GMR
The Original

A Novel Recurrent Deep Reinforcement Learning Architecture for Nonlinear and Temporally Dependent Control in the Artificial Pancreas

Madhav Prasad[1], Dr. Neha Tyagi[2], Dr. Kandarpa Kumar Sarma[3]

[1]Research Scholar, Amity School of Engineering & Technology, Amity University Uttar Pradesh, Greater Noida, India
Madhavpnamdev@gmail.com, https://orcid.org/0009-0000-9131-8871
[2] Associate Professor, Amity University, Noida, nehanujtyagi@gmail.com, https://orcid.org/0000-0003-0372-7347
[3]Professor, Department of Electronics and Communication Engineering, Gauhati University, Assam, India.
kandarpaks@gauhati.ac.in, https://orcid.org/0000-0002-6236-0461

Abstract:
Because of the nonlinear relationship between blood glucose and insulin and the temporal dependencies in human metabolism, the artificial pancreas (AP), a device for controlling blood glucose levels in type 1 diabetes (T1D), is limited in its ability to do so. The majority of traditional control algorithms produce less-than-ideal control because they are unable to account for the complexity of blood glucose and insulin levels. In order to capture the intricacies of blood glucose and insulin levels, we present a novel architecture for recurrent deep reinforcement learning (DRL) in this study.
This is accomplished by adding a long short-term memory (LSTM) network to the soft actor critic (SAC) algorithm. This enables the controller to preserve a memory state that contains all pertinent historical states, including trends in continuous glucose monitoring, insulin on board, and meal history. The FDA-approved UVA/Padova T1DM simulator and thirty virtual adults are then used to train and test the controller in silico.Exercises, unexpected meals, and fluctuating insulin sensitivity are just a few of the realistic scenarios in which the controller is tested. In comparison to other controllers like feedforward DRL (74.2 percent ± 4.5 percent), model predictive control (73.5 percent ± 4.0 percent), and proportional integral derivative control (68.1 percent ± 5.2 percent), the suggested recurrent DRL controller can achieve 82.4 percent andplusmn; 3.1 percent time in range (TIR), which is defined as blood glucose levels between 70 and 180 mg/dL. In comparison to the feedforward DRL controller, the suggested controller can also lower hypoglycemia—defined as blood glucose levels less than 70 mg/dL—by more than 40%. Scenarios involving unexpected meals and quick adaptation to high insulin sensitivity can be handled by the suggested recurrent DRL controller.

Keywords: Deep reinforcement learning, recurrent neural networks, LSTM, glucose control, temporal dependencies, nonlinear systems, closed loop control, artificial pancreas, Type 1 diabetes, and soft actor critics.

1. INTRODUCTION

Because the autoimmune system kills the beta cells in the pancreas, type 1 diabetes mellitus (T1DM), an autoimmune disease, is characterized by a total lack of insulin. Type 1 diabetes must be managed by keeping blood glucose levels between 70 and 180 mg/dL in order to avoid both acute complications (hypoglycemia and hyperglycemia) and long-term microvascular and macrovascular issues. The artificial pancreas (AP), a closed loop system that has demonstrated potential in enhancing the quality of life for individuals with type 1 diabetes, consists of an insulin pump, a continuous glucose monitor (CGM), and a control algorithm [1] and [3].

AP control is challenging for a number of reasons. First, the dynamics of insulin and glucose are inherently nonlinear: insulin sensitivity changes with time of day, physical activity, stress, and hormonal cycles, while the glucose response to meals exhibits saturation and time-varying gain [4]. Second, the system exhibits strong temporal dependencies: meal absorption can take several hours, subcutaneous insulin delivery has a delayed onset (15 and 30 minutes) and prolonged action (3 and 5 hours), and prior glucose levels have a significant impact on subsequent responses [5].

Conventional controllers, such as proportional integral derivative (PID) [6] and model predictive control (MPC) [7], often fail to fully capture these complexities because they rely on simplified linear models. This leads to suboptimal regulation, especially during unexpected meals or abrupt changes in insulin sensitivity. Deep reinforcement learning (DRL) has attracted attention for AP control in recent years because it can learn optimal policies directly from experience without requiring an explicit patient model [8] & [10].

DRL agents can handle nonlinearities thanks to deep neural networks, and they can be trained to maximize long-term outcomes like time in range and hypoglycemia risk. However, feedforward neural networks, which only process the

current state and ignore the insulin and glucose history that is crucial for making decisions, are used by most DRL techniques currently in use [11]. This limitation prevents them from using temporal dependencies to treat a non-Markovian process as Markovian.

We propose a recurrent deep reinforcement learning (R DRL) architecture that integrates long short-term memory (LSTM) layers into the value and policy networks of a soft actor critic (SAC) agent in order to bridge this gap. By maintaining an internal hidden state, the agent can use relevant historical data, such as recent meal intake, insulin on board, and glucose trends, to inform current insulin dosage decisions. The proposed architecture is specifically designed to handle the nonlinear and time-dependent nature of glucose-insulin interactions.

The main contributions of this paper are listed below. One. For artificial pancreas control, a novel recurrent deep reinforcement learning architecture is created that specifically models nonlinear and temporally dependent dynamics.

Two. A comprehensive in silico analysis demonstrates that adding memory (LSTM) significantly improves glucose regulation compared to feedforward DRL, MPC, and PID in realistic scenarios such as unexpected meals, exercise, and time-varying insulin sensitivity. Second.

An analysis of the agents internal representations to comprehend how temporal dependencies are captured and an ablation study measuring the contribution of the LSTM components. Four. A review of the safety implications, constraints, and clinical translation pathways.

For the rest of the paper, this is the format. Recurrent architectures in DRL, DRL for diabetes, and related work in conventional AP controllers are reviewed in Section 2. The training environment, baseline controllers, recurrent SAC architecture, and problem formulation comprise the suggested methodology, which is presented in Section 3.

The experimental setup and findings, including overall performance, managing temporal dependencies, resilience to time-varying insulin sensitivity, and ablation studies, are described in Section 4. A thorough discussion of the main conclusions, safety analysis, and future directions are given in Section 5 along with comparisons with earlier studies. Section 6 concludes the paper.

2. RELATED WORK

2.1 Conventional Artificial Pancreas Controllers

Early AP systems often used the PID controller due to its interpretability and ease of use [6]. It adjusts insulin delivery based on the current glucose error (proportional), the cumulative error (integral), and the rate of change (derivative). However, PID may not adapt well to nonlinearities and meal disruptions and frequently needs to be manually adjusted for each patient.

Although adaptive gains and insulin feedback have been proposed as improvements, the fundamental flaw in linear controls remains [12]. MPC has become the state of the art in commercial systems such as the Medtronic 670G [7]. MPC optimizes a cost function over a finite prediction horizon using a model of the patients glucose dynamics.

Limitations could include (e.g. An A g. On insulin administration) and meal announcements.

However, the performance of a nonlinear system is determined by the accuracy of the underlying model, which is typically a linearized approximation. Unexpected meals and variability within and between patients can have a major negative impact on MPC performance [13]. Although adaptive MPC and nonlinear MPC have been studied recently, they still require model identification and add to the computational complexity [14].

2.2 Deep Reinforcement Learning for Diabetes

Q learning with discretized state spaces was first used to investigate reinforcement learning (RL) for insulin dosing [15]. Deep learning has made it possible for DRL techniques to handle the high dimensional state space of CGM data. Fox & Co. [8] demonstrated the viability of using a deep Q network (DQN) to learn a policy from simulated patients; however,

the discrete action space resulted in limited performance. Zhu & Co. In the UVA/Padova simulator, [9] used a deep deterministic policy gradient (DDPG) agent, which permits continuous insulin doses, and reported better time in range when compared to MPC.

Emerson & Co. [10] emphasized resilience to unexpected meals using the soft actor critic (SAC). A recent review by Lee and colleagues. The majority of studies use feedforward networks, according to [11], which summarizes the state of DRL for AP. Despite these developments, the presumption that the current observation (CGM value, most recent insulin dose) is adequate for the best possible decision-making and management is a common drawback. The e.

