

ORIGINAL RESEARCH

Open access

Advancing Automated Insulin Delivery Systems through Deep Reinforcement Learning: Current Progress and Emerging Opportunities

Lukas Weber^{1*}, Sophie Janssen², Jonas Richter¹, Elena Fischer³

Abstract

Progress in continuous glucose monitoring (CGM) technologies alongside wearable electronics is driving the evolution of automated insulin delivery systems (AIDs) toward fully automated closed-loop configurations. The ultimate goal is to provide safe, individualized, and optimized management of blood glucose concentration (BGC) for patients living with diabetes. Although model predictive control offers an adaptable architecture for constructing AIDs control logic, accurate and efficient BGC regulation demands models capable of accounting for inter- and intra-patient variability as well as perturbation uncertainties. Innovations in artificial intelligence offer novel pathways for creating data-driven, fully closed-loop AIDs. Among these approaches, deep reinforcement learning (DRL) has gained substantial interest owing to its inherent robustness against disturbances. Consequently, this article presents a comprehensive literature review focused on DRL-based BGC control methodologies within AIDs. Initially, this paper systematically evaluates the advantages of deploying DRL frameworks in AIDs. Following this, it delivers an extensive analysis of diverse DRL methods and enhancements designed to surmount obstacles linked to their integration into AIDs, addressing factors such as restricted sample availability, personalization, and safety protocols. Furthermore, an application-centric assessment of DRL-based AIDs control algorithms is provided, highlighting critical hurdles encountered during practical implementation. Lastly, the paper addresses potential remedies for relevant BGC control challenges, outlines prospects for real-world application, and identifies future research trajectories.

Keywords Deep reinforcement learning, Diabetes, Automated insulin delivery systems, Glycemic control, Insulin infusion

*Correspondence:

Lukas Weber

lukas.weber@uni-stuttgart.de

¹ Department of Diabetes Technology and Reinforcement Learning, Faculty of Medicine, University of Stuttgart, Stuttgart, Germany

² Department of Automated Insulin Delivery Systems, Faculty of Engineering, Eindhoven University of Technology, Eindhoven, Netherlands

³ Department of Deep RL for Medical Devices, Faculty of Medicine, University of Basel, Basel, Switzerland

Introduction

Diabetes is a chronic metabolic condition that is broadly classified into four primary categories according to the Classification and Diagnosis of Diabetes 2023: type 1 diabetes (T1D), type 2 diabetes (T2D), diabetes resulting from alternative etiologies, and gestational diabetes [1]. At present, approximately one in eleven individuals globally is affected by diabetes, with projections indicating this proportion will increase to one in ten over the coming decade. Within this population, T1D is characterized by a complete or near-complete deficit in beta-cell function, thereby mandating exogenous insulin replacement therapy [2]. Individuals managing T1D typically execute an average of 180 health-related decisions daily to control their blood glucose concentration (BGC) [3], which imposes a persistent psychological strain. As

the disease advances, insulin regimens are also frequently introduced for individuals with T2D who manifest diminished beta-cell functionality and/or profound insulin resistance [4].

Methods for administering insulin can be categorized into manual injection techniques and automated delivery protocols. Manual administration (frequently designated as multiple daily injections) relies on syringes or insulin pens to deliver insulin multiple times per day [5–7], representing the therapeutic modality utilized by the majority of diabetic patients [8]. Conversely, insulin pump usage can be either manual—referred to as sensor-augmented pump (SAP) therapy—or automated, which is characterized as an automated insulin delivery system (AIDs) [9]. Because T1D frequently manifests during childhood or adolescence [10], this demographic often depends on caregivers for injection management and demonstrates compromised treatment adherence. Furthermore, young pediatric patients (<6 years of age) face an elevated risk of experiencing both hyperglycemia and hypoglycemia [11]. This group also displays greater glycemic volatility compared to older cohorts [12]. Consequently, manual injection regimens may prove inadequate for these patients, given their inability to respond dynamically to ongoing glycemic fluctuations and their typical restriction to daytime administration. Recent clinical evaluations demonstrate that among adolescents and young adults, the implementation of continuous glucose monitoring (CGM) systems lowers hemoglobin A1c (HbA1c) levels alongside mean BGC over a 2–3 month period, while simultaneously maximizing the time in range (TIR) for BGC (defined as 70–180 mg/dL of blood glucose) [13, 14]. Additionally, AIDs have demonstrated a measurable benefit in mitigating the occurrence of severe hypoglycemic events (BGC < 54 mg/dL) across pediatric and adult populations [15]. Clinical trials and data provided by AIDs developers reveal that these systems markedly improve the TIR of glucose levels and decrease time-below-range (hypoglycemia) instances, thereby alleviating hypoglycemia-associated anxiety and elevating overall quality of life [4]. Although current AIDs control algorithms have successfully increased TIR from the 50–60% seen with SAP up to 70–80%, AIDs are experiencing growing adoption due to their capacity to alleviate the daily demands of BGC tracking.

An AIDs setup comprises three fundamental elements: a CGM sensor, an insulin pump, and a BGC controller that dynamically calculates requisite insulin quantities [16] (**Figure 1**). This configuration is also frequently termed an artificial pancreas system [17]. To replicate the physiological insulin secretion of a healthy pancreas, the AIDs modulates the subcutaneous insulin infusion rate upward or downward based on glucose data collected by the CGM [18]. Multiple investigations have been carried out across adult and pediatric cohorts utilizing diverse AIDs platforms equipped with varied algorithms, hardware pumps, and sensors [19–22]. Commercially accessible AIDs at present function as hybrid closed-loop networks, which mandate manual data entry regarding carbohydrate consumption for mealtime bolus dosing to handle postprandial BGC, as well as exercise logging to dial back insulin delivery, mitigate hypoglycemia risks, and maximize TIR [23, 24]. Evidence indicates that AIDs usage correlates with enhanced user satisfaction and reduced diabetes management stress [25]. Currently, fully automated AIDs that remove the necessity for carbohydrate calculation, manual meal declarations, or pre-scheduled exercise modifications remain under active development [4].

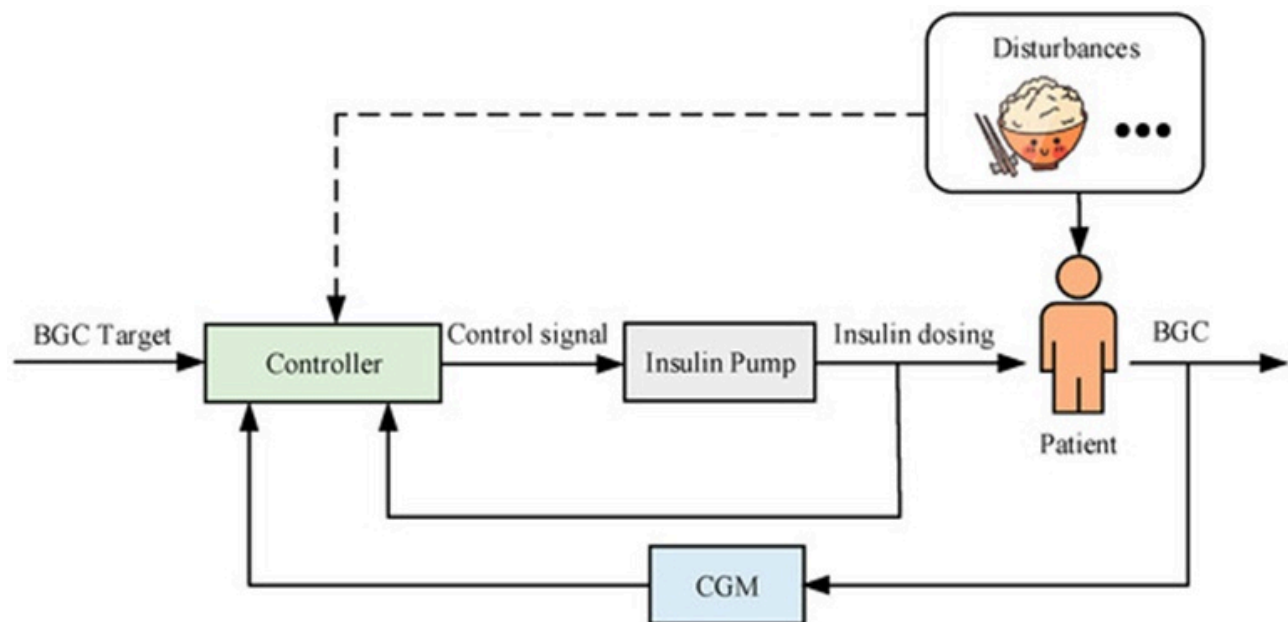


Figure 1. Basic structure of the AIDs.

The human physiological environment represents a highly intricate network, and AIDs control architectures must navigate multiple challenges stemming from the factors outlined below.

- (1) **Non-linearity:** Owing to the synergistic interactions of endogenous hormones alongside complex biochemical pathways, managing blood glucose in individuals with diabetes constitutes a fundamentally non-linear control problem.
- (2) **Inter- and intra-patient variability:** Marked disparities in glucose metabolism exist between separate patients. Furthermore, within a single individual, both long-term and short-term glycemic dynamics fluctuate continually due to circadian rhythms, emotional shifts, external variables, and other confounding events.
- (3) **Time lag:** A physiological latency of up to one hour can elapse between the administration of an insulin infusion and the corresponding shift in BGC [26], which demands predictive adjustments during dosage computations.
- (4) **Uncertain disturbances:** A patient's BGC is governed not only by insulin delivery but is also heavily driven by external elements such as physical exertion (including its type, intensity, and duration), emotional stress, and nutritional intake. Many of these elements are exceptionally difficult to detect, categorize, or quantify solely through CGM signals and insulin records. They introduce anonymous perturbations to closed-loop AIDs, creating substantial systemic uncertainty.
- (5) **High safety requirements:** Over-delivery of insulin can precipitate profound hypoglycemic episodes that may become life-threatening; hence, guaranteeing absolute systemic safety is the paramount constraint in designing and executing any control algorithm.

These compounding elements elevate the complexity of designing control logic and present formidable obstacles to the optimal performance of AIDs. Numerous literature reviews have evaluated diverse control frameworks [27-31]. Traditional Proportional Integral Derivative (PID) architectures have been deployed to counteract steady-state errors in BGC metrics [32-35]. Fuzzy logic (FL) control schemes, which deliver a degree of robustness via the inherent ambiguity of their rule sets, have likewise been investigated for BGC stabilization [36, 37]. Currently, the most prevalent control methodology is Model Predictive Control (MPC), which has demonstrated favorable outcomes in both simulated trials and clinical evaluations [38-43].

Nevertheless, commercially accessible hybrid closed-loop AIDs generally exhibit TIR outcomes bounded between 65–80%, leaving considerable room to enhance TIR in real-world scenarios by integrating artificial intelligence (AI) into control strategies. The widespread adoption of state-of-the-art sensors and insulin pumps has led AIDs to generate substantial volumes of interconnected data, establishing a foundation for data-driven AI methodologies like Deep Reinforcement Learning (DRL). DRL represents a highly advanced branch of Reinforcement Learning. While both methodologies utilize reward mechanisms to acquire optimal behavioral patterns through trial-and-error processes, DRL integrates deep neural networks (NNs) to navigate high-dimensional state and action parameters, yielding more sophisticated and scalable architectures. The autonomous learning capacity, adaptive calibration, and optimized decision-making behavior of DRL [44] render it a viable candidate for AIDs implementation, either as a standalone controller or hybridized with conventional control laws. To facilitate a more profound comprehension of the benefits and limitations characterizing traditional control algorithms like MPC within AIDs setups, as well as their direct comparison to DRL, an exhaustive comparative analysis is provided in Section 6. Fundamentally, DRL serves as a sequential learning loop that executes periodic decision-making, evaluates the consequences of these inputs, and autonomously refines its operating strategy to achieve an optimal policy. DRL has demonstrated exceptional capability across multiple domains, including mastering the game of Go [45], autonomous vehicular guidance [46], optimizing smart grid infrastructure operations [44], and automated pharmaceutical administration [47]. Over recent years, DRL-based control mechanisms have increasingly been explored for BGC stabilization [48, 49].

In a broader context, DRL operates as a computational paradigm where machines pursue predefined objectives by actively interfacing with an environment. A single interaction cycle entails the machine analyzing a specific environmental state, executing an action decision, enforcing this action upon the environment, and capturing feedback via a reward metric alongside the subsequent environmental state. This sequence repeats across numerous iterations to maximize the expected cumulative reward. Within the DRL framework, the decision-making entity is designated as the agent. Distinguishing it from the static "model" used in supervised learning, the "agent" in DRL signifies not just an aptitude for extracting data from its surroundings, but the specific capacity to directly manipulate and reshape the environment via its actions, rather than merely outputting predictive data.

In general, the DRL methodology provides several distinct benefits.

- (1) **Potential resistance to perturbations:** The agent accounts for all potential states when calculating expected values, allowing the DRL algorithm to adapt dynamically to unexpected external disturbances within the environment.
- (2) **Inter- and intra-patient variability:** Diabetic individuals function as the environment within the DRL paradigm. The agent determines optimal interventions based on its direct experiences with this individualized environment, offering distinct advantages over rigid, model-based control frameworks when accommodating environmental fluctuations.
- (3) **Adaptation to time lag:** The reward function in DRL is structured around cumulative feedback, which effectively mitigates the challenges imposed by latency inside the environment.
- (4) **Adaptation to sequential decision-making tasks:** Controlling BGC is inherently a sequential decision-making problem, wherein an action executed at the current time step alters the succeeding state, which subsequently dictates the necessary action for the following interval. DRL maintains a natural synergy with such sequential tasks.

Conversely, DRL architectures are still constrained by a collection of major difficulties.

- (1) **Low sample availability:** When dealing with high-dimensional and continuous action configurations, DRL algorithms frequently struggle to optimize sample data efficiently, causing diminished learning rates and an over-reliance on massive datasets. This limitation compromises the training velocity of DRL configurations and curtails their real-world utility, given that gathering vast sample volumes in physical clinical settings is frequently restricted by financial and temporal barriers.

(2) **Personalization:** When an initial or pre-trained DRL control agent is deployed for a new patient, it confronts challenges related to distributional shift. Throughout training, DRL architectures often only interact with localized data distributions, which can cause erratic or failed performance when encountering unfamiliar environments or novel clinical situations. Furthermore, the DRL controller must accommodate highly customized therapeutic goals due to the distinct inter-patient differences in blood glucose physiology and unique medical care strategies.

(3) **Security:** Because DRL systems operate within highly dynamic and volatile environments, their action pathways can be compromised by external threats or unforeseen noise, potentially triggering unintended interventions or hazardous clinical outcomes.

While several DRL reviews have been compiled, there remains a distinct scarcity of comprehensive assessments addressing recent implementations of DRL within AIDs [26, 49-53]. Specifically, earlier literature surveys concentrated either on generic DRL techniques or general AIDs technology without exploring the concrete challenges of blending DRL logic into AIDs. The rapid expansion of interest in AI-supported medical systems has sparked a surge in studies utilizing DRL, leading to the creation of many cutting-edge DRL architectures specialized for AIDs. This rapid progress highlights the necessity for a deep analysis of prospective DRL systems. Accordingly, this article seeks to present a thorough synthesis of the various DRL-driven approaches introduced for AIDs.

The primary contributions of this review are outlined below:

(1) To deliver a highly structured synthesis of the methodology and practical application domains, encompassing core theories alongside exhaustive breakdowns and deductions regarding DRL components used in AIDs. Particular emphasis is placed on establishing a more effective harmony between DRL logic and AIDs parameters regarding the state space, action space, and reward function configuration.

(2) To classify the primary obstacles of AIDs into three central thematic groups, illustrating the hurdles DRL faces regarding data scarcity, personalized adaptation, and operational safety.

(3) To offer critical insights into the prevailing limitations, prospective remedies, and future research pathways for DRL-driven AIDs, evaluating expansion possibilities regarding data infrastructure and hardware processing capabilities.

A comprehensive search on the Web of Science using the keywords “Reinforcement learning” and “Blood glucose” yielded a total of 106 publications. Filtering for relevant literature published from 2020 onward narrowed the scope to roughly 60 papers, from which 32 specific articles implementing DRL methodologies were ultimately preserved for closer analysis. The literature extraction was conducted on 1 January 2025. **Figure 2** charts the evolution of this publication volume over time. The annual output demonstrates a steady upward trend from 2020 through 2024.

