Supervised Learning-Guided Reinforcement Learning for Personalized Treatment: A Comparative Analysis of Multiple Reinforcement Learning Algorithms in Diabetes and Hypertension Management

Reyhan Zada Virgiwibowo (2206081723) and
Samuel Farrel Bagasputra (2206826614)

Faculty of Computer Science, Universitas Indonesia

May 24, 2025

Abstract

This study investigates the application of reinforcement learning (RL) for personalized treatment recommendation in managing diabetes and hypertension, with a particular focus on how supervised learning (SL) guidance can enhance treatment optimization. We evaluate three distinct RL algorithms—Proximal Policy Optimization (PPO), Deep Q-Network (DQN), and Advantage Actor-Critic (A2C)—in environments both with and without SL guidance. Our approach incorporates advanced supervised learning techniques including feature engineering, hyperparameter optimization, and sampling strategies to address class imbalance. Results demonstrate consistent improvements across all algorithms when SL guidance is incorporated, with time-in-target-range metrics showing substantial clinical benefits. The framework provides a comprehensive, data-driven solution to chronic disease management that leverages both supervised and reinforcement learning paradigms to personalize interventions based on individual patient characteristics and real-time clinical feedback.

Contents

1 Introduction

Type 2 diabetes and hypertension have become pressing global health issues due to their increasing prevalence and significant impact on morbidity, mortality, and healthcare costs. Managing these chronic diseases requires careful adjustment of treatment regimens, often complicated by patient-specific variability and multimorbidity. Traditional approaches relying on static clinical protocols frequently fall short in addressing the dynamic nature of these conditions. This project aims to develop and evaluate a hybrid machine learning approach that integrates supervised learning with reinforcement learning to optimize personalized treatment recommendations for patients suffering from type 2 diabetes and hypertension.

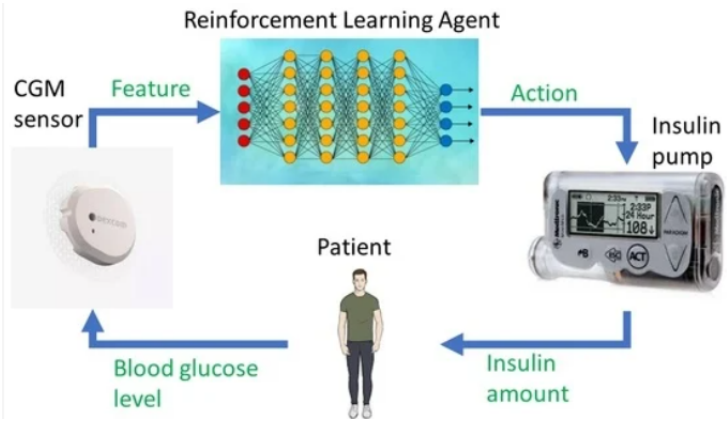

Figure 1: Schematic representation of a reinforcement learning agent managing personalized blood glucose control using continuous glucose monitoring (CGM) sensor data and insulin pump interventions.

Figure 1 depicts a typical RL-driven closed-loop system for diabetes management. In this framework, a continuous glucose monitoring (CGM) sensor continuously measures the patient's blood glucose levels, which are fed as input features to the reinforcement learning agent. The agent processes these data through a neural network to decide on optimal treatment actions, such as adjusting insulin delivery via an insulin pump. This dynamic feedback loop enables personalized, adaptive control of blood glucose by continuously responding to the patient's physiological state in real time. This concept and implementation are adapted from recent studies [1] that emphasize the

paradigm shift RL introduces to personalized blood glucose management.

In this work, we explore the impact of incorporating a supervised learning model, specifically a RandomForest Classifier with advanced optimization techniques, as an additional source of guidance in the RL environment. We comprehensively evaluate the benefits of this hybrid approach in improving training efficiency and treatment personalization. This dual-learning strategy represents a novel integration of two powerful machine learning paradigms aimed at advancing personalized medicine.

2 Related Works

Recent advancements in machine learning have demonstrated promising applications in chronic disease management. Reinforcement learning (RL), in particular, offers an adaptive framework capable of learning optimal treatment policies by interacting with patient data over time.

For example, Sun et al. [2] developed an RL-based treatment recommendation system for type 2 diabetes patients, utilizing real-world clinical datasets. Their approach surpassed traditional clinical guidelines by dynamically adjusting therapy to improve glycemic control, demonstrating the ability of RL to personalize treatments based on patient response.

Furthermore, Zheng et al. [3] applied RL to electronic health record (EHR) data encompassing patients with comorbid diabetes and hypertension. Their study highlighted RL's capacity to simultaneously optimize treatment across multiple interacting chronic conditions, addressing the complexities inherent in multimorbidity.

