biomedicines

MDPI

Article

Reinforcement Learning: A Paradigm Shift in Personalized Blood Glucose Management for Diabetes

Lehel Dénes-Fazakas [1,2,3], László Szilágyi [1,2,4], Levente Kovács [1,2], Andrea De Gaetano [1,5,6] and György Eigner [1,2,*]

[1] Physiological Controls Research Center, University Research and Innovation Center, Obuda University, 1034 Budapest, Hungary; denes-fazakas.lehel@uni-obuda.hu (L.D.-F.); szilagyi.laszlo@uni-obuda.hu (L.S.); kovacs@uni-obuda.hu (L.K.); andrea.degaetano@biomatematica.it (A.D.G.)
[2] Biomatics and Applied Artificial Intelligence Institute, John von Neumann Faculty of Informatics, Obuda University, 1034 Budapest, Hungary
[3] Doctoral School of Applied Informatics and Applied Mathematics, Obuda University, 1034 Budapest, Hungary
[4] Computational Intelligence Research Group, Sapientia Hungarian University of Transylvania, 540485 Tîrgu Mureș, Romania
[5] CNR-IASI Institute for Systems Analysis and Computer Science, National Research Council of Italy, 00185 Rome, Italy
[6] CNR-IRIB, Institute of Biomedical Research and Innovation, National Research Council of Italy, 90146 Palermo, Italy
* Correspondence: eigner.gyorgy@uni-obuda.hu

Abstract: Background/Objectives: Managing blood glucose levels effectively remains a significant challenge for individuals with diabetes. Traditional methods often lack the flexibility needed for personalized care. This study explores the potential of reinforcement learning-based approaches, which mimic human learning and adapt strategies through ongoing interactions, in creating dynamic and personalized blood glucose management plans. **Methods:** We developed a mathematical model specifically for patients with type IVP diabetes, validated with data from 10 patients and 17 key parameters. The model includes continuous glucose monitoring (CGM) noise and random carbohydrate intake to simulate real-life conditions. A closed-loop system was designed to enable the application of reinforcement learning algorithms. **Results:** By implementing a Policy Optimization (PPO) branch, we achieved an average Time in Range (TIR) metric of 73%, indicating improved blood glucose control. **Conclusions:** This study presents a personalized insulin therapy solution using reinforcement learning. Our closed-loop model offers a promising approach for improving blood glucose regulation, with potential applications in personalized diabetes management.

Keywords: blood glucose levels; diabetes; reinforcement learning; artificial intelligence; glucose control; personalized management; dynamic strategies; patient profiles; closed-loop insulin delivery systems; predictive models; monitoring

check for updates

Citation: Dénes-Fazakas, L.; Szilágyi, L.; Kovács L.; De Gaetano, A.; Eigner, G. Reinforcement Learning: A Paradigm Shift in Personalized Blood Glucose Management for Diabetes. *Biomedicines* **2024**, *12*, 2143. https://doi.org/10.3390/biomedicines12092143

Academic Editor: Tadesse Melaku Abegaz

Received: 22 August 2024
Revised: 15 September 2024
Accepted: 16 September 2024
Published: 21 September 2024

1. Introduction

Diabetes mellitus (DM) is a persistent and currently untreatable metabolic disease resulting from either a complete lack of insulin (Type 1 Diabetes Mellitus, T1DM) or a partial lack of and/or an insufficient effect of insulin (Type 2 Diabetes Mellitus, T2DM). The exact pathophysiology is still unclear; TDM1 is believed to be the result of a cascade of autoimmune reactions that destroy the insulin-producing β cells in the Langerhans islets in the pancreas [1]. This manuscript focuses on managing blood sugar levels under the circumstances of T1DM using reinforcement learning methods of artificial intelligence techniques for external insulin administration.

T1DM typically develops rapidly and mostly affects children and adolescents, possibly genetically prone individuals. However, lifestyle and external circumstances an also accelerate the manifestation of the condition [2,3].

The fundamental molecular role of insulin is the insulin-stimulated entry of glucose into insulin-sensitive tissues (mainly muscle and adipose tissue) and the inslulin-mediated suppression of excessive glucose production (gluconeogenesys, glycogenolysis) mainly in the liver and the kidneys [4].

Patients with TDM1 need external insulin throughout their life to maintain their glycemia [5]. Without external insulin administration, these patients cannot survive due to the energy breakdown of their body's household [6]. High-quality robust and personalized blood glucose management is of significant importance for patients with T1DM to maintain normal glycemia levels and to decrease the expression of side effects. In recent years, it was proven that semi-automated insulin administration in the case of T1DM is able to provide satisfactory glycemia levels over the years, and it is able to reduce the risk of developing serious side effects [7,8]. However, there is a strong need for further development in the area of control algorithm candidates that can be employed for insulin administration. Such solutions need to be robust at the beginning of usage but need to have self-tuning capabilities to adapt to the patient needs over time to provide personalized treatment and insulin administration [9].

Usually, semi-automatic glycemic control operates according to the artificial pancreas (AP) concept. An AP system consists of three components: a Continuous Glycemia Monitoring System (CGMS) to measure the glycemia, an insulin pump as an actuator to inject insulin, and an (advanced) control algorithm [9–12].

Control algorithms in artificial pancreas systems play a crucial role in determining insulin dosing based on real-time glucose data [13–16]. These algorithms continuously analyze glucose sensor readings and use mathematical models to predict future glucose levels. The primary goal is to maintain blood sugar levels within a target range while minimizing the risk of hypoglycemia (low blood sugar) and hyperglycemia (high blood sugar).

Various types of control algorithms have been developed and investigated over the two decades of AP developments, including for example proportional–integral–derivative (PID) controllers, model predictive controllers (MPCs), and fuzzy logic controllers. Each algorithm has its strengths and limitations, and researchers continue to refine and optimize them for better performance and safety [17–22].

The choice of control algorithm depends on factors such as the individual's insulin sensitivity, meal intake, physical activity, and overnight glucose patterns. Additionally, advancements in machine learning and artificial intelligence have led to the development of more adaptive and personalized algorithms that can adjust insulin dosing based on individual variability and changing circumstances [23,24]

Current methods of blood glucose management in type 1 diabetes rely on regular self-monitoring of blood glucose levels and insulin administration. However, these methods require significant dedication from patients. As a result, there is increasing interest in utilizing machine learning techniques, particularly reinforcement learning (RL) [25–27].

Reinforcement learning (RL) belongs to the advanced machine learning methods, where an agent-based algorithm (so-called neural network-based model) learns the appropriate policy for acting upon training by receiving feedback in the form of rewards or penalties. Its effectiveness has been showcased in diverse fields such as game playing [28], robotics [29], and healthcare [30]. In the context of blood glucose management, RL can be utilized to formulate a personalized strategy for adjusting insulin dosages based on the analysis of both present and past blood glucose readings, injected insulin, and other possible features [31].

In this paper, we present our reinforcement learning-based solution for blood glucose control, which uses a PPO agent to control blood glucose levels. The experiments were performed using an IVP mathematical model. From this mathematical model, we created a closed-loop simulator that corresponds to a reinforcement learning environment. For this mathematical model, we had 17 parameter sets from 10 patients.

2. Related Works

Ref. [32] conducted a comprehensive study on the application of offline RL algorithms for glucose control, particularly focusing on Batch-Constrained Q-learning (BCQ) and Conservative Q-Learning (CQL). Offline RL is advantageous because it allows for training models on pre-existing datasets, thereby minimizing risks associated with real-time patient interactions during the learning process. The study reported a significant improvement in TIR, increasing from 61.6% to 65.3%, a result that is both statistically significant and clinically relevant. The research also emphasized safety using CVGA to confirm that the majority of glucose predictions fell within Zone A, demonstrating high clinical accuracy. This work is critical, as it paves the way for the safer deployment of RL-based glucose control in real-world settings, where patient safety is paramount.

Ref. [33] explored the use of the Soft Actor-Critic (SAC) algorithm, which is a model-free RL approach known for its ability to handle continuous action spaces effectively. The SAC was applied to optimize insulin dosing in a closed-loop system designed for type 1 diabetes management. The algorithm was specifically tailored to dynamically adjust insulin delivery in response to varying blood glucose levels, addressing the inherent variability in diabetic patient responses. The study achieved an impressive TIR exceeding 75%, which is a significant improvement over traditional methods. Furthermore, the CVGA results indicated that most glucose predictions were within the safe Zones A and B, confirming the algorithm's reliability and safety. This research is particularly notable for demonstrating the feasibility of using advanced RL algorithms in real-time clinical applications.

Ref. [34] introduced a multi-step deep RL strategy designed to tackle the complexities of glucose regulation, particularly the delayed effects of insulin and meal intake on blood glucose levels. The study utilized advanced RL techniques such as Dueling DQN and Prioritized Experience Replay (PER), which enhance the algorithm's ability to learn from important past experiences while stabilizing learning in the presence of delayed rewards. The results were remarkable, with a TIR of 85.62%, representing a 28.7% improvement over the baseline methods. The CVGA analysis further validated the model's safety, with more than 90% of predictions falling within Zones A and B. This study highlights the importance of considering temporal dependencies in glucose control and showcases the potential of multi-step learning in addressing these challenges.

Ref. [35] focused on evaluating various extensions of the Deep Q-Network (DQN) algorithm for blood glucose control in in silico models of type 1 diabetes patients. The study compared different DQN variants, including Double DQN and NoisyNet-DQN, to determine which configurations were most effective in managing blood glucose levels. The research showed that these advanced DQN algorithms could significantly improve the TIR, reaching 80%, which is a substantial enhancement over traditional approaches. The CVGA analysis confirmed the safety of the control decisions, with most predictions within Zones A and B. This study is particularly important for its in-depth analysis of different DQN architectures and their implications for safe and effective glucose control.

