Data-Driven Control of Type 2 Diabetes Progression via Personalized Physical Activity

Giada Lops * **Pierluigi Francesco De Paola** *,**,*** **Vito Andrea Racanelli** * **Gioacchino Manfredi** * **Luca De Cicco** * **Saverio Mascolo** *

* *Politecnico di Bari, Italy*
** *Consiglio Nazionale delle Ricerche (CNR), Istituto di Elettronica e di Ingegneria dell'Informazione e delle Telecomunicazioni (IEIIT), Italy*
*** *Consiglio Nazionale delle Ricerche (CNR), Istituto di Analisi dei Sistemi ed Informatica (IASI)*

Abstract: This work investigates long-horizon regulation of Type 2 Diabetes progression through daily physical activity using a data-driven controller based on Proximal Policy Optimization. A five-state physiological model (comprising glucose, insulin, β-cell mass, insulin sensitivity, and IL-6 dynamic) is embedded in a custom environment enabling closed-loop simulations over a two-year horizon. The framework introduces realistic variability through parameter and initial-condition perturbations ($\pm 5\%$), bounded exogenous glucose perturbations ($\pm 20\,\mathrm{mg/dL}$), and mid-episode degradation of the insulin-sensitivity target, providing a physiologically consistent and challenging benchmark. The Proximal Policy Optimization agent learns adaptive daily exercise policies that preserve glucose homeostasis and robustness against uncertainty. Across a 200-patient virtual evaluation cohort, the controller achieves a 66% success rate in maintaining final glucose levels below $126\,\mathrm{mg/dL}$, demonstrating the feasibility of reinforcement learning for long-term, personalized physical activity regulation and its potential to support model-based digital therapeutics in Type 2 Diabetes management.

Keywords: Type 2 Diabetes; Reinforcement Learning; Proximal Policy Optimization; Physical activity; Long-horizon control; Robust simulation; Physiological modelling.

1. INTRODUCTION

Type 2 diabetes (T2D) is a chronic disease posing a severe challenge worldwide due to its rising prevalence, with an important impact on healthcare systems. T2D onset is mainly driven by peripheral insulin resistance and progressive β-cell mass failure, which can result in several long-term complications potentially affecting the human body at a systemic level, such as blindness or stroke. It is well acknowledged that, in light of its aetiology, T2D can be prevented or delayed through physical exercise and lifestyle interventions (Zimmet et al., 2014). Recently, glucose–insulin homeostasis has received increasing attention that motivated the design of the *artificial pancreas* (AP), encompassing a corpus of model-free and model-based control techniques aimed at regulating glucose levels (Cobelli et al., 2011). Within the context of T2D control, these control strategies have traditionally relied on insulin administration, in the absence of suitable mathematical models describing the physiological mechanisms through which exercise benefits the disease progression. Consequently, the role of physical exercise remains largely underexplored from a control perspective. Only very recently, several studies have attempted to fill this gap, proposing models that formalize the role of myokines in driving the complex physiological machinery primarily responsible for the benefits of exercise on T2D progression (De Paola et al., 2023, 2025a). In particular, these works describe the beneficial effects of Interleukin-6 (IL-6), a protein released during exercise that acts as a primary agent mediating improvements in insulin sensitivity and β-cell mass function over long-term horizons. In this paper, we propose a reinforcement learning (RL) framework that leverages this physiological model to design adaptive exercise-based control strategies to maintain glucose homeostasis over multi-year periods. Traditional model-based controllers are constrained by long-term model mismatch and the gradual evolution of patient physiology. Furthermore, to the best of our knowledge, no existing work has proposed a *model-based controller for physical-activity regulation in T2D grounded in reinforcement learning*, leaving open the question of how data-driven control methods may support personalized long-term intervention design.