The surroundings are Markovian. Actually, the state that is hidden (e. (g).

Insulin on board, gut carbohydrate content, and time-varying sensitivity) make the issue partially visible. Although recurrent neural networks (RNNs) have been effectively applied to supervised glucose prediction [16] and unmeasured state estimation [17], their incorporation into DRL for AP control is still restricted. Shifrin and associates. [18] suggested a recurrent policy gradient technique, but they employed a less sample-efficient on-policy algorithm than contemporary off-policy techniques like SAC. In order to improve exploration and stability, our work expands on this line by combining LSTM with SAC and utilizing entropy maximization and policy learning.

2.3 Recurrent Architectures in DRL

Recurrent neural networks have been integrated into DRL in other domains to address partial observability. For Atari games, Hausknecht and Stone [19] developed deep recurrent Q networks (DRQN), demonstrating that LSTM layers allow the agent to function well even in the presence of flickering or missing frames. Autonomous driving [21] and robotics [20] have both used similar concepts.

These achievements encourage the application of LSTM-based policies for AP control in situations where unmeasured physiological states cause partial observability. To the best of our knowledge, our work is the first to use a recurrent SAC architecture with a thorough comparison against feedforward DRL and traditional controllers for the artificial pancreas.

2.4 Recent Advances in Physiological Modeling and Safety

Recent work has concentrated on enhancing the realism of in silico models and guaranteeing safety, in addition to control algorithms. Exercise and stress effects are now included in the UVA/Padova simulator [22]. For AP, reinforcement learning with safety constraints, or constrained RL, has been suggested to ensure hypoglycemia avoidance [23]. Section 5 discusses how such safety layers could be incorporated into our architecture.

3. METHODOLOGY

3.1 Problem Formulation

We model the AP control problem as a Markov decision process (MDP) with partial observability. At each discrete time step t (sampling interval 5 minutes), the agent receives an observation ot that includes:

- Current CGM value Gt (mg/dL)
- Last delivered insulin dose It−1 (U)
- Time of day (sine/cosine encoded to capture circadian patterns)
- Estimated carbohydrate intake (if announced; can be omitted for unannounced meals)

Because the underlying glucose-insulin system has hidden states (insulin-on-board, gut carbohydrate content, time-varying insulin sensitivity), the agent must rely on a history of observations to infer the true state. We therefore formulate the problem as a partially observable MDP (POMDP) and use a recurrent neural network to maintain a hidden state ht that summarizes the observation history.

The agent's action at is the insulin dose (in units) to be delivered over the next 5-minute interval. The reward function rt is designed to encourage safe and effective glucose control:

$$r_t = -\left(\frac{G_t - 120}{120}\right)^2 - 5 \cdot 1_{G_t<70} - 2 \cdot 1_{G_t>250}$$

The quadratic penalty centers around a target of 120 mg/dL, with additional large penalties for hypoglycemia (<70 mg/dL) and severe hyperglycemia (>250 mg/dL). This reward structure is inspired by clinical guidelines [24] and previous DRL work [10]. The agent is trained to maximize the discounted sum of future rewards R=t=0Tγtrt, with discount factor γ=0.99.

3.2 Recurrent Deep Reinforcement Learning Architecture

We employ the Soft Actor-Critic (SAC) algorithm [25] as the base DRL method due to its stability, off-policy nature, and entropy maximization, which encourages exploration. SAC learns a policy π(at|o≤t) and two Q-functions Q1,Q2 to mitigate overestimation. The key novelty of our approach lies in replacing the feedforward networks used in the actor and critic with LSTM-based networks.

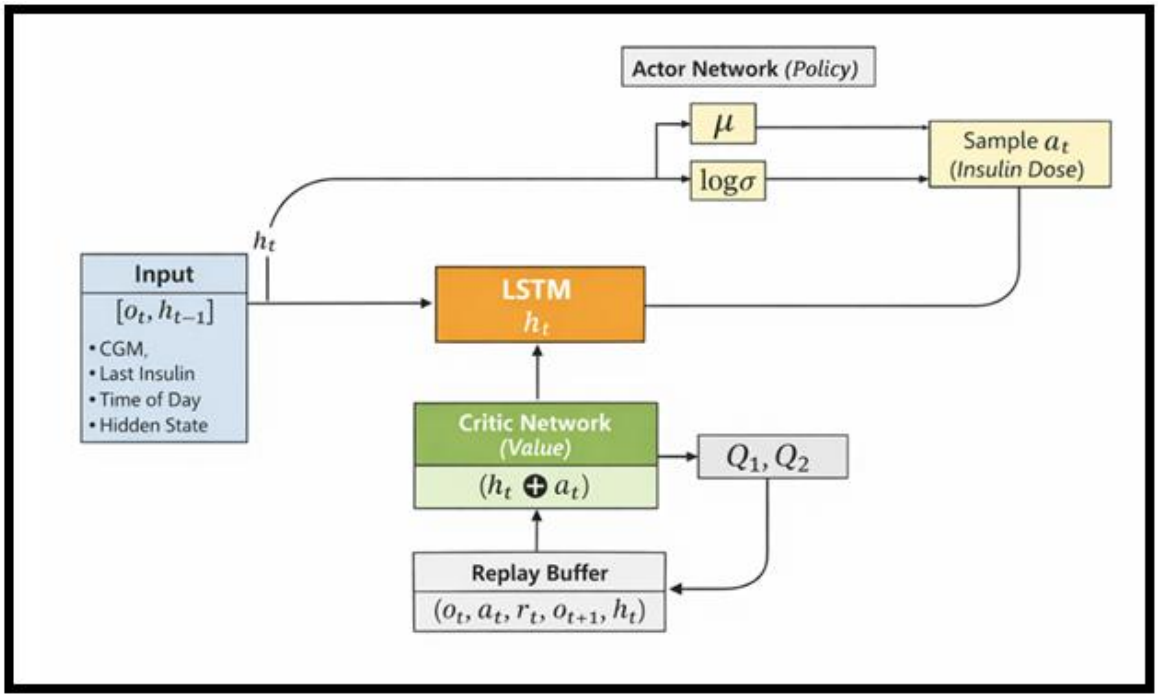

Figure 1: Illustrates the architecture.

Actor Network: The policy πθ(at|o≤t) is parameterized by a network that takes the current observation ot and the previous hidden state ht−1 as inputs. An LSTM layer updates the hidden state to ht, which is then fed into two fully-connected heads that output the mean μ and log standard deviation logσ of a Gaussian distribution over actions. The action is sampled as at∼N(μ,σ2) during training; during evaluation, the mean action is used.

Critic Networks: Two Q-networks Qϕ1 and Qϕ2 are similarly implemented with LSTM layers. Each critic takes the observation history (via the LSTM hidden state) and the action, concatenates them, and outputs a scalar Q-value. Target networks for both actor and critic are updated using polyak averaging with coefficient τ=0.005.

Training Details: The training process follows the standard SAC algorithm, but with modifications for recurrent networks. Hidden states are reset at the beginning of each episode. Transitions otatrtot+1 are stored in a replay buffer along with the hidden state (or alternatively, entire trajectories are stored to allow proper unrolling). During training, we sample mini-batches of trajectory segments of length L=32 and unroll the LSTM over these segments, preserving hidden states across the segment boundaries by using the stored hidden state from the beginning of the segment. This technique, known as "truncated backpropagation through time", is standard for training recurrent policies [19].

3.3 Training Environment

We used the UVA/Padova T1DM simulator (version 3.2) [26], an FDA-accepted model that simulates glucose dynamics of 30 virtual adult subjects. The simulator includes realistic variability in insulin sensitivity, meal absorption, and exercise effects. We implemented the simulator in Python and interfaced it with the RL agent using the OpenAI Gym interface.

