ISSN: 1692-7257 - Volumen 2 – Número 46 - 2025

RCTA
Revista Colombiana de Tecnologías de Avanzada
UNIPAMPLONA

Digital Object Identifier: 10.24054/rcta.v2i46.4141

Control industrial por aprendizaje reforzado

Industrial control through reinforcement learning

Ing. Yessica Cindy Vannesa Mora Cubides[1], Ing. Daniel Steven Arias Otálora[1], PhD. José Antonio Tumialan Borja[2], PhD. Hugo Fernando Velasco Peña[1]

[1] ***Universidad de la Salle,*** *Programa de ingeniería mecatrónica, grupo de automatización de procesos industriales*
[2] ***Universidade São Paulo-USP,*** *Escola de Engenharia de São Carlos, Laboratório de Escoamentos Multifásicos Industriais.*

Correspondencia: *jtumialan@usp.br*

Recibido: *24 abril 2025.* ***Aceptado:*** *03 julio 2025.* ***Publicado:*** *09 agosto 2025.*

Cómo citar: *Y. C. Mora Cubides, D. S. Arias Otálora, J. A. Tumialan Borja, y H. F. Velasco Peña, «Control industrial por aprendizaje reforzado», RCTA, vol. 2, n.º 46, pp. 201–208, ago. 2025.*
Recuperado de https://ojs.unipamplona.edu.co/index.php/rcta/article/view/4141

Resumen: En este artículo se presenta la implementación de la técnica de control de caudal con inteligencia artificial (DDPG) en un prototipo funcional completamente instrumentado con sensores y actuadores industriales, simulando la recirculación de caudal a través de tres tanques. Se presenta la metodología utilizada para la identificación del proceso (modelo de primer orden más tiempo muerto (FOPDT)) mediante comunicación OPC Cliente Servidor con Matlab®. También se presenta el diseño del algoritmo de aprendizaje por refuerzo y su adaptación en el entorno de aprendizaje con datos experimentales. Los resultados de la simulación fueron satisfactorios en comparación con las técnicas de control tradicionales, demostrando robustez frente a perturbaciones forzadas. Finalmente, se evaluó la implementación del control de aprendizaje reforzado integrando TIA Portal y Matlab (a través de un controlador PLC-S7-1500) con una referencia de 600 l/h, logrando un sobre impulso del 0% con un tiempo de asentamiento de 22s. Comparado con otros sistemas de control, se observó una mejor respuesta en el tiempo de asentamiento y un control libre de sobre impulso. Finalmente, se aplicaron perturbaciones al sistema, observando su efecto con relación al flujo.

Palabras clave: aprendizaje reforzado, eficiencia, inteligencia artificial, proceso industrial.

Abstract: In this article, the implementation of the flow control technique with artificial intelligence (DDPG) is presented in a fully instrumented functional prototype with industrial sensors and actuators, simulating flow recirculation through three tanks. The methodology used for process identification (first-order plus dead time model (FOPDT)) through ClientServer OPC communication with Matlab® is presented. The design of the reinforcement learning algorithm and its adaptation in the learning environment with experimental data are also presented. The simulation results were satisfactory compared to traditional control techniques, demonstrating robustness against forced disturbances. Finally, the implementation of reinforced learning control integrating TIA Portal and Matlab (trough a PLC-S7-1500 controller) was evaluated with a reference of 600 l/h, achieving 0% overshoot with a settling time of 22s. Compared to other control systems, a better response in settling time and overshoot-free control was observed. Finally, perturbations were applied to the system, observing their effect in relation to the flow.

Keywords: reinforcement learning, efficiency, artificial intelligence, industrial processes.

1. INTRODUCCIÓN

El aprendizaje por refuerzo es una técnica en la que un agente aprende a realizar una tarea mediante interacciones repetidas de ensayo y error en un entorno dinámico. La esencia de esta técnica consiste en incorporar comportamientos mediante la interacción con el entorno sin una programación explícita de la solución del problema [1]. Basado en este concepto se han generado múltiples usos de estos entornos de aprendizaje, como la jugabilidad o el desarrollo de controles para robótica. [2], e incluso la optimización energética en los edificios [3].

