A Deep Reinforcement Learning Framework for Efficient and Resilient Power Management system in Medium Voltage DC Ship Power Systems

Klaudia Bis[1*], Laxman Timilsina[1], Christopher S. Edrington[1], Behnaz Papari[2], Asif Ahmed Khan[2] Gokhan Ozkan[1]
Emails: {kbis*;ltimils;cedring;bpapari;asifahm;gokhano}@clemson.edu
[1]*Holcombe Department of Electrical and Computer Engineering, Clemson University, Clemson 29634, SC, USA*
[2]*Department of Automotive Engineering-Clemson University, Greenville, SC 29607 USA*

2025 IEEE Electric Ship Technologies Symposium (ESTS) | 979-8-3315-2517-0/25/$31.00 ©2025 IEEE | DOI: 10.1109/ESTS62818.2025.11152475

***Abstract*—Next-generation naval vessels require highly efficient and robust power management systems to accommodate the dynamic and mission-critical demands of on-board loads. Traditional deterministic control methodologies often struggle to adapt to rapidly changing conditions, leading to suboptimal resource allocation and increased susceptibility to faults. To address these challenges, this paper presents a novel power management framework for a two-zone electric ship power system using a Deep Deterministic Policy Gradient (DDPG) agent implemented in MATLAB's Reinforcement Learning Toolbox. The simplified two-zone Simulink model captures the essential interactions between power generation and load distribution in an electric ship environment, facilitating both agent training and performance evaluation. Preliminary results demonstrate that the proposed DDPG-based power manager achieves enhanced power allocation, fast adaptation to load variations, and improved resilience compared to conventional control strategies.**

***Index Terms*—power management, reinforcement learning, DDPG agent, neural network**

I. Introduction

The growing electrification of naval vessels places stringent demands on power management systems, which must ensure reliable and efficient operation across a wide array of mission profiles and load dynamics. Modern electric ship designs typically incorporate multiple power sources, energy storage systems, and complex distribution networks, all of which contribute to the complexity of real-time power allocation. As shipboard systems become increasingly reliant on electricity for propulsion, mission-critical sensors, and other operational loads, the reliability and adaptiveness of power management strategies become paramount [1], [2].

Traditional power management (PM) systems in ship power systems (SPS) rely on deterministic control methods such as rule-based logic, proportional-integral-derivative (PID) controllers, and model predictive control (MPC). These methods, while effective for steady-state operations, often struggle with dynamic load variations, system uncertainties, and unforeseen disturbances [3]. Rule-based approaches lack adaptability, as they require predefined conditions that may not account for all possible scenarios. PID controllers, while effective for local stability, are limited in handling multi-variable interactions. On the other hand, MPC, though predictive, demands high computational resources and depends on an accurate system model, which can be challenging to obtain in complex SPS environments [4], [5].

Droop control, a traditional decentralized technique, is commonly used for voltage regulation and load sharing in ship power management systems. It adjusts the output voltage of power sources based on load current, enabling automatic load distribution without requiring fast communication. In DC ship power systems, voltage droop control ensures that as the load increases, voltage decreases according to a predefined droop characteristic. Although simple and effective, droop control has several limitations, including steady-state voltage deviations, slow dynamic response, inefficient load sharing, and a fixed droop coefficient that cannot adapt to real-time conditions [6], [7].

Machine learning (ML) techniques, particularly Reinforcement Learning (RL), offer a more adaptive and data-driven approach to power management. RL-based PM systems continuously learn from operational data, enabling real-time decision-making without requiring an explicit system model. Techniques such as Deep Deterministic Policy Gradient (DDPG) can optimize power distribution by dynamically adjusting power flow and energy storage utilization based on system conditions. Unlike traditional methods, RL can proactively respond to load changes, improve system resilience, and enhance energy efficiency through intelligent adaptation [8], [9]. By leveraging RL, SPS can achieve autonomous, self-optimizing power management, reducing the need for manual tuning and enhancing overall system reliability in dynamic maritime environments.

RL agents learn optimal control policies through interaction with the environment, receiving rewards or penalties based on the outcomes of their actions [10], [11]. This trial-and-error learning allows RL to handle uncertainties, nonlinearities, and dynamic variations that challenge traditional control methods. RL has been successfully applied to a variety of control

This material is based upon research supported by the US Office of Naval Research (ONR) under award number N00014-22-1-2558, and it is approved for public release under DCN# XXX-XXX-XX.

problems, from robotic manipulation [12], [13] to autonomous driving [14].

Among various RL approaches, deep reinforcement learning (DRL) methods have garnered significant attention due to their ability to handle large state and action spaces and to learn complex control policies. The Deep Deterministic Policy Gradient (DDPG) algorithm is particularly appealing for continuous control problems such as real-time power management. It employs a hybrid actor-critic strategy, leveraging deep neural networks to approximate both the policy (actor) and the value function (critic) [10]. Recent advances in deep reinforcement learning (DRL), combining RL with deep neural networks, have enabled RL to scale to high-dimensional state and action spaces [11]. DRL agents can learn complex control policies directly from raw sensor data, such as images, without requiring hand-engineered features or state estimators. This end-to-end learning capability makes DRL particularly well-suited for complex, data-rich environments like ship power systems.

In the context of ship power management, DDPG offers several advantages over traditional methods. Its ability to learn complex, nonlinear control policies allows it to adapt to dynamic load variations and system uncertainties. The model-free nature of DDPG enables it to learn directly from operational data, without requiring an explicit system model. Furthermore, the off-policy learning of DDPG allows it to learn from a replay buffer of past experiences, improving sample efficiency and stability.