Additionally, recent research has explored hybrid methods combining supervised learning (SL) with RL to enhance learning efficiency. Incorporating SL models as auxiliary signals or guides can accelerate convergence and improve policy quality by providing prior knowledge during RL training [1].

3 Methodology

3.1 Reinforcement Learning

Reinforcement Learning (RL) is a computational approach where an agent learns to make decisions by interacting with an environment to maximize cumulative rewards. Unlike supervised learning, RL does not rely on labeled

input-output pairs; instead, the agent learns from the consequences of its actions through trial and error, guided by feedback in the form of rewards or penalties [4].

RL differs from traditional machine learning models in that it doesn't require labeled data for training. Instead, the agent learns through trial and error, exploring different actions and adjusting its behavior based on the rewards it receives. This makes RL particularly well-suited for dynamic environments like healthcare, where treatment strategies must continuously adapt to the changing conditions of patients.

RL algorithms are broadly categorized into three types: value-based methods, policy-based methods, and actor-critic methods. Value-based methods, like Q-learning, focus on estimating the value of different actions to inform decision-making. Policy-based methods, such as Proximal Policy Optimization (PPO), directly optimize the policy function, adjusting it based on the feedback received. Actor-critic methods combine the strengths of both approaches by using a value function to guide policy optimization.

In healthcare applications, RL's ability to adapt to the evolving needs of patients, such as adjusting treatment plans in real-time based on the patient's condition, makes it a powerful tool for managing chronic diseases like diabetes and hypertension.

3.2 Supervised Learning

Supervised Learning (SL) is another key machine learning paradigm, where the model is trained using labeled data. In supervised learning, the goal is to learn a mapping from input features to output labels. The model is provided with a dataset that includes both the inputs and their corresponding correct outputs (i.e., labels). The model's task is to learn from this data and make predictions or decisions based on new, unseen inputs.

There are two main types of supervised learning tasks: classification and regression. In classification tasks, the model predicts discrete labels (e.g., diagnosing whether a patient has diabetes based on medical features), while in regression tasks, the model predicts continuous values (e.g., predicting blood pressure levels based on various factors).

Supervised learning models can be linear, such as logistic regression, or more complex, such as decision trees and neural networks. One popular method in supervised learning is the Random Forest Classifier, which creates an ensemble of decision trees to make predictions. The power of Random

Forest lies in its ability to handle high-dimensional data, deal with missing values, and avoid overfitting, making it a strong candidate for use in medical applications, where the data is often complex and noisy.

When combined with reinforcement learning, supervised learning can provide a helpful signal to guide the RL agent, improving its ability to learn optimal policies. In such hybrid models, the supervised learning component can be used to provide prior knowledge or additional context, which accelerates the RL agent's learning process, improving both the efficiency and quality of its decisions

3.3 Data and Preprocessing

We utilized two key datasets for this study:

- **Diabetes Dataset**: This dataset contains clinical features such as *Pregnancies*, *Glucose*, *BloodPressure*, *SkinThickness*, *Insulin*, *BMI*, *DiabetesPedigreeFunction*, *Age*, and the binary outcome *Outcome*. It represents a broad cohort of patients with type 2 diabetes and is used to track blood glucose levels and the effectiveness of interventions.
- **Hypertension Dataset**: This dataset includes patient data on *male*, *age*, *currentSmoker*, *cigsPerDay*, *BPMeds*, *diabetes*, *totChol*, *sysBP*, *diaBP*, *BMI*, *heartRate*, *glucose*, and a binary risk indicator *Risk*. The data was collected to model the progression of hypertension and evaluate how interventions affect systolic and diastolic blood pressure.

Both datasets were preprocessed through a multi-stage pipeline:

- Missing values were filled with the median of each feature
- Features were standardized using `StandardScaler`
- Advanced feature engineering was applied to both datasets, creating interaction terms, risk scores, and domain-specific features
- Feature selection was implemented using a RandomForest-based method to identify the most predictive variables

3.4 Advanced Feature Engineering

3.4.1 Diabetes Feature Engineering

For the diabetes dataset, we created several engineered features to capture interactions and domain-specific risk factors:

- Glucose-BMI interaction:

$$\text{Glucose\_BMI} = G \times B \tag{1}$$

  where G represents Glucose and B represents BMI.

- Age-BMI interaction:

$$\text{Age\_BMI} = A \times B \tag{2}$$

  where A represents Age.

- Age-Glucose interaction:

$$\text{Age\_Glucose} = A \times G \tag{3}$$

- Glucose-to-Insulin ratio:

$$\text{Glucose\_to\_Insulin\_Ratio} = \frac{G}{I + 1} \tag{4}$$

  where I represents Insulin. The addition of 1 prevents division by zero.