En general, existen diferentes técnicas de inteligencia artificial que se encargan de resolver tareas complejas basadas en información sensorial inexplorada y de alta dimensión, lo que la convierte en una herramienta poderosa para el desarrollo de tareas complejas. A diferencia de otras ramas de la IA, el aprendizaje por refuerzo recibe retroalimentación retardada, donde el agente recibe retroalimentación tras generar una decisión y una predicción [3]. El aprendizaje por refuerzo ha surgido como un enfoque muy poderoso para la toma de decisiones automatizada en múltiples campos de sistemas de control. [4], representando una metodología de largo alcance y prometedora.

En el campo de control de procesos, [5] Se presenta una comparación entre dos técnicas de ajuste de controladores Proporcional-Integrales (PI): el método tradicional de cancelación de polo cero y una estrategia innovadora basada en aprendizaje por refuerzo para el ajuste adaptativo de un controlador PI en un sistema de refrigeración (HVAC). Los resultados demostraron que el método innovador permite optimizar el consumo energético y reducir los costos operativos. Para abordar este problema de forma más eficiente, se están explorando nuevas técnicas, incluyendo aquellas que aprovechan la Inteligencia Artificial (IA) y la optimización. Estas estrategias adoptan enfoques dinámicos y continuos utilizando herramientas como las redes neuronales. [6]-[10], algoritmos genéticos [11]-[13], controladores difusos [14][15], y otros métodos de optimización [16][17]. El objetivo principal de aplicar técnicas de ajuste basadas en IA es automatizar y refinar los parámetros característicos de un controlador PID, logrando así un mejor rendimiento que los métodos tradicionales.

Si bien el control PID sigue siendo ampliamente utilizado debido a su simplicidad, enfoques como el aprendizaje por refuerzo no se limitan a un único modelo lineal del sistema. Estas técnicas avanzadas pueden aplicarse a escenarios más complejos, como lo demuestra [18]. Además, estas estrategias permiten incorporar las preferencias del usuario en la toma de decisiones [19], la optimización energética [20] y la detección de fallas [21].

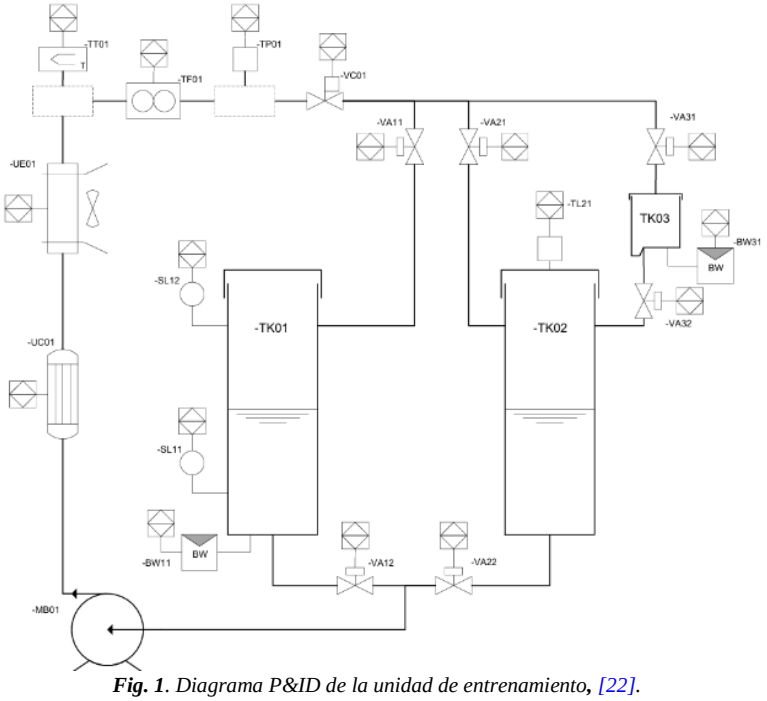

***Fig. 1**. Diagrama P&ID de la unidad de entrenamiento, [22].*

Este artículo se centra en el desarrollo de un sistema de control de aprendizaje por refuerzo para una planta de capacitación en procesos industriales, utilizando un entorno de aprendizaje con un modelo matemático de la planta. El objetivo es controlar un caudal constante en un sistema de tres tanques interconectados, modificando las frecuencias de una bomba controlada hidráulicamente mediante un convertidor de frecuencia y un PLC Siemens S7-1500 [22].

