Journal of Engineering, Project, and Production Management
2026, 16(6), 2026-125

Steady-State Control and Optimization of Process Quality of High-Speed Unit in Wrapping Workshop

Shuxi Guo[1], Jinghui Guo[2], Zhijun Xu[3], and Shaolong Han[4]
[1] Engineer, Process Quality Department, Hebei Baisha Tobacco Co., Ltd, Shijiazhuang 052165, China, E-mail: shuxi@hpu123.uu.me
[2] Director, Process Quality Department, Hebei Baisha Tobacco Co., Ltd, Shijiazhuang 052165, China, E-mail: Jinghui@hpu123.uu.me
[3] Chief Inspector, Process Quality Department, Hebei Baisha Tobacco Co., Ltd, Shijiazhuang 052165, China, E-mail: zhijun@hpu123.uu.me
[4] Senior Engineer, Silk Making Workshop, Hebei Baisha Tobacco Co., Ltd, Shijiazhuang 052165, China, E-mail: shaolong15@outlook.com (corresponding author).

Production Management
Received January 25, 2026; revised March 31, 2026; accepted June 1, 2026
Available online July 2, 2026

Abstract: High-speed equipment in the wrapping workshop exhibits strong nonlinearity, multi-variable coupling, and frequent disturbances in actual production. This makes it difficult for the traditional control method to maintain the stable consistency of tension, material supply and rolling pressure for a long time at high speed and is prone to steady-state deviation and quality fluctuations. In response to the above shortcomings, this study proposes a steady-state quality control method based on deep reinforcement learning. This method achieves adaptive adjustment of the high-speed wrapping process by constructing a dynamic prediction model, a policy network, a value evaluation network, and a steady-state deviation suppression mechanism. Experiments showed that in typical steady-state tests, the average steady-state deviation of tension, feed amount, and rolling pressure of the proposed method was only 0.04, which was 70% lower than that of the traditional control method. Under continuous disturbance conditions, the drift amplitude of the proposed variables was always controlled within 0.10. In the multi-source disturbance scenario, the Root Mean Square Error (RMSE) of the proposed method remained within a low range (0.15-0.23), maintaining high output consistency under strong disturbance conditions. The constructed intelligent control framework could automatically learn control rules without the need for precise mathematical models and achieve steady-state quality maintenance of high-speed wrapping equipment under complex disturbances. This study provides a feasible technical path for smart manufacturing and stabilization of cigarette production.

Keywords: High-speed wrapping process; steady-state quality control; deep reinforcement learning; multi-variable coupling; disturbance suppression; intelligent manufacturing.

DOI 10.32738/JEPPM-2025-125

1. Introduction

As the level of intelligence and high-speed in the cigarette industry continues to improve, the High-Speed Unit (HSU) in the rolling workshop plays an increasingly prominent role in the production line. HSU exhibits typical characteristics such as strong nonlinearity, Multi-Variable Coupling (MVC) and rapid changes in working conditions during operation, making it difficult for traditional empirical or fixed parameter control methods to maintain stable and high-quality production status within the full range of working conditions (Giannopoulos et al., 2023a; Cui et al., 2023a). Under high-speed conditions, factors such as raw material fluctuations, mechanical wear, and environmental disturbances will further amplify the uncertainty of system dynamics, causing deviations in key process variables such as paper feeding tension, tobacco supply, and rolling pressure (Groumpos et al., 2023a; Li et al., 2024a). In recent years, some research has been carried out on intelligent control methods for complex industrial systems. He et al. (2022) proposed a three-point symmetrical eccentric planetary gear train and improved a special clutch to solve the high failure rate and fragile cigarette packs in the soft box packaging machine cigarette pack steering mechanism. It analyzed wear to find defects, optimized structure and overload protection, thereby improving operational stability and reducing faults and quality risks. To solve the increasing quality and energy consumption requirements of cigarette production lines, Jin et al. (2023) proposed a silk-making parameter optimization scheme based on big data and Artificial Intelligence (AI). This program collected and analyzed massive data

by building an AI detection system, focusing on optimizing the leaf threshing and re-baking process, thus improving the leaf yield rate. Moreno et al. (2022) proposed a multi-variable generalized superhelical algorithm based on Lyapunov to solve multi-variable systems containing time-varying uncertain control matrices and disturbances. The algorithm designed a set of constant gains to stabilize the algorithm's global finite-time output within a reasonable uncertainty bound, thereby improving the robustness and convergence speed of the industrial system.

Deep Reinforcement Learning (DRL) shows performance advantages over classical control due to its self-learning ability, strong nonlinear representation ability, and adaptive ability to unknown environments. DRL can automatically learn optimal strategies through interaction with complex systems without the need for explicit mathematical models, and is suitable for handling multi-variable, multi-constraint, and strongly coupled industrial scenarios (Boute et al., 2022). Degrave et al. (2022) proposed a full-coil autonomous control architecture based on RL to address difficult high-frequency closed-loop control of high-temperature plasma shape and large configuration switching design in tokamak magnetic confinement fusion. It solved magnetic field instructions in real-time through an end-to-end policy network under high-level goal constraints, thereby achieving RL control of extremely complex fusion systems. De Marco et al. (2023) proposed an AI flight control based on deep deterministic policy gradient to address the strong coupling and nonlinear control of high-performance fighter jets in complex maneuvers and full autonomous waypoint tracking. A dedicated reward function was designed and trained on a 6-DoF high-fidelity model, which achieved fully autonomous completion of maneuvers such as rapid and continuous turns at high/low altitudes and within a large Mach number range. Bao et al. (2021) proposed an improved deep deterministic actor-critic predictor to address the poor controller adaptability caused by strong nonlinearity in industrial processes and frequent model maintenance. It decoupled the immediate reward from the action value function and derived the expected policy gradient under the assumption of normal state distribution to achieve early reliable updates and online accelerated learning, improving the stability and convergence speed of the control strategy.

In summary, although existing research has made significant progress in cigarette equipment structure optimization, intelligent control of silk-making process parameters, and robust control and DRL strategy design of complex industrial systems. However, for the HSU in the roll wrap shop, a typical discrete manufacturing process with multiple variables, strong coupling, and frequent disturbances, there is still a lack of a unified intelligent Steady-State Control (SSC) framework for all working conditions. At present, existing methods mostly focus on mechanical structure improvement or local process optimization to verify the feasibility of DRL in continuous industrial processes. An adaptive control strategy that can directly solve the multi-variable synergy and stability of tension, material supply and winding in the High-Speed Rolling (HSR) process has not yet been formed. At the same time, the influence of raw material fluctuations, equipment wear and disturbance accumulation on process variables under high-speed working conditions shows a non-linear amplification trend. Existing controllers generally rely on model accuracy or parameter tuning, and are difficult to support real-time learning, online adjustment, and Steady-State Quality (SSQ) consistency assurance. Therefore, this study constructs an SSC method for HSU process quality in the wrapping workshop that integrates DRL to provide a feasible new intelligent control paradigm for intelligent manufacturing and high-quality production in the wrapping workshop. This method achieves high precision and highly adaptable SSC of the multi-variable wrapping process by introducing a dynamic prediction model to expand the training data and using reward and value function design for Steady-State Error (SSE) suppression.

The research contributions of this paper are as follows:

(1) Addressing the strong coupling characteristics of tension, yarn supply, and pressure during high-speed winding, a multi-variable SSC method based on DRL is constructed to achieve adaptive adjustment under complex working conditions.

(2) A strategy optimization mechanism integrating Steady-State Deviation (SSD) guidance and motion smoothing constraints is proposed to improve the convergence stability and output smoothness of the control strategy.

(3) A robust value update method considering multi-source disturbances is designed to enhance the model's resistance to disturbances under conditions of raw material fluctuations and environmental changes.

(4) A training and verification system driven by digital twins is constructed, and the effectiveness of the method is verified through actual industrial systems. Experimental results show that the key performance indicators are improved compared with traditional methods.

2. Research Methodology

First, in view of the strong nonlinearity and MVC characteristics of the high-speed wrapping process, a data-driven dynamic model based on real working conditions is constructed, and a DRL control framework that integrates the strategy network and the value network is built. Secondly, by introducing mechanisms such as action smoothing, SSD guidance, and robust disturbance correction, the convergence stability and environmental adaptability of the strategy under complex disturbances and changes in all working conditions are further improved.

