Explainable Recurrent Proximal Policy Optimization: Determining Proper Insulin Dosage

Pak Hop Chan (chanpa25), Bo Gong (gongbo1), Parsa Youssefpour (youss104)
(Contributions are listed in the Appendix)

Abstract

Type 1 diabetes is a common disease that has potentially severe consequences. Patients require insulin injections to keep blood glucose levels in a safe range. However, accurate insulin dosing management in Type 1 diabetes is a complex task due to delayed physiological responses and nonlinear glucose-insulin dynamics. In this work, we explore reinforcement learning (RL) for autonomous insulin administration using the SimGlucose environment. We compare three models based on the Proximal Policy Optimization (PPO) algorithm: a feedforward PPO baseline, a recurrent PPO model with Long Short-Term Memory (LSTM), and a recurrent PPO model (including LSTM) with an additional feature-level attention mechanism. Models were trained on simulated data from five virtual patients and evaluated on both seen and unseen patients (five each) over a 24-hour period with fixed meal timing. Results show that both LSTM-based models significantly outperformed the baseline in terms of time-in-range glucose control (baseline: $25.7\% \pm 5.9\%$, LSTM: $63.8\% \pm 9.7\%$, LSTM+Attention: $61.4\% \pm 9.4\%$, average $\pm$ SEM), with fewer dangerous glucose episodes and smoother insulin dosing. The addition of feature-level attention provided interpretability by quantifying the importance of clinical features (e.g., meal intake, CGM readings), aligning with physiological expectations such as elevated postprandial sensitivity. Our findings demonstrate that recurrent RL models with attention not only improve performance but also offer greater transparency, which is critical for clinical deployment in high-risk settings.

Introduction

Diabetes is a common disease characterized by persistently high blood glucose levels. In healthy humans, the insulin hormone dynamically released by the pancreas regulates the blood glucose level within a target range of 70 - 180 mg/dL (Klaff et al., 2015). In diabetic patients, impaired insulin production (type 1) or insulin resistance (type 2) disrupts glucose regulation, leading to serious health problems such as life-threatening emergency conditions (diabetic ketoacidosis) and long-term complications that include cardiovascular disease, vision impairments, and chronic kidney diseases. For type 1 diabetes, insulin injection is the most common treatment, traditionally done several times a day where real-time glucose level is not taken into account. Glucose monitoring is now favoured since it allows for better medication planning and thus reduces the risk of hypoglycemia (low blood glucose) due to overdosage and hyperglycemia (high blood glucose). Thus, the artificial pancreas, which consists of automated insulin pumps with continuous glucose monitoring and control algorithms, provides precise glucose management.

Determining the correct insulin dose is a complex task. Blood glucose levels can be influenced by factors such as diet, activity, stress, and insulin sensitivity (Majety, 2022). The delayed effect of insulin further complicates the adjustments (King et al., 2016). The current insulin-to-carbohydrate ratio method is often inaccurate, as it overlooks these variables, increasing the risk of hypoglycemia or hyperglycemia (Diabetes Teaching Center, n.d.; Vanstone et al., 2015). Performing accurate and timely dosage adjustments are crucial to maintaining optimal blood glucose levels and preventing complications.

There has been many approaches to establish an effective control algorithm. Early attempts include interval analysis for basal-bolus methods (Revert et al., 2011) and proportional-integral-derivative (PID) control systems (Chee et al., 2003), which essentially uses differential equations. Model Predictive Control (MPC), which frames the problem as an iterative finite-horizon optimization (Pinsker et al., 2016), outperforms earlier methods but

CSC 2516 Project. University of Toronto. April 2025.

comes with higher computational cost. In recent years, reinforcement learning has been another promising approach due to its ability of adaptive personalization and automatic handling of disturbances, as well as its superior performances.

Despite their benefits, standard reinforcement learning models, including PPO, do not have an internal memory mechanism, which can be a limitation in settings where past information affects future outcomes. This is particularly relevant in insulin dosing, where the delayed effect of insulin means that recent glucose trends and actions must be considered when making decisions. To address this, we explore the use of a recurrent PPO model that incorporates Long Short-Term Memory (LSTM) units, allowing the agent to retain and leverage past observations.

Another challenge with such models, specially in the medical field, is that they are often black-box in nature, making them difficult to interpret. To address this shortcoming, we propose incorporating an attention mechanism into the PPO model at the feature layer. This allows the model to focus on the most relevant patient observations, improving accuracy by emphasizing key state variables while de-emphasizing less important ones. Additionally, reporting attention weights provides transparency into the model's reasoning. By modifying the PPO algorithm, we identify the most important feature at each time step, which can be interpreted as the driving reason behind the model's decisions. We also investigate the performance of the recurrent PPO architecture both with and without the attention layer, and compare it to the baseline PPO model.

Formal Description

Problem Setup

Task Description and Environment

Deep Reinforcement Learning (RL) algorithms are well suited for controlling the basal insulin dosage for patients with Type 1 diabetes, to keep the blood glucose level within the target range of 70–180 mg/dL. The delayed and nonlinear effects of insulin make the problem challenging, as they complicate the prediction of how a given dosage will influence future glucose levels. RL is designed to maximize future rewards, which in this case is achieved by staying in the safe zone for as long as possible; this addresses the problem of delayed insulin effects. Additionally, deep neural networks can model the nonlinear relationship between insulin dosage and its effect on blood glucose levels.

The SimGlucose environment is a Python implementation of the FDA-approved UVA/Padova simulator, which models the dynamics of blood glucose in response to insulin delivery and carbohydrate intake (Xie, 2020). The simulator was developed to create an environment for testing and evaluating diabetes treatment strategies (Man et al., 2014). The simulation is a rule based model of human glucose insulin physiology, with parameters derived from clinical studies and validated against real patient data (Man et al., 2014). It offers flexibility in customizing meal profiles, patient characteristics, and medical devices such as blood glucose sensors and insulin pumps. This paper focuses on a simplified insulin delivery approach, consistent with the convention used in the simulator adapted to OpenAI's Gym framework (Xie, 2020).