This capacity to operate in continuous action spaces aligns well with the challenges of power flow control, where the power commands to generators, storage units, or other devices must be smoothly and continuously managed.

A. Statement of Contribution

This paper presents a novel reinforcement learning-based power management framework for an MVDC SPS, utilizing the Deep Deterministic Policy Gradient (DDPG) algorithm. To effectively test the developed power management algorithm, the study simplifies a four-zone SPS into a two-zone SPS model. Unlike traditional deterministic or rule-based control methods, which struggle to adapt to dynamic load variations and unforeseen disruptions, the proposed approach enables intelligent, real-time power allocation and enhances system resilience. The framework is implemented using MATLAB's Reinforcement Learning Toolbox and validated through simulations, demonstrating its ability to improve power distribution efficiency, stabilize bus voltages, and optimize energy storage utilization. By systematically evaluating the agent's performance under various operational conditions, this work highlights the potential of deep reinforcement learning for next-generation naval power management, paving the way for more adaptive and autonomous control strategies in electric ship systems. The simulation is conducted in MATLAB/Simulink to validate the proposed power management system.

The rest of this paper is organized as follows. Section II provides an overview of the system model and its core components. Section III details the design and implementation of the proposed DDPG agent within the MATLAB Reinforcement Learning Toolbox framework. Section IV presents simulation results and performance evaluations under various operational conditions. Finally, Section V concludes the paper, summarizing key findings and outlining directions for future work.

II. Ship Power System Control Architecture

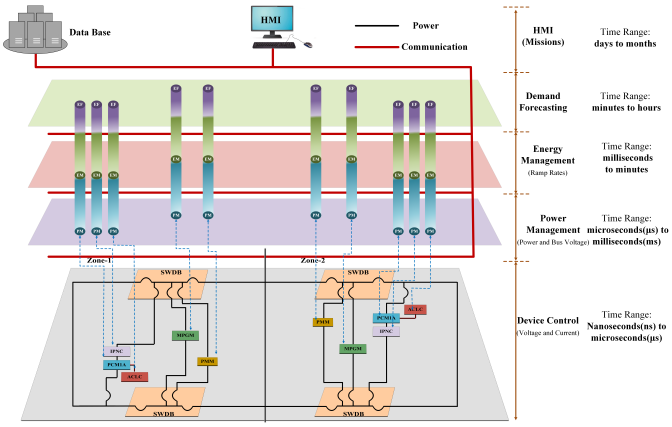

Fig. 1. Control architecture of a 2-zone ship power system.

The SPS control architecture is structured as a hierarchical multi-layered framework designed to ensure efficient power management, stability, and adaptability in an MVDC ship system as seen in Figure 1. The system is divided into multiple control layers, each operating at different time scales to address various aspects of power distribution, energy management, and system resilience. At the highest level, there is an Energy Management (EM) Layer that plays a crucial role in dynamically allocating the percentage of load to the sources present in the system. This layer functions on a time scale of milliseconds to minutes, optimizing power distribution based on demand and availability. Directly beneath it, the Power Management (PM) Layer is responsible for real-time power flow and bus voltage regulation, reacting to fluctuations in microseconds to milliseconds to maintain system stability. The lowest level, the Device Control Layer (DCL), governs voltage and current at individual components, responding within nanoseconds to microseconds to ensure fine-grained control over the sources, power electronics and distribution units.

A. Two-Zone Configuration

Naval power systems increasingly adopt a zonal approach, as depicted in references such as [15]. Each zone houses local generation, distribution, and loads, compartmentalizing damage and fostering redundancy:

- Operational Flexibility: Zones can be isolated if faults or battle damage occur, maintaining power flow in the remaining zones.

- Scalability: Additional zones can be appended as the vessel's requirements grow, without redesigning the entire system.
- Distributed Control: Power management tasks (e.g., voltage regulation, load balancing) can be handled locally within each zone or globally via a supervisory controller.

In a fully realized MVDC vessel, four or more zones might be arranged along the ship's length, each featuring multiple generators and loads. However, to streamline initial investigations, this paper focuses on two representative zones, each capturing the fundamental dynamics of droop load sharing, energy-storage buffering, and switchboard distribution.

1) Power Generation Module (PGM): Each zone has a Power Generation Module that models the behavior of a generator unit. This study assumes an idealized DC source with a small slope of voltage reduction as the current increases. Practical PGMs have rated power and current constraints, which are incorporated into the simulation by capping the maximum generator output.

2) Power Conversion Module (PCM): Realized as an energy storage block modeled by a controllable current source, each PCM supports transient and peak load demands. In a real shipboard power system, such modules might be batteries or other forms of energy storage. Their main function is to smooth out rapid changes in power demand and to support bus voltage stability.

3) Switchboard (SWBD): Each zone has a Switchboard (SWBD) or DC bus that interconnects the local PGM, PCM, and share of the load. Buses are modeled as ideal electrical nodes with minimal line impedances. In full-scale MVDC vessels, switchboards would include breakers, sensors, and reconfiguration logic to route power around damage or faults. For RL training, this work prioritizes power flow equations over detailed protection schemes.