2.1. System Characteristic Analysis and Dynamic Modeling of HSR Wrapping Process

The HSR wrapping unit shows typical MVC and strong nonlinear characteristics during operation. The key process variables interact with each other and are affected by multiple disturbances such as raw materials, equipment and environment at high-speed (Suzuki et al., 2005; Zhortuylov et al., 2021). Therefore, to construct a control strategy suitable for DRL, this study first systematically describes the variable structure, disturbance path and dynamic behavior of the wrapping process. A state space expression is formed for algorithm learning. The HSR process is defined as four main variables: paper feeding force T_t, tobacco supply amount F_t, rolling pressure P_t and linear speed v_t. Tension, wire

supply, and pressure directly determine the molding stability of the tobacco rod, while speed, as a unified driving factor, will amplify the dynamic deviation of all variables (Seo et al., 2024; Tang et al., 2025). The system is also subject to comprehensive disturbances d_t caused by fluctuations in raw material properties, mechanical wear, and changes in workshop temperature and humidity. Therefore, the key variables and disturbance transmission structure of the HSR process are shown in Fig. 1.

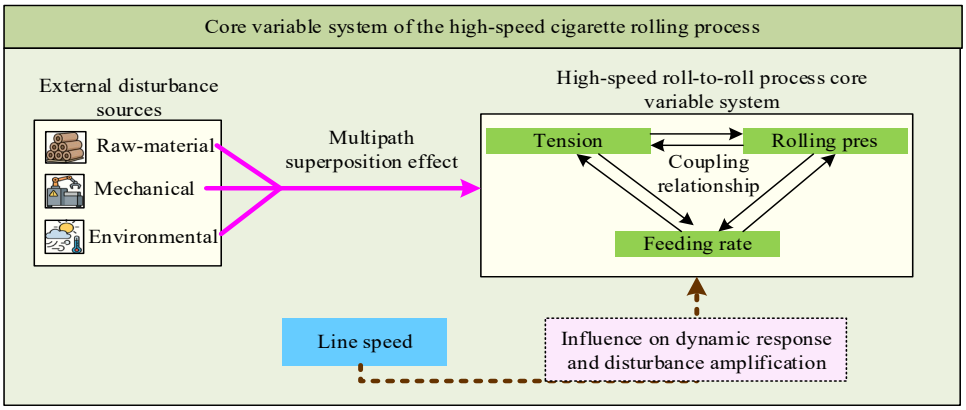


Fig. 1. Key variables and disturbance transmission structure in HSR process

In Fig. 1, tension T_t, yarn supply amount F_t and rolling pressure P_t are connected in two directions to form a strong coupling relationship. A shift in any one variable affects other variables through coupling links. Linear speed v_t, as an external amplification factor, will enhance the fluctuation and response sensitivity of all variables during high-speed operation. Peripheral raw materials, machinery and environmental disturbances simultaneously affect the three core variables through multiple action paths, forming a complex source of compound disturbances in the wrapping process. To enable the DRL model to accurately describe the process, this study organizes the above key variables into state vectors, as shown in Eq. (1).

$$s_t = [T_t, F_t, P_t, v_t, d_t] \tag{1}$$

In Eq. (1), $\boldsymbol{s}_t$ is a multidimensional state vector. To clarify the composition of the state space, this study defines the physical meaning, sampling unit, and typical fluctuation range of five categories of states, as shown in Table 1.

Table 1. Definition and physical meaning of state variables for the high-speed cigarette making and packing system

State variable	Physical meaning	Unit	Typical range
T_t	Tensile force applied to the paper during traction and forming reflects stability of paper shaping	N	2.5-4.0
F_t	Mass flow rate of tobacco shreds delivered per unit time; reflects density and filling uniformity	g/s	5.0-7.5
P_t	Compression pressure applied in the rolling section reflects rod tightness and structural stability	kPa	10-20
v_t	Machine operating speed, determining system dynamic response and disturbance amplification	m/s	2.0-4.5
d_t	Equivalent representation of unpredictable external factors (raw material, mechanical, environmental); models external disturbances	-	No fixed range

The value ranges of each state variable in Table 1 are derived from the statistical results of actual production operation data, reflecting the variation range of the system under typical operating conditions, and are used to guide the model state space construction and training process. This range is an empirical statistical range rather than a hard control constraint, and the variables can fluctuate dynamically within a certain range in actual operation. The dynamic evolution of the packaging process is abstracted into a nonlinear state transition process, as shown in Eq. (2).

$$s_{t+1} = f(s_t, a_t, w_t) \tag{2}$$

In Eq. (2), $\boldsymbol{a}_t$ is the control action vector. $\boldsymbol{w}_t$ is the composite disturbance term vector composed of raw materials, machinery and environmental disturbances. Since the nonlinear mechanism of the real system is difficult to accurately express through explicit models, the function $f(\cdot)$ is modeled, driven by operating data in the study (Pattanayak et al., 2024). To further characterize the sensitivity of the system to disturbances, a partial derivative matrix is used to describe the intensity of the influence of each disturbance on core process variables, as shown in Eq. (3).

$$\frac{\partial s_t}{\partial w_t} = \begin{bmatrix} \frac{\partial T_t}{\partial w_t} & \cdots \\ \frac{\partial F_t}{\partial w_t} & \cdots \\ \frac{\partial P_t}{\partial w_t} & \cdots \end{bmatrix} \tag{3}$$

This matrix represents the coupling sensitivity of the system and can explain why the high-speed wrapping process is prone to SSD under the action of disturbance. In addition, considering that the current state of the high-speed wrapping system basically determines the behavior trajectory at the next moment, the system has typical Markov characteristics. Therefore, this study formalizes the process into a DRL interactive Markov Decision Process (MDP), as shown in Eq. (4) (Jalali Khalil Abadi et al., 2024).

$$M = \{S, A, \rho, r, \gamma\} \tag{4}$$

In Eq. (4), S is the set of state variables corresponding to MDP. A is an executable control action. ρ is an unknown dynamic transfer structure in space. r is the evaluation function of process quality. γ is the discount factor. Through systematic modeling, the high-speed wrapping process can form a state space and dynamic description suitable for DRL learning.

2.2. Design of DRL Control Framework Integrating Dynamic Prediction

Based on the aforementioned dynamic modeling and state space construction of the HSR wrapping process, this study further designs a DRL control framework suitable for complex disturbance environments. This framework takes real-time process data as input and learns the optimal control strategy through interaction with the environment, allowing the system to automatically maintain steady-state consistency of tension, yarn supply amount, and rolling pressure under all working conditions. To ensure learning efficiency and practical use. This paper constructs a control framework based on the Deep Deterministic Policy Gradient (DDPG) algorithm, and improves the network structure, policy update method, and reward function construction to address the characteristics of HSU. The specific structure is shown in Fig. 2.