Observation Space

The default observation space in SimGlucose includes only the continuous glucose monitor (CGM) reading, which reflects the patient's blood glucose level as measured by a small sensor typically inserted under the skin. While the environment provides access to other features, we selected only 4 features that are commonly available in real-world clinical setting.

The first and most critical feature is the CGM value, as it is directly tied to the agent's objective and closely related to the reward function. Although the raw CGM values range from 20 to 600 mg/dL, values outside 54–300 mg/dL are clinically dangerous (this is when patients are advised to go to the emergency department). To improve training efficiency and keep the agent focused on physiologically safe states, we modified the environment to terminate an episode if blood glucose falls outside this range.

The second feature is carbohydrate intake, or meals (in grams), which ranges from 0 to 200 grams, though typical meals are below 100 grams. Carbohydrate intake is both observable and highly influential on glucose dynamics. In addition, we included the last insulin dosage rate (0–1.0 U/min) to help the model account for the delayed effect of insulin, and the time of day/hour, which can influence glucose fluctuations due to circadian rhythms; they are the third and fourth features.

Since these four features span different numerical ranges, we first scaled all input values to a comparable range. This helps improve training stability by preventing certain features from dominating due to magnitude differences.

Action Space

The environment has a continuous action space representing the basal insulin dosage rate, which by default ranges from 0 to 30 U/min with 0.5-unit increments. However, this range is not clinically realistic, as typical basal insulin rates fall between 0 and 1.0 U/min. To focus the agent on physiologically meaningful behavior and improve training efficiency, we modified the environment to restrict the action space to this narrower, medically realistic range (0 to 1.0 U/min).

Reward Function

The reward function used in this environment is based on the risk index, which quantifies the health risk associated with a given blood glucose (BG) level proposed by Kovatchev et al. (2006). It assigns higher penalties to dangerous ranges, especially hypoglycemia, where the immediate health risks are more severe. The risk for a given blood glucose level is computed as follows:

$$f(\text{BG}) = 1.509 \times [\ln(\text{BG})]^{1.084} - 5.381 \tag{1}$$

$$\text{Risk}(\text{BG}) = 10 \times (f(\text{BG}))^2 \tag{2}$$

The reward at each timestep is defined as:

$$\text{Reward}_t = \text{Risk}_{t-1} - \text{Risk}_t \tag{3}$$

An increase in the risk index results in a negative reward, while a decrease yields a positive reward. This reward function provides a continuous and differentiable reward signal that is also clinically meaningful. It encourages the agent to minimize the overall health risk by keeping blood glucose levels within the safe range, where the risk is lowest.

Baseline Model (PPO)

Proximal Policy Optimization (PPO) is a reinforcement learning algorithm proposed by Schulman et al. (2017) at OpenAI, widely regarded for its balance between training stability and sample efficiency. PPO collects trajectories by running the current policy for a specified number of steps, then uses that data to update the model. The model takes the current state as input and produces two outputs. The value head estimates the value of the state $V(t)$, representing the expected sum of future discounted rewards. The policy head outputs the mean μ of a one dimensional Gaussian distribution. During training, an action is sampled from this distribution using the formula $a = \mu + \sigma \cdot \varepsilon$, where $\varepsilon \sim \mathcal{N}(0, 1)$ and σ comes from a separate learning parameter log standard deviation $\log \sigma$. This stochastic sampling enables exploration, which is essential for effective policy learning. During inference, the action is just the mean μ.

During training, the collected trajectory is divided into smaller mini batches based on the `batch_size` hyperparameter. For each timestep in a mini-batch, the advantage is computed using *Generalized Advantage Estimation (GAE)*, which quantifies how much better or worse the taken action was compared to the expected value:

$$\hat{A}_t = \delta_t + (\gamma\lambda)\hat{A}_{t+1} \tag{4}$$

$$\delta_t = r_t + \gamma V(s_{t+1}) - V(s_t) \tag{5}$$

where γ is the *discount factor*, λ is the *GAE smoothing parameter*, and $V(s_t)$ is the estimated value of state s_t.

Using the estimated advantages, the policy loss is calculated using the *Clipped Surrogate Objective*, which helps prevent overly large policy updates:

$$\mathcal{L}_{\text{clip}}(\theta) = \mathbb{E}_t \left[\min \left(r_t(\theta)\hat{A}_t,\ \text{clip}(r_t(\theta), 1-\epsilon, 1+\epsilon)\hat{A}_t \right) \right] \tag{6}$$

where

$$r_t(\theta) = \frac{\pi_\theta(a_t \mid s_t)}{\pi_{\text{old}}(a_t \mid s_t)} \tag{7}$$

This clipping mechanism ensures stable training by preventing updates that deviate too far from the old policy.

The value loss is computed using the mean squared error between the predicted value and the discounted return for each timestep, averaged over the batch:

$$\mathcal{L}_{\text{vf}} = \frac{1}{N} \sum_{t} \left(V(s_t) - \hat{R}_t \right)^2 \tag{8}$$

The third component is the entropy bonus, which encourages exploration by penalizing certainty. For a Gaussian policy, the entropy is given by:

$$\mathcal{H} = \frac{1}{2} \sum_{i} \log(2\pi e \sigma_i^2) \tag{9}$$

Higher entropy corresponds to more randomness in actions, helping the policy avoid premature convergence.

The total loss combines all three components:

$$\mathcal{L}_{\text{total}} = \mathcal{L}_{\text{clip}} + c_v \cdot \mathcal{L}_{\text{vf}} - c_e \cdot \mathcal{H} \tag{10}$$

where c_v is the value loss coefficient and c_e is the entropy coefficient.

The total loss is used to backpropagate through the model and update all parameters. Once the parameters have been updated, a new trajectory is generated using the updated model, and this process is repeated until the specified number of training steps is reached.

In our baseline PPO model shown in **Figure 1**, we use two separate feedforward neural networks: one for the policy and one for the value function. Each network consists of four layers: an input layer with 4 units (corresponding to the observation), followed by three hidden layers with sizes 256, 64, and 64, respectively, using the Tanh activation function.

