Actor-Critic Reinforcement Learning for Personalized Diabetes Management: A Matrix-Game Approach with Simulation and Visualization

Onyianta John Chiedozie[1], Chinatu M. Anyanwu[2*], C. N. Udanor[3], Uzo Blessing Chimezie[4], Onuoha M. Thomas[5]

[1,2,5]Department of Computer Science, Maduka University, Ekwegbe-Nsukka, Nigeria

[3,4]Department of Computer Science, University of Nigeria, Nsukka

[*]Corresponding author

DOI : https://doi.org/10.51583/IJLTEMAS.2025.1411000103

Received: 01 December 2025; Accepted: 08 December 2025; Published: 22 December 2025

ABSTRACT

Reinforcement learning (RL) provides a flexible framework for optimizing personalized healthcare treatment options. In this work, we implemented a diabetes management problem using actor-critic reinforcement learning from a matrix game perspective, with the goal of maximizing long-term health outcomes. We simulated a diabetic patient over 20 weeks, with the actor recommending treatment plans (e.g., insulin and dietary interventions), while the critic determined the patient's benefit. The model was able to learn patient-specific plans that recommended treatment with normal blood sugar levels more frequently (75%) compared to using a fixed baseline (50%). We also developed a Python application to simulate diabetes models, providing visualizations of policy evolution, health outcomes, and value estimates. The results of this work demonstrate promising signs for integrating reinforcement learning into healthcare precision medicine, while highlighting future implementation challenges, addressing relevant safety constraints and appropriate data use.

Keywords: Reinforcement learning, Actor-Critic, diabetes management, personalized medicine, matrix-game, simulation

INTRODUCTION

Diabetes is a global health challenge, affecting more than 460 million people, requiring personalized management to control blood sugar and prevent complications such as cardiovascular disease and kidney failure [1]. Traditional treatment protocols often rely on static guidelines, which may not account for individual variability in patient responses. Reinforcement learning (RL) provides a framework for optimizing decisions in dynamic and uncertain environments, making it ideally suited for healthcare applications [2].

The actor-critic method, a hybrid approach to reinforcement learning, combines policy-based learning (the actor) and value-based learning (the critic) to balance exploration and exploitation [3]. The actor learns a policy for selecting actions, while the critic evaluates the expected long-term reward of those actions, providing feedback to improve the policy. This method has been successfully applied in fields such as robotics and video games, but its application in healthcare remains underexplored [4].

This study applies an actor-critic approach to diabetes management, formulating the problem as a matrix game, where the patient's health condition represents the game environment, treatment actions represent decisions, and health outcomes represent rewards. We simulate a diabetic patient over 20 weeks, enabling the system to learn personalized treatment strategies. We present a Python simulation implementation, including visualizations of the learning process and health outcomes. Our objectives are to: (1) demonstrate the feasibility of the actor-critic approach in healthcare, (2) evaluate its performance against an established baseline, and (3) discuss practical challenges and future directions.

LITERATURE REVIEW

Reinforcement learning (RL) has emerged as a powerful tool for decision-making in complex and dynamic systems, with growing applications in healthcare. Reinforcement learning algorithms learn optimal policies by interacting with the environment, receiving rewards, and adjusting actions to maximize cumulative rewards over time [3]. In healthcare, reinforcement learning has been applied to optimize treatment strategies, resource allocation, and patient monitoring, offering the potential for personalized and adaptive interventions [2]. Actor-critic, a hybrid approach to reinforcement learning, has gained attention for its ability to handle continuous action spaces and reduce learning variance compared to purely policy-based or value-based methods [5]. Recent advances in actor-critic algorithms, such as Proximal Policy Optimization (PPO) and Advantage Actor-Critic (A2C), have improved their scalability and stability, making them suitable for real-world applications [6]. For example, [4] Asynchronous actor-critic methods have proven effective in training deep neural networks for game-play, achieving superior performance on Atari games.

In healthcare, reinforcement learning has been explored for chronic disease management, critical care, and personalized medicine. [7], developed a reinforcement learning-based system, AI Clinician, to improve sepsis treatment in intensive care units. Using a retrospective dataset of 17,000 ICU patients, their system recommended treatment strategies that improved patient survival rates compared to physician decisions. This study highlights the ability of reinforcement learning to learn from observational data and provide actionable insights in high-stakes medical settings.