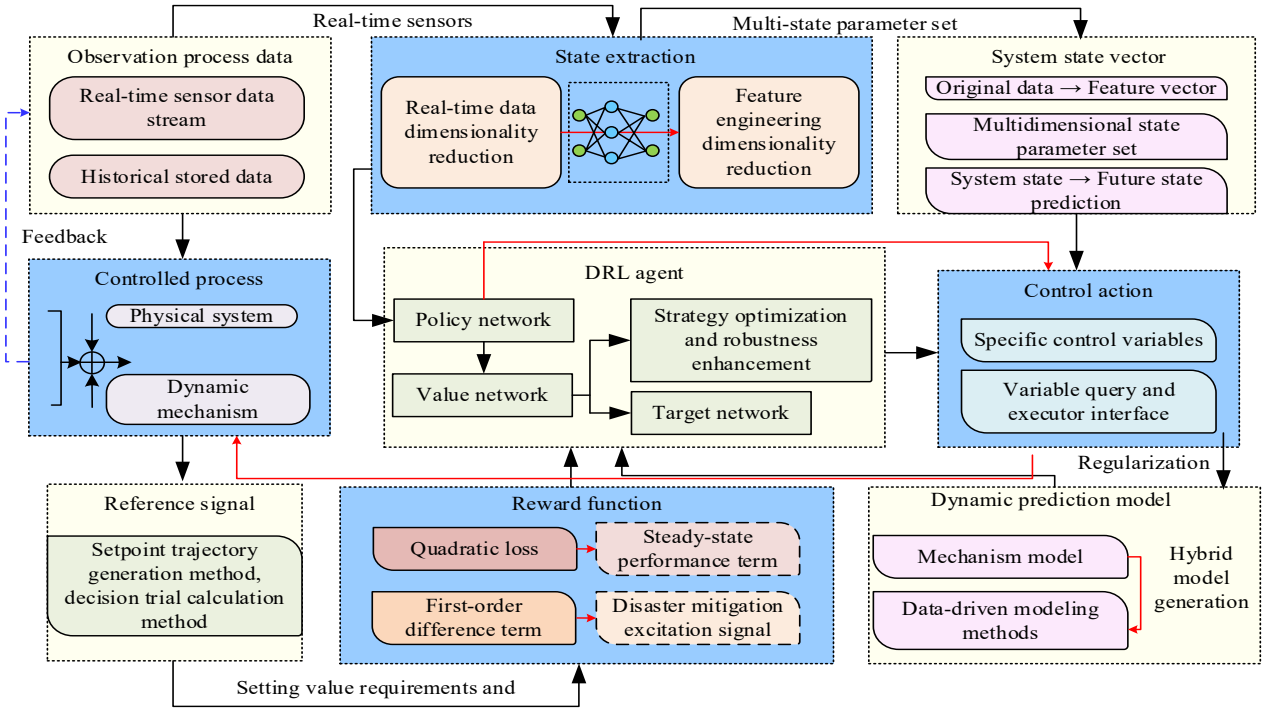


Fig. 2. DRL process quality SSC framework

In Fig. 2, the system state consists of variables such as tension, silk supply amount, pressure, speed, and disturbance, and is input into the strategy network through the state extraction module to generate control actions. The environment feeds back new status and rewards based on actions, which are used to update the value network and strategy network. Through the closed-loop structure of state input, action output, and reward learning, adaptive SSC of high-speed wrapping process is achieved. The DRL control framework constructed in this study is based on the DDPG algorithm, where the policy network and value network correspond to the Actor and Critic structures in DDPG, respectively (Kakooee et al., 2024). The core of the DRL controller is to generate action vector $\boldsymbol{a}_t$ through state vector $\boldsymbol{s}_t$, so that the system can adjust toward the steady-state (Yue, 2026). Therefore, the mapping relationship of the policy network can be expressed as Eq. (5).

$$a_t = \pi_\theta s_t \tag{5}$$

In Eq. (5), π_θ is the policy network with parameter θ. To improve the learning stability in high-speed nonlinear systems, this study adopts a dual network structure based on policy network and value network. The value network is used to evaluate the long-term benefits of actions under the current strategy. The estimated function is shown in Eq. (6).

$$Q_\phi(s_t, a_t) \tag{6}$$

In Eq. (6), Q_ϕ is the action value function network represented by parameter ϕ, consistent with the Critic network in DDPG, which is used to evaluate the long-term reward of the current state-action pair. The estimated function output is used to guide policy updates to prevent the policy from oscillating directly based on sampling rewards. In the dynamic update process, the goal is to maximize the action value, so the policy gradient is shown in Eq. (7).

$$\nabla_\theta J(\theta) = \mathbb{R}_{s_t}[\nabla_\theta \pi_\theta(s_t) \nabla_{a_t} Q_\phi(s_t, a_t)] \tag{7}$$

In Eq. (7), $J(\theta)$ is the strategy objective function. $J(\theta)$ is the gradient of the policy parameters. $\pi_\theta(\boldsymbol{s}_t)$ is the action generated by the current policy network. $\nabla_{\boldsymbol{a}_t} Q_\phi(\boldsymbol{s}_t, \boldsymbol{a}_t)$ is the gradient of the action value function to the action, which is used to guide the strategy on how to adjust actions to achieve higher returns. $\mathbb{R}_{s_t}[\cdot]$ is the expectation for the state distribution, and the strategy update is based on the average result of multiple environmental samplings. Based on the standard DDPG reward function, and considering the disturbance sensitivity of high-speed winding machines in actual operation, the study further constructed three reward components: SSD penalty, motion smoothing constraint, and disturbance robustness term (Al-Kaf et al., 2025; Zhang et al., 2023). The complete reward form is shown in Eq. (8).

$$r_t = -\alpha \left\| e_t \right\|^2 - \beta \left\| \boldsymbol{a}_t - \boldsymbol{a}_{t-1} \right\|^2 - \lambda \left\| d_t \right\| \tag{8}$$

In Eq. (8), e_t is the deviation of the current variable from the target steady-state. α, β, and λ are all adjustable weight coefficients. To further improve the numerical stability of the strategy, this study introduces the target network and soft update mechanism to make the iteration of the value network smoother. Its update strategy is shown in Eq. (9).

$$\phi_{target} \leftarrow \tau\phi + (1-\tau)\phi_{target} \tag{9}$$

In Eq. (9), ϕ_{target} is the updated strategy. τ is the soft update coefficient, with a value of 0.005-0.01. In summary, the proposed DRL control framework adopts real-time collection of state vectors, continuous learning of policy networks, and stable evaluation of value networks. It enables the HSR unit to achieve highly adaptive process quality control in complex disturbance environments and has the ability to independently learn optimal adjustment rules from data.

2.3. Strategy Optimization and Robust Enhancement Mechanism for SSQ

Based on the aforementioned control framework, high-speed wrapping systems will still face strategic oscillation challenges caused by parameter uncertainty, disturbance amplification effects and nonlinear coupling in real operations. This makes it possible that the basic DRL strategy cannot maintain stable output under all operating conditions. Therefore, this study proposes an optimization and robust enhancement method for SSQ from two aspects: the policy training process and the value estimation mechanism, to improve the controllability, stability, and environmental adaptability of the policy. First, based on the Actor update mechanism of the DDPG algorithm, to reduce the abrupt actions of high-speed actuators during the policy update phase, action smoothing constraints are introduced into the Actor update. This mechanism prevents the policy from learning rapidly changing control methods during the training process by applying inhibition terms to the action gradients of consecutive time steps. Its gradient correction expression is shown in Eq. (10).

$$\nabla_\theta J_{smooth} = -\delta(a_t - a_{t-1}) \tag{10}$$

In Eq. (10), $\nabla_\theta J_{\text{smooth}}$ is the gradient correction. δ represents the smoothing coefficient, with a value range of 0.1-0.5. Smooth operation can reduce the mechanical impact of the rolling pressure regulator, wire supply motor and other executive components, and improve the stability of the operation under high-speed operation. The corresponding optimization process is shown in Fig. 3.

In Fig. 3, the current state and the action input at the previous moment work together on the policy network through the SSD module and the action smoothing module to output smooth control actions. Secondly, to improve the control accuracy in the steady-state region of the roll-up process, an SSD guidance mechanism was constructed based on the Temporal-Difference (TD) error update of the DDPG algorithm. Instead of being directly used as a reward splitting term, it modifies the error propagation direction of the value function during the Critic update stage, making the value estimation more sensitive to SSD. The idea c to add an SSD correction term to the TD error, as shown in Eq. (11).

$$\sigma_t^{\text{steady}} = \sigma_t - \kappa \left\| \boldsymbol{s}_t - \boldsymbol{s}^* \right\|^2 \tag{11}$$

In Eq. (11), σ_t^{steady} represents SSD. σ_t is the traditional TD error. $\boldsymbol{s}^*$ is the steady-state target value. κ is the steady-state reinforcement factor. Finally, to enhance the adaptability of the strategy to uncertainties such as raw material fluctuations, mechanical wear and environmental changes, a perturbation robust correction mechanism was designed based

on the Critic function update of the DDPG algorithm. By estimating the amplitude of the disturbance proxy variable d_t It serves as a damping factor for value updates, making the Critic's estimation in high disturbance areas smoother and reducing gradient oscillations. Specifics are shown in Eq. (12).