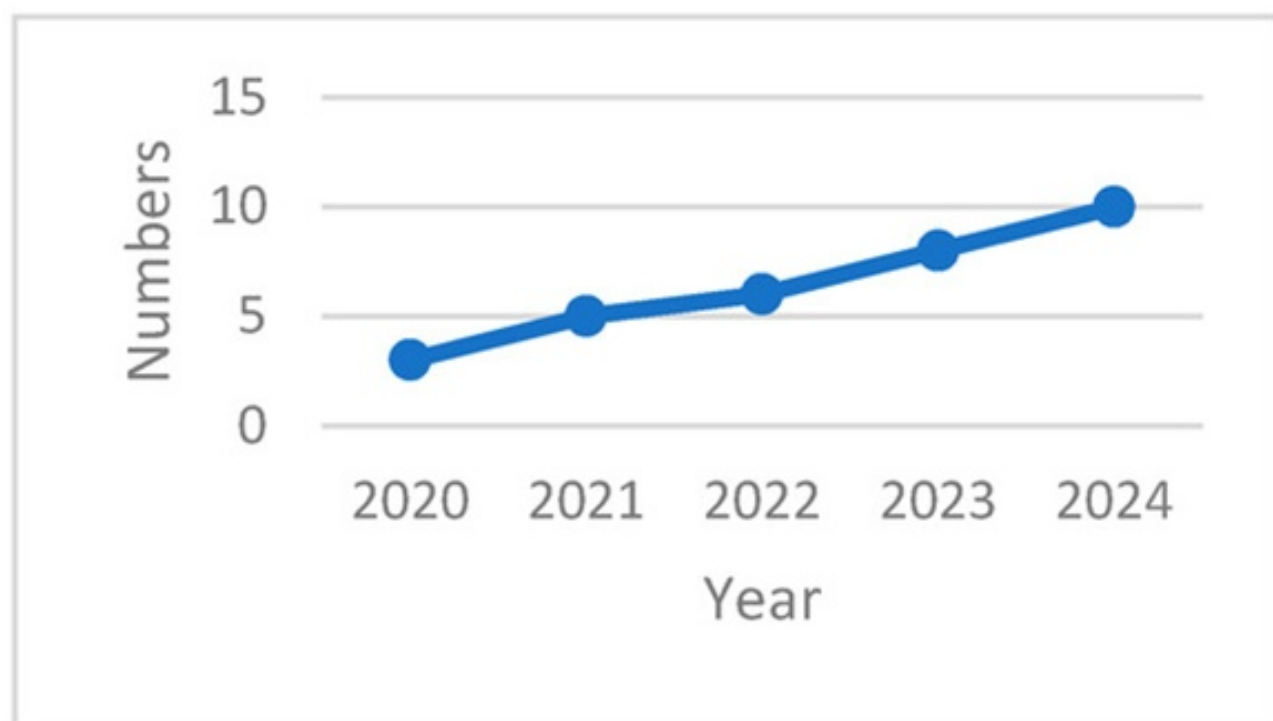


Figure 2. Number of relevant articles over years.

This review aims to deliver an exhaustive analysis of DRL-driven AIDs architectures, with an explicit emphasis on studies published from 2020 to the present. More crucially, prospective research vectors are highlighted for investigators by compiling, clarifying, and breaking down DRL traits and their execution in AIDs. The remainder of this manuscript is structured as follows.

Overview of DRL

The basic structure of DRL

DRL algorithms represent a potent category of learning models that determine how to maximize returns within a specific environment to reach defined objectives. They achieve this by interfacing directly with the environment and receiving feedback rewards, without necessitating a heavy reliance on antecedent experience [54]. This section delineates the core DRL framework, illustrated in **Figure 3**, alongside its primary learning and decision-making workflows, intended to offer an accessible introduction for new researchers.

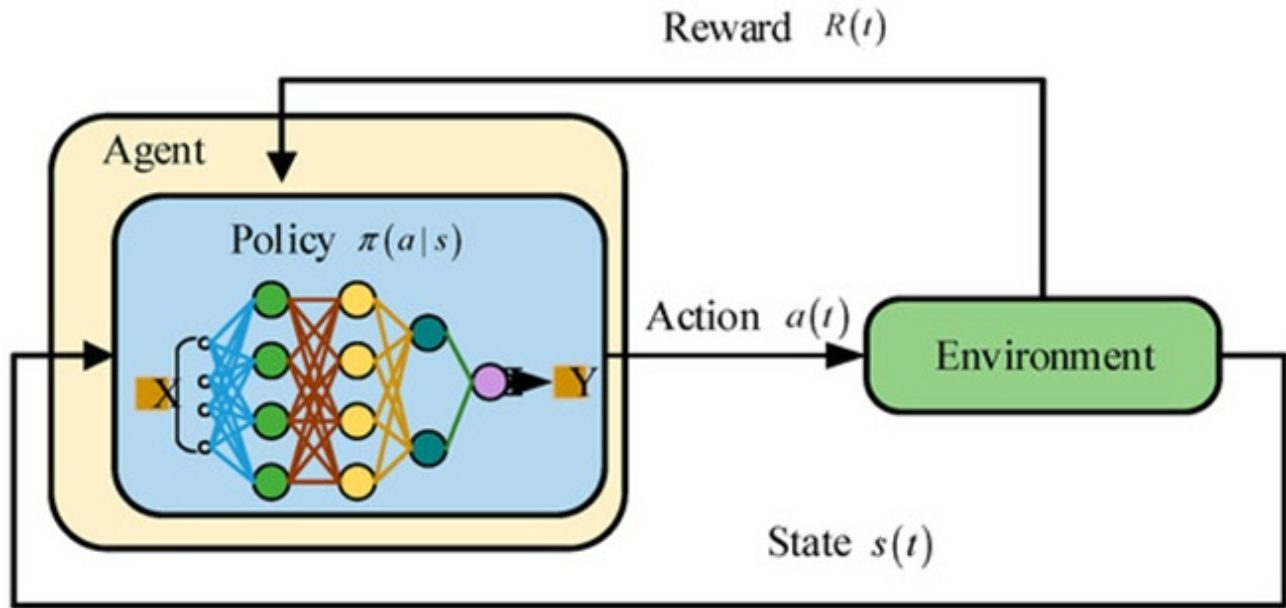


Figure 3. Basic structure of DRL.

DRL conceptualizes the environment as a Markov Decision Process (MDP), utilizing a five-tuple (S, A, P, R, γ) to define the structural components of decision-making problems. In this notation, S denotes the state space, A signifies the action space, $P(s'|s, a)$ characterizes the probability distribution for moving from the current state s to a subsequent state s' via action a , $R(s, a)$ represents the immediate reward, and γ serves as the discount factor, which modulates the relative priority of immediate rewards versus long-term returns. The objective of the agent is to develop a policy $\pi(a|s)$ through continuous interaction with the environment, thereby maximizing the anticipated cumulative reward.

The fundamental principle of DRL involves acquiring optimal behavioral patterns through successive interaction cycles. In a standard interaction sequence, the agent detects the current state s_t at time step t and selects an action a_t based on the policy $\pi(a|s)$. Once this action is executed within the environment, the environment provides a transition to a new state s_{t+1} and an immediate reward r_t . This cyclic process continues as the agent strives to maximize the expected future cumulative return. The cumulative return (Return) is defined as:

$$G_t = \sum_{k=0}^{\infty} \gamma^k r_{t+k} \quad (1)$$

In this expression, the discount factor $\gamma \in [0, 1]$ dictates the weight of future rewards regarding current decision-making.

To inform the agent's choices, RL identifies two primary value functions: the state-value function and the action-value function. The state-value function $V(s)$ denotes the expected cumulative reward initiated from state s under policy π :

$$V^{\pi}(s) = E_{\pi}[G_t | s_t = s] \quad (2)$$

The action-value function $Q(s, a)$ further assesses the expected cumulative reward following the execution of action a from a specific state s :

$$Q^{\pi}(s, a) = E_{\pi}[G_t | s_t = s, a_t = a] \quad (3)$$

Temporal-Difference (TD) learning serves as a vital technique for estimating these value functions. The TD approach integrates the recursive architecture of dynamic programming with the sampling principles of Monte Carlo methods, fundamentally utilizing the TD error δ_t to refine estimates. For the Q-value, the TD error is expressed as:

$$\delta_t = r_t + \gamma Q(s_{t+1}, a') - Q(s_t, a_t) \quad (4)$$

where a' is the action selected for the subsequent state s_{t+1} (based on policy or optimality criteria). The corresponding update rule is:

$$Q(s_t, a_t) \leftarrow Q(s_t, a_t) + \alpha \delta_t \quad (5)$$

where α represents the learning rate.

During the decision-making phase, RL strategies are generally categorized into value-based approaches and policy-based optimization.

1. Value-based algorithms, such as Deep Q-learning (DQN), utilize a greedy or ϵ -greedy policy to derive actions from the value function. This technique is typically restricted to environments with discrete actions or low-dimensional requirements:

$$a_t = \arg \max_a Q(s_t, a) \quad (6)$$

2. Policy-based algorithms, including policy gradient methods, explicitly develop a parameterized policy ($a|s$) and sample actions directly from it. The goal of policy gradient optimization is to maximize the expected cumulative reward of the policy, with the gradient defined as:

$$\nabla_{\theta} J(\pi_{\theta}) = E_{\pi}[\nabla_{\theta} \log \pi_{\theta}(a|s) \cdot Q^{\pi}(s, a)] \quad (7)$$

This approach naturally accommodates continuous action spaces and bypasses the action-search difficulties inherent in value-function methods.

In the context of AIDs, the environment is represented by the patient, while the action usually corresponds to the insulin dose for an insulin pen or the infusion rate for an insulin pump. By executing actions within this environment and calculating the resulting rewards, the agent learns strategies to minimize a specific loss function [55]. The overarching process is depicted in [Figure 4](#).

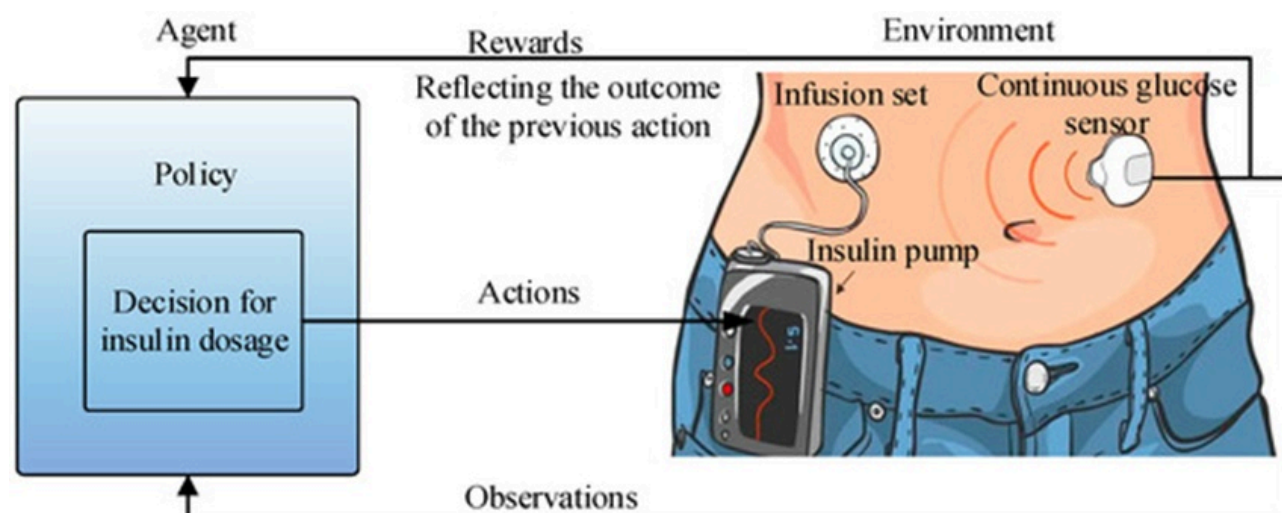


Figure 4. DRL-based AIDs.

In the clinical application of DRL-based AIDs, the agent determines insulin dosages based on its internal Policy and acts upon the Environment (patient) via the insulin pump. The CGM provides BGC data as Observations to the agent for subsequent decision cycles. The system utilizes Rewards to communicate the efficacy of the previous action. If the dosage decision maintains the BGC within the target parameters, a positive reward is assigned to signify a gain; if the BGC exits the target range, a negative reward is issued—proportional to the magnitude and duration of the deviation—to signify a loss, thereby guiding the agent toward optimized decision-making.

Classification of DRL and their characteristics

Since its development, DRL has been the subject of intensive research focused on creating algorithms specialized for various functional requirements. Within the scope of AIDs and contemporary research trends, DRL algorithms are broadly grouped into three classifications: (1) value-based and policy-based, (2) on-policy and off-policy strategies, and (3) model-based and model-free. This section offers a thorough analysis of the traits defining different algorithm types and their relevance to AIDs. The taxonomy of DRL algorithms is illustrated in [Figure 5](#).

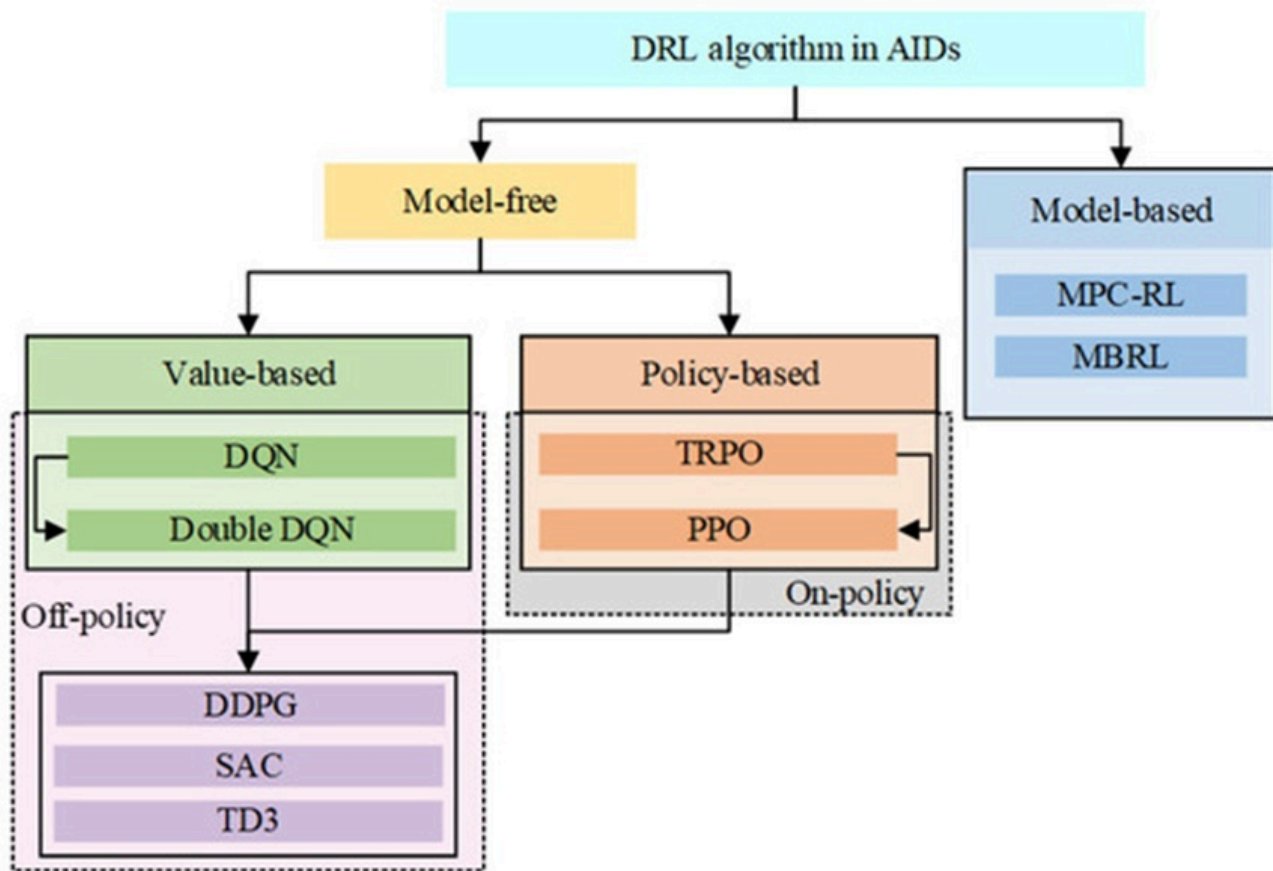


Figure 5. Classification of DRL algorithms (thick-bordered boxes represent different categories, while other boxes indicate specific algorithms).

1. **Value-based and Policy-based:** The value-based approach determines optimal interventions by directly maximizing the action-value function $Q(s,a)$. This framework is highly suited for discrete action spaces and has been utilized to formulate decision tree strategies based on blood glucose states. As a foundational approach, DQN integrates two distinct Q-networks (the current network and the target network) to evaluate values and select actions, respectively, which guarantees stable Q-value estimation [56-60]. Double DQN represents an upgraded iteration of DQN that embeds two sets of Q-networks (comprising two current Q-networks and two target Q-networks, totaling four) to mitigate the challenge of value overestimation. Specifically, during Q-value computation, it systematically selects the minimum value derived from one set of networks [61].

However, architectures relying solely on value functions face significant limitations within continuous action parameters, because actions cannot be directly derived via a continuous value function, ultimately degrading the control precision of AIDs. To surmount this obstacle, policy optimization techniques (such as PPO [62-68] and TRPO [69]) resolve intricate continuous control operations by directly optimizing the policy function $\pi_{\theta}(a|s)$. PPO serves as an enhanced variation of TRPO; both algorithms constrain policy updates using a trust region boundary to prevent the policy degradation often caused by excessive updates in deep networks.

