Optimizing Blood Glucose Control through Reward Shaping in Reinforcement Learning

Fatemeh Sarani Rad
Department of Computer Science
North Dakota State University
Fargo, USA
fatemeh.saranirad@ndsu.edu

Juan Li
Department of Computer Science
North Dakota State University
Fargo, USA
j.li@ndsu.edu

Abstract— **Achieving optimal blood glucose control is a complex challenge for individuals with diabetes, necessitating a delicate balance among insulin dosage, food consumption, physical activity, and stress management. This paper introduces an innovative approach utilizing reinforcement learning (RL) to develop personalized and effective strategies for blood glucose regulation. Specifically, we employ the state-of-the-art soft actor-critic (SAC) RL algorithm, which concurrently maximizes anticipated rewards and policy entropy. We devise an entropy-driven reward function to incentivize diverse action exploration while ensuring a secure and consistent blood glucose profile. This reward function considers both the policy's entropy and the deviation of the blood glucose level from the target range, thus optimizing blood glucose control and minimizing the risk of complications. Our methodology is applied, trained, and assessed using a sophisticated blood glucose dynamics simulator based on the UVA/Padova model. The results demonstrate that our proposed method, SAC with entropy-based reward shaping (SAC+RS), outperforms a comparative approach, SAC with Magni's risk-based reward function (SAC+MRS), in terms of risk scores, glucose levels, insulin levels, and reward values.**

Keywords—diabetes management, blood glucose control, machine learning, reinforcement learning, reward shaping

I. INTRODUCTION

Diabetes has risen to the forefront as a significant and widespread health challenge in modern times. Its increasing prevalence has prompted heightened awareness and research efforts to better understand and manage this complex condition. Among the many critical facets of diabetes management, achieving effective blood glucose control stands out as a pivotal objective [1]. Maintaining optimal blood glucose levels is paramount due to its direct impact on overall health and well-being. Proper glucose regulation not only mitigates immediate health risks, such as hypoglycemia or hyperglycemia, but also plays a substantial role in preventing long-term complications [2]. The intricacies of diabetes management are further underscored by the intricate interplay of various factors, including dietary choices, insulin dosing, physical activity, stress management, and individual responses to treatment.

In the face of these multifaceted considerations, the significance of accurate blood glucose control cannot be overstated. It is the linchpin that connects various aspects of diabetes care and significantly contributes to the quality of life for individuals living with diabetes. As a result, innovative approaches to enhancing blood glucose regulation hold immense promise for not only improving day-to-day management but also for positively influencing the long-term health outcomes of those affected by diabetes. However, mastering the intricate task of blood glucose regulation poses a formidable challenge due to the intricate interplay of various influential factors. The equilibrium required to strike the ideal balance in blood glucose levels involves a complex dance among elements like precise insulin dosing, mindful dietary choices, varying degrees of physical activity, and adept stress management. This intricate interplay underscores the multifaceted nature of diabetes management. In response to this challenge, a surge of dedicated research initiatives has emerged, driven by the collective goal of empowering individuals to navigate and control their glucose levels effectively. These endeavors span a broad spectrum of innovative approaches, each striving to address a distinct aspect of the complex glucose regulation puzzle. Predictive modeling initiatives [3, 4], for instance, seek to anticipate and forecast blood glucose trends based on historical data, enabling proactive interventions and informed decision-making. These models leverage advanced algorithms and machine learning techniques to extrapolate future glucose levels, thereby providing individuals with actionable insights to fine-tune their diabetes management strategies. Closed-loop control systems [5, 6], another pioneering avenue of research, embody the concept of real-time automated glucose regulation. These systems utilize continuous glucose monitoring technology to feed data to an automated insulin delivery system, dynamically adjusting insulin dosages to maintain optimal blood glucose levels. This technological advancement promises to relieve individuals from constant vigilance while ensuring stable glucose control. Further enriching this landscape are personalized interventions [7], which recognize and respond to the inherent variability among individuals. Tailored approaches acknowledge that each person's response to insulin, food, activity, and stress is unique. By customizing treatment plans based on an individual's specific physiological characteristics and lifestyle choices, personalized interventions enhance the precision and efficacy of blood glucose management.