El tipo de aprendizaje por refuerzo aplicado es DDPG (Gradiente de Política Determinista Profunda), un algoritmo crítico-actor con una política que maximiza la recompensa a largo plazo [23]. Durante la etapa de entrenamiento del agente, se realizan los siguientes pasos:

En la ecuación (1), se inicializan los parámetros ϕ, se toma la observación S del crítico y se ejecuta la acción A.

$$Q(S, A; \phi) \quad (1)$$

Inicialmente, el actor (2) toma la observación S y devuelve la acción que maximiza la recompensa a largo plazo.

$$\pi\ (S; \phi) \quad (2)$$

Para la observación actual S (3), se selecciona la acción A, donde *N* representa el ruido modelado.

$$A = \pi(S; \phi) + N \quad (3)$$

Se ejecuta la acción A y se calcula la recompensa R y la siguiente observación S′

La información (S, A, R, S') es almacenada en el Buffer de experiencia

Se genera aleatoriamente un pequeño lote de experiencias M a partir del búfer de experiencias (S_i, A_i, R_i, S')

Si S_i s un estado terminal, el valor objetivo y_i para la función de valor se establece en R_i de lo contrario, se establece de acuerdo con la ecuación (4)

$$y_i = R_i + \gamma Q_t + (S_i', \pi_t(S_i'; \theta_t); \phi_t \quad (4)$$

El objetivo de la función de valor es sumar la recompensa inmediata R_i Con la recompensa futura descontada. Para calcular la recompensa acumulada, el agente primero calcula la siguiente acción y la siguiente observación S'_i A partir de las muestras de experiencia que utilizan el actor objetivo, el agente estima la recompensa acumulada utilizando la siguiente acción para el crítico objetivo. Los parámetros del crítico se actualizan minimizando la pérdida *L* en todas las muestras de experiencia (5).

$$L = \frac{1}{2M}\sum_{i=1}^{M}(y_i - Q(S_i, A_i;)\phi)^2 \quad (5)$$

Los parámetros del actor se actualizan utilizando el gradiente de política, con el objetivo de maximizar la recompensa descontada esperada (6)

$$\nabla_\theta J \approx \frac{1}{M}\sum_{i=1}^{M} G_{ai} + G_{ai} \quad (6)$$
$$G_{ai} = \nabla_A Q'(S_i, A; \phi) where A = \pi(S_i; \theta)$$
$$G_{\pi i} = \nabla_\theta \pi(S_i; \theta)$$

Para actualizar los valores de los parámetros del crítico y del actor, se utilizan factores de suavizado τ (7).

$$\phi_t = \tau\phi + (1-\tau)\phi_t\ (Parametros\ cirticos\ del\ objetivo)$$
$$\phi_t = \tau\phi + (1-\tau)\theta_t\ (Parametrosdel\ actor\ objetivo) \quad (7)$$

2. METODOLOGÍA DE TRABAJO

2.1 Comunicación por OPC

A través de la arquitectura OPC y el servidor KepServerEX, se estableció una comunicación cliente-servidor entre el PLC Siemens S7-1500 y la plataforma MATLAB®, lo que permitió la adquisición, supervisión y control de variables del proceso en tiempo real, como se muestra en la Figura 2. Esta integración facilitó el registro del comportamiento dinámico del caudal ante variaciones en la frecuencia de operación de la bomba. A partir de los datos experimentales obtenidos, se procedió a la identificación del modelo dinámico de la planta, el cual fue representado mediante un modelo de primer orden con tiempo muerto (FOPDT), comúnmente empleado en sistemas de control industrial por su simplicidad y capacidad de aproximación de procesos reales.

***Fig. 2**. Diagrama de conexión por KEP Server*

2.2 Modelo matemático

Utilizando la respuesta mostrada en la Figura 3, se obtuvieron datos de la respuesta al escalón (Orden

al variador: 46.875 Hz) y la curva de respuesta, Figura 4, para identificar el modelo matemático propuesto por Alfaro [24].