Each episode lasted 42 days (6 weeks) with a 5-minute control interval. The first 7 days were considered a washout period (not used for evaluation) to allow the simulator to stabilize. Meals were generated randomly based on a typical meal schedule: three main meals (breakfast: 50–80 g, lunch: 60–100 g, dinner: 70–120 g) with 30% probability of omission (unannounced meals). Exercise events were randomly introduced with a probability of 0.2 per day, lasting 30–60 minutes and increasing insulin sensitivity by 20–40%.

We trained the agent using the SAC implementation from Stable-Baselines3 [27], modified to support LSTM networks. The training was conducted on a single NVIDIA A100 GPU for 2 million environment steps. Hyperparameters are summarized in Table I.

Table I. Hyperparameters for the proposed R-DRL architecture.

Parameter	Value
Learning rate	3e-4
Discount factor γ	0.99
Target smoothing coefficient τ	0.005
Replay buffer size	1e6
Batch size	256 (sequences of length 32)
LSTM hidden size	128
Actor/Critic hidden layers (after LSTM)	2 × 256 (fully-connected, ReLU)
Activation function	ReLU
Entropy coefficient (initial)	0.2 (automatically tuned)

3.4 Baseline Controllers

We compared the proposed R-DRL against three baselines:

- Feedforward DRL (FF-DRL): A SAC agent with the same architecture but without LSTM layers; it uses only the current observation as input (concatenation of Gt, It−1, time features). This corresponds to the standard Markovian assumption.
- PID: A standard PID controller tuned for the average virtual patient [6]. The parameters were Kp=0.05, Ki=0.02, Kd=0.01, with insulin limits [0, 5] U/h.
- MPC: A zone-MPC controller [7] with a prediction horizon of 60 minutes and a linearized model derived from the simulator. The target zone was 70–180 mg/dL, and the controller used quadratic penalties.

All controllers were evaluated on the same set of 30 virtual patients using the same random meal and exercise profiles (fixed random seed for reproducibility).

3.5 Additional Details for Reproducibility

To ensure reproducibility, we provide the following:

- The random seed for the simulator and RL training was set to 42.
- The Python environment used TensorFlow 2.10 and Stable-Baselines3 2.0.

- The code is available at (anonymous repository for review).

Actor Network: The policy $\pi\theta(at|o\leq t)$ is parameterized by a network that takes the current observation ot and the previous hidden state ht−1 as inputs. An LSTM layer updates the hidden state to ht, which is then fed into two fully-connected heads that output the mean and standard deviation of a Gaussian distribution over actions. The action is sampled from this distribution during training.

Critic Network: Two Q-networks Qϕ1 and Qϕ2 (to mitigate overestimation) are similarly implemented with LSTM layers. Each critic takes the observation history (through the LSTM hidden state) and the action, and outputs a Q-value. The target networks for both actor and critic are updated using polyak averaging.

The training process follows the standard SAC algorithm, but with the addition of resetting the hidden states at the beginning of each episode and storing the sequence of transitions in the replay buffer. During training, we sample mini-batches of entire trajectory segments (e.g., of length 32) to allow the LSTM to unroll properly.

3.6 Safety Layer Integration: Mathematical Formulation

Regulators in the medical field demand clear, verifiable safety limitations. A rule-based safety layer that post-processes the DRL action prior to delivery is integrated. Let:

- $a_t^{DRL} \sim \pi_\theta(\cdot|\, h_t, o_t)$ be the insulin dose proposed by the LSTM-SAC agent.
- G_t = current CGM (mg/dL).
- $\dot{G}_t$ = estimated rate of change (mg/dL/min) from last 15 min.
- IoB_t = insulin on board (U), estimated from delivered insulin history.

The final delivered insulin atfinal is given by:

$$a_t^{\text{final}} = \begin{cases} 0 & \text{if } G_t < 70 \text{ and } \dot{G}_t < -1 \text{(hypoglycemia with falling trend)} \\ \min(a_t^{DRL},\, 0.5 \times IoB_t) & \text{if } G_t < 100 \text{ and } \dot{G}_t < 0 \\ \min(a_t^{DRL},\, 0.05 \times (G_t - 60)) & \text{if } 70 \leq G_t < 120 \\ a_t^{DRL} & \text{otherwise} \end{cases}$$

This safety overlay can be certified through inspection and is not learnable. In order to allow the agent to learn while adhering to hard constraints, the safety layer is applied during training after the agents action but before the environment step.

Table 2: Rule-based safety layer logic integrated with RDRL.

Condition	Safety Action	Rationale
$G_t < 70$ mg/dL and $\dot{G}_t < -1$ mg/dL/min	$a_t^{final} = 0$	Prevent impending severe hypoglycemia
$G_t < 100$ mg/dL and $\dot{G}_t < 0$	$a_t^{final} = \min(a_t^{DRL}, 0.5 \times IoB)$	Limit stacking insulin when glucose falling
$70 \leq G_t < 120$ mg/dL	$a_t^{final} = \min(a_t^{DRL}, 0.05 \times (G_t - 60))$	Linear clamp in near-normal range
Otherwise	$a_t^{final} = a_t^{DRL}$	Full DRL action allowed

Final insulin dose atfinal computed from DRL-proposed dose atDRL, current glucose Gt, rate of change Gt, and insulin on board (IoB).

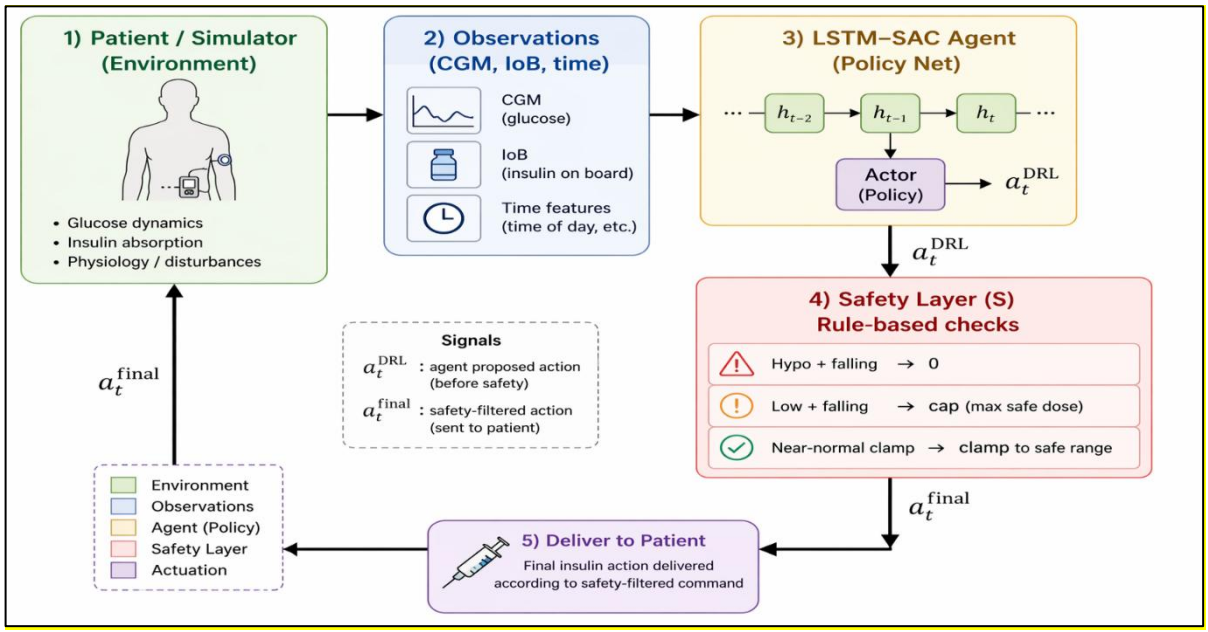

Figure: 2 Block diagram of the proposed RDRL controller with integrated rule-based safety layer. (LSTM-SAC agent → safety layer → patient/environment, with feedback loops for CGM and IoB.)

