A Deep Deterministic Policy Gradient control algorithm for Automatic Insulin Delivery

Giada Lops, Vito Andrea Racanelli, Gioacchino Manfredi, Luca De Cicco, Saverio Mascolo

Abstract: This study investigates the use of Reinforcement Learning (RL) to optimize insulin infusion in Automated Insulin Delivery (AID) systems. The Deep Deterministic Policy Gradient (DDPG) algorithm is evaluated as a potential control strategy and compared to Model Predictive Control (MPC) and PID controllers using a highly realistic *in silico* simulation model. Simulations for adult Type 1 Diabetes (T1D) patients demonstrate that the DDPG controller effectively reduces hypo- and hyperglycemic episodes and improves glucose stability.

Keywords: New control system algorithms for Automatic Insulin Delivery, AI for Control and Optimization, Diabetes Management, Optimized Insulin Delivery, Reinforcement Learning, AID Systems, MPC, PID, DDPG

1. INTRODUCTION

Diabetes is a chronic disease characterized by high blood sugar levels, leading to long-term damage to the heart, blood vessels, eyes, kidneys, and nerves. Type 2 diabetes, the most common form, occurs when the body becomes resistant to insulin or does not produce enough to maintain normal glucose levels. Its prevalence has been rising sharply worldwide. On the other hand, type 1 diabetes is an autoimmune condition in which the immune system attacks insulin-producing beta cells in the pancreas, leading to little or no insulin production. As a result, patients with type 1 diabetes require lifelong insulin therapy to regulate blood sugar levels. Nowadays, 422 million people, primarily in low- and middle-income countries, have diabetes, with 1.5 million deaths attributed to it each year. Due to the constant growth of cases, a global target to halt the rise in diabetes and obesity by 2025 has been set [WHO (2024)]. Diabetes management aims at maintaining blood glucose levels near normal while avoiding hypoglycemia or hyperglycemia. By achieving well-controlled blood glucose levels, individuals can, in fact, significantly reduce the frequency and severity of complications. Moreover, an effective management helps prevent long-term health issues and improves overall quality of life.

Automated Insulin Delivery (AID) systems monitor blood sugar levels and deliver insulin to keep them within a target range, mainly helping patients with type 1 diabetes.. These systems consist of a Continuous Glucose Monitor (CGM), a control algorithm that calculates insulin needs, and a pump to administer the dose, as shown in Fig. 1.

Some systems are fully automated (e.g. Fully Closed Loop, FCL, see Boughton (2021)), while others need user input for adjustments (e.g. Predictive low-glucose suspend, PLGS, see Joubert et al. (2022)).

⋆ Giada Lops, Vito Andrea Racanelli, Gioacchino Manfredi, Luca De Cicco, and Saverio Mascolo are with the Dipartimento di Ingegneria Elettrica e dell'Informazione at Politecnico di Bari, Via Orabona 4, 70125, Bari, Italy.
Emails: g.lops@phd.poliba.it, name.surname@poliba.it

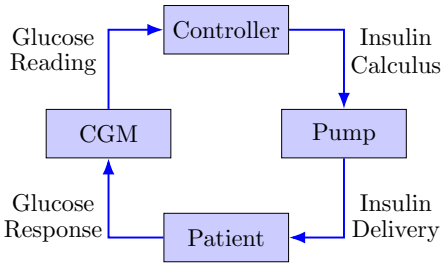


Fig. 1. Automatic Insulin Delivery closed-loop system.

The algorithms used in commercial AID system are generally undisclosed. However, in open-source systems (e.g. OpenAPS (2015)), both the code and the algorithm are publicly accessible. In general, all algorithms perform the same basic function: they collect data from the CGM and, based on predicted glucose levels and the user's personal settings (such as basal rates, insulin sensitivity, and carbohydrate-to-insulin ratios), recommend the insulin dose needed to maintain or achieve desired glucose levels. Depending on the system, users and physicians may be able to adjust the target glucose range and set different parameters to regulate the overall insulin delivery.

2. RELATED WORKS

The automatic bolus correction system for artificial pancreas control is specifically designed to improve patient safety and enhance the effectiveness of the insulin pump by ensuring an optimized insulin infusion that dynamically responds to the patient's changing needs. By employing an appropriately designed controller, the system monitors the patient's condition and activities, thereby detecting any errors or underestimations in the bolus dosage during meals.