Despite these advancements, certain challenges persist. For instance, achieving stable glucose levels across diverse physiological contexts remains elusive. Managing the trade-off between hyperglycemia and hypoglycemia episodes, ensuring patient comfort, and optimizing insulin use further compound

This work was supported by the National Science Foundation (NSF) with award numbers: 1722913 and 2218046.

the complexity. Moreover, the inherent variability in individual responses to treatments calls for tailored solutions that adapt to individual needs.

In response to these challenges, this paper introduces an effective approach to blood glucose control by harnessing the capabilities of reinforcement learning (RL), specifically the soft actor-critic (SAC) algorithm [8]. SAC, a cutting-edge RL algorithm, offers distinct advantages such as off-policy learning, model independence, real-time adaptation, and robust exploration-exploitation balance. This choice is rooted in the belief that RL, particularly SAC, can serve as a potent tool for the dynamic and personalized glucose regulation required in diabetes management [9]. Central to our approach is the novel formulation of a reward function based on reward shaping, an innovative technique that modifies the reward structure to guide the agent toward improved policies. This strategic entropy-based shaping of rewards adds a layer of finesse to our method, enhancing the fine-tuning of blood glucose control policies.

Our contribution introduces a multifaceted reward function that addresses a spectrum of goals encompassing safety, comfort, and efficiency. The reward function incorporates three distinct terms: an exploration term based on policy entropy, an exploitation term quantifying glucose level quality, and an efficiency term associated with insulin infusion rate. These terms align with diverse objectives, shaping an effective framework for comprehensive glucose control. We emphasize that the coefficients governing these terms are systematically determined through a rigorous grid search process, optimizing performance against a range of critical metrics. These metrics include risk evaluation through the Clarke Error Grid Analysis (CEGA) and Magni's Risk Analysis (MRA), average glucose levels, insulin usage efficiency, and the percentage of time spent within specific glucose level ranges.

II. Related Work

The endeavor to enhance blood glucose control for individuals with diabetes has spurred a diverse range of research efforts, encompassing various methodologies and technological advancements. In this section, we provide an overview of key studies and initiatives that have contributed to the field, highlighting their distinct approaches and contributions.

A. Predictive Modeling for Glucose Regulation

Predictive modeling has emerged as a prominent avenue to anticipate and manage blood glucose levels. Researchers have leveraged machine learning algorithms, statistical methods, and physiological models to develop predictive models capable of forecasting glucose trends. Striving to empower individuals with timely information, these models enable proactive interventions and informed decision-making. Notable contributions include the work by Zaidi et al. [9], which introduces BG-Predict, a novel deep learning model designed to forecast blood glucose levels ahead in multiple time steps. The proposed tool aids Type-1 diabetes patients in administering insulin and managing food intake for optimal BG control. The model's effectiveness is demonstrated through quantitative and qualitative evaluation on real-world data from 97 patients. Another study [10] introduces a personalized glucose prediction model, utilizing deep learning, to aid medical staff in managing Type-2 diabetes patients in hospitals. The model employs recurrent neural networks (RNNs), specifically testing simple RNN, gated recurrent unit (GRU), and long-short term memory (LSTM) architectures for optimal performance.

B. Closed-Loop Control Systems

Advances in technology have paved the way for closed-loop control systems, which offer real-time automated glucose regulation. These systems integrate continuous glucose monitoring devices with automated insulin delivery mechanisms to maintain glucose levels within target ranges. The research conducted by O'Grady et al. [11] showcases the potential of closed-loop systems in achieving stable glucose control while minimizing the burden on individuals. The research involves testing a fully automated portable system called the Medtronic Portable Glucose Control System (PGCS), utilizing a smartphone platform, subcutaneous glucose sensors, and an insulin pump. Weaver and Hirsch [12] tested Medtronic's 670G insulin pump with Guardian 3 sensor exemplifying progress, maintaining glucose levels near targets. Initial studies show improved HbA1c, safety, and reduced risk of ketoacidosis or hypoglycemia. Yet, challenges remain in replicating natural islet function for fully automated, multi-hormonal blood glucose control.