3. Materials and Methods

3.1. Reinforcement Learning

Reinforcement learning is a flexible framework known for its ability to adjust to complex process demands, functioning through semi-supervised learning. In this study, we applied the actor-critic architecture [36]. The objective of this model was to identify the sequence of actions that maximizes the reward value by the combined actions of the actor and the critic (both are neural networks). Embodied within the actor-critic models is a neural duality, which is realized as two distinct neural networks: the first one is the value network aiming to quantify the intrinsic quality of states via a value function. The second one is the policy network that sets probabilistic values for actions (vector or scalar) in order to facilitate the comprehensive objective of maximizing rewards [36]. In this study, we applied the Proximal Policy Optimization (PPO) paradigm [37]—below, we explain why we chose PPO by presenting our preliminary results. The PPO is an evolution of the

Stable Baselines3 library [38]. Policy gradient methods such as TRPO [39], GAE [40], and A2C/A3C [41,42] have recently come to attention. We worked within the OpenAI gym [43], which is an environment for the development of reinforcement learning situations.

3.2. Preliminary Results

In this section, we summarize the progress made so far through three sub-section, as we have published three articles on the subject. This manuscript draws on the conclusions of these studies and implements them together to achieve better results for both the TIR and CVGA metrics achieved in these articles. We describe the important conclusions of these manuscripts, but the full network we used in this research is described in the later Section 3.6.

3.2.1. Testing Multiple Agents and Varying Simulation Lengths

In the previous study[44], the feature vector was calculated using blood glucose measurements and insulin administrations over one hour in the past. Since the measurement interval of the CGM sensor was 5 min, 12 samples of blood glucose were used.

During feature extraction, 11 features were extracted from the blood glucose data, and 11 features were also extracted from the insulin data. From the blood glucose data, we calculated 11 point deviations from the existing 12 measured points. Then, we also used the same method, only in the second case, we divided by five, since the sensor measures at every 5 min. We also extracted 11 features from the insulin data. This was also a pointwise variation. Also, there was a calculation of the difference between two endpoints for the insulin characteristics.

$$\Delta G(i) = G(i+1) - G(i), \quad i \in [0, 11] \tag{1}$$

$$\frac{\Delta G(i)}{\Delta t} = (G(i+1) - G(i))/5, \quad i \in [0, 11] \tag{2}$$

$$\Delta u(i) = u(i+1) - u(i), \quad i \in [0, 11] \tag{3}$$

$$\Delta u = u(11) - u(0) \tag{4}$$

In that first study, we examined several types of agents, including both discrete and continuous action space agents. The algorithms we investigated were the following:

- PPO [37];
- SAC [45];
- DDPG [46];
- DQN [47];
- A2C [48];
- TD3 [49].

We implemented a closed-loop system that included a virtual patient acting as the environment [50]. As the controller, our system included a neural network-based agent: this agent sends the insulin data to be “injected” into the environment and thus modify the environment; the environment, in turn, returns a blood glucose value based on the injected insulin, on set patient parameters, and on a given carbohydrate intake. The action space spanned insulin administration rates from 0 to 18 U/h.

We examined several types of reward functions for the optimization of the controller: the tests showed that a two-step reward function provided the best performance for our problem. In the first step, we calculated a temporary reward, and then in the second step, we averaged the temporary rewards. The tests were evaluated using the Time in Range (TIR) [51] and Continuous Variability Grid Analysis (CVGA) metrics [52]. Since the PPO algorithm yielded the best performance in the first tests, we then continued to work with this algorithm for the one-day test analysis. For the one-day test, we set up three different test cases, which are introduced here in order to provide a full overview of the fundaments of the current research results.

Preliminary Results Scenario I: 1 Day of Training, 1 Day of Testing

In our initial trial, we conducted a one-day simulation for instructional purposes, followed by a one-day assessment phase. Among the patients, five exhibited extreme glucose level fluctuations that led us to terminate the testing protocol. For the remaining patients, no remarkable deviations were observed in their glycemia time course. The average Time in Range (TIR) recorded was 77.55%, which is a satisfactory outcome considering the broader patient cohort; see Table 2 from [44]. The distribution pattern of blood glucose (BG) levels across all patients also aligned well with the TIR criterion.

By evaluating the results through the Control Variability Grid Analysis (CVGA) metric, we identified Zone B control actions for a total of seven patients. Notably, patient number 7 showcased the most commendable performance, maintaining a flawless 100% adherence within the target range; see Table 2 from [44].

Preliminary Results Scenario II: 1 Day of Training, 10 Days of Testing

For the next evaluation, we engaged the agent in a one-day training period followed by a comprehensive ten-day simulation period. The objective was to test the agent's capacity to extrapolate from a single day of training towards an extended simulation period. During this trial, we encountered four instances of extreme blood glucose (BG) levels. In contrast to the previous trial, shifts in the agent's performance out of the middle range were discernible, resulting in higher blood glucose (BG) levels.

According to the CVGA diagram in Table 3 from [44], only two instances were felt in Zone B. Considering the Time in Range (TIR) metrics, the average was 72%. Notably, patient 10 emerged as the standout performer in terms of the results. An intriguing scenario emerged with patient 13, wherein the agent managed to steer the patient away from extreme conditions while leaning towards higher BG levels within the non-extreme zone.

The 180–250 BG range was occupied on average 17% of the time, which is well within the stipulated acceptable threshold (25%). This alignment with the TIR metric confirms its consistency. However, for patient 13, the allowed percentage was exceeded, raising questions on the general applicability of the agent.

Preliminary Results Scenario III: 10 Days of Learning, 1 Day of Testing

For our third evaluation, we inverted the conditions of the second test: the agent underwent a ten-day training phase, which was followed by a condensed one-day testing period. The aim was to ascertain whether the agent's ability to generalize from extensive training could be translated into shorter testing scenarios.

The outcomes of this test do not support the agent's capacity to formulate control strategies for individual patients that effectively avert extreme zones: the Time in Range (TIR) metrics failed to align with the predefined criteria, as the blood glucose (BG) levels showed shifts towards the higher range.

These results suggest that we may have a good result if we use the same simulation duration for both training and testing. In this test, no CGM noise was used in the patients.

3.2.2. Testing with Different Reward Functions

In our previous study [53], we investigated which reward functions would give the best performance in terms of blood glucose regulation. For the first time, we examined continuous functions as well. One of the main challenges here was that we used an extended mathematical model, namely, the virtual patient model was completed by a stochastic sensor noise (continuous glucose monitoring sensor—also known as CGMS). We applied the model output: the noisy blood glucose level in the reward functions. Because of this, the deviations of the different reward functions were much larger, which made training extremely difficult, as violations of the termination criterion boundaries occurred more often due to the deviations. On the other hand, this procedure increased the viability of the methodology, since in practice, the CGM signal is available as input for control actions.

The simulation time for both training and testing was one day. During training, carbohydrate inputs were randomly generated during exercise, and carbohydrate intake was randomly generated according to the Section 3.4. Thirteen possible test days were defined with different carbohydrate intake patterns. Ten test days had randomly generated carbohydrate intake, while three test days had predefined carbohydrate intake: (i) one had no carbohydrate intake, (ii) one had 12 g of carbohydrate intake every hour, (iii) one had an additional 5 g of carbohydrate at 12 h in addition to the carbohydrate amount generated by the meal function as described in Section 3.4. The control agent we used was the PPO algorithm, since this provided the best performance in our first tests according to Section 3.2.1. The architecture was not complex: both the actor and critic networks were outgoing networks. However, they were structurally identical and contained two hidden layers with a Relu activation functions in the hidden layer. Moreover, the observation space was changed, as there were now only two elements for the observation of blood glucose and insulin concentration at the previous time point.

We tried four different reward functions shown in Figure 1, one of which was not continuous. All functions were maximized around the 120 mg/dL blood glucose level. We measured the performances of the reward functions by applying concrete termination criterion during training (and stopping the simulation when it was violated, namely, when the blood glucose level felt outside the given ranges) and by not terminating the simulation at all. We applied CVGA and the TIR for performance measurements as metrics. Furthermore, we also calculated the mean squared error of the blood glucose level that occurred during the test. This error was calculated both for 90 mg/dL and 150 mg/dL. In the following, we introduce the tested reward functions.

Bump function [54]:

$$score(i) = \begin{cases} \exp\left(\frac{-1}{1-((CGM-90)/45-1)^2}\right) & 90 < CGM < 180 \\ 0 & \text{otherwise} \end{cases} \tag{5}$$

Piecewise constant function:

$$score(i) = \begin{cases} -1 & CGM < 70 \\ 1 & 70 \leq CGM \leq 180, \\ 0 & 180 < CGM \end{cases} \tag{6}$$

$$R = \frac{1}{n}\sum_{i=1}^{n} score(i), \tag{7}$$

where R is the final reward, and n is the simulation length.

Cosine function:

$$score(i) = \begin{cases} -\cos\left(\frac{CGM}{45}\right) & 0 < CGM < 300 \\ -1 & otherwise \end{cases}, \tag{8}$$

Mexican hat wavelet [55]:

$$\begin{aligned} score(i) = {} & \frac{2}{\sqrt{3}\cdot\pi^{\frac{1}{4}}} \cdot \left(1 - \left(\frac{CGM-140}{140}\right)^2\right) \\ & \cdot \exp\left(-\left(\frac{1}{2}\cdot\left(\frac{CGM-140}{140}\right)^2\right)\right) \end{aligned} \tag{9}$$

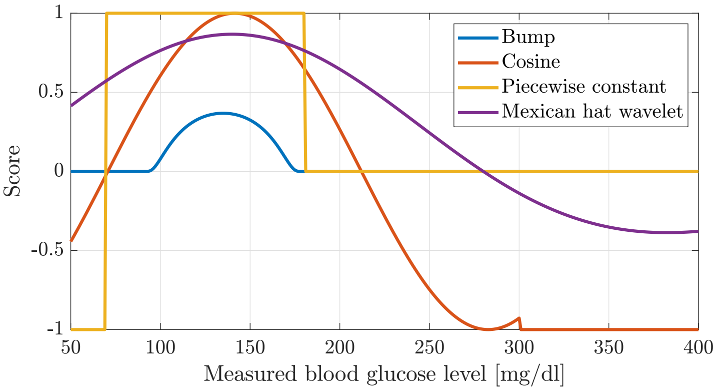


Figure 1. Investigated reward functions.