Currently, the literature highlights several control algorithms for automating insulin delivery such as Proportional-Integral-Derivative (PID, [Steil et al. (2006)]), Model Predictive Control (MPC, [Hovorka et al. (2010)]), Hypoglycemic Predictive Algorithms (HPA, [Dassau et al.

(2010)]), Fuzzy Logic [Nimri et al. (2014)], and Neural Networks [Pérez-Gandía et al. (2010)]. Importantly, these algorithms operate in such a way that they compute the insulin infusion rate as a function of both time and the continuously evolving glucose concentration, thus ensuring that the insulin delivery remains effective and responsive to the patient's real-time metabolic state.

Among all the control algorithms mentioned, the key differences lie in how they consider factors such as insulin sensitivity, carbohydrate intake, physical activity, stress levels, and age. The PID algorithm offers simple, real-time control adjustments and is well-suited for systems with linear dynamics, but it struggles with complex, non-linear dynamics like those seen in large glucose fluctuations post-meals or during physical activity, though it is still used in commercial systems, i.e. Medtronic MiniMed 670G, 770G, and 780G. In contrast, Model Predictive Control (MPC) algorithms excel at managing complex dynamics and handling factors like insulin absorption delays and glucose measurement inaccuracies. MPC's ability to solve constrained optimization problems over a time horizon makes it widely used in systems such as CamAPS FX, iAP, Diabeloop DBLG1, Tandem CONTROL IQ, and Insulet Omnipod 5 [Thomas and Heinemann (2022)]. Given the multiple variables influencing patient conditions, machine learning techniques, particularly neural networks (ANN), may play an essential role in training systems to make more personalized decisions.

Reinforcement Learning (RL) has shown success in recent AID control algorithms. In particular, DDPG is an off-policy reinforcement learning algorithm specifically designed for continuous control tasks. It was introduced by Lillicrap et al. (2019) and has gained significant attention due to its capability to handle high-dimensional action spaces using deep neural networks. The robustness of DDPG has made it a preferred choice for various applications in robotics and control systems, surpassing the performance of other algorithms in several scenarios [Sumiea et al. (2024), Tan (2021)]. Specifically, DDPG has been also successfully applied in training humanoid robots [Chun et al. (2023)], drones [Bouhamed et al. (2020)], and 7 Degrees of Freedom (DOF) surgical robots [Hou et al. (2021)].

Baldisseri et al. (2024) discusses a significant study on insulin regulation for type 1 diabetes, utilizing DDPG and the Bergman (2003) model of glucose metabolism. Compared to traditional methods (PID, Nonlinear MPC), this approach demonstrated superior computational efficiency. Atanasious et al. (2024) utilizes DDPG to demonstrate the potential of RL in maintaining normoglycemia in virtual T1D patients without the direct knowledge of the amount of ingested carbohydrates.

The main contribution of this study is to employ the DDPG algorithm to optimize insulin delivery in AID systems, offering a personalized treatment approach based on a highly realistic model for T1D and to benchmark results against PID and MPC algorithms.

The rest of the manuscript is structured as follows: Section 3 details the proposed control methodologies and the selected simulation scenario; Section 4 presents the *in silico* simulations conducted to validate the proposed controllers; lastly, Section 5 concludes and suggests future research directions.

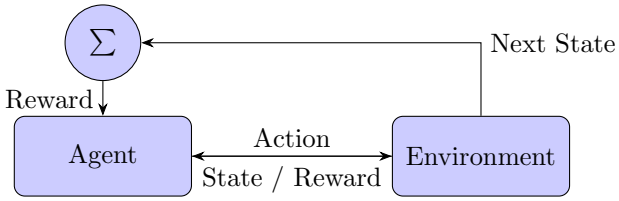


Fig. 2. General scheme of Reinforcement Learning.

3. REINFORCEMENT LEARNING FOR AID CONTROL APPROACH

RL refers to machine learning techniques where an agent learns a strategy to maximize rewards through environment interaction without explicit action instructions [AlMahamid and Grolinger (2024)]. The agent receives feedback, positive or negative, based on its actions, echoing psychological principles (e.g., Skinner's experiments on mice and pigeons, see Leeder (2022)) and cybernetic roots, first introduced by Marvin Minsky (1954).

RL is related to dynamic programming and optimal control but does not rely on a predefined environmental model. A model derived from human interaction can complement RL in what is known as *reinforcement learning through human feedback* (RLHF) [Hugging Face (2023)].