1.1 Related Work

Reinforcement Learning has been successfully employed to design personalized control algorithms in biomedical sys-

⋆ Corresponding author: g.lops@phd.poliba.it

⋆⋆This work has been partially co-funded by the EU and the Swiss State Secretariat for Education, Research and Innovation, in the frame of the SHIELD project (grant No. 101156751).

tems, including diabetes management. Early work focused on closed-loop insulin control for Type 1 Diabetes (T1D), demonstrating that model-based and deep RL methods can outperform traditional PID and model predictive control (MPC) strategies (Viroonluecha et al., 2022; Emerson et al., 2023; Lops et al., 2025b,a; Baldisseri et al., 2025). These algorithms were commonly validated using the UVA/Padova simulator (Dalla Man et al., 2014) and its Python derivative *Simglucose* (Xie, 2018), achieving improved time-in-range and reduced hypoglycemic exposure. Beyond pharmacological control, RL has been increasingly used to promote *lifestyle interventions*, such as physical activity (PA) and diet. Digital-health platforms employing contextual bandits and reinforcement learning for PA recommendation in adults with T2D have demonstrated measurable increases in daily step counts and improved engagement (Yom-Tov et al., 2017; Rabbi et al., 2020; Forman et al., 2022). Complementary design research has also emphasized the role of mobile apps and wearables in strengthening patients' self-efficacy for chronic disease management (Wulfovich et al., 2019). More recently, the *PEARL* study (Lee et al., 2025) reported a large-scale randomized evaluation of an RL-based exercise assistant, validating personalization benefits over static guidance. In parallel, open simulation environments such as *StepCountJITAI* (Karine and Marlin, 2024) have been proposed for training adaptive activity interventions, while systematic reviews continue to highlight the growing integration of machine learning and RL in diabetes self-management (Mir et al., 2025). Despite these advances, most PA-focused RL studies remain short-term (6–12 weeks) and rely on behavioral or self-reported endpoints rather than metabolic indicators. We stress that, to the best of our knowledge, there is no *model-based RL control framework* for physical-activity regulation in T2D that leverages a mechanistic representation of glucose–insulin–β-cell dynamics.

Our work addresses this gap by introducing a control-oriented simulation environment that couples glucose–insulin–β-cell dynamics with IL-6-mediated adaptation, and by training a PPO controller for personalized exercise regulation over two-year trajectories.

Contributions. This study advances personalized control in T2D management through the following key contributions:

(1) Development of a custom environment implementing a five-dimensional physiological model (glucose, insulin, β-cell mass, insulin sensitivity, and IL-6) suitable for long-term simulations.
(2) Design of a robust simulation framework incorporating parameter perturbations, heterogeneous initial conditions, bounded exogenous glucose perturbations, and mid-episode degradation of target insulin sensitivity.
(3) Training of a PPO controller that learns daily exercise policies and is evaluated *in silico* over a 200-patient cohort.

Paper Organization. The remainder of the paper is structured as follows. Section 2 introduces the physiological model and its mathematical formulation. Section 3 presents the PPO-based reinforcement learning framework, including the simulation environment, uncertainty modeling, training setup, and evaluation metrics. Section 4 presents and discusses the simulation outcomes, while Section 5 concludes the paper and outlines future research directions.

2. THE CONTROL MODEL

The control-oriented physiological model employed in this work builds upon the formulation in Topp et al. (2000), extended as in De Paola et al. (2025b) to capture long-term effects of physical exercise. The resulting system describes the coupled evolution of glucose–insulin regulation and disease progression over multi-year horizons.

The model operates on a *daily time scale* and captures *fasting glucose levels* and associated metabolic variables. Intra-day dynamics and circadian effects are not explicitly modeled; the system represents long-term metabolic trends rather than short-term fluctuations.

The formulation combines the qualitative disease progression dynamics of Topp et al. (2000) with selected mechanisms from De Paola et al. (2025b), including the modulation of insulin sensitivity and β-cell dynamics via exercise-induced inflammatory responses (modeled through V_l).

This hybrid choice preserves interpretability while maintaining a compact structure suitable for reinforcement learning.

The system is defined as:

$$\dot{G} = R_0 - (E_{g0} + S_I I)\, G, \tag{1}$$

$$\dot{I} = \beta\, \sigma \frac{G^2}{\alpha + G^2} - k\, I, \tag{2}$$

$$\dot{\beta} = \big(P_{\text{prolif}}(G, V_l) - A_{\text{apopt}}(G, V_l)\big)\, \beta, \tag{3}$$

$$\dot{S}_I = -c\big(S_I - S_{I,\text{target}}\big) \left(1 - \frac{\zeta_{si} V_l}{k_{n,si} + V_l}\right), \tag{4}$$

$$\dot{V}_l = \frac{SR}{K_{IL6}}\, u - k_s V_l. \tag{5}$$

where

$$P_{\text{prolif}}(G, V_l) = (r_{1r} G - r_{2r} G^2) \left(1 + \frac{\zeta_p V_l^2}{k_p^2 + V_l^2}\right), \tag{6}$$

$$A_{\text{apopt}}(G, V_l) = (d_0 - r_{1a} G + r_{2a} G^2) \left(1 - \frac{\zeta_a V_l^2}{k_a^2 + V_l^2}\right). \tag{7}$$