During the evaluation, we had to use a median value when evaluating the tests, as the models performed very poorly for patients with lower body weight. Our tests showed that the Bump function performance was the best when the simulation was not stopped. The second best was the Cosine-type function. For the Bumps function, we obtained a TIR result of 68%. Also, the box plot shows very clearly the several outlier values. From the glucose curves, it turned out that there was a steady increase in the blood glucose curve. The CVGA plot shows that although the Bump function produced fewer Zone A control values, there were more Zone B and C values and fewer Zone D values.

In conclusion, CMG noise highly influences the performance of the models. However, a positive result is that these models avoid hypoglycemia, albeit at the expense of significant hyperglycemia. It also appears that the continuous functions perform much better than the original two-step reward function. The steady slope of the glucose curves suggests that one day of simulation is not enough, as this short period of time is not sufficient for the models to learn the transition between days. The best result was not far short of the expected 70% performance in terms of the TIR metric.

3.2.3. Testing Differenct Avtivation Function and Hyperparameters

In our third experiment [56], we investigated how the changes in the hyperparameters affect the performance in terms of the CVGA and TIR metrics. The goal was to determine which changes in the hyperparameters are the most sensitive with respect to the control performance. To this end, we ran tests varying the number of neurons in the hidden layers between 64 and 512. In addition, we applied another test, where the number of neurons in the hidden layers was fixed (64 neurons), but the activation functions were modified simulation by simulation (Sigmoid, Relu, ELU, and No activation function). The environment was the same as in our second experiment with two characteristic inputs and one output, which was the amount of insulin.
Relu:

$$f(x) = \max(0, x) \tag{10}$$

Sigmoid:

$$f(x) = \frac{1}{1 + e^{-x}} \tag{11}$$

ELU:

$$f(x) = \begin{cases} x & x \geq 0 \\ \alpha(\exp(x) - 1) & x < 0 \end{cases} \tag{12}$$

The training and test cases were the same as those used in the second test [53]. Also, the reward function was fixed (piecewise) to the one we originally used in our investigation, and we also used median values to summarize the results. Apparently, changing the number of neurons did not determine a large variation in the control ability of the neural network models, even if larger meshes performed marginally better than smaller meshes (the 512-neuron network achieved the best performance in the tests). Conversely, different activation functions resulted in significantly different performance outcomes of the controllers, with the ELU and the Sigmoid functions outperforming the others. For the ELU activation, a TIR metric above 70% was achieved. These results suggest that deeper meshes and ELU activation functions in the hidden layer should be used.

3.3. Virtual Patient Model

The basis of the environment to be controlled was provided by the Identifiable Virtual Patient model (IVP) [57]. We did not collect human data in our experiments. This model was extended with a model of sensor noise for our simulation environment. Our approach included a cohort of 10 specifically characterized patients, with additional parameters for night-time periods. This represents a total of 17 different virtual patient profiles. In total, we had 10 validated patient records to work with. However, as we also had daytime and night-time parameter sets for some patients, we had a total of 17 parameter sets. These parameter sets can be examined in Table 1.

Table 1. Patient parameter sets, in cases where the VG and BW parameters are the same, are the same patient.

BW	GEZI	EGP	CI	SI	tau1	tau2	p2	VG
89	3.87×10^{-8}	1.4	2010	4.93×10^{-4}	49	47	1.06×10^{-2}	253
89	2.20×10^{-3}	1.33	2010	8.11×10^{-4}	49	47	1.06×10^{-2}	253
63	4.38×10^{-3}	0.6	1281	9.64×10^{-5}	41	10	1.16×10^{-2}	261
65	3.50×10^{-3}	0.856	909	1.70×10^{-4}	71	70	2.33×10^{-2}	199
65	1.64×10^{-3}	1.07	909	4.63×10^{-4}	71	70	2.33×10^{-2}	199
116	7.58×10^{-8}	2.59	1813	3.77×10^{-4}	91	70	8.14×10^{-3}	337
116	1.64×10^{-5}	0.98	1813	3.77×10^{-4}	91	70	8.14×10^{-3}	337
64	4.33×10^{-3}	0.6	1535	2.05×10^{-4}	46	46	9.63×10^{-3}	188
51	1.01×10^{-3}	0.603	588	4.12×10^{-4}	68	30	9.15×10^{-3}	104
77	2.30×10^{-3}	1.11	1806	8.16×10^{-4}	60	60	1.01×10^{-2}	263
65	1.00×10^{-8}	1.3	540	3.68×10^{-4}	95	37	1.03×10^{-2}	137
100	6.39×10^{-3}	1.27	875	2.56×10^{-4}	131	21	1.03×10^{-2}	193
64	1.04×10^{-3}	0.611	1309	6.03×10^{-4}	53	53	1.02×10^{-2}	204
51	3.79×10^{-3}	0.603	588	9.48×10^{-4}	68	30	9.15×10^{-3}	104
65	1.00×10^{-8}	0.601	540	5.40×10^{-4}	95	37	1.03×10^{-2}	137
100	6.39×10^{-3}	3.45	875	6.89×10^{-4}	131	21	1.03×10^{-2}	193
64	1.04×10^{-3}	0.611	1309	1.73×10^{-3}	53	53	1.02×10^{-2}	204

The following equations describe the dynamics of a patient:

$$\dot{G}(t) = -(GEZI + I_{EFF}(t)) \cdot G(t) + EGP + R_A(t) \tag{13}$$

$$\dot{I}_{EFF}(t) = -p_2 \cdot I_{EFF}(t) + p_2 \cdot S_I \cdot I_P(t) \tag{14}$$

$$\dot{I}_P(t) = -\frac{1}{\tau_2} IP(t) + \frac{1}{\tau_2} I_{SC}(t) \tag{15}$$

$$\dot{I}_{SC}(t) = -\frac{1}{\tau_1} I_{SC}(t) + \frac{1}{\tau_1 C_I} u(t) \tag{16}$$

$$R_A(t) = \sum_{i}^{m} \frac{d_i}{V_G \cdot \tau_{D_i}^2} t_i \cdot e^{-\frac{t_i}{\tau_{D_i}}} \tag{17}$$

Blood glucose concentration is symbolized as $G(t)$ in units of milligrams per deciliter (mg/dL), and the effectiveness of insulin is denoted by $I_{EFF}(t)$ (min^{-1}). Subcutaneous and plasma insulin concentrations are represented as $I_{SC}(t)$ and $I_P(t)$, respectively, with both expressed in microunits per milliliter (µU/mL). The agent indirectly influences blood glucose level through the infusion of insulin. The R_A encapsulates disturbances in the form of carbohydrate intake d_i in grams.

The τ_1 in minutes and τ_2 in minutes are the timing parameters, and the rate constant p_2 (min^{-1}) defines the absorption kinetics of insulin. Insulin clearance is C_I, in milliliters per minute (mL/min), while insulin sensitivity S_I is expressed in milliliters per microunit per minute (mL/µU/min).

Endogenous glucose production, denoted by EGP, in units of milligrams per deciliter per minute (mg/dL/min), represents liver glucose output. Glucose effectiveness at zero insulin concentration ($GEZI$) describes insulin-independent glucose consumption. V_G is the apparent glucose distribution volume in deciliters (dL). The time of meal absorption is described by the time constants τ_{D_i}.

We extended the IVP model based on [58]. This augmentation consists of integrating into the IVP a CGM noise model, considering diverse factors such as temporal delay [59], sensor drift, additive sensor noise, and calibration imprecision. In order to avoid affecting the training process, we excluded the sensor drift from the model.

The temporal delay inherent to CGM measurements, a result of their interstitial measurement site, was incorporated into the extended model as an additional interstitial glucose compartment I_G. The sensor noise was modeled as an autoregressive process of second order with a stochastic white noise term, w, distributed as $\mathcal{N}(0, \sigma^2)$.

$$\dot{IG}(t) = -\frac{1}{\tau_{IG}} IG(t) + \frac{1}{\tau_{IG}} G(t), \tag{18}$$

$$v(t) = \alpha_1 v(t - T_s) + \alpha_2 v(t - 2T_s) + w(t), \tag{19}$$

$$CGM(t) = IG(t) + v(t), \tag{20}$$

3.4. Meal Generator

We utilized the meal generation schema of Wang et al. [60]. The algorithm generates six meals spread across the entire day, each having random carbohydrate content and timing.

- Assume that the patient's body weight is known;
- Probability of meal appearance: $p = [0.95, 0.3, 0.95, 0.3, 0.95, 0.3]$;
- Meal time upper boundary: $up = [9, 10, 14, 16, 20, 23] \cdot 60$;
- Meal time lower boundary: $lo = [5, 9, 10, 14, 16, 20] \cdot 60$;
- Meal time: $\mu_t = [7, 9.5, 12, 15, 18, 21.5] \cdot 60$;
- Meal time variance: $\sigma_t = [60, 30, 60, 30, 60, 30]$;
- Amount of meal: $\mu_a = [0.7, 0.15, 1.1, 0.15, 1.25, 0.15] \cdot BW$;
- Amount of meal variance : $\sigma_a = \mu_a \cdot 0.15$;
- Meal set: $E = \emptyset$;

- For $k \in [1, 2, 3, 4, 5, 6]$ do
- Generate a random value for p_{tmp} between 0–100;
- if $p_{tmp} \leq p[k]$
- Calculate the meal amount for kth timestamp: $e_k = Round(max(0, Normal(\mu_a[k], \sigma_a[k])))$
- Calculate the meal time for kth timestamp: $\zeta_k = Round(TruncNorm(\mu_t[k], \sigma_t[k], lb[k], op[k]))$
- Make a union with the meal set: $E \cup \{e_k, \zeta_k\}$;
- return E.

Since the mathematical model allows for carbohydrate intake, a method is needed to generate the carbohydrate input for the simulations. To enable the model to learn on more realistic simulations, it was important to implement this generator in a reinforcement learning environment. This solution is able to generate random carbohydrate inputs for all simulations based on a normal distribution. It takes the patient's body weight into account and calculates carbohydrate intakes based on it, which is again realistic, since in real life patients determine their carbohydrate intake based on their body weight. In addition, it has 6 meals of which the main meal 95% is possible, and the side meal 30% is possible. This generator randomly generates the daily carbohydrate intake. When we used it for a simulation of several days, we generated as many days as we simulated, and we added the meals one after the other. When trainning the network, we always used random carbohydrate intake. We also regenerated the meal for each simulation.

