs only in limited intervals, while diesel generation is activated mainly during critical periods.

Discussion and Conclusion

The results demonstrate that the proposed RL framework successfully captures the temporal relationship among renewable generation, electricity price, and load demand. Rather than relying heavily on grid electricity, the agent learns to store renewable energy during low-cost or high-generation periods and use it strategically during peak-price hours. The integration of Prioritized Experience Replay accelerates convergence by emphasizing critical operational experiences, while Double DQN improves learning stability and reduces value overestimation. The learned operational behavior is consistent with practical microgrid requirements. Battery charging during midday and discharging during evening peak demand reduces dependence on expensive grid imports and limits diesel usage. This coordinated control strategy improves both economic and environmental performance while maintaining supply reliability. Overall, the proposed Double DQN-based energy management framework provides an effective and scalable solution for renewable-rich microgrids. Simulation results confirm significant reductions in operating cost and carbon emissions together with complete elimination of load shedding. These findings indicate that reinforcement learning can be a promising approach for real-time energy management of future smart microgrids operating under dynamic and uncertain conditions.



انتشارات دانشگاه تهران

فصلنامه سیستم‌های انرژی پایدار

شاپا الکترونیکی: ۸۶۹۳-۲۹۸۰

سایت نشریه: <https://ses.ut.ac.ir>

توسعه یک چهارچوب یادگیری تقویتی عمیق برای مدیریت انرژی ریز شبکه‌ها در سناریوی نفوذ بالای منابع تجدیدپذیر

مونا میررضوی^۱ | حسین یوسفی^{۲*} | احمد حاجی‌نژاد^۳

۱. دانشجوی کارشناسی ارشد مهندسی سیستم‌های انرژی، دانشکده مهندسی انرژی و منابع پایدار، دانشکده‌های میان‌رشته‌ای دانشگاه تهران، تهران، ایران. رایانامه: mona.mirrazavi@ut.ac.ir
۲. نویسنده مسئول، استاد تمام مهندسی سیستم‌های انرژی، دانشکده مهندسی انرژی و منابع پایدار، دانشکده‌های میان‌رشته‌ای دانشگاه تهران، تهران، ایران. رایانامه: hosseinyousefi@ut.ac.ir
۳. دانشیار مهندسی سیستم‌های انرژی، دانشکده مهندسی انرژی و منابع پایدار، دانشکده‌های میان‌رشته‌ای دانشگاه تهران، تهران، ایران. رایانامه: hajinezhad@ut.ac.ir

اطلاعات مقاله

چکیده

نوع مقاله:
پژوهشی

تاریخ‌های مقاله:

تاریخ دریافت: ۱۴۰۴/۱۱/۲۰

تاریخ بازنگری: ۱۴۰۴/۱۲/۱۵

تاریخ پذیرش: ۱۴۰۵/۰۲/۲۲

تاریخ انتشار: ۱۴۰۵/۰۷/۰۱

کلیدواژه:

ریز شبکه،
مدیریت انرژی،
منابع تجدیدپذیر،
یادگیری تقویتی عمیق.

با افزایش نفوذ منابع انرژی تجدیدپذیر در ریز شبکه‌ها، مدیریت هوشمند انرژی به دلیل ماهیت نوسانی تولیدات خورشیدی و بادی اهمیت فزاینده‌ای یافته است. در این پژوهش، یک چهارچوب مدیریت انرژی مبتنی بر یادگیری تقویتی عمیق برای ریز شبکه‌ای متصل به شبکه شامل پل خورشیدی، توربین بادی، سامانه ذخیره‌ساز باتری، دیزل ژنراتور و شبکه سراسری ارائه شده است. نوآوری اصلی پژوهش در ترکیب الگوریتم Double DQN با بافر تجربه اولویت‌دار^۲ و طراحی یک تابع پاداش چندهدفه است که به طور هم‌زمان هزینه بهره‌برداری، آلاینده‌گی، قابلیت اطمینان تأمین بار و مدیریت سطح شارژ باتری پیش از ساعت‌های اوج مصرف را در فرایند یادگیری لحاظ می‌کند. این ساختار به عامل اجازه می‌دهد علاوه بر یادگیری سیاست اقتصادی، تصمیم‌های خود را با الگوی زمانی تولید تجدیدپذیر و قیمت برق هماهنگ سازد. مدل پیشنهادی در افق بهره‌برداری ۲۴ ساعته و پس از ۲ هزار اپیزود آموزش ارزیابی شد. نتایج نشان داد الگوریتم به همگرایی مناسب دست یافته و هزینه کل بهره‌برداری روزانه را از حدود ۱۰ دلار به ۳/۴۴ دلار کاهش داده است که معادل ۶۶ درصد صرفه‌جویی اقتصادی است. همچنین، آلاینده‌گی کل سامانه از حدود ۴۱ کیلوگرم دی‌اکسیدکربن به ۱۶/۶ کیلوگرم دی‌اکسیدکربن رسید که نشان‌دهنده ۶۰ درصد کاهش انتشار است. علاوه بر این، بار در تمام ساعت‌های روز به طور کامل تأمین شد. تحلیل رفتار باتری نشان داد عامل در ساعت‌های تولید بالای تجدیدپذیر باتری را شارژ و در ساعت‌های اوج قیمت برق آن را دشارژ می‌کند. نتایج، کارایی و مقیاس‌پذیری روش پیشنهادی را برای مدیریت انرژی ریز شبکه‌های با نفوذ بالای منابع تجدیدپذیر تأیید می‌کند.

استناد: میررضوی، مونا؛ یوسفی، حسین و حاجی‌نژاد، احمد (۱۴۰۵). توسعه یک چهارچوب یادگیری تقویتی عمیق برای مدیریت انرژی ریز شبکه‌ها در سناریوی نفوذ بالای منابع تجدیدپذیر. فصلنامه سیستم‌های انرژی پایدار، ۵ (۴) ۷۶۳-۷۸۶.

DOI: <http://doi.org/10.22059/ses.2026.413067.1228>

ناشر: مؤسسه انتشارات دانشگاه تهران.

© نویسندگان.

DOI: <http://doi.org/10.22059/ses.2026.413067.1228>



1. Double Deep Q-Network
2. Prioritized Experience Replay Buffer

۱. مقدمه

به دلیل پیامدهای گسترده‌ی زیست‌محیطی ناشی از مصرف منابع فسیلی، نوسانات قیمتی و محدودیت منابع آن‌ها، توجه جهانی به سوی بهره‌گیری از انرژی‌های تجدیدپذیر به عنوان راهکاری پایدار و نهایی برای مقابله با این چالش‌ها جلب شده است. بر اساس آمار منتشرشده، نرخ رشد نصب سامانه‌های انرژی تجدیدپذیر از سال ۲۰۱۰ تا کنون به طور متوسط سالانه بین ۸ تا ۹ درصد افزایش داشته و تا سال ۲۰۲۰، ظرفیت تجمعی آن‌ها به حدود ۲/۹ تراوات رسیده است [۱]. مطالعات اخیر نشان داده‌اند بهره‌گیری از سامانه‌های مبتنی بر انرژی‌های تجدیدپذیر، علاوه بر مزایای زیست‌محیطی، از منظر فنی و اقتصادی نیز در بسیاری از مناطق دارای توجیه‌پذیری قابل قبولی است [۲]. با این حال، افزایش سهم انرژی‌های تجدیدپذیر متغیر (مانند باد و خورشید) در سیستم‌های قدرت سنتی به دلیل نوسانات تولید، عدم قابلیت پیش‌بینی و محدودیت ذخیره‌سازی با پیچیدگی‌های فنی روبه‌رو است [۳]. برای مقابله با این بحران‌ها، کشورهای توسعه‌یافته در حال انجام پژوهش‌ها و نوآوری‌های نوینی هستند که می‌تواند به تولید انرژی پاک‌تر، کارآمدتر و به‌صرفه‌تر منجر شود [۴]. در سال‌های اخیر، کاهش انتشار دی‌اکسیدکربن به یکی از محورهای اصلی سیاست‌گذاری در سامانه‌های انرژی تبدیل شده است و توسعه فناوری‌های نوین مدیریت انرژی در کنار راهکارهای کاهش و جذب کربن، نقشی تعیین‌کننده در حرکت به سمت سامانه‌های انرژی پایدار ایفا می‌کند [۵].

در این میان، میکروگریدها به عنوان یکی از فناوری‌های کلیدی برای یکپارچه‌سازی منابع مطرح شده‌اند. ریزشبکه‌ها از مجموعه‌ای از منابع تولید پراکنده^۱ تشکیل شده‌اند که ممکن است شامل منابع انرژی تجدیدپذیر مانند باد، خورشید و منابع غیرتجدیدپذیر مانند دیزل ژنراتور یا گاز طبیعی و یا مجموعه‌ای از باتری‌ها برای ذخیره انرژی اضافی باشند [۶]. منابع انرژی تجدیدپذیر دارای معایبی از جمله نوسانات، تصادفی بودن و قابل پیش‌بینی نبودن هستند و در ریزشبکه‌هایی که سهم بالایی از منابع تجدیدپذیر و تقاضای بار تصادفی دارند، نادیده گرفتن این عدم قطعیت‌ها ممکن نیست، زیرا این امر باعث می‌شود بهره‌برداری اقتصادی و پایدار از سیستم دشوار شود [۷]. در چنین ساختاری، وجود یک سیستم مدیریت انرژی برای استفاده‌ی بهینه، هوشمندانه، ایمن، قابل اعتماد و هماهنگ از منابع انرژی ضروری است. سیستم مدیریت انرژی^۲ هسته اصلی عملکرد هوشمند ریزشبکه‌ها است. اهداف کلیدی EMS عبارت‌اند از: بهینه‌سازی عملکرد، زمان‌بندی انرژی (میزان تولید، ذخیره‌سازی و تبادل) و افزایش قابلیت اطمینان سیستم تا هزینه، آلودگی و ریسک ناپایداری به حداقل برسد [۸]. با این حال، اغلب روش‌های سنتی مدیریت انرژی که بر پایه مدل‌های بهینه‌سازی ریاضی یا الگوریتم‌های فراابتکاری مانند NSGA-II^۳ و MOPSO^۴ طراحی شده‌اند انعطاف‌پذیری کافی برای شرایط پویا و نامطمئن شبکه را ندارند. این روش‌ها معمولاً نیازمند مدل‌سازی دقیق هستند و نمی‌توانند در محیط‌هایی با تغییرات سریع در تولید و مصرف عملکرد بالادرنگ داشته باشند.

راه حل برای چنین مشکلاتی در عصر حاضر، یکی از شاخه‌های هوش مصنوعی یعنی یادگیری ماشین است. یادگیری تقویتی یک تکنیک از یادگیری ماشین است که برای یادگیری تصمیم‌گیری‌های ترتیبی در مسائل پیچیده به کار می‌رود. این روش تلاش می‌کند تا با ایجاد تغییراتی در فرایند یادگیری، محدودیت‌های عملکردی موجود در روش‌های پیشین را برطرف کند [۹ و ۱۰]. هدف اصلی در RL این است که عامل بتواند اعمالی را انتخاب کند که به بیشترین پاداش ممکن در آینده منجر شوند [۱۱]. در چند سال گذشته، یادگیری تقویتی عمیق^۵ به طور فزاینده‌ای قدرتمند و مهم شده است، چرا که از مدل‌سازی شبکه‌های عصبی در الگوریتم‌های یادگیری تقویتی سنتی استفاده می‌کند. به طور خاص، DRL برای حل مسائل بهینه‌سازی تصمیم‌گیری می‌کند [۱۲]. یادگیری تقویتی عمیق در محیط‌های واقعی مانند ریزشبکه‌ها هماهنگی بین منابع انرژی، باتری‌ها و مصرف‌کنندگان را انجام می‌دهد و این کاربرد به عنوان یک چشم‌انداز برای پژوهش‌های آینده مطرح شده است [۱۳].

هدف این پژوهش، ارائه یک سیستم مدیریت انرژی چندهدفه مبتنی بر یادگیری تقویتی عمیق است که بتواند در شرایط واقعی و غیرقطعی، بین سه هدف اصلی یعنی هزینه، انتشار و قابلیت اطمینان تعادل برقرار کند. در این چهارچوب، عامل

1. Distributed Generators – DGs

2. Energy Management System

3. Non-dominated Sorting Genetic Algorithm II

4. Multi-Objective Particle Swarm Optimization

5. Deep Reinforcement Learning-DRL

یادگیرنده^۱ با دریافت داده‌های بلادرنگ از محیط و یادگیری مداوم، سیاست تصمیم‌گیری خود را به صورت پویا به‌روزرسانی می‌کند. نوآوری این پژوهش در ترکیب رویکرد یادگیری تقویتی با مفهوم بهینه‌سازی چندهدفه نهفته است؛ به این معنا که سیستم پیشنهادی قادر است ضمن یادگیری مستمر از بازخورد محیط، تصمیماتی اتخاذ کند که چندین هدف مانند کاهش مصرف انرژی، افزایش بازده عملکرد، کاهش هزینه‌های عملیاتی و بهبود پایداری سیستم را به صورت هم‌زمان مورد توجه قرار دهد. این ترکیب نوآورانه موجب می‌شود سیستم توانایی بالایی در انطباق با شرایط پویا و غیرقطعی محیطی (نظیر تغییرات اقلیمی، نوسانات بار، یا اختلالات منابع انرژی تجدیدپذیر) داشته باشد و در نتیجه، عملکرد کلی آن به طور چشمگیری بهبود یابد. افزون بر این، بهره‌گیری از الگوریتم‌های یادگیری تقویتی امکان یادگیری خودکار سیاست‌های تصمیم‌گیری بهینه طی زمان را فراهم می‌کند، بدون آنکه نیاز به مدل‌سازی دقیق یا نظارت انسانی مداوم وجود داشته باشد. این پژوهش تأثیر بسزایی در گسترش راه حل‌های بهینه‌سازی هوشمند در مدیریت انرژی ریزشکه‌ها دارد.

