

Article

Industrial Process Control Based on Reinforcement Learning: Taking Tin Smelting Parameter Optimization as an Example

Yingli Liu ^{1,2}, Zheng Xiong ^{1,2}, Haibin Yuan ³, Hang Yan ³ and Ling Yang ^{1,2,*} ¹ Faculty of Information Engineering and Automation, Kunming University of Science and Technology, Kunming 650093, China² Yunnan Key Laboratory of Computer Technologies Application, Kunming University of Science and Technology, Kunming 650500, China³ Tin Branch of Yunnan Tin Company Limited, Gejiu 661017, China

* Correspondence: yangling@kust.edu.cn

Abstract

To address the issues of parameter setting, reliance on human experience, and the limitations of traditional model-driven control methods in handling complex nonlinear dynamics in the tin smelting industrial process, this paper proposes a data-driven control approach based on improved deep reinforcement learning (RL). Aiming to reduce the tin entrainment rate in smelting slag and CO emissions in exhaust gas, we construct a data-driven environment model with an 8-dimensional state space (including furnace temperature, pressure, gas composition, etc.) and an 8-dimensional action space (including lance parameters such as material flow, oxygen content, backpressure, etc.). We innovatively design a Dual-Action Discriminative Deep Deterministic Policy Gradient (DADDPG) algorithm. This method employs an online Actor network to simultaneously generate deterministic and exploratory random actions, with the Critic network selecting high-value actions for execution, consistently enhancing policy exploration efficiency. Combined with a composite reward function (integrating real-time Sn/CO content, their variations, and continuous penalty mechanisms for safety constraints), the approach achieves multi-objective dynamic optimization. Experiments based on real tin smelting production line data validate the environment model, with results demonstrating that the tin content in slag is reduced to between 3.5% and 4%, and CO content in exhaust gas is decreased to between 2000 and 2700 ppm.

Keywords: reinforcement learning; industrial process optimization; tin smelting; data-driven

Academic Editor: Paulo Santos

Received: 22 December 2025

Revised: 22 January 2026

Accepted: 29 January 2026

Published: 30 January 2026

Copyright: © 2026 by the authors.

Licensee MDPI, Basel, Switzerland.

This article is an open access article distributed under the terms and conditions of the [Creative Commons Attribution \(CC BY\)](https://creativecommons.org/licenses/by/4.0/) license.

1. Introduction

In the industrial process control of tin smelting, one of the core objectives is to achieve precise tracking of key process parameters, such as smelting temperature, oxygen-enriched lance flow rate, and fuel coal flow setpoint in closed-loop regulation. However, with the rapid development of green metallurgy and intelligent manufacturing, merely pursuing parameter tracking can no longer meet production demands. It is necessary to expand control objectives to multidimensional performance optimization, which integrates product quality, environmental constraints, and other factors, aiming to achieve a comprehensive enhancement of process stability and economic-environmental performance. This is referred to as smelting process operational optimization control. Nevertheless, when factory operations deviate from the optimized targets, the consequences are often severe. Based on

field surveys of tin smelting plants and discussions with management personnel, each furnace run deviation can result in refining cost losses of up to hundreds of thousands of RMB (data sourced from a survey conducted in September 2024 at a tin factory). Additionally, this may lead to equipment failures such as furnace scaling and lance clogging and even production interruptions due to exceeding environmental standards, causing even greater economic losses.

To address the complex dynamic trade-offs among “quality, green, and safety,” traditional control methods face significant limitations in the tin smelting process. For instance, factories currently rely heavily on manual experience to configure key parameters (such as fuel coal flow), but this approach is constrained by subjective judgment and historical data, making it difficult to respond in real-time to dynamic furnace condition fluctuations. This results in unstable product quality (e.g., tin metal entrainment rate), exceeded environmental indicators (e.g., CO concentration), and equipment safety risks (e.g., lance clogging). Similarly, rule-based control depends on fixed empirical rules, which cannot adapt to nonlinear chemical reactions or multi-objective optimization needs. In contrast, data-driven methods based on Deep Reinforcement Learning (DRL) [1,2] learn through interaction with the environment, enabling adaptation to complex dynamic environments and optimization of long-term benefits, thus offering a promising solution. DRL is an intelligent control method that learns optimal strategies through environmental interaction, characterized by its ability to handle high-dimensional nonlinear systems, adapt to dynamic changes, and optimize long-term gains. In the tin smelting process, DRL, through data-driven environmental modeling, dynamically adjusts lance parameters and fuel ratios, ensuring metallurgical reaction stability while maximizing overall economic benefits [3]. Therefore, the purpose of this study is to develop a data-driven deep reinforcement learning (DRL)-based operational optimization control method for the tin smelting process, aiming to simultaneously reduce the Sn content in slag and the CO concentration in exhaust gas under complex and highly dynamic furnace conditions, while maintaining process stability and safe operation.

The tin smelting process has long faced significant challenges in control and optimization due to its high-temperature chemical reactions, complex material composition, and highly dynamic operating conditions. Existing research can be mainly categorized into two types. The first type comprises traditional mechanistic and statistical methods, such as slag-metal separation modeling based on thermodynamic equilibrium [4] and parameter optimization using statistical techniques [5]. These approaches heavily rely on static mechanistic models or offline-designed empirical rules. Once the furnace condition deviates from the designed operating point (e.g., fluctuations in raw material composition or sudden changes in furnace temperature), their optimization performance deteriorates rapidly, making real-time multi-objective regulation difficult to achieve. The second type consists of supervised learning and shallow data-driven methods that have emerged in recent years. Compared with traditional mechanistic and statistical methods, these approaches demonstrate superior performance in nonlinear representation capability, prediction accuracy, and robustness to process disturbances. For example, in copper smelting, neural network models have been employed to simulate and predict the distribution of hazardous elements, thereby optimizing operational parameters such as blast volume and oxygen supply [6]. In aluminum die-casting processes, researchers have achieved high-precision quality prediction of castings by combining shallow neural networks with feature selection techniques [7]. Other scholars have proposed hybrid soft-sensor frameworks integrating shallow and deep learning, significantly enhancing the predictive capability for quality variables in complex industrial processes [8].

However, the aforementioned methods essentially remain at the level of static mapping, making it difficult to achieve closed-loop automatic regulation in continuous time series and dynamically balance conflicting multiple objectives. In recent years, research has introduced deep reinforcement learning (DRL) into similar high-temperature smelting and precision manufacturing scenarios. For example, in intelligent manufacturing scheduling, DRL has significantly reduced downtime by optimizing production line task allocation [9]. Similarly, DRL has achieved precise automation of spraying and assembly in industrial robot control [10]. In semiconductor manufacturing, DRL has optimized Run-to-Run control, improving yield stability [11]. DRL has also been used to optimize machine scheduling, reducing production delays [12]. More specifically, in metallurgical scenarios such as steelmaking, studies have shown that DRL can be used for industrial process control, achieving closed-loop control of multi-objective optimization strategies through real-time interaction learning of high-dimensional complex systems [13]. At the same time, DRL-based maintenance optimization methods have demonstrated their potential for adaptive control and long-term benefit optimization in dynamic environments on steel production lines [14]. Focusing on rare metal smelting (such as tantalum, niobium, and lithium), which, like tin smelting, involves high energy consumption and nonlinear coupling characteristics, the application of DRL in these fields provides insights for optimizing tin smelting. For example, machine learning has been used in the ligand screening of rare earth element extraction, accelerating the discovery of new ligands and separation optimization by predicting distribution coefficients through deep neural networks [15], demonstrating the capability of data-driven methods to handle black-box processes. Similarly, RL has optimized multi-stage parameters in digital material design, focusing on the development of rare metal (such as lithium) composite materials [16], significantly improving design efficiency. These studies show that DRL can directly learn the optimal closed-loop strategy from high-dimensional states to continuous actions through continuous interaction with the environment, effectively overcoming the fundamental limitations of supervised learning in dynamic multi-objective optimization.

Nevertheless, the current mainstream continuous-control DRL algorithms still exhibit clear applicability limitations when confronted with the specific scenario of tin smelting. Although Deep Deterministic Policy Gradient (DDPG), Twin Delayed Deep Deterministic Policy Gradient (TD3), Soft Actor-Critic (SAC), and Proximal Policy Optimization (PPO) are widely employed in continuous action-space tasks, they demonstrate distinct shortcomings in industrial processes characterized by high-dimensional action spaces, highly sparse rewards, strongly nonlinear state transitions, and large-scale differences across multiple objectives (Table 1).

Table 1. Comparison of Mainstream Deep Reinforcement Learning Algorithms for Continuous Control.

Algorithm	Exploration Mechanism	Suitability for High-Dimensional and Multi-Objective Industrial Processes
DDPG	Additive OU/Gaussian noise	Good (good exploration capability but prone to getting trapped in local optima)
TD3	Clipped Double-Q and delayed update and target policy smoothing	Moderate (still insufficient exploration for sparse multi-objective rewards)
SAC	Maximum entropy regularization and stochastic policy	Good (but slow convergence in highly nonlinear scenarios with large reward scale differences)
PPO	Clipped probability ratio shared value function	Poor (low sample efficiency, hard to converge to high-quality policies in high-dim continuous control)
DADDPG(our)	Dual-action parallel generation Critic online selection	Excellent (best performance in high-dim, sparse-reward, multi-objective tin smelting scenarios)

The conventional DDPG suffers from limited exploration capability in high-dimensional action spaces and is highly prone to converging to local optima. Although TD3 and SAC alleviate Q-value overestimation through clipped double-Q learning and maximum-entropy regularization, respectively, they still rely on passive noise injection or inherently stochastic policies for exploration. Consequently, in “black-box” industrial processes such as tin smelting—characterized by highly sparse rewards and strongly nonlinear state transitions—they struggle to rapidly identify globally optimal parameter combinations. While PPO offers training stability, its extremely low sample efficiency renders it impractical when only limited offline factory data are available, often failing to converge to a deployable control policy within a reasonable timeframe. In contrast, the DADDPG proposed in this paper enables the online Actor to simultaneously generate a deterministic action and a high-quality random action, with the Critic performing real-time selection of the higher-valued candidate. This mechanism substantially improves effective exploration efficiency while preserving low estimation bias, making DADDPG particularly well-suited for high-dimensional continuous-action, multi-objective, and strongly nonlinear industrial processes such as tin smelting.

The tin metal carryover rate in tin smelting slag and the CO concentration in exhaust gas emissions are key process performance indicators in the tin smelting industrial process. Key parameters, such as total material feed rate, fuel coal flow, coal air velocity, oxygen flow, and spray gun backpressure, are dynamically related to performance indicators. Since the tin smelting process involves complex chemical reactions and the furnace itself is a black-box model, it is difficult to build an accurate model, and therefore, model-driven control methods are not applicable. In this case, where the mechanism is unclear and operating conditions vary, model-free data-driven control methods become an effective solution [17–19]. By collecting useful process data and then using machine learning methods to build controllers, the mapping relationship between parameters and performance indicators can be learned from the process data, thus enabling dynamic optimization and control of complex systems. This offline training-online inference (OTOP) model has been applied in several fields [20–23]. For example, in the steelmaking industry, a digital twin-based full-process monitoring and intelligent management system has significantly improved smelting efficiency and reduced material waste, demonstrating the potential of digital twin technology for real-time optimization in complex metallurgical processes [24]. Furthermore, an IoT-enabled digital twin framework has been developed to integrate multi-source production data across upstream and downstream rolling lines, enabling high-precision key-parameter prediction and collaborative online optimization, thereby mitigating quality defects such as flatness and thickness deviations in practical steel production [25]. On the other hand, in batch and complex industrial processes, offline reinforcement learning using static historical data can achieve process optimization without real-time interaction [26]. Similarly, in steelmaking processes, machine learning models trained with historical production data can be used to predict key steelmaking indicators (such as impurity content) and assist with quality control [27]. These studies further demonstrate the effectiveness and development direction of data-driven methods based on offline historical data and digital twin models for optimizing black-box complex metallurgical processes.

