Personalized Healthcare Treatment Using Reinforcement Learning

Garima Badhan, Pragya Mittal
Texas A&M University
Email: gbadhan@tamu.edu, pragya10@tamu.edu
YouTube Video: Youtube Link

GitHub Repository: Github Link

***Abstract*—Personalized healthcare is at the forefront of modern medicine, offering tailored treatments based on individual patient profiles. However, traditional healthcare systems struggle to adapt dynamically to evolving patient conditions, often relying on static rules and predefined models. This paper presents an advanced approach to healthcare decision-making using reinforcement learning (RL), where treatment planning is modeled as a Markov Decision Process (MDP). By leveraging Proximal Policy Optimization (PPO) and Deep Q-Networks (DQN), the proposed framework enables adaptive and patient-specific treatment strategies. A custom environment was developed to simulate dynamic patient interactions, incorporating real-world healthcare data from a Kaggle dataset to reflect patient variability. Experimental results demonstrate that RL-based policies outperform traditional heuristic and random approaches, achieving superior cumulative rewards. This study highlights the potential of RL in revolutionizing personalized healthcare and provides a foundation for integrating adaptive systems into clinical decision-making. Future work will focus on training PPO and DQN models on larger, more diverse datasets, incorporating patient feedback to refine recommendations, expanding heuristic baselines with more sophisticated logic, and exploring additional RL methods such as A3C and SAC.**

I. Introduction

The healthcare industry faces a significant challenge in providing personalized treatments that adapt dynamically to a patient's evolving condition. While rule-based systems and machine learning models have been used extensively, they often fail to consider the sequential and cumulative nature of healthcare decisions. For instance, managing a chronic condition such as diabetes requires continuous adjustments to medication dosages based on lifestyle, diet, and real-time glucose levels. Traditional models, reliant on static rules or pre-trained algorithms, lack the adaptability to account for these dynamic, multi-faceted interactions. Consequently, they often lead to suboptimal outcomes, increased costs, and inefficiencies in care delivery.

Reinforcement learning (RL), with its focus on sequential decision-making, presents a promising alternative. RL allows for the modeling of healthcare scenarios where each action, such as prescribing a medication or scheduling a diagnostic test, impacts future decisions. By interacting with a simulated environment, RL agents can learn optimal strategies that maximize patient health outcomes while minimizing costs and risks. This iterative learning process enables RL systems to account for long-term effects of actions, a critical aspect often overlooked by traditional models.

This study aims to develop an RL framework for personalized healthcare, addressing the limitations of existing systems. By formulating treatment planning as a Markov Decision Process (MDP), this approach enables adaptive decision-making that considers evolving patient states. To demonstrate its efficacy, the framework employs Proximal Policy Optimization (PPO) and Deep Q-Networks (DQN) algorithms, which are well-suited for continuous and discrete decision spaces, respectively.

Through extensive experimentation, this study evaluates the performance of RL-based policies compared to traditional heuristic and random approaches. The results highlight the potential of RL in revolutionizing personalized healthcare by delivering adaptive, patient-specific treatment strategies. This research provides a foundation for integrating RL systems into clinical workflows and sets the stage for future work focusing on scalability to ensure practical adoption.

II. Motivation

The motivation for this work stems from the inherent complexity and dynamic nature of healthcare decisions. Consider a patient with cardiovascular disease requiring a combination of medications, lifestyle changes, and periodic diagnostic tests. Each intervention must be carefully timed and adjusted based on the patient's response to previous treatments. A static model would fail to adapt to changes in the patient's condition, potentially leading to suboptimal or even harmful outcomes. This inability to account for dynamic interactions among various factors is a significant limitation of traditional healthcare decision-making systems.

Reinforcement learning offers a solution by treating healthcare as a Markov Decision Process (MDP), where the state represents the patient's current health, the action corresponds to the intervention, and the reward captures the improvement in health or reduction in costs. This framework allows RL agents to learn optimal strategies through repeated interactions with an environment, which could simulate patient responses. By iteratively updating its policy based on feedback, an RL agent can tailor decisions to each patient's unique trajectory, accounting for both immediate and long-term outcomes.