Here, G and I denote blood glucose and plasma insulin concentrations; β the β-cell mass; S_I the insulin sensitivity; V_l encodes the cumulative IL-6-mediated effect of exercise on insulin sensitivity and β-cell mass.

The control input $u \in [0, 3]$ is the *equivalent exercise intensity*, representing the average daily effect of physical activity. This abstraction enables integration of lifestyle interventions without explicitly modeling individual sessions.

Following De Paola et al. (2025b), this constant input distributes the effect of repeated exercise over time, yielding a compact representation of long-term adaptations. Assuming fixed session intensity and frequency, u can be mapped to explicit prescriptions (e.g., session duration) via a static inverse relation.

Table 1. Nominal physiological parameters.

Param.	Value	Unit	Param.	Value	Unit
R_0	864	mg/dL/day	E_{g0}	1.44	day^{-1}
σ	43.2	μU/mL/day	α	2.0×10^{4}	mg^2/dL^2
k	432	day^{-1}	d_0	0.06	day^{-1}
c	0.05	day^{-1}	r_{1r}	4.2×10^{-4}	dL/mg/day
r_{2r}	1.2×10^{-6}	$\text{dL/mg}^2\text{/day}$	r_{1a}	4.2×10^{-4}	dL/mg/day
r_{2a}	1.2×10^{-6}	$\text{dL/mg}^2\text{/day}$	ζ_p	10^{-4}	–
ζ_a	10^{-4}	–	k_p	6.94×10^{2}	pg/mL·day
k_a	6.94×10^{2}	pg/mL·day	$S_{I,\text{target}}$	0.028	mL/μU/day
ζ_{si}	1.4	–	$k_{n,si}$	3.47×10^{3}	pg/mL·day
SR	64.8	pg/mL/day	K_{IL6}	5.76	day^{-1}
k_s	3.99×10^{-3}	day^{-1}			

The range $u \in [0, 3]$ corresponds to weekly exercise durations up to approximately 500 minutes of moderate-to-vigorous activity, consistent with saturation levels reported in De Paola et al. (2025a). Through V_l, the model captures exercise effects on β-cell proliferation/apoptosis and insulin sensitivity, assuming quasi-stationary fast dynamics (De Paola et al., 2023, 2025a,b). This reflects the focus on long-term adaptations rather than acute responses.

Overall, the equivalent exercise input is a modeling abstraction capturing cumulative physiological effects while preserving tractability for control design. Table 1 reports the nominal parameters adopted from Topp et al. (2000) and De Paola et al. (2025b).

3. REINFORCEMENT LEARNING CONTROLLER

This section presents the reinforcement learning framework adopted to derive personalized physical-activity strategies based on the proposed physiological model.

3.1 Overview of Reinforcement Learning

Reinforcement Learning (RL) provides a framework for sequential decision-making under uncertainty (Sutton and Barto, 2018). At each time step t, an agent observes the system state s_t, selects an action a_t according to a policy $\pi_\theta(a_t|s_t)$, and receives a scalar reward r_t from the environment. The objective is to find policy parameters θ that maximize the expected cumulative discounted return:

$$J(\theta) = \mathbb{E}_{\pi_\theta}\left[\sum_{t=0}^{T} \gamma^t r_t\right] \tag{8}$$

where $\gamma \in (0, 1]$ controls the emphasis on long-term rewards. Policy-gradient methods optimize $J(\theta)$ via stochastic gradient ascent using sampled trajectories, avoiding explicit modeling of system dynamics.