The remainder of this paper is organized as follows. In Section 2, the tin smelting process is introduced, and the data-driven environment modeling is formulated. In Section 3, the improved DDPG-based control method for the tin smelting process is discussed. In Section 4, the experimental design and result analysis are presented. In Section 5, the advantages of the proposed method, limitations of the current work, and future research directions are discussed. Section 6 concludes the paper.

2. Tin Smelting Process and Data-Driven Environmental Description

One of the most critical steps in the tin smelting process is the Ausmel furnace smelting, where desulfurized crude tin ore is converted into concentrated refined tin, accompanied by the generation of by-products such as slag and dust. The operation of the Ausmel furnace is regulated through lances, which supply oxygen, combustion coal, reduction coal, and crude tin ore to the furnace while also enabling rotational stirring via pressurization. Therefore, effective control of the lances is the primary focus of this study. Here, we systematically describe the tin smelting process, define the role and function of the agent, and develop a data-driven environment model to support subsequent optimization control research.

2.1. Description of the Tin Smelting Process

The tin smelting process, as shown in Figure 1, comprises raw ore proportioning, desulfurization in a fluidized bed furnace, Ausmel furnace smelting, and purification via a crystallizer. Initially, raw ore is crushed in a storage bin and proportioned. The crushed ore is then sent to a fluidized bed furnace for desulfurization, where it is converted into crude tin. Subsequently, the crude tin undergoes smelting in the Ausmel furnace to remove impurities such as arsenic, iron, and lead [28,29]. Finally, the refined tin from the smelting process is further purified through electrolytic refining in a crystallizer.

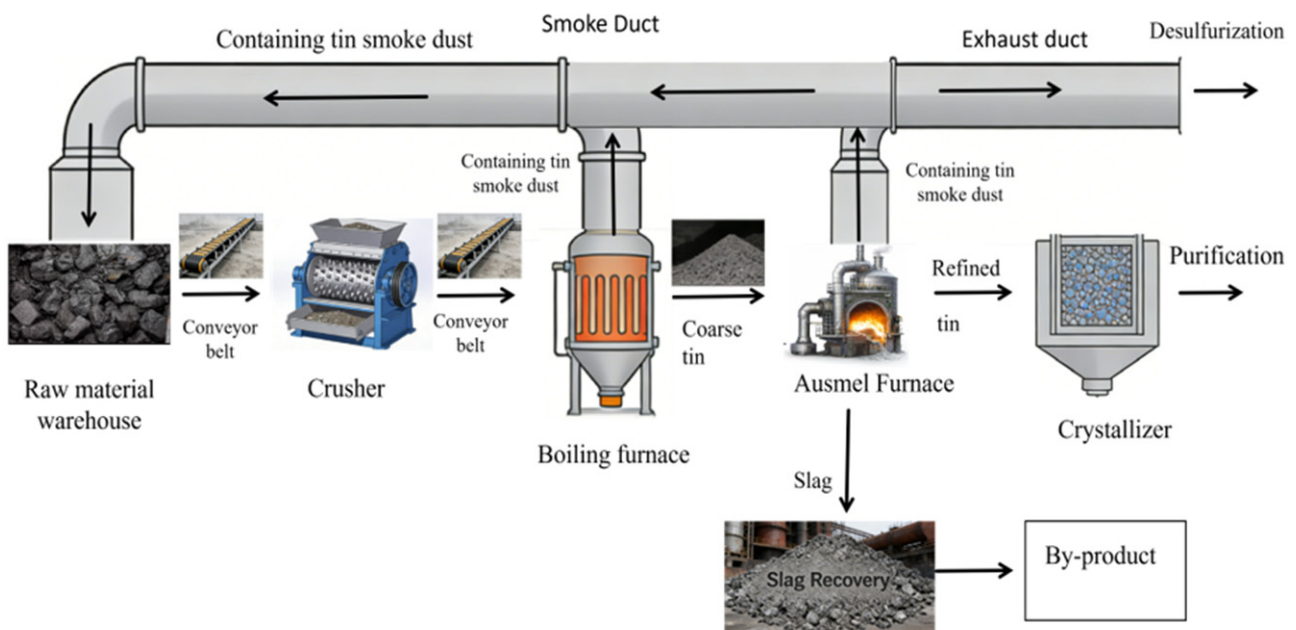


Figure 1. Tin Smelting Process Flow Chart.

Figure 2 presents the overall schematic of the Ausmel furnace. The smelting process involves feeding desulfurized crude tin raw material and flux into the furnace through the inlet. Lances inject oxygen-enriched air and combustion coal into the molten pool, with the main reactions facilitated by adjusting the oxygen concentration and lance stirring intensity. Within the Ausmel furnace, stepwise reduction reactions of tin oxide and iron oxide (from higher to lower oxidation states) occur, with carbon monoxide serving as the primary reducing agent, accompanied by the combustion of carbon. Impurity elements (e.g., As, Fe) are oxidized to form slag primarily composed of FeO-SiO_2 and CaO , as well as dust containing volatile oxides. Ultimately, refined tin, due to its density difference, settles at the furnace bottom and is collected through the outlet, achieving effective separation of tin metal from impurities.

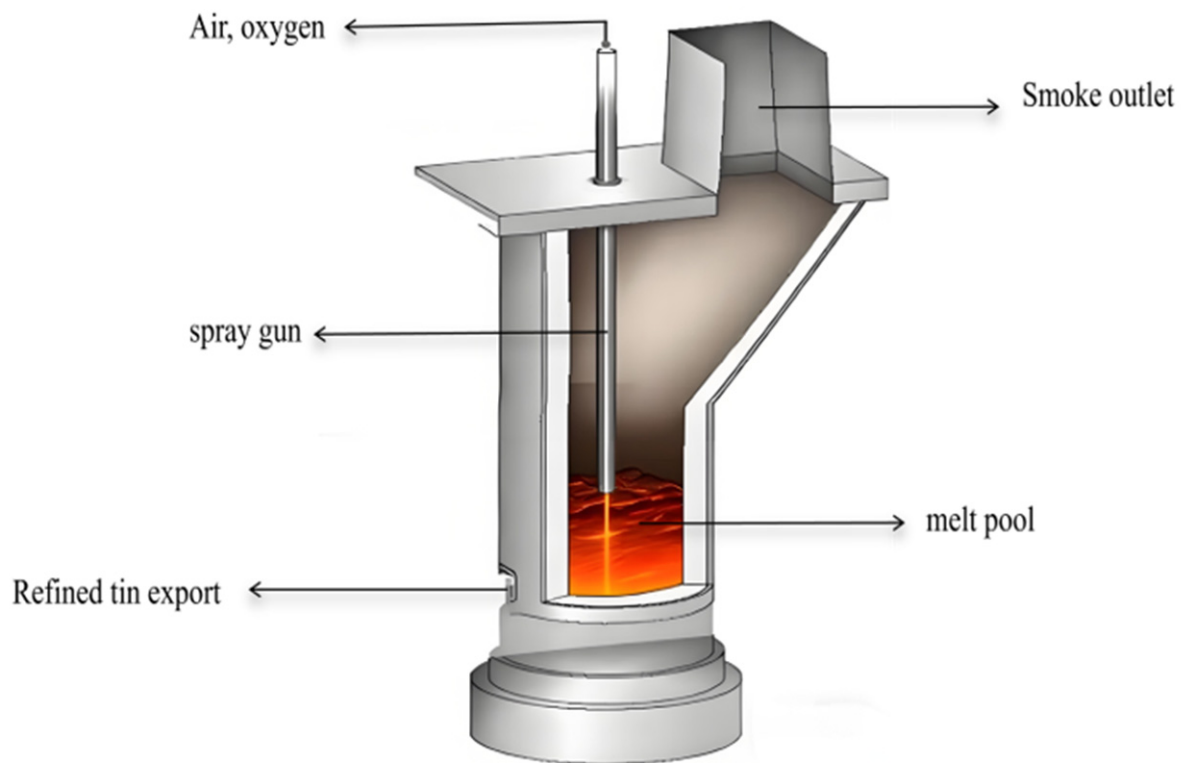


Figure 2. Structure diagram of Ausmel furnace.

In the Ausmel smelting industrial process, operational performance is quantitatively evaluated through a series of performance indicators that reflect product quality, environmental protection, and other factors. Key performance indicators in the Ausmel smelting process include the tin content in the emitted slag and the CO emission levels in the exhaust gas, both of which are influenced by lance operating parameters. In addition to these measurable lance operating parameters, various unmeasurable parameters also affect CO production and the tin content in the slag, such as oxidation and reduction reactions within the Ausmel furnace. The completeness of these reactions directly impacts the slag and CO emissions.

2.2. Definition and Role of the Agent

In the reinforcement learning framework, the agent is a decision-making entity that learns an optimal strategy through continuous interaction with the environment to achieve the goal of optimizing tin smelting parameters. In this study, the agent refers to an operational optimization decision-making unit centered on the oxygen-enriched lance, which can be understood as an intelligent controller/virtual operator for on-site lance manipulation. Its design is based on the Actor-Critic architecture, where the Actor generates actions (e.g., adjustments to spray gun parameters), and the Critic evaluates the value of those actions; this architecture enables the agent to dynamically adapt to the demands of complex industrial processes. The agent observes a multidimensional state space from the environment, generates a multidimensional action space, and dynamically updates its strategy based on a reward function (integrating quality, environmental, and safety objectives). Its operational mechanism involves a trade-off between exploration and exploitation: initially, it uses random actions to cover the action space and avoid local optima; subsequently, it leverages Critic feedback to optimize deterministic actions, achieving multi-objective dynamic adjustment. This study further innovates the agent by enabling it to simultaneously generate deterministic and random actions, selecting the optimal ones for execution, and learning from an experience pool to gradually master the optimal control strategy under

fluctuating furnace conditions, thereby consistently improving resource utilization and reducing pollution.

2.3. Data-Driven Environmental Modeling

The environment serves as the backdrop for an agent's actions and learning, representing an abstract model of the external world with which the agent can observe and interact. Defining the environment is equivalent to establishing a set of rules for the agent, encompassing the actions the agent can perform, the state of the environment, and the rewards or penalties the agent receives upon executing actions.

Several commonly used methods for constructing environments include simulation based on real physical environments [30,31], virtual environments based on games [32–34], abstract environments based on tasks [35], and data-driven environment modeling [36,37]. Given that the Ausmel furnace in this experiment is a black-box model, with unclear chemical reactions and dynamic mechanisms, and certain parameters being difficult to collect online, it is challenging to establish an accurate model of the tin smelting process. Consequently, model-driven control methods are not applicable for the practical control of parameter optimization in tin smelting. To address this issue, this paper employs a data-driven control approach, developing an RL controller that can achieve control over operational performance indicators without requiring knowledge of the system's dynamic model.