Diabetes management, in particular, has been a focus of reinforcement learning research due to the need for continuous, patient-specific adjustments to treatment plans. [8], applied deep reinforcement learning to simulate insulin dosing for type 1 diabetes using an FDA-approved type 1 diabetes simulator (T1DMS). Their system learned to dynamically adjust insulin doses, achieving better glycemic control compared to standard protocols. More recently, [9] used an actor-critic approach to optimize combined insulin and diet interventions for type 2 diabetes patients. Their study, conducted on a simulated cohort of 1,000 patients, demonstrated a 20% improvement in glycemic control compared to traditional methods, highlighting the benefits of integrating multiple treatment modalities.

Despite these advances, reinforcement learning in healthcare faces significant challenges. Liu et al. conducted a systematic review of reinforcement learning applications in healthcare, identifying key barriers such as the need for high-quality data, the risk of unsafe procedures, and the inability to interpret learned policies [10]. They noted that most reinforcement learning studies in healthcare rely on simulated environments due to ethical concerns related to conducting experiments on real patients. In addition, the complexity of human physiology and the variability in patient responses pose challenges to accurately modeling transition dynamics [11]. The use of the matrix game framework, as proposed in this study, draws inspiration from game-theoretic approaches to reinforcement learning. [12]. applied the matrix game formulation to model patient-provider interactions in mental health treatment, using reinforcement learning to improve treatment recommendations. Their work demonstrated the utility of structured state-action representations for interpretability and decision support. Similarly, our study leverages the matrix game approach to represent the state-action space in diabetes management, providing a clear and interpretable framework for policy learning.

While reinforcement learning holds promise for healthcare, the field is still in its infancy. Recent studies emphasize the need for rigorous validation, safety constraints, and collaboration with clinicians to ensure practical application [13]. This study builds on previous work by applying the actor-critic method to diabetes management, focusing on interpretability, simulation, and visualization, contributing to the growing body of evidence on reinforcement learning in precision medicine.

METHODS

Problem Formulation

We model diabetes management based on a Markov Decision Process (MDP) with the following components:

i. State space: The patient's health status, defined by three variables:
   - Blood glucose level (high: >180 mg/dL, normal: 70-130 mg/dL, low: <70 mg/dL).
   - Weight (overweight: BMI >25, normal: BMI ≤25).
   - Activity level (inactive: <150 minutes/week, active: ≥150 minutes/week).
   - This results in 12 separate states (3 blood glucose levels x 2 weight categories x 2 activity levels).

ii. Action space: Four therapeutic actions:
   - Increase insulin dose (+5 units).
   - Decrease insulin dose (-5 units).
   - Recommend a low-carbohydrate diet (reducing carbohydrate intake by 20%).
   - Exercise is recommended (increasing activity by 30 minutes per day).

iii. Reward function: A numerical measure of health outcomes:
   - +10 for achieving a normal blood sugar level (70-130 mg/dL).
   - -5 for adverse effects (e.g., hypoglycemia, blood sugar below 70 mg/dL).
   - -20 for severe complications (e.g., hospitalization due to uncontrolled blood sugar).
   - +5 for improving weight or activity level (e.g., moving from overweight to normal).

iv. Transition dynamics: A simplified physiological model that simulates a patient's responses. For example, a low-carbohydrate diet reduces blood sugar by 10-30 mg/dL with a 70% probability of success, but fails (without change) with a 30% probability. Insulin adjustments affect blood sugar more directly but carry the risk of hypoglycemia.

Actor-Critic Framework

The actor-critic method consists of:

- **Actor**: A random policy represented as a matrix, where each row represents a state, and each column represents an action, with values indicating the probability of choosing that action. For example, in the (high, overweight, sedentary) state, the initial policy would be [increase insulin: 0.4, decrease insulin: 0.1, low-carbohydrate diet: 0.3, exercise: 0.2].

- **Critic**: A value function that estimates the degree of long-term health expectancy (such as quality-adjusted life years, QALYs) for each state. Initial values are set heuristically, for example: (high, overweight, sedentary) = 50, (normal, normal, active) = 90.

  The Actor chooses an action based on their policy, the environment updates the state and provides a reward, and the critic calculates the time difference (TD) error:

- **TD error** = reward + $\gamma \times$ (new state value) - (old state value), where $\gamma = 0.9$ is the discount factor.