3.2 Proximal Policy Optimization (PPO)

Among policy-gradient algorithms, Proximal Policy Optimization (PPO) (Schulman et al., 2017) achieves a balance between stability and sample efficiency in continuous-control tasks. PPO constrains the update between successive policies π_θ and $\pi_{\theta_{\text{old}}}$ by maximizing a clipped surrogate objective:

$$L^{\text{CLIP}}(\theta) = \mathbb{E}_t\Big[\min\big(r_t(\theta)\hat{A}_t,\ \text{clip}(r_t(\theta), 1-\epsilon, 1+\epsilon)\hat{A}_t\big)\Big] \tag{9}$$

where

$$r_t(\theta) = \frac{\pi_\theta(a_t|s_t)}{\pi_{\theta_{\text{old}}}(a_t|s_t)} \tag{10}$$

is the probability ratio and $\hat{A}_t$ is the advantage estimate, computed via Generalized Advantage Estimation (GAE), which reduces variance while preserving low bias in long-horizon return estimation.

The critic network $V_\phi(s_t)$ approximates the value function and is trained jointly with the actor. Entropy regularization is used to encourage exploration, while normalization of observations and rewards improves numerical stability.

3.3 Simulation Environment

Each training episode simulates a two-year (730-day) horizon with a daily time step. The observation vector is

$$o_t = [\, G_{\text{obs}}(t),\ t/730 \,], \tag{11}$$

where $G_{\text{obs}} \in [0, 1000]\,\text{mg/dL}$ denotes the observable glucose concentration. The control action $u \in [0, 3]$ represents the daily equivalent physical activity intensity. System dynamics are integrated using a Dormand–Prince (DOPRI5) adaptive solver, ensuring accurate trajectories over long horizons. This formulation provides a consistent mapping between daily control inputs and long-term physiological evolution. The agent–environment interaction is illustrated in Fig. 1.

Uncertainty and Disturbances To assess controller robustness under physiological variability, each episode introduces: *(i)* parameter perturbations within $\pm5\%$ of nominal values; *(ii)* randomized initial conditions $G_0 \in [70, 115]\,\text{mg/dL}$ and $I_0 \in [5, 20]$; *(iii)* bounded exogenous perturbations on glucose dynamics (amplitude $\pm20\,\text{mg/dL}$), capturing daily variability without explicitly modeling circadian mechanisms; and *(iv)* a gradual reduction (up to 5%) of $S_{I,\text{target}}$ to represent long-term metabolic decline. The chosen $\pm5\%$ variability provides controlled deviations around nominal conditions while preserving numerical stability over long simulation horizons.

Reward formulation The reward function modulates the controller's behaviour by encouraging higher intervention intensity during an initial 150-day adaptation phase and more conservative actions thereafter:

$$r_t = \begin{cases} 2\,\mathbb{1}\{\,1.0 \le u_t \le 2.5\,\}, & t \le 150, \\ \mathbb{1}\{\,u_t \le 1.0\,\}, & t > 150. \end{cases} \tag{12}$$

This shaping-based formulation promotes structured intervention patterns over time but does not explicitly penalize metabolic states.

Consequently, the learned policy reflects the imposed temporal incentives rather than direct metabolic optimization, highlighting the role of reward design.

3.4 Implementation Details

The controller was implemented using the Stable-Baselines3 PPO framework (Raffin et al., 2021). Both the actor and critic share a multilayer perceptron architecture with two hidden layers of 64 units and ReLU activations. Table 2 reports the main hyperparameter configuration.