Hybrid architectures combining value and policy mechanisms (such as DDPG, SAC, and TD3) have demonstrated outstanding capabilities. These schemes evaluate the value function while simultaneously refining the policy based on that evaluation, as illustrated in Figure 6. This configuration is highly adaptable and can assimilate powerful elements

such as Double Q mechanisms (SAC, TD3), deterministic policies (DDPG), exploratory stochastic policies (SAC), and delayed updates (TD3), displaying robust versatility within AIDs setups. Certain investigations implement both the Actor and Critic components in DDPG using DNN architectures to achieve superior outcomes [70-72]. Other studies apply the normalized advantage functions approach, which streamlines the DDPG network configuration, to AIDs frameworks in pursuit of enhanced computational efficiency [73]. Nonetheless, because DDPG operates as a deterministic gradient algorithm, it can suffer from inadequate exploration capabilities in highly complex environments. Furthermore, within a continuous and infinite space, the network topology of DDPG exhibits high sensitivity to hyper-parameter configurations and suffers from poor convergence [70]. To rectify these two shortcomings of DDPG, SAC [48, 65, 70, 74-79] introduces an entropy regularization term, managing the priority of entropy (exploration) via an entropy regularization coefficient. Secondly, SAC incorporates a total of five networks: one Actor, two Critics (the current Q-network and its corresponding target Q-network), and another pair of Critic networks, selecting the network that yields the smaller Q-value at any given step to alleviate the network Q-overestimation problem. Methods to enhance the SAC algorithm's capacity for state utilization have also been explored by deploying a Dual Attention network to expand the state inputs and assign them an attention score [74]. Recently, SAC configurations utilizing a dual-agent framework have been evaluated for AIDs [80]. The TD3 framework was introduced almost concurrently with SAC to overcome the specific vulnerabilities present in DDPG [81-84].

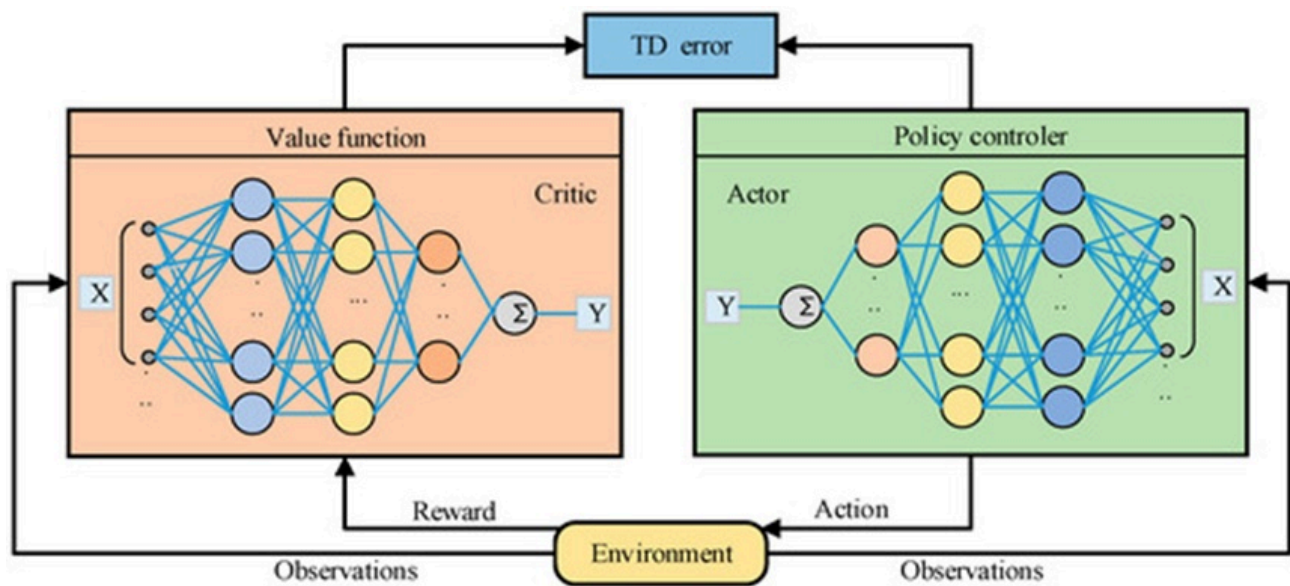


Figure 6. Actor-Critic algorithm.

2. **On-policy and Off-policy:** On-policy paradigms (such as TRPO and PPO) enforce strict alignment between the behavioral policy (the data collection strategy) and the target policy (the idealized operational strategy), dictating that the agent updates its weights exclusively using interaction data sampled from its current policy distribution. This method provides superior policy stability and convergence because On-policy frameworks consistently employ samples produced by the most recent policy for optimization, ensuring that policy updates match the current policy trajectory and avoiding distortion from historical or divergent policy distributions, thereby maximizing update accuracy [85]. Furthermore, On-policy methods confine the scope of exploration, anchoring it around the active policy distribution, which effectively mitigates the hazards linked to excessive exploration [85]. Within the context of dynamic blood glucose regulation, this attribute prevents abrupt, massive shifts in insulin dosing, thereby dampening unmanageable blood glucose volatility.

In contrast, Off-policy paradigms (such as DQN, DDPG, and SAC) decouple the behavioral policy from the target policy, focusing directly on discovering the optimal target policy. Consequently, these algorithms can exploit historical datasets or data generated by alternative policies during updates, dramatically improving sample efficiency. This feature makes Off-

policy methods exceptionally well-suited for repurposing retrospective data and optimizing simulated testing environments within AIDs. For instance, historical data collected from patients managed under different manual insulin injection strategies, alongside data generated via patient simulation software, can be utilized to pre-train Off-policy algorithms, minimizing dependency on live patient data while compressing the policy learning timeline. However, Off-policy strategies can be vulnerable to distribution shifts triggered by shifting patient dynamics (such as intra-patient variability or simulation model discrepancies), which mandates dependable environment modeling or targeted algorithmic counter-measures for resolution.

3. Model-based and Model-free: Model-free frameworks depend on directly learning policies or value functions from raw interaction data collected from the environment, thereby bypassing the requirement for highly precise mathematical modeling of a patient's internal insulin-glucose dynamics. This provides them with a major advantage when managing high-dimensional and complex assignments, particularly for individualized glycemic regulation. Nonetheless, model-free techniques demand an extensive volume of training interactions, which can introduce ethical dilemmas and prohibitive data collection overheads in real-world clinical deployment.

Conversely, model-based frameworks (such as MPC-RL [75] or MBPO [60]) combine explicit mathematical modeling of the patient's metabolic pathways with RL algorithms to drive policy optimization. The core advantage of this strategy centers on its capacity to explicitly map environmental dynamics, which reduces the necessity for live interaction data and delivers higher sample efficiency. For example, model-based frameworks can simulate a patient's postprandial glucose transitions, guiding the optimization of both insulin injection timing and dosage. However, structural model errors can drastically degrade policy execution, necessitating remediation through uncertainty estimation techniques or real-time model calibration.

DRL-Based AIDs

State space variables selection

Table 1 provides a summary of the state variables utilized in current research and examines the distinction between fully closed-loop and hybrid closed-loop AIDs.

Table 1. Selection of state space variables in the study.

Authors (References)	State Variables Used in the DRL Environment	Closed-Loop Configuration
Ian Fox [48]; Chirath Hettiarachchi [62, 66]; Dénes-Fazakas Lehel [68]; Senquan Wang [56]; Miriam Chlumsky-Hartmann [79]	Blood glucose concentration (BGC), insulin	Fully Closed Loop
Seunghyun Lee [63]	BGC, rate of change of BGC, insulin on board (IOB)	Fully Closed Loop
Francesco Di Felice [70]; Alan Mackey [81]; Phuwadol Viroonluecha [65]; Silvia Del Giorno [58]; Dénes-Fazakas Lehel [64]; Jingchi Jiang [78]	BGC	Fully Closed Loop
Zackarie Ellis [72]	BGC, insulin, error between BGC and 4.5 mmol/L	Fully Closed Loop

Sayyar Ahmad [57]	BGC, maximum value, minimum value, area under the curve	Fully Closed Loop
Mohammad Ali Raheb [73]	BGC, IOB	Fully Closed Loop
Guangyu Wang [86]	BGC, demographic information, diagnosis, symptoms, medications, laboratory test indices	Fully Closed Loop
Jonas Nordhaug Myhre [69]	BGC, insulin, IOB	Hybrid Closed Loop
Taiyu Zhu [71]	BGC, carbohydrate intake (CHO), IOB, bolus injection timing	Hybrid Closed Loop
Min Hyuk Lim [74]; Jinhao Zhu [77]; Mehrad Jaloli [80]; Anas El Fathi [67]	BGC, insulin, CHO	Hybrid Closed Loop
Zihao Wang [87]	BGC, timestamp, CHO	Hybrid Closed Loop
Taiyu Zhu [59]	BGC, CHO, basal insulin rate, bolus insulin, glucagon	Hybrid Closed Loop
Tianhao Li [60]	Dynamic and static variables extracted from electronic medical records	Hybrid Closed Loop
Harry Emerson [82]	BGC, IOB, CHO	Hybrid Closed Loop
Wenzhou Lv [75]	BGC, bolus insulin, CHO, IOB	Hybrid Closed Loop
Giulia Noaro [61]	BGC, CHO, rate of glucose change, carbohydrate ratio, standard postprandial insulin dose	Hybrid Closed Loop
Xuehui Yu [76]	Daily minimum and maximum BGC values, posterior distribution probabilities derived from prior and individualized information (e.g., age and weight)	Hybrid Closed Loop
Taiyu Zhu [83]	BGC, mean glucose, maximum glucose, minimum glucose, largest difference between adjacent measurements, percentage of hyperglycemic and hypoglycemic readings, timestamp, hours since last CHO intake, bolus insulin	Hybrid Closed Loop
Tristan Beolet [84]	BGC, insulin, IOB, total daily dose (TDD), carbohydrates on board (COB), timestamp, weight	Hybrid Closed Loop

On one hand, a fully closed-loop AIDs configuration assumes the omission of extra external inputs, thereby avoiding potential control biases that arise from imprecise estimations. On the other hand, hybrid closed-loop architectures incorporate external data (such as nutritional intake) into the state variables. This allows the agent to directly access information regarding external disturbances rather than requiring it to infer them. Consequently, if an external disturbance

estimator maintains sufficient precision, a hybrid closed-loop system may be considered superior for increasing agent efficiency.

In the aforementioned research on DRL-based AIDs, several studies have increased the dimensionality of the state space. This enables the agent to determine insulin interventions using a broader information set without necessitating the development of supplementary mathematical models, illustrating a key benefit of AI integration. Furthermore, tailored to the specific circumstances of different patients—for instance, those who can accurately report dietary data versus those who cannot—diverse state space configurations can be designed to maintain BGC regulation under varying conditions.

Action space variables selection

Physiological insulin secretion is typically categorized into two distinct components: first, a continuous, low-level release independent of nutritional intake, known as basal insulin, which is secreted in pulses over a 24 h period to stabilize BGC during fasting; and second, a substantial release triggered by the rise in BGC following a meal, known as mealtime or bolus insulin, which creates a curvilinear secretion wave [9]. Eq. 8 outlines the insulin bolus dose determined by the standard bolus calculator model [88], which is fundamentally linked to the computational objectives of DRL-based AIDs. The action variables identified in DRL-based AIDs are summarized in Table 2.

Bolus insulin = Meal insulin + Correction insulin – IOB (8)

Table 2. Selection of action space variables in the study.

Authors (References)	Basal Insulin Action	Bolus Insulin Action	Insulin Dose Action	Continuous Action Space	Discrete Action Space	Unbounded Action Space	Bounded Action Space
Ian Fox [48]; Chirath Hettiarachchi [62, 66]; Francesco Di Felice [70]; Zihao Wang [87]; Taiyu Zhu [83]; Guangyu Wang [86]; Anas El Fathi [67]; Miriam Chlumsky-Hartmann [79]; Dénes-Fazakas Lehel [68]	—	—	✓	✓	—	—	✓
Phuwadol Viroonluecha [65]; Wenzhou Lv [75]; Alan Mackey [81]; Mehrad Jaloli [80]; Tristan Beolet [84]	✓	—	—	✓	—	—	✓
Taiyu Zhu [71]; Mehrad Jaloli [80]	—	✓	—	✓	—	—	✓
Dénes-Fazakas Lehel [64]; Mohammad Ali Raheb [73]; Jinhao Zhu [77]; Min Hyuk Lim [74]; Zackarie Ellis [72]; Jingchi Jiang [78]	—	—	✓	✓	—	✓	—
Giulia Noaro [61]; Xuehui Yu [76]	—	✓	—	✓	—	✓	—

Jonas Nordhaug Myhre [69]; Harry Emerson [82]; Xuehui Yu [76]	✓	—	—	✓	—	✓	—
Silvia Del Giorno [58]; Sayyar Ahmad [57]	—	✓	—	—	✓	—	✓
Tianhao Li [60]; Taiyu Zhu [59]; Seunghyun Lee [63]; Senquan Wang [56]	—	—	✓	—	✓	—	✓

Supplement: Bounded and Borderless refer to the existence and absence of boundaries in the action space respectively; Continuous and Discrete indicate that the action space is continuous and discrete, respectively.

In **Table 2**, the parameters defining the action space encompass various behaviors, including insulin quantities (basal and bolus), administration timing, carbohydrate intake, or the refinement of bolus delivery by optimizing the underlying calculation parameters [89]. Research that did not differentiate between basal and bolus insulin combined these into a single output (Insulin doses) to manage the patient's BGC volatility.

In cases where insulin optimization is required only for mild conditions with steady intraday variations or for patients with long-term therapeutic experience, optimizing a limited number of daily mealtime doses is feasible, as their daily and postprandial requirements remain relatively stable. This minimizes the physiological and mental stress of data tracking while allowing the daily dose to be adjusted for long-term BGC trends. For patients experiencing severe intraday BGC volatility, particularly adolescents, daily optimization is insufficient. Glucose fluctuations must be monitored every 1, 3, or 5 min, and insulin delivery must be refined more frequently than once per day to manage dramatic shifts. Under such high-frequency administration, AIDs generate vast datasets, allowing DRL control algorithms to perform effectively due to their data-driven nature. Furthermore, DRL can identify specific patterns of blood glucose volatility by expanding the state space, effectively mitigating risks without needing to rebuild the mathematical model.

Table 2 indicates that whether an action space is discrete or continuous depends on the algorithm type. Most investigations have also introduced constraints to the action space to establish a safe “trust region” for insulin output, which accelerates algorithmic convergence and improves operational efficiency.

Reward function selection

This section examines the reward functions utilized in control algorithm development, which quantify the agent's performance based on actions that alter the environmental state. These functions can be customized based on task objectives and clinical expertise. **Table 3** lists the diverse reward functions identified in the literature.

Table 3. Selection of the reward function in the study.

Reward Function Type	Representative Studies and Example Formulation	Advantages	Disadvantages
Logarithmic (Risk-Based) Function	Ian Fox [48], Harry Emerson [82], Alan Mackey [81], Guangyu Wang [86], Chirath Hettiarachchi [66], Jinhao Zhu [77], Anas El Fathi [67], Tristan Beolet [84], and Miriam Chlumsky-Hartmann [79].	Quantifies blood glucose risk directly and encourages the agent to focus on	Does not explicitly account for glucose variability or fluctuations. In addition,

	Example: a logarithmic blood glucose risk function based on the transformation $(3.35506\log(g)^{0.8353}-3.7932)$, with an additional penalty term (α) applied outside the range of 10–1000 mg/dL . The value of the penalty parameter differs across studies.	high-risk glucose states, thereby helping prevent severe hyperglycemic and hypoglycemic events.	because patients exhibit different risk profiles, the function cannot easily provide individualized reward assessments.
Gaussian Function	Jonas Nordhaug Myhre [69], Mehrad Jaloli [80], Min Hyuk Lim [74], and Dénes-Fazakas Lehel [64, 68] . Example: $(R(g)=\exp(-\frac{1}{2}\frac{(g-g_{ref})^2}{\sigma^2}))$, where $(\sigma>0)$ and $(g_{ref}=127)$ mg/dL.	$g-g_{ref}$	$)),$ where $(\sigma>0)$ and $(g_{ref}=127)$ mg/dL.
Power Function	Francesco Di Felice [70], Silvia Del Giorno [58], Wenzhou Lv [75], and Jingchi Jiang [78] . Example: $(R(g)=-0.001(g-100)^2)$.	Directly penalizes deviations from the target glucose value, with larger deviations receiving increasingly severe penalties.	Produces symmetric penalties for high and low glucose deviations, limiting the agent's ability to distinguish between the different clinical risks of hyperglycemia and hypoglycemia.
Linear / Piecewise Linear Function	Taiyu Zhu [59, 71, 83], Tianhao Li [60], Mohammad Ali Raheb [73], Phuwadol Viroonluecha [65], Giulia Noaro [61], Seunghyun Lee [63], Zihao Wang [87], Zackarie Ellis [72], Sayyar Ahmad [57], and Senquan Wang [56] . Example: a piecewise reward structure assigning different rewards and penalties across glucose intervals such as 70–90 mg/dL, 90–140 mg/dL, 140–180 mg/dL, 180–300 mg/dL , and values outside these ranges.	Simplifies reward computation and provides explicit rewards or penalties for specific blood glucose intervals.	Discontinuities at interval boundaries may lead to unstable behavior near thresholds. Furthermore, interval selection can be subjective and may not generalize across patients or clinical conditions.
Other Reward Functions	Xuehui Yu [76] . Example: a reward function based on the maximum and minimum blood glucose values measured 1 day after insulin infusion , combining distance-to-target and distance-from-target terms within a safe glucose range of 70–180 mg/dL . The formulation includes a fixed margin parameter (β), a discount factor ($0<\alpha<1$), and distance metrics that quantify deviations inside and outside the target interval.	Provides a more detailed and fine-grained evaluation of blood glucose outcomes compared with standard reward formulations.	Does not explicitly consider glucose fluctuations over time and does not accommodate patient-specific glucose targets, despite differences in safe glycemic ranges among individuals.