In the policy network, the final hidden layer outputs a 64 dimensional latent vector, which is passed to a single fully connected layer that maps it to a 1 dimensional output representing the mean of a Gaussian distribution over actions. In the value network, the final layer similarly outputs a scalar representing the estimated value of the state.

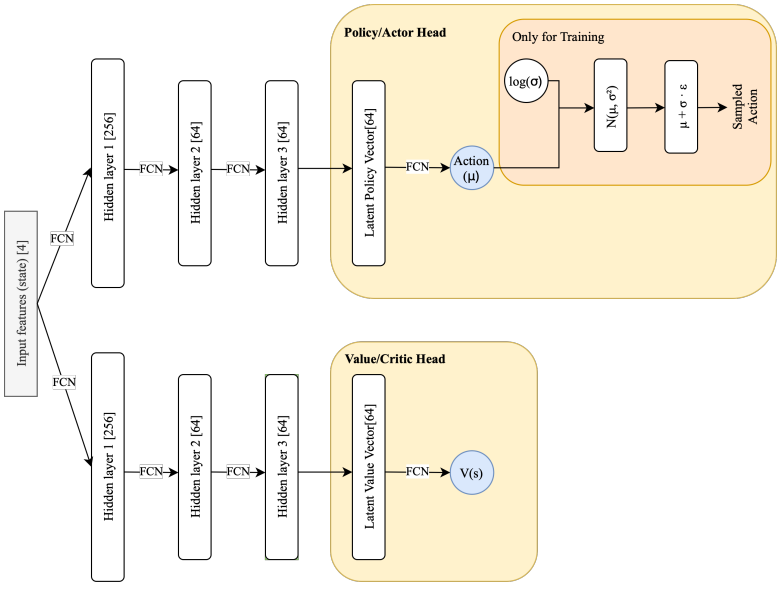

Figure 1: Baseline PPO Architecture

Recurrent PPO

As mentioned earlier, the delayed effect of insulin is a key challenge in predicting the appropriate dosage rate. While RL algorithms optimize for long-term rewards, they typically lack an internal memory mechanism. In our baseline PPO model, the only temporal information available is the previous insulin dosage, provided as a feature. To more effectively capture temporal dependencies and improve decision making, we extend the model with a Long Short Term Memory (LSTM) layers. This addition allows the agent to retain and leverage past observations, enabling more accurate predictions in a partially observable environment.

The Recurrent PPO model retains the overall structure of PPO but replaces the initial feedforward layers with two stacked LSTM layers to incorporate memory (Stable-Baselines-Team, 2023). As shown in **Figure 2**, the input feature vector (of size 4) is passed through the LSTMs, each with a hidden size of 128. The LSTM output is then fed into a neural network consisting of three fully connected layers: from 256 to 64, then 64 to 64, and finally 64 to 1. This architecture is used for both the policy and value heads, enabling the model to make decisions based on temporal context.

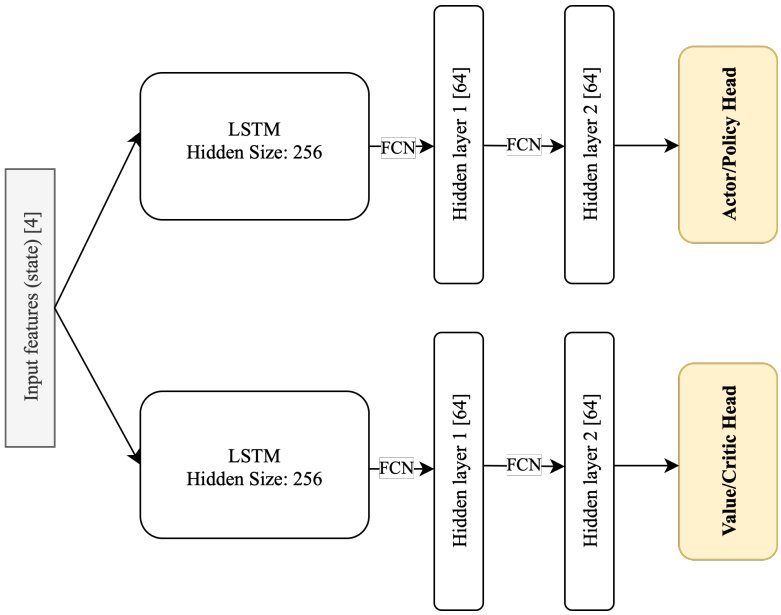

Figure 2: Recurrent PPO Architecture

Attention Recurrent PPO

Since this model is intended for medical decision making, interpretability is critical to ensure that its recommendations can be trusted by clinicians. To address the black box nature of reinforcement learning models, we introduce a feature level attention mechanism that highlights which input signals are most influential at each timestep. We used an Multi-Layered Perceptron to compute feature level attention weights instead of a transformer attention mechanism due to the small number of input features (CGM, insulin, meal, time). The transformer attention mechanism is designed to model complex interactions across high dimensional or sequential data, which is unnecessary in this case. Since temporal dependencies are already captured by the LSTM in our Recurrent PPO model, we only require per-timestep feature weighting. An MLP provides a simpler, more interpretable alternative, well suited for highlighting the influence of individual features in a transparent and clinically relevant way.

This attention module consists of three dense layers: the input feature vector $\mathbf{x} \in \mathbb{R}^n$ is first passed through two fully connected layers with ReLU activation, followed by a third layer that outputs raw attention scores. These scores are normalized using a softmax function to produce the final attention weights:

$$\mathbf{a} = \text{Softmax}(W_2(\text{ReLU}(W_1\mathbf{x} + \mathbf{b}_1)) + \mathbf{b}_2)$$

The resulting attention vector $\mathbf{a} \in \mathbb{R}^n$ is then applied element-wise to the original input:

$$\mathbf{x}_{\text{attn}} = \mathbf{a} \odot \mathbf{x}$$

The weighted feature vector $\mathbf{x}_{\text{attn}}$ is passed through two fully connected layers (4→64→64) before being fed into the Recurrent PPO model described earlier, as shown in **Figure 3**. A direct mapping from the 4 feature input to the 256 LSTM hidden state would be too abrupt and may hinder learning. The intermediate layers allow the model to learn more expressive representations, enabling better temporal reasoning. During inference, the attention weights $\mathbf{a}$ are extracted to interpret which input features (CGM readings, meal intake, previous insulin dose, or time of day) were most influential in the model's decision making. This mechanism provides greater transparency and improves trustworthiness in high stakes clinical applications.