The mathematical foundation of RL revolves around the agent, environment, states, actions, and rewards. The environment is modeled as a *Markov Decision Process* (MDP), (S, A, ρ, P_0), where S is the set of states, A the set of actions, ρ the dynamic system, and P_0 the initial distribution.

At each step $t \in \mathbb{N}_0$, the agent selects an action A_t based on a policy π_t, receives a reward R_t, and makes a transition to the next state S_{t+1}, as shown in Fig. 2. The agent's goal is to maximize the expected return $\mathbb{E}[G_t]$ of the discounted sum of rewards:

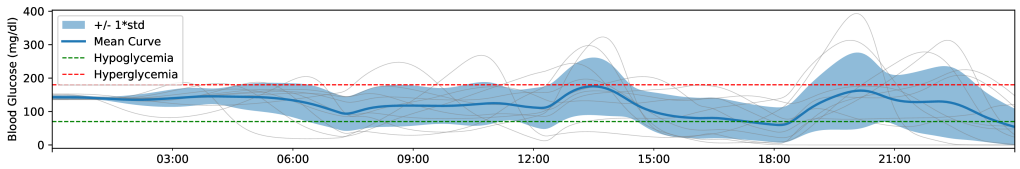
$$\mathbb{E}[G_t] = \mathbb{E}\left[\sum_{k=0}^{T} \gamma^k R_{t+k}\right] \quad \text{with} \quad 0 \leq \gamma \leq 1 \qquad (1)$$

where γ is the discount factor. In finite-horizon problems ($T \in \mathbb{N}$), $\gamma = 1$ treats future rewards equally, while in infinite-horizon problems ($T = \infty$), $\gamma < 1$ ensures convergence. When $\gamma = 0$, only immediate rewards are considered; as γ approaches 1, the agent becomes more future-oriented.

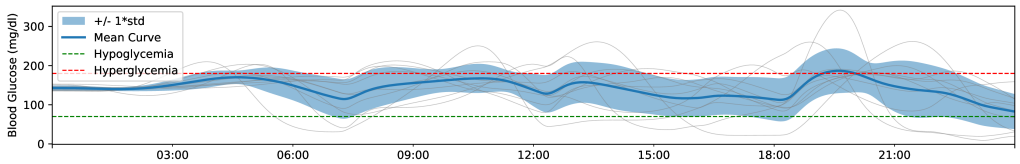
Deep Deterministic Policy Gradient (DDPG) This algorithm operates with four distinct neural networks: the *Actor Current*, *Actor Target*, *Critic Current*, and *Critic Target* [Sumiea et al. (2024)].

The *Actor Current Network* is responsible for updating policy parameters and selecting actions based on the current state of the system, while the *Actor Target Network* selects optimal actions based on the next state. The *Critic Current Network* estimates the Q-value, representing the expected return of state-action pairs, whereas the *Critic Target Network* computes the target Q-value.

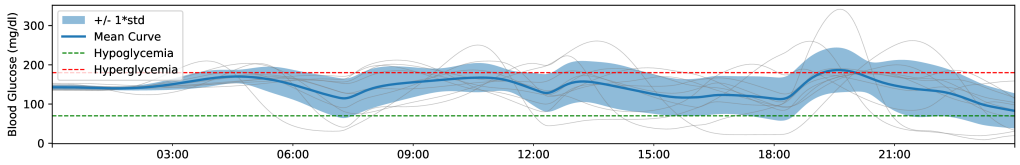
These target networks are periodically updated through a mechanism, where the parameters of the target networks are adjusted using a weighted average of the current networks' parameters. The structure of DDPG enables efficient updates of both the policy and value functions while leveraging experience replay and target networks



(a) PID Controller



(b) MPC Controller



(c) DDPG Controller

Fig. 3. Blood glucose traces for the PID (a), MPC (b), and DDPG (c) controllers: light gray lines represent the individual glucose curves for each of the 10 patients.

to enhance training stability. The algorithm operates in a loop over episodes, during which actions are selected, transitions are stored in a replay buffer, and both the actor and critic networks are updated using minibatches of sampled experiences. The target networks are updated using a soft update rule, controlled by the parameter τ.