In reward function design, the acceptable BGC range is defined as 70 mg/dL to 180 mg/dL (Time-in-range—TIR) [4]. The clinical risks of hypoglycemia and hyperglycemia are markedly different, with hypoglycemia generally presenting a more immediate threat. While hyperglycemia typically results in serious short-term consequences only in the case of ketoacidosis, hypoglycemia can lead to headaches, coma, or sudden death within hours. Furthermore, the magnitude of BGC volatility serves as a secondary metric for evaluating recommended insulin doses. Assuming BGC stays within the

70–180 mg/dL target, minimizing fluctuations improves control and reduces vascular damage. For iterative frameworks like DRL, a robust BGC-related reward function is essential to evaluate insulin dosing and define the problem's objectives.

Magni et al. [90] established a risk function where BGC values of 70 and 280 mg/dL correspond to a risk of 25, while 50 and 400 mg/dL correspond to a risk of 50. This function shows risk rising sharply during hypoglycemia and more gradually during hyperglycemia. Based on this, researchers have formulated various new reward functions that consistently yield negative values [48, 59, 77, 81, 82, 86]. Additionally, several studies integrated Gaussian functions into their reward structures [64, 69, 74, 89, 91–93]. Other researchers utilized the absolute relative error between actual BGC and mean normal BGC, resulting in a step reward between 0 and 1 [94]. Some reports analyzed additional glycemic characteristics, combining them with weighted factors for hypoglycemic and hyperglycemic components [95–98]. For instance, an exponential function using the constant e as a base served as a reward function, utilizing a target value of 127 mg/dL—the midpoint between hypoglycemia and hyperglycemia—to generate rewards between 0 and 1 [99]. Furthermore, many studies focused on extreme glucose events, utilizing penalty values (e.g., -3 , -2 , -1 , 0 , or weighted coefficients) across different concentration intervals (high, low, and normal glucose) [59–61, 64, 65, 69, 71, 73, 83, 87]. Other investigations set a specific target (e.g., 80 or 90 mg/dL) and used the negative absolute error from this value as the reward function [100, 101].

Poorly defined reward functions can result in "reward hacking," where the agent achieves high rewards through unintended or risky behavior. For example, if glucose fails to stabilize, an agent might maximize cumulative rewards by terminating the episode prematurely. To prevent this, termination penalties are applied to trajectories that enter hazardous glucose states (levels beyond established limits) [48]. In critical scenarios, this penalty should ideally be set to infinity, or a value sufficiently low to represent the impact on a patient's life expectancy [65]. Regarding safety, besides the action constraints mentioned earlier, some studies incorporate insulin doses directly into the reward function to discourage instability [99] and prevent overly aggressive dosing strategies.

Due to varied patient requirements, DRL-based AIDs research often necessitates customized reward functions to align with specific diagnostic and therapeutic needs. As the reward function is the primary driver of policy optimization, its design must prioritize safety to minimize dangerous exploration by the agent. Although current research has made progress, ensuring the safety of every individual action remains a challenge. Relying solely on reward signals may not entirely prevent hazardous actions; therefore, future work must further restrict the action space to ensure outputs remain within established safety boundaries.

Challenges to Be Addressed

The preceding synthesis of control algorithms, state spaces, action configurations, and reward functions demonstrates that integrating DRL into AIDs advances the footprint of AI within medical care. Nonetheless, several critical bottlenecks must be resolved to successfully deploy DRL frameworks in real-world AIDs applications.

Low sample availability

Substantial barriers exist when deploying NN-based DRL to clinical trials for automated insulin regulation [74]. A primary limitation stems from low sample availability; these algorithms typically require a massive volume of direct interactions with an environment, navigating successive successes and failures to synthesize policy adjustments from trial-and-error occurrences. In a clinical setting, however, patients cannot be deliberately exposed to metabolic failures like severe hypoglycemia or hyperglycemia. While simulation frameworks offer an alternative training ground, they necessitate highly precise, personalized patient models. Figure 7 illustrates the diverse strategies leveraged in current literature to surmount this obstacle. The substantial financial overhead of engineering high-fidelity models, executing extensive simulations to

encompass a broad spectrum of real-life behaviors, and gathering data over prolonged timelines remain major hurdles to applying DRL within AIDs.

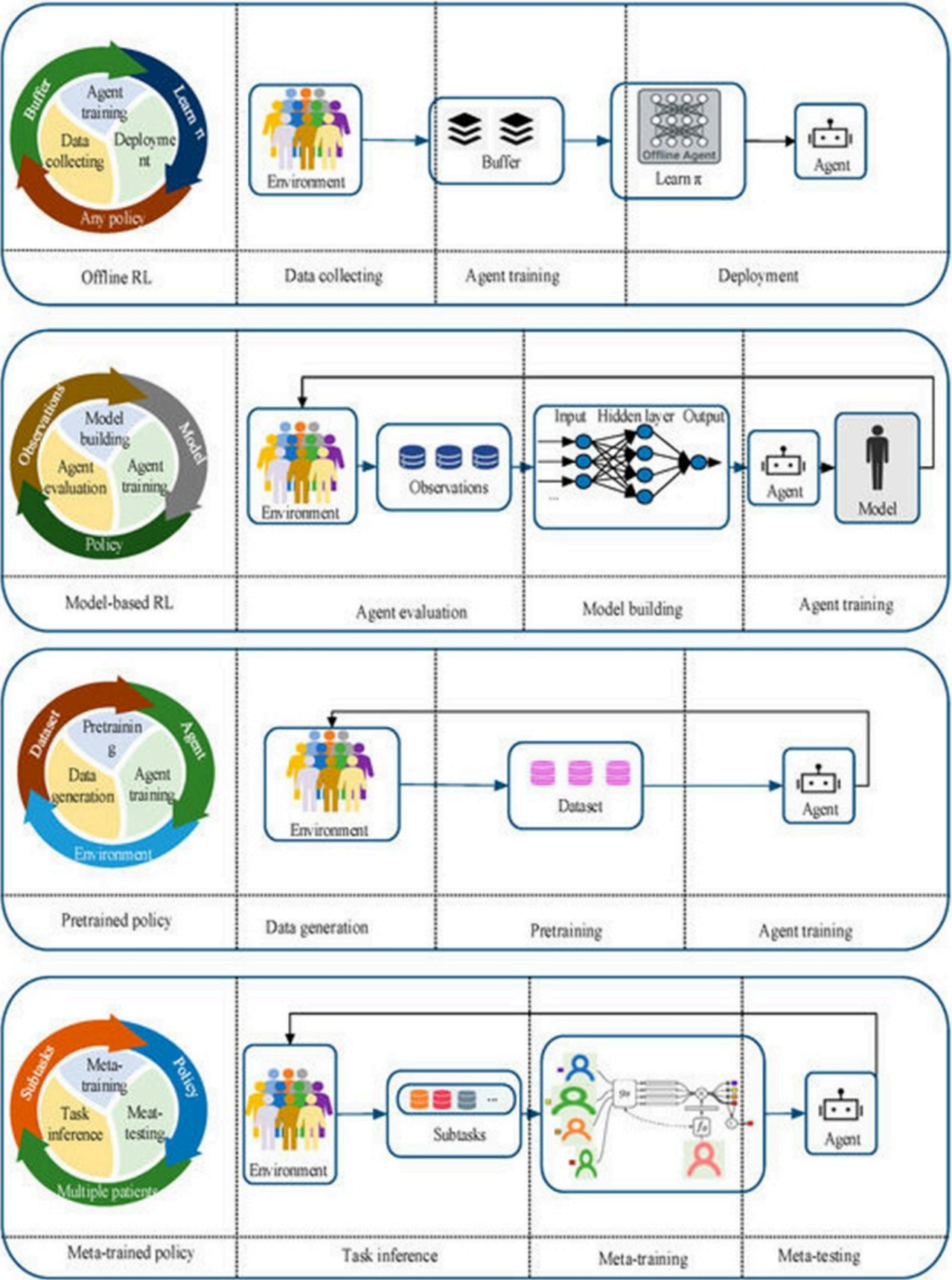


Figure 7. Existing solutions to the problem of low sample utilization in DRL.

As diagrammed in **Figure 7**, several investigations implement specialized NN architectures within DRL to optimize sample utilization, including offline DRL frameworks, model-based DRL variants, data-driven pre-training, and meta-training methodologies. Offline RL [60, 82, 83] trains the control agent entirely offline using historical datasets composed of pre-recorded state-action trajectories. Once this offline phase concludes, the agent is deployed directly into the live environment, bypassing the need for risky real-world exploration—a feature highly beneficial for AIDs given their strict safety constraints. Model-based RL [58, 102] utilizes the state-action pairs generated during the agent's interaction with the environment to construct an internal dynamics model. The agent then interacts with this learned model to output actions or continues its training using the generated synthetic data. A classic architecture like Dyna-Q [58] executes n Q-planning steps alongside a single standard Q-learning update after each environmental interaction. In model-based RL, the agent transitions to interacting primarily with the learned model following an initial real-world baseline period, lessening the necessity for subsequent empirical trial-and-error iterations.

During the pre-training phase, certain configurations [48, 59, 71, 102] (representing transfer learning in a narrower sense) initialize a portion or the entirety of the DRL neural network weights using an existing dataset. Following this pre-training step, the network is connected to the live environment for fine-tuning. For instance, developers can choose to fine-tune every layer of a generalized model or, alternatively, freeze the weights of early feature-extraction layers while optimizing only the higher-level decision components to mitigate overfitting. It has been noted that the early network layers capture broad, generalized physiological principles—such as suspending insulin delivery during downward glycemic trends—which are advantageous across all BGC regulation methodologies in T1D [59]. While the objective of pre-training is to discover parameter profiles that transfer effectively across tasks, there is no absolute guarantee that inheriting these trained parameters will yield optimal performance in distinct target tasks. Nevertheless, utilizing a dataset to establish network parameters prior to migrating them to a novel task provides a greater margin of safety than deploying a completely uninitialized network directly into a live patient environment.

Conversely, meta-training [75, 76] leverages multiple distinct tasks containing limited sample sizes to uncover shared underlying mechanisms, allowing the model to adapt swiftly to an unfamiliar task. For example, a patient's personalized clinical profile can be treated as a collection of separate sub-tasks through a probabilistic encoder embedding algorithm [63] to map common metabolic trends, accelerating adaptation when the architecture is deployed for a new diabetic individual. Unlike standard pre-training, the goal of meta-training is not to optimize network parameter performance on each individual sub-task, but rather to maximize the adaptive capability of the collective network parameters honed across those sub-tasks [103]. Although meta-training also constructs a model in advance, it differs from narrow pre-training by being inherently task-centric. It demands a smaller data footprint than purely data-driven pre-training, which further improves overall sample utility. The loss evaluated during meta-training represents the post-adaptation test loss after the sub-task is processed, whereas traditional pre-training evaluates the direct loss on top of the original baseline without a dedicated adaptation step.

Deploying meta-training techniques in tandem with DRL serves a safety function similar to pre-training by preventing a completely untrained network from driving the patient environment. However, narrow pre-training methods cannot adapt rapidly to novel tasks because disparities inevitably exist between a generic population baseline and a specific individual's physiology—a scenario that routinely occurs in personalized medicine. Consequently, some investigators integrate a probabilistic encoder to extract critical insights from fresh real-world experiences and individual patient parameters to navigate new tasks [76], offering a potential pathway to resolve distribution error challenges.

To further address low sample availability, future architectures could harness imitation learning to train models using datasets generated by expert medical clinicians. This approach enables the system to produce actions that closely mirror expert clinical strategies in real-world scenarios—a concept already validated in alternative fields [104] that holds promise for future DRL-based AIDs controller designs.

Personalization

Offline RL, model-based RL, and traditional transfer learning focus on training a control policy in advance using historical sample data. When this initial policy is introduced into a real-world setting, it frequently encounters the problem of distributional bias. Specifically, if the agent encounters a state that was absent from its training dataset, it may select an arbitrary action because the policy was never trained to navigate that specific state or its analogues. This arbitrary action can push the subsequent state even further away from the data distribution managed by the expert strategy. In behavioral cloning, this phenomenon is termed a "compounding error," while in offline RL, it is classified as a "distributional shift." Furthermore, a model optimized with an extensive dataset for a specific task typically requires a complete retraining cycle when transitioned to a different patient or condition, as illustrated in **Figure 8**, representing a highly time-consuming and labor-intensive process.

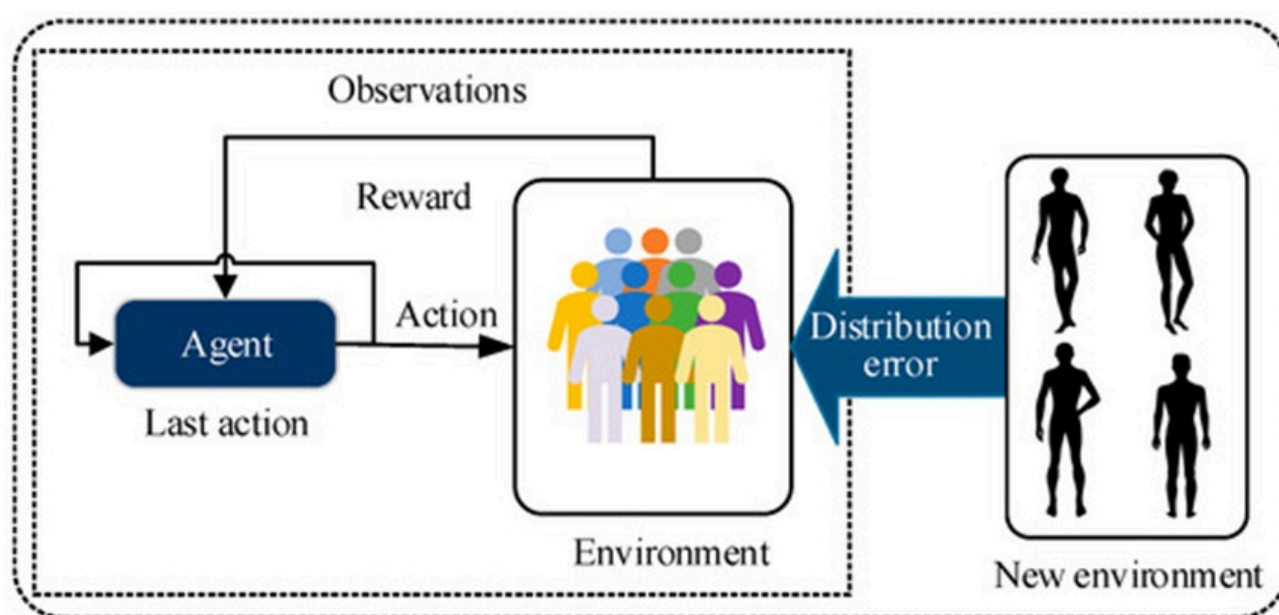


Figure 8. Distribution error problem.

The distributional bias depicted in **Figure 8** represents the core challenge of personalization: whether a control methodology can adapt rapidly when deployed across diverse patient profiles. Various solutions proposed in the literature are outlined below.

First, meta-learning methodologies [76] can effectively alleviate the computational overhead caused by extensive manual parameter tuning and the need to retrain models during task switching. This approach aims to equip the model with the capacity to "learn how to learn" tuning parameters, allowing it to rapidly assimilate new tasks by leveraging prior knowledge. Because it is structured around task-based training, it can master new concepts or metabolic profiles using a very limited number of samples, removing the need for total retraining when facing a new patient. Compared to the three methods mentioned previously, meta-trained models are expected to exhibit superior efficacy and rapid adaptation when deployed in real-world environments.

However, meta-learning demands that the training and testing tasks share highly similar distributions, which initially appears to conflict with the goal of resolving distribution errors. To mitigate generalization failures caused by divergent distributions, researchers have proposed using active learning to select and label the specific samples that are most representative of the testing task [76]. Thus, meta-learning provides a unique mechanism to buffer distribution errors due to its foundational "learning to learn" trait. Second, it specializes in mitigating extrapolation errors (the root of distributional errors) within offline RL frameworks [82].

To address model adaptation across different individuals, the "classification before training" paradigm [48, 61] offers an alternative solution by clustering patient populations into smaller, homogeneous cohorts to enhance personalization. However, this approach does not fundamentally increase the personalization capacity of an individual model, nor does it guarantee that all unique physiological traits are accommodated. The human body is highly complex; it is impossible to perfectly differentiate patients through categorical classification alone or to construct a separate mathematical model for every individual.

The "population-then-migration-to-individual" strategy [59, 71, 83] also suffers from distributional errors and has been shown to elevate catastrophic failure rates—specifically via "overfitting"—when training is executed on individual patient profiles. Other alternatives include customizing the reward function, action space, or observation frequency [65, 102], which serve as valuable complementary measures due to the highly adjustable nature of these parameters, as well as personalizing hyperparameters [63].

Furthermore, for advanced personalization, inverse DRL can be explored to decipher individual patient preferences and metabolic goals directly from historical data without requiring explicit patient classification. By inferring the underlying objective function from observed clinical behaviors, the system can automatically learn the target goals of the glycemic control task without needing manually defined, personalized parameters. This framework also illustrates how the model learns from the patient, thereby improving clinical interpretability, and can facilitate the design of personalized action spaces using clinically verified individual parameters [62].

