

Accepted Manuscript / Manuscrito Aceptado

Title Paper/Título del artículo:

Plataforma de Control de Insulina en Tiempo Real y de Bajo Costo para Diabetes Mellitus Tipo 1 Usando Deep Q-Learning

Real-Time, Low-Cost Insulin Control Platform for Type 1 Diabetes Mellitus Using Deep Q-Learning

Authors/Autores: Vidrios-Serrano, C., Navarrete Guzmán, A., García-Rodríguez, J. A., Velarde Alvarado, P.

ID: e1832

DOI: <https://doi.org/10.15741/revbio.13.e1832>

Received/Fecha de recepción: November 08th 2024

Accepted /Fecha de aceptación: June 04th 2026

Available online/Fecha de publicación: August 21th 2026

Please cite this article as/Como citar este artículo: Vidrios-Serrano, C., Navarrete Guzmán, A., García-Rodríguez, J. A., Velarde Alvarado, P. (2026). Real-Time, Low-Cost Insulin Control Platform for Type 1 Diabetes Mellitus Using Deep Q-Learning. *Revista Bio Ciencias*, 13, e1832. <https://doi.org/10.15741/revbio.13.e1832>

This is a PDF file of an unedited manuscript that has been accepted for publication. As a service to our customers we are providing this early version of the manuscript. The manuscript will undergo copyediting, typesetting, and review of the resulting proof before it is published in its final form. Please note that during the production process errors may be discovered which could affect the content, and all legal disclaimers that apply to the journal pertain.

Este archivo PDF es un manuscrito no editado que ha sido aceptado para publicación. Esto es parte de un servicio de Revista Bio Ciencias para proveer a los autores de una versión rápida del manuscrito. Sin embargo, el manuscrito ingresará a proceso de edición y corrección de estilo antes de publicar la versión final. Por favor note que la versión actual puede contener errores de forma.

Artículo original

Plataforma de Control de Insulina en Tiempo Real y de Bajo Costo para Diabetes Mellitus Tipo 1 Usando Deep Q-Learning

Real-Time, Low-Cost Insulin Control Platform for Type 1 Diabetes Mellitus Using Deep Q-Learning

Control de Glucosa con Deep Q-Learning/

Glucose Control Using Deep Q-Learning

Vidrios-Serrano, C.¹ (<https://orcid.org/0000-0003-4963-4349>), Navarrete Guzmán, A.^{2*} (<https://orcid.org/0000-0003-3084-7707>), García-Rodríguez, J. A.³ (<https://orcid.org/0000-0003-0339-0545>), Velarde Alvarado, P.⁴ (<https://orcid.org/0000-0002-1211-1061>)

¹Unidad académica de Ciencias Básicas e Ingenierías. Universidad Autónoma de Nayarit. Ciudad de la cultura Amado Nervo S/N. C.P 63000, Tepic, Nayarit, México.

²Unidad académica de Ciencias Básicas e Ingenierías. Universidad Autónoma de Nayarit. Ciudad de la cultura Amado Nervo S/N. C.P 63000, Tepic, Nayarit, México.

³Departamento de Eléctrica y Electrónica. Tecnológico Nacional de México/Instituto Tecnológico de Ciudad Guzmán. Av. Tecnológico 100. C.P 49100, Zapotlán el Grande, Jalisco, México.

⁴Unidad académica de Ciencias Básicas e Ingenierías. Universidad Autónoma de Nayarit. Ciudad de la cultura Amado Nervo S/N. C.P 63000, Tepic, Nayarit, México.

*Corresponding Author:

Carlos Vidrios-Serrano. Unidad Académica de Ciencias Básicas e Ingenierías. Universidad Autónoma de Nayarit. Ciudad de la cultura Amado Nervo S/N, C.P. 63000, Tepic, Nayarit, México. Teléfono (311) 2118800 Ext. 8757. E-mail: carlos.vidrios@uan.edu.mx

RESUMEN

El objetivo de esta investigación es desarrollar una plataforma que controle la cantidad de insulina necesaria para mantener los niveles de glucosa en pacientes con diabetes mellitus tipo 1. El estudio se basa en la implementación del modelo matemático de Sorensen y un controlador basado en aprendizaje por refuerzo llamado (Reinforcement Learning-Based Controller, RLBC) Deep Q-Learning en una tarjeta de desarrollo Raspberry Pi 2B, con una ejecución en tiempo real. Los resultados obtenidos son comparados cuantitativamente con un controlador robusto basado en μ -síntesis, concluyendo que el controlador basado en aprendizaje por refuerzo presenta un mejor desempeño. La implicación principal de esta investigación es la posibilidad de mejorar el control de la glucosa en pacientes con diabetes mellitus tipo 1 a través de un sistema automatizado y de bajo costo. La contribución original de este estudio radica en la integración de un modelo

matemático complejo con hardware accesible para el control de la diabetes. En conclusión, este trabajo sienta las bases para el desarrollo de dispositivos de control de insulina más eficientes y accesibles, aunque se requiere investigación adicional para validar su efectividad en entornos clínicos.

PALABRAS CLAVE:

Aprendizaje por refuerzo, Diabetes, Sorensen.

ABSTRACT

The objective of this research is to develop a platform that controls the insulin dose required to maintain glucose levels in patients with type 1 diabetes mellitus. The study is based on the implementation of the Sorensen mathematical model and a Reinforcement Learning-Based Controller (RLBC) named Deep Q-Learning on a Raspberry Pi 2B development board, with real-time execution. The obtained results are quantitatively compared with a μ -synthesis-based robust controller, concluding that the RLBC exhibits superior performance. The main implication of this research is the potential to improve glucose control in type 1 diabetes mellitus patients through an automated and low-cost system. The original contribution of this study lies in the integration of a complex mathematical model with accessible hardware for diabetes management. In conclusion, this work provides a foundation for the development of more efficient and accessible insulin control devices, although further research is required to validate its effectiveness in clinical settings.

KEY WORDS:

Reinforcement Learning-Based Controller, Diabetes, Sorensen.

Introducción

La diabetes mellitus tipo 1 (*Type 1 Diabetes Mellitus*, T1DM) es una enfermedad crónica que se manifiesta como la incapacidad del páncreas para producir insulina, provocando con esto la elevación de niveles de glucosa en la sangre, que con el tiempo puede desencadenar daños graves en órganos y sistemas del cuerpo humano (Katsarou *et al.*, 2017; Syed, 2022). Debido a que el páncreas no produce la insulina necesaria, los niveles de la glucosa en sangre para las personas que padecen diabetes mellitus tipo 1 dependen de la administración externa de insulina, control dietético, actividad física, estrés psicológico y enfermedades secundarias. La diabetes mellitus es una enfermedad cuya incidencia va en aumento, según la Federación Internacional de Diabetes (*International Diabetes Federation*, IDF) (Magliano & Boyko, 2022) en 2019 había 463 millones de personas que vivían con diabetes a nivel mundial y se proyecta que para 2045 la cifra

alcance los 700 millones. En 2020 fallecieron más de 151 mil personas en México por causas asociadas con la diabetes, lo que representa aproximadamente un 14 % del total de las defunciones ocurridas en este año. El costo económico asociado a la T1DM es difícil de cuantificar, pero la IDF estima que para el año 2030 la cifra a nivel mundial podría alcanzar los 850 mil millones de dólares. Estos datos reflejan que la diabetes mellitus tipo 1 no solo afecta a nivel individual, sino que representa un problema de salud pública global que requiere urgentemente soluciones eficaces y accesibles para mejorar la calidad de vida de los pacientes y reducir los costos sanitarios asociados.