3.5. Closed Loop in Terms of RL

In our approach, the virtual patient is the IVP model, while the RL agent behaves as a closed-loop controller administrating insulin (to the environment). The IVP model provides the dynamic output of blood glucose (BG) levels (noisy BG level) depending on an array of inputs including both carbohydrates and insulin. The RL agent computes the insulin dosage for the upcoming time step (5 min) by integrating prevailing BG levels and the accumulated history of administered insulin. The resulting closed-loop design is shown in Figure 2 and detailed in Equations (1)–(4).

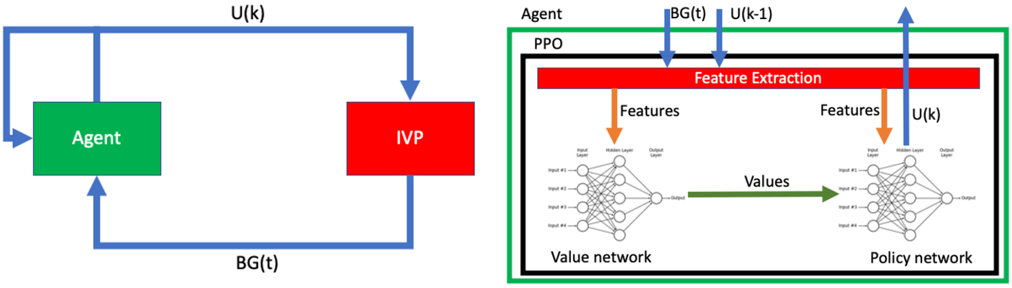


Figure 2. The Left side figure represents the block diagram of the reinforcement learning-based closed-loop algorithm. The right side figure details the control architecture of the system. BG(t) means blood glucose value measured at the t-th time point. U(k) insulin administered at time k.

3.6. Applied Neural Network

We used a Bump reward function to build our neural network. Similarly to our previous tests, our input feature vector consisted of two elements: the blood glucose level measured at the previous time and the amount of insulin administered at the previous time. The features were normalized to values between 0 and 1: for glucose, this was achieved by dividing the blood glucose level by 1000. For insulin, the action space in the environment was between $[-1$ and $1]$; therefore, in the design of the feature vector, we added one to the current decision and then divided it by 2. The action space $[-1, 1]$ represented insulin administration values from 0 up to 25 U/h. When the network made a decision, we added 1 to that value, then multiplied that value by 25, and finally divided it by 2, which was the actual administered insulin value. Since our previous tests showed that the models perform better when they do not stop the simulation time if values exceed blood glucose limit values, we simulated the whole time period in any case. Since tests showed that a longer stimulation time is needed, we considered 10 days of simulation.

Our current network shown in Figure 3 maintained the separation between value and policy as it was used in previous test. In this study, we applied large neural networks as blocks, consisting of Linear and ELU layers. The use of the ELU layer was justified by the good results obtained in the third test in Section 3.2.3, while the use of the Linear layer was applied to have continuous output. A variable number of these blocks appeared in the policy and value networks: while there were only five blocks in the value network, the policy network contained seven blocks. The value network was made smaller to allow the critical part of our model to learn faster, whereas the policy network was set deeper to give the actor more of a chance to learn the right strategy. Each block consists of 2048 neurons. The exceptions are the first block in both networks (the input layer of that block has 2 features) and the last block (which has an output of 64 values). Both networks had a terminal block consisting of an Identity layer and a Linear layer. The output of both networks is a single value: in the case of the value network, the value is a prediction of how good the given step was; for the policy network, the value, between −1 and 1, indicates the insulin quantity to be administered. Figure 4 shows the network architecture. The Linear layer was needed to obtain a linear combination in the output. In addition, in Pytorch 2.4.0 [61], where the Linear layer was used to vary the output number from the input. For other layers, the input number is the same as the output number. This output is then passed to the Identity layer.

All other hyperparameters Table 2 we used were the same as those used in the first test Section 3.2.1. We checked our agent's reward value after each simulation and we considered the average reward achieved by the model during training, respectively. This test was performed for the last 100 simulations. If this average value increased from one iteration to the next. Then, we saved the weights of the model to archive the best model that performed the best. In this way, we were able to extract the best model, on average. We obtained these 100 simulations in our first experiments illustrated in Section 3.2.1 so that with this many simulations, we could obtain a good performance for insulin control. The first results were also published on this basis. So, we did not change this value. Also, we realized that the simulation time has a big impact on learning: we thus increased the simulation time, training the model over 50 million steps. We chose 50 million steps because we only had 5 million steps in the first test described in Section 3.2.1, while the second test also had 5 million steps, as described in Section 3.2.2. And in our third test, we also had 5 million steps Section 3.2.3. We observed that when we gave the agent more time to learn, it can achieve better results. Therefore, we decided that 50 million steps will be 10 times the learning time step so far. With this change, convergence (in the ELKH cloud [62]) took 2 or 3 weeks per patient.

Table 2. Hyperparameters of the agent.

Hyperparameter	Value	Description
learning rate	0.0003	
number of steps	288	The number of steps to run for each environment per update
batch size	64	
gamma	0.99	
lambda	0.95	Factor for trade-off of bias vs variance for Generalized Advantage Estimator.
clip range	0.2	Clipping parameter—it is a function of the current progress remaining.
entropy coefficient	0	Entropy coefficient for the loss calculation.
value function coefficient	0.5	Value function coefficient for the loss calculation.
max gradient norm	0.5	The maximum value for the gradient clipping.
use SDE	False	Whether to use generalized State Dependent Exploration (gSDE) instead of action noise exploration.
target KL	None	Limit the KL divergence between updates, because the clipping is not enough to prevent large.

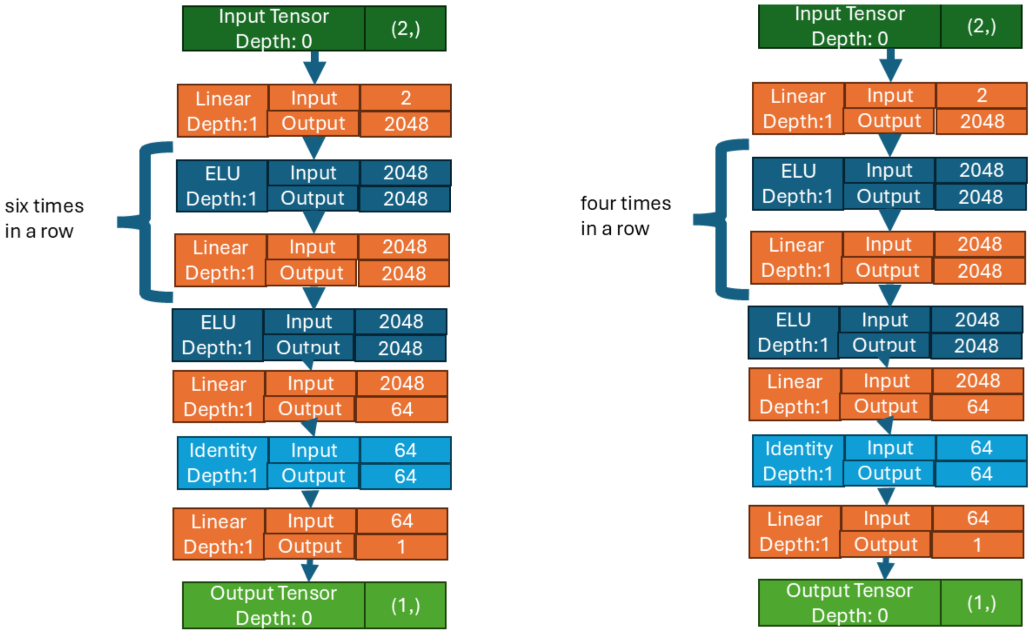


Figure 3. A graphical representation of the policy and value networks as depicted by Pytorch. The image on the left is the policy network, which is larger than the value network. This network learns the control strategy. On the right is the value network, which gives an estimate of how good the policy network was.

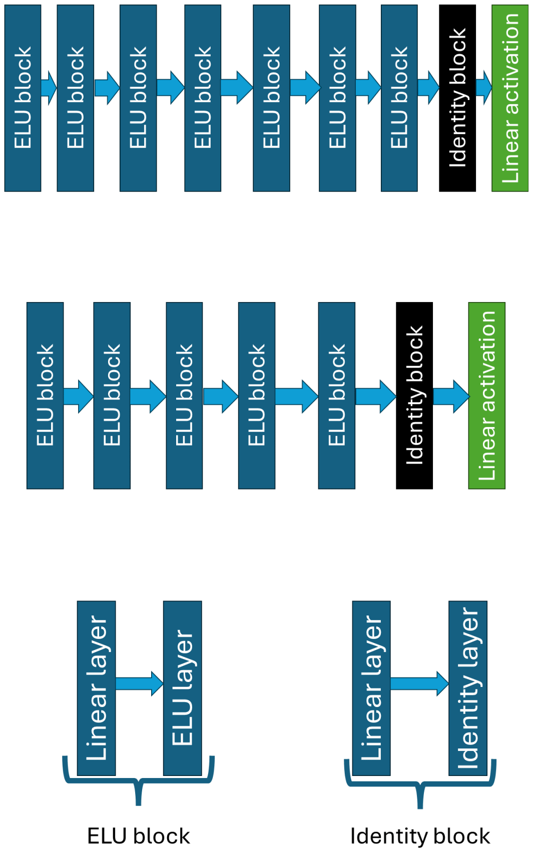


Figure 4. The **upper** figure shows the Policy network architecture. The **middle** figure shows the Value network architecture. The Block architectures are represented by the **bottom** figure.

3.7. Training Phase

Throughout the iterative training phase, a series of 10 consecutive day scenarios served as the test set to assess the evolving performance of the neural network. These scenarios incorporated randomized meal schemes interspersed between consecutive one-day periods, providing heterogeneous patterns for the networks to learn from. Different networks were trained for each subject, with constant patient parameters but varying timing and carbohydrate content of meals. The training was completed over 50 million iterative steps and was conducted on the ELKH research cloud.