4) Loads: The dominant load on most electric-driven ships is propulsion. In this simplified model, all the main loads of the ship - including the propulsion, auxiliary systems and 'hotel' loads - are aggregated into a single constant power load. Although real-world naval vessels incorporate multiple propulsion motors and a variety of auxiliary loads, the use of one high-power load captures the primary challenge of managing large power draws under different operating conditions. Both zones supply this load, with their individual contributions determined by the droop characteristics of the PGMs and by the dispatch decisions of the power manager. While real ships have multiple drives and a variety of smaller loads, combining them simplifies the RL environment, maintaining enough complexity to test load sharing and dynamic bus regulation.

B. Energy Management

Energy management (EM) in a two-zone MVDC ship power system serves as the supervisory layer that coordinates distributed generation modules, energy storage devices, and load demands to maintain system reliability, efficiency, and resilience under dynamic operating conditions. Unlike local power management—which focuses on regulating voltage and current sharing in real time—EM deals with broader scheduling and resource allocation tasks such as generator dispatch, state-of-charge (SOC) maintenance, and ramp-rate constraints [16], [2].

A primary challenge in shipboard EM is the presence of large pulse loads, which can demand steep power ramps beyond the capability of conventional generators. To address this, the EM layer orchestrates the charging and discharging of energy storage modules so that they provide fast transient support, thereby mitigating the impact on the generators' ramp-rate limits. This coordinated approach ensures that the bus voltage and overall power flow remain stable, even as propulsion loads or pulse loads vary rapidly.

The EM layer is essential to ensure that the total load demand is met during mission operations, with each generator providing optimal output power. The optimization process aims to minimize the cost associated with power generated by each generator, as well as the power transferred from neighboring areas to each node. This study considers a non-convex and nonlinear cost function as described below taken from the previous study [17]:

$$C_i(P_i) = \left(\alpha_0 + \alpha_1 P_i + \alpha_2 P_i^2 + \cdots + \alpha_m P_i^m\right) P_i \qquad (1)$$

where P_i is the generated power of the i^{th} generator and α_k $(k = 1, 2, 3, \ldots, m)$ are scalar coefficients assigning both positive and negative amounts.

Here, a distributed control is implemented for the EM in this study. For N numbers of generators, each generator has its own controller that regulates the power it generates and has to share this data with neighboring nodes. The cost of the power exchanged with a neighboring area is the amount node i needs to share node j to supply P_{ji} can be expressed as follows:

$$\sum_{j \in N_i} f_i(P_{ji}) = C\left(\sum_{j \in N_i} P_{ji} + \sum_{j \in N_i} \alpha_{ji} P_{ji}^2\right) \qquad (2)$$

where P_{ji} is the shared power, C is the price factor, and α_{ji} is related to power loss. Due to the size of the electric network for a simplified SPS, the power loss of transmission lines is neglected. Thus, finally, the total cost function and the optimization problem in each controller can be represented as:

$$\min_{P_i, P_{ji}} \left(C_i(P_i) + \sum_{j \in N_i} f_i(P_{ji})\right) \qquad (3)$$

$$\begin{aligned} \text{subject to} \quad & \sum_{j \in N_i} P_{ji} + P_i = P_{L,i} \\ & P_i^{\min} \le P_i \le P_i^{\max} \\ & P_{ji}^{\min} \le P_{ji} \le P_{ji}^{\max} \\ & P_{ji}^{\min,1} \le \sum_{j \in N_i} P_{ji} \le P_{ji}^{\max,1} \end{aligned}$$

where P_i, P_{ji}, and $P_{L,i}$ are the generator output power, the power transferred from node j to node i, and the load demand in each generator area, respectively.

Various methods can be used to find the optimal solution to the nonlinear 3, including particle swarm optimization (PSO), genetic algorithm (GA), ant colony optimization (ACO), and crow search algorithm (CSA). For this study, the CSA method is selected because of its simplicity, flexibility, fast convergence speed, and ability to avoid getting trapped in local optima, which has been successfully used in one of the previous study [17].

In CSA model, each crow's position corresponds to a subset of features from the global set. Each variable in the optimization problem represents a dimension of the general variable X_l. For example, if there are three variables in the problem, then the dimension of X_l will be three. In this method, X_l represents a crow's position, with N crows in the group. Each crow has its own location, which is updated throughout the process based on the flight length f_l. The optimal solution is determined by the crow's location with the minimum or maximum cost function value. The update process of the crow's location X_l is described as [18]:

$$X_l^{(i,\text{iteration}+1)} = \begin{cases} X_l^{(i,\text{iteration})} + r_i \cdot f_l \cdot \left(m^{(j,\text{iteration})} - X_l^{(i,\text{iteration})} \right), \\ \qquad \text{if } r^{(j,\text{iteration})} \geq AP^{(j,\text{iteration})} \\ \text{a random number,} \qquad \text{otherwise} \end{cases} \tag{4}$$

where $X_l^{(i,\text{iteration})}$ is the location of the i-th crow in the current iteration, f_l is the flight length, $m^{(j,\text{iteration})}$ is identified as the memory location of the j-th crow, which is the best location for the j-th crow to hide food after the iteration. $\text{AP}^{(j,\text{iteration})}$ represents the awareness probability of the j-th crow in the iteration, and r_i is a random number subject to uniform distribution in $[0, 1]$.

The details of the CSA implementation can be found in the study [17]. In this study, after implementing the CSA in the EM layer of the SPS, two parameters are provided as input to the PM layer: the percentage of the load to be shared between the two generators considered in this study.