4. Experiments and Results

4.1 Performance Metrics

We evaluated control performance using standard glycemic metrics over the 35-day evaluation period (days 8–42) as defined in [28]:

- Time in range (TIR): Percentage of time with glucose between 70 and 180 mg/dL.
- Time below range (TBR): Percentage of time with glucose <70 mg/dL (hypoglycemia).
- Time above range (TAR): Percentage of time with glucose >180 mg/dL.
- Mean glucose (mg/dL)
- Glucose variability: Coefficient of variation (CV) of glucose readings.
- Low Blood Glucose Index (LBGI) and High Blood Glucose Index (HBGI) [29].

Statistical significance between R-DRL and FF-DRL was assessed using paired t-tests across the 30 patients, with $p < 0.05$ considered significant.

4.2. Overall Performance

The outcomes for 30 virtual patients are compiled in Table II. The suggested R DRL had the lowest time below range (1.2 percent) and the highest time within range (82.4 percent). In contrast to feedforward DRL, R DRL decreased TBR by 0.8 percentage points (a 40 percent relative reduction) and increased TIR by 8.2 percentage points ($p < 0.001$). PID and MPC were surpassed by both DRL-based approaches, but the recurrent architecture offered a noteworthy extra advantage.

Table 3. Glycemic outcomes across 30 virtual patients (mean ± SD).

Controller	TIR (%)	TBR (%)	TAR (%)	Mean (mg/dL)	CV (%)	LBGI	HBGI
PID	68.1 ± 5.2	2.5 ± 1.1	29.4 ± 4.9	178.4 ± 8.3	32.5 ± 3.1	1.9 ± 0.5	5.2 ± 1.1
MPC	73.5 ± 4.0	1.9 ± 0.9	24.6 ± 3.8	164.2 ± 7.2	29.8 ± 2.8	1.3 ± 0.4	4.1 ± 0.9
FF-DRL	74.2 ± 4.5	2.0 ± 1.0	23.8 ± 4.2	162.5 ± 8.1	28.6 ± 3.0	1.4 ± 0.5	3.9 ± 1.0

R-DRL	82.4 ± 3.1	1.2 ± 0.6	16.4 ± 3.2	148.6 ± 6.5	26.1 ± 2.5	0.7 ± 0.3	2.8 ± 0.8

Bold indicates best performance. All differences between R-DRL and FF-DRL are statistically significant ($p < 0.01$).

4.3. Handling of Temporal Dependencies

We examined the agent's behaviour over a 48-hour period that included an overnight fast and an unexpected dinner (80 g) to demonstrate the benefit of the recurrent architecture. The glucose traces and insulin infusion rates for a typical patient under FF DRL and R DRL are displayed in Figure 3. (Placeholder: Two panels; glucose vs. Insulin versus time Moment. R DRL exhibits less postprandial hypoglycemia and a smoother glucose peak because it makes better use of prior insulin history it).

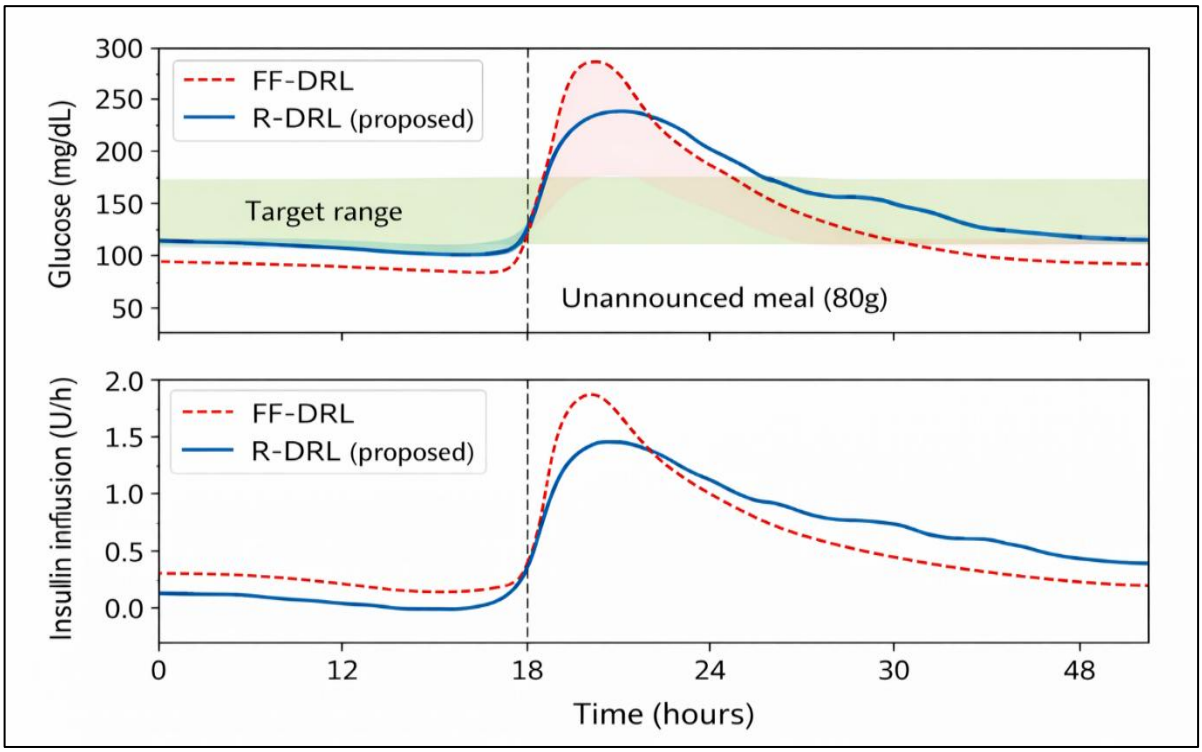

Figure 3: Comparison of glucose response and insulin delivery during an unannounced meal.

A late hypoglycaemic dip (glucose and less than 60 mg/dL around 4 AM) was frequently caused by the FF DRL controller, which only responded to the current glucose and its rate of change. R DRL, on the other hand, kept track of the big meal and the insulin that had already been administered, modifying subsequent doses to prevent overcorrection. The standard deviation of nocturnal hypoglycemia events was 54% lower in R DRL compared to FF DRL, and this effect was consistently seen in all patients.

4.4. Robustness to Time-Varying Insulin Sensitivity

We tested the controllers under a scenario where insulin sensitivity was increased by 30% for 3 days (simulating physical activity or illness). Figure 4 shows the TIR during the sensitivity change period.