3.8. Testing Phase

The phase of evaluation was carried over 13 different meal schemes—both random (10 schemes) and predefined (3 schemes) [53,56]. This makes the evaluation of performance possible on a comprehensive spectrum of challenges in terms of the meal intake. Furthermore, the objective here was to avoid overfitting; thus, the test cases were defined manually the as same a in the Section 3.2.3 testing scenarios. The virtual patient parameters were kept constant for both training and testing phases.

Among the 13 test scenarios, 10 replicated the methodology employed during the training phase, while three were manually designed to represent particular situations: the first was a scenario without any meal intake (0 g); and in the second, 12 g of carbohydrates were administered each hour. A single carbohydrate intake at 12 h is defined as the maximum amount of carbohydrate the generator can deliver plus 5g of carbohydrate.

3.9. Applied Metrics

We applied two different metrics for analyzing the result, which are the two most popular metrics for diabetic control:

- Control Variability Grid Analysis (CVGA);
- Time in Range (TIR) .

Control Variability Grid Analysis (CVGA) Figure 5 control an algorithm's performance. It provides insights into stability, responsiveness, and overall efficacy by assessing system behavior across various scenarios.

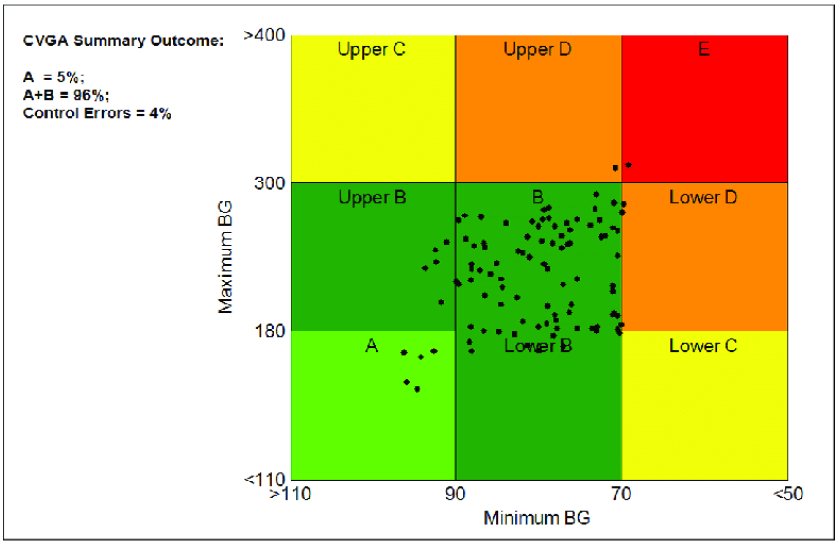


Figure 5. Control variability grid analysis (CVGA) metric.

CVGA subjects control systems to different inputs or disturbances and measures responses, which are displayed on a grid. This visualization helps evaluate trade-offs between control objectives and optimizes system parameters.

Smith et al. offered an extensive overview of CVGA [63]. Johnson et al. demonstrated CVGA's use in comparing control strategies [64]. Chen et al. showed the CVGA's effectiveness in optimizing industrial process control systems [65].

The CVGA is essential for comprehensively assessing and optimizing control algorithms as the field evolves.

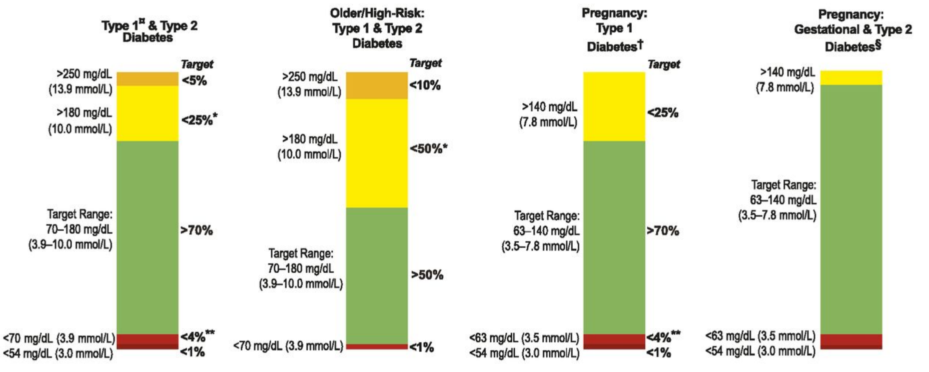
The Time in Range (TIR) Figure 6 metric offers a comprehensive assessment of blood glucose control over time. Unlike traditional metrics focusing on extremes like hypoglycemia or hyperglycemia, the TIR quantifies the time spent within a target blood glucose range.

The TIR is expressed as the percentage of time blood glucose levels stay within a specified range, often 70–180 mg/dL reflecting optimal glycemic control. This range may vary based on individual factors and treatment goals.

The TIR captures the dynamic fluctuations in blood glucose levels, providing a holistic view of glycemic variability. It helps assess the effectiveness of treatment regimens, lifestyle changes, and therapy adjustments.

The International Consensus on Time in Range (ICTR) highlights the TIR's importance in glycemic control. Battelino et al. showed the TIR's association with reduced diabetes complications [66]. Beck et al. linked the TIR to quality of life in diabetes [67]. The ADA and ATTD provide guidelines on the TIR's clinical use [68,69].

In summary, the TIR is a vital metric that offers a detailed evaluation of blood glucose control, supporting personalized diabetes management strategies.



¤ For age <25 yr., if the A1C goal is 7.5%, then set TIR target to approximately 60%. (See *Clinical Applications of Time in Ranges* section in the text for additional information regarding target goal setting in pediatric management.)
† Percentages of time in ranges are based on limited evidence. More research is needed.
§ Percentages of time in ranges have not been included because there is very limited evidence in this area. More research is needed. Please see *Pregnancy* section in text for more considerations on targets for these groups.
* Includes percentage of values >250 mg/dL (13.9 mmol/L).
** Includes percentage of values <54 mg/dL (3.0 mmol/L).

Figure 6. Time in range (TIR) metric.

4. Results

Table 3 lists the aggregated results in different glucose ranges. In the 70–180 mg/dL range the median value is high, 86% which surpasses previous results. It can be seen in 75% of the tests the method is able to achieve better than 44% TIR. The lowest value in this range is 18%, showing that there are still patients whose glucose control was not feasible with the current method. The outlier patients had low body weights in all cases. Overall, a TIR average greater than 70% was achieved, which is higher compared to our previous studies.

Based on the 180–250 mg/dL range, more than 75% of the patients remained under 180 mg/dL throughout the evaluation period. Similar observations can be seen to the >250 column as well, which indicates that it was only in certain scenarios where the blood glucose tended towards the larger values, but even those cases are within the permissible limit which limit is 5%.

The RMSE150 variable showed the root mean squared error from the 150 mg/dL glucose concentration; 50% of the data deviated from 150 by less than 37 mg/dL. Whereas if we also look at the upper quartile, 65 mg/dL is a reasonable result, as it can be said that a quarter of the test values range between 85 mg/dl and 215 mg/dL.

Table 3. Aggregated table for all investigated simulation days, including mean, maximum, minimum, standard deviation, and quarterlies values in terms of TIR zones.

	<50	**50–70**	**70–180**	**180–250**	**>250**	**RMSE90**	**RMSE150**
mean	11.925	14.013	73.261	0.512	0.289	94.777	69.905
std	21.647	17.288	28.350	1.482	1.466	79.771	67.590
min	0.000	0.000	18.056	0.000	0.000	20.016	17.106
25%	0.000	0.000	44.792	0.000	0.000	43.382	31.384
50%	0.000	8.333	86.806	0.000	0.000	61.408	37.472
75%	11.458	22.222	100.000	0.000	0.000	109.674	65.366
max	73.264	73.264	100.000	9.375	11.458	360.283	309.865

Figure 7 shows median and standard deviation of the 221 days simulated in the test. The median value was calculated for the blood glucose values of this simulation. This median value is also plotted on two sub-bars one, with only the median value and the other with the median value and values between the 10th and 90th percentiles. In addition, the carbohydrate intake is also shown. The median value is in the right range which shows satisfying performance. It is also a satifying result that the 10th percentile values are also in this range. However, there are patients whose simulated values go up to 400 mg/dL that indicating that the robustness of the system should be increased (this will be the focus of our further studies). These are mainly light-weight patients.

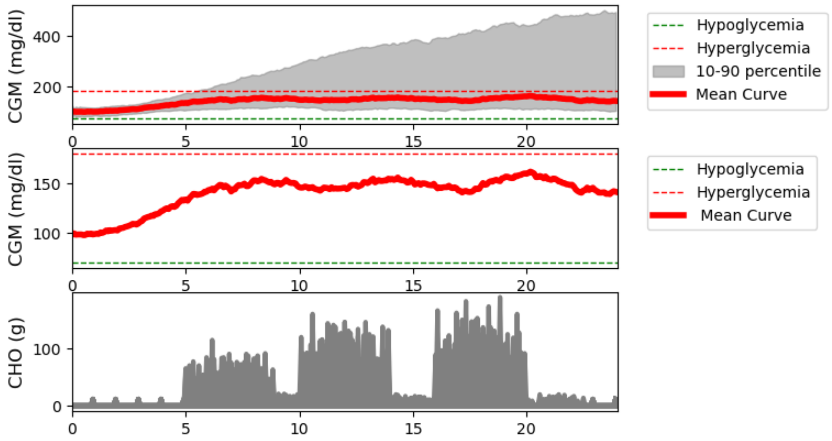


Figure 7. Blood glucose median curve with standard deviations and meal intakes for all the investigated days.

The Figure 8 plot is similar to the Figure 7 plot. The difference is that the mean values are plotted for the data simulated during the test. Also, it is not the percentile difference that is shown, but the standard deviation +/−2. Notice that the mean also moved very much into the correct range until it left the range toward the end of the simulation. There is a noticeable steady increase in the blood glucose curve. This is due to the fact that there are simulations where the neural network model does not dose insulin, so there is a continuous increase in the blood glucose value. Since the average is sensitive to this, the simulation was in the high blood sugar range. The variance being large it is completely apparent. This also proves that the robustness of the RL controller is not satisfying in case of patients with lower weights. However, it is a satisfying result in that there is no high difference between median and mean blood glucose values.