C. Reinforcement Learning for Glucose Control

More recently, reinforcement learning (RL) has emerged as a promising paradigm for blood glucose regulation. RL algorithms, such as the soft actor-critic (SAC) algorithm employed in this paper, hold the potential to learn effective glucose control policies through interactions with the environment. Numerous studies have proposed RL-based algorithms for controlling blood glucose, often incorporating model predictive control (MPC) [7, 13, 14] as a component. MPC predicts future states and optimizes a cost function, accommodating system uncertainties and constraints. However, MPC's reliance on accurate models poses challenges. Tejedor et al. [15] and Fox et al. [6] utilized MPC within RL-based actor-critic methods. Tejedor et al. employed fixed parameters for MPC, while Fox et al. introduced an adaptive model using Gaussian processes. Both approaches improved glucose level maintenance, reducing hypoglycemia and personalizing control. Yet, both studies focused solely on blood glucose levels as objectives, overlooking patient comfort, safety, energy usage, insulin consumption, and preferences. Furthermore, their evaluations relied on simulated data, potentially differing from real-world complexities of blood glucose control.

III. Methodology

Our proposed system revolves around the constructing an insulin treatment framework by training a Soft Actor-Critic (SAC) Reinforcement Learning (RL) agent on simulated data from different patients with diabetes. The objective is to establish a closed-loop insulin delivery system capable of dynamically adjusting insulin dosages based on real-time glucose readings, effectively optimizing diabetes management.

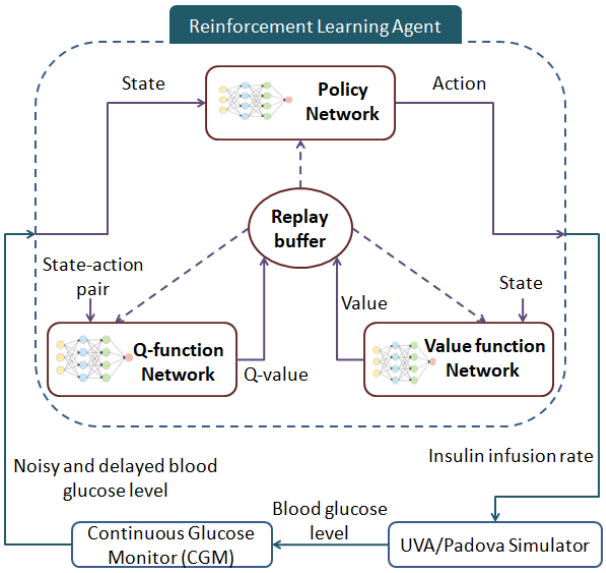

Fig.1 Framework of the proposed methodology process

A. Overview

Glucose control in closed-loop systems can be seen as a problem of making decisions under uncertainty, which can be formalized by a mathematical model called a partially observable Markov decision process (POMDP). A POMDP consists of a finite set of states, actions, observations, and functions that describe the relationships between them. At each time step, the agent faces a situation that represents the current state of the environment and chooses an action that affects the environment and produces a reward that indicates how good the action was. Then, the agent moves to a new state with a certain probability that depends on the previous state and action. However, the agent cannot directly observe the new state but only gets a clue that is related to the state with some probability [16]. In the context of glucose control, the state and action are defined by the blood glucose level of the patient and by the amount of insulin that is delivered at that time. The uncertainty of the state comes from the noise in the devices that measure the blood glucose level and from the influence of past data, such as the carbohydrates that were eaten, the insulin that was injected, and the blood glucose levels that were recorded. Fig.1 illustrates the conceptual architecture of our methodology process. The RL agents are responsible for dynamically regulating insulin dosages in response to real-time glucose readings, thereby automating and optimizing the treatment process.