Security

The challenge of low error tolerance (i.e., safety), which represents a major feasibility hurdle in the clinical integration of DRL with AIDs, involves two primary obstacles: the scarcity of training samples during the initial phases of controller development and the implementation of post-protection safeguards. The former issue was previously analyzed. The latter problem, characterized as "late-blooming measures," refers to the capacity of a clinician to intervene with both the controller and the patient when predictions are erroneous, thereby providing contextual insights into the internal states of the system. Most investigations have addressed this through various strategies, such as "imposing termination penalties on reward functions," "narrowing the action space" [48, 61, 70, 71, 87, 89, 105], "restricting the search strategy" [83], and utilizing a "threshold pause" [59] to prioritize security. Others have introduced "switching control" concepts [58, 74, 75] to transition to a more conservative strategy when DRL performance degrades.

In the future, hierarchical DRL may be implemented to manage distinct safety tiers—for instance, utilizing a primary control policy for standard operations and triggering a high-level emergency protocol upon detecting hazardous states. Safe Reinforcement Learning also offers a pathway to enhance AIDs security by incorporating constraints directly into the policy optimization process to prevent the agent from entering unsafe states. This approach has yielded positive results in areas such as robotics [106] and autonomous driving [107]. Furthermore, prior to practical deployment, comprehensive evaluation is mandatory, alongside the establishment of rigorous constraints to ensure the automated decision-making process is as secure as possible. Crucially, collaboration with regulatory bodies is required to define extensive testing protocols and safety criteria for every stage of the proposed methodologies.

Practical Applications

Recently, DRL algorithms have been introduced as a prospective control technology for AIDs, particularly for T1D management. The efficacy of these methods is typically demonstrated through simulation studies. Real-world clinical data remain difficult to acquire due to the precision required by healthcare professionals during sensor data collection and the ethical complexities involved in its use [49, 81]. Furthermore, AIDs face practical challenges that hinder the seamless implementation of DRL, including high costs that restrict widespread use without insurance coverage. Data from current users indicate that mean time-in-range for glucose concentration has improved compared to traditional insulin dosing methods. Early AIDs models were limited in their ability to interpret certain dynamic glucose variations, leading to

suboptimal insulin delivery; however, user feedback has facilitated algorithmic refinements. Usability has also progressed via enhanced user interfaces. Current systems depend on glucose and insulin data and require manual logging of meals and exercise. Multivariable AIDs utilizing additional sensors may eventually remove the need for manual entries and improve situational awareness. Technological progress and market competition are expected to further refine AIDs and reduce device costs. **Table 4** summarizes recent literature concerning the practical application of DRL in AIDs, with "✓" indicating the adoption of specific features.

Table 4. Practical applications of DRL algorithms in AIDs.

Authors (References)	Virtual Platform	Mathematical Model	Electronic Medical Records	Real Subjects	Virtual Subjects	Offline Clinical Datasets	T1DM	T2DM	No Diabetes- Type Distinction
Xuehui Yu [76]; Harry Emerson [82]; Chirath Hettiarachchi [66]; Mehrad Jaloli [80]; Wenzhou Lv [75]; Giulia Noaro [61]; Zihao Wang [87]; Ian Fox [48]; Chirath Hettiarachchi [62]; Jinhao Zhu [77]; Seunghyun Lee [63]; Alan Mackey [81]; Phuwadol Viroonluecha [65]; Min Hyuk Lim [74]; Taiyu Zhu [83]; Taiyu Zhu [59]; Anas El Fathi [67]; Sayyar Ahmad [57]; Jingchi Jiang [78]; Miriam Chlumsky- Harttmann [79]; Senquan Wang [56]	✓	—	—	—	✓	—	✓	—	—
Taiyu Zhu [71]	✓	—	✓	—	✓	✓	✓	—	—
Silvia Del Giorno [58]	—	—	✓	—	✓	—	✓	—	—

Jonas Nordhaug Myhre [69]; Francesco Di Felice [70]; Dénes-Fazakas Lehel [64, 68]	—	✓	—	—	✓	—	✓	—	—
Mohammad Ali Raheb [73]	—	✓	—	—	✓	—	—	✓	—
Zackarie Ellis [72]	—	✓	—	—	✓	—	—	—	✓
Tianhao Li [60]	—	—	✓	—	—	✓	—	—	✓
Tristan Beolet [84]	—	—	✓	—	—	✓	✓	—	—
Guangyu Wang [86]	—	—	✓	✓	—	—	—	✓	—

As shown in **Table 4**, the majority of current research utilizes simulators to generate data for agent training, with virtual patients serving as the primary test subjects. This creates a significant gap between simulated and real-world environments, as these studies often employ fixed patient parameters and meal settings while ignoring disruptive factors like physical activity. A few studies use mathematical models to generate data directly; while similar to simulators, these models often fail to represent real-world complexity accurately. Recently, there has been a gradual shift toward using real electronic medical record (EMR) datasets for training and evaluation, signaling future clinical potential. One recent investigation even implemented DRL for clinical evaluation under highly restrictive safety constraints [86].

Data

The prevalence of simulated data over real-world data in the studies listed in **Table 4** may be attributed to several factors:

(1) The collection of extensive clinical data is typically costly and time-intensive [75, 108], with only a limited number of studies conducting clinical validations [86, 94]. The algorithms in **Table 4** utilized simulated data spanning 5 days [76], 6 days [75], 30 days [59], 180 days [61, 71], 1000 days [48], and 1050 days [77] to develop dosing strategies. It is considered unethical to allow unproven DRL algorithms to manage BGC in vivo for several years without guaranteed safety [82]. To test the scalability of offline DRL in complex environments, algorithms should be trained and assessed using authentic retrospective patient samples (e.g., the OhioT1DM dataset) [83], highlighting the necessity of out-of-strategy evaluations for offline models [83, 109].

(2) Testing algorithms on human subjects without prior preclinical validation or safety constraints involves high risks [71]. The nature of DRL makes simulated environments particularly suitable, as the model learns through trial-and-error; conducting these iterations on virtual subjects is vital to prevent dangerous outcomes for patients [61].

However, reliance on in-silico data presents several drawbacks:

(1) Simulations frequently overestimate the efficacy of interventions because they do not account for all the uncertainties and disturbances present in daily life [89].

(2) Patterns extracted from large simulated datasets do not always align with clinical reality [76].

Regarding the first point, several studies argue that if a system does not rely on dietary announcements to function, accurate simulation is sufficient and real-world data are not required for the learning phase—provided that precise mathematical models with patient-specific parameters are used [73]. Nevertheless, the lack of real-world data remains a hurdle for adapting this approach to other conditions. A recent proof-of-concept trial demonstrated the feasibility of DRL for hospitalized patients with T2D, showing that the safety and effectiveness of AI protocols were comparable to those of physicians [86].

Regarding the second point, the distribution error challenge is detailed. While simulators may not perfectly capture inter-patient differences or long-term physiological changes, the FDA-approved UVa/Padova simulator [110] represents a significant milestone [48]. This simulator is the most frequently used, followed by the minimal models of AIDA and Bergman [49, 111, 112]. While acquiring real-world data remains challenging, there is a clear need to transition from simulated to clinical data [94] to facilitate algorithm validation [49]. To improve sample utilization, several studies [59, 61, 62, 71, 75, 76, 82, 87] used simulators or datasets for "pre-training" models before testing. Recent efforts have also combined clinical electronic health record (EHR) data with DRL to recommend dosages across various physiological states. This approach incorporates more complex variables and mortality data than typical glucose management simulations [60] and is better suited for inpatient management and other EHR-based research [101, 113–115].

Instrument noise also affects data quality; previous work has shown that CGM noise impacts model performance [71]. Real-world uncertainties—including CGM artifacts, physical exertion, and health fluctuations—can alter BGC levels and reduce control effectiveness. Future research should prioritize clinical validation and the integration of stochastic noise, physical activities, acute stress, and varying insulin sensitivity into simulators to better approximate real-world scenarios.

Computational power

A primary drawback of DRL is the computational intensity of the training process [69]. Studies indicate that over 1000 days of simulated training are often required to achieve superior BGC control [48, 59, 61, 62, 77, 116], creating high hardware demands. Using mid-level hardware (Intel i9-10920X, 64 GB RAM, 2 NVIDIA RTX 2080 GPUs), hyper-parameter training for a single patient typically requires 7 to 10 days, even when processing 4 to 8 patients simultaneously [65]. Consequently, maintaining a compact state space model is essential to avoid memory issues during multi-day simulations. Some estimates suggest that as the state space expands, GPU requirements could increase by a factor of 10 [102]. Training can be accelerated using methods such as prioritized memory replay [117, 118] or nonlinear action mapping [62], though efficiency remains a concern.

Regarding algorithmic embedding, technological progress in medical devices and communications has linked mobile applications, CGMs, and insulin pumps more closely, facilitating AIDs portability. Researchers have begun integrating simple control logic into apps to manage or recommend insulin infusions for clinical evaluation [119–122]. While current models are typically trained on cloud servers, future work may involve migrating these into smartphone operating systems (e.g., iOS, Android) to allow for local, real-time DRL training using the device's CPU. Such data-driven algorithms could also benefit individuals with T2D on insulin therapy. Once matured, these models can be embedded into smartphone applications for clinical trials [71]. For instance, DRL models developed in TensorFlow can be deployed on embedded devices via TensorFlow Lite. Furthermore, embedded algorithms can be continuously optimized using new data from CGMs, insulin pumps, activity monitors, and user inputs [59].

Comparison of Main Control Algorithms for AIDs

Model predictive control

The MPC algorithm forecasts future BGC trajectories with the explicit objective of driving the current BGC toward the targeted physiological range. The runtime execution of MPC involves two sequential operations: constructing the predictive mathematical model and solving the constrained optimization problem. Three distinct paradigms exist for

building the underlying predictive model: data-driven modeling, physiological first-principles (typically compartmental) modeling, and hybrid modeling architectures. Various model-based control frameworks rely on capturing the physiological glucose-insulin kinetics of the human body, and substantial research has focused on fitting these physiological equations [38, 123]. Solving the constrained optimization problem aims to minimize the deviation between the predictive model outputs and reference target targets, with the generalized structural architecture depicted in Figure 9.

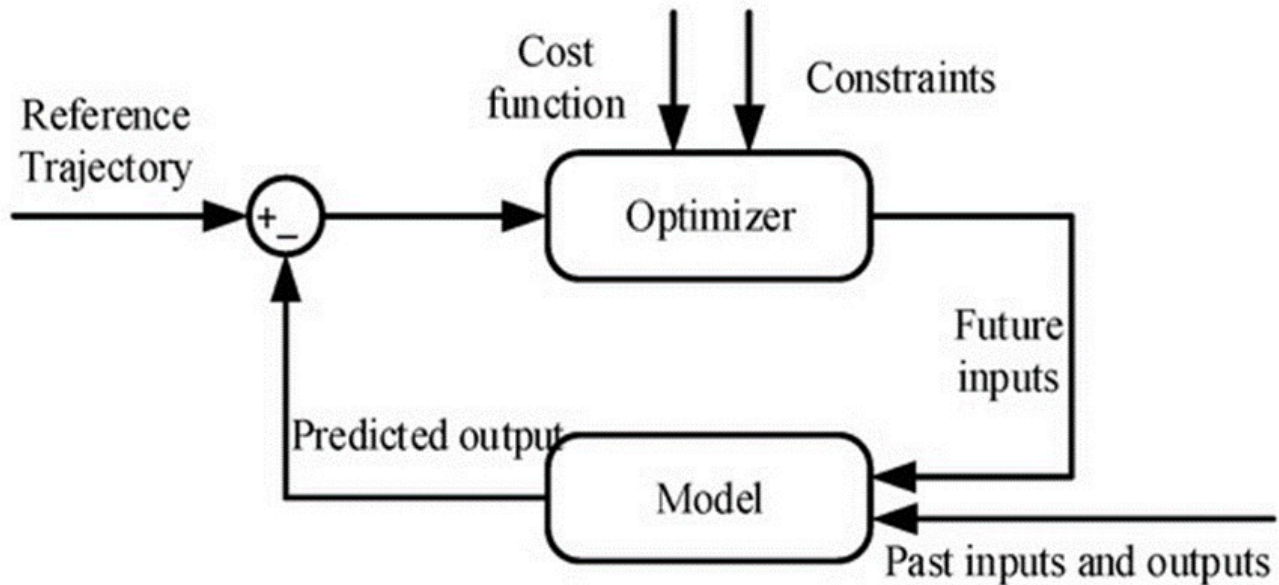


Figure 9. Model predictive control structure diagram.

Predictive modeling is traditionally executed utilizing the state space equations formulated below [90]:

$$\begin{cases} x(k+1) = Ax(k) + Bu(k) + Md(k) \\ y(k) = Cx(k) \end{cases} \quad (9)$$

where $x(k)$, $u(k)$, $d(k)$, and $y(k)$ represent the state vectors, control inputs, external disturbances, and measured outputs of the system, respectively, while A , B , M , and C denote the corresponding parameter matrices. The optimization control objective is typically formalized as a quadratic polynomial expression [124]:

$$J(u) = \|y_k - y_r\|^2 + \lambda \|\Delta_k - \Delta\|^2 \quad (10)$$

where y_k is the current measured BGC value of the patient at that sampling moment, y_r is the target BGC reference value, Δ represents the parameters linked to the regularization term, and λ acts as the regularization coefficient bounding the range 0–1.

The MPC algorithm operates based on historical BGC trends (derived from recorded CGM signals or estimated from raw CGM data) along with active insulin infusion records. BGC forecasting is achieved via the established predictive model, which subsequently determines the most appropriate current insulin infusion via the numerical optimizer and repeats this computation at each successive sampling interval to generate the optimized insulin infusion rate for the next step (Figure 10).

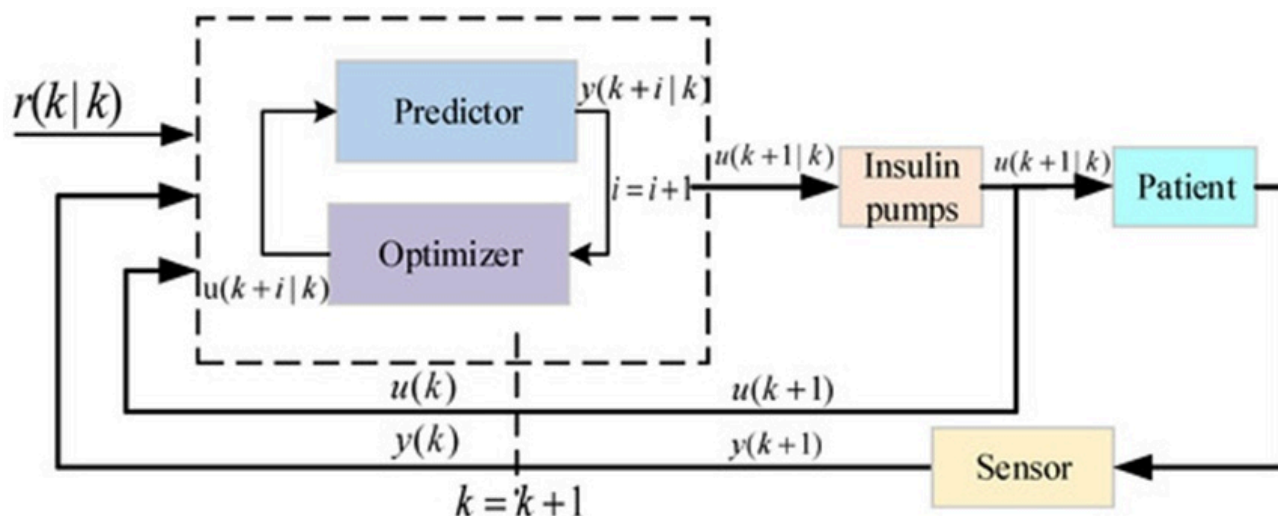


Figure 10. Structure of model predictive control algorithm for Artificial Pancreas.

The DRL control framework presents numerous structural advantages over the frequently utilized MPC methodologies in contemporary AIDs. These distinct advantages are detailed below:

1. Adaptation to Potential Perturbations

The DRL control algorithm possesses an inherent advantage when managing uncertain disturbances, which can be contrasted with the MPC algorithm through the following observations. First, because MPC relies on explicit models for forecasting, and explicit equations cannot capture all environmental perturbations, MPC remains highly vulnerable to unmodeled disturbances. Furthermore, the majority of these models are linear (data-driven) formulations structured as follows:

$$y(k) + a_1 y(k-1) + \dots + a_{n_a} y(k-n_a) = b_0 u(k-1) + \dots + b_{n_b} u(k-n_b) + f_0 \varepsilon(k) + \dots + f_{n_c} \varepsilon(k-n_c) \quad (11)$$

where y represents the discrete-time (k) BGC concentration data of patients with type 1 diabetes, u is the insulin infusion rate, ε is the model prediction error, and $a_1, \dots, a_{n_a}, b_0, \dots, b_{n_b}, f_0, \dots, f_{n_c}$ constitute the unknown parameters requiring estimation within the model [125].

To replace the linear framework described above, non-linear models can be developed. These may incorporate physiological first-principles kinetics or neural network (NN) topologies to map the non-linear relationship between inputs and outputs. **Figure 11** illustrates a standard feedforward NN architecture.