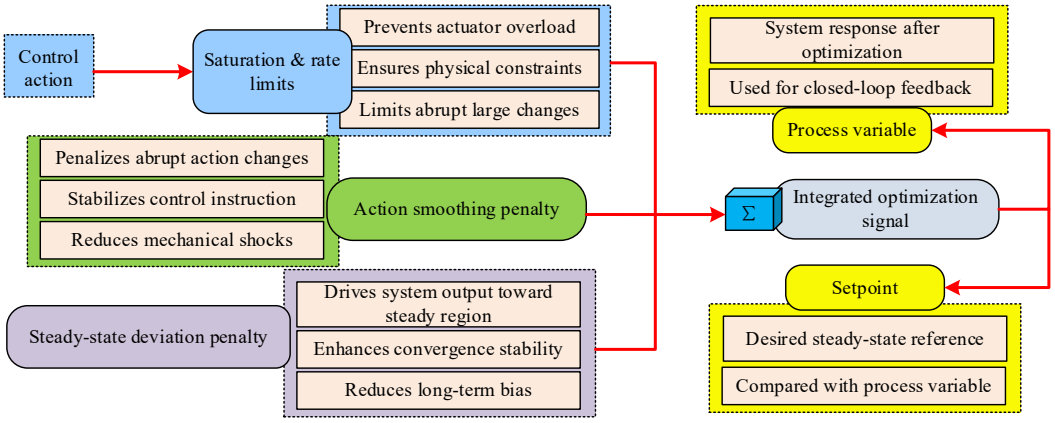


Fig. 3. Strategy smoothing and SSD optimization mechanism

$$\sigma_t^{\text{robust}} = \sigma_t^{\text{steady}} - \vartheta d_t \tag{12}$$

In Eq. (12), σ_t^{robust} is the perturbation robust correction result. ϑ is the disturbance sensitivity coefficient. The overall structure and influence path of the robust correction mechanism are shown in Fig. 4.

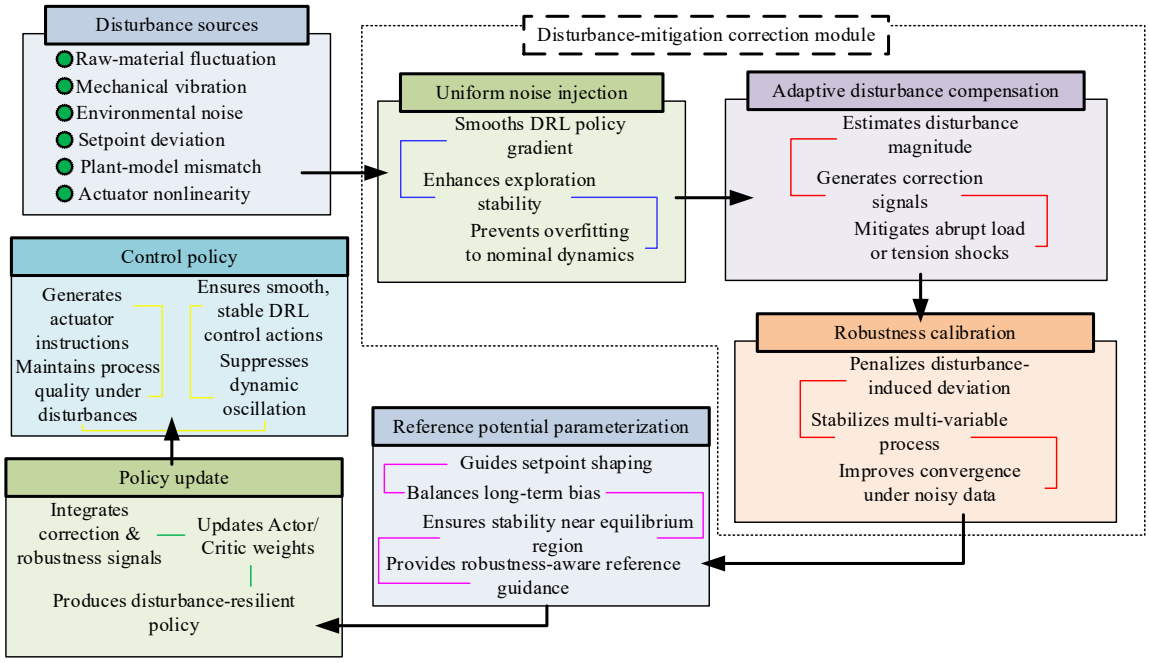


Fig. 4. Policy robustness enhancement mechanism and perturbation suppression path

In Fig. 4, external disturbances first enter the disturbance correction module, and the strategy instability caused by the disturbance is corrected through steps such as unified noise injection, adaptive disturbance compensation, and robust calibration. Secondly, the correction signal and the parameterization result of the reference potential function work together in the strategy update process, so that the generated control strategy has both steady-state tracking capability and interference immunity performance, thereby maintaining stable and smooth control behavior in the HSR MVC process. Based on the above strategy update mechanism, the final value network update form is shown in Eq. (13).

$$\sigma_t^{enhanced} = \sigma_t - \kappa \|s_t - s^*\|^2 - \vartheta d_t \tag{13}$$

The TD error effectively suppresses update oscillations caused by nonlinear coupling and disturbance amplification, allowing Actor-Critic to maintain stable convergence in the full working range and generate smoother and more reliable action output. Based on the nature of training, this policy optimization and robust enhancement mechanism can not only improve the stability and convergence speed of policy learning but also enhance the adaptability of the control strategy to external disturbances. This enables the DRL controller to meet the strict requirements for SSQ consistency of high-speed

wrapping machines, thereby achieving online intelligent control.

3. Results

First, through multiple sets of experiments such as SSE convergence, dynamic response speed, and action smoothness, the proposed improved DRL control method is compared and verified. Secondly, combined with the robustness test of three typical working conditions: sudden disturbance, continuous disturbance, and compound disturbance, the disturbance suppression capability and long-term stability in complex environments are further verified.

3.1. Experimental Setup

To validate the proposed DRL-based steady-state control method, a comprehensive experimental platform consisting of real operational data, a digital twin simulation environment, and a high-performance computing system was constructed. The dataset was collected from a high-speed wrapping unit under multiple shifts and operating speeds, with a total volume of approximately 10^5 samples and a sampling interval of about 0.1 s. The features include tension, tobacco supply rate, rolling pressure, line speed, and disturbance-related variables. The dataset was divided into training, validation, and test sets in a ratio of 70%/15%/15%.

A digital twin environment was developed using a dynamic prediction model, enabling the DRL controller to perform high-frequency interaction learning under extended operating conditions. To validate the consistency between the digital twin and the real system, representative real operational data were used as reference inputs, and the dynamic responses of key variables, including tension, tobacco supply rate, and rolling pressure, were compared between simulation and actual measurements. Quantitative evaluation was conducted using RMSE and relative error. The results show that the RMSE of key variables is approximately within 0.05-0.10, with relative errors controlled within 5%, indicating high fidelity between the simulation and real system.

The model adopts a DDPG-based Actor-Critic framework, including a policy network, a value network, and a target network, with additional action smoothing and SSD guidance mechanisms. The training process consists of 2 million interaction steps with a soft update coefficient of 0.005, and the value network is optimized using mean squared error. To improve generalization, domain randomization is introduced during training by applying stochastic variations in disturbance magnitude, material properties, and environmental conditions. Moreover, the operating speed range of 2.0-4.5 m/s is uniformly sampled to ensure coverage of the full working condition space. The test phase includes three types of typical working conditions. (1) Steady-state working conditions are used to verify that the strategy maintains steady-state consistency of tension, wire supply amount, and rolling pressure. (2) Sudden disturbance working conditions, which simulates step disturbances in silk supply amount, tension impact, and sudden changes in paper humidity to evaluate rapid recovery performance. (3) Continuous disturbance working conditions, including slow changes in raw material quality and accumulation of mechanical wear, were used to verify the robustness of the strategy under long-term operation. The experimental information is listed in Table 2.

Table 2. Experimental hardware and software environment configuration

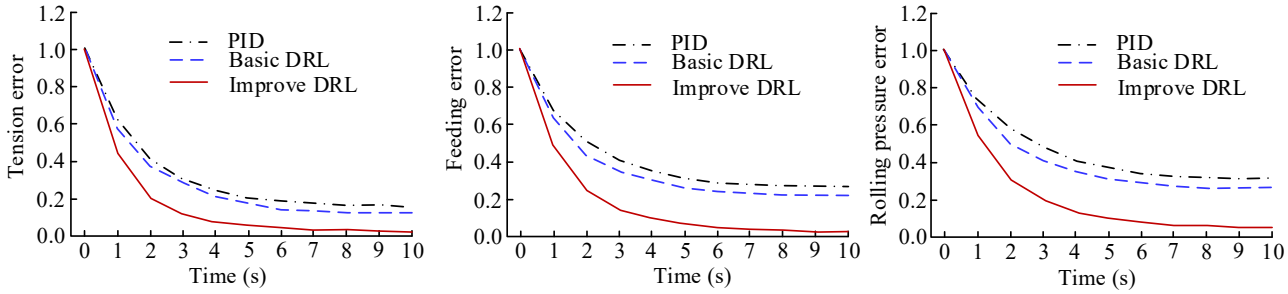
Category	Configuration
Operating system	Ubuntu 22.04 LTS (64-bit)
CPU	Intel Core i9-12900K (16 Cores / 24 Threads, up to 5.2 GHz)
GPU	NVIDIA RTX 3090 (24 GB GDDR6X)
Memory	64 GB DDR5
Deep learning framework	PyTorch 2.2 + CUDA 12.1
Simulation environment	Digital twin model of the high-speed cigarette-packing process trained on operational data (Python 3.10)
Reinforcement learning library	Custom Actor–Critic module + Stable-Baselines3 (for baseline comparison)
Data processing tools	NumPy, Pandas, SciPy, Matplotlib
Deployment verification platform	Cigarette-packing machine PLC + industrial data acquisition system (desensitized production data)

3.2. SSC Performance and Dynamic Response Results

To verify the advantages of the improved DRL control method in maintaining steady-state, this study first compares the SSE convergence process of three types of typical process variables. The comparison methods include a traditional Proportional-Integral-Derivative (PID) controller. Basic DRL strategy without adding a robust enhancement mechanism, and an improved DRL control strategy. All methods are run under the same initial deviation conditions, and the controller's ability in steady-state consistency is evaluated by comparing the error decline rate over time, the final SSD, and the residual oscillation amplitude. The SSE convergence process of the three types of control methods on key process variables is shown in Fig. 5.