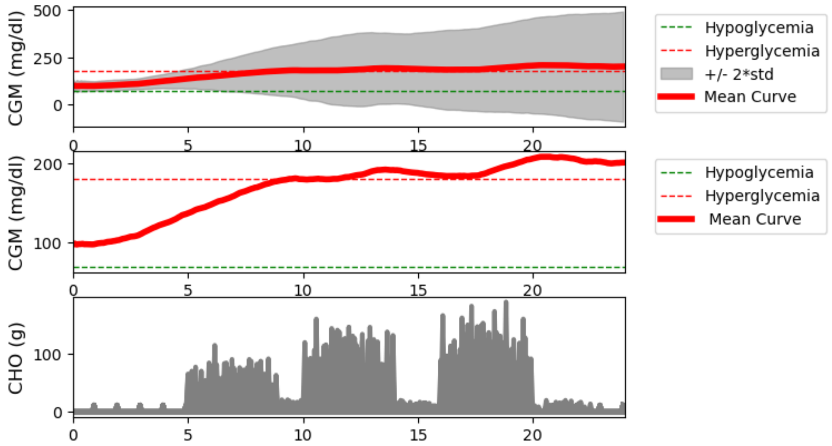


Figure 8. Mean blood glucose curve with standard deviation and meal intakes for all the investigated days.

In Figure 9 the CVGA is given. A large portion of the samples lie in the Zone A and Zone B. Compared to around 1% Zone A in our previous study, Sections 3.2.2 and 3.2.3, the method now reached 22% fir Zone A, which is a significant improvement. Similarly, the number of Zone B occurrences also increased compared to our preliminary results in Sections 3.2.2 and 3.2.3 from around 30% to almost 50%. Zone C occurrences remained at the same level as before. Samples in Zone Upper D were greatly reduced, approximately halved. Also, zone E samples were minimized. The spikes can be seen in both the high blood sugar level section and the low blood sugar level section. At each of the two extremes (above 400 mg/dL and lower than 50 mg/dL alongside the axis) produced by the virtual patients with lower weights (lower than cca. 65 kg). In our further study, we will investigate this phenomenon. In general, the model mostly avoids hypoglycemia and hyperglycemia, however, in some cases, both hypo- and hyper glycemia occur.

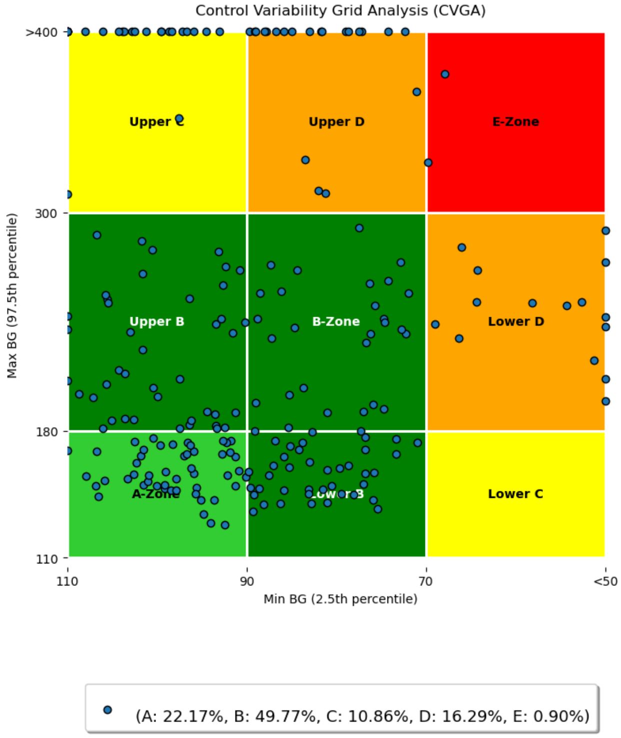


Figure 9. CVGA diagram for all tested days

5. Discussion

After illustrating our results, we will now look in more depth at all the results we have achieved. Let us start with the Table 3, where it was already mentioned that the aggregate results are shown for all simulated days. For the value <50, notice that half of the simulated days have a value of 0, since the value of 50% has a value of 0. For the value 75% there is a value of 11. This means that 25% of the simulated days fell into the very hypoglycemic zone. There were 55 days where the blood sugar went below 50 mg/dL during the simulation. What is worse is that in the max row, you can see a value of 73, which shows that there are patients for whom our model absolutely cannot work. These cases occur in patients with a lighter body weight. The model suddenly injects a large amount of insulin into the patient, causing a very low blood glucose level, and the patient is unable to get back into the correct range. This can be avoided by not administering insulin when blood glucose levels are already very low. Looking at the next column, which is 50–70, we see that at least half of the simulations are below 9. However, a quarter of the simulated days have no value in this zone. In row, 75% the value doubled. So for many patients, blood glucose levels move in this zone during the simulation. According to the TIR metrics, you should not spend more than 5 min in this zone. However, it can be said that if the value does not go below 50 mg/dL, it is not a serious problem. Now the most important column is the TIR metric value. It is very good that more than 50% of the simulated days satisfy the TIR metric with a value greater than 70 in column 70–180. In fact, the 75% row shows that a quarter of the simulated days are 100% in this zone. The average value is also greater than 70%, which was not seen in our previous tests. Furthermore, 25% should show that a quarter of the simulated days can only achieve less than 40%. The next two columns will be discussed together. To the extent that our previous studies were large, the values of these columns are minimal here to show that the models do not learn by not injecting insulin. However, it can be seen that you can have days where there were high blood glucose values. These also tend to happen in lightweight patients, Because the model reacts too late to sudden spikes in blood sugar when there is carbohydrate intake.

Next are Figures 7 and 8, where the simulation results for all days are shown, including the mean and median curves calculated from these days. For the mean, +/−2 standard deviations, and for the median, the value between the 10th and 90th percentiles are plotted. In addition, we examined the carbohydrate intakes. Starting with the carbohydrate intakes, the graphs clearly show that the three main meals appear in larger sizes. And the shifts are also clearly visible. This would prove that very random days were used to perform the tests. In the case of the median curve, it is clearly visible that our model structure picture of staying in the TIR zone even at the end of the day did not allow us to leave this zone. However, it can be seen that at the beginning. the models allowed the blood glucose level to rise and then stabilize around 150 mg/dL. Examining the picture on which the percentiles are also shown, we can say that the TIR range is not allowed to be exceeded even for the 10th percentile. The bottom of the variance is still within the range. However, for the 90th percentile, we can see that the upper range reached 400 mg/dL, which is very high. So, our model structure still does not handle hyperglycaemia in the best way. On the other hand, it handles hypoglycaemia very well. The vast majority of tests are not even in the hypoglycemic range. What you will notice is that there is a steady increase in blood glucose. This proves that 10 days of study is too long. The following should reduce the simulation time used for teaching. Let us turn to the figure showing the average. Unlike the median value, the average value here is out of the TIR range by the end of the day. And you can see a steadily increasing slope. This proves our previous claim that the learning simulation time is too long. However, the average curve did not exceed 200 mg/dL. Looking at the scatter plot, it can be seen that the maximum value can reach 500. So, we can say that there are patient cases for which our solution does not work absolutely well. This is true for patients with light weight. Because the model in their case reacts late or not at all—because for these patients, if the insulin is administered before the time of administration, it causes a very big change in their blood glucose. Next is the CVGA Figure 9 shows all the test days. It can be seen that in one A aand in Zone B which is the most important in terms of the goodness of control our value at 72% which means almost three-quarters of the days that we can solve with good control. It can also be seen that the days are shifted to the upper range for bad controls, as there being much more for Upper C and Upper D than for Lower C and Lower D. For our Zone E control, which is the completely bad control, we have hardly any value of days 1% is there. Looking at the graph, this means specifically one day and two are on the borderline. So we can say for the vast majority of days our solution works well. However, in the cases where it does not work well, most of the time the control moves toward the high blood glucose side. To a lesser extent, it moves to the low blood sugar.

6. Conclusions

We developed a reinforcement learning-based closed-loop controller that is able to bring the median TIR metric value towards 80%. In addition, the number of Zone A and Zone B samples in the CVGA analysis was substantially increased compared to our previous studies. This new approach is able to generally stabilize the glycemic curves, greatly improving the TIR values at the cost of a slight increase in the TBR values. Hypoglycemias appeared to occur in patients weighing less than 65 kg: finding methods to effectively and safely control patients with lower body weight is a goal in future studies, since apparently the current model does not generalize well in their case. Simulation time is also of concern and should be reduced in further studies. Other issues to be tackled concern the policy algorithm and the fact that during tests the PPO model could fail by getting stuck in a local minimum. In addition, learning with a much larger feature vector should be evaluated, since this could solve the problem of administering excessive doses of insulin at once.

Author Contributions: Conceptualization, L.D.-F. and G.E.; Methodology, L.D.-F. and L.S.; Validation, L.D.-F.; Resources, L.S.; Data curation, L.S.; Writing—original draft, L.D.-F. and A.D.G.; Writing—review & editing, A.D.G.; Visualization, L.D.-F. and A.D.G.; Supervision, L.K. and G.E.; Project administration, L.K. and G.E.; Funding acquisition, L.S. and G.E. All authors have read and agreed to the published version of the manuscript.

Funding: Project no. 2019-1.3.1-KK-2019-00007. has been implemented with the support provided from the National Research, Development and Innovation Fund of Hungary, financed under the 2019-1.3.1-KK funding scheme. This project has been supported by the National Research, Development, and Innovation Fund of Hungary, financed under the TKP2021-NKTA-36 funding scheme. The work of L. Szilágyi, A. de Gaetano and G. Eigner were partially supported by the Researcher Program of Óbuda University.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: The original contributions presented in the study are included in the article, further inquiries can be directed to the corresponding author.

Acknowledgments: On behalf of the AI development for diabetic and brain MRI scans project, we are grateful for the possibility to use the ELKH Cloud (see Héder et al. 2022; https://science-cloud.hu/, accessed on 20 February 2024), which helped us achieve the results published in this paper.

Conflicts of Interest: The authors declare no conflicts of interest.