Based on a data-driven approach, this paper proposes a predictive model utilizing triplets (current state, current action, next state). This model enables the prediction of the next state based on the current furnace state and lance operating parameters.

The data description of the studied tin smelting line is presented in Table 2. The input to the RL controller consists of eight parameters representing the current state of the Ausmel furnace, denoted as s_t , including (furnace bottom center temperature, furnace bottom external temperature, furnace bottom internal temperature, furnace upper temperature, furnace pressure, CO content in exhaust gas, Sn content in emitted slag). The output is a_t , representing eight actions executable by the lance, which influence the reactions within the Ausmel furnace, including (total material flow, combustion coal flow, coal-carrying air flow, coal combustion backpressure, lance backpressure, lance position, oxygen flow, air flow). The ultimate goal of the RL controller is to reduce both the CO content in the exhaust gas and the Sn content in the emitted slag. The reward function is defined as follows Equation (1):

$$R = w_1 * R_{CO} + w_2 * R_{Sn} + w_3 * R_{\Delta CO} + w_4 * R_{\Delta Sn} - P_{CO} - P_{Sn} \quad (1)$$

The total reward R consists of two components: positive reward terms and penalty terms. R_{CO} and R_{Sn} are the reward components related to current CO and Sn concentrations, reflecting the degree of optimization in the present state. $R_{\Delta CO}$ and $R_{\Delta Sn}$ represent the reward components for the change in CO and Sn concentrations relative to the previous state, designed to promote continuous improvement. P_{CO} and P_{Sn} are the corresponding penalty terms for CO and Sn, used to constrain deviations from target values. This reward function design not only accounts for the instantaneous CO and Sn levels but also incorporates their temporal variations, enabling finer-grained reward or penalty assignment at each step. During implementation, Gaussian functions combined with exponential decay are employed to achieve a continuous penalty mechanism. Compared with abrupt penalties, this approach keeps operations within safe bounds, provides smooth gradients, prevents sudden reward drops, and delivers more refined feedback to the agent. Finally, the Actor network model yielding the highest cumulative reward is saved and deployed

in the plant. Once deployed, the RL controller can optimize lance operating parameters simply by invoking this frozen model.

Table 2. Data Description of Status and Actions.

	Parameters	Values
Actions	Total Feed PV	0 to 86.5 t/h
	Fuel Coal Flow PV	0 to 9960.0 kg/h
	Coal Carrier Air Flow	0 to 634.0 Nm ³ /h
	Fuel Coal Back Press	0 to 106.3 kPa
	Shroud Air Back Press	0 to 22.5 kPa
	Lance Hoist Position	410.0 to 15,574.0 mm
	Lance Oxygen Flow	0 to 7031.0 Nm ³ /h
	Lance Air Flow	0 to 17,832.0 Nm ³ /h
States	Calc O ₂ Enrich	21.0 to 37.3%
	Furnace Base Mid Temp	54.7 to 65.8 °C
	Furnace Base Outer Temp	46.5 to 57.5 °C
	Furnace Base Inner Temp	66.9 to 82.1 °C
	Furnace Heatup Temperature	31.3 to 67.9 °C
	Furnace Pressure	−94.5 to 21.9 Pa
	Offgas CO Analyzer	0 to 48,334 ppm
	Tin content in slag	1.96 to 10.3%

3. Research Methods

In this section, we propose an RL algorithm called Dual Action Generation, which builds upon the Deep Deterministic Policy Gradient (DDPG) algorithm. Through the online Actor network, two actions are generated: one based on a deterministic policy and one random action. The Critic network compares their values, introducing an exploration mechanism to the DDPG algorithm, thereby enhancing training performance or policy diversity in certain scenarios.

3.1. DDPG

This paper adopts DDPG as the foundational algorithm, given that the data collected from the tin factory consists of continuous data series, and this algorithm excels in RL tasks with continuous action spaces. DDPG is an improvement over the Deterministic Policy Gradient (DPG) algorithm [38], integrating the expressive power of deep learning by incorporating deep neural networks to enhance performance. It employs an Actor-Critic framework, comprising two Actor networks and two Critic networks: a main Actor network generates deterministic actions, while a target Actor network facilitates smooth updates to improve stability; a main Critic network evaluates the value function of actions, and a target Critic network computes target Q-values to reduce variance during training. This dual-network design uses a soft update mechanism to allow the target networks to slowly track the online networks, mitigating drastic fluctuations during training and enhancing algorithm stability. The objective of the online Actor network is to optimize the policy to maximize the Q-value estimated by the Critic network. The gradient update formula is as follows Equation (2):

$$\nabla_{\theta\mu} J \approx E_{s \sim p^\beta} \left[\nabla_{\theta\mu} \mu(s; \theta^\mu) * \nabla_a Q(s, a; \theta^Q) \Big|_{a=\mu(s; \theta^\mu)} \right] \quad (2)$$

Here, J represents the expected cumulative reward, E denotes taking the average over a batch of states sampled from the replay buffer to compute the policy gradient. with the state distribution p^β generated by the behavioral policy β . The term $\nabla_a Q(s, a; \theta^Q) \Big|_{a=\mu(s; \theta^\mu)}$ denotes the gradient of the Q-value with respect to action a , computed by the Critic network,

evaluating the quality of the current action. The term $\nabla_{\theta^{\mu}} \mu(s; \theta^{\mu})$ represents the gradient of the Actor network parameters, used to adjust the policy μ . Through the chain rule, the Critic's feedback is propagated to the Actor, optimizing the policy to produce better actions.

The online Critic network updates by minimizing the temporal difference error, and the loss function during training is Equation (3):

$$L = E_{s,a,r,s' \sim D} [(Q(s,a;\theta^Q) - y)^2] \quad (3)$$

Here, the target value y is Equation (4):

$$y = r + \gamma Q'(s', \mu'(s', \theta^{\mu'}); \theta^{Q'}) \quad (4)$$

Here, L represents the loss function of the Critic network, which measures the discrepancy between the predicted Q -value and the target Q -value. E denotes taking the average over a mini-batch of transitions randomly sampled from the replay buffer to compute the mean-squared-error (MSE) loss. D denotes the experience replay buffer, storing samples of state s , action a , reward r , and next state s' . r is the immediate reward, γ is the discount factor, and $Q'(s', \mu'(s', \theta^{\mu'}); \theta^{Q'})$ represents the Q -value estimation by the target Critic network for the next state s' and the target action $\mu'(s')$. By minimizing L , the online Critic network's Q -values are driven to approximate the true value.

3.2. Dual Action Selective DDPG

Figure 3 illustrates the overall design framework of this study. The system first collects data from the Ausmel furnace via sensors and stores the data in a database. Based on this data, the system constructs an environment model, which is then used to train and optimize an improved model, ultimately providing operational recommendations to factory operators. The core of the model improvement lies in generating two types of actions through the online Actor network: one based on a deterministic policy and the other being a random action. Subsequently, the Critic network evaluates the value of these two actions, selects the action with higher value, and stores high-reward random actions in the experience replay buffer.

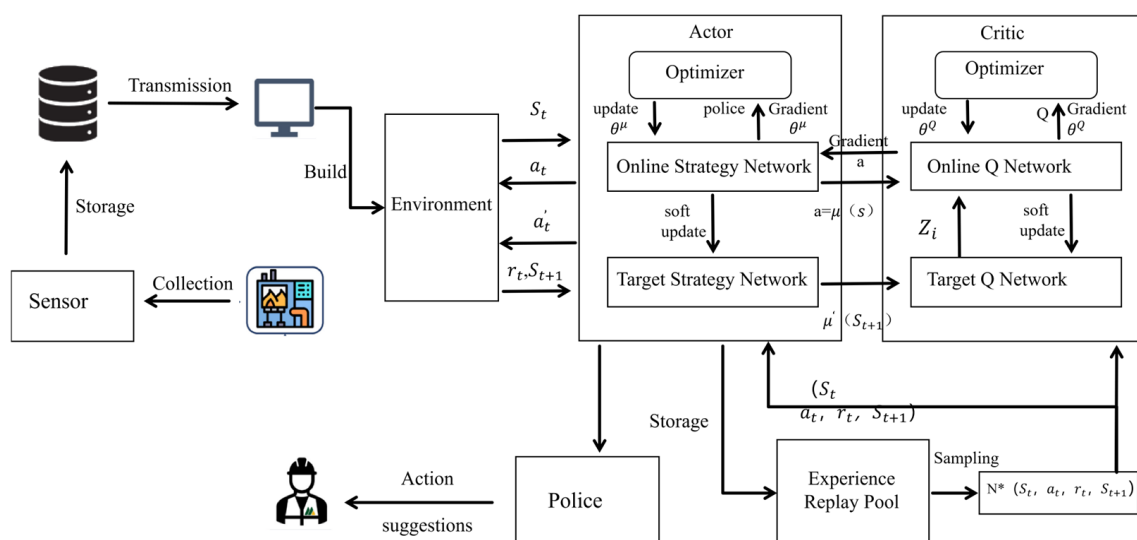


Figure 3. Data collection, environmental modeling, improved model training.

Compared to the traditional DDPG algorithm, DADDPG consistently enhances exploration efficiency. In the traditional DDPG algorithm, the deterministic policy is defined as Equation (5):

$$\pi(s) = \mu_{\theta}(s) + \epsilon \quad (5)$$

where ϵ represents noise, resulting in a low overall policy entropy. Exploration primarily relies on action noise, which has limited noise variance, leading to low exploration efficiency and a tendency to get trapped in local optima, especially in high-dimensional continuous action spaces—such as the spray gun parameters in the tin smelting process studied here. In contrast, the effective policy of DADDPG is Equation (6):

$$\pi_{eff}(s) = \operatorname{argmax}_{a \in \{a_{det}, a_{rand}\}} Q_i(s, a) \quad (6)$$

By dynamically integrating deterministic and random actions, this approach consistently enhances the diversity of exploration in the action space, thereby increasing the probability of discovering high-value actions. The introduction of random actions resembles a “lottery” mechanism, which, in certain cases, can identify actions superior to those from the deterministic policy, particularly in environments with local optima or sparse rewards. $\pi_{eff}(s)$ compares Q_{det} and Q_{rand} to select the action, ensuring the executed action is closer to the optimal a^* . When $Q_{rand} > Q_{det}$, the random action introduces external perturbations to break the local optima trap; otherwise, the deterministic action leverages current experience.

Mathematically, the exploration contribution of random actions is quantified through variance, with a larger standard deviation σ expanding the sampling range and reducing the dependence of gradient updates on local optima. From the perspective of the Bellman optimality equation, Equation (7) can be written as follows:

$$Q^*(s, a) = r(s, a) + \gamma E_{s'} [V^*(s')] \quad (7)$$

The Actor update gradient in DADDPG is Equation (8):

$$\nabla_{\theta} J(\theta) = E[\nabla_a Q(s, a|\theta) \nabla_{\theta} \mu_{\theta}(s)] \quad (8)$$

Through the selection mechanism, low-value actions are filtered out, preventing gradient explosion dominated by noise.