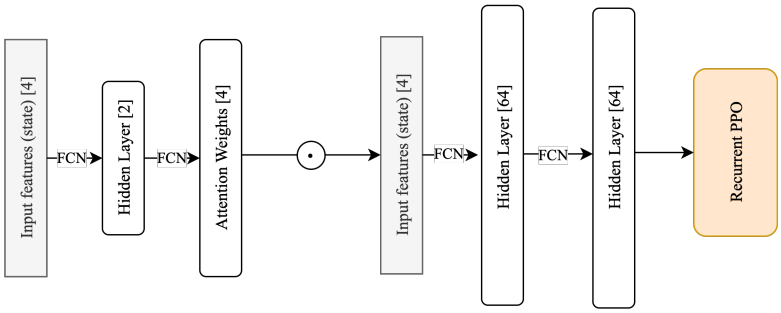

Figure 3: Attention Architecture of Attention Recurrent PPO

Related Work

The use of reinforcement learning in insulin dosage prescription gained prominence in the 2010s due to possibility of fully automated algorithms (Bothe et al., 2013). For example, actor-critic reinforcement learning algorithms used by Daskalaki et al. (2013) and Q learning algorithms used by Patil et al. (2015) allow for real-time personalization. De Paula et al. (2015) proposed a model-based RL framework with Gaussian Processes to handle inter-patient variability. These initial models gave rise to recent model-free deep reinforcement learning methods. Although somewhat reliant on simulations (being *in silico*) and relevant lack of patient control studies, they are currently among the best methods for achieving fully automated insulin delivery. Oroojeni Mohammad Javad et al. (2019) was among the few studies that used real clinical data. Recommendations of their Q-learning algorithm matched the actual dose prescribed by the physician 88% of the time, but did not take lifestyle information into account and lacked real time control. Most advances in deep reinforcement learning are based on simulations. For example, Ngo et al. (2018) used reinforcement learning with feedforward controller to calculate insulin doses. Kalman filter was used to estimate unmeasurable state variables. These allowed the controller to be robust to meal time uncertainties and carb count inaccuracies. Nordhaug Myhre et al. (2020) used the Trust-Region Policy Optimization algorithm, which is a deep policy gradient algorithm, to have a continuous action space, resulting in a better performance. Yamagata et al. (2020) proposed to use model-based recurrent neural networks to predict blood glucose levels, then combined them with MPC controllers to give insulin dosage. Long term dependencies could be learned in an online manner, but parameters for each person and meal were fixed, which was unrealistic. There were more recent models as well. The offline learning model involving partially observable Markov decision processes proposed by Rachim et al. (2025) employs the soft actor-critic algorithm to achieve preclinical validation. Meanwhile, double deep Q learning used by Noaro et al. (2023) is stable and fast by using approximations and prevents overestimation, but might not involve insulin sensitivity variability and confounding factors. Tejedor et al. (2023) used DQN algorithms that are model-free, achieving better performance. Viroonluecha et al. (2022) employed PPO methods to achieve longer periods of safe glycemic state. Hettiarachchi et al. (2024) augmented the PPO algorithm in continuous control by adding a model learning phase to learn a glucose dynamics model, and a planning phase to fine-tune the learned control policy, but only focusing on adult patients.

Yet, these reinforcement learning methods often lack interpretability. Yang (2022) suggested that explainable AI (XAI) methods help physicians apply predictive modeling in practical situations. In addition, Prendin et al. (2023) pointed out that reinforcement learning algorithms should be physiologically sound and explainable to ensure patient safety, and used SHapley Additive exPlanation (SHAP) - a black box model explainer - to evaluate different long-short term (LSTM) neural network models and showed that some models would be preferred over others by correctly learning physiological relationships. Duckworth et al. (2024) used SHAP to identify personal risk factors in a gradient boosted tree–based model that predicts imminent hypoglycemia and hyperglycemia. However, SHAP, despite its potential to explain many models, could give errorneous or misleading interpretations (Huang & Marques-Silva, 2024). On the model side, Lim et al. (2021) included dual attention networks with soft actor-critic networks, using attribution scores to obtain generalized interpretations on glucose predictions but

with worse performance. De Bois et al. (2021) used a two-level attention mechanism to achieve an interpretable RETAIN model to predict glucose levels, although it doesn't make efficient use of long histories. Yet, most interpretable solutions that does not use SHAP in literature involve simpler methods such as random forests or regression, lacking the advantages of deep learning algorithms.

Analysis and Discussion

Model Training

The training environment was set up with insulin doses and associated blood glucose levels recorded every 3 minutes over a 24-hour period (a total of 480 intervals) with randomized timing of meals. Three models were trained based on five simulated adult patients (#1-5): a baseline PPO model, a Recurrent PPO model, and an Attention Recurrent PPO. Due to frequent instability during training—where models often collapsed in later stages—we employed a callback mechanism to evaluate the model every 5000 steps and save the model that achieved the best result (i.e the highest reward)

Blood Glucose Control

Model	**Patients 1–5** (Training Patients)			**Patients 6–10** (New Patients)			**All Patients**		
	Glucose in Range	Total Rewards	Dangerous Episodes	Glucose in Range	Total Rewards	Dangerous Episodes	Glucose in Range	Total Rewards	Dangerous Episodes
PPO (baseline)	32.9% (9.5%)	-12.5 (7.5)	2	16.6% (3.5%)	-28.0 (6.9)	4	25.7% (5.9%)	-19.4 (5.6)	6
Recurrent PPO (baseline+LSTM)	69.9% (16.1%)	-2.1 (1.9)	1	57.8% (12.1%)	-3.9 (2.9)	1	63.8% (9.7%)	-3.0 (1.7)	2
Attn Recurrent PPO (baseline+LSTM+Attn)	71.6% (15.4%)	-4.2 (3.6)	1	51.1% (10.4%)	-7.4 (3.7)	2	61.4% (9.4%)	-5.8 (2.5)	3