Algorithm Overview: During each episode, the agent interacts with the environment by selecting actions based on the current state, using an exploration strategy that incorporates noise $\mathcal{N}$. The transitions (s_t, a_t, r_t, s_{t+1}) are stored in the replay buffer D, and at each time step, a minibatch of transitions is sampled to update the critic and actor networks. The critic network is updated by minimizing the loss function, while the actor network is updated using the policy gradient method. The process continues, as outlined in Algorithm 1, until either T steps have been completed in the episode or an irreversible state is reached. In the algorithm, θ represents the trainable parameters of the neural networks: θ^Q corresponds to the critic network $Q(s, a|\theta^Q)$, while θ^μ corresponds to the actor network $\mu(s|\theta^\mu)$.

Simulation scenario: To develop effective Reinforcement Learning strategies, a realistic simulation scenario is essential to accurately capture the complex dynamics of glucose metabolism. To the purpose, the SimGlucose simulator [Xie (2018)] has been integrated with the FDA-approved UVA/Padova model [Dalla Man et al. (2007)] which simulates the physiological characteristics of adults with type 1 diabetes. In total, ten adults have been considered. A single policy was trained for all 10 adults, each with their specific insulin sensitivity provided by the simulator. The scenario specifics are detailed in Table 1. To ensure variability across episodes and improve policy generalization, the meal schedule was randomized by introducing a standard deviation of 30 minutes in meal timing and

Algorithm 1 DDPG Algorithm

```
Initialize critic network Q(s, a|θ^Q), actor network μ(s|θ^μ), and target networks with random weights
Initialize replay buffer D
for episode = 1 to M do
    Initialize random process 𝒩 for exploration, receive initial state s_1
    for time step = 1 to T do
        Select action a_t = μ(s_t|θ^μ) + 𝒩_t
        Execute action, observe reward r_t and new state s_{t+1}
        Store transition (s_t, a_t, r_t, s_{t+1}) in D
        Sample a minibatch of N transitions from D
        Update critic by minimizing loss ℒ = (1/N) Σ_i (y_i − Q(s_i, a_i|θ^Q))^2
        Update actor using policy gradient
        Update target networks: θ^{Q'} ← τθ^Q + (1 − τ)θ^{Q'}
    end for
end for
```

±10 g in carbohydrate intake per meal. This resulted in a total daily carbohydrate intake ranging from 170 g to 270 g for each patient, ensuring diversity in meal scenarios across the total training episodes. To tackle the issue of

Table 1. Simulation Settings

Parameter	**Value**
Sensor	Dexcom
Pump	Insulet
Meal schedule	
7:00 AM	45 grams of carbohydrates
12:00 PM	70 grams of carbohydrates
4:00 PM	15 grams of carbohydrates
6:00 PM	80 grams of carbohydrates
11:00 PM	10 grams of carbohydrates

safely regulating blood glucose levels, we created a custom environment suitable for the Rllab library by Duan et al. (2016). This environment models the blood glucose dynamics over a daily cycle with a time step of $\Delta t = 3\,\text{min}$, dictated by the sensor's sampling interval. The observation space is defined by: $\mathcal{O} = \{\text{BG} \in [0, 500]\,\text{mg/dL}\}$, where BG represents the blood glucose level at each step. The action space corresponds to the insulin dose: $\mathcal{A} = \{\textit{Insulin Dose} \in [0.0, 1.5]\,\text{units/min}\}$.

The agent's behavior is guided by a discrete reward function, which encourages maintaining blood glucose within a specified range. Given $T_{\text{low}} = 70\,mg/dL$ and $T_{\text{high}} = 180\,mg/dL$, the reward function is defined as follows:

$$r_t(\text{BG}) = \begin{cases} -3, & \text{if BG} < T_{\text{low}} \\ -1, & \text{if } T_{\text{low}} \leq \text{BG} \leq 90 \\ 2, & \text{if } 90 < \text{BG} < T_{\text{high}} \\ -1, & \text{if } T_{\text{high}} \leq \text{BG} < 200 \\ -2, & \text{if BG} \geq 200 \end{cases} \tag{2}$$

The results were obtained over a total of 250 episodes, i.e., simulated days, with each episode representing a full 24-hour period divided into 480 time steps at intervals of 3 minutes each. Moreover, our DDPG controller determines insulin dosages based on current inputs, employing the Adam optimizer to minimize the loss function.
Action noise is introduced during training to foster exploration, using Gaussian noise ($\sigma = 0.1$), as it shows similar performance to Ornstein-Uhlenbeck noise [OpenAI (2021)]. The performance of the DDPG controller was tested on the 10 patients, ensuring that the test meal schedule, randomly selected from the unseen meal schedules outlined in Table 1, was the same for all 10 patients. The results showed in Figure 3c illustrate glucose control outcomes, emphasizing effective management during both fasting and meals.