Debido a la complejidad que implica para un paciente saber la cantidad exacta de insulina que se debe suministrar para el control adecuado de su enfermedad, se han intentado crear diversos sistemas que ayuden a realizar esta tarea de forma automática y precisa, estos dispositivos son incluyen monitores continuos de glucosa (*Continuous Glucose Monitors*, CGM) y bombas de insulina. Para el correcto funcionamiento de estos sistemas se han propuesto diversos algoritmos de control que van desde el control PID o técnicas como el Model Predictive Control (Sánchez-Peña *et al.*, 2017; Fox & Wiens, 2019), hasta sistemas de control robusto como los propuestos en (Ramprasad *et al.*, 2004; Kovács *et al.*, 2009; Ruiz-Velázquez *et al.*, 2020; Parker *et al.*, 2000). Un área de gran interés para el control de las bombas de insulina son los sistemas basados en Aprendizaje por Refuerzo (*Reinforcement Learning*, RL), la razón principal de este interés es debido a que los agentes de RL han demostrado tener un buen desempeño en sistemas altamente no lineales, como el sistema glucosa-insulina del cuerpo humano, además de que estos agentes no requieren de conocer la dinámica del sistema a controlar. Ejemplos de aplicación de agentes de RL en el control de la glucosa se presentan en (Dénes-Fazakas *et al.*, 2022; Wang *et al.*, 2021; Viroonluecha *et al.*, 2022).

Este trabajo tiene como objetivo determinar la viabilidad de implementar en una tarjeta Raspberry Pi un sistema de control de glucosa basado en aprendizaje por refuerzo, que en una primera etapa se implemente sobre un modelo matemático para evaluar su correcto funcionamiento, sentando las bases para que en un trabajo futuro esta plataforma pueda servir como medio de control portátil para de forma personalizada y automatizada subministre la cantidad precisa de insulina que permita mantener los niveles óptimos de glucosa de pacientes con T1DM, los resultados obtenidos son comparados con los mostrados por el controlador robusto basado en μ -síntesis (Navarrete *et al.*, 2024) mostrando resultados prometedores.

Material y Métodos

Implementación en hardware

Como se mencionó anteriormente, la intención del presente proyecto es sentar las

bases para el desarrollo de una plataforma que pueda controlar la cantidad de insulina necesaria para controlar los niveles de glucosa de un paciente con T1DM, por lo tanto, el modelo matemático del paciente y el controlador se implementaron en una tarjeta de desarrollo Raspberry Pi 2B, que es una computadora de bajo costo y tamaño reducido con las siguientes características: procesador ARM Cortex de cuatro núcleos a 900 Mhz, 1 GB de RAM, conexión a ethernet, 4 puertos USB, 40 pines GPIO, entre otras. El sistema operativo disponible para esta tarjeta es Raspbian, que es una variante del sistema operativo Linux y tiene la capacidad necesaria para ejecutar los programas diseñados en el lenguaje de programación Python, que se utilizó para este desarrollo.

Cabe hacer mención, que dada la alta demanda de procesamiento de cómputo requerida para el entrenamiento del controlador basado en RL, este se llevó a cabo en una computadora de escritorio con las siguientes características: procesador Intel Core i5 a 2 GHz, 8 GB de RAM y GPU de 4 GB. Una vez realizado el entrenamiento, el cual lleva aproximadamente 40 horas en el equipo descrito, se exporta la Q-network ya entrenada en la tarjeta Raspberry Pi 2B.

Modelado matemático de la diabetes mellitus tipo 1

Como en muchos ámbitos existen diversos modelos matemáticos que describen el comportamiento de algún sistema, el modelo utilizado en este proyecto para describir el comportamiento del cuerpo humano de un adulto promedio que padece T1DM es el desarrollado por John T. Sorensen.

El modelo de Sorensen es un modelo compartimental de fisiología basado en ecuaciones diferenciales que representa varios aspectos del metabolismo de la glucosa y la insulina. Este modelo incluye varios compartimentos que simulan diferentes órganos y tejidos del cuerpo humano, como el páncreas, el hígado, los músculos y el tejido adiposo, así como la circulación sistémica. A través de estas ecuaciones diferenciales, el modelo representa la producción y eliminación de glucosa, la secreción de insulina y la captación y liberación de glucosa por los tejidos. Considera procesos fisiológicos como la gluconeogénesis y la glucogenólisis en el hígado, así como la captación de glucosa por los músculos y el tejido adiposo en respuesta a la insulina. También aborda la secreción de insulina, que está severamente comprometida en la diabetes tipo 1, y su eliminación principalmente por el hígado y los riñones. El modelo incluye numerosos parámetros fisiológicos que deben ser ajustados para representar con precisión la fisiología individual del paciente (Sorensen, 1985).

Como ya se mencionó el modelo Sorensen divide el cuerpo humano en compartimentos, esta técnica es utilizada para dividir procesos complejos en subprocesos de estructura más simple pero relacionados entre sí. Sorensen propone su modelo considerando la dinámica del metabolismo de la glucosa, esto es, el comportamiento de la

glucosa y las hormonas responsables de su regulación (insulina y glucagón) en cada uno de los compartimentos. Como resultado se obtuvo un sistema de 19 ecuaciones diferenciales dividido en los siguientes tres subsistemas:

1. Compartimento de la glucosa.

Los órganos principales que participan en el comportamiento dinámico de la glucosa en el cuerpo humano se presentan en la Figura 1, las flechas gruesas representan la dirección del flujo sanguíneo mientras que las flechas delgadas representan consumo o producción de glucosa, las flechas que apuntan hacia adentro del órgano representan la extracción de glucosa y las que apuntan hacia afuera del órgano significan la producción de esta, la descripción de las variables involucradas se presenta en la Tabla 1. A continuación, se presentan las ecuaciones del modelo de glucosa.

Cerebro

$\frac{dG_{BV}}{dt} = \frac{Q_B^G}{V_{BV}^G} (G_H - G_{BV}) - \frac{V_{BI}}{V_{BV}^G T_B} (G_{BV} - G_{BI})$	Ecuación 1
--	------------

$\frac{dG_{BI}}{dt} = \frac{V_{BI}}{V_{BI} T_B} (G_{BV} - G_{BI}) - \frac{\Gamma_{BGU}}{V_{BI}}$	Ecuación 2
--	------------

Corazón y pulmones

$\frac{dG_H}{dt} = \frac{1}{V_H^G} (Q_B^G G_{BV} + Q_L^G G_L + Q_K^Q G_K + Q_P^G G_{PV} - Q_H^G G_H - \Gamma_{RBCU})$	Ecuación 3
---	------------

Intestino

$\frac{dG_G}{dt} = \frac{Q_G^G}{V_G^G} (G_H - G_G) + \frac{1}{V_{GG}^G} (\Gamma_{meal} - \Gamma_{GGU})$	Ecuación 4
---	------------

Hígado

$\frac{dG_L}{dt} = \frac{1}{V_L^G} (Q_A^G G_H + Q_G^G G_G - Q_L^G G_L + \Gamma_{HGP} - \Gamma_{HGU})$	Ecuación 5
---	------------

Riñón

$\frac{dG_K}{dt} = \frac{Q_K^G}{V_K^G} (G_H - G_K) - \frac{\Gamma_{KGE}}{V_K^G}$	Ecuación 6
--	------------

Periferia

$\frac{dG_{PV}}{dt} = (G_H - G_{PV}) - \frac{V_{PI}}{T_P^G V_{PV}^G} (G_{PV} - G_{PI})$	Ecuación 7
---	------------

$\frac{dG_{PI}}{dt} = \frac{V_{PI}}{T_P^G V_{PI}} (G_{PV} - G_{PI}) - \frac{\Gamma_{PGU}}{V_{PI}}$	Ecuación 8
--	------------