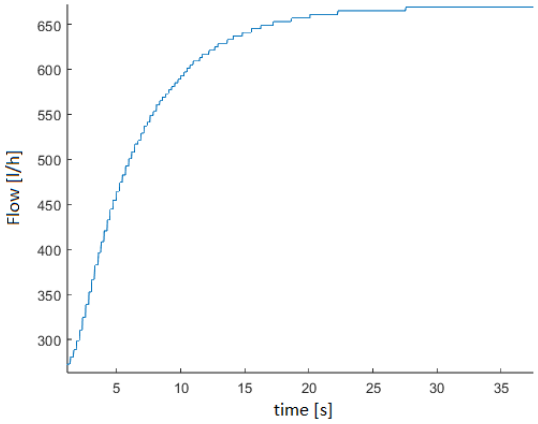

***Fig. 3**. Diagrama de curva de proceso*

La función de transferencia obtenida se presenta en la ecuación (8).

$$G_{p1}(S) = \frac{0.02575e^{-1.6s}}{5.17s+1} \qquad (8)$$

La Figura 4 compara la respuesta temporal de la planta con la función de transferencia dada en la ecuación (8). El error promedio de los modelos obtenidos fue del 0,30 %.

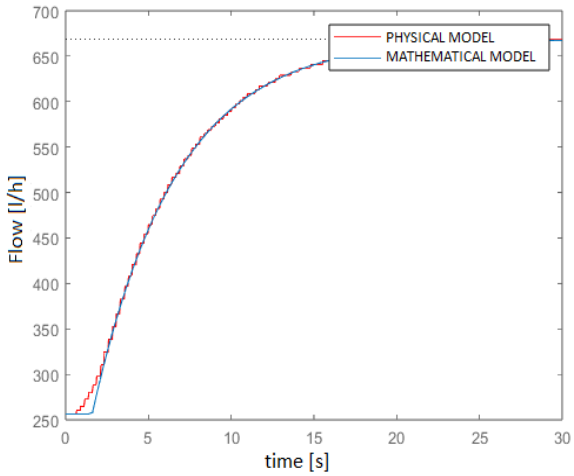

***Fig. 4**. Modelo físico vs modelo matemático*

2.3 Tabla de recompensa del ambiente de aprendizaje

Los rangos y recompensas proporcionados por el ambiente se presentan en la Tabla I. El rango de variación del caudal a través de los tanques, 200-700 l/h, se utilizó como condición de terminación de la planta.

2.4 Diagrama general del entorno de aprendizaje

Simulink® Se utilizó para representar el modelo a través de diagramas de bloques, como se muestra en la Figura 5. El integrador H representa la condición inicial de la frecuencia de la bomba hidráulica.

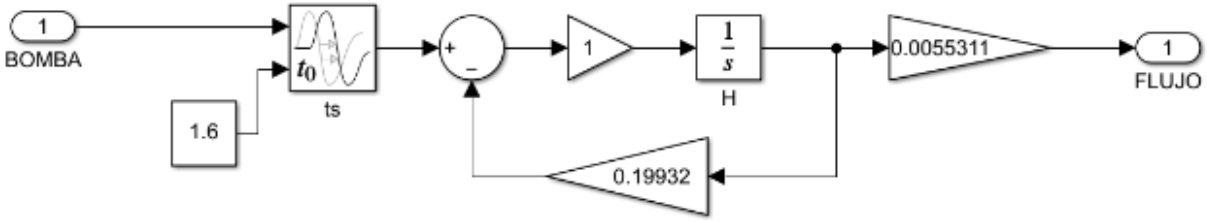

***Fig. 5**. Bloque de diagrama del modelo matemático*

2.5 Diagrama General del Entorno de Aprendizaje

El entorno de aprendizaje se construyó utilizando un agente RL, implementando la tabla de recompensas de la Tabla 1 (Recompensa calculada), el entorno empleado (modelo matemático) y las condiciones de terminación como los límites físicos de la planta y el valor de referencia de flujo (Ver Figura 6).