۲. مرور ادبیات

در دهه اخیر، پژوهش‌های گسترده‌ای در زمینه مدیریت انرژی در ریزشکه‌ها انجام شده است. هدف اصلی این مطالعات، دستیابی به بهره‌برداری اقتصادی، پایدار و قابل اعتماد از منابع تولید پراکنده در شرایط نامطمئن و متغیر محیطی بوده است. نگاه به ریزشکه‌ها از منظر سامانه‌های انرژی یکپارچه به گونه‌ای تقویت شده که هماهنگ‌سازی منابع تولید، ذخیره‌سازی و مصرف در قالب ساختارهای انرژی هاب، به یکی از رویکردهای مهم در طراحی و بهره‌برداری هوشمند تبدیل شده است [۱۴]. روش‌های متنوعی برای حل این مسئله پیشنهاد شده است که این روش‌ها را بر اساس الگوریتم‌های بهینه‌سازی می‌توان به صورت کلی به ۳ دسته زیر تقسیم کرد.

دسته اول شامل روش‌های مبتنی بر الگوریتم‌های فراابتکاری است که با استفاده از الگوریتم‌های الهام‌گرفته از طبیعت، مسائل پیچیده و چندهدفه مدیریت انرژی را حل می‌کنند. برای مثال، در پژوهش Ali و Sharma [۱۵] الگوریتم MOGWO^۲ برای مدیریت انرژی ریزشکه‌های مجهز به خودروهای برقی به کار گرفته شد و پژوهشگران یادشده توانستند هزینه‌ها را بیش از ۲۰ درصد کاهش دهند. در پژوهش Yousri و همکاران [۱۶] MOEMPA^۳ برای مدیریت اشتراک انرژی در ریزشکه‌ای متصل به شبکه توسط نویسندگان به کار رفت و هزینه بهره‌برداری را در سه سناریو به ترتیب ۱۶، ۳۴/۵ و ۲۶/۹ درصد و میزان آلاینده‌ها را ۱۳/۳۷ و ۱۴/۷ درصد کاهش داد. در پژوهش Vijaykumar [۱۷] ترکیب الگوریتم SOA^۴ با شبکه عصبی مصنوعی توسط نویسندگان برای ردیابی نقطه حداکثر توان به راندمان ۹۴/۹۹ درصد و راندمان ردیابی ۹۶/۷۲ درصد منجر شد. در پژوهش Mei و همکاران [۱۸] الگوریتم ASAPSO^۵ موجب بهبود هم‌زمان عملکرد اقتصادی و زیست‌محیطی شد. همچنین، در پژوهش Babu و همکاران [۱۹] الگوریتم ژنتیک همراه با پیش‌بینی داده‌ها هزینه خرید برق از شبکه را ۴۱/۴۶ درصد و هزینه فرسایش باتری را ۵۰/۹۸ درصد کاهش داد. در پژوهش Aeggegn و همکاران [۲۰] الگوریتم MOMVO^۶ در یک سیستم شامل چهار ریزشکه متصل، هزینه عملیاتی روزانه را ۴۰/۳۲ درصد کاهش داد و مقدار LPSP^۷ برابر ۰/۱۶ درصد به دست آمد. در پژوهش Osama و همکاران [۲۱] روش MOMRFO^۸ بیشترین صرفه‌جویی هزینه عملیاتی یعنی ۸۱/۸۵۹ درصد را حاصل کرد. در پژوهش Rochd و همکاران [۲۲] نسخه بهبودیافته PSO^۹ هزینه خالص انرژی را ۴۲ درصد کاهش داد و در پژوهش Lamari و همکاران [۲۳] الگوریتم MOFEPso^{۱۰} توانست هزینه کل را ۷۰/۳۵ درصد و انتشار آلاینده‌ها را ۳۱/۹۱ درصد کاهش دهد. در

1. DRL agent

2. Multi-Objective Grey Wolf Optimizer

3. Multi-Objective Marine Predator Algorithm

4. Combined Seagull Optimization Algorithm

5. Adaptive Simulated Annealing Particle Swarm Optimization

6. Multi-Objective Multi-Verse Optimizer

7. Loss of Power Supply Probability

8. Multi-Objective Manta Ray Foraging Optimization

9. Particle Swarm Optimization

10. Improved Multi-Objective Particle Swarm Optimization with Feasibility Evaluation Capability

پژوهش Ali و همکاران [۲۴] نیز نویسندگان با استفاده از PSO نشان دادند تغییرات نرخ تنزیل ($0.03/\%$) و قیمت برق ($4/5$ درصد) اثر قابل قبولی بر نتایج دارند. در پژوهش Kamarposhti و همکاران [۲۵] از NSGA-II برای کاهش هم‌زمان هزینه و آلودگی استفاده شد. در پژوهش Momen و همکاران [۲۶] الگوریتم MSFL^۱ در صورت تمرکز بر کاهش تلفات، هزینه‌ها را $25/21$ درصد و با اجرای برنامه‌های پاسخگویی بار تا $31/56$ درصد کاهش داد. همچنین در پژوهش Wang و همکاران [۲۷] نیز یک مدل MOCO^۲ باعث کاهش $6/12$ درصد آلودگی و $11/91$ درصد هزینه‌های عملیاتی شد.

دسته دوم شامل مدل‌های برنامه‌ریزی و بهینه‌سازی ریاضی است که از روش‌هایی مانند MILP^۳، MINLP^۴، تجزیه بندرز و برنامه‌ریزی مقاوم برای بهینه‌سازی ریزشکه‌ها استفاده می‌کنند. در پژوهش Wei و همکاران [۲۸] نویسندگان یک مدل دوسطحی مبتنی بر تجزیه بندرز ارائه کردند که ضمن حفظ دقت، پیچیدگی و زمان حل را به طور محسوسی کاهش داد. در پژوهش Ahmadi و همکاران [۲۹] یک مدل MINLP توانست 22 درصد صرفه‌جویی بیشتر در هزینه سالانه برق و 37 درصد کاهش انتشار دی‌اکسیدکربن ایجاد کند. در پژوهش Niknami و همکاران [۳۰] مدل مقاوم روز پیش با الگوریتم C&CG^۵ حداقل 8 درصد هزینه عملیاتی را نسبت به روش‌های مرجع کاهش داد. در پژوهش Veisi و همکاران [۳۱] نیز راهبرد دوسطحی مبتنی بر شرایط KKT^۶ و برنامه‌ریزی تصادفی - مقاوم، تلفات انرژی را حدود 30 درصد کاهش داد و شاخص‌های قابلیت اطمینان و پایداری ولتاژ را بهبود بخشید.

دسته سوم شامل مدل‌های مبتنی بر یادگیری ماشین و یادگیری تقویتی است که هدف آن‌ها تصمیم‌گیری بلادرنگ و سازگار با عدم قطعیت‌ها است. در پژوهش Hosseini و همکاران [۳۲] روش RL-MOF-EMS^۷ هزینه عملیاتی را از 0.767 h/€ به 0.704 h/€ و هزینه گرمایش را از $1/484$ h/€ به $1/327$ h/€ کاهش داد، در حالی که نرخ استفاده از هیدروژن از $63/15$ درصد به $71/43$ درصد افزایش یافت. در پژوهش Cui و همکاران [۳۳] یک روش یادگیری تقویتی بلادرنگ نرخ خودتوازی سیستم را $22/97$ درصد و سود اپراتور را $33/74$ درصد افزایش داد. در پژوهش Alabdullah و Abido [۳۴] الگوریتم DQN با هزینه‌ای نزدیک به بهینه، بیش از 200 برابر سریع‌تر از حل‌کننده MILP عمل کرد. در پژوهش Zheng و همکاران [۳۵] الگوریتم MOATD3^۸ پایداری پاسخ سیستم را تا $31/4$ درصد افزایش و افت عمر باتری را $46/7$ درصد کاهش داد. در پژوهش Yao و همکاران [۳۶] نیز نویسندگان نشان دادند روش یادگیری تقویتی عمیق پیشنهادی در ریزشکه آمریکا، پاداش سالانه را در حالت پیش‌بینی دقیق تا $139/33$ درصد و در حضور نویز تا $125/45$ درصد نسبت به مدل پایه افزایش داده است. تفاوت و ویژگی‌های پژوهش‌های بالا را می‌توان در جدول ۱ خلاصه کرد.

جدول ۱. خلاصه اهداف و شکاف‌ها در مرور ادبیات

ردیف	دسته روش‌ها	نمونه الگوریتم‌ها	اهداف اصلی	شکاف‌ها
۱	فرابتناری	MOGWO, MOEMPA, PSO, GA, NSGA-II, MOMVO...	کاهش هزینه و آلودگی، افزایش بهره‌وری انرژی	نیازمند تکرار زیاد، وابسته به داده ایستا، واکنش کند در شرایط پویا، تمرکز بیشتر بر هزینه تا آلودگی، عدم پاسخ بلادرنگ به نوسانات
۲	برنامه‌ریزی کلاسیک	بندرز، MILP, MINLP، برنامه‌ریزی مقاوم و تصادفی	کاهش هزینه و افزایش قابلیت اطمینان	دقیق اما سنگین و زمان‌بر، نیاز به حل مجدد با تغییر شرایط، نیاز به مدل دقیق انتشار، تعادل سخت بین هزینه و آلوده‌ها، پایداری خوب ولی انعطاف‌پذیری پایین در عدم قطعیت‌ها
۳	یادگیری ماشین و تقویتی	RL, DRL, DQN, MOATD3	تصمیم‌گیری بلادرنگ، کاهش هزینه و آلودگی، افزایش پایداری	یادگیری بلادرنگ، تصمیم اقتصادی سریع، یادگیری تطبیقی، حفظ پایداری در تغییرات شدید، عدم بررسی هم‌زمان سه هدف کاهش آلودگی، کاهش هزینه و افزایش پایداری

1. Modified Shuffled Frog Leaping
2. multi-objective optimization
3. Mixed-Integer Linear Programming
4. Mixed-Integer Nonlinear Programming
5. Cut-and-Column Generation
6. Karush-Kuhn-Tucker conditions
7. Reinforcement Learning-Multi-Objective Function-Energy Management System
8. Multi-Objective Deep Reinforcement Learning Algorithm based on TD3

با وجود پیشرفت‌های قابل توجه در حوزه مدیریت انرژی ریزشکه‌ها، مرور مطالعات اخیر نشان می‌دهد هنوز چند شکاف پژوهشی مهم باقی مانده است. نخست آنکه بخش قابل توجهی از روش‌های فراابتکاری و برنامه‌ریزی کلاسیک، اگرچه در کاهش هزینه یا آلاینده‌گی عملکرد مناسبی دارند، اما عمدتاً ماهیت آفلاین و ایستا دارند و برای تصمیم‌گیری بلادرنگ در محیط‌های دارای عدم قطعیت و تغییرات سریع بار و تولید تجدیدپذیر مناسب نیستند. دوم آنکه در بسیاری از مطالعات مبتنی بر یادگیری تقویتی، تمرکز اصلی بر کمینه‌سازی هزینه بهره‌برداری بوده و معمولاً سه شاخص کلیدی هزینه، آلاینده‌گی و قابلیت اطمینان به صورت هم‌زمان در قالب یک چارچوب یکپارچه مورد توجه قرار نگرفته‌اند. افزون بر این، در پژوهش‌های موجود کمتر به رفتار پیش‌نگرانه باتری و مدیریت سطح شارژ پیش از ساعات اوج قیمت برق پرداخته شده است؛ در حالی که این عامل می‌تواند نقش تعیین‌کننده‌ای در عملکرد اقتصادی و پایداری ریزشکه داشته باشد.