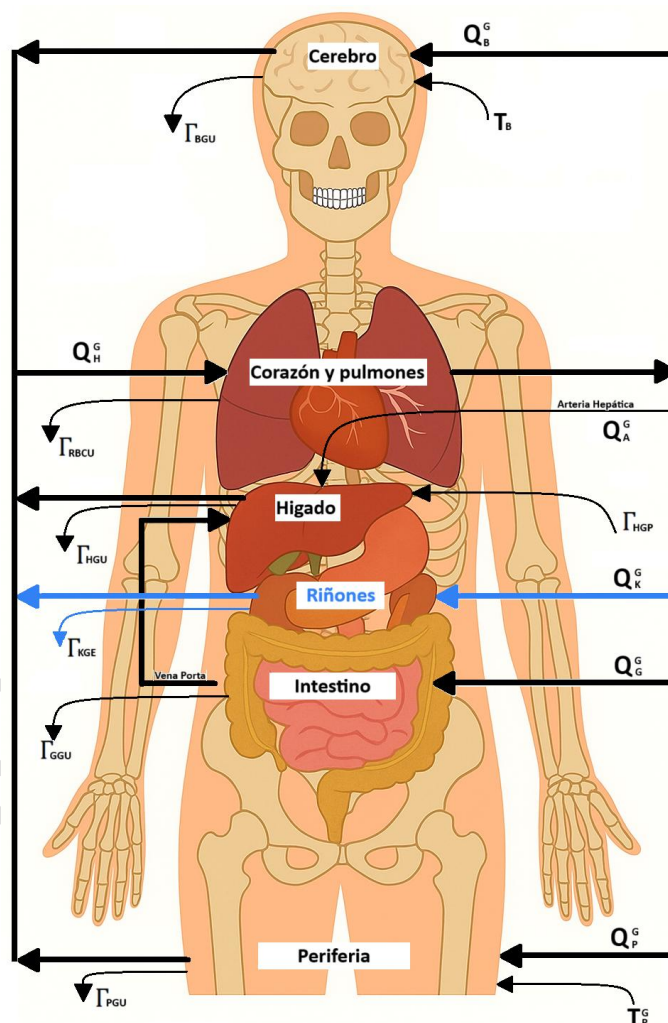


Figura 1. Compartimento de la glucosa.

2. Compartimento de la insulina.

El diagrama del comportamiento de la insulina en el cuerpo humano se presenta en la Figura 2, enseguida se muestran las ecuaciones que describen su comportamiento,

seguido de la Tabla 1 con la descripción de las variables utilizadas.

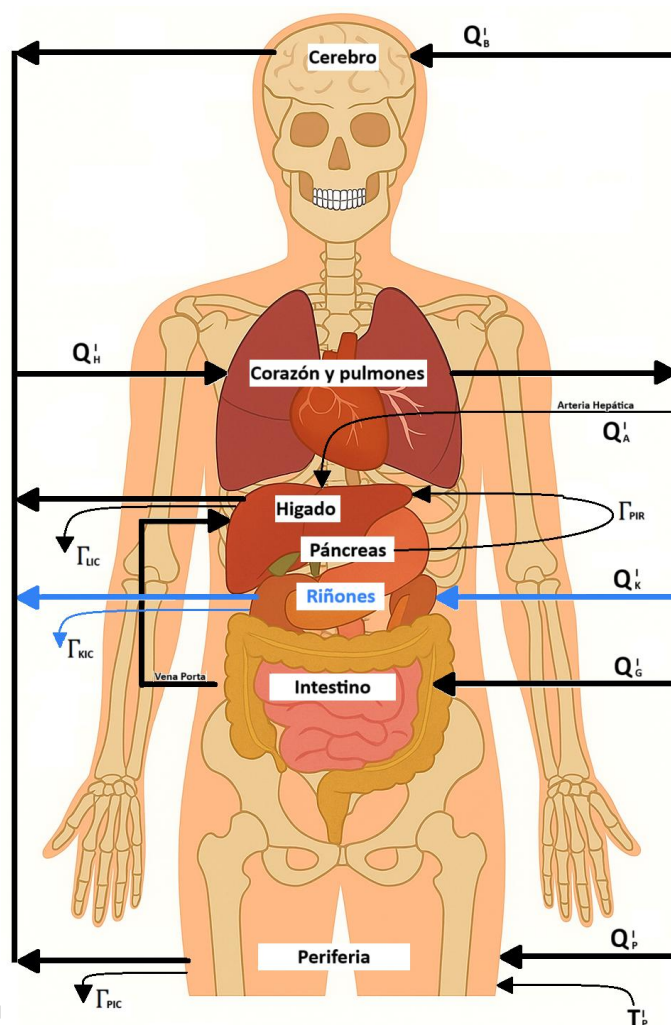


Figura 2. Compartimento de la insulina.

Cerebro

$\frac{dI_B}{dt} = \frac{Q_B^I}{V_B^I} (I_H - I_B)$	Ecuación 9
---	------------

Corazón y pulmones

$\frac{dI_H}{dt} = \frac{1}{V_H^I} (Q_B^I I_B + Q_L^I I_L + Q_K^I I_K + Q_P^I I_{PV} - Q_H^I I_H + i(t))$	Ecuación 10
---	-------------

Intestino

$\frac{dI_G}{dt} = \frac{Q_G^I}{V_G^I} (I_H - I_G)$	Ecuación 11
---	-------------

Hígado

$\frac{dI_L}{dt} = \frac{1}{V_L^I} (Q_A^I I_H + Q_G^I I_G - Q_L^I I_L + \Gamma_{PIR} - \Gamma_{LIC})$	Ecuación 12
---	-------------

Riñón

$\frac{dI_K}{dt} = \frac{Q_K^I}{V_K^I} (I_H - I_K) - \frac{\Gamma_{KIC}}{V_K^I}$	Ecuación 13
--	-------------

Periferia

$\frac{dI_{PV}}{dt} = \frac{Q_P^I}{V_{PV}^I} (I_H - I_{PV}) - \frac{V_{PI}}{T_P^I V_{PV}^I} (I_{PV} - I_{PI})$	Ecuación 14
--	-------------

$\frac{dI_{PI}}{dt} = \frac{V_{PI}}{T_P^I V_{PI}^I} (I_{PV} - I_{PI}) - \frac{\Gamma_{PIC}}{V_{PI}^I}$	Ecuación 15
--	-------------

La descripción de variables utilizadas en las ecuaciones anteriores se presenta en la Tabla 1.

3. Compartimento del glucagón.

El efecto contra regulatorio del glucagón del sistema glucosa-insulina se modela con una única ecuación

$\frac{dG_C}{dt} = 0.0916(\Gamma_{PIR} - G_C)$	Ecuación 16
--	-------------

donde:

G_C → Concentración de glucagón (pg/ml)

Γ_{PIR} → Tasa de liberación pancreática de glucagón (pg/min)

Tabla 1. Descripción de variables.

<u>Variables:</u>	<u>Segundo subíndice:</u> Subcompartimento fisiológico
G → Concentración de glucosa (mg/dl)	J → Espacio de fluido intersticial (valido para el compartimento de glucosa)
I → Concentración de insulina (mU/dl)	I → Espacio de fluido intersticial (valido para el compartimento de insulina)
Q → Tasa de flujo de agua en la sangre vascular (dl/min, valido para el compartimento de glucosa)	V → Espacio acuoso sanguíneo vascular (valido para el compartimento de glucosa)
Q → Tasa de flujo de plasma vascular (l/min, valido para el compartimento de insulina)	

Γ → Tasa metabólica (mg/min) T → Tiempo de difusión transcilar (min) V → Volumen (dl) t → Tiempo (min)	V → Espacio de plasma vascular (valido para el compartimento de insulina)
<u>Primer subíndice:</u> Compartimento fisiológico B → Cerebro G → Intestino H → Corazón y pulmones L → Hígado P → Periferia A → Arteria hepática	<u>Subíndice de tasa metabólica:</u> BGU → Captación de glucosa en el cerebro GGU → Utilización de glucosa en el intestino HGP → Producción de glucosa en el hígado HGU → Captación de glucosa en el hígado KGE → Excreción renal de glucosa PGU → Captación de glucosa en la periferia $RBCU$ → Captación de glucosa por los glóbulos rojos KIC → Depuración renal de insulina LIC → Depuración hepática de insulina PIC → Depuración periférica de insulina PIR → Liberación de insulina pancreática
	<u>Superíndice:</u> G → Modelo de glucosa I → Modelo de insulina

Control basado en aprendizaje por refuerzo (RBLC)

Un RBLC se subdivide en varios elementos que lo componen (ver Figura 3), enlistados a continuación en términos del problema abordado:

- *Un agente*, es el sistema de control de la glucosa.
- *Un entorno*, es el paciente con T1DM.
- *Estado*, son las variables que provienen del entorno con las que cuenta el agente para tomar las decisiones de control necesarias.
- *Acción*, es la cantidad de insulina inyectada por el agente al paciente con T1DM.
- *Recompensa*, es el parámetro que le indica al agente que tan bien o no está realizando su trabajo.