The pseudocode for the proposed method (DADDPG) is presented in Algorithm 1. Table 3 provides definitions for the key notations first appearing in the pseudocode to aid understanding of the algorithm. Initially, the Actor and Critic networks, along with the experience replay buffer, are initialized. The system starts with an arbitrary initial state, and the online Actor network interacts with the environment, outputting an action based on the current state. The environment receives the action from the online Actor network, updates the model and reward function based on the state, and outputs a new state and the reward associated with the action. The Critic network takes the current state and action as input and outputs a Q-value, representing the expected cumulative reward for executing the action in the current state, which indicates the “quality” of the action. In the original DDPG algorithm, each online Actor network generates only a single action, resulting in limited exploration, susceptibility to local optima, sensitivity to noise, and constrained training stability. In contrast, DADDPG’s online Actor network produces not only a deterministic action but also a randomly perturbed action. The Critic network then evaluates the Q-values of these two actions, storing only the action with the higher Q-value, along with its corresponding state-reward-new state tuple, in the experience replay buffer for subsequent network updates. Additionally, to prevent the generation of unsafe actions,

constraints are applied to randomly generated actions, restricting them to a safe range. DADDPG mitigates the performance instability and tendency to converge to local optima in the original DDPG algorithm, which stem from insufficient exploration.

Algorithm 1. DDPG with Dual Action Candidates and Value-Based Selection.

```

Initialize  $\mu_\theta, \mu'_{\theta'}, Q_i, Q'_{i'}$ 
Initialize replay buffer B
For episode = 1 to M do
   $s \leftarrow \text{env.reset}()$ 
  For  $t = 1$  to T do
     $a_{det} \leftarrow \mu_\theta(s)$ 
     $a_{rand} \sim \text{Uniform}(a_{min}, a_{max})$ 
     $a_{rand} \leftarrow \text{clip}(a_{rand}, a_{min}, a_{max})$ 
     $Q_{det} \leftarrow Q_i(s, a_{det})$ 
     $Q_{rand} \leftarrow Q_i(s, a_{rand})$ 
     $a \leftarrow a_{rand}$  if  $Q_{rand} > Q_{det}$  else  $a_{det}$ 
     $s', r, done \leftarrow \text{env.step}(a)$ 
    Store  $(s, a, r, s')$  in B
    If  $|B| \geq \text{batch}_{size}$  then
      Sample  $\{(s_i, a_i, r_i, s'_i)\}$  from B
       $a'_i \leftarrow \mu'_{\theta'}(s'_i)$ 
       $y_i \leftarrow r_i + \gamma Q'_{i'}(s'_i, a'_i)$ 
      Update  $Q_i$  via  $\min 1/N \sum (y_i - Q_i(s_i, a_i))^2$ 
      Update  $\mu_\theta$  via  $\max 1/N \sum (Q_i(s_i, \mu_\theta(s_i)))$ 
       $\mu'_{\theta'} \leftarrow \tau \times \mu_\theta + (1 - \tau)\mu'_{\theta'}$ 
       $Q'_{i'} \leftarrow \tau \times Q_i + (1 - \tau)Q'_{i'}$ 
       $s \leftarrow s'$ 
    If done then break
  End for
End for

```

Table 3. Symbol Definition Table.

Symbols	Meaning
μ_θ	Main Policy Network
$\mu'_{\theta'}$	Target Policy Network, used for stabilizing training
Q_i	The i -th action value function network
$Q'_{i'}$	The i -th target action value function network
a_{det}	Deterministic action
a_{rand}	Random action
Q_{det}	The value of deterministic action
Q_{rand}	The value of random actions
s, s'	Current state and next state
y_i	Target value for updating the Q_i
τ	Soft update factor

3.3. Relation to Prior Exploration Mechanisms

Exploration is a key challenge in industrial continuous-control tasks such as tin smelting, where the action space is high-dimensional, the reward is sparse, and the state transition is highly nonlinear. Most mainstream DRL algorithms rely on passive stochasticity injection to encourage exploration, which may be inefficient and unstable in black-box

metallurgical processes. This subsection clarifies how the proposed DADDPG differs fundamentally from prior exploration mechanisms.

- (1) Action-noise exploration in DDPG. In DDPG, exploration is typically achieved by perturbing the deterministic policy output with additive noise (e.g., OU or Gaussian noise). Such exploration is unstructured and local, and its effectiveness depends heavily on the noise scale. In high-dimensional action spaces, small noise often fails to escape local optima, while large noise may produce unsafe actions and destabilize training.
- (2) Stochastic policy and entropy regularization in SAC. SAC encourages exploration by learning a stochastic policy and maximizing an entropy-augmented objective. Although this increases exploration diversity, it also introduces higher gradient variance and may slow down convergence in industrial processes with large-scale differences between multi-objective rewards, where stable and consistent long-term optimization is required.
- (3) Target policy smoothing in TD3. TD3 alleviates Q-value overestimation by clipped double-Q learning and target policy smoothing. However, the injected noise mainly serves as a regularizer for target updates rather than a mechanism for efficient exploration in the behavior policy. In practice, TD3 still relies on action noise similar to DDPG for exploration and thus remains limited under sparse-reward conditions.
- (4) PPO-style on-policy exploration. PPO relies on stochastic policies and on-policy rollouts, which tend to be sample-inefficient when only limited offline factory data are available, making it difficult to obtain high-quality exploration within a reasonable training budget.
- (5) DADDPG: Critic-guided dual-action discriminative exploration (ours). Different from the above methods, DADDPG introduces a value-aware exploration-and-selection mechanism. At each time step, the online Actor generates two candidate actions in parallel: a deterministic action and an exploratory random action sampled from a broader distribution within safety bounds. The Critic then evaluates both candidates and executes the action with a higher estimated value.

Therefore, exploration in DADDPG is not performed by blindly injecting noise into the policy output but by generating diverse candidates and selecting the most promising one using Critic evaluation. This discriminative selection provides two major advantages: (1) it enables occasional large exploration jumps beyond local perturbations, improving the probability of discovering globally beneficial parameter combinations under sparse rewards; (2) it filters out low-value random actions online, preventing training instability caused by excessive noise and improving convergence robustness.

4. Guidelines Simulation Experiment of Tin Smelting Process

In this section, the experiment is divided into three steps. First, a data-driven state update model, also referred to as the RL environment in this paper, is constructed. This study employs a data-driven, model-free environment approach to simulate the operational characteristics of the Ausmel furnace. Second, the improved DDPG algorithm interacts with the RL environment to train the agent to learn furnace operation expertise. Specifically, the model-free data-driven method outlined in Algorithm 1 is applied to the Ausmel furnace smelting process described in Section 4.2 (RL training), where eight process parameters of the furnace's spray gun directly influence the tin content in the slag and the CO content in the exhaust gas. After training, the RL agent learns an optimal action generation strategy, which enables it to provide the best spray gun operation recommendations under different environmental states. Finally, the RL environment is used to evaluate this strategy.

4.1. RL Environment

The Ausmel furnace currently monitors eight state variables that reflect the furnace's internal conditions. States 1 to 6 are six basic states, namely bottom center temperature, oxygen content percentage, furnace temperature variation, bottom external temperature, bottom internal temperature, and furnace pressure. States 7 and 8 are two indicator states, specifically CO content in exhaust gas and Sn content in slag. To achieve data-driven state prediction, we selected different models tailored to the characteristics of each state variable to predict the next time step's state, categorizing them into three types. First, for the six basic features, which exhibit small variations and high predictability, a regression model based on a multilayer perceptron (MLP) was employed. Figure 4 illustrates the prediction performance of this model, demonstrating that it effectively captures the dynamic changes in the state variables, with predicted values closely aligning with actual values and remaining within reasonable ranges. For the CO state, a Gated Recurrent Unit (GRU) model was adopted, as GRUs are well-suited for capturing long-term and short-term patterns in time series, making them ideal for handling the exhaust gas CO content, which is heavily influenced by historical data. Figure 5 presents a comparison between the predicted and actual CO concentrations using the GRU model, with the horizontal axis representing the number of samples and the vertical axis indicating CO concentration (in ppm). Finally, for the Sn content in slag, which is a low-frequency measurement state (measured once per furnace cycle), an XGBoost model was utilized. XGBoost is particularly effective for handling discrete, sparse, or discontinuous data, aligning well with the measurement characteristics of Sn content. Figure 6 illustrates the performance of the XGBoost model in predicting the Sn content in the slag discharged from the Ausmel furnace. However, due to the limited number of Sn measurement samples, the model's prediction accuracy is somewhat constrained, resulting in certain deviations between predicted and actual values. Nevertheless, the model successfully captures the overall trend of Sn content variations.

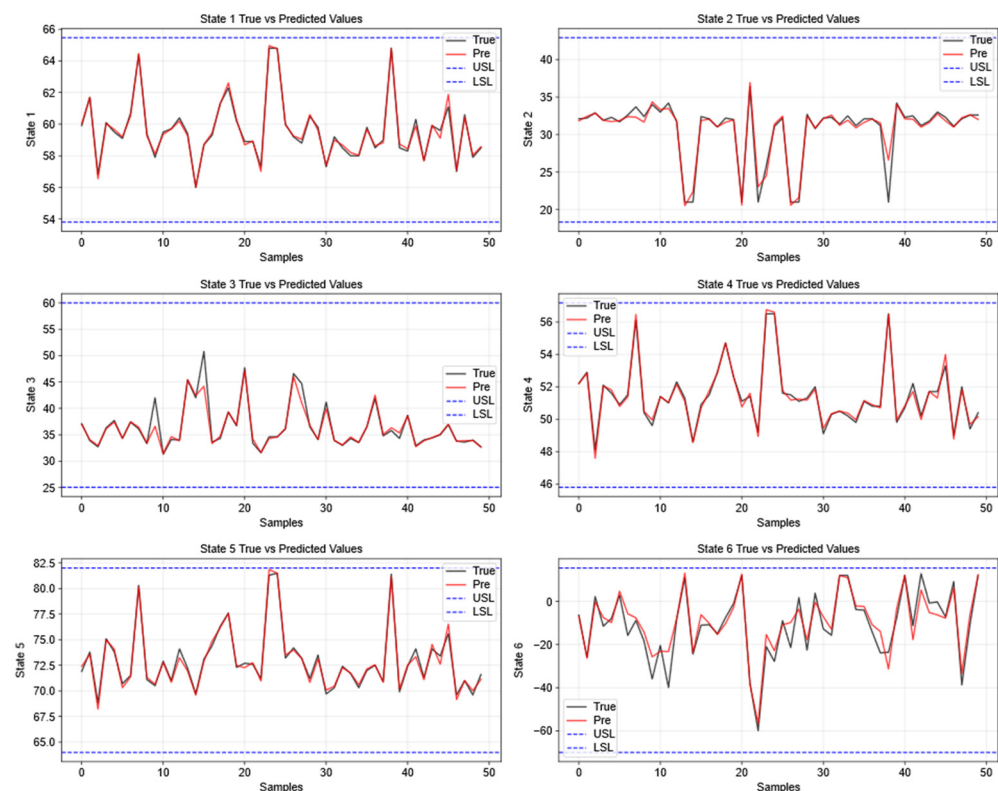


Figure 4. Comparison of basic state prediction based on MLP model and actual values.