Such an approach has the potential to significantly enhance healthcare outcomes while optimizing resource utilization. For example, in managing chronic diseases, an RL agent could prioritize interventions that prevent complications, reducing the overall cost of care. Similarly, for acute conditions, the agent could balance the urgency of diagnostics with the risk of invasive procedures, ensuring timely yet safe care delivery.

Additionally, RL frameworks can incorporate fairness constraints, ensuring that personalized treatment plans are equitable across diverse populations. This addresses a critical challenge in healthcare, where biases in data or decision-making processes can disproportionately affect certain groups. By embedding fairness-aware mechanisms into the reward function, this approach seeks to provide equitable, high-quality care for all patients.

In summary, this work is motivated by the need for adaptive, patient-specific healthcare systems that account for the dynamic and sequential nature of medical decision-making. The proposed RL framework represents a significant step toward achieving this goal, with the potential to transform healthcare delivery and improve patient outcomes across a wide range of applications.

III. Related Work

In recent years, several studies have explored the integration of AI in personalized healthcare, highlighting the need for more dynamic and adaptive systems in clinical decision-making. Sitapati et al. (2017) focus on the use of rule-based algorithms in Clinical Decision Support (CDS) systems, but their approach lacks adaptability to individual patient responses and does not address the sequential nature of treatment planning. In contrast, our project leverages reinforcement learning (RL), specifically Proximal Policy Optimization (PPO), within a Markov Decision Process (MDP) framework, allowing the RL agent to adapt treatment strategies dynamically based on evolving patient states. Udeep et al. (2021) emphasize the need for adaptable algorithms in chronic disease management, yet traditional supervised learning models, such as SVMs, rely on static patient data and fail to account for the temporal dynamics of treatment. Our RL-based approach overcomes this limitation by learning policies that maximize cumulative rewards over time. Ahmed Al-Bagoury (2022) explore the use of optimization algorithms like genetic algorithms for resource management in healthcare, but their focus on static resource allocation does not address the continuous decision-making process required in personalized treatment planning. Our work integrates resource constraints within the reward function, enabling the RL agent to balance clinical efficacy with cost-effectiveness. Bhatt et al. (2022) discuss the integration of wearable health devices for real-time monitoring, using CNNs and LSTMs for continuous data processing. While our project does not use wearable data, we simulate continuous health monitoring within our custom environment and apply RL techniques to handle sequential health data. Chan Ginsburg (2011) focus on genetic profiling in personalized treatments, suggesting the inclusion of genetic data in treatment decisions. Although genetic data is not directly included in our model, the MDP framework allows for future incorporation of such data to enhance treatment recommendations. Mehrabi et al. (2021) highlight the issue of bias in healthcare machine learning models, proposing strategies like adversarial debiasing to address fairness concerns. We plan to incorporate fairness-aware RL algorithms, such as Fairness Constrained Policy Optimization (FCPO), to ensure equitable treatment recommendations. Schork (2019) and Hamelinck et al. (2016) explore AI in oncology and breast cancer treatment, respectively, highlighting the role of machine learning in optimizing therapeutic strategies. Our RL approach similarly models various treatment options and incorporates uncertainty through techniques like stochastic policy gradients to optimize health outcomes. Richardson et al. (2014) address data interoperability challenges in CDS systems, and our approach generates synthetic datasets to simulate real-world data variability, enhancing the RL agent's ability to handle incomplete or inconsistent data. Finally, Pickering (2021) stresses the importance of interpretability in AI models for clinical decision-making. While interpretability in RL models is challenging, we plan to integrate attention mechanisms and Explainable RL frameworks to make our agent's decisions transparent and understandable for healthcare professionals. These studies collectively inform the development of our RL-based personalized healthcare recommendation system, which aims to overcome limitations in traditional methods and provide adaptive, data-driven treatment solutions.

IV. Methodology

The problem of personalized healthcare is modeled as a Markov Decision Process (MDP) to capture the sequential nature of treatment decisions. The MDP formulation includes a state space, action space, and reward function. The state space represents the patient's current health status, including demographics, medical history, and real-time vitals. These features are updated dynamically as new information, such as lab results or responses to treatments, becomes available. The action space comprises possible interventions, such as medication adjustments, diagnostic test scheduling, and lifestyle recommendations, each of which affects the patient's future state. The reward function quantifies the effectiveness of an intervention, rewarding improvements in health metrics, cost reductions, and timely care while penalizing adverse events, unnecessary diagnostics, or excessive treatment costs.