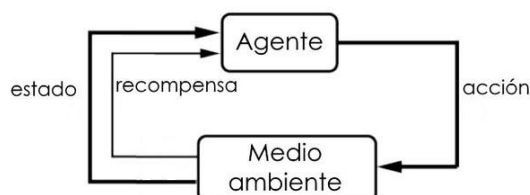


Figura 3. Estructura de un RLBC.

Un algoritmo de aprendizaje por refuerzo se encarga de que un agente aprenda una política, es decir, una estrategia de toma de decisiones de tal manera que maximice la recompensa acumulada. Un algoritmo comúnmente utilizado es Q-learning que estima el valor de la recompensa futura esperada para cada acción en cada estado, es decir, la función $Q(s, a)$ estima la recompensa acumulada esperada de tomar una acción a en un estado s (Watkins & Dayan, 1992). La ecuación de actualización de la función Q es:

$Q(s, a) \leftarrow Q(s, a) + \alpha[r + \gamma \max_{a'} Q(s', a') - Q(s, a)]$	Ecuación 17
---	-------------

donde:

- $Q(s, a)$ es el valor Q actual para el estado s y la acción a .
- α es la tasa de aprendizaje, un parámetro que determina que tanto se actualizan los valores Q .
- r es la recompensa recibida después de tomar la acción a en el estado s .
- γ es el factor de descuento, que pondera la importancia de las recompensas futuras.
- $\max_{a'} Q(s', a')$ es el valor Q máximo para el próximo estado s' sobre todas las posibles acciones a' .

Un punto clave para el correcto funcionamiento del algoritmo Q-learning es un buen diseño de la función de recompensa, en esta implementación se toman en cuenta varios puntos para la función de recompensa. El primer factor a considerar es el tiempo de simulación máximo, ya que los algoritmos de RL son iterativos y se busca que en cada iteración su desempeño mejore al buscar obtener una mejor recompensa, por lo tanto, la función de recompensa propuesta le da una recompensa de 1 a cada iteración que el algoritmo mantiene al paciente dentro de los rangos de glucosa aceptable. Por lo tanto, cada época de entrenamiento del algoritmo concluye cuando el paciente cae en hipoglucemia o hiperglucemia más un margen de error. Además, al factor anterior se le resta un factor que depende de la cercanía con un valor de glucosa objetivo, por lo tanto, la función de recompensa total está dada por la siguiente ecuación:

$Recompensa[k] = 1 + \frac{Recompensa[k-1]}{\sqrt{(glucosa\ objetivo - glucosa\ actual)^2}}$	Ecuación 18
--	-------------

donde k representa el número de iteración presente.

Para este trabajo se eligió el algoritmo de RL Deep-Q-Learning (DQL), este método hace uso de una red neuronal para estimar los valores de la función Q . Esta red, conocida como Q-network, toma un estado s como entrada y produce un vector de valores Q para todas las acciones posibles en ese estado $Q(s, a; \theta)$, donde θ son los parámetros de la red neuronal (Van et al., 2016).

Adicionalmente el algoritmo DQL emplea un buffer de memoria llamado *experience replay* para romper las correlaciones en la secuencia de datos observados y estabilizar el entrenamiento. Las experiencias se almacenan como tuplas (s, a, r, s') y se muestrean aleatoriamente durante el entrenamiento. Para reducir las oscilaciones y la inestabilidad en el entrenamiento, DQL utiliza una copia separada de la red Q-network llamada *target network*, con parámetros θ^- . Esta red se actualiza periódicamente copiando los parámetros de la red principal

$\theta^- = \theta$	Ecuación 19
---------------------	-------------

Los parámetros de la Q-network se actualizan minimizando la diferencia entre el valor Q actual y el valor objetivo y :

$y = r + \gamma \max_{a'} Q(s', a'; \theta^-)$	Ecuación 20
--	-------------

La función de perdida se define como:

$L(\theta) = \mathbb{E}_{(s,a,r,s') \sim \mathcal{D}} [(y - Q(s, a; \theta))^2]$	Ecuación 21
--	-------------

donde $\mathbb{E}_{(s,a,r,s')}$ es el valor esperado y $\mathcal{D}$ es la distribución de experiencias almacenadas en el buffer de replay.

Por otra parte, la política ϵ -greedy es la estrategia donde el agente, para cada estado s , decide entre explorar nuevas acciones o explotar las acciones que ya conoce como las mejores según las experiencias previas. Comúnmente el parámetro ϵ decrece de forma lineal o exponencial, conforme el agente gana confianza en el conocimiento que ha adquirido, lo que aumenta la probabilidad de explotación y disminuye la de exploración.

En resumen, el algoritmo de entrenamiento para el algoritmo DQL se resume en los siguientes pasos:

1. Iniciar la Q-network con parámetros aleatorios θ .
2. Iniciar la target network con los mismos parámetros $\theta^- = \theta$.
3. Iniciar la memoria de replay $\mathcal{D}$.
4. Para cada episodio.
 - Inicializar el estado s .
 - Para cada paso del episodio:
 - a. Seleccionar una acción a usando la política ϵ -greedy.
 - b. Ejecutar la acción a en el entorno y observar la recompensa r y el nuevo estado s' .

- c. Almacenar la transición (s, a, r, s') en el buffer de replay $\mathcal{D}$.
- d. Muestrear un minibatch de transiciones (s_i, a_i, r_i, s'_i) de $\mathcal{D}$.
- e. Calcular el valor objetivo para muestra del minibatch:

$y_i = r_i + \gamma \max_{a'} Q(s'_i, a'; \theta^-)$	Ecuación 22
--	-------------

- f. Actualizar los parámetros θ minimizando la función de pérdida:

$\theta \leftarrow \theta - \alpha \nabla_{\theta} L(\theta)$	Ecuación 23
---	-------------

- g. Cada cierto número de pasos, actualizar la target network según la Ecuación (19)

Resultados

A continuación, se presentan los resultados obtenidos al simular el modelo de Sorensen en la tarjeta Raspberry Pi, cabe destacar que la frecuencia de muestreo del controlador es la misma que la utilizada en la simulación del modelo de Sorensen, es decir, ambos operan con un intervalo de muestreo de $\delta=0.425$ minutos. Esta decisión se justifica debido a que el sistema fisiológico involucrado presenta una dinámica lenta, lo cual implica que los cambios significativos en los niveles de glucosa e insulina ocurren en escalas de tiempo relativamente largas. Por lo tanto, no se requiere una frecuencia de muestreo alta, ya que una mayor granularidad no aportaría beneficios sustanciales al desempeño del sistema de control. Esta característica permite además reducir la carga computacional del sistema, lo cual es especialmente relevante al implementarlo en plataformas de bajo costo como la Raspberry Pi.

Las simulaciones se realizaron considerando las condiciones iniciales correspondientes al estado basal de una persona con T1DM (definidas en la Tabla 2), entendiendo por estado basal aquel en el que el individuo no ha ingerido alimentos recientemente, no se encuentra bajo estrés físico o emocional, y sus niveles hormonales y metabólicos se encuentran en equilibrio. Este estado representa un punto de partida fisiológicamente estable, y se utiliza comúnmente como referencia en simulaciones para evaluar el comportamiento de sistemas de control sin interferencias externas iniciales.