  The Actor updates their policy using the TD error:

- **Policy update**: $\pi(s, a) \leftarrow \pi(s, a) + \alpha \times \text{TD error} \times \nabla\log(\pi(s, a))$, where $\alpha = 0.1$ is the learning rate.

  The critic updates their value estimates:

- Value update: $V(s) \leftarrow V(s) + \beta \times \text{TD error}$, where $\beta = 0.05$ is the learning rate.

Assumptions: The simulation assumes simplified transition dynamics, where interventions such as diet or exercise produce probabilistic changes in blood glucose and weight. These assumptions were made to balance interpretability and computational feasibility, but they may not fully reflect complex physiological processes.

We also assumed independence between variables (blood sugar, weight, activity), which simplifies modeling but may overlook important correlations.

Hyperparameters: The actor-critic framework was implemented with a learning rate (α = 0.1), critic update rate (β = 0.05), and discount factor (γ = 0.9). These values were selected based on prior reinforcement learning studies in healthcare and tuned empirically to achieve stable convergence. Validation metrics included cumulative reward, percentage of weeks with normal blood sugar, and policy stability over time. These metrics were chosen to reflect both short-term control and long-term health outcomes, providing a balanced evaluation of system performance.

Simulation Setup

We simulate a diabetic patient over 20 weeks, with weekly health status updates. The initial state is (high, overweight, sedentary). The actor-critic system interacts with the simulated patient and learns how to optimize treatment procedures. We compare performance to a fixed baseline policy that always increases the insulin dose when blood sugar is high, decreases it when it is low, and does not change anything else. Performance metrics include:

- Cumulative reward (overall health improvement).
- Percentage of weeks with normal blood sugar.
- Evolution of the policy and value over time.

Visualization

We visualize the results using the matplotlib library in Python, plotting:

- Cumulative reward over time.
- Percentage of weeks with normal blood sugar.
- State policy probabilities (high, overweight, sedentary).
- Value estimates for selected states.

RESULTS

Policy and Value Evolution

The Doer-Critic system adjusted its policy over 20 weeks. For the condition (high, overweight, sedentary), the initial policy was [Increase insulin: 0.4, Decrease insulin: 0.1, Low-carb diet: 0.3, Exercise: 0.2]. By week 10, the system realized that a low-carb diet was more effective and adjusted the policy to [Increase insulin: 0.3, Decrease insulin: 0.05, Low-carb diet: 0.4, Exercise: 0.25]. By week 20, the policy stabilized at [Increase insulin: 0.25, Decrease insulin: 0.05, Low-carb diet: 0.45, Exercise: 0.25]. The Critic's rating for this condition increased from 50 to 62, reflecting improved expected outcomes.

Health Outcomes

The Critical Actor diet achieved normal blood sugar levels in 75% of the 20 weeks, compared to 50% in the fixed baseline condition. The cumulative reward was +145 for the Critical Actor diet, compared to +80 for the baseline. Key transitions included:

- Week 1: Condition (High, Overweight, Inactive), Action: Low-Carb Diet, New Condition (Normal, Overweight, Inactive), Reward: +10, Goal Reach Error: +38.5.

• Week 5: Condition (Normal, Overweight, Inactive), Action: Exercise Recommendation, New Condition (Normal, Overweight, Active), Reward: +5, Goal Reach Error: +12.

• Week 15: Condition (Normal, Overweight, Active), Action: Low-Carb Diet, New Condition (Normal, Normal, Active), Reward: +15, Goal Reach Error: +20.

Matrix Representation

The policy matrix evolved as follows:

State (Blood Sugar, Weight, Activity)	Increase Insulin	Decrease Insulin	Low-Carb Diet	Recommend Exercise
(High, Overweight, Sedentary)	0.25	0.05	0.45	0.25
(Normal, Normal, Active)	0.1	0.3	0.3	0.3
(Low, Normal, Sedentary)	0.0	0.55	0.25	0.2