***Tabla 1:** recompensas del entorno de aprendizaje*

Recompensa	Estado
+10	Por mantener el valor dentro de una variación menor a 1 l/h
-1	Por mantener el valor del caudal dentro de los límites del sistema
-650	Por exceder de los límites de flujo

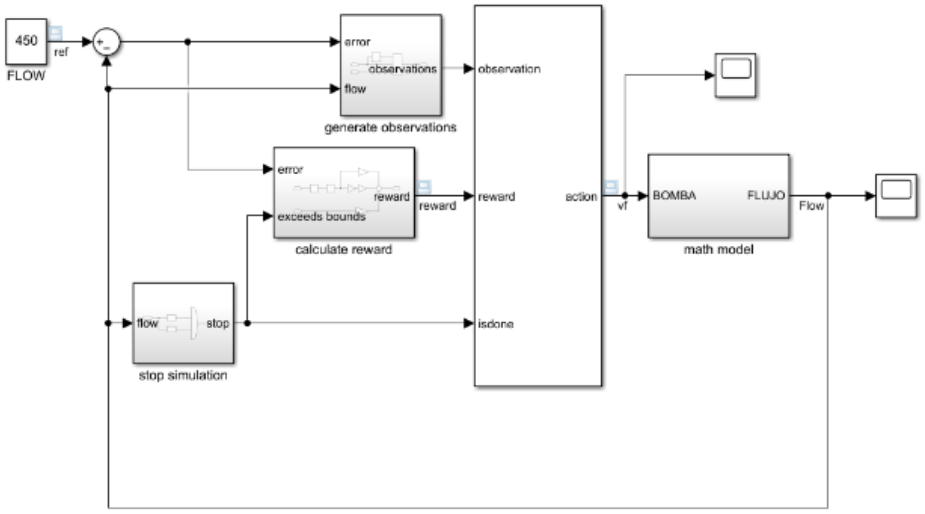

***Fig. 6**. Diagrama de bloques del entorno de aprendizaje*

2.6 Programación del entorno de aprendizaje

El Algoritmo 1 y la Tabla 2 presenta las diferentes características utilizadas en el entorno de aprendizaje. Con base en ella, se desarrolló el modelo de aprendizaje.

Los resultados del entrenamiento se presentan gráficamente en la Figura 7, con un total de 880 épocas de entrenamiento, una recompensa superior a 600 y un tiempo de entrenamiento de 2 horas y 12 minutos.

Algorithm 1 Learning Environment

$RLAgent \leftarrow GetAgentCharacteristics$
$RLAgent \leftarrow InitializeLearningEnvironment$
$InitializeParameters \leftarrow (Flow = 0, Pump = 0)$
while $Flow < 200 || Flow > 700$ **do**
  $Rand \leftarrow GenerateRandomNumbersBetween(A, B)$
  $Flow \leftarrow Rand(200, 700)$
end while
while $Pump < 7000 || Pump > 25000$ **do**
  $Rand \leftarrow GenerateRandomNumbersBetween(A, B)$
  $Pump \leftarrow Rand(7000, 25000)$
end while
$RLAgent \leftarrow TrainAgent$ =0

***Tabla 2:** características del entorno de aprendizaje*

Criterio	Valor
Aproximadores utilizados por la política	Crítico, Actor
Tiempo de muestreo	1s
Tiempo de simulación	200s
Numero de neuronas del critico	25
Numero de neuronas del actor	25
Tipo de agente usado para	Deep Deterministic Policy Gradient (DDPG)
Factor del objetivo	1.00E-03
Factor de aprendizaje del crítico	1.00E-03
Factor de aprendizaje del actor	1.00E-04
Umbral del gradiente (Critico, Actor)	1.0
Buffer de experiencia del agente	1.00E+06
Tamaño de lote del agente	6.40

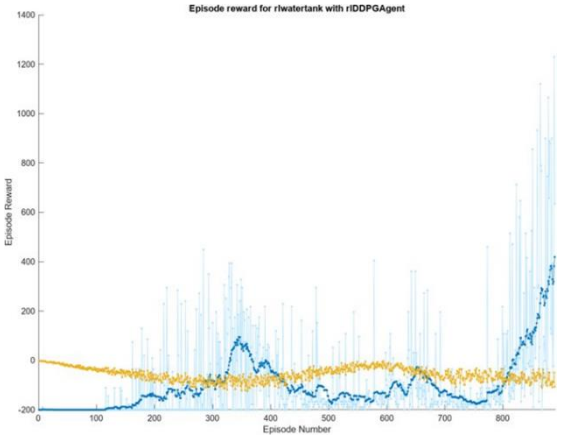

***Fig. 7**. Diagrama de recompensa por episodio (Recompensa por episodio en azul y Valor del Crítico en amarillo)*

3. IMPLEMENTACIÓN PROTOTIPO

La comunicación entre el agente entrenado y la planta física (PLC Siemens S7-1500) se estableció a través de KepServer (Figura 8).