Abbreviations

The following abbreviations are used in this manuscript:

DM	Diabetes mellitus
T1DM	Type 1 Diabetes Mellitus
T2DM	Type 2 Diabetes Mellitus
AP	Artificial pancreas
CGMS	Continuous Glycemia Monitoring System

PID	Proportional–Integral–Derivative
MPCs	Model Predictive Controllers
RL	Reinforcement Learning
IVP	Identifiable Virtual Patient model
GEZI	Glucose effectiveness at Zero Insulin concentration
EGP	Endogenous Glucose Production
PPO	Proximal Policy Optimization
BG	Blood Glucose Levels
TIR	Time In Range
CVGA	Continuous Variability Grid Analysis

References

1. Da Silva Xavier, G. The Cells of the Islets of Langerhans. *J. Clin. Med.* **2018**, *7*, 54.
2. Ogrotis, I.; Koufakis, T.; Kotsa, K. Changes in the Global Epidemiology of Type 1 Diabetes in an Evolving Landscape of Environmental Factors: Causes, Challenges, and Opportunities. *Medicina* **2023**, *59*, 668.
3. Abela, A.G.; Fava, S. Why is the Incidence of Type 1 Diabetes Increasing? *Curr. Diabetes Rev.* **2021**, *17*, e030521193110.
4. Holt, R.I.; Cockram, C.; Flyvbjerg, A.; Goldstein, B.J. *Textbook of Diabetes*; John Wiley & Sons: Chichester, UK, 2017.
5. Janež, A.; Guja, C.; Mitrakou, A.; Lalic, N.; Tankova, T.; Czupryniak, L.; Tabák, A.G.; Prazny, M.; Martinka, E.; Smircic-Duvnjak, L. Insulin Therapy in Adults with Type 1 Diabetes Mellitus: A Narrative Review. *Diabetes Ther.* **2020**, *11*, 387–409.
6. Mendez, C.E.; Umpierrez, G. Management of the hospitalized patient with type 1 diabetes mellitus. *Hosp. Pract. (1995)* **2013**, *41*, 89–100.
7. Bassi, M.; Franzone, D.; Dufour, F.; Strati, M.F.; Scalas, M.; Tantari, G.; Aloi, C.; Salina, A.; d'Annunzio, G.; Maghnie, M.; et al. Automated Insulin Delivery (AID) Systems: Use and Efficacy in Children and Adults with Type 1 Diabetes and Other Forms of Diabetes in Europe in Early 2023. *Life* **2023**, *13*, 783.
8. Sherr, J.L.; Heinemann, L.; Fleming, G.A.; Bergenstal, R.M.; Bruttomesso, D.; Hanaire, H.; Holl, R.W.; Petrie, J.R.; Peters, A.L.; Evans, M. Automated Insulin Delivery: Benefits, Challenges, and Recommendations. A Consensus Report of the Joint Diabetes Technology Working Group of the European Association for the Study of Diabetes and the American Diabetes Association. *Diabetes Care* **2022**, *45*, 3058–3074. https://doi.org/10.2337/dci22-0018.
9. Stavdahl, Ø.; Fougner, A.L.; Kölle, K.; Christiansen, S.C.; Ellingsen, R.; Carlsen, S.M. The artificial pancreas: A dynamic challenge. *IFAC-PapersOnLine* **2016**, *49*, 765–772.
10. Tagougui, S.; Taleb, N.; Molvau, J.; Nguyen, É.; Raffray, M.; Rabasa-Lhoret, R. Artificial pancreas systems and physical activity in patients with type 1 diabetes: challenges, adopted approaches, and future perspectives. *J. Diabetes Sci. Technol.* **2019**, *13*, 1077–1090.
11. Cobelli, C.; Renard, E.; Kovatchev, B. Artificial pancreas: past, present, future. *Diabetes* **2011**, *60*, 2672–2682.
12. Moon, S.J.; Jung, I.; Park, C.Y. Current Advances of Artificial Pancreas Systems: A Comprehensive Review of the Clinical Evidence. *Diabetes Metab. J.* **2021**, *45*, 813–839.
13. Hovorka, R. Closed-loop insulin delivery: from bench to clinical practice. *Nat. Rev. Endocrinol.* **2011**, *7*, 385–395.
14. Turksoy, K.; Cinar, A. Adaptive control of artificial pancreas systems - a review. *J. Healthc. Eng.* **2014**, *5*, 1–22.
15. Quiroz, G. The evolution of control algorithms in artificial pancreas: A historical perspective. *Annu. Rev. Control* **2019**, *48*, 222–232. https://doi.org/https://doi.org/10.1016/j.arcontrol.2019.07.004.
16. Jørgensen, J.B.; Boiroux, D.; Mahmoudi, Z. An artificial pancreas based on simple control algorithms and physiological insight. *IFAC-PapersOnLine* **2019**, *52*, 1018–1023. https://doi.org/https://doi.org/10.1016/j.ifacol.2019.06.196.
17. Batmani, Y.; Khodakaramzadeh, S.; Moradi, P. Automatic Artificial Pancreas Systems Using an Intelligent Multiple-Model PID Strategy. *IEEE J. Biomed. Health Inform.* **2022**, *26*, 1708–1717.
18. Matamoros-Alcivar, E.; Ascencio-Lino, T.; Fonseca, R.; Villalba-Meneses, G.; Tirado-Espín, A.; Barona, L.; Almeida-Galárraga, D. Implementation of MPC and PID Control Algorithms to the Artificial Pancreas for Diabetes Mellitus Type 1. In Proceedings of the 2021 IEEE International Conference on Machine Learning and Applied Network Technologies (ICMLANT), Soyapango, El Salvador, 16–17 December 2021; pp. 1–6. https://doi.org/10.1109/ICMLANT53170.2021.9690529.
19. Huyett, L.M.; Dassau, E.; Zisser, H.C.; Doyle, F.J. 3rd. Design and Evaluation of a Robust PID Controller for a Fully Implantable Artificial Pancreas. *Ind. Eng. Chem. Res.* **2015**, *54*, 10311–10321.
20. Kang, S.L.; Hwang, Y.N.; Kwon, J.Y.; Kim, S.M. Effectiveness and safety of a model predictive control (MPC) algorithm for an artificial pancreas system in outpatients with type 1 diabetes (T1D): systematic review and meta-analysis. *Diabetol. Metab. Syndr.* **2022**, *14*, 187.
21. Mauseth, R.; Hirsch, I.B.; Bollyky, J.; Kircher, R.; Matheson, D.; Sanda, S.; Greenbaum, C. Use of a "fuzzy logic" controller in a closed-loop artificial pancreas. *Diabetes Technol. Ther.* **2013**, *15*, 628–633.
22. Atlas, E.; Nimri, R.; Miller, S.; Grunberg, E.A.; Phillip, M. MD-logic artificial pancreas system: a pilot study in adults with type 1 diabetes. *Diabetes Care* **2010**, *33*, 1072–1076.
23. Lee, S.; Kim, J.; Park, S.W.; Jin, S.M.; Park, S.M. Toward a Fully Automated Artificial Pancreas System Using a Bioinspired Reinforcement Learning Design: In Silico Validation. *IEEE J. Biomed. Health Inform.* **2021**, *25*, 536–546.