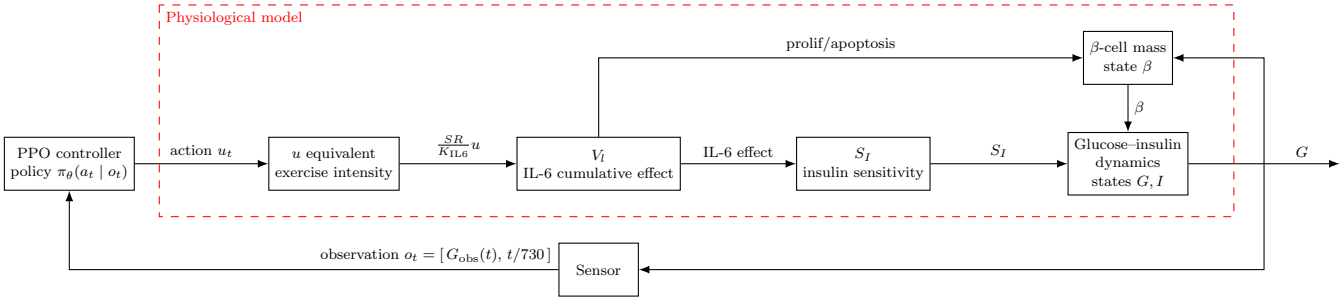

Fig. 1. Reinforcement learning closed-loop structure. A PPO controller implements the policy $\pi_\theta(a_t \mid o_t)$, receiving the observation $o_t = [G_{\text{obs}}(t), t/730]$ and generating the daily action u_t (equivalent exercise intensity). The dashed block represents the nonlinear physiological model, where V_l is the cumulative IL-6 effect, S_I the insulin sensitivity, β the β-cell mass, and (G, I) the glucose–insulin subsystem. The glucose concentration G is both the observable output and part of the internal feedback within the model.

Table 2. PPO controller architecture and hyperparameters.

Category	Parameter / Symbol	Value
Architecture		
Policy	`MlpPolicy`	2×64 (ReLU) shared actor–critic
Algorithmic		
Learning rate	α	3×10^{-4}
Discount factor	γ	0.99
GAE parameter	λ	0.95
Clip range	ϵ	0.2
Entropy coeff.	β_{ent}	0.01
Exploration	`use_sde`	True (freq. 4)
Init. log-std	$\log(\sigma_0)$	0.5
Training setup		
Rollout length	n_{steps}	2048
Batch size	B	64
Epochs/update	n_{epochs}	10
Normalization	`VecNormalize`	Obs. + reward

Training was performed on 500 simulated patients, each spanning a two-year horizon (730 days), to capture long-term physiological adaptation under uncertainty.

3.5 Evaluation Protocol

Evaluation was performed using fixed normalization statistics obtained during training (Raffin et al., 2021), ensuring consistent scaling of observations and rewards. The trained controller was deployed in deterministic mode to assess steady-state performance. An independent cohort of 200 virtual patients was simulated.

Each cohort incorporates the same sources of uncertainty used during training, including parameter perturbations (±5%), randomized initial conditions, bounded exogenous perturbations on glucose dynamics, and gradual degradation of insulin sensitivity. This setup mirrors the training distribution and enables assessment of policy generalization under controlled variability.

3.6 Performance Metrics

The primary endpoint is the fraction of patients achieving a final fasting glucose concentration below 126 mg/dL after the 730-day simulation:

$$P_{\text{norm}} = \frac{1}{N}\sum_{i=1}^{N} \mathbb{1}\{G_i(T) < 126\}. \tag{13}$$

This threshold corresponds to the diagnostic boundary for diabetes defined by the *American Diabetes Association (ADA)* and the *World Health Organization (WHO)* (American Diabetes Association Professional Practice Committee, 2025; World Health Organization, 1999).

In addition to this binary endpoint, secondary metrics are used to assess stability and physiological plausibility of the learned control policy:

- mean and standard deviation of final glucose levels;
- glucose variability during the final 100 days (range and 95th percentile);
- action smoothness, quantified as the variance of daily control inputs.

4. RESULTS

This section evaluates the PPO controller introduced in Section 3 on an independent cohort of $N = 200$ virtual patients. All testing conditions, including ±5% parameter perturbations, randomized initial states, bounded exogenous glucose perturbations, and mid-episode reductions in insulin sensitivity, follow the protocol of Section 3.3. During evaluation, the policy acts deterministically using the fixed normalization statistics computed at training time.

4.1 Clinical Endpoint: Normoglycemia Achievement

The primary clinical endpoint is the proportion of individuals achieving a final fasting glucose below 126 mg/dL after the 730-day horizon. The PPO controller attains a *66% success rate*.

Figure 2 exhibits a distinctly bimodal distribution: one subgroup converges near the clinical target, whereas another stabilizes at markedly elevated values (500–650 mg/dL). This separation reflects heterogeneous metabolic responses arising from parameter perturbations and long-term physiological drift. The emergence of this second mode indicates that a subset of virtual patients is driven toward a deteriorating metabolic regime not sufficiently counteracted by the learned policy. This behavior

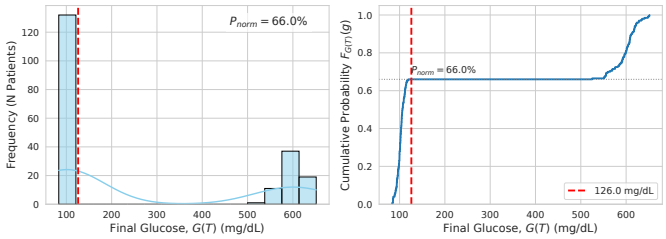

Fig. 2. Final glucose distribution across the evaluation cohort: (a) histogram of $G(T)$ with diabetes threshold indicator; (b) empirical cumulative distribution function (CDF) of $G(T)$.