3.1. Análisis del Control por Aprendizaje Reforzado

Se analizó la respuesta del controlador para dos puntos de referencia (350 y 600) l/h, y se evaluó su comportamiento frente a perturbaciones forzadas artificialmente (apertura y cierre de las electroválvulas de la planta). generando variaciones abruptas en el flujo del sistema con el fin de evaluar

la robustez del controlador ante cambios no programados en las condiciones operativas, en donde se obtuvo respuestas satisfactorias, como se muestra en la Figura 9.

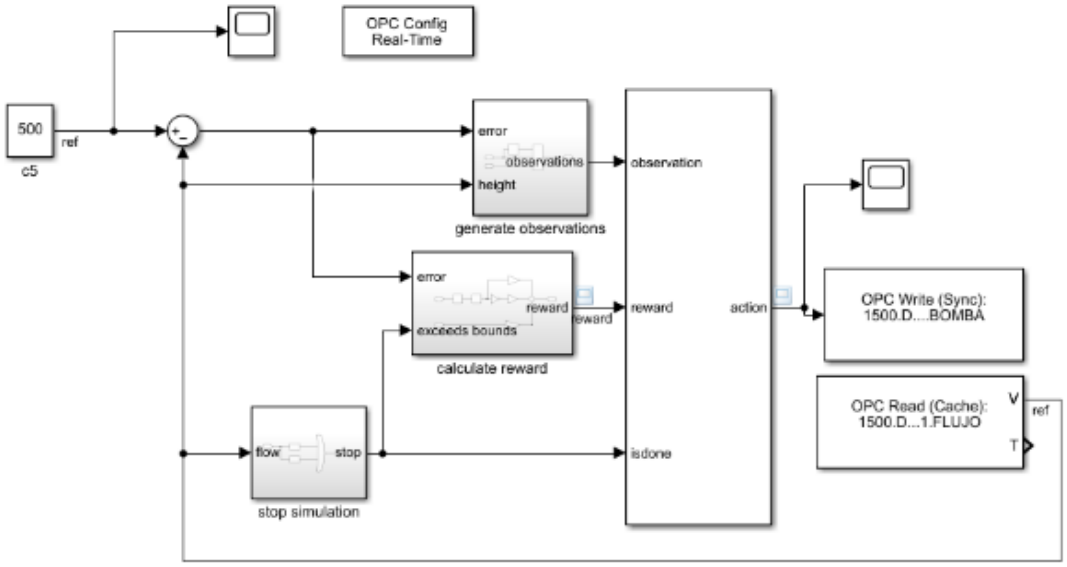

***Fig. 8**. Conexión de bloques por OPC*

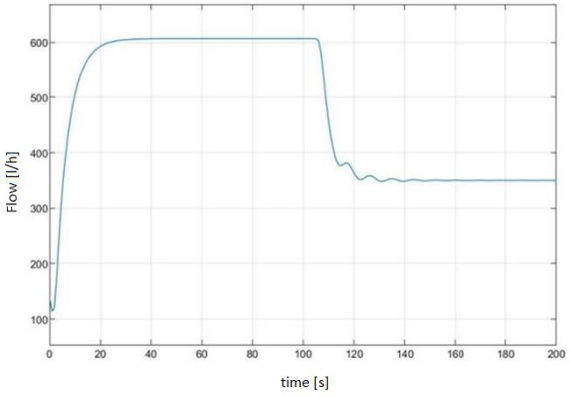

***Fig. 9**. Respuesta de la planta al sistema de control propuesto con dos valores de referencia*

A partir de la Figura 9, se calcularon el tiempo de asentamiento y el sobre impulso para una referencia de 600 l/h, con un resultado de 22 segundos y 0 % respectivamente. Para una referencia de 350 l/h, se obtuvo un tiempo de asentamiento de 19 segundos y un sobre impulso de 0 %.