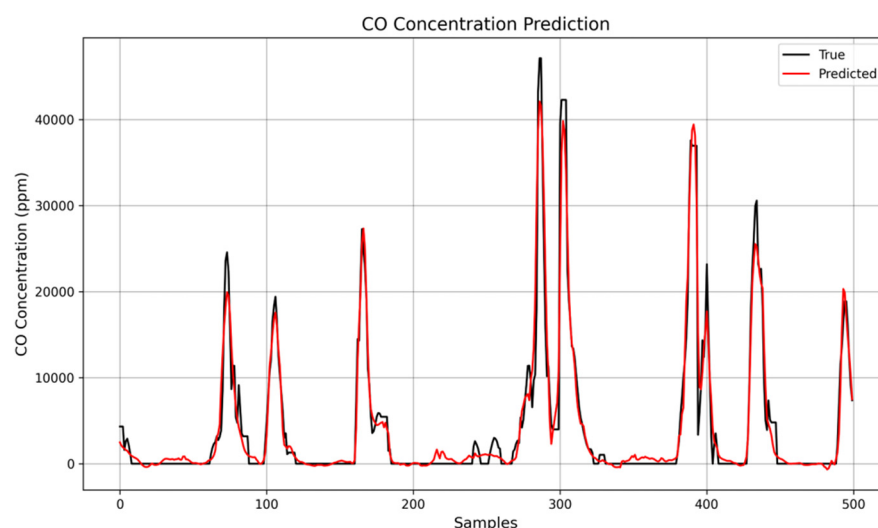


Figure 5. Comparison of CO concentration prediction based on GRU model and real values.

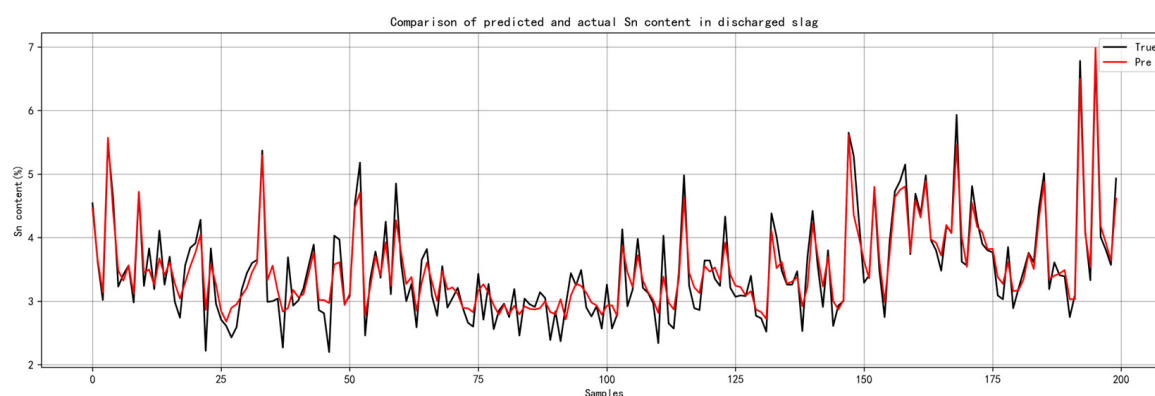


Figure 6. Comparison of Sn content prediction based on XGBoost model and actual values.

As shown in Table 4, the model performs excellently across most state dimensions. The R^2 value for State 1 (where R^2 is the coefficient of determination, measuring the goodness of fit of a regression model, ranging from 0 to 1, with values closer to 1 indicating a better fit) reaches a maximum of 0.9906, indicating near-perfect prediction accuracy, followed by state 4 (0.9899), state 7 (0.9881), state 5 (0.9876), state 3 (0.9690), and state 2 (0.9353). However, the R^2 values for state 6 and state 8 are lower, at 0.8690 and 0.8780, respectively, suggesting weaker predictive performance in these dimensions. Analysis indicates that the lower performance for state 6 may be attributed to its high variance and predominantly negative values, while the reduced prediction accuracy for state 8 is likely due to insufficient sample size.

Table 4. R^2 Scores Predicted For Each State Dimension.

State	R^2	Model
State 1	0.9906	MLP
State 2	0.9353	MLP
State 3	0.9690	MLP
State 4	0.9899	MLP
State 5	0.9876	MLP
State 6	0.8690	MLP
State 7	0.9881	GRU
State 8	0.8780	XGBoost

4.2. RL Training

This study is conducted in a data-driven RL environment and compares five representative algorithms, including DDPG, SAC [39], TD3 [40], PPO [41], and the proposed DADDPG. These methods are among the most widely adopted approaches in reinforcement learning and have been extensively applied to optimization control problems in continuous action spaces and complex dynamic systems, making them suitable baselines for multi-objective optimization in the tin smelting process. To ensure training efficiency, fairness, and reproducibility, all algorithms were trained and evaluated in the same data-driven RL environment, using identical state/action normalization, action bounds, and shared hyperparameters as reported in Table 5. Specifically, common parameters (e.g., learning rate $\eta = 0.0001$, discount factor $\gamma = 0.99$, and replay buffer size 10^4) were selected as optimal configurations through experimental comparisons over different values ($\eta \in [0.0001, 0.01]$, $\gamma \in [0.9, 0.99]$), while model-specific hyperparameters were directly adopted from the default settings recommended in the original papers. This strategy was adopted because (i) it ensures consistency with literature baselines and supports reproducibility, (ii) the high-dimensional continuous-control nature of tin smelting is comparable to typical benchmark tasks where the default settings have demonstrated robustness, and (iii) sensitivity analysis indicates that performance differences caused by default model-specific hyperparameters are negligible, whereas learning rate and discount factor are the most influential factors. Moreover, each algorithm was independently repeated five times, and in each run the environment initialization procedure, network architecture, and parameter initialization scheme were kept identical across algorithms, with randomness controlled by fixed seeds to guarantee consistent initial conditions. The following assumptions were made for the data-driven simulated tin smelting RL environment.

Table 5. RL Algorithm Hyperparameter Settings.

Shared		PPO		SAC	
Hyperparameters	Value	Hyperparameters	Value	Hyperparameters	Value
gamma	0.99	CLIP_PARAM	0.2	Polyak	0.95
lr_actor	0.0001	GAE_LAMBDA	0.95	alpha	0.2
lr_critic	0.0001	ENTROPY_COEF	0.01	target_entropy	−1
batch_size	128	MAX_GRAD_NORM	0.5		
Total steps	1,000,000				
buffer	10,000	TD3		DADDPG/DDPG	
learn_every	20	Hyperparameters	Value	Hyperparameters	Value
Optimizer	Adam	policy_delay	2	ou_noise_theta	0.15
		noise	0.2	ou_noise_mu	0.0
		noise_clip	0.5	ou_noise_sigma	0.2
				DADDPG actor	2

- (1) The materials in the furnace are perfectly mixed.
- (2) The density of the crude tin ore is approximately uniform.
- (3) The reaction delay in the furnace is minimal.

Figure 7 presents the average reward performance of the five algorithms in the data-driven RL environment. These algorithms induce changes in the process state by generating different actions, and reward values are computed based on the variations in CO concentration and tin content in the discharged slag. Higher reward values indicate closer alignment with the optimization objectives, thereby enabling the agent to progressively acquire lance control expertise. The reward function is designed to be stringent, with rewards increasing as CO and tin concentrations approach zero.

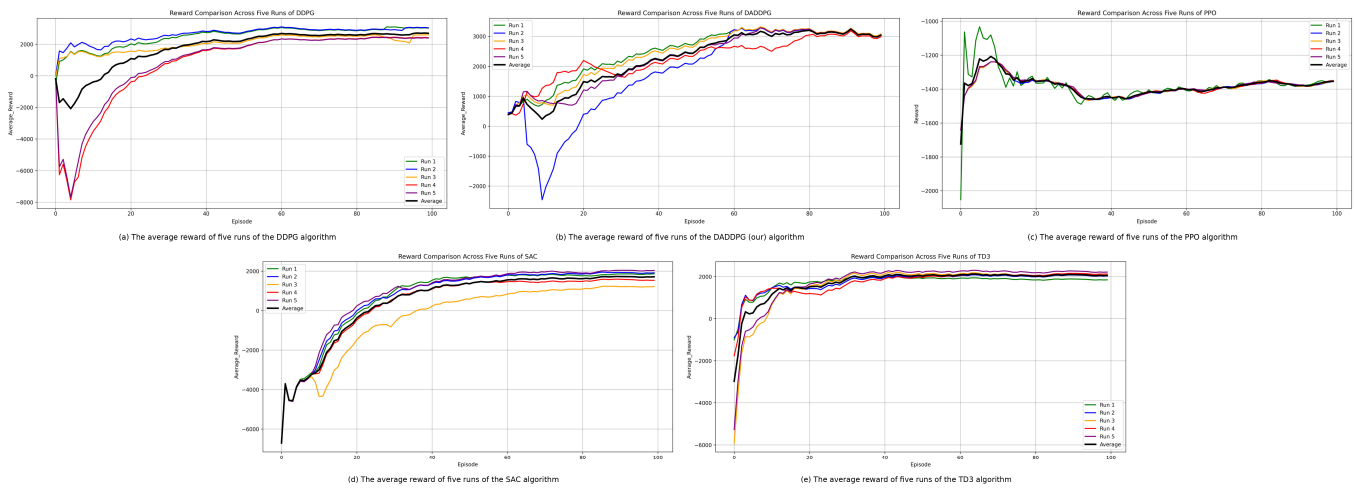


Figure 7. Average reward performance of the five algorithms.

Specifically, Figure 7a shows the reward curves for five runs of DDPG, where the average reward gradually improves from approximately -8000 and stabilizes around 2200 , demonstrating that DDPG ultimately acquires a certain level of control capability. Figure 7b illustrates the reward curves for five runs of DADDPG, with the average reward rising from -1000 to approximately 3000 , indicating superior control expertise compared to the original DDPG. Figure 7c displays the reward curves for five runs of PPO, where the average reward stabilizes around -1000 after starting from -1800 , reflecting relatively poor learning performance. Figure 7d presents the reward curves for five runs of SAC, with the average reward increasing from -6000 to approximately 2000 , showing strong learning ability on the tin smelting task yet still falling short of DDPG. Figure 7e depicts the reward curves for five runs of TD3, where the average reward follows a trajectory similar to SAC, rising from -6000 to around 2000 , indicating stable convergence but no significant advantage. Overall, the proposed DADDPG algorithm achieves the highest final reward performance.