Proportional–integral–derivative controller (PID): In this work, the PID controller parameters were tuned using the Optuna library to tune the gain values K_p, K_i, and K_d over 100 iterations. The optimization criterion was the Risk Index (RI), described in (3), which measures the risk of hypoglycemia (LBGI) and hyperglycemia (HBGI)

$$RI = LBGI + HBGI, \tag{3}$$

The successful trial identified $K_p = 2.120 \times 10^{-3}$, $K_i = 1.100 \times 10^{-6}$, and $K_d = 3.004 \times 10^{-3}$ as the optimal gains. The simulations maintained normal blood glucose levels for most of the day, but out of 10 simulated patients, as shown in Fig. 3a, some experienced extreme glucose levels (below 70 mg/dL or above 250 mg/dL), highlighting the need for further optimization.

Model Predictive Control (MPC): MPC is an advanced method for controlling complex processes with multiple variables, based on predicting the future behavior of a system using a dynamic process model. The goal is to optimize control inputs to achieve desired outputs while satisfying constraints on inputs, outputs, and system states Hovorka et al. (2010). At each time step, MPC solves an optimization problem over a prediction horizon, determining the best control action by minimizing a cost function subject to constraints. It is used extensively in industries requiring high-quality control, such as power plants, waste management, and manufacturing, where traditional controllers (e.g. PID or *fuzzy logic*) fall short.

The MPC can be formulated as a constrained receding time horizon control as follows:

$$\min_{\mathbf{u}} \sum_{i=0}^{N-1} \left(\|\mathbf{y}(k+i) - \mathbf{y}_{\text{ref}}(k+i)\|_Q^2 + \|\mathbf{u}(k+i)\|_R^2 \right) \tag{4}$$

subject to:

$$\mathbf{x}_{\min} \leq \mathbf{x}(k+i) \leq \mathbf{x}_{\max}, \quad \mathbf{u}_{\min} \leq \mathbf{u}(k+i) \leq \mathbf{u}_{\max},$$

where N is the prediction horizon, y_{ref} is the reference for the outputs, and Q and R are weighting matrices Baldisseri et al. (2024), Schwenzer et al. (2021).

In our development case, we chose to implement a cost function that severely penalizes hypoglycemia, $G \leq 70\,mg/dL$ (G_{low}), through λ and general deviations from the target of 140 mg/dL (G_{target}). This approach is specifically designed to prevent potential *hypoglycemic crises* in the patient, particularly during nighttime fasting or before meals. The chosen prediction horizon $N = 60$ corresponds to 3 hours, which aligns with the typical full digestion time after a meal, ensuring that the insulin administration model accounts for post-meal glucose fluctuations.

The cost function is formulated as follows:

$$\begin{aligned} \min_{u} \sum_{i=0}^{N-1} & \left(\|G(k+i) - G_{\text{target}}\|_Q^2 + \|u(k+i)\|_R^2 \right) \\ & + \lambda \cdot \max(0, G_{\text{low}} - G(k+i))^2 \end{aligned} \tag{5}$$

subject to:

$$\begin{gathered} G_{\min} \leq G(k+i) \leq G_{\max}, \\ |\Delta u(k+i)| \leq 1.50\,\text{U/min}, \\ 0 \leq u_{\text{basal}}(k+i) \leq 0.25\,\text{U/min}, \end{gathered}$$

where $u(k+i)$ represents the total insulin infusion, which is the sum of the basal (u_{basal}) and bolus (u_{bolus}) insulin infusions. The total insulin variation $\Delta u(k+i)$ is limited to a maximum of 1.50 U/min per time step to prevent sharp fluctuations whereas basal insulin infusion is constrained to a maximum of 0.25 U/min. The glucose levels are constrained within the safe range, where $G_{\min} = 40\,\text{mg/dL}$ and $G_{\max} = 600\,\text{mg/dL}$.
The results obtained are presented in Fig. 3b. Managing diabetic patients with the MPC controller presents notable challenges, particularly in handling overshoots and undershoots around meal times and fasting periods. Simulations show that the controller effectively maintains glucose levels within the normal range for most of the day, with no incidents of nocturnal hypoglycemia or dawn hyperglycemia.