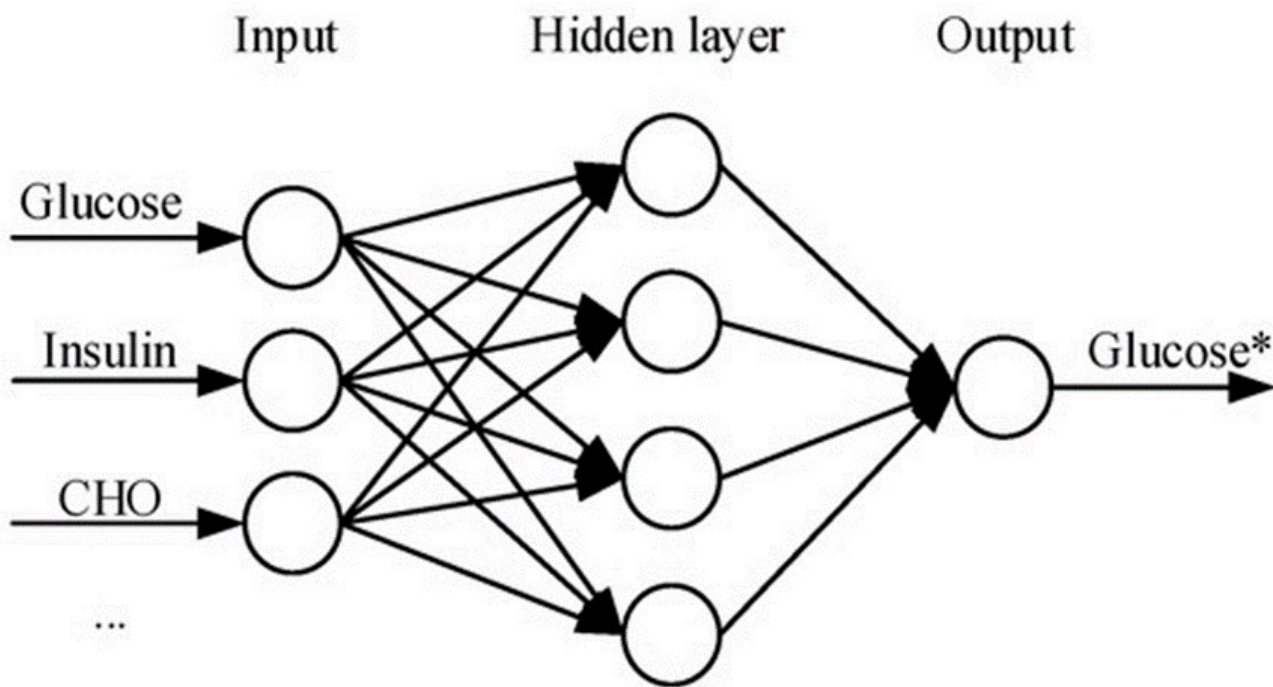


Figure 11. Feedforward Neural Network Glucose Prediction Model. (Glucose* represents the predicted value of blood glucose).

When an NN is deployed, calculating the state variables cannot be formulated as a direct equation embedded within the optimization loop. This limitation arises because the function $f(\cdot)$ in the expression $y = f(x,u)$ cannot be stated explicitly. Consequently, it is impossible to solve directly for the control output u using the state space equations, meaning subsequent optimization must rely on an iterative traversal across all conforming control variations, as shown in [Figure 12](#).

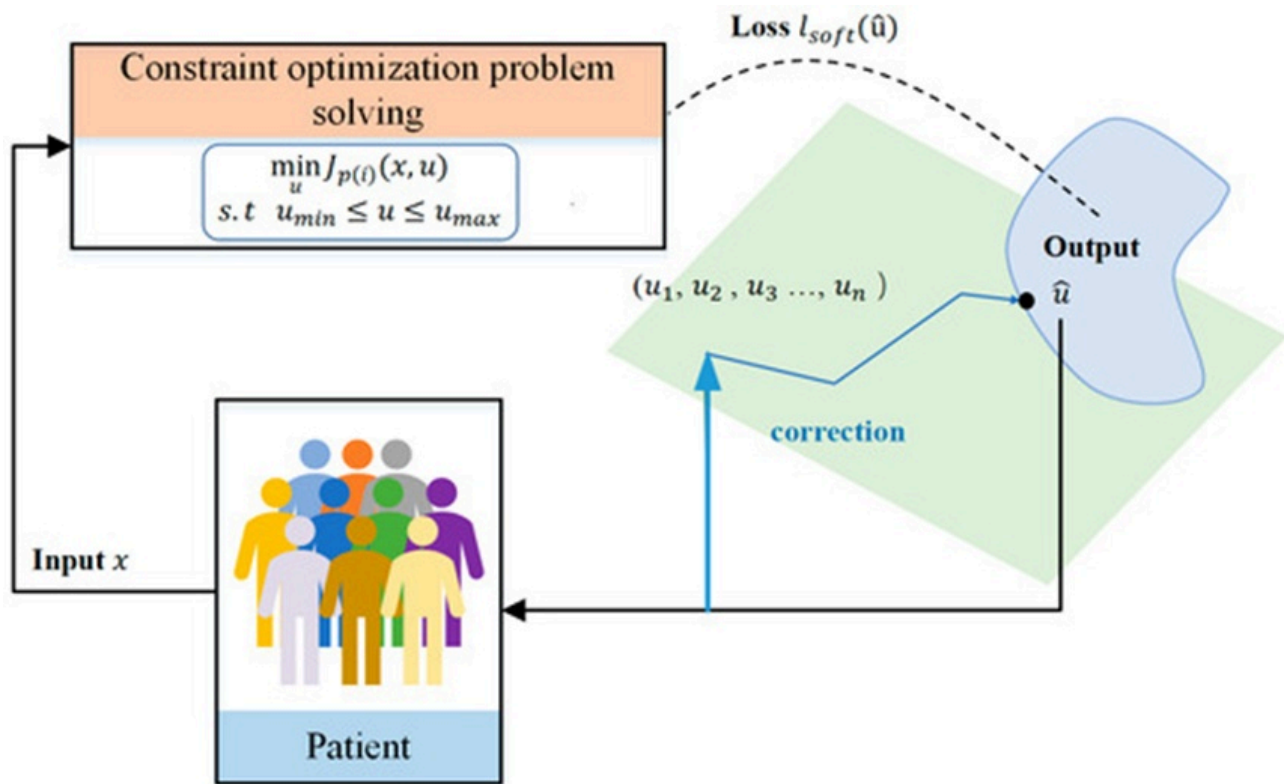


Figure 12. Ergodic solution in AIDs using MPC.

Although predictive frameworks utilizing non-linearities capture non-linear trends more effectively and handle multi-dimensional inputs, two core challenges persist:

- (1) The model cannot adapt dynamically to unmodeled perturbations because the predictor inputs are restricted to factors already explicitly known to alter BGC.
- (2) Even when utilizing a neural network base, MPC must resolve the optimization problem at every individual time step (via traversal solving), which can become highly complex in volatile, uncertain clinical environments.

In contrast, since DRL functions as a value-based paradigm, the product of the state transition probabilities of all potential future states and their corresponding values is natively integrated into the value estimation equations, as outlined in [Eqs. 2 and 3](#). Thus, the DRL framework adapts fluidly to latent perturbations, whereas the MPC algorithm cannot adapt until the disturbance is either explicitly quantified or its physiological impact manifests in the measured output (the BGC reported by the CGM), introducing significant latency into the control loop.

2. Responding to patient variability

AIDs control logic must adapt swiftly to distinct patient profiles as well as the inter-day and intra-day variations within a single individual, which defines the challenge of inter- and intra-patient variability.

For the MPC algorithm, although continuous optimization of the predictive model has shown promising clinical outcomes for glycemic control, the controller may fail to deliver the desired accuracy because it can never perfectly replicate actual human physiological pathways. The structural boundaries of a predictive model with static parameters lead to performance degradation when large, sudden disturbances occur [\[126\]](#). Moreover, while MPC parameters can be

customized for individual patients, personalized parametric tuning remains highly invasive and costly [127]. Additionally, executing this intensive online optimization problem on embedded medical hardware increases computational overhead.

Conversely, DRL control algorithms, by virtue of their environment-interacting architecture, can differentiate between distinct patient behaviors or the changing dynamics of the same patient across separate temporal intervals. This is accomplished by implicitly modulating state transition probabilities in a closed-loop manner through active environmental feedback.

3. Computational efficiency

Within the DRL framework, once the underlying neural network completes its training phase, there is no requirement to execute optimization algorithms at every discrete time step like MPC, yielding a significant increase in real-time operational efficiency.

4. Exploratory capacity

MPC utilizes predictive principles to partially mitigate the challenges of physiological time delays. Similarly, DRL incorporates predictive forecasting via its value function, which represents an estimation of the cumulative future reward. However, MPC evaluates state quality based on y_p —the projected future BGC trajectory, which is entirely dictated by the precision of the predictive model—and derives the control input u directly from y_p . This means the performance of MPC is tightly coupled to model accuracy.

In contrast, DRL evaluates state quality via the value metric Q , representing the overall return expected from future states. The control input u is not derived directly from Q via a rigid equation; instead, it utilizes an exploratory strategy to prevent the agent from becoming trapped in localized sub-optimal policies.

5. Adaptation to delayed insulin action

Because DRL prioritizes maximizing the long-term cumulative reward, its value function captures the expected value of future rewards. Consequently, an executed action can account for long-term physiological impacts, aligning with the inherent delay in insulin metabolic action, an area where several studies have introduced innovations [63, 102].

$$\begin{aligned}
 V(s) &= E[G_t | S_t = s] \\
 &= E[R_t + \gamma R_{t+1} + \gamma^2 R_{t+2} + \dots | S_t = s] \\
 &= E[R_t + \gamma(R_{t+1} + \gamma R_{t+2} + \dots) | S_t = s] \\
 &= E[R_t + \gamma V(S_{t+1}) | S_t = s]
 \end{aligned} \quad (12)$$

Table 5 outlines the primary benefits and drawbacks of both methodologies while identifying prospective research vectors.

Table 5. Comparison between MPC and DRL.

Aspect	Model Predictive Control (MPC)	Deep Reinforcement Learning (DRL)
Advantages	1. Strong capability for predicting future system behavior. 2. Effective management of operational constraints. 3. Well suited for multivariable control problems.	1. Does not require an accurate mathematical model of the system. 2. Eliminates the need for complex optimization calculations during operation. 3. Exhibits strong adaptability to changing conditions. 4. Capable

		of operating in highly complex and uncertain environments.
Disadvantages	1. Performance is highly dependent on the accuracy of the underlying model.2. Computational requirements can be substantial.3. Limited ability to adapt to significant changes in system dynamics.	1. Requires large amounts of training data.2. Decision-making processes are often difficult to interpret.3. Safety and reliability may be compromised in scenarios not represented in the training data.
Future Research Directions	1. Development of more adaptive and robust predictive models.2. Integration with advanced detection and classification techniques.3. Reduction of computational burden to facilitate real-time implementation.	1. Enhancement of data efficiency and learning performance.2. Improvement of model transparency and interpretability.3. Strengthening of safety frameworks and reliability mechanisms.

Based on the preceding analysis, two primary alternatives exist for constructing AIDs control architectures capable of resolving the challenges examined in this paper. The first paradigm relies on MPC configurations that deploy adaptive models driven by recursive system identification alongside multivariable control networks. These architectures leverage real-time data to identify diverse disturbances (such as physical exertion) and project their impact on BGC well before those changes manifest physiologically [128,133]. While this framework requires real-time optimization that imposes a specific computational load, smartphone-based applications are currently undergoing clinical testing. The second paradigm involves substituting the traditional MPC module with a DNN-based controller trained on extensive simulation data generated via the UVA/Padova simulator [134]. Both technologies can be successfully hybridized with DRL to capitalize on their respective strengths and engineer robust, high-performance AIDs.

Proportional integral derivative

The PID algorithm leverages three distinct mathematical terms to replicate, as closely as possible, the physiological transport dynamics of insulin naturally secreted by human \beta cells [135]. Within the architecture of an AID controller, the proportional component computes the volume of insulin to be delivered relative to the instantaneous deviation between the measured BGC and the baseline target BGC. This term is vital for modulating insulin adjustments to stabilize BGC at the desired threshold. Meanwhile, the derivative component is engineered to accelerate or decelerate insulin delivery in response to rapid fluctuations in the BGC profile, formulated as follows:

$$PID(t) = K_P(G - G_0) + K_I \int (G - G_0)dt + K_D(\frac{dG}{dt}) \tag{13}$$

where $PID(t)$ represents the calculated insulin actions, G_0 denotes the target BGC, G is the real-time continuous measurement of BGC, and K_P , K_I , and K_D represent the control gains assigned to the proportional, integral, and differential terms, respectively.

Table 6 details the primary advantages and disadvantages of both methodologies while exploring prospective domains for future research.

Table 6. Comparison between PID and DRL.

Aspect	Proportional–Integral–Derivative (PID) Control	Deep Reinforcement Learning (DRL)
--------	--	-----------------------------------

Advantages	1. Requires only a small number of parameters and is straightforward to implement.2. Supported by a well-established and mature theoretical foundation.	1. Demonstrates strong adaptability to changing conditions.2. Does not rely on precise mathematical models of the system.3. Capable of operating effectively in complex and uncertain environments.
Disadvantages	1. Restricted by the fixed structure of the controller equation.2. Limited adaptability to dynamic physiological changes.3. Unable to explicitly manage operational constraints.4. Less effective in addressing complex system dynamics and nonlinear behaviors.	1. Requires large quantities of training data.2. Limited interpretability of learned decision-making policies.3. Potential safety concerns when encountering situations not represented in the training data.
Future Research Directions	1. Further development of adaptive parameter-tuning strategies.2. Increased integration with intelligent and data-driven algorithms.	1. Enhancement of data-efficient learning approaches.2. Improvement of model transparency and interpretability.3. Development of stronger safety and reliability mechanisms.

Fuzzy logic

In the framework of AIDs, the foundational mechanism of FL relies on codifying long-term clinical experience in diabetes management into a comprehensive expert knowledge database. The specific conditions, thresholds, and operational decisions involved in medical treatment are mapped using fuzzy sets, and these logical rules are preserved within the expert system. By utilizing mathematical fuzzy reasoning, the controller derives appropriate parameter configurations for the insulin infusion dosage. The FL algorithm eliminates the prerequisite of constructing intricate physiological models, which grants it greater structural compatibility and simplifies its integration into embedded hardware systems.

Nonetheless, following its deployment into hardware, the FL algorithm lacks the operational flexibility to modify its parameters dynamically. Given the highly pronounced individual variations across diabetic patient cohorts, the control efficacy of the FL algorithm degrades significantly when the underlying physiological parameters undergo substantial changes, occasionally yielding worse control outcomes than a standard PID system. Furthermore, the selection of fuzzy rules and membership functions lacks rigorous theoretical or analytical guidance, meaning their derivation and realization rely entirely on heuristic strategies.

Table 7 presents the primary advantages and disadvantages of both control frameworks and concurrently identifies prospective lines of future research.

Table 7. Comparison between FL and DRL.

Aspect	Fuzzy Logic (FL)	Deep Reinforcement Learning (DRL)
Advantages	1. Incorporates expert clinical knowledge and experience into the control strategy.2. Operates without requiring complex mathematical models of the physiological system.	1. Exhibits strong adaptive capabilities.2. Reduces reliance on predefined mathematical models.3. Responds effectively to complex, dynamic, and uncertain environments.
Disadvantages	1. Limited flexibility when faced with changing conditions.2. Lacks a rigorous theoretical framework	1. Requires substantial amounts of training data.2. Decision-making processes are often

	for systematic design and analysis.3. Adaptability is generally restricted compared with learning-based approaches.	difficult to interpret.3. Safety concerns may arise when the system encounters scenarios not represented in the training data.
Future Research Directions	1. Development of advanced methods for optimizing fuzzy rules and membership functions.2. Greater integration with complementary control and decision-support technologies.3. Expansion of hybrid approaches combining fuzzy systems with artificial intelligence techniques.	1. Improvement of data-efficient learning methods.2. Enhancement of interpretability and transparency.3. Strengthening of safety, robustness, and reliability mechanisms.

Conclusion

The volume of relevant research published from 2020 to the present exceeds the cumulative sum of all publications from prior years, demonstrating that the unique opportunities and barriers in this domain have captured the attention of machine learning and medical research communities globally. In this review, we have presented a comprehensive evaluation of the deployment of DRL within AIDs. First, a foundational overview of DRL was established. Second, the methodological frameworks identified in recent literature were structured into two primary categories of DRL techniques based on optimization mechanics: value-based and policy-based algorithms. A detailed theoretical walkthrough of classic value-based DRL models, including DQN and its structural variants, was provided. Several widespread policy-based frameworks, encompassing both stochastic and deterministic paradigms, were also detailed. Third, we executed an exhaustive survey, cross-comparison, and structural analysis of the three foundational components—namely state space, action configuration, and reward function engineering—utilized by DRL applications in AIDs. Fourth, core operational challenges, potential mitigation pathways, and future research opportunities were examined through the lens of integrating DRL into real-world AIDs. Fifth, several classical control algorithms deployed in AIDs were compared in detail, mapping their respective advantages and limitations to clarify their individual contributions and real-world efficacy. Finally, two critical bottlenecks—data availability and algorithmic optimization—along with their corresponding deployment opportunities were synthesized based on recent data and hardware resource trends in the literature.