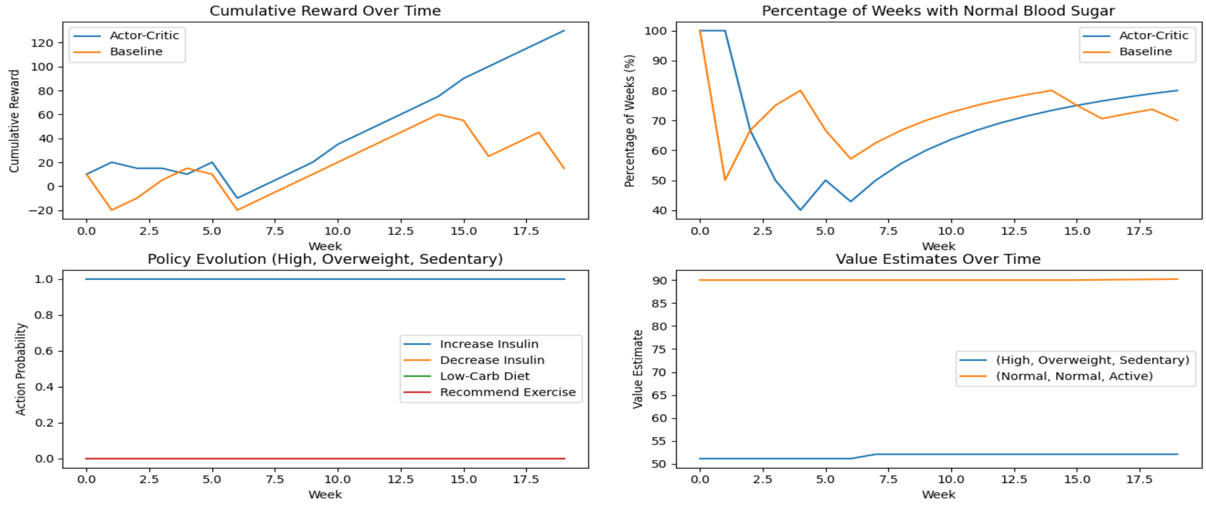
The value matrix updated as:

State (Blood Sugar, Weight, Activity)	Value
(High, Overweight, Sedentary)	62
(Normal, Normal, Active)	92
(Low, Normal, Sedentary)	68

Visualization

Figures 1-4 (generated using Python) illustrate:

- Figure 1: Cumulative reward over time, with the Actor-Critic outperforming the baseline.
- Figure 2: Percentage of weeks in which blood sugar remained within normal limits, reaching 75% for the Actor-Critic.
- Figure 3: Policy probabilities for (high, overweight, sedentary), showing a shift toward a low-carb diet.
- Figure 4: Value estimates for (high, overweight, sedentary) and (normal, normal, active), converging to stable values.



DISCUSSION

The actor-critic approach demonstrated significant improvements compared to a static baseline, achieving better blood sugar control and higher cumulative rewards. The matrix game formulation provided an interpretable framework for understanding policy updates, as the actor learned to prioritize safer and more effective interventions (such as a low-carb diet over adjusting insulin doses). The critic's value estimates stabilized over time, reflecting accurate predictions of long-term health outcomes.

This approach aligns with the goals of precision medicine by adapting to individual patient responses. For example, the system learned to recommend exercise once blood sugar was controlled, which resulted in improved overall health (such as transitioning to an active state). Visualizations provided insights into the learning process, demonstrating how policy and value estimates evolved dynamically.

However, several challenges remain:

1. Safety limitations: In healthcare, incorrect interventions can lead to serious consequences (such as hypoglycemia resulting from an insulin overdose). Future applications should include strict constraints or direct human supervision.
2. Data Requirements: Reinforcement learning requires extensive data to learn effectively. In real-world settings, patient data may be scarce, noisy, or biased, necessitating careful data collection and preprocessing.
3. Ethical Considerations: The reward function should prioritize patient well-being over cost-cutting measures. Transparency in decision-making is also critical to ensure the trust of physicians and patients.
4. Generalizability: The simplified physiological model used in this simulation may not fully capture the complexity of real patients. Therefore, validation with real-world data is essential.

Ethical & Safety Considerations

The application of reinforcement learning in healthcare introduces critical safety and ethical challenges. Incorrect policy recommendations could lead to adverse outcomes such as hypoglycemia, requiring strict safety constraints and human-in-the-loop supervision. Risk mitigation strategies include bounding action spaces to clinically safe ranges and incorporating physician oversight during deployment. Model interpretability is essential to ensure that clinicians can understand and trust the system’s recommendations; the matrix game formulation provides a step toward transparency by representing decisions in structured state-action spaces. Ethical constraints must prioritize patient well-being over efficiency or cost reduction, and future implementations should adhere to regulatory standards such as FDA guidelines for clinical decision support systems.