In the tension error in Fig. 5(a), the improved DRL quickly reduces the error from 1.00 to 0.20 in 2s, and stabilizes to 0.02 in 10s. The final SSE is 0.12 for the basic DRL and 0.15 for the PID. Compared with other methods, the improved DRL is reduced by 83.33% and 86.67%. In Fig. 5(b), the improved DRL compresses the error to 0.15 in 3s, and the final steady-state remains at 0.03. Compared with the basic DRL and PID, the yarn supply error of the improved DRL is reduced by 86.36% and 88.89%. Regarding the rolling pressure error in Fig. 5(c), the improved DRL quickly drops to 0.13 within

4 s and finally converges to 0.05, which is 80.77% and 83.87% lower than the basic DRL and the PID, respectively. The improved DRL performs optimally in terms of convergence speed, SSD, and oscillation suppression. The dynamic response and overshoot characteristics of the three methods are shown in Fig. 6.

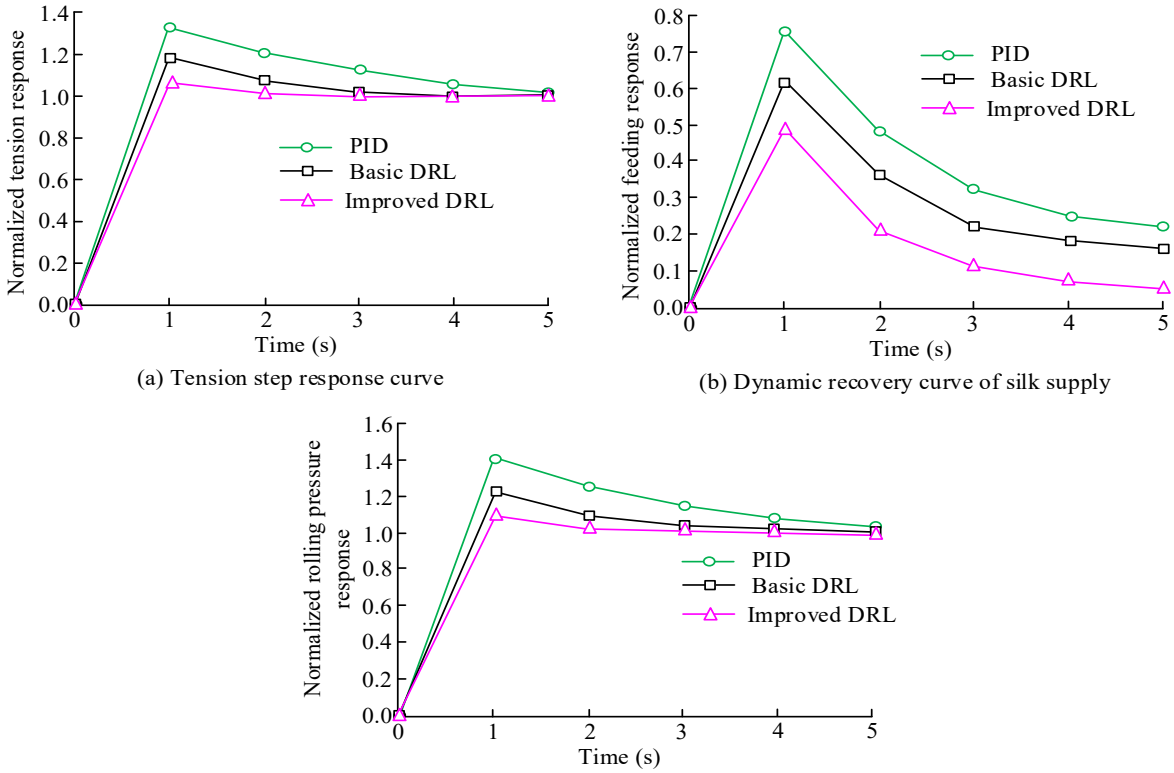


(a) Comparison of tension error convergence

(b) Comparison of feeding error convergence

(c) Comparison of rolling pressure error convergence

Fig. 5. Comparison of SSE convergence



(a) Tension step response curve

(b) Dynamic recovery curve of silk supply

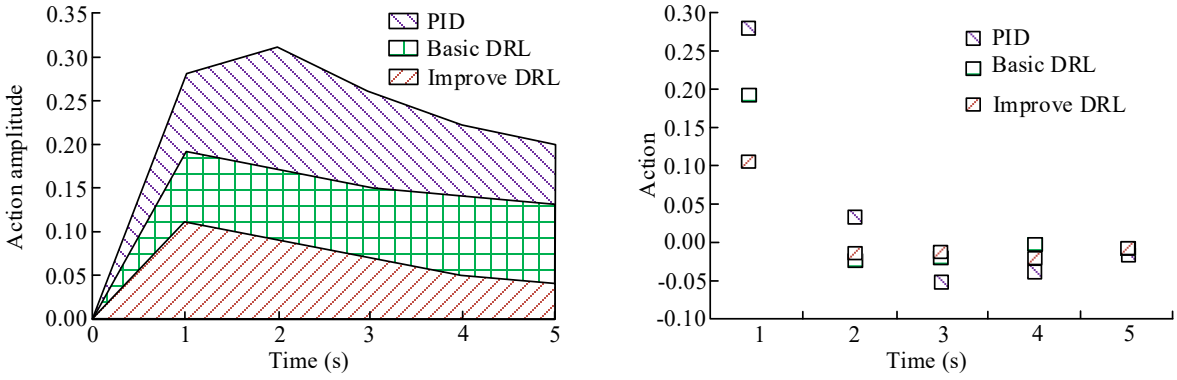
(c) Comparison of rolling pressure response speed

Fig. 6. Comparison of dynamic response performance under typical step disturbances

The tension response data in Fig. 6(a) show that the PID curve fluctuates greatly in the early stage, and the process of converging to the target value of 1.00 is relatively slow. The basic DRLs are 1.18 and 1.07 at 1s and 2s, respectively. The response process has improved, but there is still a certain degree of overshoot and adjustment. In contrast, the improved DRL has faster overall response, shorter adjustment process and smoother curves. In the silk supply amount recovery process in Fig. 6(b), the improved DRL reduces the error to less than 0.10 within 2.70s, which is significantly faster than the recovery time of basic DRL and PID. The rolling pressure response in Fig. 6(c) also reflects the high-speed characteristics of the improved DRL. Compared with the basic DRL and PID, the rolling pressure response speed of the improved DRL is increased by 21.43% and 42.11%. Overall, the improved DRL performs best in overshoot suppression, dynamic response speed, and recovery ability. After introducing the motion smoothing mechanism, the system effectively suppresses overshoot while maintaining a fast response speed. However, as the smoothing coefficient increases, the response process will be slightly prolonged, reflecting the trade-off between response speed and control smoothness. The performance of the three types of control strategies in action smoothness is shown in Fig. 7.