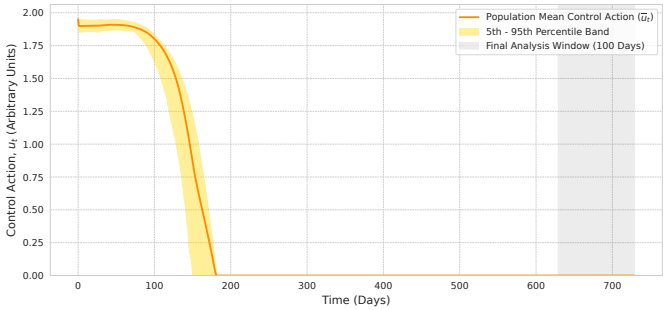

Fig. 3. Population-level trajectories of recommended daily physical activity.

is associated with unfavorable combinations of initial conditions and parameter realizations within the uncertainty set. The empirical CDF further highlights this structure. A sharp rise to $P_{\text{norm}} = 66\%$ occurs at the clinical threshold, followed by a plateau with no intermediate outcomes. The remaining probability mass accumulates at high glucose values, confirming the presence of a pathological cluster failing to achieve glycemic control.

4.2 Action Patterns and Behavioural Adaptation

Because the reward does not explicitly include metabolic state variables, behavioural adaptation is primarily driven by the imposed temporal incentives. Figure 3 shows that during the early phase ($t \leq 150$), the agent recommends moderate activity ($u_t \approx 1.7$–1.9). After the reward reshaping, the policy converges toward minimal action, reaching $u_t \approx 0$ for nearly all patients by day 200. The tight clustering of actions indicates stable PPO updates and consistent policy behavior across the population. However, the lack of state-dependent penalties implies that similar control inputs are applied across heterogeneous physiological conditions.

4.3 Performance Metrics

Figure 4 reports secondary indicators for the final 100 days. Panel (a) shows that intra-patient variability remains low (typically 10–20 mg/dL), indicating stable long-term dynamics without oscillatory behavior. Panel (b) highlights the separation in 95th-percentile glucose values between controlled and uncontrolled individuals, consistent with the bimodal final-glucose distribution. Table 3 shows that cohort-level statistics are strongly influenced by the separation between the two subgroups. Normoglycemic

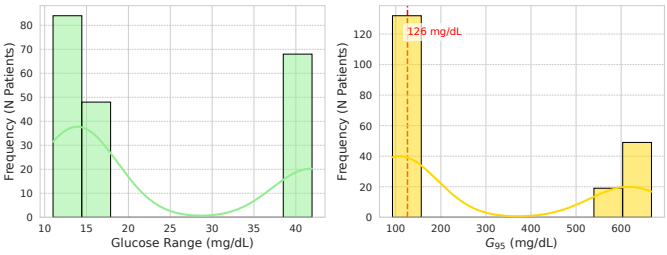

Fig. 4. Performance metrics over the final 100 simulation days: (a) glucose-range ΔG; (b) 95th-percentile glucose $P_{95}(G)$.

Table 3. Glucose outcomes and performance metrics across the evaluation cohort.

Metric	Value (Mean ± SD)	Unit
A. Entire cohort		
Initial glucose $G(0)$	92.6 ± 14.0	mg/dL
Final glucose $G(T)$	270.9 ± 236.6	mg/dL
Normoglycemia success	66%	–
B. Normoglycemic subgroup ($G(T) < 126$ mg/dL)		
Final glucose $G(T)$	92.5 ± 15.2	mg/dL
Glucose range ΔG	17.2 ± 6.1	mg/dL
95th percentile G_{95}	118.4 ± 19.7	mg/dL
C. Dysglycemic subgroup ($G(T) \geq 126$ mg/dL)		
Final glucose $G(T)$	585.3 ± 42.8	mg/dL
Glucose range ΔG	35.8 ± 12.4	mg/dL
95th percentile G_{95}	612.1 ± 48.3	mg/dL