C. Reinforcement Learning based Power Management System

A central feature of this work is the Reinforcement Learning (RL) block, implemented using MATLAB's Reinforcement Learning Toolbox. This RL "agent" sends setpoint commands to both the PGMs and PCMs in real time:

1) PGMs (Generators): In the control strategy implemnented in this paper, the RL agent computes power references for the generators based on the total load demand provided by the energy management layer. The resulting power references are then constrained to lie between 0 and the maximum rated power of 34.8 MW, ensuring that the generators operate within their safe operational bounds.

2) PCMs (Energy Storage): In this framework, the RL agent also modulates the energy storage modules to support transient load variations and maintain bus voltage stability. Two elements of the agent's action vector are mapped to battery power commands. Each action directly determines the charging or discharging rate of the corresponding energy storage module. A positive command indicates discharge (supplying power to the system), while a negative command signifies charging (absorbing power). This rapid adjustment capability enables the energy storage modules to compensate for sudden load swings, thereby smoothing power fluctuations and relieving stress on the generators during peak demand.

By continuously monitoring bus voltages, load demands, and other critical state variables, the RL agent develops an optimal control policy that coordinates the operation of both the generators and energy storage modules. Its primary objective is to maintain a stable power flow by keeping the bus voltage within tight limits, while dynamically adjusting the power outputs to match fluctuating load conditions. In doing so, the agent ensures that the generators provide the baseline power required for normal operations, and the energy storage units are reserved to respond to transient peak loads, thereby maximizing overall system efficiency and reliability across a broad range of operating scenarios.

D. Model Simplifications and Assumptions

Although the two-zone MVDC system reflects the principal dynamics of a zonal shipboard power system, several simplifications are adopted to reduce computational overhead and streamline training for the RL agent:

1) Ideal Source Representation: The PGMs are treated as ideal voltage sources with droop control instead of modeling prime movers, generator windings, and detailed power electronics.

2) Aggregated Load: All significant ship loads are approximated by a single constant-power propulsion load, which captures the high-power demand behavior while omitting secondary load complexities.

3) Reduced Zonal Complexity: Each zone contains only one generator, one energy storage module, and one switchboard. This level of abstraction preserves the essential control and power-sharing dynamics while avoiding the additional complexity of multiple distributed loads or ancillary systems.

These assumptions preserve the key features of zonal power management—voltage regulation, load sharing, and transient response—while enabling more rapid simulations and facilitating practical RL training. As a result, the two-zone framework serves as a tractable yet representative test environment for validating advanced reinforcement learning approaches to shipboard power management.

III. Control Methodology Using a DDPG Agent

Reinforcement Learning (RL) demonstrates strong capabilities in handling high-dimensional and partially observed control problems, thanks in part to advances in deep neural networks. RL provides a normative framework for sequential

decision making, in which an agent learns an optimal policy by interacting with an environment and maximizing cumulative future rewards. In the context of ship power management, RL can allow the power controller to adapt in real time to dynamic load changes, varying generator performance, and other uncertainties, thereby improving overall system resilience. Below three key RL concepts are briefly reviewed: the Markov property, action–value functions, and reward design. Next, the Deep Deterministic Policy Gradient (DDPG) approach that we adopt in this work is introduced.

A. Reinforcement Learning Formulation

In this study, ship power management is formulated as a Markov Decision Process (MDP) defined by the tuple

$$\langle \mathcal{S}, \mathcal{A}, P, r, \gamma \rangle \tag{5}$$

where $\mathcal{S}$ denotes the state space (e.g., bus voltages, generator output power, energy storage state-of-charge), $\mathcal{A}$ the continuous action space (e.g., power setpoint commands to generators and energy storage modules), $P(s_{t+1} \mid s_t, a_t)$ the state transition probabilities, $r(s_t, a_t)$ the immediate reward function, and $\gamma \in (0, 1)$ the discount factor. The Markov property ensures that the future state depends solely on the current state and control action:

$$P(s_{t+1} \mid s_t, a_t, s_{t-1}, a_{t-1}, \dots) = P(s_{t+1} \mid s_t, a_t) \tag{6}$$

This assumption simplifies the representation of ship power system dynamics, meaning that the next operating condition (e.g., bus voltage or load sharing) is determined by the present measurements and applied power commands without the need to consider the full past history.

A key concept in reinforcement learning is the action–value function $Q^\pi(s_t, a_t)$, which estimates the expected cumulative future reward when taking action a in state s and following policy π thereafter. The theoretical foundation for this approach is provided by the Bellman equation:

$$Q^\pi(s_t, a_t) = \mathbb{E}\left[r(s_t, a_t) + \gamma Q^\pi(s_{t+1}, \pi(s_{t+1}))\right]. \tag{7}$$

which establishes a recursive relationship between the value of the current state–action pair and that of subsequent pairs. This equation is critical in the presented implementation: the critic network is trained to minimize the temporal-difference (TD) error between its estimated Q-values and the targets computed using the Bellman equation [10]

Deep Deterministic Policy Gradient (DDPG) is an off-policy actor–critic method that effectively handles continuous control problems such as real-time power management in ship power systems. In our implementation, the actor network $\mu_\phi(s)$ deterministically maps states to continuous control actions, while the critic network $Q_\theta(s, a)$ evaluates the expected return of these actions. Specifically, the critic is updated to minimize the loss

$$L_{\text{critic}}(\theta) = \mathbb{E}_{(s_t, a_t, r_t, s_{t+1})}\left[\left(Q_\theta(s_t, a_t) - y_t\right)^2\right], \tag{8}$$

The actor, in turn, is updated by following the policy gradient:

$$\nabla_\phi J = \mathbb{E}_{(s_t)}\left[\nabla_a Q_\theta(s_t, a)|_{a=\mu_\phi(s_t)} \nabla_\phi \mu_\phi(s_t)\right]. \tag{9}$$

These updates ensure that the actor learns to select actions that maximize the Q-value estimates, with the critic's learning fundamentally driven by the Bellman equation [10].