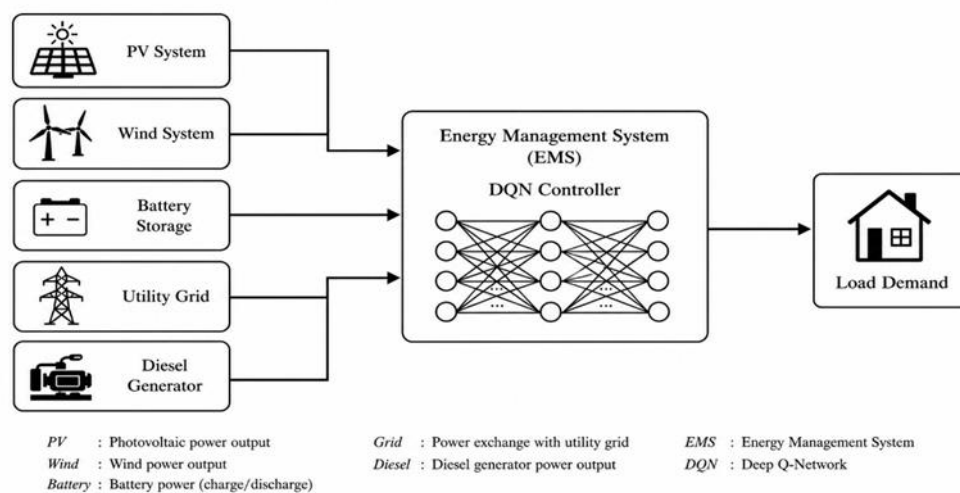
بر این اساس، نوآوری اصلی پژوهش حاضر در توسعه یک چارچوب یادگیری تقویتی عمیق چندهدفه برای مدیریت انرژی ریزشکه‌های با نفوذ بالای منابع تجدیدپذیر است که در آن از ترکیب Double DQN و PER استفاده شده است. همچنین، یک تابع پاداش چندمعیاره طراحی شده که به صورت هم‌زمان هزینه بهره‌برداری، انتشار آلاینده‌ها، قابلیت اطمینان تأمین بار و مدیریت سطح شارژ باتری پیش از ساعات‌های اوج مصرف را در فرایند یادگیری لحاظ می‌کند. این ساختار به عامل اجازه می‌دهد تا بدون نیاز به مدل‌سازی دقیق و حل مکرر مسئله بهینه‌سازی، سیاستی تطبیقی، پایدار و قابل اجرا در زمان واقعی بیاموزد. از این رو، پژوهش حاضر تلاش می‌کند خلأ موجود در ادبیات را با ارائه رویکردی یکپارچه برای بهینه‌سازی هم‌زمان عملکرد اقتصادی، زیست‌محیطی و قابلیت اطمینان ریزشکه پوشش دهد.

۳. مواد و روش‌ها

سامانه مورد مطالعه، یک میکروگرید هیبریدی متناوب پیوسته است که شامل مجموعه‌ای از منابع تولید تجدیدپذیر، واحدهای تأمین توان پشتیبان و تجهیزات ذخیره‌سازی انرژی است. این میکروگرید در سطح شبکه توزیع محلی قرار گرفته و با اهدافی نظیر تأمین مطمئن بار الکتریکی روزانه، کاهش هزینه بهره‌برداری، افزایش قابلیت اطمینان و کاهش آلاینده‌گی زیست‌محیطی طراحی شده است. این سامانه از شش جزء اصلی تشکیل می‌شود: (۱) واحد تولید فتوولتائیک؛ (۲) واحد تولید بادی؛ (۳) سیستم ذخیره‌سازی انرژی باتری؛ (۴) واحد دیزل ژنراتور؛ (۵) شبکه برق سراسری؛ (۶) بار مصرفی. در هر گام زمانی، توان تولیدی منابع مختلف همراه با توان باتری، دیزل و شبکه برای تأمین بار ترکیب می‌شود.

منابع PV و بادی به عنوان منابع ارزان اما غیرقابل پیش‌بینی در نظر گرفته می‌شوند. به منظور جبران نوسانات و عدم قطعیت منابع تجدیدپذیر، از باتری به عنوان ابزار ذخیره‌سازی و مدیریت هوشمند انرژی استفاده شده است. به دلیل محدودیت‌های عملیاتی و هزینه باتری، تصمیم‌گیری در مورد زمان‌های مناسب شارژ، دشارژ یا حالت بی‌عمل اهمیت ویژه‌ای دارد. دیزل ژنراتور نیز برای افزایش قابلیت اطمینان سامانه در شرایط کمبود تولید تجدیدپذیر و پایین بودن SOC^۱ باتری به عنوان منبع پشتیبان به کار گرفته می‌شود؛ با این حال به دلیل هزینه سوخت و آلاینده‌گی بالا لازم است تنها در شرایط ضروری مورد استفاده قرار گیرد. شبکه برق سراسری به عنوان یک منبع قابل اعتماد با تعرفه زمان متغیر^۲ مدل شده است. در ساعات‌های اوج مصرف، قیمت انرژی شبکه بالاتر بوده و واردات انرژی از شبکه موجب افزایش هزینه بهره‌برداری می‌شود. بار الکتریکی به صورت یک پروفایل روزانه مشخص در نظر گرفته شده است که باید تا حد امکان به طور کامل تأمین شود. این پروفایل تعیین می‌کند عامل یادگیرنده چگونه منابع مختلف را مدیریت کند تا ضمن تأمین بار، هزینه و آلاینده‌گی سامانه به حداقل برسد. در بخش‌های بعد، مدل ریاضی هر یک از اجزا و نحوه تبدیل مسئله مدیریت انرژی به یک محیط یادگیری تقویتی تشریح می‌شود. ساختار کلی میکروگرید و ارتباط بین منابع در شکل ۱ نشان داده شده است.

1. State of Charge
2. Time of Use Pricing



شکل ۱. نمودار شماتیک سیستم مدیریت انرژی پیشنهادی برای یک ریزشبه مبتنی بر شبکه Q عمیق

۳-۱. توضیحات مدل

میکروگرید مورد بررسی، شامل منابع انرژی تجدیدپذیر (فتوولتائیک و بادی)، یک واحد دیزل ژنراتور، یک باتری ذخیره‌ساز انرژی و شبکه برق سراسری است.

۳-۱-۱. مدل‌سازی سامانه میکروگرید

در این سامانه، جریان توان در هر گام زمانی (یک ساعت) با توجه به تراز توان در معادله ۱ برقرار می‌شود [۳۷].

$$P_{\text{load}}(t) = P_{\text{PV}}(t) + P_{\text{wind}}(t) + P_{\text{batt}}(t) + P_{\text{diesel}}(t) + P_{\text{grid}}(t) + LS(t) \quad (1)$$

که در آن $P_{\text{load}}(t)$ توان بار مصرفی، $P_{\text{PV}}(t)$ توان تولیدی پنل خورشیدی، $P_{\text{wind}}(t)$ توان تولیدی توربین بادی، $P_{\text{batt}}(t)$ توان ذخیره در باتری، $P_{\text{diesel}}(t)$ توان تولیدی دیزل، $P_{\text{grid}}(t)$ توان شبکه برق است.

۳-۱-۲. مدل پنل خورشیدی و توربین بادی

توان تولیدی PV و Wind به صورت داده ورودی (پروفایل روزانه) در معادلات ۲ و ۳ در نظر گرفته شده و عدم قطعیت آن‌ها در فرایند یادگیری لحاظ می‌شود. برای لحاظ کردن عدم قطعیت تولید تجدیدپذیر، روی پروفایل‌های PV/Wind نویز تصادفی کوچک اعمال می‌شود تا عامل یادگیرنده رفتار نوسانی تولید را تجربه کند [۳۷].

$$E_{\text{PV}}^{\text{total}} = \sum_t P_{\text{PV}}(t) \Delta t \quad (2)$$

$$E_{\text{Wind}}^{\text{total}} = \sum_t P_{\text{Wind}}(t) \Delta t \quad (3)$$

۳-۱-۳. مدل باتری

باتری مدل‌شده یک باتری لیتیوم یون ساده در نظر گرفته شده است. پارامترهای باتری شامل ظرفیت کل، محدوده مجاز SOC، نرخ شارژ، دشارژ و بازده انرژی در معادلات ۴ تا ۶ لحاظ می‌شود [۳۸].

$$SOC(t+1) = SOC(t) + \frac{\eta_c P_{\text{ch}}(t) - \frac{1}{\eta_d} P_{\text{dis}}(t)}{E_{\text{cap}}} \quad (4)$$

که در آن E_{cap} ظرفیت کل باتری، $SOC(t)$ سطح شارژ باتری در زمان t ، $P_{\text{ch}}(t)$ نرخ شارژ، $P_{\text{dis}}(t)$ نرخ دشارژ و η_d بازده انرژی است.

$$SOC_{\text{min}} \leq SOC(t) \leq SOC_{\text{max}} \quad (5)$$

$$0 \leq P_{ch}(t) \leq P_{bat}^{max} \quad 0 \leq P_{dis}(t) \leq P_{bat}^{max} \quad (6)$$

۳-۱-۴. مدل دیزل ژنراتور

در این مدل‌سازی، هزینه روشن و خاموش کردن دیزل، نرخ تغییر توان و محدودیت تعداد سیکل‌ها لحاظ نشده و دیزل در هر ساعت می‌تواند در محدوده توان نامی خود کار کند. توان خروجی دیزل طبق رابطه ۷ در بازه مجاز قرار دارد [۳۹].

$$P_{DG}^{min} \leq P_{DG}(t) \leq P_{DG}^{max} \quad (7)$$

میزان انرژی تولیدی دیزل به صورت معادله ۸ و آلاینده‌گی آن نیز به صورت معادله ۹، خطی مدل شده است [۴۰].

$$E_{DG} = \sum_t P_{DG}(t) \quad (8)$$

$$E_{DG}(t) = \kappa \cdot P_{DG}(t) \quad (9)$$

که در آن κ_{DG} ضریب انتشار دیزل و $P_{DG}(t)$ توان تولیدی واحد دیزل ژنراتور است.

۳-۱-۵. شبکه برق

قیمت انرژی خریداری‌شده از شبکه در ساعت‌های مختلف متفاوت است و یک پروفایل قیمت زمان متغیر برای آن در نظر گرفته شده است. شبکه تنها نقش واردکننده انرژی را دارد (تبادل برگشتی و فروش لحاظ نشده). انرژی وارداتی از شبکه به صورت معادله ۱۰ و محدودیت آن به صورت معادله ۱۱ مدل شده است [۳۷].

$$E_{Grid}(t) = P_{Grid}(t) \Delta t \quad (10)$$

$$|P_{grid}(t)| \leq P_{grid,max} \quad (11)$$

آلاینده‌گی شبکه در معادله ۱۲ آمده است [۴۰].

$$E_{Grid}^{CO_2}(t) = \kappa_{Grid} \cdot P_{Grid}(t) \quad (12)$$

که در آن κ_{grid} ضریب انتشار متوسط شبکه و $P_{grid}(t)$ توان وارداتی از شبکه است.

۳-۲. تعریف مسئله به صورت یک محیط یادگیری تقویتی

در این پژوهش، مسئله مدیریت انرژی میکروگرید به صورت یک مسئله یادگیری تقویتی مدل شده است. هدف عامل یادگیرنده، یافتن یک سیاست کنترلی است که با تصمیم‌گیری در هر گام زمانی، موجب حداقل‌سازی هزینه بهره‌برداری، کاهش آلاینده‌گی و جلوگیری از عدم تأمین بار شود. برای این منظور، ساختار محیط RL شامل سه مؤلفه اصلی فضای حالت، فضای عمل و تابع گذار تعریف شده است.

۳-۲-۱. فضای حالت

حالت سامانه در هر ساعت از روز با مجموعه‌ای از متغیرهای فیزیکی و اقتصادی تعیین می‌شود. این متغیرها در قالب یک بردار حالت در رابطه ۱۳ بر اساس ادبیات RL انرژی توسعه داده شده به عامل ارائه می‌شوند.

$$s_t = [SOC(t), P_{Load}(t), P_{PV}(t), P_{Wind}(t), Price(t), Hour(t)] \quad (13)$$

این ساختار حالت باعث می‌شود عامل بتواند رفتار سیستم را در آینده پیش‌بینی کرده و تصمیمات مؤثر بر هزینه و آلاینده‌گی را اتخاذ کند.

۳-۲-۲. فضای عمل

فضای عمل یک مجموعه گسسته است که تصمیم عامل درباره وضعیت باتری را نشان می‌دهد. در هر گام زمانی، عامل می‌تواند یکی از عمل‌های شارژ با نرخ کم، شارژ با نرخ زیاد، دشارژ با نرخ کم، دشارژ با نرخ زیاد و حالت بی‌عمل را اتخاذ کند. این طراحی باعث می‌شود عامل کنترلی ساده، اما مؤثر داشته باشد و بتواند رفتار باتری را هوشمندانه تنظیم کند. تمام محدودیت‌های عملیاتی SOC، نرخ شارژ / دشارژ و ظرفیت توان در مدل محیط اعمال شده‌اند.

۳-۲-۳. تابع گذار

تابع گذار سیستم، وضعیت جدید s_{t+1} را با توجه به وضعیت فعلی، اقدام عامل و قیود فیزیکی تعیین می‌کند. این شامل به‌روزرسانی SOC باتری، محاسبه تراز توان میان منابع مختلف، تأمین بار یا ثبت عدم تأمین، محاسبه انرژی وارداتی از شبکه یا توان مورد نیاز دیزل، تابع گذار به صورت قطعی است و به شکل مستقیم از مدل فیزیکی میکروگرید به دست می‌آید.