patients reach near-target levels (92.5 ± 15.2 mg/dL), with a narrow glucose range of $\Delta G \approx 17$ mg/dL, indicating stable closed-loop behaviour. Patients who fail to reach normoglycemia stabilise at markedly elevated glucose (≈ 585 mg/dL) and display larger variability ($\Delta G \approx 36$ mg/dL). This divergence highlights a limitation of the current formulation: the reward does not penalize adverse metabolic states, allowing trajectories with extreme hyperglycemia to persist. Addressing this limitation through state-dependent penalties or safety constraints represents a key direction for improving clinical plausibility.

5. CONCLUSION

This work presented a reproducible reinforcement-learning framework for long-horizon regulation of glucose dynamics in a nonlinear and uncertain model of Type 2 Diabetes progression. By incorporating parameter perturbations, bounded exogenous glucose disturbances, and progressive metabolic drift, the proposed environment enabled the evaluation of closed-loop robustness under physiologically motivated uncertainty. A PPO controller trained through behavioural shaping achieved stable, low-variance behaviour and reached normoglycaemia in 66% of a 200-patient virtual cohort, while producing smooth and coherent low-intensity physical-activity actions over multi-year horizons. The results suggest that reinforcement learning can generate structured physical-activity policies for long-term metabolic regulation. However, the dysglycaemic subgroup observed in the evaluation highlights the need for explicit state-dependent safety penalties and more clinically targeted reward design. Future work will extend the framework toward higher-fidelity physiological mod-

els, broader inter-patient variability, safety-aware learning mechanisms, and benchmarking against established clinical and model-based control strategies.

REFERENCES

American Diabetes Association Professional Practice Committee (2025). 2. diagnosis and classification of diabetes: Standards of care in diabetes—2025. *Diabetes Care*, 48(Supplement$_1$), $S27--S49.doi : 10.2337/dc25-S002$.

Baldisseri, F., Atanasious, M.M.H., Becchetti, V., Di Paola, A., Lops, G., Menegatti, D., Wrona, A., Mascolo, S., and Priscoli, F.D. (2025). A quantitative comparison of deep reinforcement learning algorithms for type 1 diabetes control. In *2025 11th International Conference on Control, Decision and Information Technologies (CoDIT)*, volume 1, 1073–1078. doi: 10.1109/CoDIT66093.2025.11321730.

Cobelli, C., Renard, E., and Kovatchev, B. (2011). Artificial pancreas: past, present, future. *Diabetes*, 60(11), 2672–2682.

Dalla Man, C., Micheletto, F., Lv, D., Breton, M., Kovatchev, B., and Cobelli, C. (2014). The uva/padova type 1 diabetes simulator: New features. *Journal of Diabetes Science and Technology*, 8(1), 26–34. doi:10.1177/1932296813514502. URL `https://doi.org/10.1177/1932296813514502`.

De Paola, P.F., Borri, A., Dabbene, F., Keshavjee, K., Palumbo, P., and Paglialonga, A. (2025a). Modeling the cumulative benefits of regular physical activity on type 2 diabetes progression. *Computers in Biology and Medicine*, 198, 111194.

De Paola, P.F., Borri, A., Paglialonga, A., Palumbo, P., and Dabbene, F. (2025b). A model-based approach for glucose control via physical activity. *arXiv preprint arXiv:2503.09435*.

De Paola, P.F., Paglialonga, A., Palumbo, P., Keshavjee, K., Dabbene, F., and Borri, A. (2023). The long-term effects of physical activity on blood glucose regulation: a model to unravel diabetes progression. *IEEE Control Systems Letters*, 7, 2916–2921.

Emerson, H., Guy, M., and McConville, R. (2023). Offline reinforcement learning for safer blood glucose control in people with Type 1 Diabetes. *Journal of Biomedical Informatics*, 144, 104376. doi:10.1016/j.jbi.2023.104376.

Forman, E.M., Kerrigan, S.G., Butryn, M.L., Crosby, R.D., Herbert, J.D., and Lowe, M.R. (2022). Personalized digital health behavior change interventions: Reinforcement Learning framework and clinical application. *Annals of Behavioral Medicine*, 56(5), 498–511. doi: 10.1093/abm/kaab094.