Respecto a la Q-network se eligieron dos distintas configuraciones de red, la primera con una capa de entrada con 10 neuronas, correspondiente a los últimos 10 valores de glucosa que describen la tendencia comportamental del modelo de Sorensen, 128 neuronas en una capa oculta y una capa de salida con tres neuronas, correspondiente al número de acciones posibles aplicables al modelo. Las posibles acciones que puede tomar el sistema

de control de RL son aumentar en una unidad el suministro de insulina al paciente, disminuir en una unidad o permanecer sin cambio. Todas las neuronas utilizan la función de activación ReLu, dado el tipo de dato de la capa de salida de la red, se implementó una función que elige la acción a tomar basado en la máxima recompensa que se pueda obtener dada esa acción. La política ϵ -greedy elegida tiene un decrecimiento de forma exponencial, ya que en pruebas realizadas esta fue la que mejor desempeño presento para esta implementación. La segunda configuración de red utilizada tiene como única diferencia respecto a la primera que cuenta con dos capas ocultas de 128 neuronas.

Tabla 2. Condiciones iniciales de simulación.

$G_{BV} = 75.5819$	$I_G = 32.1744$
$G_{BI} = 42.9091$	$I_L = 19.3030$
$G_H = 87.4477$	$I_K = 24.7497$
$G_G = 85.4673$	$I_{PV} = 27.3504$
$G_L = 98.0168$	$I_{PI} = 12.32$
$G_K = 87.4477$	$G_C = 0.7811$
$G_{PV} = 83.9350$	$M_{HGPI} = 1.1910$
$G_{PI} = 85$	$f_2 = -0.0969$
$I_B = 32.1744$	$M_{HGUI} = 0.9162$
$I_H = 32.1744$	

El entrenamiento de las redes neuronales se realizó por 2,000 épocas, se utilizó un buffer replay de 10,000 elementos de almacenamiento y un tamaño de batch de 32 elementos. Respecto a la función de recompensa se utiliza un valor de glucosa objetivo de 85 gm/dL, los resultados de ambos controladores se comparan con el desempeño del controlador basado en μ -síntesis (Navarrete *et al.*, 2024). Para este trabajo, como se mencionó anteriormente, se consideran que cada época de entrenamiento concluye cuando los niveles de glucosa exceden los valores de hipoglucemia menos la tolerancia en 50 gm/dL o menos y la hiperglucemia más la tolerancia por arriba de las 300 gm/dL.

La cantidad de muestras utilizadas durante el entrenamiento, así como la diversidad de escenarios simulados, permiten que el agente de RL explore un amplio rango de condiciones fisiológicas relevantes para el control de la glucosa. Esto incluye distintos niveles iniciales de glucosa, respuestas a ingestas de carbohidratos en distintos momentos del día, y transiciones entre estados de euglucemia, hipoglucemia e hiperglucemia. Esta variabilidad en las muestras contribuye significativamente al proceso de aprendizaje, ya que ofrece al agente ejemplos ricos y representativos de las dinámicas del sistema, facilitando la generalización y la selección de políticas de control más robustas y efectivas.

Se plantean tres escenarios de simulación, presentados en la Tabla 3, para probar el desempeño de los controladores basados en RL y compararlo con el basado en μ -síntesis, los escenarios se eligen de forma similar a los propuestos en (Navarrete *et al.*,

2024), con la diferencia de que los controladores se prueban con niveles de exigencia mayor. En la siguiente tabla se muestran las características de cada uno de estos tres escenarios de simulación, donde se busca tener variedad en las cantidades de carbohidratos ingeridos, la hora de las ingestas y el número de las mismas, en todos los casos se elige simular periodos de 24 horas.

Tabla 3. Escenarios de simulación propuestos.

Horario de comidas	Cantidad (gramos de carbohidratos)	Duración de la comida (minutos)
ESCENARIO 1		
7:00 hrs	40	25
14:00 hrs	60	30
20:00 hrs	30	15
ESCENARIO 2		
7:00 hrs	40	25
10:00 hrs	20	10
14:00 hrs	60	30
17:00 hrs	20	10
20:00 hrs	30	15
ESCENARIO 3		
7:00 hrs	40	25
10:00 hrs	40	10
14:00 hrs	60	30
17:00 hrs	40	10
20:00 hrs	30	15

En las Figuras 4, 5 y 6 se observan los resultados obtenidos en los tres escenarios de simulación propuestos, en todas las figuras es posible observar gráficamente las cantidades y momentos de la ingesta de carbohidratos, la elevación de la glucosa y como los controladores suministran insulina para intentar llevar la glucosa de los pacientes a niveles aceptables. Datos presentados por la Asociación Americana de la Diabetes (ADA) aseguran que los niveles de glucosa en la sangre en estado de euglucemia deben permanecer entre los 70 y los 180 mg/dL, aunque esto puede variar dependiendo del momento del día. En la Figura 4 se puede observar cómo después de la primera ingesta de alimentos la glucosa regulada por el controlador basado en μ -síntesis sobrepasa estos límites, mientras que la de los controladores basado en RL permanecen dentro de los límites mencionados, resaltados con una franja verde en todas las gráficas. En importante notar el comportamiento de la glucosa después de la ingesta de carbohidratos de las 14:00 hrs, ya que la glucosa del paciente con todos los controladores sobrepasa considerablemente el límite máximo mencionado anteriormente, situación que si bien no es deseada, tampoco se considera un riesgo mayor para la salud de los pacientes, ya que según la ADA, este nivel

máximo se puede sobrepasar siempre y cuando la glucosa regrese a sus rangos normales en un tiempo no mayor a dos horas después de la ingesta de carbohidratos, situación que se cumple para todos los controladores.

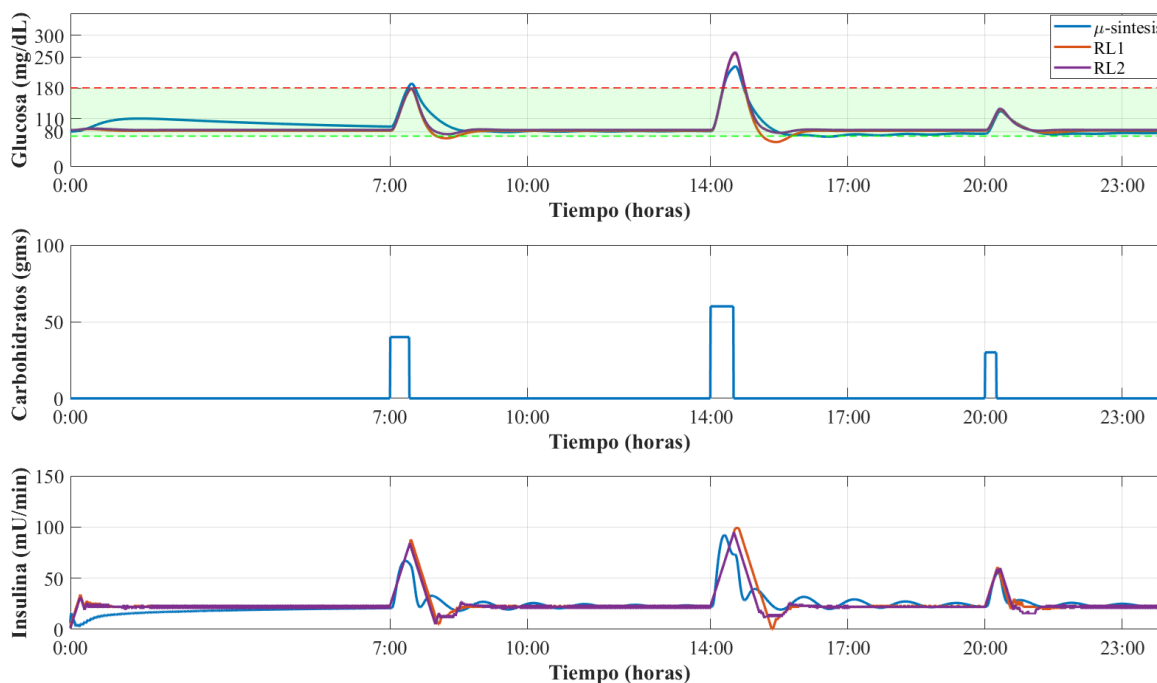


Figura 4. Comparativa de comportamiento de paciente con T1DM controlado por algoritmo μ -síntesis vs. controladores basados en RL, según las condiciones del escenario 1.