۳-۳. طراحی تابع پاداش^۱

تابع پاداش به منظور هدایت عامل به سمت عملکرد بهینه طراحی شده و اجزای آن در معادله ۱۴ آورده شده است. تابع پاداش بر اساس ساختارهای رایج در مدیریت انرژی مبتنی بر یادگیری تقویتی توسعه داده شده است.

$$R = -C_{grid} - C_{diesel} - \lambda_{uns} \cdot P_{uns} - \lambda_{emi} \cdot E(t) + \lambda_{soc} \cdot f(SOC) \quad (14)$$

که در آن C_{grid} هزینه توان خریداری شده از شبکه، C_{diesel} هزینه تولید دیزل، P_{uns} توان تأمین نشده، $E(t)$ آلاینده‌ی دیزل، $f(SOC)$ تابع تشویقی برای نگه داشتن SOC در بازه مناسب جهت عبور از ساعت‌های پیک است. تابع تشویق SOC میانی نیز به صورت معادله ۱۵ تعریف شده است.

$$f(SOC) = -|SOC(t) - SOC_{ref}| \quad (15)$$

که در آن SOC_{ref} سطح مطلوب شارژ و $OC(t)$ سطح شارژ فعلی است. همچنین، وزن‌ها به گونه‌ای تنظیم شده‌اند که عدم تأمین بار بسیار شدید جریمه شود، واردات در ساعت‌های اوج قیمت جریمه شود و هم‌زمان عامل به استفاده هوشمندانه از باتری تشویق شود.

۳-۴. الگوریتم Double DQN

معادله ۱۶ بلمن Q^* در رابطه ۱۶ آمده است و نشان می‌دهد ارزش انجام عمل a در حالت s برابر است با مجموع پاداش فوری R_t و پاداش آینده به همراه ضریب تخفیف [۴۱].

$$Q^*(s, a) = \mathbb{E} \left[R_t + \gamma \max_{a'} Q^*(s_{t+1}, a') \right] \quad (16)$$

الگوریتم Double DQN نسخه بهبودیافته روش کلاسیک DQN است که برای رفع مشکل بیش‌برآوردی^۲ مقادیر Q پیشنهاد شده است. در روش DQN اصلی، مقدار هدف با استفاده از یک شبکه واحد محاسبه می‌شود، در نتیجه عمل انتخاب اکشن و ارزیابی آن توسط یک مدل انجام شده و این موضوع باعث می‌شود مقدار Q برای برخی اعمال به صورت نادرست بزرگ‌تر برآورد شود. در Double DQN این مشکل با تفکیک دو نقش انتخاب اکشن و ارزیابی اکشن حل می‌شود [۴۲]. به این ترتیب انتخاب اکشن با شبکه اصلی^۳ در معادله ۱۷ و محاسبه مقدار هدف با شبکه هدف^۴ در معادله ۱۸ آمده است [۴۳].

$$a^* = \arg \max_a Q_{\theta}(s_{t+1}, a) \quad (17)$$

$$y = r_t + \gamma Q_{\theta^-}(s_{t+1}, a^*) \quad (18)$$

که در این رابطه θ پارامترهای شبکه اصلی، θ^- پارامترهای شبکه هدف، γ ضریب تخفیف و r_t پاداش گام زمانی t است.

۳-۴-۱. دلیل استفاده در مدیریت انرژی میکروگرید

در سامانه مدیریت انرژی، رفتارهای نامنظم و نوسانی منابع تجدیدپذیر باعث می‌شود الگوریتم‌های RL نسبت به خطاهای برآورد حساس باشند. با استفاده از Double DQN پایداری یادگیری افزایش می‌یابد، از انتخاب اشتباه اکشن‌های با Q اغراق‌شده جلوگیری می‌شود و عامل به یک سیاست اقتصادی و پایدار همگرا می‌شود. به دلیل این ویژگی‌ها، Double DQN یک انتخاب ایده‌آل برای مسائل EMS است که نیازمند سیاست‌های دقیق در حضور عدم قطعیت هستند [۴۳].

1. Reward Function
2. Overestimation
3. Online Network
4. Target Network

۳-۵. بافر تجربه اولویت‌دار^۱

در الگوریتم‌های یادگیری تقویتی، عامل در هر گام زمانی یک تجربه از نوع (s_t, a_t, r_t, s_{t+1}) تولید می‌کند. در روش‌های سنتی، تمام تجربه‌ها با احتمال برابر انتخاب می‌شوند. با این حال، بعضی از تجربه‌ها اطلاعات بسیار ارزشمندتری برای یادگیری دارند؛ مخصوصاً تجربه‌هایی که در آن‌ها عامل اشتباه بزرگی مرتکب شده و TD error بالایی دارد. برای رفع این مشکل از بافر تجربه اولویت‌دار استفاده می‌شود. از جمله مزایای استفاده از PER در پروژه حاضر می‌توان به افزایش سرعت همگرایی، کاهش نیاز به اپیزودهای زیاد، تمرکز بیشتر یادگیری روی تجربه‌های بحرانی مانند کمبود تولید تجدیدپذیر، ساعت‌های پیک قیمت برق و کاهش شدید SOC، بهبود پایداری و کیفیت سیاست نهایی اشاره کرد. PER به‌خصوص در مدیریت انرژی میکروگرید بسیار کارآمد است، زیرا رخدادهای نادر اما مهم مانند کمبود شدید تولید PV یا افزایش بار اهمیت بالایی در یادگیری دارند. ایده PER بر این اساس است که تجربه‌های مهم‌تر با احتمال بیشتری انتخاب شوند. به هر تجربه یک مقدار اولویت نسبت داده می‌شود و رابطه کمی آن در معادله ۱۹ آمده است [۴۴].

$$p_i = (|TD\ error| + \epsilon)^\alpha \quad (19)$$

که α میزان تأثیر اولویت و TD error تجربه‌هایی است که عامل اشتباه بزرگی مرتکب شده است. PER احتمال انتخاب تجربه را نیز بالا می‌برد. رابطه ۲۰ باعث می‌شود تجربه‌هایی که TD error بزرگ‌تری دارند، بیشتر در فرایند یادگیری ظاهر شوند [۴۴].

$$P(i) = \frac{p_i}{\sum_k p_k} \quad (20)$$

که در آن $P(i)$ احتمال انتخاب تجربه i و p_k احتمال انتخاب تجربه k است.

۳-۵-۱. وزن اهمیت^۲

برای جلوگیری از بایاس ناشی از انتخاب نامتقارن، وزن‌های اصلاحی در معادله ۲۱ آورده شده است [۴۴].

$$w_i = \left(\frac{1}{N \cdot P(i)} \right)^\beta \quad (21)$$

که در آن β از ۰/۴ شروع شده و به تدریج تا ۱/۰ افزایش یافته و N نیز تعداد کل تجربه‌ها در حافظه است. پس از هر مرحله یادگیری، مقادیر TD error جدید محاسبه شده و اولویت تجربه‌های انتخاب‌شده به‌روزرسانی می‌شود.

۳-۶. روند آموزش عامل^۳

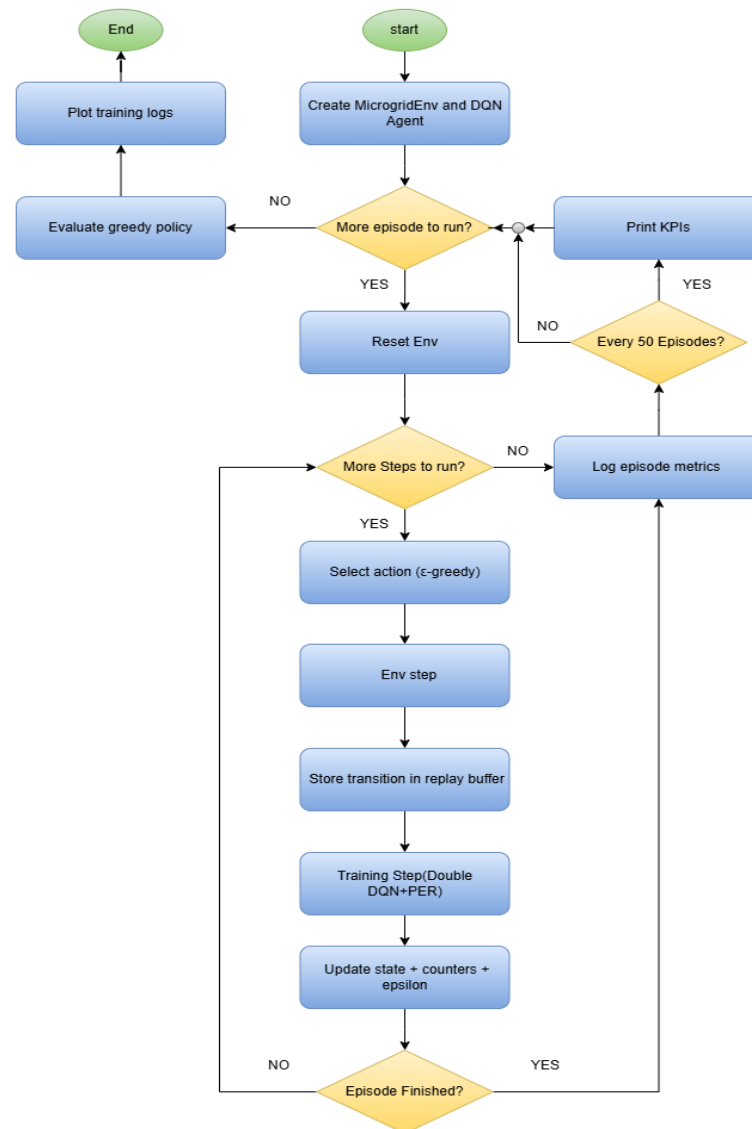
فرایند آموزش شامل تکرار چند هزار اپیزود است که در هر اپیزود، عامل ۲۴ گام زمانی (ساعت‌های روز) را طی می‌کند. تمام ورودی‌های شبکه (بار، تولید، قیمت، SOC) قبل از ورود به شبکه عصبی در بازه $[0, 1]$ نرمال‌سازی شده‌اند. مراحل اصلی آموزش عبارت از چندین مرحله است. نخست در انتخاب اکشن با سیاست ϵ -greedy عامل ابتدا با احتمال ϵ اقدام به کاوش کرده و سپس به تدریج از دانش یادگرفته‌شده استفاده می‌کند. مقدار ϵ از ۱/۰ تا ۰/۰۱ طی ۱۲ هزار گام کاهش می‌یابد. فرمول کاهش اپسیلون^۴ در معادله ۲۲ آمده است [۴۵].

$$\epsilon_k = \max \left(\epsilon_{\text{end}}, \epsilon_{\text{start}} - \frac{k}{N_{\text{decay}}} (\epsilon_{\text{start}} - \epsilon_{\text{end}}) \right) \quad (22)$$

سپس با انتخاب اکشن، حالت جدید و مقدار پاداش محاسبه می‌شود و تجربه همراه با مقدار TD-error در بافر ذخیره می‌شود. در مرحله بعدی با استفاده از PER، تجربه‌های مهم‌تر برای به‌روزرسانی شبکه انتخاب شده و مقدار هدف با استفاده از دو شبکه اصلی و هدف محاسبه می‌شود تا مشکل بیش‌برآوردی کاهش یابد. در هر ۵۰۰ مرحله پارامترهای شبکه هدف با شبکه اصلی

1. Prioritized Experience Replay
2. Importance Sampling Weight
3. Training Procedure
4. ϵ -greedy decay

همانگ شده و پاداش تجمعی برای تحلیل عملکرد ذخیره می‌شود. این روند تا اپیزود ۲۰۰۰ ادامه یافته و در نهایت سیاست بهینه برای اجرا روی یک روز تست استخراج می‌شود. در پایان آموزش، سیاست بهینه روی یک روز تست اجرا شده و معیارهایی از جمله هزینه کل بهره‌برداری، انرژی وارداتی از شبکه، مصرف دیزل و میزان آلاینده‌گی، میزان Load Shedding (باید صفر باشد) و رفتار SOC اندازه‌گیری می‌شوند. فلوچارت الگوریتم DQN و مراحل آن به طور کلی در شکل ۲ آمده است.



شکل ۲. فلوچارت الگوریتم DQN

۳-۷. هزینه کل

هزینه کل شامل دو بخش هزینه انرژی وارداتی از شبکه و هزینه مصرف دیزل ژنراتور تعریف شده است. بنابراین، هزینه کل به صورت معادله ۲۳ تعریف می‌شود [۳۷].

$$C_{total} = \sum_t (C_{grid}(t) + C_{diesel}(t)) \quad (23)$$

که در آن هزینه شبکه به صورت رابطه ۲۴ است [۳۷].

$$C_{grid}(t) = Price(t) \times E_{grid}(t) \quad (24)$$

که در آن $Price(t)$ قیمت لحظه‌ای برق شبکه طبق تعرفه زمان متغیر و $E_{grid}(t)$ انرژی وارداتی از شبکه برحسب کیلووات ساعت است. همچنین، هزینه دیزل ژنراتور نیز به صورت رابطه ۲۵ در نظر گرفته شده است [۳۷].

$$C_{diesel}(t) = \lambda_{DG} \times E_{DG}(t) \quad (25)$$

که در آن λ_{DG} هزینه سوخت و نگهداری دیزل بر حسب (\$/kWh) و $E_{DG}(t)$ انرژی تولیدی دیزل (kWh) است.