The architecture consists of three main components; the first is a Reinforcement learning agent that learns a policy to control the insulin infusion rate based on the glucose sensor readings and the patient's preferences. The agent uses a soft actor-critic (SAC) algorithm, an actor-critic method that maximizes both the expected reward and the entropy of the policy. Our reward function is defined as a combination of the entropy of the policy, the squared discrepancy between the blood glucose level and the target value, and the insulin infusion rate. The reinforcement learning agent consists of three neural networks: a policy network, a Q-function network, and a value function network. The policy network is a stochastic actor that outputs a Gaussian distribution over actions given the current state. The Q-function network is a critic that estimates the Q-value of a state-action pair. The value function network is another critic that estimates the value of a state. The reinforcement learning agent updates its networks using gradient descent and experience replay. It samples transitions from a replay buffer and computes the target values for the Q-function and the value function. The second component of the architecture is the UVA/Padova Simulator which models the glucose-insulin dynamics of a person with type 1 diabetes. It takes the insulin infusion rate as input and outputs the blood glucose level. The last component of the architecture is Continues Glucose Monitor (CGM)/Glucose sensor, which measures the blood glucose level from the UVA/Padova simulator and adds some noise and delay to simulate real-world sensor errors.

B. Reinforcement Learning with Soft Actor-Critic (SAC)

RL provides a computational framework for learning optimal decisions in uncertain environments. RL is well-suited in blood glucose management due to its capability to handle complex, high-dimensional state spaces and stochastic dynamics. SAC, our chosen RL algorithm, bridges stochastic policy optimization and Deep Deterministic Policy Gradient (DDPG)-style methods. It emphasizes maximizing both expected return and policy entropy, promoting exploration and preventing premature convergence. SAC consists of three core components: an actor, a critic, and an entropy temperature. The actor employs a stochastic policy, the critic estimates state-action value functions, and the entropy temperature controls exploration-exploitation balance. We update the actor and critic through well-defined objectives, optimizing policy and value estimation.

The SAC RL model represents an advanced version of the actor-critic algorithm, designed to learn a stochastic policy that maximizes cumulative rewards. The policy, a function of the current state, is learned through an iterative actor-critic process, where the actor formulates the policy, and the critic evaluates state-value functions. A key feature of the SAC RL model is its maximum entropy formulation, which promotes exploration and prevents premature convergence to suboptimal policies. The SAC RL model's objective function involves maximizing both expected reward and policy entropy, while the critic network aims to minimize the mean squared error between predicted and actual state-value functions [8]. Integrating the SAC RL model into the metabolic model entails training the RL agent using real-time simulated data. Subsequently, the trained model is employed to automate insulin delivery within the simulator. The RL agent's learning process involves adjusting bolus insulin dosages based on real-time glucose readings, meal intake, and past agent actions. This adaptation adheres to an optimal policy π, designed to maximize the objective function $J(\pi)$. Our proposed MDI therapy framework combines cutting-edge reinforcement learning with a robust metabolic model, offering a promising avenue for enhancing blood glucose control and diabetes management.

C. Reward Function and Reward Shaping

To foster effective blood glucose management, we introduce a meticulously crafted reward function coupled with a reward shaping technique. This section delineates the design

of our reward function, underscored by the application of reward shaping to holistically address varied objectives in blood glucose control.

1) Designing a Reward Function for Blood Glucose Control

We harness reinforcement learning's potential by constructing a reward function tailored to the intricacies of blood glucose regulation. Our approach integrates reward shaping, a potent technique that fine-tunes the reward function to steer the RL agent toward optimal policies [17]. By amalgamating prior knowledge, expediting learning, and enhancing performance, reward shaping offers an adaptable framework. This integration is performed with prudence to circumvent potential disruptions to optimal policy attainment.

2) Proposing a Multi-Objective Reward Function for Blood Glucose Control

We propose a versatile reward function calibrated to fulfill multiple objectives, including the preservation of safe and comfortable blood glucose levels, prevention of hypoglycemic episodes, and judicious utilization of energy and insulin resources. Our novel reward function is mathematically described as:

$$r_s(g,u,\pi) = \alpha H(\pi) - \beta \frac{(g-125)^2}{100} - \gamma \frac{u}{10} \quad (1)$$

Where:

- $r_s(g,u,\pi)$ denotes the reward shaping component catering to blood glucose level (g), insulin infusion rate (u), and policy (π).
- H(π) signifies the entropy of policy π, capturing its randomness or uncertainty.
- α, β, and γ are positive coefficients that control the trade-off between exploration and exploitation.