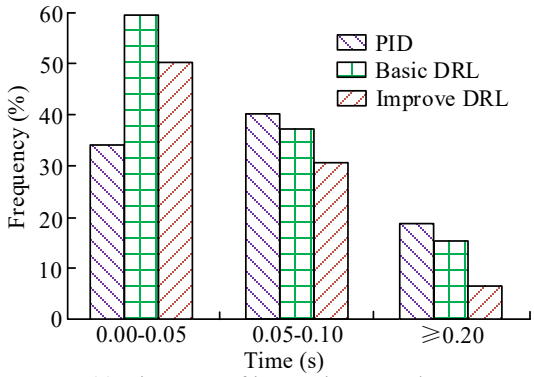
Fig. 7(a) shows the standard deviation of action signal fluctuations. The fluctuation intensity of PID is 0.24, the basic DRL is reduced to 0.15, and the improved DRL is further reduced to 0.08, and the fluctuation amplitude is 66.67% smaller than that of PID. Fig. 7(b) shows the action change rate. The average action accuracy of the improved DRL is only 0.03, which is significantly lower than that of the basic DRL and PID, indicating that its control command changes are more

continuous and without mutations. The instruction smoothness statistics in Fig. 7(c) also confirm this trend. The improved DRL can effectively suppress motion oscillations, reduce the mechanical stress on the actuator, and improve control stability under high-speed conditions. As the smoothing coefficient increases, motion ripple and overshoot decrease significantly, but the system response speed decreases slightly, reflecting the trade-off between overshoot and response time.



(a) Amplitude of motion signal fluctuation

(b) Comparison of motion change rates



(c) Histogram of instruction smoothness

Fig. 7. Comparison of motion smoothness

3.3. Disturbance Suppression Capability and Robustness Verification

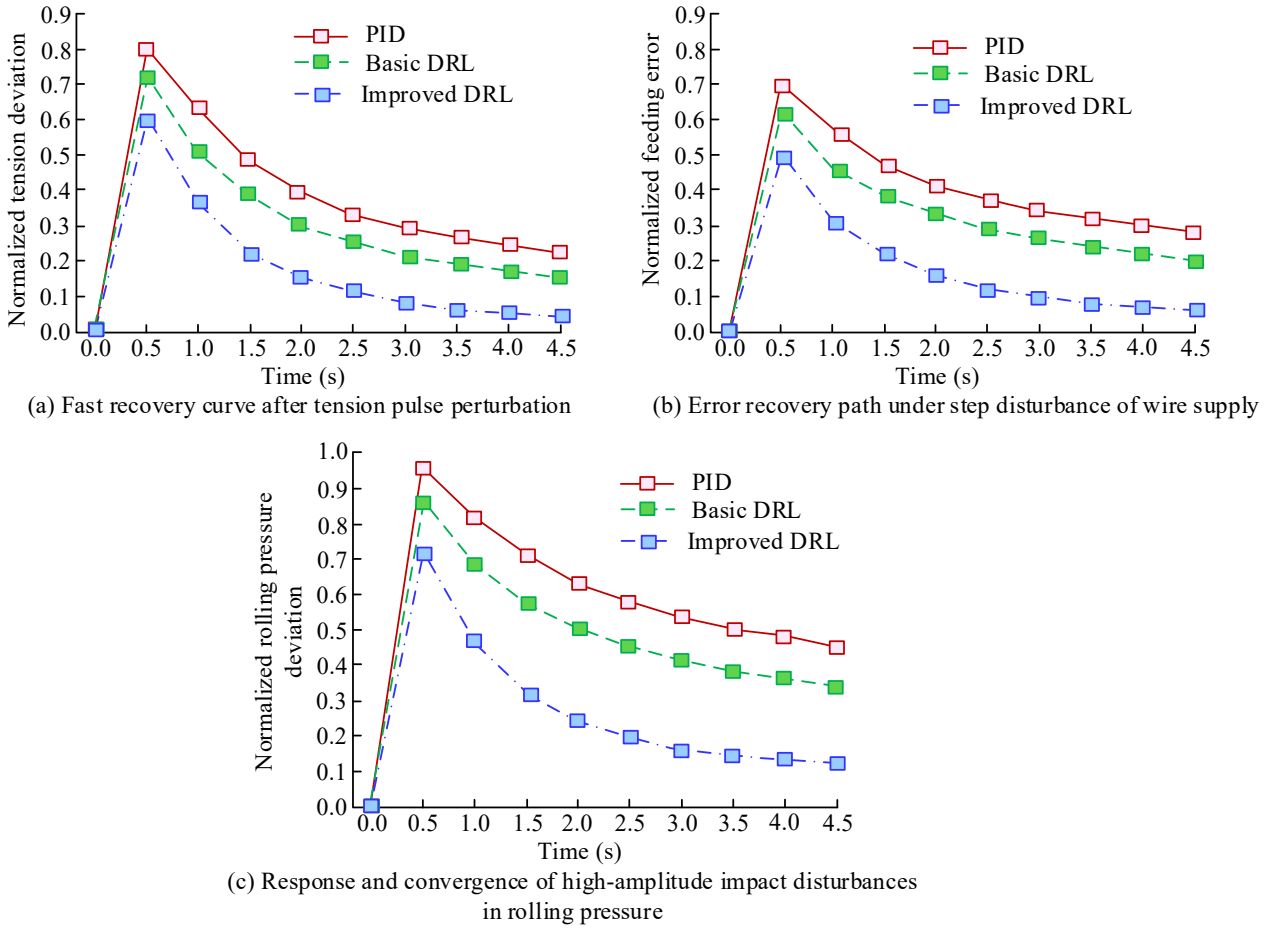
In actual high-speed wrapping conditions, the system not only needs to maintain steady-state accuracy, but also must face multi-source disturbance factors such as raw material fluctuations, mechanical wear, and paper humidity changes. Therefore, this study further compares the anti-disturbance performance of PID, basic DRL, and improved DRL from three dimensions: sudden disturbance recovery capability, stability under continuous disturbance, and robust output characteristics under compound disturbance situations. The transient recovery conditions of the three methods under step and pulse disturbances are shown in Fig. 8.

In Fig. 8(a), the PID deviation is still 0.62 at 1s, the basic DRL is reduced to 0.50, and the improved DRL is further reduced to 0.36. The recovery speed of the improved DRL is significantly accelerated, and it is close to that of SSD in 2s to 3s. In the step disturbance of the wire supply amount in Fig. 8(b), at 4.5s, the remaining errors of PID and basic DRL are still 0.28 and 0.20, while the error of the improved DRL is reduced to 0.06. The error recovery speed of the improved DRL is 78.57% higher than that of PID. Fig. 8(c) further illustrates the superiority of bias convergence of the improved DRL in three typical disturbance scenarios. This may be because this study introduces action smoothing, SSD guidance and disturbance robust correction mechanisms, which enables the improved DRL to effectively suppress the transient offset caused by sudden disturbances and restore the steady-state in a shorter time. The comparison between the steady-state retention capabilities of the three methods under continuous disturbance and SSD is shown in Fig. 9.

In Fig. 9(a), the SSD of PID continues to accumulate over time, rising from 0.16 at 10 s to 0.31 at 60s, and the SSD of basic DRL remains at 0.13-0.18 in the same range. The SSD of the improved DRL always remains in the low deviation range of 0.08-0.09, with minimal fluctuations. In Fig. 9(b), the wire supply offset results of PID and basic DRL are always higher than those of the improved DRL. At 60 s, the offset results of the improved DRL are 67.65% and 50.00% lower than those of PID and basic DRL. In Fig. 9(c), the improved DRL has the smallest SSD, the narrowest fluctuation range, and the smallest accumulated drift over time under continuous disturbance conditions, and has long-term stability. The robustness of the three methods under compound disturbance and random disturbance scenarios is shown in Fig. 10.

Fig. 10(a) shows the multi-source disturbance RMSE of the three methods. The RMSE of PID under disturbance levels one, two, and three are 0.28, 0.36, and 0.45, and the basic DRL is 0.22, 0.28, and 0.34, while the improved DRL is the lowest, only 0.15, 0.19, and 0.23. Under the random noise condition in Fig. 10(b), the overall change of the improved DRL is the smallest, indicating that its action output is more stable and less sensitive to noise. In the system stability index in

Fig. 10(c), under disturbance level three, the stability indices of PID, basic DRL, and improved DRL are 0.49, 0.68, and 0.82. The improved DRL is 67.35% and 20.59% higher than that of PID and basic DRL. Overall, the improved DRL maintains the lowest error, lowest action variance, and highest stability index under conditions of compound disturbance, strong noise, and high uncertainty, showing optimal robustness and environmental adaptability. To further evaluate the application effect of the proposed method in actual production, the operating status of different control strategies under step changes in yarn supply was compared from the production process. The details are shown in Table 3.