The MATLAB implementation integrates the DDPG agent with a Simulink model of a two-zone ship power system. The environment is constructed with a 16-dimensional observation space and a 4-dimensional action space.

During training, the agent interacts in a closed-loop with the Simulink model, receiving measurements such as bus voltage, generator outputs, and battery SOC. The reward function is carefully designed to balance multiple objectives—including voltage regulation, load sharing, and energy storage management—so that the RL agent learns to use the generators primarily for baseline support while reserving the batteries for peak loads.

By combining the theoretical formulation based on the Bellman equation with a robust DDPG implementation and an accurately modeled SPS environment, the control methodology presented in this study offers an adaptive and efficient solution for real-time ship power management. This framework is implemented and validated using MATLAB's Reinforcement Learning Toolbox, demonstrating its potential to advance next-generation naval power systems.

In this paper, exploration is incorporated by adding noise directly to the deterministic actions generated by the actor network. In particular, the agent's exploration is governed by the noise options defined in the DDPG agent options, where the mean attraction constant is set to 1 and the variance is set to 0.1. These parameters ensure that the noise is centered around zero while maintaining sufficient variability to encourage exploration in the continuous action space. By perturbing the actor's output in this manner, the agent is able to sample a diverse set of actions during training, thereby avoiding premature convergence to suboptimal policies. This approach is crucial in the ship power management application, where the control inputs (power setpoint commands) must be finely tuned to achieve stable operation. The chosen exploration settings strike a balance between exploring the state–action space and preserving the stability of the deterministic policy, as detailed in our agent configuration

IV. Simulation Results

As previously explained, the operating data from the SPS simulation are transmitted to the control layer. Initially, all parameters are sent to the EM layer, where the CSA algorithm is executed within a distributed control framework. The algorithm determines the optimal power distribution between the generators, and this output is then forwarded to the PM layer. The PM layer ensures compliance with the constraints set by the EM, allocates the required power to each generator, and maintains voltage stability throughout the system. The flow of

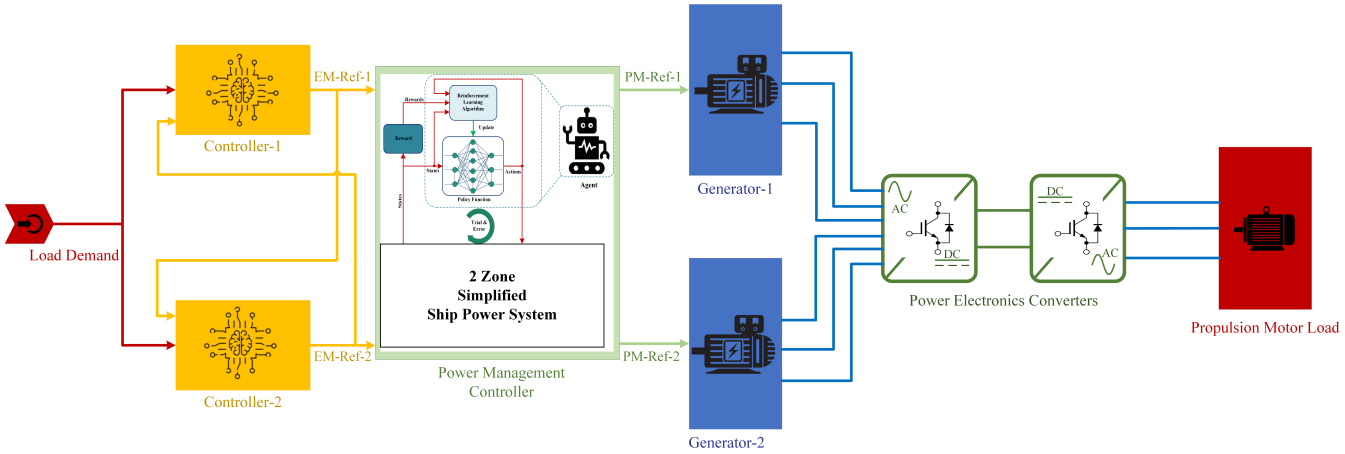

Fig. 2. Simplified SPS with controllers for EM and PM considered for this study.