Table 1: Performance on training, new, and all patients. Each cell shows average (standard error of the mean, SEM)

In a similar 24-hour environment, insulin dosing and the resultant glucose control was tested on the training patients (#1–5) and new adult patients (#6–10) across all three models (see **Table 1**, where results are presented as averages and SEMs). When analyzing performance using the percentage of time where blood glucose remained within the target range (70–190 mg/dL) averaged across all 10 patients, the Recurrent PPO and the Attention Recurrent PPO displayed no significant difference (Recurrent PPO: $63.8 \pm 9.7\%$, Attention Recurrent PPO: $61.4 \pm 9.4\%$, p = 0.86, Student's t-test). However, both models significantly outperformed the baseline PPO model (PPO: $25.7 \pm 5.9\%$, $p < 0.05$), highlighting the effectiveness of incorporating temporal memory.

These results suggest that the addition of LSTM layers enables the agent to better capture the delayed effects of insulin and the temporal dependencies inherent in blood glucose regulation. Unlike feedforward networks, which treat each decision independently, the recurrent model leverages historical context—such as prior insulin doses, meal events, and glucose trends which is crucial in a physiological system with slow, time-lagged responses. The improved performance across both training and unseen patients also indicates that the temporal dynamics learned by the LSTM model generalize well beyond the training set.

Similar patterns were observed when analyzing the training cohort (patients #1–5) and the new cohort (patients #6–10) separately. As expected, performance was slightly higher in the training cohort for all models, but the recurrent architectures still maintained strong generalization. A comparable trend was also seen in the total reward values, reinforcing that the LSTM enhanced models not only improved glycemic control but also led to more favorable cumulative outcomes under the reward framework.

When the averaged (across 10 patients) blood glucose levels and insulin doses were examined over time (as shown in **Figure 4**), the Recurrent models (with or without attention) demonstrated similar trends of glucose control (generally within target range) and overall smooth insulin dosing. The baseline model, on the other hand, demonstrated a prolonged period of elevated glucose level, and insulin doses that are predominantly at extreme values (0 and 1.0 U/min). This pattern reinforces the inability of feedforward models to manage the lag between insulin action and glucose response, and demonstrates the clear advantage of recurrent architectures in this delayed feedback environment.

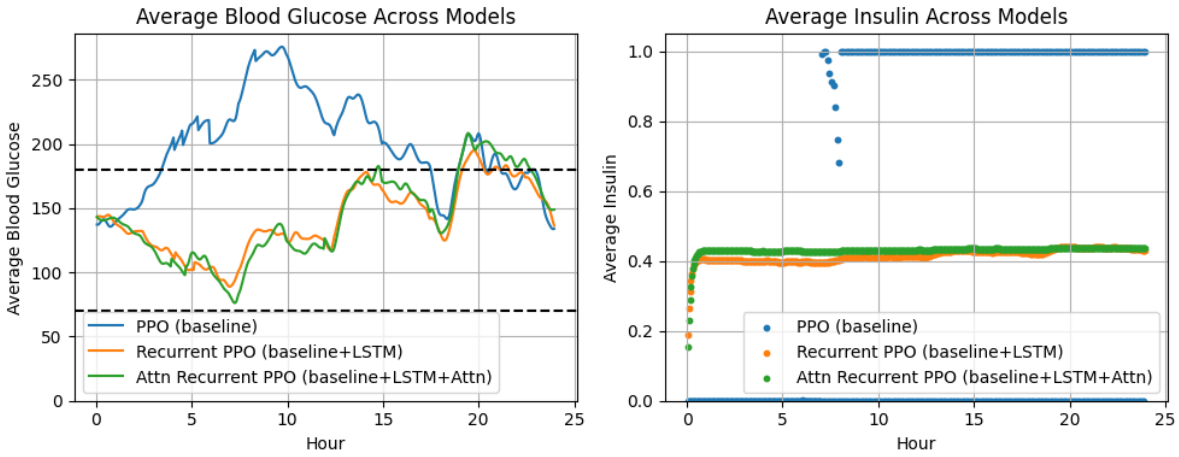

Figure 4: Blood glucose level and insulin dosing averaged across 10 patients over 24 hours in three models during testing.

Explainability with Attention

The LSTM model with attention confers the unique advantage of allowing quantified feature importance to suggest a certain level of clinical explainability. In **Figure 5**, the relative importance of four analyzed features (blood glucose level, last insulin dose, meal intake, hour of the day) was illustrated. Overall, "meal" carried the greatest weight over the 24-hour period. There is an noticeable immediate spike of feature importance (over feature "meal") after the administration of each meal, with the spike amplitude largely corresponding to the amount of carbohydrate intake in each meal. Furthermore, following each meal, there is a prolonged period of sustained elevation of "meal" feature importance, which is consistent with the physiology of delayed absorption of carbohydrate in the hours after meals. Notably, the insulin dosing itself does not undergo dramatic alterations after the meals, which is perhaps a calculated result in a model taking into account of temporal dependence and trained over scenarios with meals to produce an overall smooth insulin dosing.

The feature of hour of the day carried the lowest importance, which generally lowered further in response to elevated meal importance. The other two features (last insulin dose, blood glucose level) generally demonstrated no significant variation around 0.25 (1 in 4).