Los resultados obtenidos a partir de la simulación llevada a cabo con las ingestas mencionadas en el escenario 2 se presentan en la Figura 5, se puede observar que el comportamiento es muy similar a lo presentado en la Figura 4, ya que las ingestas realizadas a las 10:00 y 17:00 hrs no son muy grandes, por lo que no representan estrés para el sistema y todos los controladores pueden realizar su trabajo sin que estas ingestas provoquen que la glucosa del paciente exceda los niveles normales.

Por último, en la Figura 6 se observan los resultados obtenidos en condiciones similares a las presentadas en el escenario 2, únicamente con ingestas mayores que en el caso anterior realizadas a las 10:00 y 17:00 hrs, se puede observar que para el caso del controlador basado en μ -síntesis la glucosa excede los valores esperados en la ingesta de las 10:00 hrs, el resto del comportamiento para todos los controladores es muy similar al caso anterior.

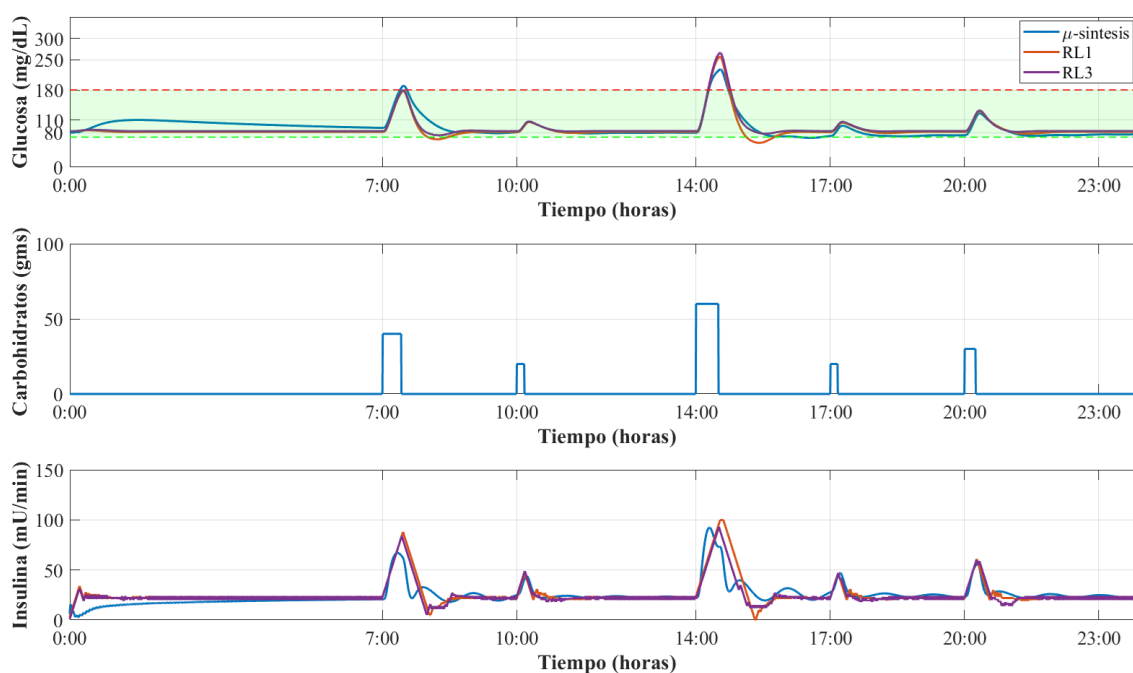


Figura 5. Comparativa de comportamiento de paciente con T1DM controlado por algoritmo μ -síntesis vs. controladores basados en RL, según las condiciones del escenario 2.

Es necesario mencionar, que para los tres escenarios de simulación presentados, el caso del controlador basado en RL con una capa oculta (identificado en las figuras como RL1), presenta un desempeño similar al del controlador basado en RL con dos capas ocultas (identificado en las figuras como RL2), con la principal diferencia que el RL1 presenta sobretiros negativos que hacen que el paciente entre en estado de hipoglucemia por instantes de tiempo cortos, que si bien es cierto, no ponen en riesgo la vida de un paciente, es un comportamiento indeseado.

Ya que visualmente no resulta sencillo determinar que controlador tiene un comportamiento más cercano al deseado, se realiza un análisis cuantitativo para determinar el error promedio de los niveles de glucosa obtenidos respecto a los deseados, este error promedio se realiza calculando las normas L1 y L2 por medio de las siguientes ecuaciones:

$$L_1 = \frac{1}{T} \int_0^T \|e(t)\| dt$$

Ecuación 24

$$L_2 = \sqrt{\frac{1}{T} \int_0^T e(t)^2 dt}$$

Ecuación 25

Donde $e(t)$ representa el error de la glucosa respecto al valor objetivo, las barras paralelas representan el valor absoluto y T el tiempo total de simulación. Los resultados

obtenidos de aplicar esta métrica al desempeño de los controladores son presentados en la Tabla 4.

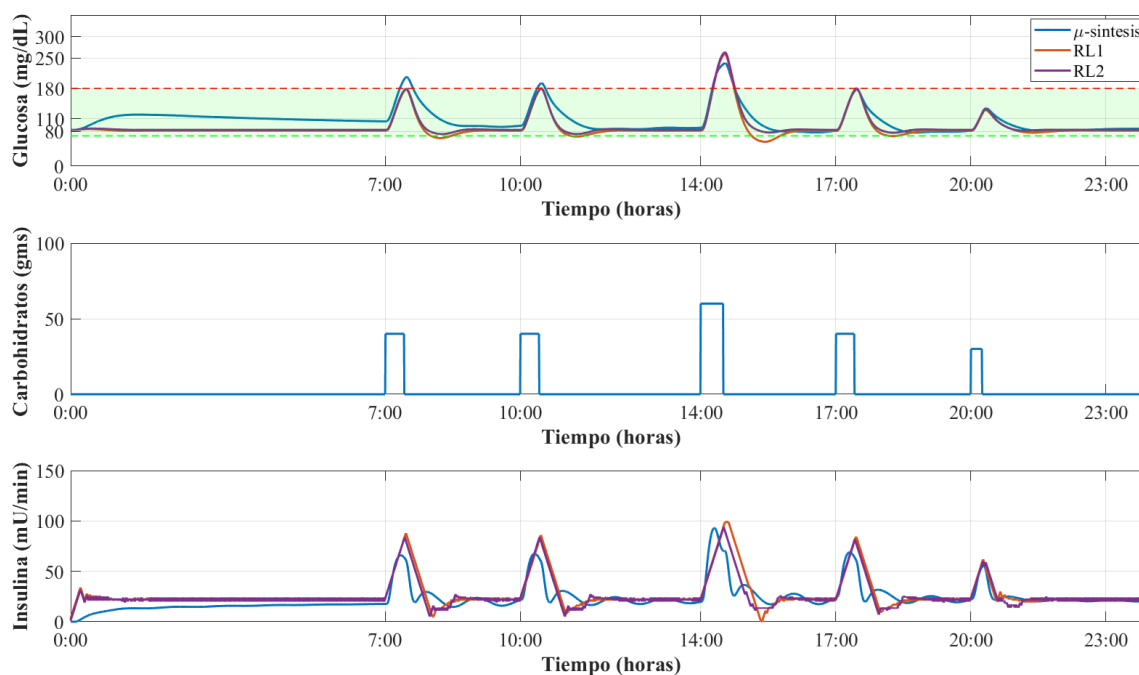


Figura 6. Comparativa de comportamiento de paciente con T1DM controlado por algoritmo μ -síntesis vs. controladores basados en RL, según las condiciones del escenario 3.