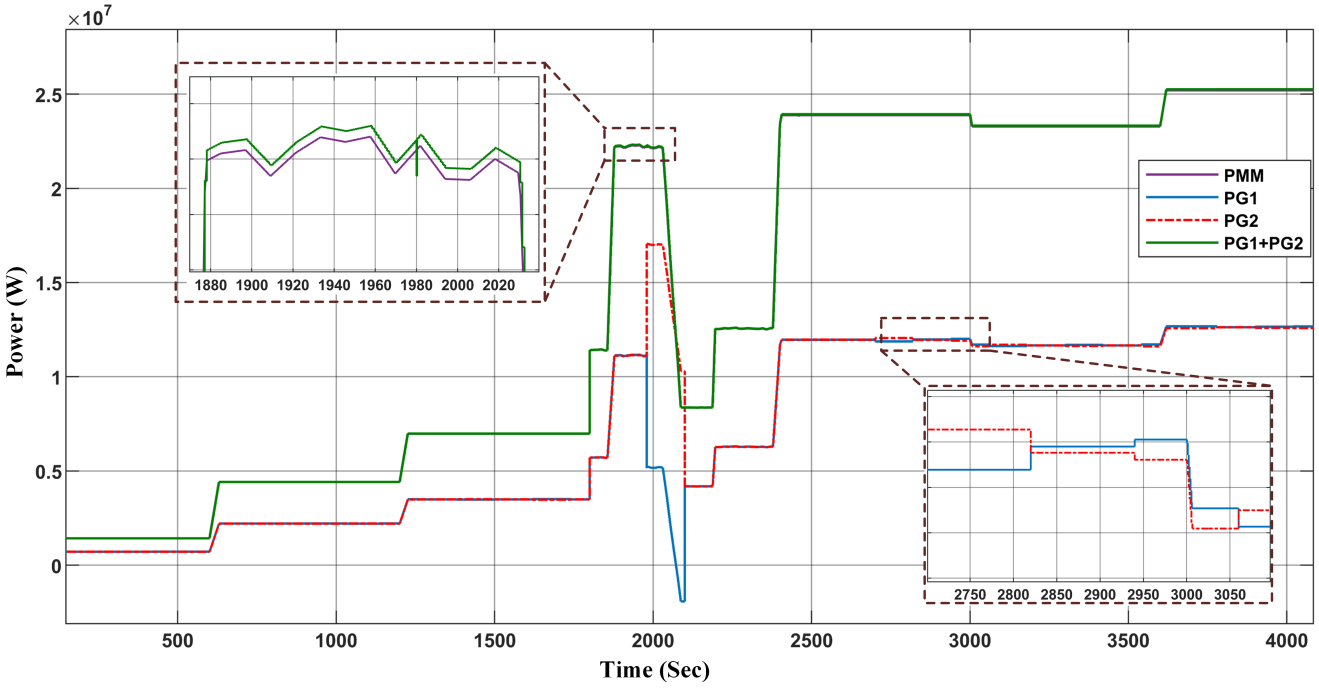

Fig. 3. Different power flows in the simplified SPS.

parameters can be visualized in the figure 2, which shows how the signals are forwarded from one layer to another.

Figure 3 illustrates the power flow dynamics in the simplified SPS considered in this study. This figure presents four distinct power profiles that evolve throughout the SPS operation. The purple line represents the total propulsion power demanded by the ship's loads, which must be supplied by two identical generators. The power contribution of Generator 1 (G_1) is shown in blue (P_{G1}), while Generator 2 (G_2) is represented by the red dashed (P_{G2}) line. The green line, which is the sum of P_{G1} and P_{G2}, represents the total supplied power. Since the total generated power must always match the load demand, the green line should ideally overlap with the purple line, confirming that the power management strategy effectively meets the demand.

A key aspect of this study is that the EM system updates power allocation every two minutes, meaning that new optimized power values are determined at these intervals. This is clearly observable in the second zoomed-in section of the figure, where power values change discretely at two-minute intervals. The CSA-based energy management strategy dynamically adjusts power generation, ensuring optimal load sharing between the two generators in response to demand fluctuations.

Additionally, the first zoomed-in section highlights a critical observation: the total generated power and the load demand

closely match, with only minor variations. This indicates that the CSA algorithm is effective in achieving real-time power balancing while maintaining high accuracy. The minor deviations observed may be attributed to transient effects, small control delays, or system constraints, but overall, the system successfully maintains power equilibrium.

To evaluate the proposed DDPG-based power management strategy, this study trained and tested the RL agent within a simplified two-zone ship power system model in Simulink. The main objectives were (1) bus voltage stability, ensuring the voltage remains near $12kV$; and (2) power sharing, distributing load demand between two Power Generation Modules (PGM1 and PGM2) according to a constant ratio commanded by the higher-level Energy Manager. Figures 3 - 6 summarize key outcomes, including the training progression, bus voltage response, generator power outputs, and load demand.

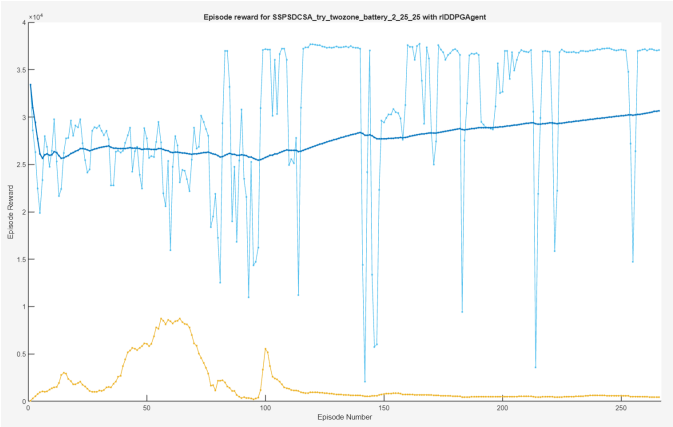

Fig. 4. Episode reward for two-zone SPS system with DDPG agent

Figure 4 depicts the training progress in terms of episode reward versus episode number. Early in training, the reward exhibits large swings as the agent explores the action space (e.g., attempting aggressive power adjustments). Over subsequent episodes, the reward curve stabilizes, indicating that the agent converges to a policy that successfully balances multiple objectives (voltage regulation, power sharing, and avoiding excessive battery usage). Occasional dips reflect ongoing exploration, where the agent tests less-familiar state–action pairs, ensuring it can adapt to rare or extreme conditions. Figure 5 shows the bus voltage profile during a load step and subsequent transients. Initially, the bus experiences an overshoot followed by damped oscillations. The voltage settles around the nominal $12kV$ within roughly $0.02s$, achieving an overall voltage stability of about 99.67% (i.e., the steady-state voltage deviates only about 0.33% from nominal). This rapid settling demonstrates the agent's ability to respond quickly to load changes, adjusting generator setpoints and battery power references to keep the bus voltage within acceptable bounds. Post-transient, the bus voltage remains within a narrow band around $12kV$. This indicates that the agent successfully balances generator outputs with any required battery contribution, minimizing long-term voltage error.