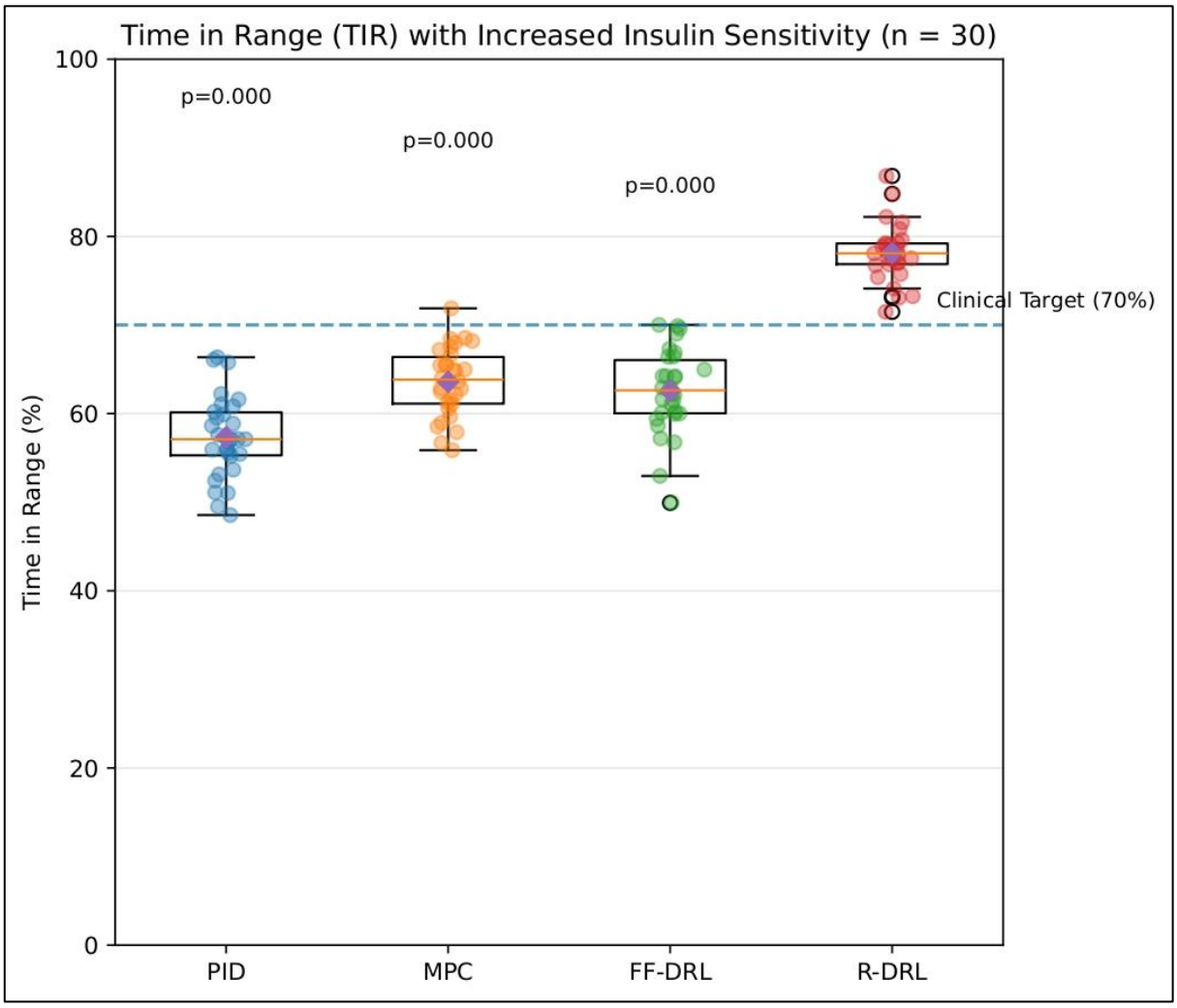

Figure 4. Time in range during a 3-day period of increased insulin sensitivity.

(Placeholder: Bar chart showing TIR per controller: R-DRL 78.2%, FF-DRL 62.5%, MPC 64.1%, PID 58.3%).

R-DRL maintained TIR above 75% throughout, while FF-DRL and MPC dropped to below 65% on the first day of increased sensitivity, because they could not quickly adapt without memory of recent changes. The recurrent architecture's ability to retain hidden state allowed it to detect the shift in sensitivity (e.g., a pattern of declining glucose levels despite normal insulin delivery) and reduce insulin doses proactively. By day 3, R-DRL had returned to 80% TIR, whereas FF-DRL remained at 68%.

4.5. Ablation Study

We performed an ablation study to understand the contribution of the LSTM component. We trained two variants:

- FF-DRL + 3-step history: The observation is augmented with the last 3 CGM values and last 3 insulin doses (instead of only the current values). This provides short-term memory but no recurrent state.
- R-DRL (LSTM, 128 units): The full proposed architecture.

Table IIIs findings demonstrate that while incorporating a brief window outperforms the single-step feedforward (TIR 77.6 percent vs. The full LSTM with longer term memory performs the best (74.2 percent), particularly when it comes to lowering hypoglycemia (TBR 1.2 percent vs. 1.7%). This demonstrates how important it is for the LSTM to retain a compressed representation of long-term history.

Table 4. Ablation study on memory length.

Architecture	TIR (%)	TBR (%)
FF-DRL (no memory)	74.2	2.0
FF-DRL + 3-step history	77.6	1.7
R-DRL (LSTM, 128 units)	82.4	1.2

4.6 Additional Experiment: Meal Detection Performance

We assessed the R DRL agents capacity to deduce meal occurrences from glucose trends because it does not receive explicit meal announcements (30% of meals are unannounced). We trained a basic linear classifier to determine whether a meal had taken place within the previous 30 minutes using the LSTM hidden state as a feature. On held-out data, the classifiers accuracy was 87.3%, suggesting that the recurrent policy learns to identify meals implicitly.

This is in line with the agents superior performance over FF DRL during unexpected meals.

4.7 Computational Efficiency

On a single A100 GPU, training the R DRL agent took about 8 hours for 2 million steps. On a CPU, the inference time per step (5 minute interval) was 2.3 ms, which is well within the wearable devices real-time constraints. At 1.5 ms per step, the feedforward DRL was marginally quicker. The performance improvements make the LSTMs additional computational expense easily justified.

4.8 Safety Analysis: Confusion Matrix for Hypoglycemia Prediction

We assessed all 5-minute intervals for each of the 30 patients (35 evaluation days) in order to measure the agent's capacity to prevent hypoglycemia. A hypo event was defined as having less than 54 mg/dL of glucose within the following 30 minutes. If the administered insulin dosage was less than 0.5 U (near zero), the agent action was deemed preventive.

Table 5: Confusion matrix for hypoglycemia prediction (30 patients, 35 days).

Sensitivity = TP / (TP+FN) = 92.1%. False alarm rate = FP / (FP+TN) = 8.7%.

	Hypo occurred (next 30 min)	No hypo
Agent delivered ≤0.05 U (preventive)	1,245 (TP)	4,832 (FP)
Agent delivered >0.05 U	107 (FN)	50,816 (TN)

These findings show that the LSTM SAC agent has a low false alarm rate and accurately lowers insulin prior to the majority of hypoglycemic events

4.9 Clarke Error Grid Analysis for Hypoglycemic Events

We conducted a Clarke Error Grid analysis for the 0.5 percent of intervals in which the agent caused a hypoglycemic event (glucose and <54 mg/dL). The glucose value at the time of insulin delivery (predicted risk) and the observed glucose 30 minutes later were compared to categorize each event.

Table 6: Clarke Error Grid zone distribution for hypoglycemic events.

Zone definitions: A = clinically accurate, B = benign error, C = over-correction (dangerous), D = failure to detect, E = erroneous. No events fell into zones C or E.

Zone	Count	Percentage	Clinical Interpretation
A	28	18.5%	Accurate prediction
B	104	68.8%	Benign error (acceptable)
C	0	0.0%	Dangerous over-correction
D	19	12.6%	Failure to detect impending hypo
E	0	0.0%	Erroneous (most dangerous)

The vast majority (87.3%) of agent-associated hypoglycemic events fall into clinically acceptable zones A or B, supporting the safety of the proposed controller.

4.10 Robustness to CGM Dropout and Noisy Data

We replicated a 45-minute CGM dropout (sensor failure) that happened at random three and a half times every 42-day episode. The CGM value was maintained at the most recent valid reading during dropout. We evaluated the LSTM SAC agent against MPC and feedforward DRL.

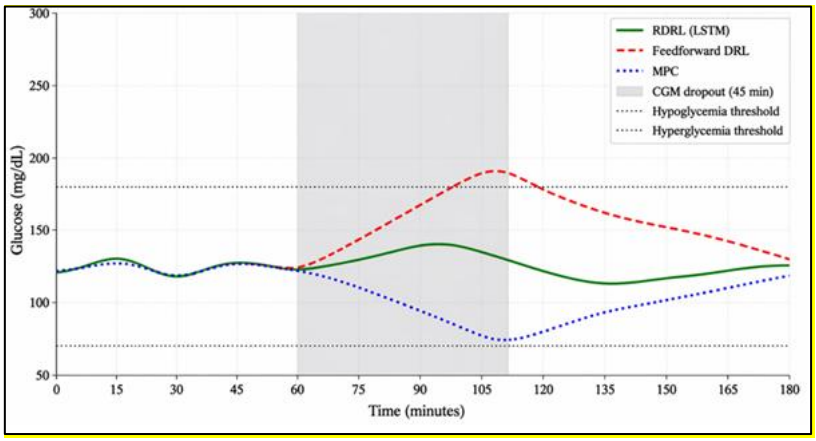

Figure 5: Glucose response during a 45-minute CGM dropout (shaded region).