(a) Fast recovery curve after tension pulse perturbation

(b) Error recovery path under step disturbance of wire supply

(c) Response and convergence of high-amplitude impact disturbances in rolling pressure

Fig. 8. Short-term sudden disturbance suppression capabilities of the three methods

As shown in Table 3, compared with PID and the basic DRL, the improved DRL reduces the recovery time by 43.48% and 23.74%, respectively. This indicates that the system can recover to a stable operating state more quickly after a disturbance, thereby reducing adjustment waiting time in the production process and improving equipment utilization efficiency. Simultaneously, the fluctuation amplitude of key variables and SSD significantly decreases, indicating that the proposed method provides a smoother control process, reduces mechanical shock and component wear, decreases maintenance frequency, improves product quality consistency, and reduces the defect rate. In summary, the proposed method not only has significant advantages in control performance but also demonstrates significant engineering decision-making value in production efficiency improvement, equipment operating stability, and quality control.

Table 3. Performance comparison before and after the production process

Metric	PID	Basic DRL	Improve DRL
Recovery time (s)	4.83	3.58	2.73
Fluctuation amplitude	0.25	0.15	0.08
SSD	0.15	0.12	0.05
Stability index	0.49	0.68	0.82

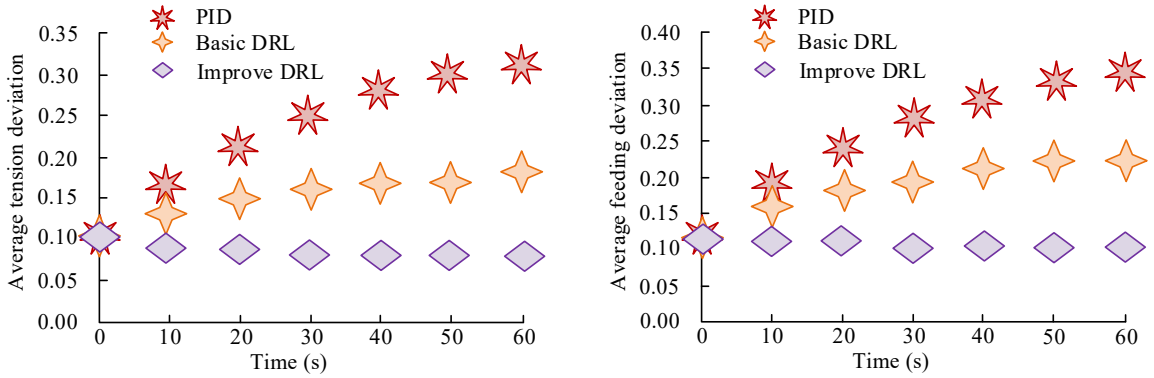
4. Conclusion

To address the challenges of multi-variable coupling and disturbance sensitivity in high-speed cigarette packaging processes, a DRL-based SSQ control method is proposed. By incorporating action smoothing constraints and a SSD guidance mechanism, the proposed method demonstrates clear advantages in dynamic response speed, overshoot suppression, and steady-state accuracy. Experimental results show that under representative disturbance conditions, the recovery time is reduced by more than 40%, and the fluctuation amplitude of key variables is decreased by over 60%,

significantly enhancing control stability and operational reliability.

From the perspective of engineering application and managerial decision-making, the findings indicate a transition from traditional experience-based parameter tuning and manual intervention toward data-driven and adaptive optimization paradigms. With the proposed method, managers can reduce reliance on real-time manual adjustments and shift control strategy selection from empirical settings to policy learning and optimization. In scenarios characterized by frequent disturbances or varying operating conditions, intelligent control methods with adaptive capabilities can be preferentially adopted. Furthermore, the improved system stability enables more flexible production scheduling decisions, such as increasing operating speed or reducing conservative safety margins, thereby improving overall production efficiency and resource utilization.

The current method still relies on the matching degree between the simulation environment and real data, and its generalization ability to extreme working conditions needs to be further verified. Future work will focus on cross-device transfer learning, online retraining mechanisms, and deep integration with digital twins to further improve the generalization, deployment efficiency, and industrial applicability of the strategy



(a) Tension steady-state deviation under slow drift of raw material quality

(b) Wire feed deviation caused by cumulative mechanical wear

(c) Stability trend of splicing pressure under long-term operation

Fig. 9. Steady-state maintenance capability and SSD under continuous perturbation

Funding

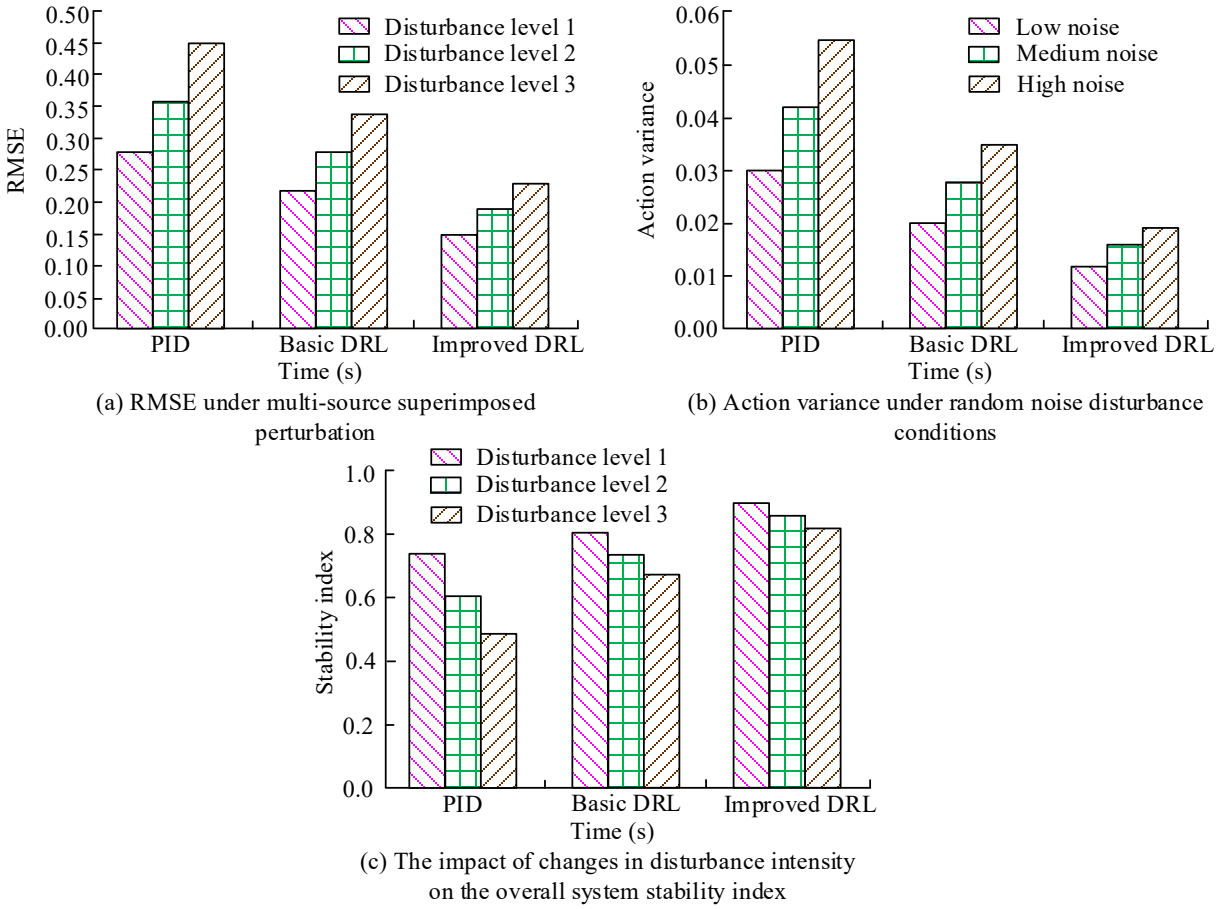
The research is supported by: Science and Technology Project of Hebei China Tobacco Industrial Co., Ltd, Research and Application of Steady-State Control Technology for Process Quality in High-Speed Equipment (ZB416-ZJ119) (NO. HBZY2025A042).

Institutional Review Board Statement

Not applicable.

Declaration of Artificial Intelligence (AI) Tools

The authors used AI tools solely for language editing and readability improvement. The authors reviewed and verified all content and take full responsibility for the accuracy and integrity of the manuscript.