Tabla 4. Evaluación cuantitativa de los controladores.

	Norma L ₁			Norma L ₂		
	μ -síntesis	RL1	RL2	μ -síntesis	RL1	RL2
Escenario 1	5285.9	3255.7	2560.7	1530.0	1457.3	1463.6
Escenario 2	5391.2	3412.1	2819.5	1527.2	1453.1	1508.3
Escenario 3	7631.4	4521.8	3805.3	2088.8	1706.3	1725.0

En la Tabla 4 se observa claramente que en todos los escenarios ambos controladores basados en RL superan en desempeño al controlador basado en μ -síntesis, recordando que un menor valor en las métricas utilizadas representa un comportamiento más cercano al ideal. Por otro lado, no es fácil concluir cuál de los controladores basados en RL tiene mejor desempeño, ya que para la norma L₁ el controlador RL2 presenta mejores resultados, pero observando la norma L₂ el controlador RL1 se desempeña mejor.

Discusión

Los resultados obtenidos muestran que el controlador basado en aprendizaje por refuerzo propuesto es capaz de regular de manera efectiva la concentración de glucosa en sangre dentro de rangos fisiológicamente aceptables, manteniendo un comportamiento estable ante perturbaciones asociadas a la dinámica metabólica del sistema. Este desempeño es consistente con el reportado en la literatura reciente sobre controladores para páncreas artificial basados en RL, los cuales han demostrado ser herramientas prometedoras para el control automático de la glucosa en pacientes con T1DM.

Diversos trabajos han explorado el uso de algoritmos de RL profundo para el control de la administración de insulina en sistemas de páncreas artificial. Por ejemplo, un controlador basado en Dual Proximal Policy Optimization (PPO) mostró mejoras significativas en el tiempo dentro del rango glucémico objetivo, reduciendo episodios de hiperglucemia y manteniendo bajos niveles de hipoglucemia en simulaciones realizadas con pacientes virtuales del simulador UVA/Padova (Marchetti *et al.*, 2025). De manera similar, otros estudios han demostrado que controladores basados en RL pueden estabilizar la glucosa en sangre y mejorar el control glucémico en comparación con estrategias convencionales como PID o terapias basales-bolo tradicionales (Ngo *et al.*, 2018). Los resultados obtenidos en este trabajo son congruentes con estos reportes, lo cual sugiere que la estrategia propuesta presenta un desempeño comparable al de otros enfoques basados en RL para el control de glucosa.

Sin embargo, una diferencia importante con respecto a gran parte de los trabajos reportados en la literatura radica en la infraestructura computacional utilizada para el desarrollo y evaluación de los controladores. Muchos de los algoritmos de RL profundo empleados en investigaciones recientes son entrenados y ejecutados en plataformas de cómputo de alto desempeño, debido a la complejidad de los modelos utilizados y al elevado número de parámetros que deben optimizarse durante el proceso de aprendizaje. Por ejemplo, controladores avanzados basados en RL incluyen múltiples agentes cooperativos, redes neuronales profundas o mecanismos de optimización complejos que incrementan considerablemente los requerimientos computacionales del sistema (Marchetti *et al.*, 2025). En contraste, el controlador propuesto en este trabajo está diseñado para operar en una plataforma embebida de bajo costo, lo que representa una ventaja significativa en términos de viabilidad práctica para la implementación en dispositivos portátiles de control automático de insulina.

Otra característica relevante del controlador presentado es que su estrategia de control se basa únicamente en el conocimiento de los valores actuales e históricos de la concentración de glucosa para determinar la acción de control. Esta característica simplifica considerablemente el diseño del sistema, ya que evita la necesidad de medir o estimar un

gran número de variables fisiológicas adicionales. En contraste, algunos enfoques reportados en la literatura requieren la estimación de múltiples parámetros o estados internos del sistema metabólico, como tasas de absorción de glucosa, sensibilidad a la insulina o parámetros dinámicos del modelo fisiológico del paciente, lo cual incrementa la complejidad del sistema de control y puede introducir incertidumbre adicional en la estimación de dichas variables. Por ejemplo, ciertos esquemas de control robusto o modelos híbridos basados en RL requieren estructuras de control con múltiples parámetros ajustables o modelos fisiológicos detallados para garantizar estabilidad y desempeño bajo incertidumbre (Tanzanakis & Lygeros, 2023).

Desde esta perspectiva, la simplicidad estructural del controlador propuesto constituye una ventaja importante, al utilizar únicamente la información proveniente del monitoreo continuo de glucosa y su historial reciente, el algoritmo mantiene un bajo requerimiento de información del sistema mientras conserva la capacidad de adaptarse a la dinámica no lineal del metabolismo de la glucosa. Esta característica es particularmente relevante para aplicaciones en sistemas de páncreas artificial portátiles, donde la disponibilidad de sensores adicionales puede ser limitada y donde la robustez frente a incertidumbre en los parámetros fisiológicos es un aspecto crítico.

En conjunto, los resultados obtenidos y la comparación con trabajos previos sugieren que el controlador propuesto logra un equilibrio favorable entre desempeño, simplicidad del modelo y requerimientos computacionales. Mientras que otros enfoques basados en RL han demostrado resultados prometedores en simulaciones y entornos de investigación, la posibilidad de implementar el algoritmo presentado en hardware embebido de bajo costo representa un paso relevante hacia el desarrollo de soluciones prácticas y accesibles para el control automático de glucosa en sistemas de páncreas artificial.

Conclusiones

En este trabajo se presentó el desarrollo e implementación de una plataforma para el control de glucosa en pacientes con T1DM basada en RL profundo. El sistema propuesto integra el modelo fisiológico de Sorensen con un controlador basado en Deep Q-Learning, ejecutado en una plataforma embebida de bajo costo.

Los resultados obtenidos en los distintos escenarios de simulación muestran que los controladores basados en aprendizaje por refuerzo presentan un mejor desempeño en comparación con el controlador robusto basado en μ -síntesis utilizado como referencia. En particular, ambos controladores basados en aprendizaje por refuerzo lograron mantener los niveles de glucosa más cercanos al valor objetivo, obteniendo menores valores en las métricas de error L1 y L2 en los tres escenarios analizados. Estos resultados indican que el enfoque basado en aprendizaje por refuerzo es capaz de adaptarse de manera efectiva

a las dinámicas no lineales del sistema glucosa–insulina y responder adecuadamente ante perturbaciones fisiológicas como la ingesta de alimentos.

La comparación entre las dos configuraciones de red neuronal utilizadas permitió observar que ambos controladores presentan un comportamiento general similar. No obstante, el controlador con dos capas ocultas (RL2) mostró una mayor estabilidad al evitar episodios de hipoglucemia transitoria observados en el controlador con una sola capa oculta (RL1). Por otro lado, el controlador RL1 presenta una menor complejidad computacional, lo cual puede resultar ventajoso en implementaciones donde los recursos de procesamiento son limitados.

Asimismo, una de las ventajas relevantes del enfoque basado en RL es que el controlador requiere únicamente la medición de los niveles actuales de glucosa y el historial reciente de los mismos para tomar decisiones de control. Esto contrasta con otros enfoques basados en modelos fisiológicos completos, que requieren estimar múltiples variables internas del sistema, lo cual puede resultar complejo en entornos clínicos reales.

Además, la implementación del modelo fisiológico y del controlador basado en RL en una plataforma embebida de bajo costo representa un avance importante hacia el desarrollo de dispositivos portátiles de control automático de insulina. Esta característica resulta especialmente relevante si se considera que muchos trabajos reportados en la literatura se limitan a evaluaciones realizadas únicamente en entornos de simulación o en plataformas de alto desempeño computacional. En este sentido, el presente trabajo demuestra la viabilidad de ejecutar tanto el modelo fisiológico como el algoritmo de control en hardware accesible, lo cual abre la posibilidad de desarrollar soluciones tecnológicas más accesibles y portátiles para el manejo de la diabetes.