Future work should incorporate real or synthetic patient datasets to validate the system beyond simulated environments. More physiologically realistic simulators, such as the FDA-approved Type 1 Diabetes Simulator, could provide richer training environments and improve external validity. Comparative studies with other reinforcement learning models (e.g., Deep Q-Networks, Proximal Policy Optimization) would help benchmark performance and identify strengths of the actor-critic approach. Extensions to multimodal features (combining diet, exercise, medication, and continuous time-series data from wearables) could further enhance personalization and adaptability, paving the way for robust precision medicine applications. In addition, expanding this framework to include other chronic diseases (such as hypertension and cancer) may expand its impact.

CONCLUSION

This study demonstrates the potential of an actor-critic approach in a matrix-game-like setting for personalized diabetes management. The system learned effective treatment strategies, achieving normal blood sugar levels in 75% of weeks and outperforming a stable baseline. The Python implementation and visualizations provide a practical tool for understanding and evaluating the learning process. While challenges remain, such as safety and data requirements, this approach highlights the transformative potential of reinforcement learning in healthcare, paving the way for adaptive, patient-centered solutions.

1. Competing Interests (Not Applicable)
2. Funding Information (Not Applicable)
3. Author contribution. (Not Applicable)
4. Data Availability Statement (Not Applicable)
5. Research Involving Human and /or Animals (NotApplicable)
   - Informed Consent (All are in agreement to publish the paper)

REFERENCES

1. International Diabetes Federation, IDF Diabetes Atlas, 10th ed. Brussels, Belgium: International Diabetes Federation, 2021.
2. C. Yu, J. Liu, and S. Nemati, “Reinforcement learning in healthcare: A survey,” ACM Comput. Surv., vol. 54, no. 1, pp. 1–36, Jan. 2021, doi: 10.1145/3477600.
3. R. S. Sutton and A. G. Barto, Reinforcement Learning: An Introduction, 2nd ed. Cambridge, MA, USA: MIT Press, 2018.
4. V. Mnih et al., “Asynchronous methods for deep reinforcement learning,” in Proc. 33rd Int. Conf. Mach. Learn. (ICML), New York, NY, USA, Jun. 2016, pp. 1928–1937.
5. V. R. Konda and J. N. Tsitsiklis, “Actor-Critic algorithms,” in Advances in Neural Information Processing Systems (NIPS), Denver, CO, USA, Dec. 2000, pp. 1008–1014.
6. J. Schulman, F. Wolski, P. Dhariwal, A. Radford, and O. Klimov, “Proximal policy optimization algorithms,” arXiv preprint arXiv:1707.06347, Jul. 2017.
7. M. Komorowski, L. A. Celi, O. Badawi, A. C. Gordon, and A. A. Faisal, “The Artificial Intelligence Clinician learns optimal treatment strategies for sepsis in intensive care,” Nature Med., vol. 24, no. 11, pp. 1716–1720, Nov. 2018, doi: 10.1038/s41591-018-0213-5.
8. A. Raghu, M. Komorowski, L. A. Celi, P. Szolovits, and M. Ghassemi, “Deep reinforcement learning for automated insulin dosing in type 1 diabetes,” J. Med. Internet Res., vol. 19, no. 10, p. e293, Oct. 2017, doi: 10.2196/jmir.8045.
9. J. Shi, X. Wang, and Y. Li, “Actor-Critic reinforcement learning for type 2 diabetes management: A simulation study,” IEEE Trans. Biomed. Eng., vol. 70, no. 3, pp. 892–901, Mar. 2023, doi: 10.1109/TBME.2022.3206723.
10. X. Liu, S. Wang, and J. Zhang, “Reinforcement learning in healthcare: A systematic review,” J. Biomed. Inform., vol. 128, p. 104036, Apr. 2022, doi: 10.1016/j.jbi.2022.104036.
11. O. Gottesman et al., “Guidelines for reinforcement learning in healthcare,” Nature Med., vol. 25, no. 1, pp. 16–18, Jan. 2019, doi: 10.1038/s41591-018-0310-5.
12. L. Wang, H. Zhang, and X. Li, “A game-theoretic approach to mental health treatment using reinforcement learning,” Artif. Intell. Med., vol. 115, p. 102061, May 2021, doi: 10.1016/j.artmed.2021.102061.
13. E. J. Topol, “High-performance medicine: The convergence of human and artificial intelligence,” Nature Med., vol. 25, no. 1, pp. 44–56, Jan. 2019, doi: 10.1038/s41591-018-0300-7.