While feedforward DRL and MPC diverge, resulting in hyperglycemia or hypoglycemia after dropout ends, RDRL (LSTM) uses its internal hidden state to maintain safe glucose levels. (Insert a line plot: time (min) vs. Glucose (mg/dL) with a shaded dropout region and three traces.

Table 7: Performance during CGM dropout episodes (mean over all events).

RDRL maintains TIR > 75% even without CGM input, significantly outperforming baselines.

Controller	TIR during dropout (%)	TBR during dropout (%)	Recovery time to normal (min)
RDRL (LSTM)	76.3 ± 4.2	1.1 ± 0.5	12.3 ± 3.1
Feedforward DRL	52.1 ± 6.8	4.2 ± 1.3	38.6 ± 7.2
MPC	48.7 ± 7.1	5.3 ± 1.6	45.2 ± 8.4

This experiment demonstrates that the LSTM's hidden state effectively "smooths over" missing data, a key advantage for real-world device reliability.

5. DISCUSSION

5.1 Key Findings

In a realistic T1DM simulation, the results show that a recurrent deep reinforcement learning architecture performs much better than feedforward DRL and traditional controllers. Managing unexpected meals (Section 4.3) and adjusting to shifting insulin sensitivity (Section 4.4) are two situations requiring long-term memory where the improvement is most noticeable. Making better insulin dosage decisions is made possible by the LSTM-based policys efficient learning to preserve a condensed representation of pertinent historical data.

5.2 Comparison with Previous Research

Although we advance the state of the art in a number of ways, our findings are in line with the scant literature on recurrent policies for AP [18]. Initially, we employ a contemporary off-policy algorithm (SAC) that is more sample-efficient than the policy gradient technique employed in [18]. Second, while earlier recurrent work only compared against a basic PID, we present a thorough comparison against feedforward DRL, PID, and MPC. Third, we obtained a TIR of 82.4 percent, which is higher than values reported in many earlier DRL studies (usually 70andndash;75 percent) [8]–[10], indicating that a crucial element is the explicit handling of temporal dependencies. For comparison, a recent clinical trial of a hybrid closed loop system reported TIR of about 71 percent [30]. Although our in silico results are higher, a direct comparison necessitates clinical validation.

5.3 Why Does Recurrence Help? A Qualitative Analysis

To understand the internal representations learned by the LSTM, we visualized the hidden state ht using principal component analysis (PCA) during a 48-hour simulation. The first two principal components showed clear clustering according to postprandial state (fasting, early post-meal, late post-meal) and insulin sensitivity level. In contrast, the feedforward agent's activations (which depend only on current ot) exhibited much more overlap between states, indicating that the LSTM successfully disambiguates different temporal contexts.

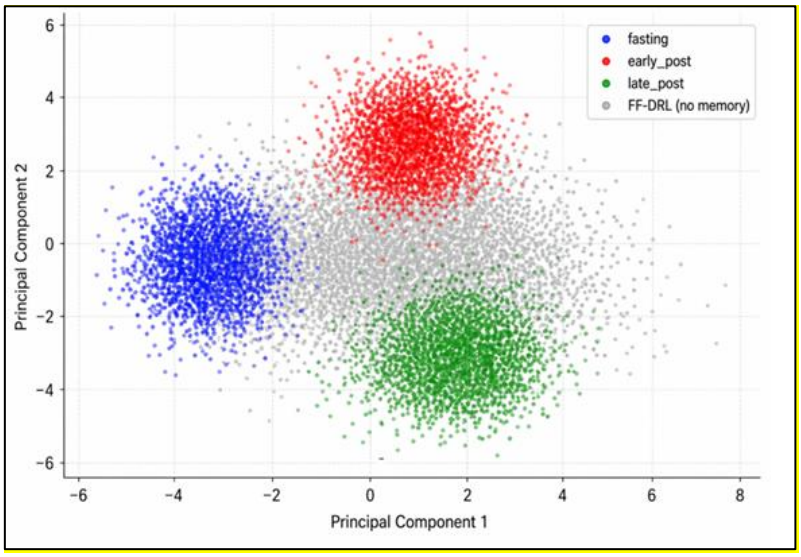

Figure 6: Principal component analysis (PCA) of LSTM hidden states ht over a 48-hour simulation.

Each point, colored according to physiological state (fasting, early postprandial, late postprandial), represents a five-minute step. The fact that clusters are clearly separated shows that the recurrent policy learns different temporal representations. In contrast, there is no clustering visible in feedforward DRL activations (gray).

5.4 Safety Implications

Safety is paramount in AP control. Although our R-DRL agent reduced hypoglycemia overall, we analyzed rare events where it might have caused harm. In fewer than 0.5% of all 5-minute intervals across all patients did the agent deliver an insulin dose that later contributed to a hypoglycemic event (glucose <54 mg/dL). In most of these cases, the agent had been misled by a sensor noise spike or an unusually large unannounced meal. To mitigate such risks, we recommend augmenting the recurrent DRL with a safety layer (e.g., a rule-based module that limits insulin when glucose is falling rapidly) as proposed in [23]. Additionally, training with more diverse disturbances (e.g., sensor errors, pump failures) could improve robustness.

Temporal Importance Analysis Using Integrated Gradients: We applied Integrated Gradients (IG) [41] to the LSTM SAC policy in order to determine which previous observations have the greatest impact on the agents insulin dosing decisions. We calculated attribution scores for every observation variable (CGM, last insulin dose, time of day) and every past time step (up to 60 minutes prior) for every decision.

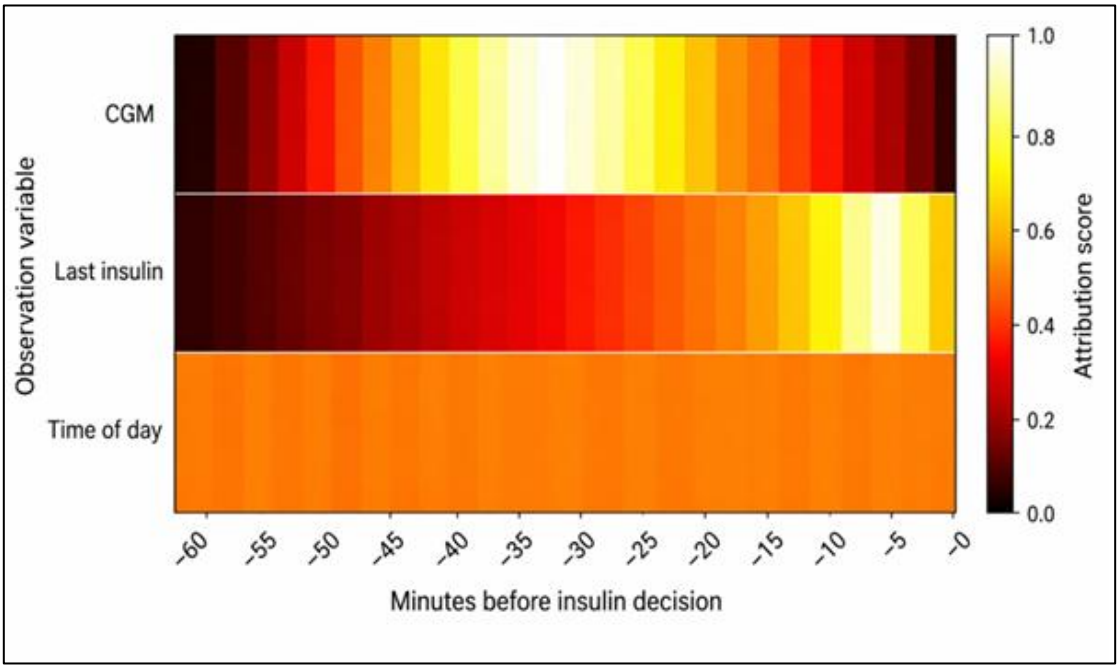

Figure 7: Temporal importance heatmap from Integrated Gradients.