The squared discrepancy between blood glucose level (g) and the target value of 125 mg/dL is encapsulated by the term $\beta \frac{(g-125)^2}{100}$, promoting meticulous glucose control and ameliorating health outcomes. This squared discrepancy term is bolstered by the coefficient β. The insulin infusion rate (u), a pivotal factor in efficiency and hypoglycemia risk, is embedded within the term $\gamma \frac{u}{10}$, fostering prudent insulin management. γ regulates the weight assigned to the insulin infusion rate term.

D. Parameter Tuning and Performance Metrics

Employing an iterative grid search method, we ascertain suitable values for coefficients α, β, and γ. Rigorous evaluation within our simulated environment involves diverse performance metrics:

- Risk: The Clarke Error Grid Analysis (CEGA) [18] calculated the average risk score, gauging the clinical acceptability of glucose predictions. CEGA is a method to assess the clinical accuracy and significance of glucose predictions or measurements compared to a reference value.
- MagniRisk: The average risk score computed using Magni's Risk Analysis (MRA) [19], quantifying hypoglycemic and hyperglycemic risks.
- Glucose: Average blood glucose level (mg/dL).
- Insulin: Average insulin infusion rate (U/h).
- Euglycemic: Percentage of time within the euglycemic range (70-180 mg/dL).
- Hypoglycemic: Percentage of time below the hypoglycemic threshold (70 mg/dL).

We adopt the iterative grid search method to fine-tune the parameters of a model or algorithm by exhaustively exploring a predefined range of values for each parameter. It involves generating a grid or matrix of different parameter combinations and evaluating the performance of the model for each combination. This process is iterative, meaning that it involves repeated cycles of adjusting the parameters, evaluating the model's performance, and refining the parameter values based on the evaluation results. This method is effective for systematically exploring the parameter space of a model and finding the best set of parameters that optimize its performance. It helps avoid manual guesswork in parameter tuning and provides a data-driven approach to finding optimal values. However, it can be computationally expensive, especially when dealing with many parameters or large parameter space. In such cases, we employed Bayesian optimization [20] to improve its efficiency. By harnessing reward shaping and a meticulously engineered reward function, our methodology bridges the gap between theoretical underpinnings and practical outcomes, offering an innovative trajectory toward refined blood glucose control in diabetes management.

IV. Evaluation

We have conducted comprehensive experiments to assess the effectiveness of our proposed approach. In this section, we present and discuss the results of our experimental evaluations.

A. Patient Data Simulation

Acquiring real-world data for evaluating medical interventions, especially in intricate and sensitive domains like diabetes management, can be a formidable challenge due to various ethical, logistical, and safety considerations. Therefore, we turn to validated and widely accepted simulation models, such as the FDA-approved UVA/Padova simulator [21] . These simulators replicate the physiological processes and dynamics of the human body, allowing us to create controlled and repeatable experimental environments. Using such simulators, we can simulate various scenarios, manipulate various parameters, and generate realistic data that closely approximates real patient responses without compromising patient privacy or safety.

To generate synthetic data for 30 virtual patients, we employed the open-source version of the UVA/Padova simulator [22, 23]. The simulator is a validated tool to create realistic and individualized data for blood glucose dynamics, insulin delivery, and carbohydrate intake. The 30 patients were divided into three groups: children, adolescents, and adults. Each group had 10 patients with different characteristics. These features are summarized in Table . We used 10 days of data for each patient, which included blood glucose measurements taken every five minutes by a continuous glucose monitor (CGM) and insulin dosages delivered every five minutes by an insulin pump.