4. ANALYSIS AND DISCUSSION OF THE RESULTS

Since the analysis of *raw data* is essential for meaningful interpretations and testing research hypotheses, this Section provides an analysis framework to evaluate the effectiveness of the methods and approaches discussed in Section 3.

Analysis of Clinical Metrics: Clinical metrics are essential for assessing the performance of the designed

Table 2. Clinical Metrics average values across the 10 adult patients for PID, DDPG, and MPC

Controller	TIR (%)	TBR (%)	TAB (%)	VH (%)	VL (%)	Mean Glucose (mg/dL)	%CV
PID	65.22	22.60	12.18	3.56	18.30	**120.84**	51.58
DDPG	**79.92**	**5.41**	14.68	**1.33**	**2.20**	133.28	**26.72**
MPC	78.17	10.98	**10.85**	2.54	5.90	130.52	34.02

controller. According to the American Diabetes Association (ADA), relying solely on Time in Range (TIR: 70-180 mg/dL) might not fully capture the performance. Additional metrics such as Time Above Range (TAR: 180-250 mg/dL), Time Below Range (TBR: 54-70 mg/dL), Very High ($\geq$ 250 mg/dL), and Very Low ($\leq$ 54 mg/dL), along with the Coefficient of Variation (%CV), are crucial for a thorough evaluation because they account for intra-day glycemic fluctuations.

Table 2 illustrates that DDPG achieves the highest TIR and the lowest TBR, VH, and VL, signifying superior glucose control with minimal variability. MPC shows a relatively high TIR and a low mean glucose, although it experiences a higher TBR and VH compared to DDPG. Its %CV (34.02%) score is less favorable, pointing to reduced stability in glucose management. PID records the lowest TIR (65.22%) and a high TBR (22.60%), indicating poorer maintenance of glucose within the desired range. The higher %CV (51.58%) suggests more variability in glucose control. In summary, DDPG shows the most effective management of hypoglycemic and hyperglycemic events, with MPC following due to its ability to maintain a lower mean glucose despite experiencing more instability. PID is the least effective in this comparison.

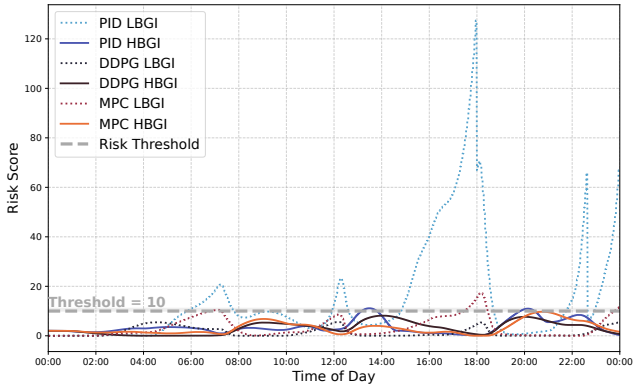


Fig. 4. Comparison of average LBGI and HBGI risk scores over time for the three controllers.

In the context of blood glucose monitoring, LBGI and HBGI metrics are crucial for evaluating the risks of hypoglycemia and hyperglycemia, respectively. The objective is to compare the effectiveness of the proposed control methods in keeping these risks below a fixed threshold of 10. The DDPG method demonstrates superior stability in both LBGI and HBGI values, as shown in Fig. 4, indicating its effectiveness in reducing high-risk periods for hypoglycemia and hyperglycemia. In contrast, the PID method shows significant variability, particularly toward the end of the simulated period, while the MPC method improves risk management compared to PID but still exhibits notable risk peaks at certain times.
Overall, DDPG provides the lowest and most stable risk values, indicating better blood glucose management and a significant reduction in high-risk periods. This suggests that advanced control strategies like DDPG may play a crucial role in optimizing diabetes therapy, reducing long-term risks, and improving patient outcomes, with potential healthcare cost benefits [Bommer et al. (2018)].

Comparison with real clinical practice: Understanding whether the simulated insulin levels are close to those administered in real clinical practice is crucial for evaluating the accuracy of the simulation model and its ability to realistically represent the glucose-insulin control system.