Los resultados obtenidos en este estudio son congruentes con lo reportado en otros trabajos recientes que aplican técnicas de RL profundo al problema del control automático de la glucosa, como los discutidos en la sección de Discusión. En dichos estudios se ha demostrado que los algoritmos de RL pueden mejorar el control glucémico al mantener los niveles de glucosa dentro del rango objetivo durante mayores periodos de tiempo y reducir la incidencia de eventos de hipoglucemia e hiperglucemia. En este contexto, el desempeño observado en el presente trabajo confirma el potencial de este tipo de algoritmos para aplicaciones en sistemas de páncreas artificial y control automatizado de la administración de insulina.

Finalmente, aunque los resultados obtenidos son prometedores, es importante señalar que las evaluaciones realizadas en este trabajo se basan en simulaciones utilizando un modelo fisiológico del sistema glucosa–insulina. Por lo tanto, como trabajo futuro se propone continuar investigando otras variantes de algoritmos de RL y realizar evaluaciones adicionales que permitan avanzar hacia la validación experimental del sistema en entornos más cercanos a condiciones reales.

Contribución de los autores

“Conceptualización del trabajo, autor 1, autor 2, autor 3.; desarrollo de la metodología, autor 2, autor 3.; manejo de software, autor 1, autor 2.; validación experimental, autor 3, autor 4.; análisis de resultados, autor 1, autor 3.; Manejo de datos, autor 3, autor 4.; escritura y preparación del manuscrito, autor 1, autor 2.; redacción, revisión y edición, autor 3, autor 4.; administrador de proyectos, autor 1, autor 2.

“Todos los autores de este manuscrito han leído y aceptado la versión publicada del mismo.”

Financiamiento

"Esta investigación fue financiada con fondos propios".

Conflicto de interés

“Los autores declaran no tener conflicto de interés”.

Referencias

Dénes-Fazakas, L., Siket, M., Kertész, G., Szilágyi, L., Kovács, L., & Eigner, G. (2022). Control of type 1 diabetes mellitus using direct reinforcement learning based controller. In *2022 IEEE International Conference on Systems, Man, and Cybernetics (SMC)*, 1512-1517. IEEE. <http://dx.doi.org/10.1109/SMC53654.2022.9945084>

Fox, I., & Wiens, J. (2019). Reinforcement learning for blood glucose control: Challenges and opportunities. *Reinforcement Learning for Real Life (RL4RealLife) Workshop in the 36th International Conference on Machine Learning*. <https://openreview.net/forum?id=ByexVzSAs4>

Katsarou, A., Gudbjörnsdottir, S., Rawshani, A., Dabelea, D., Bonifacio, E., Anderson, B. J., Jacobsen, L. M., Schatz, D. A., & Lernmark, Å. (2017). Type 1 diabetes mellitus. *Nature Reviews Disease Primers*, 3, 1-17. <http://dx.doi.org/10.1038/nrdp.2017.16>

Kovács, L., Benyó, B., Bokor, J., & Benyó, Z. (2011). Induced L2-norm minimization of glucose–insulin system for Type I diabetic patients. *Computer Methods and Programs in Biomedicine*, 102(2), 105-118. <http://dx.doi.org/10.3182/20090812-3-DK-2006.0068>

Magliano, D. J., & Boyko, E. J. (2022). IDF diabetes atlas. Federación

Internacional de Diabetes. <https://www.diabetesatlas.org>

Marchetti, A., Sasso, D., D'Antoni, F., Morandin, F., Parton, M., Matarrese, M. A. G., & Merone, M. (2025). Deep reinforcement learning for Type 1 Diabetes: Dual PPO controller for personalized insulin management. *Computers in Biology and Medicine*, 191, 110-147. <https://doi.org/10.1016/j.compbimed.2025.110147>

Navarrete, G. A., & Vidrios-Serrano, C., & Maldonado-Fregoso, B., & García, R. J. A., & Chocoteco, J. A. (2024). Implementación del modelo y control en tiempo discreto para diabetes mellitus tipo 1 en una plataforma embebida de bajo costo. In *Ingeniantes, Instituto Tecnológico Superior de Misantla*. 3(1), 97-102. <https://citt.itsm.edu.mx/ingeniantes/articulos/ingeniantes11no1vol3/13.pdf>

Ngo, P. D., Wei, S., Holubová, A., Muzik, J., & Godtliebsen, F. (2018). Control of Blood Glucose for Type-1 Diabetes by Using Reinforcement Learning with Feedforward Algorithm. *Computational and mathematical methods in medicine*, 2018, 4091497. <https://doi.org/10.1155/2018/4091497>

Parker, R. S., Doyle III, F. J., Ward, J. H., & Peppas, N. A. (2000). Robust H^∞ glucose control in diabetes using a physiological model. *AIChE Journal*, 46(12), 2537-2549. <https://aiche.onlinelibrary.wiley.com/doi/abs/10.1002/aic.690461220>

Ramprasad, Y., Rangaiah, G. P., & Lakshminarayanan, S. (2004). Robust PID controller for blood glucose regulation in type I diabetics. *Industrial & engineering chemistry research*, 43(26), 8257-8268. <http://dx.doi.org/10.1021/ie049546a>

Ruiz-Velázquez, E., García-Rodríguez, J., Quiroz, G., & Femat, R. (2020). Robust μ -synthesis: Towards a unified glucose control in adults, adolescents and children with T1DM. *Journal of the Franklin Institute*, 357(14), 9633-9653. <http://dx.doi.org/10.1016/j.jfranklin.2020.07.030>

Sánchez-Peña, R., Colmegna, P., Grosembacher, L., Breton, M., De Battista, H., Garelli, F., & Cherňavsky, D. (2017). Artificial pancreas: first clinical trials in Argentina. *IFAC-PapersOnLine*, 50(1), 7731-7736. <https://doi.org/10.1016/j.ifacol.2017.08.1151>

Sorensen, J. T. (1985). A physiologic model of glucose metabolism in man and

its use to design and assess improved insulin therapies for diabetes (Doctoral dissertation, Massachusetts Institute of Technology). MIT Libraries. <https://dspace.mit.edu/bitstream/handle/1721.1/15234/13180776-MIT.pdf?sequence=2>

Syed, F. Z. (2022). Type 1 diabetes mellitus. *Annals of internal medicine*, 175(3), ITC33-ITC48. <http://dx.doi.org/10.7326/AITC202203150>

Tanzanakis, A., & Lygeros, J. (2023). Data-Driven Robust Reinforcement Learning Control of Uncertain Nonlinear Systems: Towards a Fully-Automated, Insulin-Based Artificial Pancreas. *arXiv preprint arXiv:2312.04503*. <https://doi.org/10.48550/arXiv.2312.04503>

Van Hasselt, H., Guez, A., & Silver, D. (2016). Deep reinforcement learning with double q-learning. In *Proceedings of the AAAI conference on artificial intelligence*, 30(1). <https://doi.org/10.1609/aaai.v30i1.10295>

Viroonluecha, P., Egea-Lopez, E., & Santa, J. (2022). Evaluation of blood glucose level control in type 1 diabetic patients using deep reinforcement learning. *Plos one*, 17(9), e0274608. <http://dx.doi.org/10.1371/journal.pone.0274608>

Wang, Z., Xie, Z., Tu, E., Zhong, A., Liu, Y., Ding, J., & Yang, J. (2021). Reinforcement learning-based insulin injection time and dosages optimization. In *2021 International Joint Conference on Neural Networks (IJCNN)*, 1-8. IEEE. <http://dx.doi.org/10.1109/IJCNN52387.2021.9533957>

Watkins, C. J., & Dayan, P. (1992). Q-learning. *Machine learning*, 8, 279-292. <https://doi.org/10.1007/BF00992698>