3.2. Comparación con otras técnicas de control

El control de aprendizaje de refuerzo se comparó con otras técnicas de control bajo las mismas condiciones de operación (control PI, control en cascada, predictor Smith) en la Tabla 3, utilizando el mismo valor de referencia de 600 l/h.

***Table 3:** comparación con otros controladores sobre la planta física*

Tipo de control	Tiempo de establecimiento	Sobre impulso
Control PI	35s	15%
Control por Cascada	32s	0%
Predictor de Smith	25s	5%
Aprendizaje reforzado	22s	0%

De la figura 3, se observa que el control por aprendizaje de refuerzo no presenta sobre impulso en el flanco de subida y presenta una pequeña oscilación sin sobre impulso en el flanco de bajada, como las estructuras de control clásicas. El predictor de Smith muestra un sobre impulso del 5% con un tiempo de asentamiento de 25 segundos; el control PI presenta un sobre impulso mayor, del 15%, con un tiempo de asentamiento de 35 segundos; y el control en cascada muestra un comportamiento similar al del aprendizaje de refuerzo, sin sobre impulso, pero con un tiempo de asentamiento 10 segundos mayor. Con base en este análisis, se validó que el control por aprendizaje de refuerzo genera una mejor respuesta en términos de velocidad y sobre impulso.

4. CONCLUSIONES

La implementación del controlador de aprendizaje por refuerzo mostró resultados satisfactorios en un

prototipo funcional. Además, demostró robustez ante perturbaciones forzadas, manteniendo un punto de referencia constante con un error de estado estacionario del 0,2 %.

En comparación con otros sistemas de control clásicos, el controlador de aprendizaje por refuerzo presentó un tiempo de asentamiento más corto, de 22 segundos, sin sobre impulsos. Similar al comportamiento del control en cascada, pero mejorando el tiempo de asentamiento.

REFERENCES

[1] Kaelbling, Leslie Pack; Littman, Michael L.; MOORE, Andrew W. Reinforcement learning: A survey. Journal of artificial intelligence research, 1996, vol. 4, p. 237-285

[2] Carlos G´omez, J., Verrastro, C. A. (s/f). Aprendizaje por refuerzo y control difuso para generar comportamiento de robots. Edu.ar. from 2022, de https: //www.f rba.utn.edu.ar/wp content/uploads/2021/02/jar8submission30.pdf

[3] Wang, Zhe; Hong, Tianzhen. Reinforcement learning for buildingcontrols: The opportunities and challenges. Applied Energy, 2020, vol.269, p. 115036.

[4] Damjanovi ´C, Ivana, et al. High Performance Computing Reinforcement Learning Framework for Power System Control. En 2023 IEEEPower Energy Society Innovative Smart Grid Technologies Conference (ISGT). IEEE, 2023. p. 1-5.

[5] V. Abad-Alcaraz, M. Castilla, J.D. Álvarez (2024) A Comparison of Classical and Reinforcement Learning-based Tuning Techniques for PI controllers, IFAC-papers OnLine Volume 58, Issue 7, 2024, Pages 180-185. doi: 10.1016/j.ifacol.2024.08.031.

[6] Gunther, J., Reichensd¨orfer, E., Pilarski, P.M., and Diepold, K. (2020). Interpretable PID parameter tuning for control engineering using general dynamic neural networks: An extensive comparison. Plos One, 15(12), e0243320.

[7] Ali, M., Firdaus, A.A., Arof, H., Nurohmah, H., Suyono, H., Putra, D.F.U., and Muslim, M.A. (2021). The comparison of dual axis photovoltaic tracking system using artificial intelligence techniques. IAES Int. J. Artif. Intell, 10(4), 901.

[8] Daye Yang, Jingcheng Wang, Huihuang Cai, Jun Rao, Chengtian Cui, (2025) Intelligent control strategy for electrified pressure-swing distillation processes using artificial neural networks-based composition controllers, Separation and Purification Technology, Volume 360, Part 2, 2025, 130991, ISSN 1383-5866.

[9] Ibtihaj Khurram Faridi, Evangelos Tsotsas, Wolfram Heineken, Marcus Koegler, Abdolreza Kharaghani, Development of a neural network model predictive controller for the fluidized bed biomass gasification process, Chemical Engineering Science, Volume 293, 2024, 120000, ISSN 0009-2509.