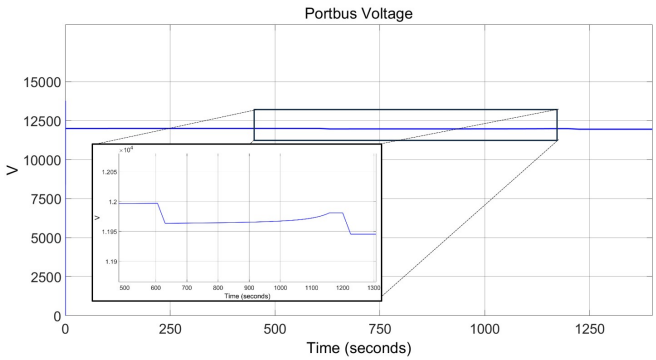

Fig. 5. Portbus voltage

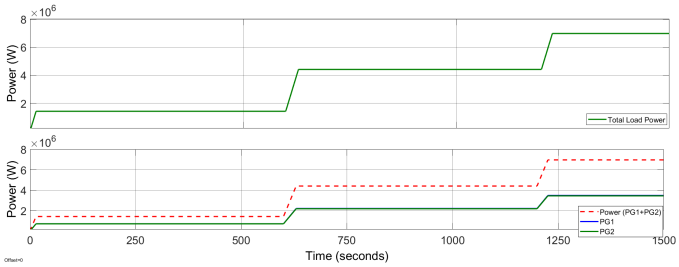

Fig. 6. Power Sharing

The plots in Figure 6 illustrate the real-time load profile and power distribution between PGM1 and PGM2 in response to increasing load demand. The top plot shows the total load power, which gradually increases over time before stabilizing. The bottom plot presents the power output of PGM1 and PGM2, as well as the total generated power. The results indicate that the Energy Manager's predefined power-sharing ratio is effectively maintained, as the RL-based controller successfully distributes the load demand between the two generators. Both PGMs exhibit a smooth increase in power output, closely following the rising load profile. At around 4.5 seconds, when the load demand stabilizes, the power outputs of PGM1 and PGM2 also reach steady-state values, demonstrating the system's ability to converge efficiently.

Notably, there are no significant transient spikes or oscillations in the power output, suggesting that the control strategy ensures a smooth response to load changes. The nearly overlapping total power output and load demand curves further confirm that the power management strategy maintains a balanced and stable operation. While not explicitly depicted in the plot, energy storage modules (PCMs) assist in mitigating transient fluctuations, preventing excessive ramp rates, and reducing mechanical stress on the generators.

The training curve suggests that the agent reliably converged to a stable policy. Empirically, the agent keeps voltage oscillations small and achieves the desired power split. During constant or slowly varying loads, the agent primarily relies on the generators, reserving the batteries for large, sudden demands. In some trials, the agent occasionally dispatches battery power

to manage minor voltage deviations. Future reward shaping could discourage battery usage for base loads, aligning more closely with practical energy-management policies. Although tested on a two-zone system, the same RL approach can be scaled to a multi-zone ship architecture by expanding the observation (bus voltages, generator outputs, battery states) and action spaces (multiple PGMs and PCMs). Additional training time and well-crafted reward functions will likely be necessary to accommodate more complex interactions. In prior deterministic or droop-based methods, voltage stability is typically guaranteed but can lead to suboptimal load sharing or excessive battery usage. The RL agent adapts to changing conditions without explicit system modeling, potentially yielding higher efficiency. The current reward function does not heavily penalize partial battery usage under quasi-steady loads, causing minor battery discharge in non-peak situations. Refining the reward function to reduce battery wear or enforce minimal battery usage under normal conditions would address this shortcoming.

Although the present model focuses on a two-zone MVDC architecture, the methodology scales to more complex ship power systems. The same RL framework—trained with suitable reward functions and state observations—can be extended to a four-zone configuration, where additional generators, energy storage modules, and load types demand a higher-dimensional control policy.

V. Conclusion

This paper presented a novel DDPG-based reinforcement learning (RL) framework for managing power flows in a simplified two-zone MVDC ship power system. By learning a policy directly from simulated operational data, the RL agent demonstrated robust performance in maintaining bus voltage near its nominal level (within $\pm 0.33\%$) and distributing load demand according to the energy manager's setpoints. Unlike conventional deterministic methods, the proposed approach is inherently adaptive, requiring no explicit modeling of complex SPS dynamics and offering improved resilience under sudden load changes.

Despite these encouraging results, future work can refine the reward function to minimize unnecessary battery usage under constant loads, thus reserving battery capacity for peak demands. Moreover, scaling the RL solution to a four-zone or larger MVDC architecture—potentially with multiple batteries and more varied loads—will require careful design of observation/action spaces and additional training episodes. Finally, incorporating generator fuel efficiency or other cost factors into the reward function would enable a more holistic optimization that aligns with real-world operational objectives.