Figure 8 presents a comparison of the average reward values for five RL algorithms—DDPG, SAC, TD3, PPO, and DADDPG—each repeated five times. The figure illustrates that all methods converge within 100 episodes. Specifically, PPO (purple) starts with an initial reward of approximately -1400 and shows no significant improvement throughout training, indicating poor adaptability. SAC (yellow) begins with the lowest performance (approximately -4000) but exhibits rapid growth within the first 0–20 episodes, stabilizing at 1800 – 2000 after 40 episodes, demonstrating fast learning and stability. TD3 (red) starts at -3000 , rises quickly to 1800 within 0–20 episodes, and ultimately stabilizes around 2000 , reflecting a strong capability to escape penalties. DDPG (blue) gradually improves from a negative value, stabilizing at 2000 – 2300 after 60 episodes, showcasing robust long-term performance. Overall, DADDPG (green), with its improved distributional adaptability, starts from a negative value and exhibits slightly slower but continuous reward growth, stabilizing at 2500 – 3000 after 60 episodes with minimal fluctuations. This highlights its superior long-term performance and robustness in the tin smelting task, surpassing DDPG and consistently outperforming other algorithms. After training, DADDPG reduces the tin content in slag to between 3.5% and 4% and the CO content in exhaust gas to between 2000 and 2700 ppm. Compared to pre-optimization levels, where slag tin content ranged from 1.96% to 10.3% and CO emissions ranged from 0 to $48,334$ ppm, DADDPG demonstrates superior performance.

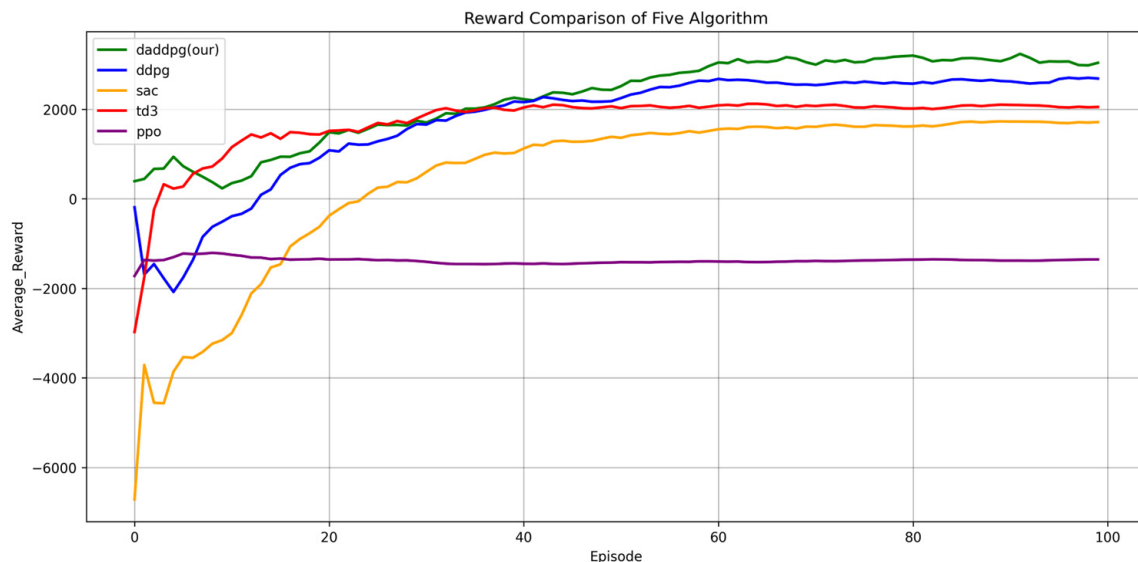


Figure 8. Comparison of Average Rewards of Five Algorithms.

4.3. Ablation Study

To systematically evaluate the contributions of the core components in the proposed DADDPG algorithm, as shown in Figure 9, we conducted an ablation study to quantitatively validate its innovation. First, the baseline model DDPG was selected as the reference; second, Variant 1 was designed by removing deterministic actions to independently assess the contribution of random actions; third, Variant 2 retained both deterministic and random actions but disabled the Critic selection mechanism to quantify the value of the selection strategy; finally, the complete DADDPG model integrated the dual-action selection mechanism, serving as a comparison to verify its overall superiority.

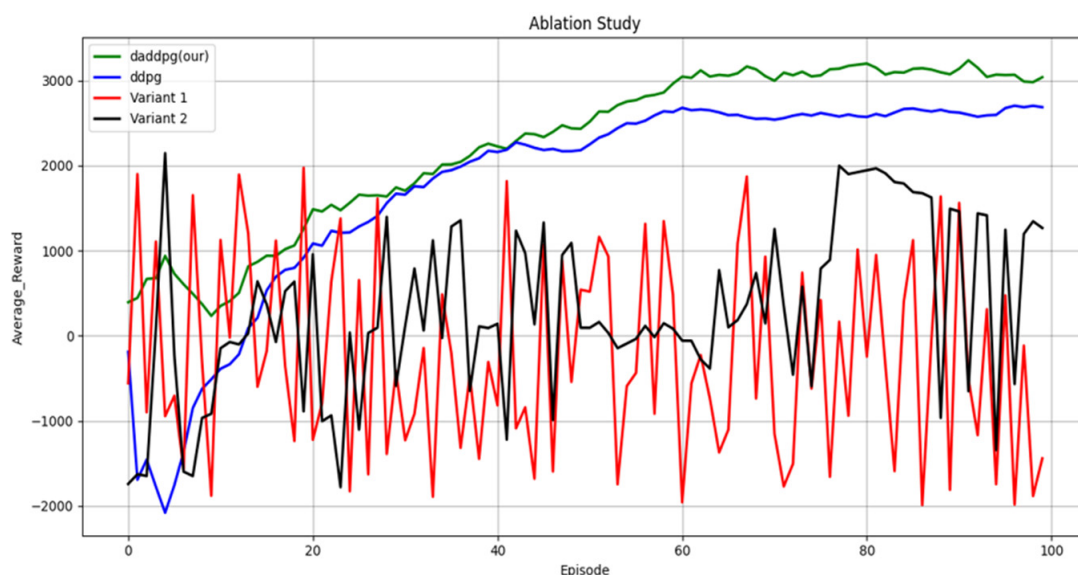


Figure 9. Cumulative Rewards for DADDPG Variants in Ablation Study.

As shown in Figure 9, we systematically evaluated the contributions of the core components of the DADDPG algorithm through an ablation study. The baseline DDPG model exhibited reward values rising from -189.66 to 2703.97 , indicating limited exploration efficiency. Variant 1 (only random actions) showed significant reward fluctuations, ranging from -1990.15 to 1899.96 , with poor convergence stability, as the majority of the values were low and only a small portion of the rewards performed well. Variant 2 (no selection mech-

anism) had reward values increasing from -1697.36 to 1999.19 , with an overall upward trend but noticeable fluctuations. The complete DADDPG model, however, demonstrated the best performance, with rewards steadily improving and consistently outperforming the other variants.

4.4. Theoretical Comparison of DADDPG with Baseline Algorithms in Terms of Estimation Bias, Variance, and Convergence

First, regarding Q-value estimation bias, the conventional DDPG suffers from severe overestimation bias because it uses a single Critic and directly maximizes the target Q-value. TD3 significantly reduces this bias through Clipped Double-Q, while SAC further suppresses overestimation via maximum-entropy regularization. The proposed DADDPG inherits the twin-network architecture of DDPG; however, by storing only the higher-valued action (as evaluated by the Critic) into the replay buffer, it effectively implements “selective bias suppression” at the experience replay level. This mechanism largely prevents low-quality random actions from introducing overestimation, yielding a bias level comparable to or even lower than that of TD3.

Second, with respect to Q-value estimation variance, exploration in vanilla DDPG relies entirely on additive noise, making variance difficult to control. Although SAC increases policy variance through its maximum-entropy framework, this also elevates Q-value variance. While the random actions in DADDPG introduce additional variance, the immediate Critic-based screening ensures that only high-Q-value samples are retained in the replay buffer, substantially reducing the propagation of harmful variance. Empirical evidence from the ablation study (Figure 9) shows that the training curve of DADDPG exhibits markedly lower fluctuation than the “random-action-only” variant.

Finally, concerning convergence and exploration efficiency in high-dimensional continuous action spaces, the exploration efficiency of both DDPG and TD3 decays exponentially with increasing dimensionality. Although SAC theoretically guarantees optimality through entropy regularization, tuning the entropy coefficient becomes extremely challenging when reward scales differ significantly across objectives, resulting in slow practical convergence. In contrast, the “parallel deterministic + random + online Critic selection” mechanism of DADDPG enables rapid discovery of potentially high-return regions in the early stage (analogous to mutation-selection in evolutionary strategies) while maintaining stable exploitation in later stages. Theoretically, this confers higher sample efficiency and stronger capability to escape local optima. Experiments show that DADDPG achieves approximately 10–15% higher final cumulative reward than DDPG and 25–30% higher than SAC/TD3 (Figure 7), empirically validating its superior convergence performance.

4.5. Comparison of Core Process Indicators

Although Figure 7 provides an effective evaluation of the learning processes and relative performance of five algorithms, relying solely on the variation in reward values cannot fully reflect the practical optimization effectiveness of reinforcement learning algorithms in industrial processes. In complex industrial scenarios such as tin smelting, the superiority of an algorithm is not only reflected in the cumulative rewards of the agent but, more critically, in its ability to improve key production metrics and enhance process stability. We have plotted Figure 10, which compares the real factory operational data before optimization with the simulated data distributions of CO content in exhaust gas and Sn content in slag from the five algorithms, intuitively validating the effectiveness of the algorithms.

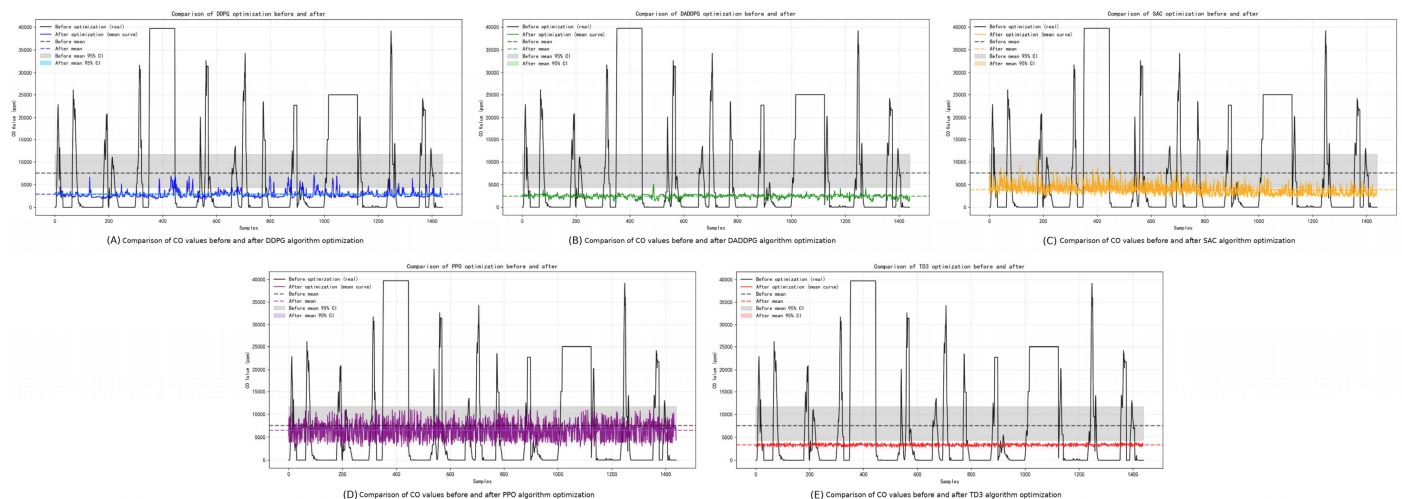


Figure 10. Comparison of CO concentration in exhaust gas before and after optimization using the Five Algorithms.