TABLE I SIMULATED PATIENTS' FEATURES

Person	Age	Total Daily Insulin Dose (TDI)-Mean (STD)
child (#001-#010)	7-12	22.84 (±8.09)
adolescent (#001-#010)	14-19	40.611 (±13.23)
adult (#001-#010)	26-68	54.469 (±11.19)

B. Model Setup

In this paper, we developed a patient-specific model for each simulated individual using deep reinforcement learning (RL). To assess the effectiveness of deep RL for blood glucose control, we trained and tested the models with different random seeds on 30 different simulated individuals. We trained each model for 300 epochs, using a batch size of 256 and an epoch length of 10 days. The architecture of model networks consisted of two GRU layers with 128 hidden units each, followed by a fully connected layer that produced the action output. In this paper, we used a discount factor of 0.99 for RL. Additionally, we employed a learning rate of 3e-4 for the policy, Q-function, and value function networks. Moreover, we incorporated domain knowledge into RL by using a reward-shaping technique. For model selection, we utilized 10 days of validation data to choose the best epoch for each model based on its performance. To avoid overfitting, we employed the model parameters from the best epoch to evaluate the model on 10 days of test data.

C. Results and Discussions

In this paper, we propose a novel method for blood glucose control using soft actor-critic (SAC) with entropy-based reward shaping (SAC+RS). We compare our method with another approach, SAC, with Magni's risk-based reward function (SAC+MRS) [6]and show that our method can achieve better performance regarding risk scores, glucose levels, insulin levels, and reward values. We evaluate our method on different person categories, including children, adolescents, and adults, and demonstrate its adaptability and robustness to the dynamics of blood glucose regulation. Our method is based on the idea of maximizing the entropy of the policy, which encourages exploration and diversity of actions. We design a reward function that incorporates the policy's entropy and the deviation of the blood glucose level from the target range. We show that this reward function can effectively shape the policy to achieve optimal blood glucose control and reduce the risk of complications. We argue that this reward function has some limitations, such as being sensitive to the choice of parameters and ignoring the uncertainty of the policy. We conducted experiments on simulated patients with mean values of each metric based on 10 simulation runs for each method. We used a validated model of blood glucose dynamics to evaluate the performance of each method for each person category. We use various metrics, such as risk scores, glucose levels, insulin levels, and the occurrence of euglycemic, hypoglycemic, and hyperglycemic states.

The average risk score is inversely proportional to the clinical accuracy and significance of the glucose predictions or measurements [18]. The results shown in Table II indicate that our proposed method (SAC+RS) achieved a lower average risk (3.45) than the other two methods. Our method can reduce the risk scores by more than 2 points compared to SAC and SAC+MRS. This implies that our method can decrease the likelihood of developing long-term complications such as cardiovascular disease, kidney failure, nerve damage, and blindness.

TABLE II PERFORMANCE COMPARISON OF DIFFERENT METHODS FOR RISK SCORES

Method	Risk ↓
SAC	5.67
SAC+MRS	4.96
SAC+RS	**3.45**

We evaluated the performance of each method using three criteria: blood glucose level, risk of hypoglycemia or hyperglycemia, and time spent in the euglycemic range (70-180 mg/dL). The euglycemic range is the optimal range of blood glucose that minimizes the complications of diabetes. Our result demonstrated that SAC+RS outperformed SAC in blood glucose control, as it had lower blood glucose levels (124.96) and risk scores (3.4) than SAC. Moreover, for insulin infusion rate, SAC+MRS was higher than SAC, but SAC+RS was the same as SAC (0.0029). This indicates that our reward-shaping method achieved better blood glucose control with less risk and more time in the euglycemic range while using the same amount of insulin as SAC. The results are presented in Table III.

TABLE III PERFORMANCE COMPARISON OF DIFFERENT METHODS FOR GLUCOSE AND INSULIN LEVEL AND STATE OCCURRENCE

Metric \ Method	SAC	SAC+MRS	SAC+RS
Glucose	128.45	124.96	124.96
Insulin	0.0029	0.0050	0.0029
Euglycemic↑	0.82	0.87	0.87
Hypoglycemic ↓	0.04	0.03	0.03
Hyperglycemic ↓	0.14	0.10	0.10

Based on the results, our method can increase the reward values by more than 20 points compared to SAC+MRS. This means that our method can generate more diverse and exploratory actions to cope with the uncertainty and variability of blood glucose dynamics. Fig.2 shows the reward values obtained by each reward function for each person category. As can be seen from results, our method outperforms the other methods regarding risk scores, glucose levels, insulin levels, and reward values. This indicates that our method can achieve better blood glucose control and reduce the risk of complications for diabetic patients.