(a) RMSE under multi-source superimposed perturbation

(b) Action variance under random noise disturbance conditions

(c) The impact of changes in disturbance intensity on the overall system stability index

Fig. 10. Robustness verification under combined and random perturbation scenarios

References

Al-Kaf, H. A. G., Lee, K. B., and Blaabjerg, F. (2025). Overview of DSP-based implementation of machine learning methods for power electronics and motor drives. *Journal of Power Electronics*, 25(2), 271-288. https://doi.org/10.1007/s43236-024-00975-2

Bao, Y., Zhu, Y., and Qian, F. (2021). A deep reinforcement learning approach to improve the learning performance in process control. *Industrial & Engineering Chemistry Research*, 60(15), 5504-5515. https://doi.org/10.1021/acs.iecr.0c05678

Boute, R. N., Gijsbrechts, J., Van Jaarsveld, W., and Vanvuchelen, N. (2022). Deep reinforcement learning for inventory control: A roadmap. *European Journal of Operational Research*, 298(2), 401-412. https://doi.org/10.1016/j.ejor.2021.07.016

Cui, J., Qin, Y., Wu, Y., Shao, C., and Yang, H. (2023). Skip connection YOLO architecture for noise barrier defect detection using UAV-based images in high-speed railway. *IEEE Transactions on Intelligent Transportation Systems*, 24(11), 12180-12195. https://doi.org/10.1109/TITS.2023.3292934

De Marco, A., D' Onza, P. M., and Manfredi, S. (2023). A deep reinforcement learning control approach for high-performance aircraft. *Nonlinear Dynamics*, 111(18), 17037-17077. https://doi.org/10.1007/s11071-023-08725-y

Degrave, J., Felici, F., Buchli, J., Neunert, M., Tracey, B., Carpanese, F., and Riedmiller, M. (2022). Magnetic control of tokamak plasmas through deep reinforcement learning. *Nature*, 602(7897), 414-419. https://doi.org/10.1038/s41586-021-04301-9

Giannopoulos, A. A., Keller, L., Sepulcri, D., Boehm, R., Garefa, C., Venugopal, P., and Buechel, R. R. (2023). High-speed on-site deep learning-based FFR-CT algorithm: evaluation using invasive angiography as the reference standard. *American Journal of Roentgenology*, 221(4), 460-470. https://doi.org/10.2214/AJR.23.29156

Groumpos, P. P. (2023). A critical historic overview of artificial intelligence: Issues, challenges, opportunities, and threats. *Artificial Intelligence and Applications*, 1(4), 197-213. https://doi.org/10.47852/bonviewAIA3202689

He, R., Tian, S., Tian, J., Zhang, N., and He, S. (2022). Improvement of cigarette package steering mechanism of YB25 soft packer. *Food and Machinery*, 38(3), 114-117. https://doi.org/10.13652/j.spjx.1003.5788.2022.90043

Jalali Khalil Abadi, Z., Mansouri, N., and Javidi, M. M. (2024). Deep reinforcement learning-based scheduling in distributed systems: a critical review. *Knowledge and Information Systems*, 66(10), 5709-5782. https://doi.org/10.1007/s10115-024-02167-7

Jin, W., Wang, L., and Zhang, C. (2023). Artificial intelligence and big data in the production process to optimise the parameters of the cut tobacco-making process. *International Journal of Information Technology and Management*, 22(3-4), 315-334. https://doi.org/10.1504/IJITM.2023.131823

Kakooee, R. and Dillenburger, B. (2024). Reimagining space layout design through deep reinforcement learning. *Journal of Computational Design and Engineering*, 11(3), 43-55. https://doi.org/10.1093/jcde/qwae025

Li, T., Wang, S., Luo, Y., Wan, J., Luo, Z., and Chen, M. (2024). 3-D vision and intelligent online inspection in SMT microelectronic packaging: A review. *IEEE Journal of Emerging and Selected Topics in Industrial Electronics*, 5(2), 779-789. https://doi.org/10.1109/JESTIE.2024.3365030

Moreno, J. A., Ríos, H., Ovalle, L., and Fridman, L. (2022). Multivariable super-twisting algorithm for systems with uncertain input matrix and perturbations. *IEEE Transactions on Automatic Control*, 67(12), 6716-6722. https://doi.org/10.1109/TAC.2021.3130880

Pattanayak, S. K., Bhoyar, M., and Adimulam, T. (2024). Deep reinforcement learning for complex decision-making tasks. *International Journal of Innovative Research in Science, Engineering and Technology*, 13(11), 18205-18220. https://doi.org/10.15680/IJIRSET.2024.1311017

Seo, J., Kim, S., Jalalvand, A., Conlin, R., Rothstein, A., Abbate, J., and Kolemen, E. (2024). Avoiding fusion plasma tearing instability with deep reinforcement learning. *Nature*, 626(8000), 746-751. https://doi.org/10.1038/s41586-024-07024-9

Suzuki, K., Kumai, S., Saito, Y., and Haga, T. (2005). High-speed twin-roll strip casting of Al-Mg-Si alloys with high iron content. *Materials Transactions*, 46(12), 2602-2608. https://doi.org/10.2320/matertrans.46.2602

Tang, C., Abbatematteo, B., Hu, J., Chandra, R., Martín-Martín, R., and Stone, P. (2025). Deep reinforcement learning for robotics: A survey of real-world successes. *Annual Review of Control, Robotics, and Autonomous Systems*, 8(1), 153-188. https://doi.org/10.1146/annurev-control-030323-022510

Yue, H. (2026). Route optimization for unmanned delivery vehicles using improved Q-learning and Lego-Loma. *Journal of Engineering, Project, and Production Management*, 16(2), 196-2025. https://doi.org/10.32738/JEPPM-2025-196

Zhang, H., Liu, R., Kaushik, A., and Gao, X. (2023). Satellite edge computing with collaborative computation offloading: An intelligent deep deterministic policy gradient approach. *IEEE Internet of Things Journal*, 10(10), 9092-9107. https://doi.org/10.1109/JIOT.2022.3233383

Zhortuylov, O. Z., Rzaliyev, A. S., Karymsakov, T. N., Zhumatay, G. S., Bekenov, U. E., and Zhortuylov, A. O. (2021). Modern harvesting in precision agriculture by wrapping mechanism of haylage rolls' packager with stretch film in a line. *International Journal of Agricultural Resources, Governance and Ecology*, 17(2-4), 381-401. https://doi.org/10.1504/IJARGE.2021.121686

Shuxi Guo works in the process and quality department of Hebei Baisha Tobacco Co., Ltd., serving as Process and Quality Administrator. She is a Certified Six Sigma Black Belt, an Intermediate Lecturer at China Tobacco Hebei Training Center and an assessor for the tobacco inspection vocational qualification examination. She has led multiple Six Sigma projects, been awarded the Demonstration and Professional Prizes by the China Association for Quality (CAQ), published several technical papers and presided over a number of scientific research projects. She has made important contributions to digital quality management and homogenization control of cigarette production. Her research includes cigarette process quality.

Jinghui Guo is the Director of the Process and Quality Department, Hebei Baisha Tobacco Co., Ltd. She specializes in cigarette process quality management. She holds the national qualification in cigarette sensory evaluation and is an expert on the evaluation panel for scientific and technological innovation, as well as for professional technical qualification assessment.

Zhijun Xu is the Chief Inspector of Auxiliary Materials, Process and Quality Department, Hebei Baisha Tobacco Co., Ltd. His research focuses on cigarette processing technology, tobacco equipment and process quality management. He owns a national cigarette sensory evaluation qualification and a Certified Six Sigma Black Belt credential. He is a certified expert in three professional fields: visual inspection of tobacco auxiliary materials, cigarette sensory evaluation and tobacco flavors and essences is a Senior Lecturer at China Tobacco Hebei Training Center and a member of the company’s tender evaluation committee. He has long been committed to technical research and on-site process quality management of tobacco leaf raw materials, cigarette manufacturing processes, tobacco production equipment and tobacco auxiliary materials and flavorings.

Shaolong Han is a senior engineer in the mechanical and electrical technician, silk making workshop department, Hebei Baisha Tobacco Co., Ltd. His responsibilities include Technical research, equipment transformation, and management and scientific and technological innovation. He is a Young Editorial Board Member, a Tobacco Science and Technology Member, a Hebei Artificial Intelligence Association Tobacco Industry Repair Craftsman, an Equipment Management Talent, an Outstanding Skilled Talent at Young Science and Technology Support Talent, an Expert of Science and Technology Committee, Expert of Tobacco Society.