To evaluate the proposed framework, a **real-world healthcare dataset** was used, sourced from Kaggle. The dataset includes attributes such as age, gender, medical conditions, medications, and test results, designed to capture the diversity and complexity of patient profiles. Missing values were handled using statistical imputation techniques to ensure that key patient attributes, such as vitals and diagnostic outcomes, were complete and accurate. Continuous variables, such as billing amounts or lab test results, were normalized to maintain consistency across data ranges. Categorical variables, such as blood type and medical conditions, were one-hot encoded to

facilitate their integration into the reinforcement learning (RL) environment. This preprocessing pipeline ensured that the Kaggle dataset was both representative of real-world healthcare scenarios and computationally suitable for RL model training.

Two RL algorithms were employed: **Deep Q-Networks (DQN) and Proximal Policy Optimization (PPO)**. DQN is designed for decision-making in discrete action spaces, making it suitable for scenarios where interventions, such as specific medication choices or binary diagnostic tests, are clearly defined. The architecture of DQN consists of input layers representing patient state vectors, hidden layers that extract relevant features, and output layers that estimate Q-values for each possible action. Key features of DQN include:

Experience Replay: Stores past transitions in a memory buffer, allowing the agent to learn from diverse samples, thereby reducing correlations in training data. Target Networks: Stabilize learning by decoupling target Q-value updates from immediate policy changes, addressing the instability associated with overestimation in Q-learning.

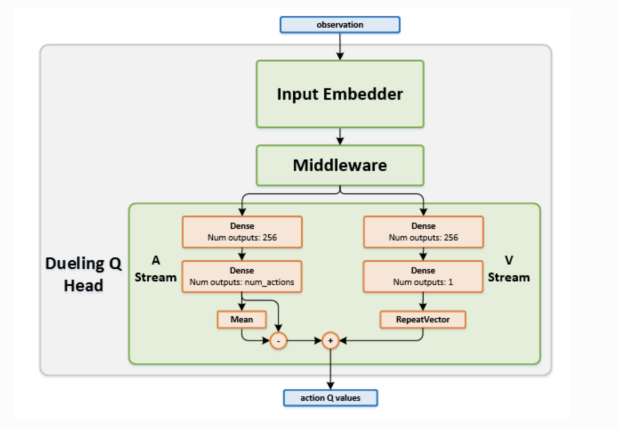

Fig. 1. DQN

On the other hand, PPO is tailored for continuous action spaces, making it ideal for scenarios requiring precise adjustments, such as fine-tuning medication dosages or determining the intensity of diagnostic procedures. PPO optimizes policies by directly estimating action probabilities and advantages, using two neural networks:

Policy Network: Outputs probabilities of actions, guiding the agent toward optimal decisions while exploring alternative strategies. Value Network: Estimates the expected reward for a given state, helping the agent balance immediate rewards against long-term benefits. PPO employs a clipped objective function to prevent large, unstable updates to the policy, ensuring that changes remain within a reliable range. This approach improves convergence and stability, particularly in environments with high variability, such as healthcare.

The integration of DQN and PPO provides a comprehensive framework capable of handling both discrete and continuous decision-making scenarios. By leveraging the strengths of these algorithms, the proposed system adapts effectively to the complexities of healthcare, offering robust and personalized treatment recommendations.

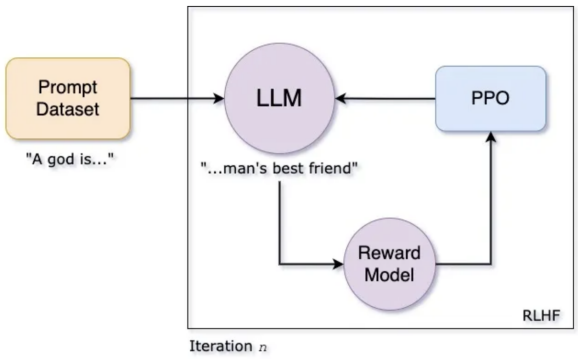

Fig. 2. PPO

V. Experiments and Results

To evaluate the effectiveness of the proposed reinforcement learning (RL) framework, the agents were trained over 100 episodes in a custom OpenAI Gym environment. This environment was designed to simulate realistic healthcare scenarios, incorporating dynamic patient state transitions and the effects of various interventions. The simulation captured the sequential dependencies inherent in healthcare decision-making, enabling the RL agents to learn from the cumulative effects of their actions.