۳-۷-۱. آلاینده‌گی

آلاینده‌گی کل سامانه برابر مجموع آلاینده‌گی شبکه و دیزل است و در رابطه ۲۶ آمده است [۴۰]. با توجه به اینکه بخش قابل توجهی از برق شبکه از نیروگاه‌های فسیلی تأمین می‌شود، واردات انرژی از شبکه نیز به انتشار CO_2 منجر می‌شود.

$$E_{total}(t) = E_{DG}(t) + E_{grid}(t) \quad (26)$$

مقدار به‌دست‌آمده از معادله ۲۶ مستقیم در تابع پاداش ظاهر شده و با یک ضریب جریمه وزنی ($W_{emission}$) وارد مدل یادگیری تقویتی در معادله ۲۷ شده است [۴۰].

$$R_t = R_t - W_{emission} \cdot E_{total}(t) \quad (27)$$

۳-۸. فرضیات مدل‌سازی و بهره‌برداری

در این پژوهش، میکروگرید تحت فروشی که شرح داده می‌شود مدل‌سازی شده است. ۱- شبکه تک‌شینه و متعادل در نظر گرفته می‌شود؛ افت ولتاژ، تلفات خطوط و توان راکتیو باصراحت مدل نشده و تمرکز بر تراز توان اکتیو در باس اصلی است. ۲- گام زمانی شبیه‌سازی برابر یک ساعت است و افق بهره‌برداری یک روز (۲۴ گام زمانی) در نظر گرفته شده است. در هر گام، سیستم در حالت شبه‌مانا^۱ فرض می‌شود. به دلیل ماهیت تکرارشونده پروفایل بار و تولید، افق بهره‌برداری یک‌روزه انتخاب شده است. برای سناریوهای چندروزه، می‌توان مدل را با افق بلندمدت بسط داد. ۳- پروفایل‌های بار، تولید PV، تولید بادی و قیمت برق شبکه به صورت سری زمانی شناخته‌شده^۲ در اختیار عامل قرار دارند و عدم قطعیت آن‌ها به طور صریح مدل نشده است. ۴- میکروگرید امکان خرید انرژی از شبکه بالادست را دارد و در این کار، فروش انرژی به شبکه در سناریوهای اصلی لحاظ نشده است؛ بنابراین توان شبکه در مدل همواره به صورت توان وارداتی ظاهر می‌شود. ۵- باتری به کمک یک مبدل DC/AC ایده‌آل به باس AC متصل است؛ راندمان شارژ و دشارژ ثابت فرض شده و اثرات حرارتی و پیری باتری در نظر گرفته نشده است. ۶- دیزل ژنراتور در محدوده توان نامی خود کار می‌کند، دینامیک سرعت و جزئیات مکانیکی آن مدل نشده و تنها رابطه توان - سوخت و آلاینده‌گی به صورت خطی استفاده شده است. ۷- قابلیت اطمینان شبکه بالادست ۱۰۰ درصد در نظر گرفته شده و خاموشی شبکه سراسری بررسی نمی‌شود؛ لذا در صورت کمبود تولید تجدیدپذیر و شارژ ناکافی باتری، شبکه و دیزل نقش منبع پشتیبان را ایفا می‌کنند. ۸- قندهای فیزیکی شامل محدوده SOC، حداکثر نرخ شارژ / دشارژ باتری، حداکثر توان دیزل و حداکثر توان تبادل با شبکه در تابع گذار محیط اعمال می‌شوند؛ هرگونه نقض این قیود با جریمه در تابع پاداش منعکس شده است. ۹- عامل یادگیرنده تنها روی توان اکتیو تصمیم‌گیری می‌کند و مسائل کیفیت توان، هارمونیک‌ها و توان راکتیو خارج از محدوده این مطالعه است.

۳-۸-۱. پارامترهای اصلی سامانه و شبیه‌سازی

پارامترهای عددی مورد استفاده در شبیه‌سازی به صورت جدول‌های ۲ تا ۵ است (برخی مقادیر از روی پیاده‌سازی نرم‌افزاری تنظیم شده‌اند).

جدول ۲. پارامترهای فنی میکروگرید

پارامتر	مقدار مفروض	رفرنس
حداکثر توان پنل خورشیدی	۵ kW	[۴۶]
حداکثر توان توربین بادی	۳ kW	[۴۷]
حداکثر توان دیزل ژنراتور	۴ kW	[۴۸]
هزینه تولید دیزل	۰.۲۵ \$/kWh	[۴۹]
ضریب آلایندگی دیزل	۰.۸ kgCO ₂ /kWh	[۵۰]
ضریب آلایندگی شبکه	۰.۵ kgCO ₂ /kWh	[۵۱]

جدول ۳. پارامترهای باتری (BESS)

پارامتر	مقدار مفروض	رفرنس
ظرفیت نامی باتری	۱۰ kWh	[۵۲]
حد پایین باتری	٪۲۰	[۵۳]
حد بالای باتری	٪۸۰	[۵۳]
میزان اولیه باتری	٪۵۰	-
حداکثر توان شارژ/دشارژ	۳ kW	[۵۳]
راندمان شارژ و دشارژ	۰.۹۵	[۵۴]

جدول ۴. پروفایل قیمت برق شبکه

پارامتر	مقدار مفروض	رفرنس
قیمت پایه در اغلب ساعت‌های روز	۰.۱۰ \$/kWh	[۵۵]
ساعت‌های میان روز (۱۰ تا ۱۵) با قیمت ارزان‌تر	۰.۰۷ \$/kWh	[۵۵]
ساعت‌های اوج عصر (۱۸ تا ۲۲) با قیمت بالاتر	۰.۱۸ \$/kWh	[۵۵]

جدول ۵. ضرایب تابع پاداش

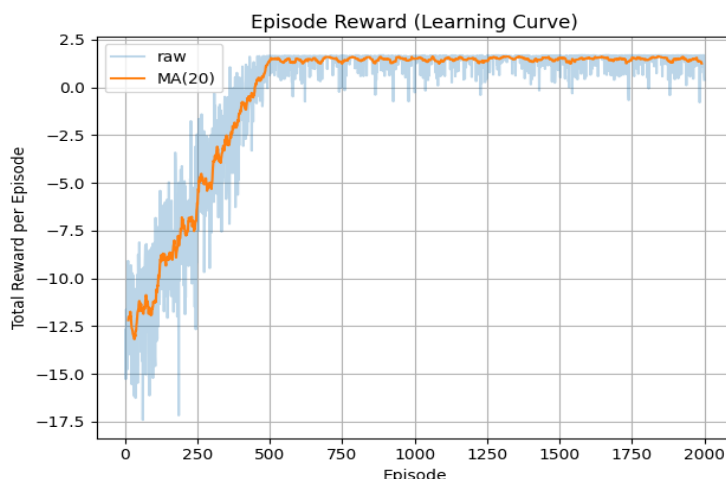
پارامتر	مقدار مفروض	رفرنس
جریمه عدم تأمین بار	۲۰۰ \$/kWh	-
جریمه نزدیک شدن SOC به حدود مینیوم/ماکسیمم	۰.۷	-
وزن آلایندگی در پاداش	۰.۰۸	-
وزن تشویق SOC مناسب قبل از ساعت‌های پیک	۰.۱۵	-
وزن تشویق SOC میانی طی روز	۰.۰۵	-
وزن جریمه واردات از شبکه در ساعت‌های اوج	۰.۲	-

۴. نتایج شبیه‌سازی

شبیه‌سازی‌ها مطابق مفروضات بخش ۳ و با افق زمانی ۲۴ ساعته انجام شده است. عامل RL پس از طی ۲ هزار اپیزود آموزش، با سیاست کاملاً حریصانه ($\epsilon=0$) روی یک روز آزمون اجرا شده و شاخص‌های هزینه، آلایندگی، انرژی وارداتی از شبکه و رفتار باتری ارزیابی شده‌اند.

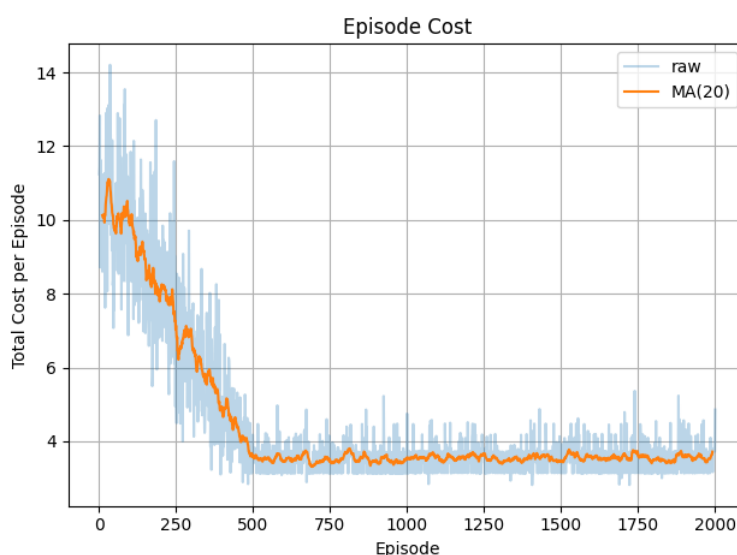
۴-۱. تحلیل فرایند یادگیری عامل RL

در شکل ۳ روند تغییر پاداش تجمعی در هر اپیزود نشان داده شده است. در ابتدای آموزش، پاداش اپیزودی مقادیر منفی قابل توجهی حدود ۱۵- تا ۱۰- دارد که ناشی از هزینه زیاد شبکه، استفاده نامناسب از باتری و احتمال وقوع جریمه‌های SOC و آلایندگی است. با پیشرفت اپیزودها، عامل با استفاده از مکانیزم PER و Double DQN، به تدریج سیاست خود را بهبود می‌دهد؛ به طوری که از حدود اپیزود ۴۰۰ به بعد، پاداش‌ها در حوالی ۱+ تا ۱/۵+ پایدار می‌شوند. این موضوع نشان‌دهنده همگرایی مناسب الگوریتم و دستیابی به سیاستی است که هزینه و آلایندگی را تا حد خوبی کاهش می‌دهد.



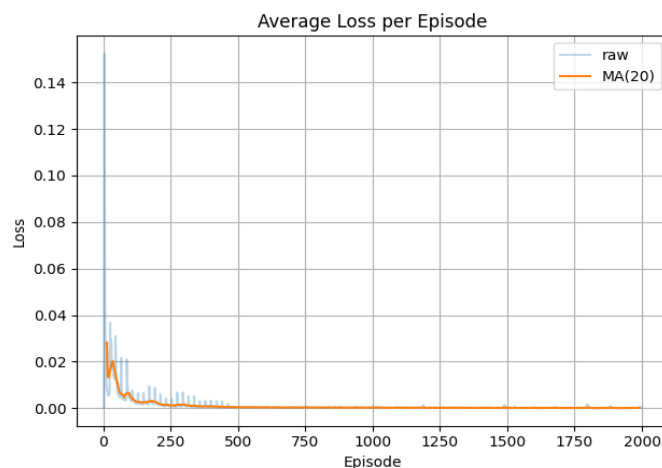
شکل ۳. روند تغییر پاداش تجمعی هر اپیزود

شکل ۴ روند تغییر هزینه کل هر اپیزود را نشان می‌دهد. در اپیزودهای اولیه، هزینه کل روزانه حدود ۱۰ تا ۱۲ واحد پولی است که بیانگر وابستگی بالا به شبکه و استفاده غیربهرینه از دیزل و باتری است. با ادامه آموزش، عامل یاد می‌گیرد در ساعت‌های ارزان، باتری را شارژ و در ساعت‌های اوج قیمت، از باتری و دیزل به جای خرید از شبکه استفاده کند. نتیجه این رفتار، کاهش هزینه اپیزودی به محدوده تقریباً ۳/۵ تا ۴ دلار و پایداری آن در انتهای دوره آموزش است. این کاهش واضح هزینه، تأییدکننده کارایی تابع پاداش و ساختار محیط RL تعریف شده در بخش ۳-۳ است.



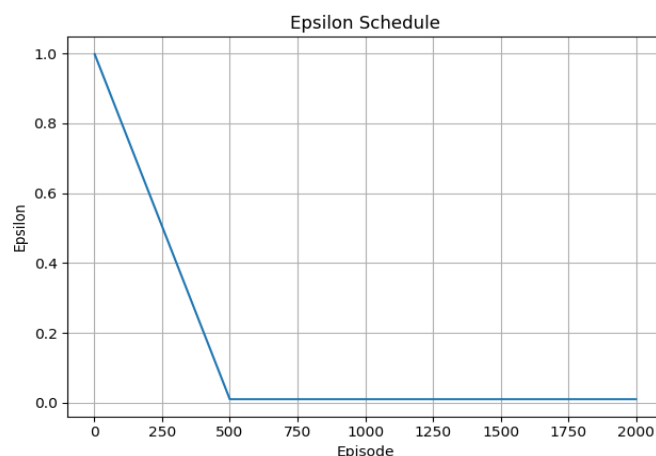
شکل ۴. روند تغییر هزینه کل هر اپیزود

شکل ۵ میانگین خطای یادگیری شبکه Q در هر اپیزود را نمایش می‌دهد. در ابتدای آموزش مقدار Loss نسبتاً بزرگ بوده و نوسانات زیادی دارد که ناشی از تصادفی بودن اکشن‌ها و عدم پایداری شبکه عصبی است. با افزایش تعداد نمونه‌های ذخیره شده در PER و به‌روزرسانی‌های مکرر شبکه، خطا به سرعت کاهش می‌یابد و از حدود اپیزود ۴۰۰ به بعد به مقادیر بسیار کوچک نزدیک صفر می‌رسد. این رفتار نشان می‌دهد تقریب تابع Q به‌خوبی صورت گرفته و شبکه در ناحیه‌ای پایدار از فضای پارامترها قرار گرفته است.