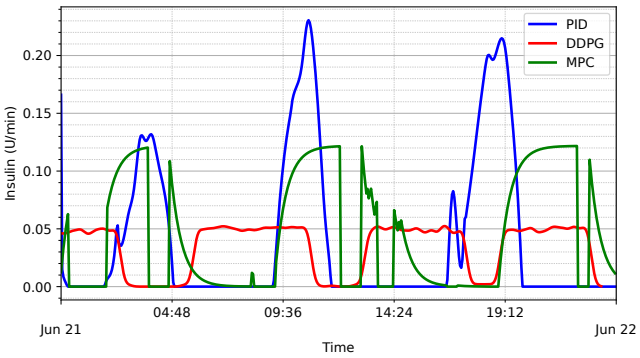


Fig. 5. Insulin infused by PID, MPC, and DDPG controllers for Adult #007

The total daily dose (TDD) of insulin is typically calculated based on patient weight and insulin sensitivity, usually in the range of 0.5-1.0 units/kg/day [Janež et al. (2020)]. Basal insulin, administered continuously via long-acting insulin (e.g., *glargine* or *detemir*, [WHO (2023)]), is delivered at a rate that typically varies between 0.1 and 2.0 U/hour (0.01 and 0.05 U/min, [Formoso and Bruttomesso (2022)]) depending on individual needs. Bolus insulin, taken before meals to cover expected increases in blood sugar due to carbohydrate intake, is adjusted based on current glucose levels and carbohydrate consumption. Fig. 5 shows the insulin levels administered to patient *Adult*#007 during the *in silico* simulation, which are consistent with typical clinical practice values. In our approach, the DDPG algorithm exhibits more stable insulin delivery compared to both PID and MPC controllers. The plot shows that the MPC and PID controllers produce higher insulin deliveries at meal times compared to DDPG, which, as evidenced by the results, may lead to a higher risk of hypoglycemia following meal consumption.

5. CONCLUSION

DDPG demonstrated superior control in the *in silico* simulations with the 10-patient cluster, reducing episodes of hypo- and hyperglycemia and maintaining glucose within safe limits more effectively than PID and MPC. The TIR achieved with DDPG was 79.92%, compared to 65.22% for PID and 78.17% for MPC. PID exhibited sub-optimal

performance, with glucose levels falling below 54 mg/dL for an average of 18.30% of the simulation time. MPC, while performing better than PID, lacked the stability and efficiency of DDPG in managing high-risk periods and adapting to insulin and carbohydrate variations (%CV: 34.02% vs. 26.72%). The simulations confirmed realistic insulin dosing and validated the models, indicating that DDPG could improve automated diabetes management by enhancing stability and reducing complications.

ACKNOWLEDGMENTS

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 (grant Nº 101156751) project.

REFERENCES

AlMahamid, F. and Grolinger, K. (2024). Reinforcement learning algorithms: An overview and classification. *Department of Electrical and Computer Engineering, Western University*.

Atanasious, M.M.H., Becchetti, V., Baldisseri, F., Menegatti, D., and Wrona, A. (2024). Deep reinforcement learning control of type-1 diabetes with cross-patient generalization. *2024 32nd Mediterranean Conference on Control and Automation (MED)*, 221–226.

Baldisseri, F., Menegatti, D., and Wrona, A. (2024). Deep deterministic policy gradient control of type 1 diabetes. In *2024 European Control Conference (ECC)*, 868–873.

Bergman, R.N. (2003). The minimal model of glucose regulation: A biography. In *Advances in Experimental Medicine and Biology*, volume 537, 1–19. Springer.

Bommer, C., Sagalova, V., Heesemann, E., et al. (2018). Global economic burden of diabetes in adults: Projections from 2015 to 2030. *Diabetes Care*, 41(5), 963–970.

Boughton, C.K. (2021). Fully closed-loop insulin delivery—are we nearly there yet? *The Lancet Digital Health*, 3(11), e689–e690.

Bouhamed, O., Ghazzai, H., Besbes, H., and Massoud, Y. (2020). Autonomous uav navigation: A ddpg-based deep reinforcement learning approach. *2020 IEEE International Symposium on Circuits and Systems (ISCAS)*.

Chun, Y., Choi, J., Min, I., Ahn, M.S., and Han, J. (2023). Ddpg reinforcement learning experiment for improving the stability of bipedal walking of humanoid robots. *2023 IEEE/SICE International Symposium on System Integration (SII)*, 1–7.

Dalla Man, C., Rizza, R.A., and Cobelli, C. (2007). Meal simulation model of the glucose-insulin system. *IEEE Transactions on Biomedical Engineering*.