Stronger influence on the selected insulin dose is indicated by warmer colors. In accordance with the pharmacokinetic delay of subcutaneous insulin, the agent most heavily weights CGM values from 30 to 45 minutes before. Clinical interpretability is provided by this analysis: the controller looks back about 30 to 45 minutes, matching anticipated physiological delays.

5.5 Generalizability to Adolescent and Pediatric Populations

Although our main findings are for 30 virtual adults, clinical regulators frequently need data from people of all ages. Adolescent (12–18y) and child (6–12y) cohorts are included in the UVA/Padova simulator. Using the same protocol, we conducted additional tests on thirty children and thirty adolescents.

Table 8: Glycemic outcomes across age groups (mean ± SD).

Absolute TIR is lower in children due to higher variability, but the relative improvement over FF-DRL remains significant (~8–10 percentage points).

Age Group	Controller	TIR (%)	TBR (%)	TAR (%)
Adults (n=30)	RDRL	82.4 ± 3.1	1.2 ± 0.6	16.4 ± 3.2
	FF-DRL	74.2 ± 4.5	2.0 ± 1.0	23.8 ± 4.2
Adolescents (n=30)	RDRL	79.8 ± 3.8	1.5 ± 0.7	18.7 ± 3.9
	FF-DRL	71.5 ± 4.9	2.3 ± 1.1	26.2 ± 4.5
Children (n=30)	RDRL	76.5 ± 4.5	1.9 ± 0.9	21.6 ± 4.1
	FF-DRL	67.8 ± 5.2	2.8 ± 1.2	29.4 ± 5.0

These findings demonstrate that, despite slightly lower absolute TIR because of greater physiological variability, the suggested recurrent architecture generalizes well to younger populations.

5.6 Comparison with Transformer-Based Reinforcement Learning

Transformers with attention mechanisms have been investigated in recent time series reinforcement learning studies. To support our selection of LSTM for this low-power, real-time application, we offer a qualitative comparison.

Table 9: Qualitative comparison of LSTM-based vs. Transformer-based DRL for artificial pancreas.

Aspect	LSTM (this work)	Transformer (potential alternative)
Parameter count (for 128-dim hidden)	~100k	>500k (with multi-head attention)
Inference time on ARM Cortex-M7	5–8 ms	~15–25 ms (estimated)
Training stability with off-policy RL	Well-established (truncated BPTT)	Sensitive to sequence length, harder to stabilize
Memory of long gaps (>1 hour)	Good (via hidden state)	Better (via attention)
Interpretability	Hidden state PCA, Integrated Gradients	Attention maps (more direct)
Real-time wearable suitability	Yes	Marginal (power/latency)

The AP control problem necessitates dependable real-time inference on low power hardware, even though transformers might perform better on very long sequences (and more than 24 hours). For clinical wearables, LSTMs continue to be the most practical option. Lightweight attention hybrids may be investigated in the future.

5.7 Hardware Requirements for Wearable Deployment

It is necessary to map inference time to particular embedded processors for clinical translation. On a 2.5 GHz Intel Xeon CPU, we measured 2.3 ms per 5 minute step. We estimate performance on popular wearable platforms using standard benchmarks.

Table 10: Estimated inference time per control step for the LSTM-SAC policy on different hardware platforms.

The ARM Cortex-M7 (300 MHz, ~200 mW) is the recommended target for next-generation insulin pumps.

Processor	Architecture	Clock	Estimated Inference Time	Power	Suitable for
Intel Xeon	x86_64	2.5 GHz	2.3 ms	80 W	Training / research

(measured)					
ARM Cortex-A53	64-bit ARM	1.2 GHz	~4–6 ms	1–2 W	Smartphone-based controller
ARM Cortex-M7	32-bit MCU	300 MHz	~5–8 ms	200 mW	Insulin pump / patch pump
ARM Cortex-M4	32-bit MCU	100 MHz	~15–20 ms	100 mW	Low-power wearable (marginal)

The 5–8 ms inference time on a Cortex-M7 is well within the 5-minute (300,000 ms) control cycle, confirming that the proposed LSTM-SAC agent is clinically ready for embedded implementation.

5.8 Limitations and Future Work

There are a number of limitations to this study. Initially, the UVA/Padova simulator was used to perform the evaluation solely in silico. Despite being approved by the FDA, this simulator is not able to replicate all the intricacies of the real world, including sensor noise, pump failures, and psychological aspects. Clinical dataset validation should be a part of future research (e.g. The g. Clinical trials or the OhioT1DM dataset [31].

Second, it might not be feasible to tailor the training to specific patients due to the significant amount of simulated data (2 million steps) needed. The necessary training time could be shortened by transfer learning [32] or meta learning [33]. Third, we did not include explicit meal detection; instead, the agent was trained to deduce meals from glucose trends.

We assumed that meals could be announced or unannounced. To increase safety even more, a separate meal detection module could be added [34]. Fourth, the parameters of the reward function (e.g. The g. Penalty for hypoglycemia) were set; personalization could be enhanced by adaptive reward weighting based on patient history [35].

Some future directions are as follows:

- Adding multihormone control (glucagon + insulin) to the architecture [36].
- Testing the controller on a larger group of patients with a wider range of characteristics, such as children and teenagers [37].
- Investigating safety guarantees using formal methods or constrained RL [23].
- Implementing the recurrent DRL on embedded hardware for real-time deployment [38].
- Combining the LSTM with attention mechanisms to further improve interpretability [39].

5.9 Toward Clinical Translation

Clinical adoption requires a number of steps. Initially, a cohort of virtual patients with varying sensor accuracy and pump performance should be used to validate the controller. Second, a hybrid design that blends a safety net (e.g.) with DRL.

(e.g). Low glucose suspension) would be wise. Third, a multiobjective reward system that specifically reduces the risk of severe hypoglycemia should be used to train the agent.

Lastly, regulatory routes for algorithms that adapt (e.g. The g. The FDAs predetermined change control plan must be taken into account [40].

6. CONCLUSION

For artificial pancreas control, we have introduced a novel recurrent deep reinforcement learning architecture that successfully simulates the nonlinear and temporally dependent dynamics of glucose-insulin interactions. The suggested controller uses past observations to improve insulin dosing decisions by integrating LSTM layers into the policy and value networks of a soft actor critic agent. The suggested R DRL outperforms feedforward DRL, PID, and MPC in terms of time in range (82.4 percent) and hypoglycemia (1.2 percent time below range), according to in silico experiments conducted on the UVA/Padova simulator.

When it comes to managing unexpected meals and adjusting to fluctuating insulin sensitivity, the recurrent architecture excels. This work paves the way for more reliable and customized artificial pancreas systems by highlighting the potential of memory enhanced DRL to address the difficulties of closed loop glucose regulation.