شکل ۵. میانگین خطای یادگیری شبکه Q

شکل ۶ سیر تغییر پارامتر ϵ را نشان می‌دهد. مقدار ϵ از $1/0$ در ابتدای آموزش شروع شده و طی حدود ۵۰۰ اپیزود به مقدار $0/01$ کاهش می‌یابد و پس از آن تقریباً ثابت می‌ماند. این روند باعث می‌شود عامل در ابتدای کار کاوش گسترده‌ای در فضای عمل انجام دهد و سپس به تدریج به سمت بهره‌برداری از دانش آموخته‌شده حرکت کند. هم‌زمانی کاهش ϵ با بهبود پاداش و کاهش Loss مؤید طراحی مناسب برنامه کاهش اپسیلون است.

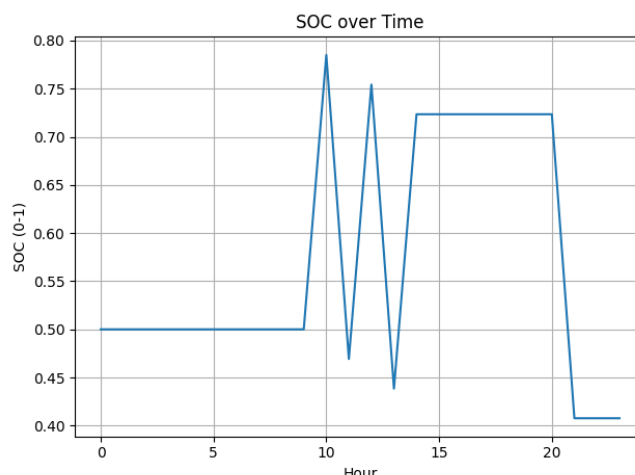


شکل ۶. سیر تغییر پارامتر ϵ

۴-۲. تحلیل بهره‌برداری روزانه میکروگرید

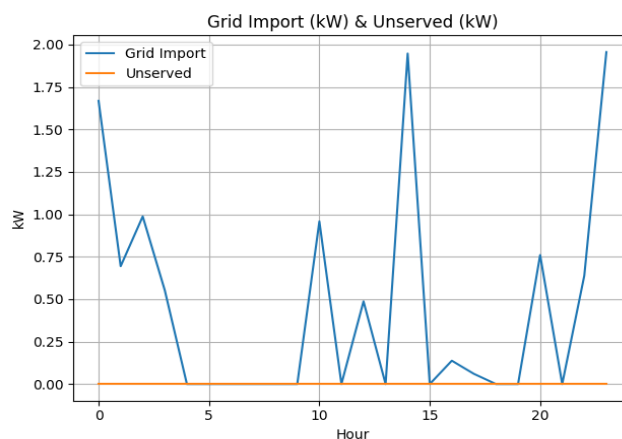
در این بخش، سیاست نهایی عامل RL روی یک روز آزمون (۲۴ ساعت) اجرا شده و رفتار متغیرهای کلیدی سیستم بررسی شده است. نمودارهای زمانی از خروجی کد پایتونی استخراج شده‌اند.

شکل ۷ تغییرات SOC باتری طی روز را نشان می‌دهد. باتری روز را با SOC حدود $0/5$ آغاز می‌کند. در ساعت‌های میانی روز که تولید PV و باد افزایش می‌یابد، عامل RL باتری را شارژ می‌کند و SOC تا حدود $0/75 - 0/8$ بالا می‌رود. در ادامه، با نزدیک شدن به ساعت‌های اوج قیمت (۱۸ تا ۲۲)، باتری به تدریج دشارژ می‌شود و بخش قابل توجهی از بار عصر را تأمین می‌کند، به گونه‌ای که در انتهای روز SOC به حدود $0/4$ می‌رسد. این الگو نشان می‌دهد عبارتهای تشویقی و جریمه‌ای مربوط به SOC در تابع پاداش (بخش ۳-۴) به درستی عامل را به ذخیره انرژی قبل از ساعت‌های پیک و جلوگیری از نزدیک شدن به حدود ممنوعه هدایت کرده‌اند.



شکل ۷. تغییرات SOC باتری طی روز

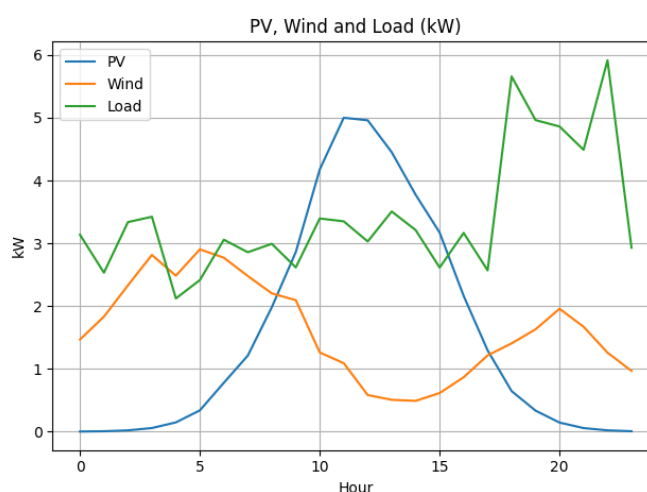
در شکل ۸ توان وارداتی از شبکه و توان تأمین‌نشده بر حسب ساعت نشان داده شده است. واردات توان از شبکه در چند بازه محدود و به صورت پیک‌های کوچک در ساعت‌های خاص دیده می‌شود، در حالی که منحنی Unserved در تمام ساعت‌ها تقریباً صفر باقی مانده است. این موضوع تأیید می‌کند که عامل RL توانسته است با ترکیب مناسب PV، باد، باتری و دیزل، بار را به طور کامل تأمین کند و از وقوع Load Shedding جلوگیری کند؛ همان‌طور که در KPI ها مقدار انرژی تأمین‌نشده برای روز آزمون برابر صفر گزارش شده است. این رفتار با نتایج مطالعات اخیر در حوزه انرژی‌هاب‌های مسکونی نیز هم‌راستا است؛ به گونه‌ای که بهره‌برداری هدفمند از سامانه‌های ذخیره‌سازی می‌تواند ضمن کاهش واردات انرژی از شبکه، سطح خوداتکایی انرژی را در مقیاس روزانه افزایش دهد، هرچند این امر ممکن است با افزایش هزینه نهایی انرژی همراه باشد [۵۶].



شکل ۸. توان وارداتی از شبکه و توان تأمین‌نشده

شکل ۹ پروفایل توان PV، توان باد و بار الکتریکی را مقایسه می‌کند. بار در ساعت‌های عصر به حدود ۶ kW می‌رسد، در حالی که تولید PV در حوالی ظهر بیشینه است و سپس کاهش می‌یابد. تولید باد در برخی ساعت‌های شب و اوایل صبح افزایش دارد و به تأمین بار در این بازه‌ها کمک می‌کند. عامل RL با بهره‌گیری از این الگوی زمانی، در ساعت‌های تولید بالا اقدام به شارژ باتری کرده و در ساعت‌های بار سنگین از ذخیره باتری و در صورت نیاز از دیزل و شبکه استفاده می‌کند؛ بنابراین، ترکیب مناسب منابع و رفتار شارژ / دشارژ باتری باعث می‌شود بار بدون نیاز زیاد به شبکه تأمین شود. این الگوی بهره‌برداری با نتایج مطالعات اخیر در حوزه

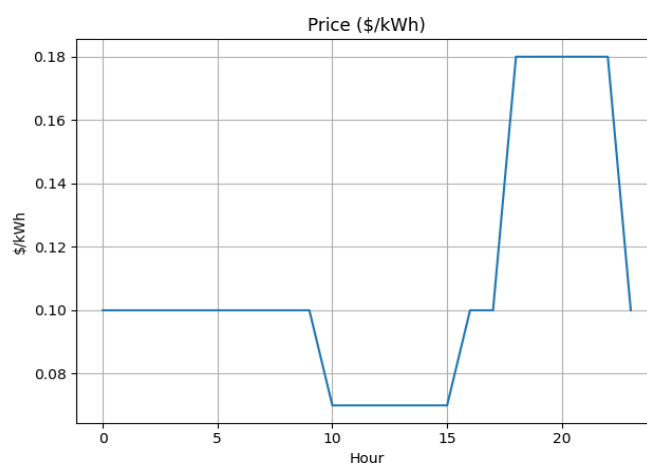
پخش اقتصادی ریزشکه نیز هم‌راستا است؛ به گونه‌ای که هماهنگی بهینه میان منابع تجدیدپذیر، واحدهای تولید پراکنده و شبکه اصلی می‌تواند ضمن تأمین کامل بار، هزینه کل بهره‌برداری را به طور محسوسی کاهش دهد [۵۷].



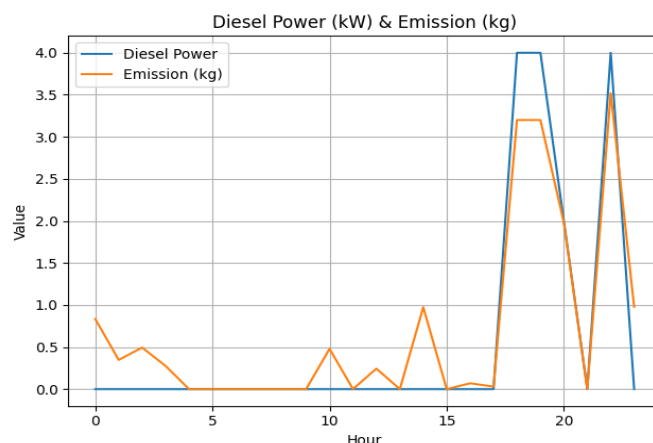
شکل ۹. پروفایل توان PV، توان باد و بار الکتریکی

برای درک بهتر تصمیم‌های عامل در واردات از شبکه، شکل ۱۰ پروفایل قیمت برق شبکه را نشان می‌دهد. قیمت شبکه در اغلب ساعت‌ها برابر $0.10 \text{ \$/kWh}$ ، در ساعت‌های میانی روز به $0.07 \text{ \$/kWh}$ کاهش یافته و در ساعت‌های اوج عصر به $0.18 \text{ \$/kWh}$ می‌رسد. مقایسه شکل‌های ۸ و ۱۰ نشان می‌دهد عامل RL عمدتاً در ساعت‌های ارزان از شبکه واردات دارد و در ساعت‌های گران، بیشتر از باتری و دیزل استفاده می‌کند؛ این دقیقاً همان رفتاری است که تابع پاداش طراحی شده به دنبال آن بوده است.

شکل ۱۱ توان خروجی دیزل ژنراتور و آلاینده‌ی لحظه‌ای را نمایش می‌دهد. همان‌طور که مشاهده می‌شود، دیزل تنها در چند ساعت محدود، عمدتاً نزدیک ساعت‌های اوج قیمت، فعال می‌شود و در سایر ساعت‌ها خاموش است. آلاینده‌ی نیز فقط در همین بازه‌ها ظاهر می‌شود و مقدار آن با توجه به ضریب انتشار $0.8 \text{ kgCO}_2/\text{kWh}$ مقدار قابل قبولی دارد. مقایسه این شکل با شکل‌های ۸ و ۱۰ نشان می‌دهد عامل RL از دیزل به عنوان منبع پشتیبان در شرایطی استفاده می‌کند که یا قیمت شبکه بالاست و یا برای حفظ SOC باتری و جلوگیری از Load Shedding نیاز به تولید اضافی وجود دارد. در عین حال، وجود ترم جریمه آلاینده‌ی در تابع پاداش باعث شده استفاده از دیزل حداقلی باقی بماند.



شکل ۱۰. پروفایل قیمت برق شبکه



شکل ۱۱. توان و آلاینده‌گی لحظه‌ای خروجی دیزل ژنراتور

۳-۴. نتایج نهایی

نتایج شبیه‌سازی نشان می‌دهد الگوریتم Double DQN توانسته است سیاست بهینه بهره‌برداری را با کارایی قابل توجهی یاد بگیرد. هزینه کل بهره‌برداری از مقدار اولیه حدود ۱۰ دلار به حدود ۳/۴۴ دلار کاهش یافته که بیانگر کاهش ۶۶ درصدی هزینه است. آلاینده‌گی کل نیز از مقدار ۴۱ kg CO₂ در اپیزودهای ابتدایی به حدود ۱۶/۶ kg CO₂ در انتهای آموزش رسیده و ۶۰ درصد کاهش داشته است. علاوه بر این، مقدار توان تأمین‌نشده که در اپیزودهای اولیه مقادیر کوچکی داشته، پس از همگرایی الگوریتم کاملاً حذف شده و مقدار Load Shedding صفر گزارش شده است. رفتار باتری نیز نشان می‌دهد عامل RL به طور مؤثر باتری را در ساعات‌های کم‌قیمت شارژ و در ساعات‌های اوج دشارژ کرده و از این طریق هزینه و آلاینده‌گی را کاهش داده است. به طور کلی، نتایج نهایی در جدول ۶ خلاصه شده است.