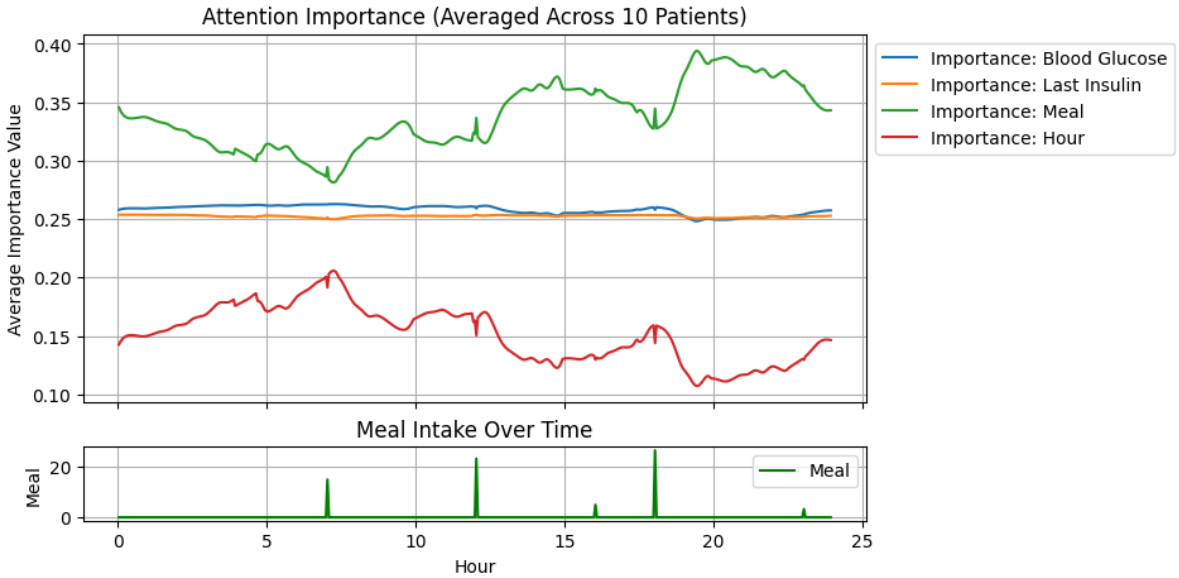

Figure 5: Feature importance averaged across 10 patients over 24 hours in the Attention Recurrent PPO model (with LSTM and attention).

One of the authors [BG] is a physician, and considers the explainability results generally useful. While ultimately it is the glucose control result that determines the medical outcome of the patients, having the ability to quantitatively assess the rationale behind the insulin dosage decision, even in a fairly limited way, is felt to be a welcome addition to the human-algorithm interactions. It helps to build confidence in the models not only for physicians, but also for patients, and could potentially be useful for improving patient education and treatment compliance.

Notably, the quantitative feature importance was provided numerically at each insulin dose decision point (480 times in 24 hours). An enhanced explainable output that synthesis human-like reasoning, perhaps through large language models, over an accumulated period at major decision points may provide further value. While interpretation of deep learning model behaviors is inherently challenging to assess, the quality of the explainability results generated by the attention model may be clinically assessed by comparing the numerical/qualitative feature importance with physician interpretation, ideally tested in a real-world clinical setting by enrolling several physicians with specialist training (endocrinologists) to account for individual subjectivity.

Overall, the results demonstrated the benefit of adding an attention layer over an LSTM model to allow explainability, without significant performance compromise. Both models performed significantly better than a baseline model without LSTM, demonstrating the benefit of analyzing temporal dependence in glucose control using the PPO model.

Limitations and Future Steps

In training the models, we encountered frequent policy crashes, towards the end of the training, where the reward did not trend upward as expected and had to manually select the best model using a callback method. This may be related to the known "policy collapse" phenomenon observed in PPO/Recurrent PPO models, particularly in complex environments such as Simglucose (Dohare et al., 2023). Although recent studies suggest that this issue can be mitigated through careful selection of optimizers and regularization techniques (Dohare et al., 2023), we were unable to implement these due to significant time constraints. Additionally, we were only able to train models for up to 500,000 steps; extending this to over 1 million steps may further improve performance in predicting insulin dosage and potentially enhance the explainability of the attention mechanism by allowing it to learn more stable and meaningful patterns over time.

Despite our progress in achieving a limited level of explainability by analyzing four features, future work could expand both the number and diversity of input features, especially when evaluating the model in more complex environments, such as across broader patient populations including children and adolescents. Additionally, incorporating a transformer style attention mechanism over the LSTM outputs could enhance the modeling of temporal dependencies and offer more interpretable insights into how past observations influence current decisions. Testing can be further enhanced by comparing the feature importance output with specialist physician human interpretation of the model behavior.

As in most previous works on insulin dosages, our work is still *in silico* without clinical trials. There is a current clinical trial at the University of Virginia that investigates the use of reinforcement learning based automated insulin delivery (Candelier & Beach, 2025). This can be compared to simulations to see whether reinforcement learning models work in practice. Another way could be using sim-to-real transfer using domain randomization (Chen et al., 2022). Also, run time of models could be an issue. For a reinforcement learning based automated insulin delivery system to be successful, ideally the output can be obtained quickly on the pump, which is likely without advanced computing resources. Our model does require some time to get results, which is to be improved. Besides, safety constraints could be added to our model. Other future work directions include transfer learning that allows quick personalization from population, integrating multi-modal data from smartwatches and smartphones, and incorporating dual-hormone control, i.e. also controlling glucagon.

Conclusions

Determining insulin dosage in type 1 diabetic patients is a complex yet clinically important scenarios that could benefit from development of improved deep learning algorithms. Building upon existing knowledge of the use of Proximal Policy Optimization (PPO) method in automated continuous glucose monitoring and control, we trained and tested a PPO model in comparison with two Recurrent PPO models featuring a LSTM structure to capture temporal information, with or without a superimposed attention layer. Despite training time constraints and technical challenges with maintaining training stability, our results as a proof of concept demonstrated clear benefits of employing recurrent LSTM structure in achieving significantly better model performance in several key patient outcome and safety measurements.