Across different person categories, including children, adolescents, and adults, our method exhibits adaptability and robustness, effectively addressing the varying physiological characteristics and preferences of diabetic patients. Compared to SAC and SAC+MRS, our approach significantly reduces risk scores by more than 2 points, lowering the probability of long-term complications. Additionally, it achieves better blood glucose control and more time spent in the euglycemic range with an equal amount of insulin infusion rate. The entropy-driven reward function enables our method to generate more diverse and exploratory actions, enhancing its ability to cope with the uncertainty and variability of blood glucose dynamics.

Overall, our novel approach using SAC+RS presents a promising solution for personalized and effective blood glucose regulation, providing insights and implications for improved diabetes management and patient outcomes.

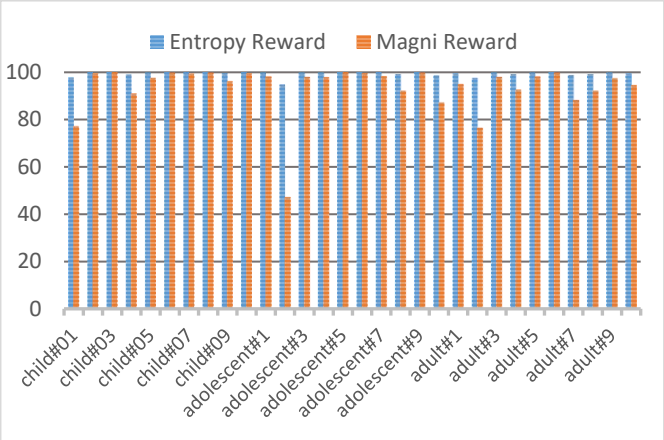

Fig.2 Average of Reward for all patient in different Epochs

V. Conclusions

In conclusion, this study introduces an effective approach to blood glucose control by leveraging the power of reinforcement learning, particularly the SAC algorithm, along with a novel reward function based on entropy-driven reward shaping. Our extensive evaluation highlights the effectiveness of this method in achieving key goals within diabetes management, underscoring its capacity to reshape glucose regulation and elevate patient health. However, it's important to acknowledge certain limitations. The proposed approach relies heavily on simulated data from validated models, which may not fully capture the complexities of real-world patient scenarios. Moreover, the coefficients in the reward function require careful tuning, and their generalizability to diverse patient populations warrants further investigation.

In the realm of future work, efforts should be directed towards the application and validation of the proposed approach using real patient data, potentially obtained through collaborations with medical institutions. Additionally, refining the reward function's coefficients through advanced optimization techniques could enhance the method's adaptability and robustness across different patient profiles. Further exploration of personalized and adaptive approaches within the reinforcement learning framework holds promise for optimizing blood glucose control tailored to individual patient needs..

References

[1] A. D. Association, "6. Glycemic targets: standards of medical care in diabetes—2021," *Diabetes Care,* vol. 44, no. Supplement_1, pp. S73-S84, 2021.

[2] A. D. Association, "Standards of Medical Care in Diabetes," *Diabetes Care,* vol. 28, no. suppl_1, pp. s4-s36, 2005, doi: 10.2337/diacare.28.suppl_1.S4.

[3] A. K. Rahimi *et al.*, "Machine learning models for diabetes management in acute care using electronic medical records: A systematic review," *International Journal of Medical Informatics,* vol. 162, p. 104758, 2022.

[4] H. Lai, H. Huang, K. Keshavjee, A. Guergachi, and X. Gao, "Predictive models for diabetes mellitus using machine learning techniques," *BMC endocrine disorders,* vol. 19, pp. 1-9, 2019.

[5] M. K. Bothe *et al.*, "The use of reinforcement learning algorithms to meet the challenges of an artificial pancreas," *Expert review of medical devices,* vol. 10, no. 5, pp. 661-673, 2013.

[6] I. Fox, J. Lee, R. Pop-Busui, and J. Wiens, "Deep reinforcement learning for closed-loop blood glucose control," in *Machine Learning for Healthcare Conference*, 2020: PMLR, pp. 508-536.