جدول ۶. جدول نتایج نهایی

شاخص عملکرد	مقدار اولیه	مقدار نهایی	درصد بهبود
هزینه کل روزانه	۱۰ دلار	۳/۴۴ دلار	۶۶٪ کاهش
آلاینده‌گی CO ₂	۴۰ Kg	۱۶/۶ kg	۶۰٪ کاهش
Load Shedding	مقدار کم در اپیزودهای اولیه (۰/۳-۰/۱)	۰	۱۰۰٪ حذف کامل
SOC نهایی	~۰/۵۰	~۰/۴۱	- (تحلیل کیفی)
پایداری پاداش	بی‌ثبات	پایدار +۱	-

۴-۴. اعتبارسنجی

در این بخش، عملکرد روش پیشنهادی مبتنی بر یادگیری تقویتی با الگوریتم بهینه‌سازی چندهدفه NSGA-II مورد مقایسه قرار گرفته است. الگوریتم NSGA-II با هدف کمینه‌سازی هم‌زمان هزینه بهره‌برداری و میزان انتشار آلاینده‌گی، مجموعه‌ای از راه‌حل‌های پارتو را تولید می‌کند که از میان آن‌ها نقطه زانو به عنوان راه‌حل نماینده انتخاب شده است. نتایج عددی نشان می‌دهد اگرچه NSGA-II قادر به دستیابی به راه‌حل‌های بهینه از نظر هزینه و آلاینده‌گی است، اما این بهبود عمدتاً از طریق تخلیه شدید باتری و کاهش سطح شارژ نهایی آن تا نزدیکی حداقل مجاز حاصل شده است. در مقابل، روش مبتنی بر یادگیری تقویتی ضمن حذف کامل Load Shedding، توانسته است سطح شارژ باتری را در محدوده‌ای ایمن‌تر و پایدارتر حفظ کند. علاوه بر این، نتایج مقایسه‌ای نشان می‌دهد روش RL نه تنها از نظر پایداری عملکرد باتری برتری دارد، بلکه در این سناریو موفق شده است هزینه کل بهره‌برداری و میزان انتشار CO₂ را نیز نسبت به راه‌حل زانوی NSGA-II کاهش دهد. این موضوع بیانگر توانایی RL در یادگیری یک سیاست کنترلی متعادل و عملی برای مدیریت انرژی ریزشکه در شرایط واقعی است.

در مجموع، اگرچه NSGA-II یک ابزار قدرتمند برای بهینه‌سازی آفلاین محسوب می‌شود، روش RL به دلیل قابلیت تصمیم‌گیری بلادرنگ، پایداری بالاتر و عملکرد متوازن‌تر، گزینه مناسب‌تری برای پیاده‌سازی عملی در سیستم‌های مدیریت انرژی هوشمند به شمار می‌رود. نتایج عددی هر دو روش در جدول ۷ نمایش داده شده است.

جدول ۷. جدول مقایسه خروجی‌ها

روش	هزینه کل (\$)	انتشار CO ₂ (kg)	Load Shedding (kWh)	SOC نهایی
NSGA-II (Knee)	۴.۷۵	۲۰.۲۸	۰	۰.۲۱
RL (Proposed)	۳.۴۴	۱۶۶۰	۰	۰.۴۱
تفاوت (RL - NSGA-II)	-۱.۳۱	-۳۶۸	۰	+۲.۰

به منظور ارزیابی اعتبار و جایگاه روش پیشنهادی، عملکرد چهارچوب ارائه‌شده با چند مطالعه منتخب منتشرشده در پنج سال اخیر در حوزه مدیریت انرژی ریزشکه مقایسه شده است. همان‌گونه که در جدول ۸ مشاهده می‌شود، روش پیشنهادی این پژوهش با بهره‌گیری از الگوریتم Double DQN همراه با PER و طراحی تابع پاداش چندهدفه، توانسته است به طور هم‌زمان کاهش قابل توجه هزینه بهره‌برداری، کاهش انتشار آلاینده‌ها و حذف کامل بار تأمین‌نشده را محقق سازد.

جدول ۸. مقایسه عملکرد روش پیشنهادی با مطالعات منتخب اخیر

ردیف	مرجع	سال	روش / الگوریتم	مهم‌ترین نتایج گزارش‌شده در مطالعه	مقایسه با پژوهش حاضر
۱	[۳۴]	2022	DQN	دستیابی به عملکرد نزدیک بهینه با سرعت محاسباتی بالا نسبت به MILP	در پژوهش حاضر، استفاده از Double DQN + PER باعث شد هزینه بهره‌برداری از حدود ۱۰ دلار به ۳.۴۴ دلار برسد (۶۶٪ کاهش) و آلایندگی نیز از ۴۱ kg CO ₂ به ۱۶.۶ kg CO ₂ کاهش یابد (۶۰٪ کاهش).
۲	[۵۸]	2023	Safe Reinforcement Learning	تمرکز اصلی بر رعایت قیود عملیاتی و امنیت بهره‌برداری	در حالی که مطالعه یادشده بیشتر بر ایمنی تمرکز دارد، روش پیشنهادی این پژوهش به صورت هم‌زمان سه هدف اصلی هزینه، آلایندگی و قابلیت اطمینان را پوشش داده و Load Shedding را به صفر رسانده است.
۳	[۵۹]	2023	Multi-objective Evolutionary Reinforcement Learning	ارائه مجموعه جواب‌های پارتو میان اهداف اقتصادی و محیط زیستی	به خلاف رویکردهای پارتو محور که نیازمند انتخاب راه‌حل نهایی هستند، روش پیشنهادی این پژوهش مستقیم یک سیاست بهره‌برداری عملیاتی و پایدار ارائه داده که هم‌زمان موجب کاهش هزینه و انتشار شده است.
۴	[۶۰]	2024	Deep Reinforcement Learning	حدود ۲٪ بهبود اقتصادی با در نظر گرفتن تلفات غیرخطی باتری	در پژوهش حاضر، بهبود اقتصادی بسیار بالاتر (۶۶٪) حاصل شده و در عین حال SOC نهایی در سطح ایمن تر حدود ۰.۴۱ حفظ شده که بیانگر بهره‌برداری متعادل‌تر از باتری است.
۵	[۶۱]	2025	Deep Reinforcement Learning	بهبود مدیریت انرژی در حضور منابع تجدیدپذیر و خودرو برقی	با وجود ساختار ساده‌تر ریزشکه در این پژوهش، نتایج نشان می‌دهد روش پیشنهادی توانسته است هزینه کمتر، آلایندگی کمتر و حذف کامل بار تأمین‌نشده را به طور هم‌زمان محقق کند.

۵. بحث و نتیجه‌گیری

نتایج شبیه‌سازی نشان داد چارچوب پیشنهادی مبتنی بر Double DQN همراه با PER توانسته است در مدیریت انرژی ریزشکه عملکرد قابل قبولی ارائه دهد. عامل یادگیرنده پس از ۲ هزار اپیزود آموزش به همگرایی مناسب رسید و توانست سیاست بهره‌برداری مؤثری میان منابع تجدیدپذیر، باتری، دیزل ژنراتور و شبکه برق ایجاد کند. از دیدگاه کمی، هزینه بهره‌برداری روزانه از حدود ۱۰ دلار به ۳/۴۴ دلار کاهش یافت که بیانگر ۶۶ درصد صرفه‌جویی اقتصادی است. همچنین، آلایندگی کل سامانه از

حدود ۴۱ کیلوگرم دی‌اکسیدکربن به ۱۶/۶ کیلوگرم دی‌اکسیدکربن کاهش یافت که معادل ۶۰ درصد کاهش انتشار است. افزون بر این، مقدار بار تأمین‌نشده در تمام ساعت‌های روز به صفر رسید و عامل توانست بدون وقوع Load Shedding، تقاضای بار را به طور کامل تأمین کند. تحلیل رفتار عملیاتی سیستم نشان داد عامل RL توانسته است رابطه زمانی میان تولید منابع تجدیدپذیر، تغییرات بار و تعرفه زمانی برق را به‌خوبی یاد بگیرد. باتری در ساعت‌های میانی روز که تولید خورشیدی و بادی بیشتر و قیمت برق کمتر است شارژ شده و در ساعت‌های اوج قیمت (۱۸ تا ۲۲) دشارژ شده است. سطح شارژ باتری از مقدار اولیه ۰/۵ تا حدود ۰/۸ در میانه روز افزایش یافته و در پایان روز حدود ۰/۴۱ باقی مانده که نشان‌دهنده بهره‌برداری اقتصادی و در عین حال ایمن از سامانه ذخیره‌ساز است. این رفتار مؤید آن است که طراحی تابع پاداش چندهدفه توانسته عامل را به سمت تصمیم‌گیری‌هایی هدایت کند که هم‌زمان اهداف اقتصادی، زیست‌محیطی و قابلیت اطمینان را برآورده سازد.

از مهم‌ترین قوت‌های این پژوهش می‌توان به چند مورد اشاره کرد. نخست، ترکیب Double DQN با PER باعث افزایش پایداری یادگیری و سرعت همگرایی شده است. دوم، طراحی تابع پاداش چندهدفه امکان در نظر گرفتن هم‌زمان هزینه، آلاینده‌گی، قابلیت اطمینان و مدیریت سطح شارژ باتری را فراهم کرده است. سوم، نتایج نشان می‌دهد روش پیشنهادی نه تنها از نظر اقتصادی و زیست‌محیطی مؤثر بوده، بلکه حذف کامل بار تأمین‌نشده را نیز محقق کرده است؛ موضوعی که از منظر بهره‌برداری عملی ریزشکه اهمیت زیادی دارد. با وجود این مزایا، مطالعه حاضر دارای محدودیت‌ها و ضعف‌هایی نیز هست که می‌تواند مسیر تحقیقات آینده را مشخص کند. نخست آنکه پروفایل‌های بار، تولید خورشیدی، تولید بادی و قیمت برق به صورت قطعی و برای یک افق ۲۴ ساعته در نظر گرفته شده‌اند؛ در حالی که در شرایط واقعی، عدم قطعیت‌های اقلیمی و خطاهای پیش‌بینی می‌توانند بر عملکرد سیستم اثرگذار باشند. دوم، مدل باتری به صورت ساده در نظر گرفته شده و اثرات تخریب، پیری سلول‌ها و محدودیت‌های حرارتی لحاظ نشده است. سوم، در این پژوهش تنها واردات انرژی از شبکه در نظر گرفته شده و امکان فروش انرژی به شبکه مدل نشده است. همچنین، مدل دیزل ژنراتور فاقد محدودیت‌هایی نظیر هزینه روشن و خاموش شدن، حداقل زمان کارکرد و قیود دینامیکی است. این ساده‌سازی‌ها اگرچه موجب تمرکز بر ارزیابی قابلیت الگوریتم شده‌اند، اما می‌توانند فاصله‌ای میان مدل و شرایط واقعی ایجاد کنند.

در مجموع، نتایج این پژوهش نشان می‌دهد چهارچوب پیشنهادی قادر است در ریزشکه‌های با نفوذ بالای منابع تجدیدپذیر، یک سیاست کنترلی هوشمند، پایدار و قابل اتکا ارائه دهد. کاهش ۶۰ درصدی هزینه بهره‌برداری، کاهش ۶۰ درصدی آلاینده‌گی و حذف کامل Load Shedding بیانگر آن است که استفاده از یادگیری تقویتی عمیق می‌تواند رویکردی مؤثر برای مدیریت انرژی ریزشکه‌های آینده باشد. در ادامه، توسعه مدل در شرایط عدم قطعیت، افق‌های زمانی بلندمدت، مدل‌سازی دقیق‌تر باتری و مقایسه با الگوریتم‌های پیشرفته‌تری نظیر SAC^۱، TD3^۲ و PPO^۳ می‌تواند تصویر دقیق‌تری از مناسب‌ترین روش یادگیری تقویتی برای سیستم‌های میکروگرید ارائه دهد. در نهایت، استفاده از چهارچوب‌های یادگیری تقویتی چندعامله، که در آن منابع مختلف (باتری، دیزل، شبکه) توسط عامل‌های مستقل اما هماهنگ کنترل می‌شوند، می‌تواند برای میکروگریدهای بزرگ‌تر بسیار کارآمدتر باشد. همچنین، افزودن شاخص‌های اقتصادی بلندمدت مانند هزینه هم‌سطح انرژی^۴ یا تحلیل چرخه عمر، کاربردپذیری صنعتی مدل را افزایش خواهد داد.