While this manuscript underscores the strong capabilities of DRL-driven AIDs, specific industrial and clinical rationales explain the ongoing reliance on MPC. The strong appeal of utilizing MPC stems from the capacity to analytically verify and guarantee the stability of its underlying models, its well-defined optimization objectives, and the ability to dynamically adapt the cost function over time based on an individual's changing physiological status. Transitioning the use of DRL in AIDs forward requires successfully addressing these stability concerns while obtaining regulatory approval from agencies like the US Food and Drug Administration and parallel global bodies. Methodical clinical trials executed across controlled hospital settings, transitional environments, and unsupervised real-world living conditions remain imperative to conclusively demonstrate first the non-inferiority, and ultimately the clinical superiority, of DRL-based AIDs before this methodology can transition from academic theory into commercially accessible medical therapies for individuals living with diabetes.

Abbreviations

The following abbreviations are used in this manuscript:

AI	Artificial intelligence
AIDs	Automated insulin delivery system

BGC	Blood glucose concentration
CGM	Continuous glucose monitoring
CHO	Carbohydrate
DDPG	Deep deterministic policy gradient
DNN	Deep neural networks
Double DQN	Double deep Q-Network
DRL	Deep reinforcement learning
FL	Fuzzy logic
HbA1c	Hemoglobin A1c
IOB	Insulin on board
MBPO	Model-based policy optimization
MPC	Model predictive control
PID	Proportional integral derivative
PPO	Proximal policy optimization
SAC	Soft actor critic
SAP	Sensor augmented pump
TD3	Twin delayed deep deterministic policy gradient
TD3-BC	Twin delayed DDPG with behavioral cloning
T1D	Type 1 diabetes
T2D	Type 2 diabetes
TIR	Time in range
TRPO	Trust region policy optimization

Acknowledgements

None

Conflict of interest

None

Financial support

This work is supported by Noncommunicable Chronic Diseases-National Science and Technology Major Project (2024ZD0532000; 2024ZD0532003), Young Scientists Fund of the National Natural Science Foundation of China (62403114) and Liaoning Provincial Natural Science Foundation of China (2023-BSBA-128).

Ethics statement

None

Received: 05 Sep 2022

Revised: 29 Oct 2022

Accepted: 22 Nov 2022

Published online: 15 March 2023

Rights and permissions

Open Access The author(s) retain copyright. This article is licensed under the Creative Commons Attribution-NonCommercial-ShareAlike 4.0 International License. It may be shared and adapted for non-commercial purposes with appropriate attribution, an indication of changes, and distribution of adaptations under the same license. Third-party material may be subject to separate terms identified in its credit line. View the license at <https://creativecommons.org/licenses/by-nc-sa/4.0/>.

References

Association AD. Diagnosis and classification of diabetes mellitus. *Diabetes Care*. 2012;36(Suppl 1):S67-S74.

ElSayed NA, Aleppo G, Aroda VR, Bannuru RR, Brown FM, Bruemmer D, et al. 9. Pharmacologic approaches to glycemic treatment: Standards of care in diabetes—2023. *Diabetes Care*. 2023;46(Suppl 1):S140-S157.

Stanford Medicine. New research keeps diabetics safer during sleep [Internet]. Stanford Medicine; 2014 [cited 2024 Oct 23]. Available from: <http://scopeblog.stanford.edu/2014/05/08/new-research-keeps-diabetics-safer-during-sleep/>

Phillip M, Nimri R, Bergenstal RM, Barnard-Kelly K, Danne T, Hovorka R, et al. Consensus recommendations for the use of automated insulin delivery technologies in clinical practice. *Endocr Rev*. 2023;44(2):254-80.

Oliveira CP, Mitchell BD, Fan LD, Garey C, Liao BR, Bispham J, et al. Patient perspectives on the use of half-unit insulin pens by people with type 1 diabetes: a cross-sectional observational study. *Curr Med Res Opin*. 2021;37:45-51.

Kamrul-Hasan ABM, Hannan MA, Alam MS, Rahman MM, Asaduzzaman M, Mustari M, et al. Comparison of simplicity, convenience, safety, and cost-effectiveness between use of insulin pen devices and disposable plastic syringes by patients with type 2 diabetes mellitus: a cross-sectional study from Bangladesh. *BMC Endocr Disord*. 2023;23:37.

Machry RV, Cipriani GF, Pedrosa HU, Nunes RR, Pires TLS, Ferreira R, et al. Pens versus syringes to deliver insulin among elderly patients with type 2 diabetes: a randomized controlled clinical trial. *Diabetol Metab Syndr*. 2021;13:64.

ElSayed NA, Aleppo G, Aroda VR, Bannuru RR, Brown FM, Bruemmer D, et al. 7. Diabetes technology: Standards of care in diabetes—2023. *Diabetes Care*. 2023;46(Suppl 1):S111-S127.

Chinese Society of Endocrinology. Chinese insulin pump treatment guidelines. *Chin J Endocrinol Metab*. 2021;37:679-701.

Tönnies T, Brinks R, Isom S, Dabelea D, Divers J, Mayer-Davis EJ, et al. Projections of type 1 and type 2 diabetes burden in the US population aged <20 years through 2060: the SEARCH for Diabetes in Youth Study. *Diabetes Care*. 2023;46:313-20.

ElSayed NA, Aleppo G, Aroda VR, Bannuru RR, Brown FM, Bruemmer D, et al. 14. Children and adolescents: Standards of care in diabetes—2023. *Diabetes Care*. 2023;46(Suppl 1):S230-S253.

Wadwa RP, Reed ZW, Buckingham BA, DeBoer MD, Ekhlaspour L, Forlenza GP, et al. Trial of hybrid closed-loop control in young children with type 1 diabetes. *N Engl J Med*. 2023;388:991-1001.

- DiMeglio LA, Kanapka LG, DeSalvo DJ, Hilliard ME, Laffel LM, Tamborlane WV, et al. A randomized clinical trial assessing continuous glucose monitoring use with standardized education with or without a family behavioral intervention compared with fingerstick blood glucose monitoring in very young children with type 1 diabetes. *Diabetes Care*. 2021;44:464-72.
- Laffel LM, Kanapka LG, Beck RW, Bergamo K, Clements MA, Criego A, et al. Effect of continuous glucose monitoring on glycemic control in adolescents and young adults with type 1 diabetes: a randomized clinical trial. *JAMA*. 2020;323:2388-96.
- Yeh HC, Brown TT, Maruthur N, Ranasinghe P, Berger Z, Suh YD, et al. Comparative effectiveness and safety of methods of insulin delivery and glucose monitoring for diabetes mellitus: a systematic review and meta-analysis. *Ann Intern Med*. 2012;157:336-47.
- Forlenza GP, Lal RA. Current status and emerging options for automated insulin delivery systems. *Diabetes Technol Ther*. 2022;24:362-71.
- Lal RA, Ekhlaspour L, Hood K, Buckingham B. Realizing a closed-loop (artificial pancreas) system for the treatment of type 1 diabetes. *Endocr Rev*. 2019;40:1521-46.
- Boughton CK, Hovorka R. New closed-loop insulin systems. *Diabetologia*. 2021;64:1007-15.
- Karageorgiou V, Papaioannou TG, Bellos I, Alexandraki K, Tentolouris N, Stefanadis C, et al. Effectiveness of artificial pancreas in the non-adult population: a systematic review and network meta-analysis. *Metabolism*. 2019;90:20-30.
- Anderson SM, Buckingham BA, Breton MD, Robic JL, Barnett CL, Wakeman CA, et al. Hybrid closed-loop control is safe and effective for people with type 1 diabetes who are at moderate to high risk for hypoglycemia. *Diabetes Technol Ther*. 2019;21:356-63.
- Forlenza GP, Ekhlaspour L, Breton M, Maahs DM, Wadwa RP, DeBoer M, et al. Successful at-home use of the Tandem Control-IQ artificial pancreas system in young children during a randomized controlled trial. *Diabetes Technol Ther*. 2019;21:159-69.
- Forlenza GP, Pinhas-Hamiel O, Liljenquist DR, Shulman DI, Bailey TS, Bode BW, et al. Safety evaluation of the MiniMed 670G system in children 7-13 years of age with type 1 diabetes. *Diabetes Technol Ther*. 2019;21:11-9.
- Sherr JL, Cengiz E, Palerm CC, Clark B, Kurtz N, Roy A, et al. Reduced hypoglycemia and increased time in target using closed-loop insulin delivery during nights with or without antecedent afternoon exercise in type 1 diabetes. *Diabetes Care*. 2013;36:2909-14.
<https://doi.org/10.2337/dc13-0010>.
- Zaharieva DP, Messer LH, Paldus B, O'Neal DN, Maahs DM, Riddell MC. Glucose control during physical activity and exercise using closed-loop technology in adults and adolescents with type 1 diabetes. *Can J Diabetes*. 2020;44:740-9.
- Carlson AL, Sherr JL, Shulman DI, Garg SK, Pop-Busui R, Bode BW, et al. Safety and glycemic outcomes during the MiniMed advanced hybrid closed-loop system pivotal trial in adolescents and adults with type 1 diabetes. *Diabetes Technol Ther*. 2022;24:178-89.
- Bothe MK, Dickens L, Reichel K, Tellmann A, Ellger B, Westphal M, et al. The use of reinforcement learning algorithms to meet the challenges of an artificial pancreas. *Expert Rev Med Devices*. 2013;10:661-73.
<https://doi.org/10.1586/17434440.2013.827515>.
- Bhonsle S, Saxena S. A review on control-relevant glucose-insulin dynamics models and regulation strategies. *Proc Inst Mech Eng Part I J Syst Control Eng*. 2020;234:596-608.
<https://doi.org/10.1177/0959651819870328>.
- Kovatchev B. A century of diabetes technology: signals, models, and artificial pancreas control. *Trends Endocrinol Metab*. 2019;30:432-44.
- Quiroz G. The evolution of control algorithms in artificial pancreas: a historical perspective. *Annu Rev Control*. 2019;48:222-32.

Thomas A, Heinemann L. Algorithm for automated insulin delivery (AID): an overview. *Diabetologie*. 2022;18:862-74.

Fuchs J, Hovorka R. Closed-loop control in insulin pumps for type 1 diabetes mellitus: safety and efficacy. *Expert Rev Med Devices*. 2020;17:707-20.

Marchetti G, Barolo M, Jovanovic L, Zisser H, Seborg DE. An improved PID switching control strategy for type 1 diabetes. *IEEE Trans Biomed Eng*. 2008;55:857-65.
<https://doi.org/10.1109/TBME.2008.915665>.

MohammadRidha T, Ait-Ahmed M, Chaillous L, Krempf M, Guilhem I, Poirier JY, et al. Model free iPID control for glycemia regulation of type 1 diabetes. *IEEE Trans Biomed Eng*. 2018;65:199-206.

Al-Hussein ABA, Tahir FR, Viet-Thanh P. Fixed-time synergetic control for chaos suppression in endocrine glucose-insulin regulatory system. *Control Eng Pract*. 2021;108:104723.

Skogestad S. Simple analytic rules for model reduction and PID controller tuning. *Model Identif Control*. 2004;25:85-120.

Nath A, Dey R, Balas VE. Closed loop blood glucose regulation of type 1 diabetic patient using Takagi-Sugeno fuzzy logic control. In: *Soft Computing Applications*. Cham: Springer; 2018. p. 286-96.

Yadav J, Rani A, Singh V. Performance analysis of fuzzy-PID controller for blood glucose regulation in type 1 diabetic patients. *J Med Syst*. 2016;40:254.

Bondia J, Romero-Vivo S, Ricarte B, Diez JL. Insulin estimation and prediction: a review of the estimation and prediction of subcutaneous insulin pharmacokinetics in closed-loop glucose control. *IEEE Control Syst Mag*. 2018;38:47-66.

Oviedo S, Vehi J, Calm R, Armengol J. A review of personalized blood glucose prediction strategies for T1DM patients. *Int J Numer Methods Biomed Eng*. 2017;33:e2833.
<https://doi.org/10.1002/cnm.2833>.

Gondhalekar R, Dassau E, Doyle FJ. Velocity-weighting and velocity-penalty MPC of an artificial pancreas: improved safety and performance. *Automatica*. 2018;91:105-17.

Shi DW, Dassau E, Doyle FJ. Adaptive zone model predictive control of artificial pancreas based on glucose- and velocity-dependent control penalties. *IEEE Trans Biomed Eng*. 2019;66:1045-54.

Birjandi SZ, Sani SKH, Pariz N. Insulin infusion rate control in type 1 diabetes patients using information-theoretic model predictive control. *Biomed Signal Process Control*. 2022;76:103635.

Williams G, Drews P, Goldfain B, Rehg JM, Theodorou EA. Information-theoretic model predictive control: theory and applications to autonomous driving. *IEEE Trans Robot*. 2018;34:1603-22.
<https://doi.org/10.1109/TRO.2018.2865891>.

Li Y, Yu C, Shahidehpour M, Yang T, Zeng Z, Chai T. Deep reinforcement learning for smart grid operations: algorithms, applications, and prospects. *Proc IEEE*. 2023;111:1055-96.
<https://doi.org/10.1109/JPROC.2023.3303358>.

Silver D, Schrittwieser J, Simonyan K, Antonoglou I, Huang A, Guez A, et al. Mastering the game of Go without human knowledge. *Nature*. 2017;550:354-9.

Kiran BR, Sobh I, Talpaert V, Mannion P, Sallab AAA, Yogamani S, et al. Deep reinforcement learning for autonomous driving: a survey. *IEEE Trans Intell Transp Syst*. 2022;23:4909-26.

Nemati S, Ghassemi MM, Clifford GD. Optimal medication dosing from suboptimal clinical examples: a deep reinforcement learning approach. In: *Proceedings of the 38th Annual International Conference of the IEEE Engineering in Medicine and Biology Society*;

2016 Aug 16-20; Orlando, FL. Piscataway (NJ): IEEE; 2016. p. 2978-81.

Fox I, Lee J, Pop-Busui R, Wiens J. Deep reinforcement learning for closed-loop blood glucose control. *Proc Mach Learn Res*. 2020;126:508-36.

Tejedor M, Woldaregay AZ, Godtliebsen F. Reinforcement learning application in diabetes blood glucose control: a systematic review. *Artif Intell Med*. 2020;104:101836.
<https://doi.org/10.1016/j.artmed.2020.101836>.

Sharma R, Singh D, Gaur P, Joshi D. Intelligent automated drug administration and therapy: future of healthcare. *Drug Deliv Transl Res*. 2021;11:1878-902.

Toffanin C, Visentin R, Messori M, Palma FD, Magni L, Cobelli C. Toward a run-to-run adaptive artificial pancreas: in silico results. *IEEE Trans Biomed Eng*. 2018;65:479-88.

Yau KLA, Chong YW, Fan X, Wu C, Saleem Y, Lim PC. Reinforcement learning models and algorithms for diabetes management. *IEEE Access*. 2023;11:28391-415.

Denes-Fazakas L, Fazakas GD, Eigner G, Kovacs L, Szilagyi L. Review of reinforcement learning-based control algorithms in artificial pancreas systems for diabetes mellitus management. In: *Proceedings of the 18th International Symposium on Applied Computational Intelligence and Informatics (SACI)*; 2024; Timisoara, Romania. p. 565-71.

Shin J, Badgwell TA, Liu KH, Lee JH. Reinforcement learning: overview of recent progress and implications for process control. *Comput Chem Eng*. 2019;127:282-94.

Daskalaki E, Diem P, Mougiakakou SG. Model-free machine learning in biomedicine: feasibility study in type 1 diabetes. *PLoS One*. 2016;11:e0158722.
<https://doi.org/10.1371/journal.pone.0158722>.

Wang S, Gu W. An improved strategy for blood glucose control using multi-step deep reinforcement learning. In: *Proceedings of the 16th International Conference on Bioinformatics and Biomedical Technology (ICBBT)*; 2024. p. 196-203.

Ahmad S, Beneyto A, Zhu T, Contreras I, Georgiou P, Vehi J. An automatic deep reinforcement learning bolus calculator for automated insulin delivery systems. *Sci Rep*. 2024;14:15245.
<https://doi.org/10.1038/s41598-024-62912-4>.

Del Giorgio S, D'Antoni F, Piemonte V, Merone M. A new glycemic closed-loop control based on Dyna-Q for type 1 diabetes. *Biomed Signal Process Control*. 2023;81:104492.

Zhu T, Li K, Herrero P, Georgiou P. Basal glucose control in type 1 diabetes using deep reinforcement learning: an in silico validation. *IEEE J Biomed Health Inform*. 2021;25:1223-32.

Li T, Wang Z, Lu W, Zhang Q, Li D. Electronic health records based reinforcement learning for treatment optimizing. *Inf Syst*. 2022;104:101878.

Noaro G, Zhu T, Cappon G, Facchinetti A, Georgiou P. A personalized and adaptive insulin bolus calculator based on double deep q-learning to improve type 1 diabetes management. *IEEE J Biomed Health Inform*. 2023;27:2536-44.

Hettiarachchi C, Malagutti N, Nolan CJ, Suominen H, Daskalaki E. Non-linear continuous action spaces for reinforcement learning in type 1 diabetes. In: *AI 2022: Advances in Artificial Intelligence*. Cham: Springer; 2022. p. 557-70.