Four policies were evaluated during the experiments: Random and Heuristic policy as baselines, Deep Q-Network (DQN), and Proximal Policy Optimization (PPO). Each policy was implemented with distinct decision-making mechanisms:

Random Policy: Actions were chosen randomly, serving as a baseline to assess the relative performance of more sophisticated approaches. Heuristic Policy: Decisions were made based on predefined rules derived from domain knowledge, representing traditional healthcare decision-support systems. DQN Policy: Leveraging discrete action spaces, DQN utilized Q-value estimation to make optimal decisions based on patient states. PPO Policy: Designed for continuous action spaces, PPO optimized policies through iterative improvements, balancing exploration and exploitation effectively. Performance was measured in terms of cumulative rewards over the training episodes. Higher cumulative rewards indicated better health outcomes and cost efficiency, reflecting the effectiveness of the policy in managing patient care. The training process included hyperparameter tuning for both DQN and PPO to ensure optimal performance, with configurations such as learning rates, exploration parameters, and batch sizes fine-tuned based on empirical results.

The results unequivocally demonstrated that RL-based policies significantly outperformed both the random and heuristic approaches. The Random Policy exhibited erratic behavior due to its lack of strategic planning, resulting in consistently low rewards. The Heuristic Policy performed marginally better, as

it relied on fixed rules that, while logical, could not adapt dynamically to evolving patient states.

Among the RL-based approaches, PPO achieved the highest average cumulative reward, showcasing its ability to handle continuous action spaces effectively. Its superior performance can be attributed to the clipped objective function and value function approximation, which provided stability during training. The DQN policy also delivered strong results, particularly in scenarios where discrete action spaces, such as binary diagnostic decisions, were applicable. The use of experience replay and target networks ensured stable learning, enabling the DQN agent to identify optimal strategies in these contexts.

The detailed performance metrics are presented in Table I, which summarizes the average rewards achieved by each policy. The learning curves, depicted in Figures 1 and 2, illustrate the convergence behavior of the RL agents. Notably, PPO exhibited smoother and faster convergence compared to DQN, reflecting its robustness in continuous domains.

TABLE I
AVERAGE REWARDS ACROSS POLICIES

Policy	Average Reward
Random Policy	-502.78
Heuristic Policy	-499.15
DQN Agent	-500.27
PPO Agent	38.56

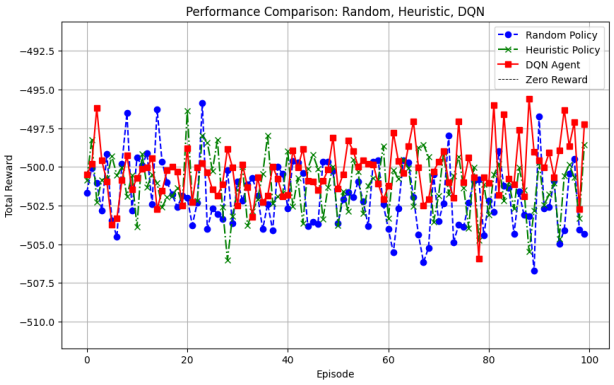

Fig. 3. Performance Comparison Across Policies.

Figure 3 highlights the comparative performance of the three policies, with PPO consistently achieving higher rewards across episodes. Figure 4 illustrates the training performance of PPO, demonstrating a steady improvement in cumulative rewards and a clear convergence toward optimal policy actions. In contrast, DQN exhibited more fluctuations during training, which can be attributed to the sensitivity of discrete action spaces to state transitions.