Overall, these findings highlight the promise of deep RL methods in shipboard power management, offering a flexible, model-free approach capable of meeting voltage and load-sharing requirements under challenging operating scenarios. By further tuning reward designs and exploring multi-zone extensions, RL-driven strategies can help pave the way toward more efficient and resilient naval power systems.

References

[1] D. Gonsoulin, T. Vu, F. Diaz, H. Vahedi, D. Perkins, and C. Edrington, "Centralized MPC for multiple energy storages in ship power systems," in *IECON 2017 - 43rd Annual Conference of the IEEE Industrial Electronics Society*. Beijing: IEEE, Oct. 2017, pp. 6777–6782.

[2] S. Faddel, M. E. Hariri, A. A. Saad, and O. Mohammed, "Intelligent power management for the hybrid energy storage of the ship power system," *IEEE Transactions on Energy Conversion*, vol. 32, no. 2, pp. 798–809, Jun. 2017.

[3] S. Paudel, J. Zhang, B. Ayalew, M. Castanier, and A. Skowronska, "A model predictive control-based operation of vehicle-borne microgrid considering battery degradation," in *Proceedings of the Ground Vehicle Systems Engineering and Technology Symposium*, pp. 15–17.

[4] P. Xie, S. Tan, J. M. Guerrero, and J. C. Vasquez, "Mpc-informed ecms based real-time power management strategy for hybrid electric ship," *Energy Reports*, vol. 7, pp. 126–133, 2021.

[5] H. Yang, T. Li, Y. Long, and Y. Xiao, "Model-predictive-based power management for pulsed power load fed with multizonal shipboard power system," *IEEE Transactions on Transportation Electrification*, vol. 10, no. 2, pp. 4119–4128, 2023.

[6] D. Perkins, T. Vu, H. Vahedi, D. Gonsoulin, and C. S. Edrington, "Distributed power management for dc distribution system with model uncertainties," in *2017 IEEE Electric Ship Technologies Symposium (ESTS)*. IEEE, 2017, pp. 534–538.

[7] T. V. Vu, D. Perkins, F. Diaz, D. Gonsoulin, C. S. Edrington, and T. El-Mezyani, "Robust adaptive droop control for dc microgrids," *Electric Power Systems Research*, vol. 146, pp. 95–106, 2017.

[8] S. Ghimire, D. Joshi, V. M. Venkatasubramanian, G. Torresan, P. Chabas, and S. Hivart, "Wide-area measurement based online oscillation alarming system at rte," in *2024 IEEE International Conference on Communications, Control, and Computing Technologies for Smart Grids (SmartGridComm)*. IEEE, 2024, pp. 412–418.

[9] D. Joshi, S. Ghimire, S. S. Shiuab *et al.*, "Computational analysis of online oscillation monitoring algorithms," in *2024 IEEE International Conference on Communications, Control, and Computing Technologies for Smart Grids (SmartGridComm)*. IEEE, 2024, pp. 366–371.

[10] T. P. Lillicrap, J. J. Hunt, A. Pritzel, N. Heess, T. Erez, Y. Tassa, D. Silver, and D. Wierstra, "Continuous control with deep reinforcement learning," Jul. 2019, arXiv:1509.02971 [cs].

[11] V. Mnih, K. Kavukcuoglu, D. Silver, A. A. Rusu, J. Veness, M. G. Bellemare, A. Graves, M. Riedmiller, A. K. Fidjeland, G. Ostrovski, S. Petersen, C. Beattie, A. Sadik, I. Antonoglou, H. King, D. Kumaran, D. Wierstra, S. Legg, and D. Hassabis, "Human-level control through deep reinforcement learning," *Nature*, vol. 518, pp. 529–533, Feb. 2015.

[12] S. Ohnishi, E. Uchibe, Y. Yamaguchi, K. Nakanishi, Y. Yasui, and S. Ishii, "Constrained deep q-learning gradually approaching ordinary q-learning," *Frontiers in Neurorobotics*, vol. 13, p. 103, 2019.

[13] Z. Zhu, P. Li, and L. Meng, "Deep deterministic policy gradient for data-driven voltage control in power systems," *IEEE Transactions on Smart Grid*, vol. 11, no. 2, pp. 1345–1356, Mar. 2020.

[14] W. Zhou, D. Chen, J. Yan, Z. Li, H. Yin, and W. Ge, "Multi-agent reinforcement learning for cooperative lane changing of connected and autonomous vehicles in mixed traffic," *Autonomous Intelligent Systems*, vol. 2, no. 5, pp. 1–14, 2022.

[15] D. Perkins and T. Vu, "Distributed power management implementation for zonal mvdc ship power systems," in *IEEE Electric Ship Technologies Symposium (ESTS)*, 2023, pp. 1–6.

[16] E. Nivolianiti, Y. L. Karnavas, and J.-F. Charpentier, "Energy management of shipboard microgrids integrating energy storage systems: A review," *Renewable and Sustainable Energy Reviews*, vol. 189, p. 114012, 2024.

[17] B. Papari, G. Ozkan, H. P. Hoang, P. R. Badr, C. S. Edrington, H. Parvaneh, and R. Cox, "Distributed control in hybrid ac-dc microgrids based on a hybrid mcsa-admm algorithm," *IEEE Open Journal of Industry Applications*, vol. 2, pp. 121–130, 2021.

[18] A. Askarzadeh, "A novel metaheuristic method for solving constrained engineering optimization problems: crow search algorithm," *Computers & structures*, vol. 169, pp. 1–12, 2016.