Lee S, Kim J, Park SW, Jin SM, Park SM. Toward a fully automated artificial pancreas system using a bioinspired reinforcement learning design: in silico validation. *IEEE J Biomed Health Inform*. 2021;25:536-46.

Lehel DF, Siket M, Szilágyi L, Eigner G, Kovács L. Investigation of reward functions for controlling blood glucose level using reinforcement learning. In: Proceedings of the 2023 IEEE 17th International Symposium on Applied Computational Intelligence and Informatics (SACI); 2023; Timisoara, Romania. Piscataway (NJ): IEEE; 2023. p. 387-92.

Viroonluecha P, Egea-Lopez E, Santa J. Evaluation of blood glucose level control in type 1 diabetic patients using deep reinforcement learning. *PLoS One*. 2022;17:e0274608.
<https://doi.org/10.1371/journal.pone.0274608>.

Hettiarachchi C, Malagutti N, Nolan CJ, Suominen H, Daskalaki E. G2P2C—a modular reinforcement learning algorithm for glucose control by glucose prediction and planning in type 1 diabetes. *Biomed Signal Process Control*. 2024;90:105839.
<https://doi.org/10.1016/j.bspc.2023.105839>.

El Fathi A, Pryor E, Breton MD. Attention networks for personalized mealtime insulin dosing in people with type 1 diabetes. *IFAC-PapersOnLine*. 2024;58:245-50.

Denes-Fazakas L, Szilagyi L, Kovacs L, De Gaetano A, Eigner G. Reinforcement learning: a paradigm shift in personalized blood glucose management for diabetes. *Biomedicines*. 2024;12:2143.

Nordhaug Myhre J, Tejedor M, Launonen IK, El Fathi A, Godtliobsen F. In-silico evaluation of glucose regulation using policy gradient reinforcement learning for patients with type 1 diabetes mellitus. *Appl Sci*. 2020;10:6350.

Di Felice F, Borri A, Di Benedetto MD. Deep reinforcement learning for closed-loop blood glucose control: two approaches. *IFAC-PapersOnLine*. 2022;55:115-20.
<https://doi.org/10.1016/j.ifacol.2023.01.058>.

Zhu T, Li K, Kuang L, Herrero P, Georgiou P. An insulin bolus advisor for type 1 diabetes using deep reinforcement learning. *Sensors*. 2020;20:5058.

Ellis Z. Application of reinforcement learning algorithm to minimize the dosage of insulin infusion [master's thesis]. Greenville (NC): East Carolina University; 2024. p. 59.

Raheb MA, Niazmand VR, Eqra N, Vatankhah R. Subcutaneous insulin administration by deep reinforcement learning for blood glucose level control of type 2 diabetic patients. *Comput Biol Med*. 2022;148:105860.

Lim MH, Lee WH, Jeon B, Kim S. A blood glucose control framework based on reinforcement learning with safety and interpretability: in silico validation. *IEEE Access*. 2021;9:105756-75.
<https://doi.org/10.1109/ACCESS.2021.3100007>.

Lv W, Wu T, Xiong L, Wu L, Zhou J, Tang Y, et al. Hybrid control policy for artificial pancreas via ensemble deep reinforcement learning. *IEEE Trans Biomed Eng*. 2025;72:309-23.
<https://doi.org/10.1109/TBME.2024.3451712>.

Yu X, Guan Y, Yan L, Li S, Fu X, Jiang J. ARLPE: a meta reinforcement learning framework for glucose regulation in type 1 diabetes. *Expert Syst Appl*. 2023;228:120156.

Zhu J, Zhang Y, Rao W, Zhao Q, Li J, Wang C. Reinforcement learning for diabetes blood glucose control with meal information. In: *Bioinformatics Research and Applications*. Cham: Springer; 2021. p. 80-91.

Jiang J, Shen R, Wang B, Guan Y. Blood glucose control via pre-trained counterfactual invertible neural networks. *arXiv*. 2024;arXiv:2405.17458.

Chlumsky-Hartmann M, Ayad A, Schmeink A. HypoTreat: reducing hypoglycemia in artificial pancreas simulation. In: Proceedings of the 2024 IEEE Conference on Evolving and Adaptive Intelligent Systems (EAIS); 2024; Madrid, Spain. Piscataway (NJ): IEEE; 2024. p. 56-62.

Jaloli M, Cescon M. Basal-bolus advisor for type 1 diabetes (T1D) patients using multi-agent reinforcement learning (RL) methodology. *Control Eng Pract*. 2024;142:105762.
<https://doi.org/10.1016/j.conengprac.2023.105762>.

Mackey A, Furey E. Artificial pancreas control for diabetes using TD3 deep reinforcement learning. In: *Proceedings of the 2022 33rd Irish Signals and Systems Conference (ISSC); 2022; Cork, Ireland*. p. 1-6.

Emerson H, Guy M, McConville R. Offline reinforcement learning for safer blood glucose control in people with type 1 diabetes. *J Biomed Inform*. 2023;142:104376.

Zhu T, Li K, Georgiou P. Offline deep reinforcement learning and off-policy evaluation for personalized basal insulin control in type 1 diabetes. *IEEE J Biomed Health Inform*. 2023;27:5087-98.
<https://doi.org/10.1109/JBHI.2023.3303367>.

Beolet T, Adenis A, Hunecker E, Louis M. End-to-end offline reinforcement learning for glycemia control. *Artif Intell Med*. 2024;154:102920.
<https://doi.org/10.1016/j.artmed.2024.102920>.

Schulman J, Levine S, Moritz P, Jordan M, Abbeel P. Trust region policy optimization. In: *Proceedings of the 32nd International Conference on Machine Learning*; 2015. p. 1889-97.

Wang G, Liu X, Ying Z, Yang G, Chen Z, Liu Z, et al. Optimized glycemic control of type 2 diabetes with reinforcement learning: a proof-of-concept trial. *Nat Med*. 2023;29:2633-42.
<https://doi.org/10.1038/s41591-023-02552-9>.

Wang Z, Xie Z, Tu E, Zhong A, Liu Y, Ding J, et al. Reinforcement learning-based insulin injection time and dosages optimization. In: *Proceedings of the 2021 International Joint Conference on Neural Networks (IJCNN); 2021 Jul 18-22; Shenzhen, China*. Piscataway (NJ): IEEE; 2021. p. 1-8.

Unsworth R, Avari P, Lett AM, Oliver N, Reddy M. Adaptive bolus calculators for people with type 1 diabetes: a systematic review. *Diabetes Obes Metab*. 2023;25:3103-13.
<https://doi.org/10.1111/dom.15204>.

Jafar A, Fathi AE, Haidar A. Long-term use of the hybrid artificial pancreas by adjusting carbohydrate ratios and programmed basal rate: a reinforcement learning approach. *Comput Methods Programs Biomed*. 2021;200:105936.

Magni L. Model predictive control of type 1 diabetes: an in silico trial. *J Diabetes Sci Technol*. 2007;1:804-12.

De Paula M, Acosta GG, Martínez EC. On-line policy learning and adaptation for real-time personalization of an artificial pancreas. *Expert Syst Appl*. 2015;42:2234-55.
<https://doi.org/10.1016/j.eswa.2014.10.038>.

De Paula M, Ávila LO, Martínez EC. Controlling blood glucose variability under uncertainty using reinforcement learning and Gaussian processes. *Appl Soft Comput*. 2015;35:310-32.

De Paula M, Martinez E. Probabilistic optimal control of blood glucose under uncertainty. In: *22nd European Symposium on Computer Aided Process Engineering*. Amsterdam: Elsevier; 2012. p. 1357-61.

Akbari Torkestani J, Ghanaat Pisheh E. A learning automata-based blood glucose regulation mechanism in type 2 diabetes. *Control Eng Pract*. 2014;26:151-9.

Daskalaki E, Diem P, Mougiakakou SG. An actor-critic based controller for glucose regulation in type 1 diabetes. *Comput Methods Programs Biomed*. 2013;109:116-25.
<https://doi.org/10.1016/j.cmpb.2012.03.002>.

Sun Q, Jankovic MV, Budzinski J, Moore B, Diem P, Stettler C, et al. A dual mode adaptive basal-bolus advisor based on reinforcement learning. *IEEE J Biomed Health Inform.* 2019;23:2633-41.
<https://doi.org/10.1109/JBHI.2018.2887067>.

Sun Q, Jankovic MV, Mougiakakou SG. Reinforcement learning-based adaptive insulin advisor for individuals with type 1 diabetes patients under multiple daily injections therapy. In: *Proceedings of the 41st Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)*; 2019; Berlin, Germany. Piscataway (NJ): IEEE; 2019. p. 3609-12.
<https://doi.org/10.1109/EMBC.2019.8857178>.

Sun Q, Jankovic MV, Stettler C, Mougiakakou S. Personalised adaptive basal-bolus algorithm using SMBG/CGM data. In: *Proceedings of the 11th International Conference on Advanced Technologies and Treatments for Diabetes*; 2018; Vienna, Austria.

Thananjeyan B, Balakrishna A, Nair S, Luo M, Srinivasan K, Hwang M, et al. Recovery RL: safe reinforcement learning with learned recovery zones. *IEEE Robot Autom Lett.* 2021;6:4915-22.

Yasini S, Naghibi-Sistani MB, Karimpour A. Agent-based simulation for blood glucose control in diabetic patients. *World Acad Sci Eng Technol.* 2009;33:672-9.

Myhre JN, Launonen IK, Wei S, Godtliebsen F. Controlling blood glucose levels in patients with type 1 diabetes using fitted Q-iterations and functional features. In: *Proceedings of the 2018 IEEE 28th International Workshop on Machine Learning for Signal Processing (MLSP)*; 2018; Aalborg, Denmark. p. 1-6.

Shifrin M, Siegelmann H. Near-optimal insulin treatment for diabetes patients: a machine learning approach. *Artif Intell Med.* 2020;107:101917.

Shu Y, Cao Z, Gao J, Wang J, Yu PS, Long M. Omni-training: bridging pre-training and meta-training for few-shot learning. *IEEE Trans Pattern Anal Mach Intell.* 2023;45:15275-91.

Reske A, Carius J, Ma Y, Farshidian F, Hutter M. Imitation learning from MPC for quadrupedal multi-gait control. In: *Proceedings of the 2021 IEEE International Conference on Robotics and Automation (ICRA)*; 2021; Xi'an, China. p. 5014-20.

Ahmad S, Beneyto A, Contreras I, Vehi J. Bolus insulin calculation without meal information: a reinforcement learning approach. *Artif Intell Med.* 2022;134:102436.
<https://doi.org/10.1016/j.artmed.2022.102436>.

Brunke L, Greeff M, Hall AW, Yuan Z, Zhou S, Panerati J, et al. Safe learning in robotics: from learning-based control to safe reinforcement learning. *Annu Rev Control Robot Auton Syst.* 2022;5:411-44.

Zhang Y, Liang X, Li D, Ge SS, Gao B, Chen H, et al. Adaptive safe reinforcement learning with full-state constraints and constrained adaptation for autonomous vehicles. *IEEE Trans Cybern.* 2024;54:1907-20.

Artman WJ, Nahum-Shani I, Wu T, McKay JR, Ertefaie A. Power analysis in a SMART design: sample size estimation for determining the best embedded dynamic treatment regime. *Biostatistics.* 2020;21(3):432-45.

Fu J, Norouzi M, Nachum O, Tucker G, Wang Z, Novikov A, et al. Benchmarks for deep off-policy evaluation. In: *International Conference on Learning Representations*; 2021.

Kovatchev BP, Breton M, Dalla Man C, Cobelli C. In silico preclinical trials: a proof of concept in closed-loop control of type 1 diabetes. *J Diabetes Sci Technol.* 2009;3:44-55.

Lehmann ED, Deutsch T. A physiological model of glucose-insulin interaction in type 1 diabetes mellitus. *J Biomed Eng.* 1992;14:235-42.
[https://doi.org/10.1016/0141-5425\(92\)90058-S](https://doi.org/10.1016/0141-5425(92)90058-S).

Bergman RN. Minimal model: perspective from 2005. *Horm Res.* 2005;64(Suppl 3):8-15.

Liu Z, Ji L, Jiang X, Zhao W, Liao X, Zhao T, et al. A deep reinforcement learning approach for type 2 diabetes mellitus treatment. In: *Proceedings of the 2020 IEEE International Conference on Healthcare Informatics (ICHI)*; 2020; Oldenburg, Germany. p. 1-9.

Lopez-Martinez D, Eschenfeldt P, Ostvar S, Ingram M, Hur C, Picard R. Deep reinforcement learning for optimal critical care pain management with morphine using dueling double-deep Q networks. In: *Proceedings of the 41st Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)*; 2019; Berlin, Germany. Piscataway (NJ): IEEE; 2019. p. 3960-3.

Weng WH, Gao M, He Z, Yan S, Szolovits P. Representation and reinforcement learning for personalized glycemic control in septic patients. *arXiv.* 2017;arXiv:1712.00654.

Fox I, Wiens J. Reinforcement learning for blood glucose control: challenges and opportunities. In: *Proceedings of the 2019 RL4RealLife Workshop*; 2019; Long Beach, CA.

Schaul T, Quan J, Antonoglou I, Silver D. Prioritized experience replay. *arXiv.* 2015;arXiv:1511.05952.

Hester T, Vecerik M, Pietquin O, Lanctot M, Schaul T, Piot B, et al. Deep Q-learning from demonstrations. In: *Proceedings of the AAAI Conference on Artificial Intelligence*; 2018; New Orleans, LA. p. 394-401.

Herrero P, El-Sharkawy M, Daniels J, Jugnee N, Uduku CN, Reddy M, et al. The bio-inspired artificial pancreas for type 1 diabetes control in the home: system architecture and preliminary results. *J Diabetes Sci Technol.* 2019;13:1017-25.

Deshpande S, Pinsker JE, Zavitsanou S, Shi D, Tompot RW, Church MM, et al. Design and clinical evaluation of the interoperable artificial pancreas system (iAPS) smartphone app: interoperable components with modular design for progressive artificial pancreas research and development. *Diabetes Technol Ther.* 2019;21:35-43.

Li K, Liu C, Zhu T, Herrero P, Georgiou P. GluNet: a deep learning framework for accurate glucose forecasting. *IEEE J Biomed Health Inform.* 2020;24:414-23.

Li K, Daniels J, Liu C, Herrero P, Georgiou P. Convolutional recurrent neural networks for glucose prediction. *IEEE J Biomed Health Inform.* 2020;24:603-13.

Cobelli C. Diabetes: models, signals and control. In: *Proceedings of the 13th IMEKO TC1-TC7 Joint Symposium: Without Measurement No Science, Without Science No Measurement*; 2010; London, UK. p. 238.

Copp D, Gondhalekar R, Hespanha J. Simultaneous model predictive control and moving horizon estimation for blood glucose regulation in type 1 diabetes. *Optim Control Appl Methods.* 2017;39:904-18.

Sun X. Study on dynamic control of blood glucose in type 1 diabetes mellitus based on GPC [master's thesis]. Boston (MA): Northeastern University; 2018.

Babar SA, Rana IA, Mughal IS, Khan SA. Terminal synergetic and state feedback linearization based controllers for artificial pancreas in type 1 diabetic patients. *IEEE Access.* 2021;9:28012-9.

Messori M, Incremona GP, Cobelli C, Magni L. Individualized model predictive control for the artificial pancreas: in silico evaluation of closed-loop glucose control. *IEEE Control Syst Mag.* 2018;38:86-104.

Turksoy K, Bayrak ES, Quinn L, Littlejohn E, Cinar A. Multivariable adaptive closed-loop control of an artificial pancreas without meal and activity announcement. *Diabetes Technol Ther.* 2013;15:386-400.

Turksoy K, Hajizadeh I, Hobbs N, Kilkus J, Littlejohn E, Samadi S, et al. Multivariable artificial pancreas for various exercise types and intensities. *Diabetes Technol Ther.* 2018;20:662-71.

Hajizadeh I, Rashid M, Samadi S, Sevil M, Hobbs N, Brandt R, et al. Adaptive personalized multivariable artificial pancreas using plasma insulin estimates. *J Process Control*. 2019;80:26-40.

Hajizadeh I, Rashid MM, Turksoy K, Samadi S, Feng J, Sevil M, et al. Incorporating unannounced meals and exercise in adaptive learning of personalized models for multivariable artificial pancreas systems. *J Diabetes Sci Technol*. 2018;12:953-66.

Sun X, Rashid M, Askari MR, Cinar A. Adaptive personalized prior-knowledge-informed model predictive control for type 1 diabetes. *Control Eng Pract*. 2023;131:105386.

Sun X, Rashid M, Hobbs N, Askari MR, Brandt R, Shahidehpour A, et al. Prior informed regularization of recursively updated latent-variables-based models with missing observations. *Control Eng Pract*. 2021;116:104933.

Castillo A, Villa-Tamayo MF, Pryor E, Garcia-Tirado JF, Colmegna P, Breton M. Deep neural network architectures for an embedded MPC implementation: application to an automated insulin delivery system. *IFAC-PapersOnLine*. 2023;56:11521-6.

Steil G, Rebrin K, Mastrototaro JJ. Metabolic modelling and the closed-loop insulin delivery problem. *Diabetes Res Clin Pract*. 2006;74(Suppl):S183-S186.