Furthermore, the addition of feature-level attention mechanism in the Attention Recurrent PPO model generated physiologically sound quantitative interpretation of insulin dosage decisions over four clinically relevant features, without compromising model performance. The feature importance results were felt to be clinically useful and relevant when assessed by one physician author, and as a future direction can be more systemically tested objectively in real-life clinical settings with enrollment of specialist physicians. If further enhanced and robustly evaluated, an attention-based recurrent PPO model based on the concept demonstrated in our paper has the potential of improving the care of diabetic patients through improved automated glucose control with clinical explainability and transparency.

References

Bothe, M. K., Dickens, L., Reichel, K., Tellmann, A., Ellger, B., Westphal, M., & Faisal, A. A. (2013). The use of reinforcement learning algorithms to meet the challenges of an artificial pancreas [Publisher: Taylor & Francis _eprint: https://doi.org/10.1586/17434440.2013.827515]. *Expert Review of Medical Devices*, *10*(5), 661–673. https://doi.org/10.1586/17434440.2013.827515

Candelier, E., & Beach, B. (2025, February 5). *New UVA clinical trial explores AI-powered insulin delivery for better diabetes care* [Center for diabetes technology]. Retrieved April 16, 2025, from https://med.virginia.edu/diabetes-technology/2025/02/05/new-uva-clinical-trial-explores-ai-powered-insulin-delivery-for-better-diabetes-care/

Chee, F., Fernando, T., Savkin, A., & van Heeden, V. (2003). Expert PID control system for blood glucose control in critically ill patients. *IEEE Transactions on Information Technology in Biomedicine*, *7*(4), 419–425. https://doi.org/10.1109/TITB.2003.821326

Chen, X., Hu, J., Jin, C., Li, L., & Wang, L. (2022). Understanding domain randomization for sim-to-real transfer. Retrieved April 16, 2025, from https://collaborate.princeton.edu/en/publications/understanding-domain-randomization-for-sim-to-real-transfer

Daskalaki, E., Diem, P., & Mougiakakou, S. G. (2013). An actor–critic based controller for glucose regulation in type 1 diabetes. *Computer Methods and Programs in Biomedicine*, *109*(2), 116–125. https://doi.org/10.1016/j.cmpb.2012.03.002

De Bois, M., El Yacoubi, M. A., & Ammi, M. (2021). Enhancing the interpretability of deep models in healthcare through attention: Application to glucose forecasting for diabetic people [Publisher: World Scientific Publishing Co.]. *International Journal of Pattern Recognition and Artificial Intelligence*, *35*(12), 2160006. https://doi.org/10.1142/S0218001421600065

De Paula, M., Ávila, L. O., & Martínez, E. C. (2015). Controlling blood glucose variability under uncertainty using reinforcement learning and gaussian processes. *Applied Soft Computing*, *35*, 310–332. https://doi.org/10.1016/j.asoc.2015.06.041

Diabetes Teaching Center. (n.d.). *Calculating insulin dose* [Calculating insulin dose]. Retrieved April 15, 2025, from https://diabetesteachingcenter.ucsf.edu/about-diabetes/type-2-diabetes/use-insulin-type-2-diabetes/calculating-insulin-dose

Dohare, S., Lan, Q., & Mahmood, A. R. (2023). Overcoming policy collapse in deep reinforcement learning. *Sixteenth European Workshop on Reinforcement Learning*. https://openreview.net/forum?id=m9Jfdz4ymO

Duckworth, C., Guy, M. J., Kumaran, A., O'Kane, A. A., Ayobi, A., Chapman, A., Marshall, P., & Boniface, M. (2024). Explainable machine learning for real-time hypoglycemia and hyperglycemia prediction and personalized control recommendations [Publisher: SAGE Publications Inc]. *Journal of Diabetes Science and Technology*, *18*(1), 113–123. https://doi.org/10.1177/19322968221103561

Hettiarachchi, C., Malagutti, N., Nolan, C. J., Suominen, H., & Daskalaki, E. (2024). G2p2c — a modular reinforcement learning algorithm for glucose control by glucose prediction and planning in type 1 diabetes. *Biomedical Signal Processing and Control*, *90*, 105839. https://doi.org/10.1016/j.bspc.2023.105839

Huang, X., & Marques-Silva, J. (2024). On the failings of shapley values for explainability. *International Journal of Approximate Reasoning*, *171*, 109112. https://doi.org/10.1016/j.ijar.2023.109112

King, A. B., Kuroda, A., Matsuhisa, M., & Hobbs, T. (2016). A review of insulin-dosing formulas for continuous subcutaneous insulin infusion (CSII) for adults with type 1 diabetes. *Current Diabetes Reports*, *16*(9), 83. https://doi.org/10.1007/s11892-016-0772-0

Klaff, L. J., Brazg, R., Hughes, K., Tideman, A. M., Schachner, H. C., Stenger, P., Pardo, S., Dunne, N., & Parkes, J. L. (2015). Accuracy evaluation of contour next compared with five blood glucose monitoring systems across a wide range of blood glucose concentrations occurring in a clinical research setting [Publisher: Mary Ann Liebert, Inc., publishers]. *Diabetes Technology & Therapeutics*, *17*(1), 8–15. https://doi.org/10.1089/dia.2014.0069

Kovatchev, B. P., Otto, E., Cox, D., Gonder-Frederick, L., & Clarke, W. (2006). Evaluation of a new measure of blood glucose variability in diabetes. *Diabetes Care*, *29*(11), 2433–2438. https://doi.org/10.2337/dc06-1085

Lim, M. H., Lee, W. H., Jeon, B., & Kim, S. (2021). A blood glucose control framework based on reinforcement learning with safety and interpretability: In silico validation. *IEEE Access*, *9*, 105756–105775. https://doi.org/10.1109/ACCESS.2021.3100007

Majety, P. (2022, December 6). *Insulin dosage: What do i need to know?* [Insulin dosage: What do i need to know?]. Retrieved April 15, 2025, from https://www.healthcentral.com/condition/type-1-diabetes/insulin-dosage

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

Ngo, P. D., Wei, S., Holubová, A., Muzik, J., & Godtliebsen, F. (2018). Control of blood glucose for type-1 diabetes by using reinforcement learning with feedforward algorithm. *Computational and Mathematical Methods in Medicine*, *2018*, 4091497. https://doi.org/10.1155/2018/4091497