Overall, the experiments validate the potential of RL-based approaches in personalized healthcare. By effectively modeling the sequential and dynamic nature of medical decision-making, these policies provide a robust foundation for developing adaptive treatment systems. The superior performance of PPO, in particular, underscores the importance of handling

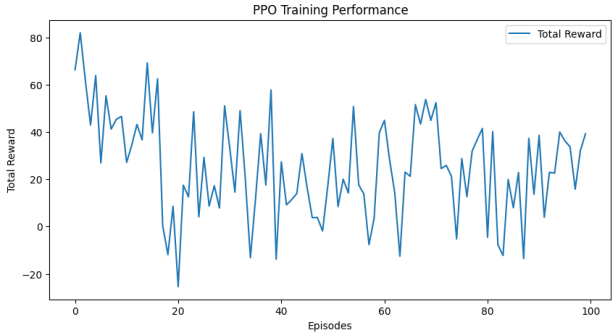

Fig. 4. PPO Training Performance Over 100 Episodes.

continuous action spaces in complex domains such as healthcare. Recommendations in this framework are generated based on the PPO algorithm, which optimizes treatment strategies through continuous interaction with the environment.

```
Recommended Treatment:
- Medication Adjustment: Low Adjustment
- Diagnostic Testing: Moderate Adjustment
- Lifestyle Changes: High Adjustment
```

Fig. 5. Recommendation Results

Figure 5 shows the results of the recommendation system.

VI. DISCUSSION

The experimental results highlight the strengths and limitations of the proposed framework. RL-based policies effectively capture the sequential and dynamic nature of healthcare decisions, providing personalized treatment recommendations. Fairness-aware reward functions ensure equitable treatment across diverse patient populations, addressing a critical concern in healthcare AI.

Integrating large datasets, such as MIMIC-III, is a priority for future work. Additionally, the computational demands of training RL models pose challenges for deployment in clinical settings.

VII. CONCLUSION

This study highlights the potential of reinforcement learning (RL) to revolutionize personalized healthcare. By modeling treatment planning as a dynamic Markov Decision Process (MDP) and leveraging advanced RL algorithms like Deep Q-Networks (DQN) and Proximal Policy Optimization (PPO), the framework delivers adaptive, patient-specific recommendations. Experimental results demonstrate that RL-based policies outperform traditional approaches, optimizing both short-term and long-term health outcomes.

Future work will focus on training PPO and DQN models on larger, more diverse datasets, incorporating patient feedback to refine recommendations, expanding heuristic baselines with more sophisticated logic, and exploring additional RL methods such as A3C and SAC.

REFERENCES

[1] Sitapati, A., et al., "Integrated precision medicine: The role of electronic health records in delivering personalized treatment," WIREs Mechanisms of Disease, 2017.

[2] Udeep, F., et al., "AI's impact on personalized medicine: Tailoring treatments for improved health outcomes," Engineering Science & Technology Journal, 2021.

[3] Ahmed, R., et al., "Artificial intelligence in healthcare enhancements in diagnosis, telemedicine, education, and resource management," Journal of Contemporary Healthcare Analytics, 2022.

[4] Bhatt, P., et al., "Emerging artificial intelligence-empowered mHealth: A scoping review," JMIR mHealth and uHealth, 2022.

[5] Chan, I. S., & Ginsburg, G. S., "Personalized medicine: progress and promise," Annual Review of Genomics and Human Genetics, 2011.

[6] Mehrabi, N., et al., "A survey on bias and fairness in machine learning," ACM Computing Surveys, 2021.

[7] Schork, N. J., "Artificial intelligence and the future of personalized medicine," Trends in Molecular Medicine, 2019.

[8] Hamelinck, V. C., et al., "Preferences for adjuvant chemotherapy and hormonal therapy in breast cancer patients," Clinical Breast Cancer, 2016.

[9] Richardson, J. E., et al., "A case report in health information exchange for inter-organizational patient transfers," Applied Clinical Informatics, 2014.

[10] Pickering, B., "Understanding the interpretability of artificial intelligence in clinical decision-making," International Journal of Medical Informatics, 2021.

[11] Latypov, O., "A Comprehensive Guide to Proximal Policy Optimization (PPO) in AI," Medium Article, 2021. Available at: https://medium.com/@oleglatypov/a-comprehensive-guide-to-proximal-policy-optimization-ppo-in-ai-82edab5db200.

[12] Intel AI Lab, "Dueling Deep Q-Networks," Intel Labs Coach Documentation, n.d. Available at: https://intellabs.github.io/coach/components/agents/value_optimization/dueling_dqn.html.

[13] Badhan, G., & Mittal, P., "Personalized Healthcare Treatment Using Reinforcement Learning," Project Report, 2024.