Karine, A. and Marlin, B. (2024). Stepcountjitai: Simulation environment for reinforcement learning with application to physical activity adaptive intervention. URL `https://arxiv.org/abs/2411.00336`.

Lee, A., Zhang, Y., and Nemati, S. (2025). Pearl: A personalized exercise assistant using reinforcement learning — results from a four-arm randomized controlled trial. URL `https://arxiv.org/abs/2508.10060`.

Lops, G., Manfredi, G., Racanelli, V.A., De Cicco, L., and Mascolo, S. (2025a). A safety aware deep reinforcement learning technique for automated insulin delivery. In *2025 33rd Mediterranean Conference on Control and Automation (MED)*, 120–125. doi: 10.1109/MED64031.2025.11073297.

Lops, G., Racanelli, V.A., Manfredi, G., De Cicco, L., and Mascolo, S. (2025b). A deep deterministic policy gradient control algorithm for automatic insulin delivery. *IFAC-PapersOnLine*, 59(2), 13–18. doi: https://doi.org/10.1016/j.ifacol.2025.06.003.

Mir, S., Ali, M., Shah, S.A., Alqahtani, T., and Hussain, F. (2025). Impact of machine learning on dietary and exercise behaviours in type 2 diabetes self-management: A systematic literature review. *PeerJ Computer Science*, 11, e2568. doi:10.7717/peerj-cs.2568.

Rabbi, M.M., Aung, M.M., Zhang, M., and Choudhury, T. (2020). Smartphone-based reinforcement learning to promote physical activity in diabetes and depression: Pilot study. *JMIR mHealth and uHealth*, 8(8), e17807. doi: 10.2196/17807.

Raffin, A., Hill, A., Gleave, A., Kanervisto, A., Ernestus, M., and Dormann, N. (2021). Stable-baselines3: Reliable reinforcement learning implementations. *Journal of Machine Learning Research*, 22(268), 1–8. URL `http://jmlr.org/papers/v22/20-1364.html`.

Schulman, J., Wolski, F., Dhariwal, P., Radford, A., and Klimov, O. (2017). Proximal policy optimization algorithms. URL `https://arxiv.org/abs/1707.06347`.

Sutton, R.S. and Barto, A.G. (2018). *Reinforcement Learning: An Introduction*. MIT Press, Cambridge, MA.

Topp, B., Promislow, K., Devries, G., Miura, R.M., and T FINEGOOD, D. (2000). A model of β-cell mass, insulin, and glucose kinetics: pathways to diabetes. *Journal of theoretical biology*, 206(4), 605–619.

Viroonluecha, P., Egea-Lopez, E., and Santa, J. (2022). Evaluation of blood glucose level control in Type 1 Diabetic patients using deep reinforcement learning. *PLOS ONE*, 17(9), e0274608. doi:10.1371/journal.pone.0274608.

World Health Organization (1999). Definition, diagnosis and classification of diabetes mellitus and its complications: Report of a who consultation. Technical report, World Health Organization, Geneva. URL `https://apps.who.int/iris/handle/10665/66040`. Part 1: Diagnosis and Classification of Diabetes Mellitus.

Wulfovich, S., Fiordelli, M., Rivas, H., Concepcion, W., and Wac, K. (2019). "I Must Try Harder": Design implications for mobile apps and wearables contributing to self-efficacy of patients with chronic conditions. *Frontiers in Psychology*, 10, 2388. doi:10.3389/fpsyg.2019.02388.

Xie, J. (2018). Simglucose v0.2.1. `https://github.com/jxx123/simglucose`. [Online]. Available: https://github.com/jxx123/simglucose.

Yom-Tov, E., Feraru, G., Kozdoba, M., Mannor, S., Tennenholtz, M., and Hochberg, I. (2017). Encouraging physical activity in patients with diabetes: Reinforcement Learning–based recommendations. *Journal of the American Medical Informatics Association*, 24(2), 320–326. doi: 10.1093/jamia/ocw127.

Zimmet, P.Z., Magliano, D.J., Herman, W.H., and Shaw, J.E. (2014). Diabetes: a 21st century challenge. *The lancet Diabetes & endocrinology*, 2(1), 56–64.