Figure 10 presents the CO concentration variations over 1440 samples, where the x -axis represents 24 h of real operational records and the y -axis denotes the exhaust-gas CO concentration. The black curve shows the CO concentration during real furnace operation before optimization, featuring strong fluctuations with a peak close to 40,000 ppm and a mean value of 7583.60 ppm (95% CI: 4346.95, 11,733.04). After optimization, Figure 10A–E reports the CO profiles optimized by five RL-based controllers. Figure 10A (blue, DDPG) achieves an average CO of 2896.48 ppm (95% CI: 2765.09, 3037.49), corresponding to a mean reduction of 4687.11 ppm (95% CI: 1435.51, 8780.18, block-bootstrap test, $p < 0.001$). Figure 10B (green, DADDPG) shows the best performance, reaching the lowest average CO of 2424.04 ppm (95% CI: 2373.99, 2477.10), with the largest reduction of 5159.40 ppm (95% CI: 1912.07, 9313.17, $p < 0.001$). Figure 10C (orange, SAC) yields an average CO of 3861.85 ppm (95% CI: 3651.72, 4074.77), with a reduction of 3721.75 ppm (95% CI: 464.82, 7820.64, $p = 0.021$). Figure 10D (purple, PPO) shows limited improvement, with an average CO of 6537.55 ppm (95% CI: 6459.99, 6611.54); the mean reduction is 1046.04 ppm and is not statistically significant (95% CI: -2163.16 , 5193.55, $p = 0.607$). Figure 10E (red, TD3) reduces CO to 3355.05 ppm (95% CI: 3343.99, 3366.61), achieving a reduction of 4228.54 ppm (95% CI: 984.87, 8379.23, $p = 0.004$). Overall, all algorithms except PPO consistently reduce CO emissions compared with the pre-optimization baseline, and the proposed DADDPG demonstrates the most effective and robust CO-emission reduction among the compared methods.

Figure 11 shows the comparison of slag Sn content before and after optimization using five RL algorithms over 50 furnace cycles (x -axis). Since Sn is measured only once per cycle in real production, one Sn value was randomly sampled from each optimized simulated cycle to mimic the practical sampling uncertainty. The black curve represents the pre-optimization industrial operation, with a mean Sn content of 4.4882% (95% CI: 4.2655, 4.7459) and a peak of approximately 7%. After optimization, Figure 11A (blue) shows that DDPG reduces the mean Sn content to 3.7773% (95% CI: 3.6702, 3.8936), corresponding to a mean decrease of 0.7109% (95% CI: 0.4499, 0.9993, block-bootstrap test, $p < 0.001$). Figure 11B (green) demonstrates that the proposed DADDPG achieves the best performance, reaching the lowest mean Sn content of 3.5886% (95% CI: 3.4920, 3.6886) and the largest reduction of 0.8564% (95% CI: 0.6851, 1.0704, $p < 0.001$). Figure 11C (orange) indicates that SAC decreases the mean Sn content to 4.0600% (95% CI: 3.8716, 4.2245), with a reduction of 0.3850% (95% CI: 0.0888, 0.7176, $p = 0.006$). Figure 11D (purple) shows that PPO provides only marginal improvement, with a mean Sn content of 4.4342% (95% CI: 4.3583, 4.5147) and

a statistically insignificant reduction of 0.0108% (95% CI: $-0.2017, 0.2291$, $p = 0.952$). Figure 11E (red) illustrates that TD3 reduces Sn to 3.9843% (95% CI: 3.8907, 4.0782), achieving a reduction of 0.5039% (95% CI: 0.2862, 0.7366, $p < 0.001$). Overall, all algorithms except PPO consistently reduce slag Sn content, and DADDPG delivers the most effective and robust optimization performance.

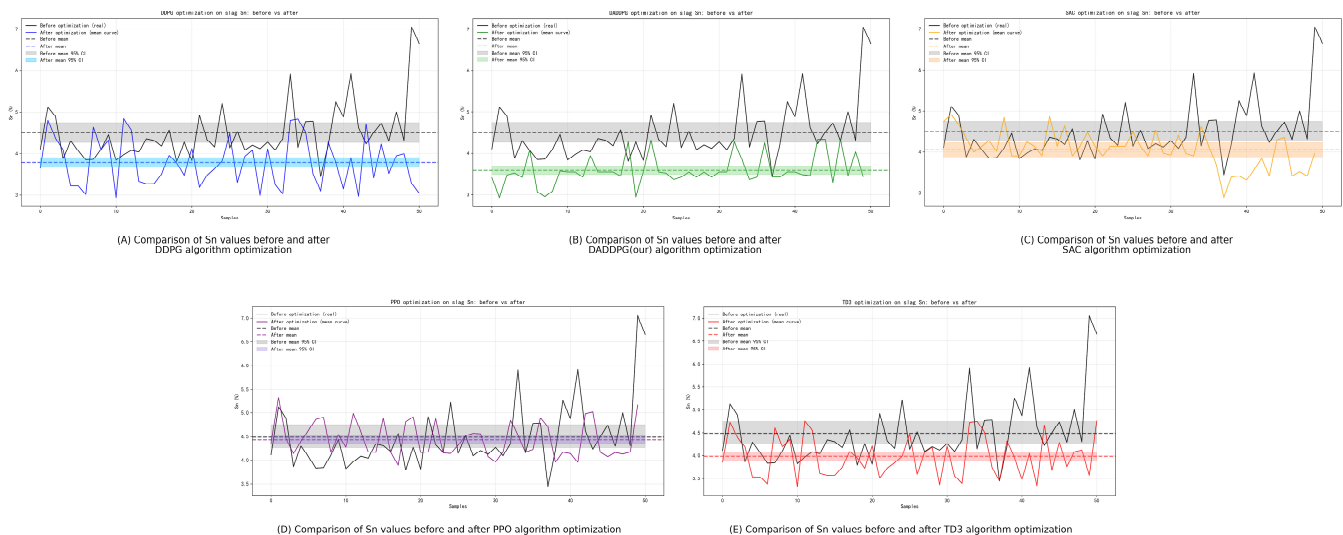


Figure 11. Comparison of tin content in slag before and after optimization using the Five Algorithms.

5. Discussion and Future Work

This paper uses the RL method to address real-world industrial control problems. The experimental results show that the RL controller can effectively reduce the concentration of CO and the tin metal carryover rate in the slag, thereby improving resource utilization and reducing environmental pollution. Additionally, based on the simulation results, the tin content in the slag decreased from 4.49% to 3.59%, which brought consistent economic benefits. According to estimates, if 1000 tons of tin ore are processed annually, the reduction in tin carryover amounts to approximately 9 tons, which corresponds to a saving of about 3,240,000 RMB in tin procurement costs. However, this method has only been tested through computer simulation experiments, and although the results from these simulations are promising, there are two shortcomings and limitations that need further research before it can be applied in actual factories.

1. The RL environment in computer simulations cannot fully replicate the complex dynamic characteristics of the actual tin smelting process, such as equipment wear, environmental disturbances (e.g., temperature and humidity fluctuations), and variations in raw material batches. These factors result in discrepancies between the simulated environment and real-world production settings, potentially leading to safety risks or economic losses if the strategies trained in the simulated environment are directly applied to actual control. In particular, the Ausmel furnace is essentially a black-box industrial process, and many critical internal variables cannot currently be measured online under the existing sensing and detection infrastructure. Therefore, a rigorous quantitative assessment of the mismatch between the simulated environment and the real production system across the full state space is still challenging at this stage.
2. The optimized performance was evaluated using a simulated RL environment, a common practice in RL research. However, for a complex, multidimensional industrial environment like tin smelting, simulated evaluations have certain limitations. As the RL environment is constructed based on specific training data, it struggles

to comprehensively cover all scenarios encountered in actual production. Consequently, the agent's performance in unseen scenarios may not be accurately predicted through simulated evaluations, leading to uncertainties regarding the applicability and robustness of the optimized performance in real-world production environments.

Despite the aforementioned limitations, the experimental results demonstrate that RL methods can effectively adapt to the control requirements of the tin smelting process, thereby helping factories reduce costs and improve efficiency. Currently, the literature rarely explores the application of RL methods to the control of rare metal smelting, with applications in tin smelting being particularly scarce, making this study pioneering in its significance. To address the above limitations, we propose the following future work plan:

1 Regarding the limitations of deviations between simulation and reality:

Collaborate closely with experts in the tin smelting domain to collect data that more accurately simulates key factors in the tin smelting process, including chemical reactions, equipment wear, environmental disturbances, and raw material variations. Based on this data, construct a high-fidelity RL environment that closely aligns with actual production conditions, thereby minimizing the discrepancy between simulation and reality.

2 Limitations of simulation evaluation:

We plan to develop a miniaturized Ausmel furnace and spray gun system with operational functions identical to those in an actual factory. The trained RL strategy will be deployed on this miniature equipment for testing to validate its control effectiveness in an environment closely resembling real production conditions, thereby addressing the shortcomings of simulated evaluations. Furthermore, multi-step rollout consistency checks will be conducted by comparing simulated trajectories with logged real production trajectories, enabling progressive reduction in the sim-to-real gap through iterative refinement and near-real validation.

6. Conclusions

This study proposes a model-free, data-driven approach based on RL for intelligent control of the tin smelting process. Utilizing real industrial production data, an RL environment was constructed, and a novel Dual-Action Discriminative Deep Deterministic Policy Gradient (DADDPG) algorithm was developed. Experimental results demonstrate that the proposed method reduces tin content in slag to 3.5–4% and CO content in exhaust gas to 2000–2700 ppm, consistently outperforming traditional RL algorithms, including DDPG, SAC, TD3, and PPO. Notably, DADDPG exhibits superior adaptability in complex, dynamic industrial environments, underscoring its substantial advantages in optimizing tin smelting parameters. By learning control strategies directly from data without relying on precise dynamic models, this approach offers an efficient and practical solution for intelligent control in this domain. However, the current simulation environment does not account for practical factors such as equipment wear and raw material variations, limiting its industrial deployment. Future work will focus on high-fidelity environment modeling and validation on small-scale industrial setups to address discrepancies between simulation and real-world production.

Author Contributions: Validation, H.Y. (Hang Yan); Resources, H.Y. (Haibin Yuan) and H.Y. (Hang Yan); Data curation, H.Y. (Haibin Yuan); Writing—original draft, Y.L. and Z.X.; Supervision, L.Y.; Funding acquisition, L.Y. All authors have read and agreed to the published version of the manuscript.

Funding: This work was supported by the Major Science and Technology Projects in Yunnan Province (202302AG050009, 202302AB080014).

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: The data presented in this study are not publicly available due to confidentiality restrictions under a non-disclosure agreement with the industrial partner.