1. Soft Actor-Critic
2. Twin Delayed Deep Deterministic Policy Gradient
3. Proximal Policy Optimization
4. Levelized Cost of Energy

منابع

- [1] Olabi A, Elsaid K, Obaideen K, Abdelkareem MA, Rezk H, Wilberforce T, et al. Renewable energy systems: Comparisons, challenges and barriers, sustainability indicators, and the contribution to UN sustainable development goals. *International Journal of Thermofluids*. 2023;20:100498.
- [2] Yajloo AB, Hamedani EA. Simulation of a solar-based small-scale green hydrogen production unit in Iran: A techno-economic-feasibility analysis. *Results in Engineering*. 2025:106734.
- [3] Erdiwansyah f, Mahidin f, Husin H, Nasaruddin f, Zaki M, Muhibbuddin f. A critical review of the integration of renewable energy sources with various technologies. *Protection and control of modern power systems*. 2021;6(1):3.
- [4] Ang T-Z, Salem M, Kamarol M, Das HS, Nazari MA, Prabakaran N. A comprehensive study of renewable energy sources: Classifications, challenges and suggestions. *Energy strategy reviews*. 2022;43:100939.
- [5] Abbasian Hamedani E, Bahrami Yajloo A, Talebi S. A comprehensive review on carbon capture, transportation, storage, and utilization technologies; part I: Carbon capture technologies. *Advances in Energy Sciences and Technologies*. 2025;1(1):119-32.
- [6] Albarakati AJ, Boujoudar Y, Azeroual M, Eliyaouy L, Kotb H, Aljarbouh A, et al. Microgrid energy management and monitoring systems: A comprehensive review. *Frontiers in Energy Research*. 2022;10:1097858.
- [7] Liu D, Zang C, Zeng P, Li W, Wang X, Liu Y, et al. Deep reinforcement learning for real-time economic energy management of microgrid system considering uncertainties. *Frontiers in Energy Research*. 2023;11:1163053.
- [8] Zia MF, Elbouchikhi E, Benbouzid M. Microgrids energy management systems: A critical review on methods, solutions, and prospects. *Applied energy*. 2018;222:1033-55.
- [9] Shakya AK, Pillai G, Chakrabarty S. Reinforcement learning algorithms: A brief survey. *Expert Systems with Applications*. 2023;231:120495.
- [10] Nian R, Liu J, Huang B. A review on reinforcement learning: Introduction and applications in industrial process control. *Computers & Chemical Engineering*. 2020;139:106886.
- [11] Matsuo Y, LeCun Y, Sahani M, Precup D, Silver D, Sugiyama M, et al. Deep learning, reinforcement learning, and world models. *Neural Networks*. 2022;152:267-75.
- [12] Wang H-n, Liu N, Zhang Y-y, Feng D-w, Huang F, Li D-s, et al. Deep reinforcement learning: a survey. *Frontiers of Information Technology & Electronic Engineering*. 2020;21(12):1726-44.
- [13] Gronauer S, Diepold K. Multi-agent deep reinforcement learning: a survey. *Artificial Intelligence Review*. 2022;55(2):895-943.
- [14] Hamedani EA, Khodaparast P, Hosseini E, Mahmudy T, Yajloo AB. A mini-review of energy hub: Concept, components, classifications, and applications. *Energy Reports*. 2026;15:108886.
- [15] Sharma S, Ali I. Efficient energy management and cost optimization using multi-objective grey wolf optimization for EV charging/discharging in microgrid. *e-Prime-Advances in Electrical Engineering, Electronics and Energy*. 2024;10:100804.
- [16] Yousri D, Ousama A, Fathy A, Babu TS, Allam D. Managing the exchange of energy between microgrid elements based on multi-objective enhanced marine predators algorithm. *Alexandria Engineering Journal*. 2022;61(11):8487-505.
- [17] Vijaykumar G. Integrated energy management and harmonic mitigation in a microgrid using sea gull-ANN MPPT and advanced multi-level inverter. *Ain Shams Engineering Journal*. 2025;16(7):103397.
- [18] Mei Y, Li B, Wang H, Wang X, Negnevitsky M. Multi-objective optimal scheduling of microgrid with electric vehicles. *Energy Reports*. 2022;8:4512-24.
- [19] Babu VV, Roselyn JP, Sundaravadeivel P. Multi-objective genetic algorithm based energy management system considering optimal utilization of grid and degradation of battery storage in microgrid. *Energy Reports*. 2023;9:5992-6005.
- [20] Aeggegn DB, Nyakoe GN, Wekesa C. An energy management system for multi-microgrid system considering uncertainties using multi-objective multi-verse optimization. *Energy Reports*. 2025;13:286-302.

- [21] Osama A, Allam D, Fathy A, Abdelaziz AY, Kim W-W, Hong J, et al. Optimal energy managing and power scheduling of microgrid in grid-connected mode using modified multi-objective manta ray technique. *Energy Reports*. 2025;13:4781-99.
- [22] Rochd A, Raihani A, Mahir O, Kissaoui M, Laamim M, Lahmer A, et al. Swarm Intelligence-driven Multi-objective optimization for microgrid energy management and trading considering DERs and EVs integration: case studies from green energy park, Morocco. *Results in Engineering*. 2025;25:104400.
- [23] Lamari M, Amrane Y, Boudour M, Boussahoua B. Multi-objective economic/emission optimal energy management system for scheduling micro-grid integrated virtual power plant. *Energy Science & Engineering*. 2022;10(8):3057-74.
- [24] Ali L, Muyeen S, Bizhani H, Ghosh A. A multi-objective optimization for planning of networked microgrid using a game theory for peer-to-peer energy trading scheme. *IET Generation, Transmission & Distribution*. 2021;15(24):3423-34.
- [25] Kamarposhti MA, Colak I, Shokouhandeh H, Iwendi C, Padmanaban S, Band SS. Optimum operation management of microgrids with cost and environment pollution reduction approach considering uncertainty using multi-objective NSGAII algorithm. *IET Renewable Power Generation*. 2025;19(1):e12579.
- [26] Momen S, Nikoukar J, Gandomkar M. Multi-objective optimization of energy consumption in microgrids considering CHPs and renewables using improved shuffled frog leaping algorithm. *Journal of Electrical Engineering & Technology*. 2023;18(3):1539-55.
- [27] Wang Z, Luo Y, Wu W, Cao L, Li Z. Multi-objective optimization models for power load balancing in distributed energy systems. *Energy Informatics*. 2025;8(1):104.
- [28] Wei W, Wang H, Hou K, Ji L. Multi-objective optimal configuration of stand-alone microgrids based on Benders decomposition considering power supply reliability. *IET Energy Systems Integration*. 2022;4(2):281-95.
- [29] Ahmadi B, Hoogsteen G, Zawadzki P, Gerards ME, Radziszewska W, Bykuć S, et al. A novel multi-objective approach to user-centric energy management systems. *Energy Reports*. 2025;14:185-204.
- [30] Niknami A, Askari MT, Ahmadi MA, Nik MB, Moghaddam MS. Resilient day-ahead microgrid energy management with uncertain demand, EVs, storage, and renewables. *Cleaner Engineering and Technology*. 2024;20:100763.
- [31] Veisi M, Adabi F, Kavousi-Fard A, Karimi M. A novel comprehensive energy management model for multi-microgrids considering ancillary services. *IET Generation, Transmission & Distribution*. 2022;16(23):4710-25.
- [32] Hosseini E, García-Triviño P, Horrillo-Quintero P, Carrasco-Gonzalez D, García-Vázquez CA, Sarrias-Mena R, et al. A novel reinforcement learning-based multi-objective energy management system for multi-energy microgrids integrating electrical, hydrogen, and thermal elements. *Electric Power Systems Research*. 2025;242:111474.
- [33] Cui Y, Xu Y, Li Y, Wang Y, Zou X. Deep reinforcement learning based optimal energy management of multi-energy microgrids with uncertainties. *CSEE Journal of Power and Energy Systems*. 2024.
- [34] Alabdullah MH, Abido MA. Microgrid energy management using deep Q-network reinforcement learning. *Alexandria Engineering Journal*. 2022;61(11):9069-78.
- [35] Zheng Y, Jia J, An D. Energy management for microgrids with hybrid hydrogen-battery storage: A reinforcement learning framework integrated multi-objective dynamic regulation. *Processes*. 2025;13(8):2558.
- [36] Yao F, Zhao W, Forshaw M, Zhou W. A unified data-driven approach under deep reinforcement learning with direct control responses for microgrid operations. *Knowledge-Based Systems*. 2025;325:113844.
- [37] Olivares DE, Mehrizi-Sani A, Etemadi AH, Cañizares CA, Iravani R, Kazerani M, et al. Trends in microgrid control. *IEEE Transactions on smart grid*. 2014;5(4):1905-19.
- [38] Rossow M. *Electrification of Aircraft: Challenges, Barriers, and Potential Impacts*. Continuing Education and Development, Inc, USA. 2021.

- [39] Das I, Cañizares CA. Renewable energy integration in diesel-based microgrids at the Canadian arctic. *Proceedings of the IEEE*. 2019;107(9):1838-56.
- [40] Basu M. Economic environmental dispatch using multi-objective differential evolution. *Applied soft computing*. 2011;11(2):2845-53.
- [41] Sutton RS, Barto AG. *Reinforcement learning: An introduction*: MIT press Cambridge; 1998.
- [42] Ye Y, Wang H, Chen P, Tang Y, Strbac G. Safe deep reinforcement learning for microgrid energy management in distribution networks with leveraged spatial-temporal perception. *IEEE Transactions on Smart Grid*. 2023;14(5):3759-75.
- [43] Van Hasselt H, Guez A, Silver D, editors. *Deep reinforcement learning with double q-learning*. *Proceedings of the AAAI conference on artificial intelligence*; 2016.
- [44] Schaul T, Quan J, Antonoglou I, Silver D. Prioritized experience replay. *arXiv preprint arXiv:151105952*. 2015.
- [45] Mnih V, Kavukcuoglu K, Silver D, Rusu AA, Veness J, Bellemare MG, et al. Human-level control through deep reinforcement learning. *nature*. 2015;518(7540):529-33.
- [46] Green MA. *Solar cells: operating principles, technology, and system applications*. 1981.
- [47] Anup K, Whale J, Urmee T. Urban wind conditions and small wind turbines in the built environment: A review. *Renewable energy*. 2019;131:268-83.
- [48] Sibley J, Richards R, Krikke R, Forney S, editors. *The caterpillar 3176 heavy duty diesel engine*. *SAE International Truck and Bus Meeting and Exposition*; 1988: SAE Technical Paper.
- [49] Aeggegn DB, Agajie TF, Workie YG, Khan B, Fopah-Lele A. Feasibility and techno-economic analysis of PV-battery priority grid tie system with diesel resilience: A case study. *Heliyon*. 2023;9(9).
- [50] Hiraishi T, Krug T, Tanabe K, Srivastava N, Baasansuren J, Fukuda M, et al. 2013 supplement to the 2006 IPCC guidelines for national greenhouse gas inventories: Wetlands: IPCC Switzerland; 2014.
- [51] Birol F. *World energy outlook 2006*. International Energy Agency. 2009.
- [52] Atcitty S, Neely J, Ingersoll D, Akhil A, Waldrip K. Battery energy storage system. *Power Electronics for Renewable and Distributed Energy Systems: A Sourcebook of Topologies, Control and Integration*: Springer; 2013. p. 333-66.
- [53] Lai X, Yao J, Jin C, Feng X, Wang H, Xu C, et al. A review of lithium-ion battery failure hazards: Test standards, accident analysis, and safety suggestions. *Batteries*. 2022;8(11):248.
- [54] Truong CN, Naumann M, Karl RC, Müller M, Jossen A, Hesse HC. Economics of residential photovoltaic battery systems in Germany: The case of Tesla's Powerwall. *Batteries*. 2016;2(2):14.
- [55] Ramdas A, McCabe K, Das P, Sigrin BO. California Time-of-Use (TOU) transition: Effects on distributed wind and solar economic potential. *National Renewable Energy Laboratory (NREL), Golden, CO (United States)*; 2019.
- [56] Hamedani EA, Araghi AR, Hajinezhad A, Yousefi H. A Scenario-Based Techno-Economic Optimization Framework for Power-to-X Residential Energy Hubs Across Diverse Climate Conditions. *Energy*. 2026:141155.
- [57] Bahrami Yajloo A, Abbasian Hamedani E, Maleki P, Hosseinpour M. Optimization of economic dispatch for distributed generation-based power networks. *Advances in Energy Sciences and Technologies*. 2025;1(3):266-79.
- [58] Wang Y, Qiu D, Sun M, Strbac G, Gao Z. Secure energy management of multi-energy microgrid: A physical-informed safe reinforcement learning approach. *Applied Energy*. 2023;335:120759.
- [59] Leite GMC, Jiménez-Fernández S, Salcedo-Sanz S, Marcelino CG, Pedreira CE. Solving an energy resource management problem with a novel multi-objective evolutionary reinforcement learning method. *Knowledge-Based Systems*. 2023;280:111027.
- [60] Domínguez-Barbero D, García-González J, Sanz-Bobi MÁ, García-Cerrada A. Energy management of a microgrid considering nonlinear losses in batteries through Deep Reinforcement Learning. *Applied Energy*. 2024;368:123435.
- [61] Xiong B, Zhang L, Hu Y, Fang F, Liu Q, Cheng L. Deep reinforcement learning for optimal microgrid energy management with renewable energy and electric vehicle integration. *Applied Soft Computing*. 2025;176:113180.