24. de Farias, J.L.C.B.; Bessa, W.M. Intelligent Control with Artificial Neural Networks for Automated Insulin Delivery Systems. *Bioengineering* **2022**, *9*, 664.
25. Viroonluecha, P.; Egea-Lopez, E.; Santa, J. Evaluation of blood glucose level control in type 1 diabetic patients using deep reinforcement learning. *PLoS ONE* **2022**, *17*, e0274608. https://doi.org/10.1371/journal.pone.0274608.
26. Tejedor, M.; Woldaregay, A.Z.; Godtliebsen, F. Reinforcement learning application in diabetes blood glucose control: A systematic review. *Artif. Intell. Med.* **2020**, *104*, 101836. https://doi.org/10.1016/j.artmed.2020.101836.
27. Tašić, J.; Takács, M.; Kovács, L. Control Engineering Methods for Blood Glucose Levels Regulation. *Acta Polytech. Hung.* **2022**, *19*, 127–152. https://doi.org/10.12700/APH.19.7.2022.7.7.
28. Perolat, J.; De Vylder, B.; Hennes, D.; Tarassov, E.; Strub, F.; de Boer, V.; Muller, P.; Connor, J.T.; Burch, N.; Anthony, T.; et al. Mastering the game of Stratego with model-free multiagent reinforcement learning. *Science* **2022**, *378*, 990–996. https://doi.org/10.1126/science.add4679.
29. Liu, Y.; Xu, H.; Liu, D.; Wang, L. A digital twin-based sim-to-real transfer for deep reinforcement learning-enabled industrial robot grasping. *Robot.-Comput.-Integr. Manuf.* **2022**, *78*, 102365. https://doi.org/10.1016/j.rcim.2022.102365.
30. Liu, S.; See, K.C.; Ngiam, K.Y.; Celi, L.A.; Sun, X.; Feng, M. Reinforcement Learning for Clinical Decision Support in Critical Care: Comprehensive Review. *J. Med. Internet Res.* **2020**, *22*, e18477. https://doi.org/10.2196/18477.
31. Mughal, I.S.; Patanè, L.; Caponetto, R. A comprehensive review of models and nonlinear control strategies for blood glucose regulation in artificial pancreas. *Annu. Rev. Control* **2024**, *57*, 100937. https://doi.org/https://doi.org/10.1016/j.arcontrol.2024.100937.
32. Emerson, H.; Guy, M.; McConville, R. Offline reinforcement learning for safer blood glucose control in people with type 1 diabetes. *J. Biomed. Inform.* **2023**, *142*, 104376. https://doi.org/https://doi.org/10.1016/j.jbi.2023.104376.
33. Fox, I.; Lee, J.; Pop-Busui, R.; Wiens, J. Deep Reinforcement Learning for Closed-Loop Blood Glucose Control. *arXiv* **2020**, arXiv:cs.LG/2009.09051.
34. Gu, W.; Wang, S. An Improved Strategy for Blood Glucose Control Using Multi-Step Deep Reinforcement Learning. *arXiv* **2024**, arXiv:cs.AI/2403.07566.
35. Tejedor, M.; Hjerde, S.N.; Myhre, J.N.; Godtliebsen, F. Evaluating Deep Q-Learning Algorithms for Controlling Blood Glucose in In Silico Type 1 Diabetes. *Diagnostics* **2023**, *13*, 3150. https://doi.org/10.3390/diagnostics13193150.
36. Konda, V.; Tsitsiklis, J. Actor-Critic Algorithms. *Soc. Ind. Appl. Math.* **2001**, *42*, 1008–1014.
37. Schulman, J.; Wolski, F.; Dhariwal, P.; Radford, A.; Klimov, O. Proximal Policy Optimization Algorithms. *arXiv* **2017**, https://doi.org/10.48550/ARXIV.1707.06347.
38. Raffin, A.; Hill, A.; Gleave, A.; Kanervisto, A.; Ernestus, M.; Dormann, N. Stable-Baselines3: Reliable Reinforcement Learning Implementations. *J. Mach. Learn. Res.* **2021**, *22*, 1–8.
39. Schulman, J.; Levine, S.; Moritz, P.; Jordan, M.I.; Abbeel, P. Trust Region Policy Optimization. *arXiv* **2015**, https://doi.org/10.48550/ARXIV.1502.05477.
40. Kitouni, R.; Kitouni, A.; Jiang, F. Generalized Critic Policy Optimization: A Model For Combining Advantage Estimates In Actor Critic Methods. In Proceedings of the 2020 IEEE International Conference on Image Processing (ICIP), Abu Dhabi, United Arab Emirates, 25–28 October 2020; pp. 3184–3188. https://doi.org/10.1109/ICIP40778.2020.9190994.
41. Huang, S.; Kanervisto, A.; Raffin, A.; Wang, W.; Ontañón, S.; Dossa, R.F.J. A2C is a special case of PPO. *arXiv* **2012**, https://doi.org/10.48550/ARXIV.2205.09123.
42. Birck, M.; Corrêa, U.; Ballester, P.; Andersson Vianna, V.; Araujo, R. Multi-Task reinforcement learning: An hybrid A3C domain approach. In Proceedings of the Conference: ENIAC—Encontro Nacional de Inteligência Artificial e Computacional, Umberlandia, Brazil, 2–5 October 2017; pp. 1–8.
43. Brockman, G.; Cheung, V.; Pettersson, L.; Schneider, J.; Schulman, J.; Tang, J.; Zaremba, W. Openai gym. *arXiv* **2016**, arXiv:1606.01540.
44. Dénes-Fazakas, L.; Siket, M.; Kertész, G.; Szilágyi, L.; Kovács, L.; Eigner, G. Control of Type 1 Diabetes Mellitus using direct reinforcement learning based controller. In Proceedings of the 2022 IEEE International Conference on Systems, Man, and Cybernetics (SMC), Prague, Czech Republic, 9–12 October 2022; pp. 1512–1517. https://doi.org/10.1109/SMC53654.2022.9945084.
45. Haarnoja, T.; Zhou, A.; Abbeel, P.; Levine, S. Soft actor-critic algorithms and applications. *arXiv* **2018**, arXiv:1812.05905.
46. Lillicrap, T.P.; Hunt, J.J.; Pritzel, A.; Heess, N.; Erez, T.; Tassa, Y.; Silver, D.; Wierstra, D. Continuous control with deep reinforcement learning. *arXiv* **2015**, arXiv:1509.02971.
47. Mnih, V.; Kavukcuoglu, K.; Silver, D.; Rusu, A.A.; Veness, J.; Bellemare, M.G.; Graves, A.; Riedmiller, M.; Fidjeland, A.K.; Ostrovski, G.; et al. Human-level control through deep reinforcement learning. *Nature* **2015**, *518*, 529–533.
48. Mnih, V.; Badia, A.P.; Mirza, M.; Graves, A.; Lillicrap, T.; Harley, T.; Silver, D.; Kavukcuoglu, K. Asynchronous methods for deep reinforcement learning. *Int. Conf. Mach. Learn.* **2016**, *48*, 1928–1937.
49. Fujimoto, S.; van Hoof, H.; Meger, D. Addressing function approximation error in actor-critic methods. *Int. Conf. Mach. Learn.* **2018**, *80*, 1582–1591.
50. Kanderian, S.S.; Weinzimer, S.; Voskanyan, G.; Steil, G.M. Identification of Intraday Metabolic Profiles during Closed-Loop Glucose Control in Individuals with Type 1 Diabetes. *J. Diabetes Sci. Technol.* **2009**, *3*, 1047–1057.
51. Yoo, J.H.; Kim, J.H. Time in Range from Continuous Glucose Monitoring: A Novel Metric for Glycemic Control. *Diabetes Metab. J.* **2020**, *44*, 828–839.

52. Clarke, W.; Kovatchev, B. Statistical tools to analyze continuous glucose monitor data. *Diabetes Technol. Ther.* **2009**, *11* (Suppl. S1), S45–S54.
53. Lehel, D.F.; Siket, M.; Szilágyi, L.; Eigner, G.; Kovács, L. Investigation of reward functions for controlling blood glucose level using reinforcement learning. In Proceedings of the 2023 IEEE 17th International Symposium on Applied Computational Intelligence and Informatics (SACI), Timisoara, Romania, 23–26 May 2023; pp. 000387–000392. https://doi.org/10.1109/SACI58269.2023.10158621.
54. Fry, R.; McManus, S. Smooth bump functions and the geometry of banach spaces: A brief survey. *Expo. Math.* **2002**, *20*, 143–183. https://doi.org/https://doi.org/10.1016/S0723-0869(02)80017-2.
55. Singh, A.; Rawat, A.; Raghuthaman, N. Mexican Hat Wavelet Transform and Its Applications. In *Proceedings of the Methods of Mathematical Modelling and Computation for Complex Systems*; Singh, J., Dutta, H., Kumar, D., Baleanu, D., Hristov, J., Eds.; Springer: Cham, Switzerland, 2022; pp. 299–317. https://doi.org/10.1007/978-3-030-77169-0_12.
56. Lehel, D.F.; Siket, M.; Szilágyi, L.; Eigner, G.; Kovács, L. Effect of Hyperparameters of Reinforcement Learning in Blood Glucose Control. In Proceedings of the 2023 IEEE International Conference on Systems, Man, and Cybernetics (SMC), Maui, HI, USA, 1–4 October 2023; pp. 1333–1340. https://doi.org/10.1109/SMC53992.2023.10393930.
57. Kanderian, S.S.; Weinzimer, S.A.; Steil, G.M. The identifiable virtual patient model: comparison of simulation and clinical closed-loop study results. *J. Diabetes Sci. Technol.* **2012**, *6*, 371–379.
58. Vettoretti, M.; Battocchio, C.; Sparacino, G.; Facchinetti, A. Development of an Error Model for a Factory-Calibrated Continuous Glucose Monitoring Sensor with 10-Day Lifetime. *Sensors* **2019**, *19*, 5320. https://doi.org/10.3390/s19235320.
59. Huyett, L.M.; Dassau, E.; Zisser, H.C.; Doyle, F.J. Glucose Sensor Dynamics and the Artificial Pancreas: The Impact of Lag on Sensor Measurement and Controller Performance. *IEEE Control Syst. Mag.* **2018**, *38*, 30–46. https://doi.org/10.1109/MCS.2017.2766322.
60. Wang, Z.; Xie, Z.; Tu, E.; Zhong, A.; Liu, Y.; Ding, J.; Yang, J. Reinforcement Learning-Based Insulin Injection Time And Dosages Optimization. In Proceedings of the 2021 International Joint Conference on Neural Networks (IJCNN), Virtual, 18–22 July 2021; pp. 1–8. ISSN: 2161-4407. https://doi.org/10.1109/IJCNN52387.2021.9533957.
61. Paszke, A.; Gross, S.; Massa, F.; Lerer, A.; Bradbury, J.; Chanan, G.; Killeen, T.; Lin, Z.; Gimelshein, N.; Antiga, L.; et al. PyTorch: An Imperative Style, High-Performance Deep Learning Library. In *Advances in Neural Information Processing Systems 32*; Curran Associates, Inc.: Red Hook, NY, USA, 2019; pp. 8024–8035.
62. Héder, M.; Rigó, E.; Medgyesi, D.; Lovas, R.; Tenczer, S.; Török, F.; Farkas, A.; Emődi, M.; Kadlecsik, J.; Mező, G.; et al. The Past, Present and Future of the ELKH Cloud. *Információs Társadalom* **2022**, *22*, 128. https://doi.org/10.22503/inftars.xxii.2022.2.8.
63. Smith, J.R.; Johnson, E.S. Control Variability Grid Analysis: A Systematic Approach for Assessing Control System Performance. *Control Syst. Mag.*
64. Johnson, M.A.; Williams, S.K. Comparative Analysis of Advanced Control Algorithms Using Control Variability Grid Analysis. *Int. J. Control Autom.*
65. Chen, L.; Zhang, W.; Wang, Q. Optimizing Process Control Strategies using Control Variability Grid Analysis. *J. Process Eng.*
66. Battelino, T.; Danne, T.; Bergenstal, R.M.; Amiel, S.A.; Beck, R.W.; Biester, T.; Bosi, E.; Buckingham, B.A.; Cefalu, W.T.; Close, K.L.; et al. Clinical Targets for Continuous Glucose Monitoring Data Interpretation: Recommendations From the International Consensus on Time in Range. *Diabetes Care* **2020**, *43*, 1593–1603.
67. Beck, R.W.; Bergenstal, R.M.; Cheng, P.; Kollman, C.; Li, Z. Time in Range as a Metric for Reporting and Clinical Targets in People with Diabetes. *Diabetes Care* **2018**, *41*, 1891–1899.
68. Association, A.D. Consensus Report: Standards of Medical Care in Diabetes—2022. *Diabetes Care* **2022**, *45*, S3–S356.
69. Danne, T.; Nimri, R.; Battelino, T.; Bergenstal, R.M.; Close, K.L.; DeVries, J.H.; Garg, S.; Heinemann, L.; Hirsch, I.; Amiel, S.A.; et al. International Consensus on Use of Continuous Glucose Monitoring. *Diabetes Care* **2017**, *40*, 1631–1640.

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.