[10] Hong-Gui Han, Lu Zhang, Hong-Xu Liu, Jun-Fei Qiao, Multiobjective design of fuzzy neural network controller for wastewater treatment process, Applied Soft Computing, Volume 67, 2018, Pages 467-478, ISSN 1568-4946.

[11] P. Siva Krishna, P.V. Gopi Krishna Rao, Fractional-order PID controller for blood pressure regulation using genetic algorithm, Biomedical Signal Processing and Control, Volume 88, Part B, 2024, 105564, ISSN 1746-8094.

[12] Marco Paz Ramos, Axel Busboom, Andriy Slobodyan, Genetic PI Controller Tuning to Emulate a Pole Assignment Design∗∗This work was supported by the Bavarian Ministry of Economic Affairs, Regional Development and Energy (StMWi) under Grant DIK0397/03., IFAC-PapersOnLine, Volume 58, Issue 7, 2024, Pages 138-143, ISSN 2405-8963.

[13] K.G Aparna, R. Swarnalatha, Dynamic optimization of a wastewater treatment process for sustainable operation using multi-objective genetic algorithm and non-dominated sorting cuckoo search algorithm, Journal of Water Process Engineering, Volume 53, 2023, 103775, ISSN 2214-7144

[14] Himanshukumar R. Patel, Optimal intelligent fuzzy TID controller for an uncertain level process with actuator and system faults: Population-based metaheuristic approach, Franklin Open, Volume 4, 2023, 100038, ISSN 2773-1863.

[15] Slavica Prvulovic, Predrag Mosorinski, Dragica Radosav, Jasna Tolmac, Milica Josimovic, Vladimir Sinik, Determination of the temperature in the cutting zone while processing machine plastic using fuzzy-logic controller (FLC), Ain Shams Engineering Journal, Volume 13, Issue 3,2022, 101624, ISSN 2090-4479.

[16] Rios H. R, Apráez B. A., Rodriguez C. J and Tumialan. B. J. A., PI tuning based on Bacterial Foraging Algorithm for flow control, (2020) IX International Congress of Mechatronics Engineering and Automation (CIIMA), Cartagena, Colombia, 2020, pp. 1-6, doi: 10.1109/CIIMA50553.2020.9290289.

[17] Ali, M., Firdaus, A.A., Arof, H., Nurohmah, H., Suyono, H., Putra, D.F.U., and Muslim, M.A. (2021). The comparison of dual axis photovoltaic tracking system using artificial intelligence techniques. IAES Int. J. Artif. Intell, 10(4), 901.

[18] Shuprajhaa, T., Sujit, S.K., and Srinivasan, K. (2022). Reinforcement learning based adaptive pid controller design for control of linear/nonlinear unstable processes. Applied Soft Computing, 128, 109450.

[19] Lei, Y., Zhan, S., Ono, E., Peng, Y., Zhang, Z., Hasama, T., and Chong, A. (2022). A practical deep reinforcement learning framework for multivariate occupant centric control in buildings. Applied Energy, 324, 119742.

[20] Yu, L., Xie, W., Xie, D., Zou, Y., Zhang, D., Sun, Z., Zhang, L., Zhang, Y., and Jiang, T. (2019). Deep reinforcement learning for smart home energy management. IEEE Internet of Things Journal, 7(4), 2751–2762.

[21] Matetic, I., Stajduhar, I., Wolf, I., and Ljubic, S. (2023). Improving the efficiency of fan coil units in hotel buildings through deep-learning-based fault detection. Sensors, 23(15), 6717.

[22] Andina de Integración Tecnológica (ANITCO), Manual de operación y mantenimiento unidad de entrenamiento en automatización, 1ra ed. Bogotá, Colombia: ANITCO, 2016.

[23] Mathworks, Reinforcement Learning Toolbox™ User's Guide, Matlab and Simulink, MathWorks, Natick, MA, USA, 2020. [En Línea] Disponible en: https://www.mathworks.com/help/reinforcement-learning/

[24] Alfaro, Víctor M. Método de identificación de modelos de orden reducido de tres puntos 123c. Escuela de Ingeniería Eléctrica Universidad de Costa Rica, 2007, p. 1-7