Noaro, G., Zhu, T., Cappon, G., Facchinetti, A., & Georgiou, P. (2023). A personalized and adaptive insulin bolus calculator based on double deep q- learning to improve type 1 diabetes management. *IEEE Journal of Biomedical and Health Informatics*, *27*(5), 2536–2544. https://doi.org/10.1109/JBHI.2023.3249571

Nordhaug Myhre, J., Tejedor, M., Kalervo Launonen, I., El Fathi, A., & Godtliebsen, F. (2020). In-silico evaluation of glucose regulation using policy gradient reinforcement learning for patients with type 1 diabetes mellitus [Number: 18 Publisher: Multidisciplinary Digital Publishing Institute]. *Applied Sciences*, *10*(18), 6350. https://doi.org/10.3390/app10186350

Oroojeni Mohammad Javad, M., Agboola, S. O., Jethwani, K., Zeid, A., & Kamarthi, S. (2019). A reinforcement learning–based method for management of type 1 diabetes: Exploratory study [Company: JMIR Diabetes Distributor: JMIR Diabetes Institution: JMIR Diabetes Label: JMIR Diabetes Publisher: JMIR Publications Inc., Toronto, Canada]. *JMIR Diabetes*, *4*(3), e12905. https://doi.org/10.2196/12905

Patil, P., Kulkarni, P., & Shirsath, R. (2015). Sequential decision making using q learning algorithm for diabetic patients. In L. P. Suresh, S. S. Dash, & B. K. Panigrahi (Eds.), *Artificial intelligence and evolutionary algorithms in engineering systems* (pp. 313–321). Springer India. https://doi.org/10.1007/978-81-322-2126-5_35

Pinsker, J. E., Lee, J. B., Dassau, E., Seborg, D. E., Bradley, P. K., Gondhalekar, R., Bevier, W. C., Huyett, L., Zisser, H. C., & Doyle, F. J., III. (2016). Randomized crossover comparison of personalized MPC and PID control algorithms for the artificial pancreas. *Diabetes Care*, *39*(7), 1135–1142. https://doi.org/10.2337/dc15-2344

Prendin, F., Pavan, J., Cappon, G., Del Favero, S., Sparacino, G., & Facchinetti, A. (2023). The importance of interpreting machine learning models for blood glucose prediction in diabetes: An analysis using SHAP [Publisher: Nature Publishing Group]. *Scientific Reports*, *13*(1), 16865. https://doi.org/10.1038/s41598-023-44155-x

Rachim, V. P., Yoo, J., Lee, J., Lee, Y., & Park, S.-M. (2025). Generalized reinforcement learning control algorithm for fully automated insulin delivery system. *Expert Systems with Applications*, *274*, 126909. https://doi.org/10.1016/j.eswa.2025.126909

Revert, A., Calm, R., Vehi, J., & Bondia, J. (2011). Calculation of the best basal–bolus combination for postprandial glucose control in insulin pump therapy. *IEEE Transactions on Biomedical Engineering*, *58*(2), 274–281. https://doi.org/10.1109/TBME.2010.2058805

Schulman, J., Wolski, F., Dhariwal, P., Radford, A., & Klimov, O. (2017). Proximal policy optimization algorithms. *CoRR*, *abs/1707.06347*. http://arxiv.org/abs/1707.06347

Stable-Baselines-Team. (2023). Stable-baselines3-contrib: Extra modules for stable-baselines3 [Accessed: 2025-04-17].

Tejedor, M., Hjerde, S. N., Myhre, J. N., & Godtliebsen, F. (2023). Evaluating deep q-learning algorithms for controlling blood glucose in in silico type 1 diabetes. *Diagnostics*, *13*(19), 3150. https://doi.org/10.3390/diagnostics13193150

Vanstone, M., Rewegan, A., Brundisini, F., Dejean, D., & Giacomini, M. (2015). Patient perspectives on quality of life with uncontrolled type 1 diabetes mellitus: A systematic review and qualitative meta-synthesis. *Ontario Health Technology Assessment Series*, *15*(17), 1–29. Retrieved April 15, 2025, from https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4664939/

Viroonluecha, P., Egea-Lopez, E., & Santa, J. (2022). Evaluation of blood glucose level control in type 1 diabetic patients using deep reinforcement learning [Publisher: Public Library of Science]. *PLOS ONE*, *17*(9), e0274608. https://doi.org/10.1371/journal.pone.0274608

Xie, J. (2020). *Simglucose: A simulator for glucose control in type 1 diabetes* [GitHub repository]. Retrieved April 17, 2025, from https://github.com/jxx123/simglucose

Yamagata, T., O'Kane, A., Ayobi, A., Katz, D. S., Stawarz, K., Marshall, P., Flach, P. A., Bristol, R. S.-R. U. o., University, T. O., & University, C. (2020). Model-based reinforcement learning for type 1 diabetes blood glucose control. Retrieved April 16, 2025, from https://www.semanticscholar.org/paper/Model-Based-Reinforcement-Learning-for-Type-1-Blood-Yamagata-O'Kane/196e31fff2701ad3748dc86f081f3a33a26410c1

Yang, C. C. (2022). Explainable artificial intelligence for predictive modeling in healthcare [Company: Springer Distributor: Springer Institution: Springer Label: Springer Number: 2 Publisher: Springer International Publishing]. *Journal of Healthcare Informatics Research*, *6*(2), 228–239. https://doi.org/10.1007/s41666-022-00114-1

Appendix

Contributions

Each author contributed collaboratively and substantially. Main contributions:

Coding	
Baseline Model - Code & Training	PY
Recurrent PPO Model - Code & Training	PY
Attention Recurrent PPO Model - Code & Training	PY
Environment Modification	PY
Model Testing	BG
Data Analysis, Plotting, Statistics	BG
Write up	
Abstract	PHC + PY
Introduction	PHC
Formal Description	PY
Related Work	PHC
Analysis	BG
Limitations and Future Steps	BG + PHC
Conclusion	BG