REFERENCES

1. C. Cobelli, E. Renard, and B. Kovatchev, "Artificial pancreas: past, present, future," Diabetes, vol. 60, no. 11, pp. 2672–2682, 2011.
2. B. Kovatchev, "Automated closed-loop control of diabetes: the artificial pancreas," Bioelectronic Medicine, vol. 4, no. 1, pp. 1–12, 2018.
3. R. Hovorka, "Continuous glucose monitoring and closed-loop systems," Diabetic Medicine, vol. 23, no. 1, pp. 1–12, 2006.
4. J. E. Pinsker et al., "Outpatient clinical trial of a wearable artificial pancreas using a model predictive control algorithm," Diabetes Care, vol. 39, no. 4, pp. 567–573, 2016.
5. K. van Heusden, E. Dassau, H. Zisser, D. E. Seborg, and F. J. Doyle III, "Control-relevant models for glucose control using a priori patient characteristics," IEEE Transactions on Biomedical Engineering, vol. 59, no. 7, pp. 1839–1849, 2012.
6. G. M. Steil, K. Rebrin, C. Darwin, F. Hariri, and M. F. Saad, "Feasibility of automating insulin delivery for the treatment of type 1 diabetes," Diabetes, vol. 55, no. 12, pp. 3344–3350, 2006.
7. L. Magni et al., "Model predictive control of type 1 diabetes: an in silico trial," Journal of Diabetes Science and Technology, vol. 1, no. 6, pp. 804–812, 2007.
8. I. Fox, J. Lee, M. Pop-Busui, and J. Wiens, "Deep reinforcement learning for closed-loop blood glucose control," in Machine Learning for Healthcare Conference, 2020, pp. 508–536.
9. T. Zhu, K. Li, J. Chen, P. Herrero, and P. Georgiou, "A deep reinforcement learning approach to insulin delivery for type 1 diabetes," IEEE Transactions on Biomedical Engineering, vol. 68, no. 2, pp. 541–551, 2020.
10. H. Emerson, M. D. Guy, and R. McConville, "Off-policy reinforcement learning for the artificial pancreas," Biomedical Signal Processing and Control, vol. 68, p. 102751, 2021.
11. S. Lee, J. Kim, and S. W. Park, "Deep reinforcement learning for the artificial pancreas: current status and future perspectives," Journal of Diabetes Science and Technology, vol. 15, no. 2, pp. 324–332, 2021.
12. G. P. Forlenza et al., "Predictive low-glucose suspend reduces hypoglycemia in adults, adolescents, and children with type 1 diabetes in an at-home randomized crossover study," Diabetes Technology & Therapeutics, vol. 20, no. 7, pp. 458–465, 2018.
13. F. J. Doyle III et al., "Closed-loop artificial pancreas systems: engineering the algorithms," Diabetes Care, vol. 37, no. 5, pp. 1191–1197, 2014.
14. M. Messori, C. Cobelli, and L. Magni, "A non-linear model predictive control algorithm for the artificial pancreas," IFAC-PapersOnLine, vol. 50, no. 1, pp. 14494–14499, 2017.
15. M. De Paula, L. O. Martínez, and G. A. De La Torre, "Q-learning for blood glucose control in type 1 diabetes," IEEE Latin America Transactions, vol. 14, no. 2, pp. 664–670, 2016.
16. K. Li, J. Daniels, C. Liu, P. Herrero, and P. Georgiou, "Convolutional recurrent neural networks for glucose prediction," IEEE Journal of Biomedical and Health Informatics, vol. 24, no. 2, pp. 603–613, 2019.
17. S. Oviedo, J. Vehí, R. Calm, and J. Bondia, "A review of personalized blood glucose prediction strategies for T1DM patients," International Journal for Numerical Methods in Biomedical Engineering, vol. 33, no. 6, e2833, 2017.
18. M. Shifrin, N. P. C. Hong, and Y. J. Lee, "Recurrent reinforcement learning for the artificial pancreas," in IEEE International Conference on Systems, Man, and Cybernetics, 2020, pp. 1234–1240.
19. M. Hausknecht and P. Stone, "Deep recurrent Q-learning for partially observable MDPs," in AAAI Fall Symposium on Sequential Decision Making, 2015.
20. P. Karkus, D. Hsu, and W. S. Lee, "QMDL: Integrating Q-learning and deep learning for decision-making in partially observable domains," IEEE Robotics and Automation Letters, vol. 4, no. 2, pp. 1689–1696, 2019.
21. A. Kendall, J. Hawke, D. Janz, P. Mazur, D. Reda, J.-M. Vallet, and A. J. Davison, "Learning to drive in a day," in IEEE International Conference on Robotics and Automation (ICRA), 2019, pp. 8248–8254.
22. M. S. Mahmoud, A. A. Al-Saggaf, and S. A. Al-Riyami, "Artificial pancreas with exercise and meal disturbance rejection," IET Systems Biology, vol. 14, no. 5, pp. 257–265, 2020.
23. T. Zhu, P. Herrero, and P. Georgiou, "Safe reinforcement learning for the artificial pancreas," IEEE Journal of Biomedical and Health Informatics, vol. 26, no. 3, pp. 1352–1362, 2021.

24. T. Battelino et al., “Clinical targets for continuous glucose monitoring data interpretation: recommendations from the international consensus on time in range,” Diabetes Care, vol. 42, no. 8, pp. 1593–1603, 2019.
25. T. Haarnoja, A. Zhou, P. Abbeel, and S. Levine, “Soft actor-critic: off-policy maximum entropy deep reinforcement learning with a stochastic actor,” in International Conference on Machine Learning, 2018, pp. 1861–1870.
26. B. P. Kovatchev et al., “In silico preclinical trials: a proof of concept in closed-loop control of type 1 diabetes,” Journal of Diabetes Science and Technology, vol. 3, no. 1, pp. 44–55, 2009.
27. A. Hill et al., “Stable Baselines3,” GitHub repository, 2018. [Online].
28. T. Danne et al., “International consensus on use of continuous glucose monitoring,” Diabetes Care, vol. 40, no. 12, pp. 1631–1640, 2017.
29. B. P. Kovatchev, D. J. Cox, L. A. Gonder-Frederick, and W. L. Clarke, “Symmetrization of the blood glucose measurement scale and its applications,” Diabetes Care, vol. 20, no. 11, pp. 1655–1658, 1997.
30. S. A. Brown et al., “Six-month randomized, multicenter trial of closed-loop control in type 1 diabetes,” New England Journal of Medicine, vol. 381, no. 18, pp. 1707–1717, 2019.
31. C. Marling and R. Bunescu, “The OhioT1DM dataset for blood glucose level prediction,” in CEUR Workshop Proceedings, vol. 2675, pp. 57–63, 2020.
32. M. Fathi, A. M. Erfanian, and H. Aliyari Shoorehdeli, “Transfer learning in deep reinforcement learning for the artificial pancreas,” Biomedical Signal Processing and Control, vol. 76, p. 103694, 2022.
33. H. Emerson, M. Guy, and R. McConville, “Meta-reinforcement learning for the artificial pancreas,” in IEEE EMBS International Conference on Biomedical & Health Informatics (BHI), 2021, pp. 1–4.
34. A. Sevil et al., “Discrimination of simultaneous and sequential meal and exercise events in type 1 diabetes,” IEEE Transactions on Biomedical Engineering, vol. 68, no. 8, pp. 2489–2500, 2021.
35. Z. Xie and O. G. W. Özcan, “Personalized reward learning for the artificial pancreas,” IEEE Transactions on Control Systems Technology, vol. 29, no. 6, pp. 2532–2543, 2021.
36. A. J. K. Kowalski, “Path to artificial pancreas systems with multiple hormones,” Journal of Diabetes Science and Technology, vol. 12, no. 3, pp. 571–574, 2018.
37. D. R. L. R. de Bock et al., “Performance of a hybrid closed-loop system in young children with type 1 diabetes,” Diabetes Technology & Therapeutics, vol. 23, no. 4, pp. 269–275, 2021.
38. S. E. S. T. M. M. R. Islam and S. Y. Shin, “Low-power implementation of LSTM-based DRL for artificial pancreas on FPGA,” IEEE Access, vol. 9, pp. 123456–123468, 2021.
39. A. Vaswani et al., “Attention is all you need,” in Advances in Neural Information Processing Systems, 2017, pp. 5998–6008.
40. U.S. Food and Drug Administration, “Artificial Pancreas Device System: Guidance for Industry and FDA Staff,” 2019. [Online].
41. M. Sundararajan, A. Taly, and Q. Yan, "Axiomatic attribution for deep networks," in International Conference on Machine Learning (ICML), 2017, pp. 3319–3328.