Conflicts of Interest: Authors Haibin Yuan and Hang Yan were employed by the company Tin Branch of Yunnan Tin Company Limited. The remaining authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

References

- Lewis, F.L.; Vrabie, D.; Vamvoudakis, K.G. Reinforcement learning and feedback control: Using natural decision methods to design optimal adaptive controllers. *IEEE Control Syst. Mag.* **2012**, *32*, 76–105. [\[CrossRef\]](#)
- Spielberg, S.P.K.; Gopaluni, R.B.; Loewen, P.D. Deep reinforcement learning approaches for process control. In Proceedings of the 2017 6th International Symposium on Advanced Control of Industrial Processes, Taiwan, Taipei, 28–31 May 2017; IEEE: New York, NY, USA, 2017; pp. 201–206. [\[CrossRef\]](#)
- Liu, D.; Wei, Q. Policy iteration adaptive dynamic programming algorithm for discrete-time nonlinear systems. *IEEE Trans. Neural Netw. Learn. Syst.* **2014**, *25*, 621–634. [\[CrossRef\]](#)
- Fallah-Mehrjardi, A.; Hidayat, T.; Abdeyazdan, H.; Shishin, D.; Shevchenko, M.; Hayes, P.C.; Jak, E. Integrated experimental and thermodynamic modeling study of slag–matte–metal equilibrium in the Pb–Fe–O–S–Si–(Al, Ca, Zn) systems at 1100–1200 °C. *J. Mater. Res. Technol.* **2024**, *31*, 1754–1768. [\[CrossRef\]](#)
- Siyambaş, Y.; Turgut, Y. Experimental investigation and optimization of the effects of manufacturing parameters on geometric tolerances in additive manufacturing of AlSi10Mg alloy. *Int. J. Adv. Manuf. Technol.* **2024**, *134*, 415–429. [\[CrossRef\]](#)
- Wang, Y.; Wen, Z.; Li, H.; Jiang, Z.; Xu, Q. Neural network model construction and analysis for simulating and predicting hazardous element allocation in copper smelting key links. *J. Hazard. Mater.* **2025**, *494*, 138725. [\[CrossRef\]](#)
- Bak, C.; Ghosh Roy, A.; Son, H. Quality prediction for aluminum diecasting process based on shallow neural network and data feature selection technique. *CIRP J. Manuf. Sci. Technol.* **2021**, *33*, 327–338. [\[CrossRef\]](#)
- Zhu, J.; Jia, M.; Zhang, Y.; Deng, H.; Liu, Y. Transductive transfer broad learning for cross-domain information exploration and multigrade soft sensor application. *Chemom. Intell. Lab. Syst.* **2023**, *235*, 104778. [\[CrossRef\]](#)
- Sui, F.P.; Guo, R.; Zhang, Z.; Gu, G.X.; Lin, L.W. Deep reinforcement learning for digital materials design. *ACS Mater. Lett.* **2021**, *3*, 1433–1439. [\[CrossRef\]](#)
- Sequeira, S.E.; Graells, M.; Puigjaner, L. Real-time evolution for on-line optimization of continuous processes. *Ind. Eng. Chem. Res.* **2002**, *41*, 1815–1825. [\[CrossRef\]](#)
- Chai, T.; Jia, Y.; Li, H.; Wang, H. An intelligent switching control for a mixed separation thickener process. *Control Eng. Pract.* **2016**, *57*, 61–71. [\[CrossRef\]](#)
- Igel'nik, B.; Pao, Y.-H. Stochastic choice of basis functions in adaptive function approximation and the functional-link net. *IEEE Trans. Neural Netw.* **1995**, *6*, 1320–1329. [\[CrossRef\]](#)
- Deng, J.; Sierla, S.; Sun, J.; Vyatkin, V. Reinforcement learning for industrial process control: A case study in flatness control in steel industry. *Comput. Ind.* **2022**, *143*, 103748. [\[CrossRef\]](#)
- Ferreira Neto, W.A.; Cavalcante, C.A.V.; Do, P. Deep reinforcement learning for maintenance optimization of a scrap-based steel production line. *Reliab. Eng. Syst. Saf.* **2024**, *249*, 110199. [\[CrossRef\]](#)
- Wang, D.; Liu, D.; Wei, Q.; Zhao, D.; Jin, N. Optimal control of unknown nonaffine nonlinear discrete-time systems based on adaptive dynamic programming. *Automatica* **2012**, *48*, 1825–1832. [\[CrossRef\]](#)
- Han, Z.; Pedrycz, W.; Zhao, J.; Wang, W. Hierarchical granular computing-based model and its reinforcement structural learning for construction of long-term prediction intervals. *IEEE Trans. Cybern.* **2022**, *52*, 666–676. [\[CrossRef\]](#)
- Zheng, J.; Jia, R.; Liu, S.; He, D. Safe reinforcement learning for industrial optimal control: A case study from metallurgical industry. *Inf. Sci.* **2023**, *649*, 119684. [\[CrossRef\]](#)
- Chen, C.; Cui, M.; Li, F.; Yin, S.; Wang, X.; Wang, X. Model-free emergency frequency control based on reinforcement learning. *IEEE Trans. Ind. Inform.* **2021**, *17*, 2336–2346. [\[CrossRef\]](#)
- Duan, J.; Shi, D.; Diao, R.; Li, H.; Wang, Z.; Zhang, B.; Bian, D.; Yi, Z. Deep reinforcement learning-based autonomous voltage control for power grid operations. *IEEE Trans. Power Syst.* **2020**, *35*, 814–817. [\[CrossRef\]](#)
- Fan, H.; Zhu, L.; Yao, C.; Guo, Y.; Jiao, X.; Jin, J. Deep reinforcement learning for energy efficiency optimization in wireless networks. In Proceedings of the 2019 IEEE 4th International Conference on Cloud Computing and Big Data Analysis, Chengdu, China, 12–15 April 2019; IEEE: New York, NY, USA, 2019; pp. 465–471. [\[CrossRef\]](#)

21. Ning, Z.; Zhang, K.; Wang, X.; Obaidat, M.S.; Guo, L.; Hu, X.; Hu, B.; Guo, Y.; Sadoun, B.; Kwok, R.Y.K. Joint computing and caching in 5G-envisioned Internet of Vehicles: A deep reinforcement learning-based traffic control system. *IEEE Trans. Intell. Transp. Syst.* **2021**, *22*, 5201–5212. [[CrossRef](#)]
22. Zhang, T.; Yu, L.; Yue, D.; Dou, C.; Xie, X.; Hancke, G.P. Two-timescale coordinated voltage regulation for high renewable-penetrated active distribution networks considering hybrid devices. *IEEE Trans. Ind. Inform.* **2024**, *20*, 3456–3467. [[CrossRef](#)]
23. Lu, R.; Jiang, Z.; Wu, H.; Ding, Y.; Wang, D.; Zhang, H.-T. Reward shaping-based actor-critic deep reinforcement learning for residential energy management. *IEEE Trans. Ind. Inform.* **2023**, *19*, 2662–2673. [[CrossRef](#)]
24. Fu, T.; Liu, S.; Li, P. Digital twin-driven smelting process management method for converter steelmaking. *J. Intell. Manuf.* **2025**, *36*, 2749–2765. [[CrossRef](#)]
25. Li, J.; Wang, X.; Yang, Q.; Sun, Q.; Zhao, C.; Mao, Z.; Qie, Z. IoT-based framework for digital twins in steel production: A case study of key parameter prediction and optimization for CSR. *Expert Syst. Appl.* **2024**, *250*, 123909. [[CrossRef](#)]
26. Jia, R.; Zhang, M.; Zheng, J.; He, D. Offline constrained reinforcement learning for batch-to-batch optimization of cobalt oxalate synthesis process. *Chem. Eng. Res. Des.* **2024**, *209*, 138–152. [[CrossRef](#)]
27. Schäfer, M.; Faltings, U.; Glaser, B. Machine learning approach for predicting tramp elements in the basic oxygen furnace based on the compiled steel scrap mix. *Sci. Rep.* **2025**, *15*, 2430. [[CrossRef](#)]
28. Huang, W. *Tin*; Metallurgical Industry Press: Beijing, China, 2001; pp. 93–99.
29. Song, X. *Tin Metallurgy*; Metallurgical Industry Press: Beijing, China, 2011; pp. 21–92.
30. Mishra, S. A Huber reward function-driven deep reinforcement learning solution for cart-pole balancing problem. *Neural Comput. Appl.* **2023**, *35*, 16705–16722. [[CrossRef](#)]
31. Dalla Libera, A.; Romeres, D.; Jha, D.K.; Yerazunis, B.; Nikovski, D. Model-based reinforcement learning for physical systems without velocity and acceleration measurements. *IEEE Robot. Autom. Lett.* **2020**, *5*, 1911–1916. [[CrossRef](#)]
32. Lukas, M.; Tomicic, I.; Bernik, A. Anticheat system based on reinforcement learning agents in Unity. *Information* **2022**, *13*, 173. [[CrossRef](#)]
33. Ranaweera, M.; Mahmoud, Q.H. Deep Reinforcement Learning with Godot Game Engine. *Electronics* **2024**, *13*, 985. [[CrossRef](#)]
34. Zhang, R.; Zong, Q.; Zhang, X.; Dou, L.; Tian, B. Game of drones: Multi-UAV pursuit-evasion game with online motion planning by deep reinforcement learning. *IEEE Trans. Neural Netw. Learn. Syst.* **2023**, *34*, 7900–7909. [[CrossRef](#)] [[PubMed](#)]
35. Xue, Y.; Kudenko, D.; Khosla, M. Graph learning-based generation of abstractions for reinforcement learning. *Neural Comput. Appl.* **2025**, *37*, 13187–13207. [[CrossRef](#)]
36. Blad, C.; Bøgh, S.; Kallesøe, C.S. Data-driven Offline Reinforcement Learning for HVAC-systems. *Energy* **2022**, *261*, 125290. [[CrossRef](#)]
37. Wu, N.; Yang, Z.; Li, R.; Wei, N.; Chen, Y.; Dong, Q.; Li, J.; Zheng, G.; Gong, X.; Gao, F.; et al. High-fidelity data-driven dynamics model for reinforcement learning-based control in HL-3 tokamak. *Commun. Phys.* **2025**, *8*, 393. [[CrossRef](#)]
38. Silver, D.; Lever, G.; Heess, N.; Degris, T.; Wierstra, D.; Riedmiller, M. Deterministic policy gradient algorithms. In Proceedings of the 31st International Conference on Machine Learning, Beijing, China, 22–24 June 2014; PMLR: Beijing, China, 2014; pp. 387–395.
39. Haarnoja, T.; Zhou, A.; Abbeel, P.; Levine, S. Soft actor-critic: Off-policy maximum entropy deep reinforcement learning with a stochastic actor. In Proceedings of the 35th International Conference on Machine Learning, Stockholm, Sweden, 10–15 July 2018; PMLR: Stockholm, Sweden, 2018; pp. 1861–1870. Available online: <https://proceedings.mlr.press/v80/haarnoja18b.html> (accessed on 20 January 2026).
40. Fujimoto, S.; van Hoof, H.; Meger, D. Addressing function approximation error in actor-critic methods. In Proceedings of the 35th International Conference on Machine Learning, Stockholm, Sweden, 10–15 July 2018; PMLR: Stockholm, Sweden, 2018; pp. 1587–1596. Available online: <https://proceedings.mlr.press/v80/fujimoto18a.html> (accessed on 20 January 2026).
41. Gu, Y.; Cheng, Y.; Chen, C.L.P.; Wang, X. Proximal policy optimization with policy feedback. *IEEE Trans. Cybern.* **2022**, *52*, 4600–4610. [[CrossRef](#)]

Disclaimer/Publisher’s Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.