Dassau, E., Atlas, E., and Phillip, M. (2010). Closing the loop. *International Journal of Clinical Practice*, 64(s166), 20–25.

Duan, Y., Chen, X., Houthooft, R., Schulman, J., and Abbeel, P. (2016). Benchmarking deep reinforcement learning for continuous control.

Formoso, G. and Bruttomesso, D. (2022). Insulin pump therapy: The state of the art. *Il Diabete Online*.

Hou, Q., Gu, C., Wang, X., Zhang, Y., and Zhao, P. (2021). Dynamic trajectory planning of a 7-dof surgical robot based on her-ddpg algorithm. *ASME International Mechanical Engineering Congress and Exposition, Volume 5: Biomedical and Biotechnology*.

Hovorka, R., Allen, J.M., Elleri, D., and et al. (2010). Manual closed-loop insulin delivery in children and adolescents with type 1 diabetes: a phase 2 randomised crossover trial. *The Lancet*, 375(9716), 743–751.

Hugging Face (2023). Illustrating reinforcement learning from human feedback (rlhf). `https://huggingface.co/`. Accessed: 2024-06-18.

Janež, A., Guja, C., Mitrakou, A., et al. (2020). Insulin therapy in adults with type 1 diabetes mellitus: a narrative review. *Diabetes Ther*, 11(2), 387–409.

Joubert, M., Briant, A.R., Kessler, L., et al. (2022). Sensor-augmented insulin pump with predictive low-glucose suspend (plgs): Determining optimal settings of pump and sensor in a multicenter cohort of patients with type 1 diabetes. *Diabetes Therapy*, 13(9), 1645–1657.

Leeder, T.M. (2022). Behaviorism, skinner, and operant conditioning: Considerations for sport coaching practice. *Strategies*, 35, 27–32.

Lillicrap, T.P., Hunt, J.J., Pritzel, A., Heess, N., Erez, T., Tassa, Y., Silver, D., and Wierstra, D. (2019). Continuous control with deep reinforcement learning.

Minsky, M. (1954). *Theory of neural-analog reinforcement systems and its application to the brain-model problem*. Dissertation, Harvard University.

Nimri, R., Muller, I., Atlas, E., and et al. (2014). Md-logic overnight control for 6 weeks of home use in patients with type 1 diabetes: randomized crossover trial. *Diabetes Care*, 37(11), 3025–3032.

OpenAI (2021). Deep deterministic policy gradient (ddpg). Accessed: 2024-07-07.

OpenAPS (2015). Openaps. `https://openaps.org/`. Accessed: 2024-06-18.

Pérez-Gandía, C., Facchinetti, A., Sparacino, G., and et al. (2010). Artificial neural network algorithm for online glucose prediction from continuous glucose monitoring. *Diabetes Technology & Therapeutics*, 12(1), 81–88.

Schwenzer, M., Ay, M., Bergs, T., et al. (2021). Review on model predictive control: an engineering perspective. *International Journal of Advanced Manufacturing Technology*, 117, 1327–1349.

Steil, G.M., Rebrin, K., Darwin, C., Hariri, F., and Saad, M.F. (2006). Feasibility of automating insulin delivery for the treatment of type 1 diabetes. *Diabetes*, 55(12), 3344–3350.

Sumiea, E.H., Abdulkadir, S.J., Alhussian, H.S., Al-Selwi, S.M., Alqushaibi, A., Ragab, M.G., and Fati, S.M. (2024). Deep deterministic policy gradient algorithm: A systematic review. *Heliyon*, 10(9), e30697.

Tan, H. (2021). Reinforcement learning with deep deterministic policy gradient. *International Conference on Artificial Intelligence, Big Data and Algorithms*, 82–85.

Thomas, A. and Heinemann, L. (2022). Algorithms for automated insulin delivery: An overview. *Journal of Diabetes Science and Technology*, 16(5), 1228–1238.

WHO (2023). *The selection and use of essential medicines 2023: web annex A: World Health Organization model list of essential medicines: 23rd list (2023)*. WHO.

WHO (2024). Diabetes. `https://www.who.int/health-topics/diabetestab= tab1`. Accessed: 2024-09-04.

Xie, J. (2018). Simglucose v0.2.1. `https://github.com/jxx123/simglucose`. Accessed on: 04-01-2024.