[7] B. A. Buckingham *et al.*, "Safety and feasibility of the OmniPod hybrid closed-loop system in adult, adolescent, and pediatric patients with type 1 diabetes using a personalized model predictive control algorithm," *Diabetes technology & therapeutics,* vol. 20, no. 4, pp. 257-262, 2018.

[8] T. Haarnoja, A. Zhou, P. Abbeel, and S. Levine, "Soft actor-critic: Off-policy maximum entropy deep reinforcement learning with a stochastic actor," in *International conference on machine learning*, 2018: PMLR, pp. 1861-1870.

[9] S. M. A. Zaidi, V. Chandola, M. Ibrahim, B. Romanski, L. D. Mastrandrea, and T. Singh, "Multi-step ahead predictive model for blood glucose concentrations of type-1 diabetic patients," *Scientific Reports,* vol. 11, no. 1, p. 24332, 2021/12/21 2021, doi: 10.1038/s41598-021-03341-5.

[10] D. Y. Kim *et al.*, "Developing an Individual Glucose Prediction Model Using Recurrent Neural Network," (in eng), *Sensors (Basel),* vol. 20, no. 22, Nov 12 2020, doi: 10.3390/s20226460.

[11] M. J. O'Grady *et al.*, "The Use of an Automated, Portable Glucose Control System for Overnight Glucose Control in Adolescents and Young Adults With Type 1 Diabetes," *Diabetes Care,* vol. 35, no. 11, pp. 2182-2187, 2012, doi: 10.2337/dc12-0761.

[12] "The Hybrid Closed-Loop System: Evolution and Practical Applications," *Diabetes Technology & Therapeutics,* vol. 20, no. S2, pp. S2-16-S2-23, 2018, doi: 10.1089/dia.2018.0091.

[13] T. Yamagata *et al.*, "Model-based reinforcement learning for type 1diabetes blood glucose control," *arXiv preprint arXiv:2010.06266,* 2020.

[14] L. Magni *et al.*, "Model predictive control of type 1 diabetes: an in silico trial," ed: SAGE Publications, 2007.

[15] M. Tejedor, A. Z. Woldaregay, and F. Godtliebsen, "Reinforcement learning application in diabetes blood glucose control: A systematic review," *Artificial intelligence in medicine,* vol. 104, p. 101836, 2020.

[16] R. D. Smallwood and E. J. Sondik, "The optimal control of partially observable Markov processes over a finite horizon," *Operations research,* vol. 21, no. 5, pp. 1071-1088, 1973.

[17] A. Y. Ng, D. Harada, and S. Russell, "Policy invariance under reward transformations: Theory and application to reward shaping," in *Icml*, 1999, vol. 99: Citeseer, pp. 278-287.

[18] M. A. Atkinson, G. S. Eisenbarth, and A. W. Michels, "Type 1 diabetes," *The Lancet,* vol. 383, no. 9911, pp. 69-82, 2014.

[19] C. f. D. C. a. Prevention, "National Diabetes Statistics Report, 2022: Estimates of Diabetes and Its Burden in the United States," *US department of health and human services. Atlanta, GA: Centers for Disease Control and Prevention,* 2020.

[20] P. I. Frazier, "A tutorial on Bayesian optimization," *arXiv preprint arXiv:1807.02811,* 2018.

[21] B. P. Kovatchev, M. Breton, C. D. Man, and C. Cobelli, "In silico preclinical trials: a proof of concept in closed-loop control of type 1 diabetes," (in eng), *J Diabetes Sci Technol,* vol. 3, no. 1, pp. 44-55, Jan 2009, doi: 10.1177/193229680900300106.

[22] J. Xie. "Simglucose v0.2.1 (2018) [Online]. Available: https://github.com/jxx123/simglucose . Accessed on: May-20-2023." (accessed.

[23] C. D. Man, F. Micheletto, D. Lv, M. Breton, B. Kovatchev, and C. Cobelli, "The UVA/PADOVA Type 1 Diabetes Simulator: New Features," (in eng), *J Diabetes Sci Technol,* vol. 8, no. 1, pp. 26-34, Jan 2014, doi: 10.1177/1932296813514502.