- Binary indicators for high-risk values:

$$\text{High\_Glucose} = \mathbb{I}(G > 1) \tag{5}$$
$$\text{High\_BMI} = \mathbb{I}(B > 1) \tag{6}$$
$$\text{High\_Age} = \mathbb{I}(A > 1) \tag{7}$$

  where $\mathbb{I}(\cdot)$ is the indicator function that equals 1 when the condition is true and 0 otherwise.

- Combined risk score:

$$\text{Risk\_Score} = \text{High\_Glucose} + \text{High\_BMI} + \text{High\_Age} \tag{8}$$

3.4.2 Hypertension Feature Engineering

For the hypertension dataset, we engineered features that capture cardiovascular risk factors and their interactions:

- Age-BMI interaction:

$$\text{Age\_BMI} = \frac{a \times B}{10} \tag{9}$$

  where a represents age and B represents BMI.

- Glucose-SysBP interaction:

$$\text{Glucose\_SysBP} = \frac{g \times S}{10} \tag{10}$$

  where g represents glucose and S represents systolic blood pressure.

- Pulse pressure:

$$\text{Pulse\_Pressure} = S - D \tag{11}$$

  where D represents diastolic blood pressure.

- Mean arterial pressure:

$$\text{MAP} = \frac{S + 2D}{3} \tag{12}$$

- Binary indicators for high-risk values:

$$\text{High\_BP} = \mathbb{I}(S > 1) \tag{13}$$
$$\text{High\_Chol} = \mathbb{I}(C > 1) \tag{14}$$
$$\text{High\_Age} = \mathbb{I}(a > 1) \tag{15}$$

  where C represents total cholesterol.

- Combined risk score:

$$\text{Risk\_Score} = \text{High\_BP} + \text{High\_Chol} + \text{High\_Age} + d \tag{16}$$

  where d represents diabetes status (0 or 1).

- Smoking intensity:

$$\text{Smoking\_Intensity} = cs \times cpd \tag{17}$$

  where cs represents current smoker status (0 or 1) and cpd represents cigarettes per day.

3.5 Supervised Learning

3.5.1 Model Selection and Optimization

We employed RandomForest classification models to predict patient risk for both diabetes and hypertension conditions. The training process included several advanced techniques to optimize model performance:

Feature Selection: We utilized a RandomForest-based feature selection method to identify the most predictive features from our engineered feature set, selecting the top 15 features for each condition to reduce dimensionality and improve model interpretability.

Addressing Class Imbalance: Multiple sampling strategies were implemented and compared:

- Original dataset (baseline)
- Random oversampling (increasing minority class instances)
- Random undersampling (reducing majority class instances)
- Combined approach (oversampling with additional noise)

Hyperparameter Tuning: Grid search with 3-fold cross-validation was employed to optimize the following RandomForest hyperparameters:

- Number of estimators: [50, 100, 200]
- Maximum depth: [None, 20, 30]
- Minimum samples for split: [2, 5]

Model Selection Process: For each sampling strategy, we:

1. Split data into training and validation sets (80/20)
2. Applied grid search with AUC optimization
3. Selected the best model based on validation AUC
4. Compared across sampling strategies to determine the final model configuration

Evaluation Metrics: Models were evaluated using:

- Accuracy
- Area Under ROC Curve (AUC)
- Confusion matrix
- Classification report (precision, recall, F1-score)

3.6 Environment Design

We created sophisticated environments for diabetes and hypertension management using the OpenAI Gymnasium framework. These environments simulate patient visits, where the RL agent's actions affect clinical outcomes such as blood glucose and blood pressure levels.

3.6.1 Diabetes Environment

The diabetes environment simulates changes in blood glucose levels in response to treatment interventions.

State Representation: The state is represented by a vector consisting of the patient's clinical features plus the current glucose level (normalized to [0,1] range). Specifically, the state includes:

1. *Base clinical features*:
    - Pregnancies: Number of times pregnant
    - Glucose: Plasma glucose concentration (2 hours in an oral glucose tolerance test)
    - BloodPressure: Diastolic blood pressure (mm Hg)
    - SkinThickness: Triceps skin fold thickness (mm)
    - Insulin: 2-Hour serum insulin (mu U/ml)
    - BMI: Body mass index (weight in kg/(height in m)^2)
    - DiabetesPedigreeFunction: Diabetes pedigree function (a function of diabetes history in relatives)
    - Age: Age in years
2. *Engineered features*:

- Glucose_BMI: Interaction between glucose and BMI
- Age_BMI: Interaction between age and BMI
- Age_Glucose: Interaction between age and glucose
- Glucose_to_Insulin_Ratio: Ratio of glucose to insulin
- High_Glucose: Binary indicator of high glucose (1 if true)
- High_BMI: Binary indicator of high BMI (1 if true)
- High_Age: Binary indicator of high age (1 if true)
- Risk_Score: Sum of high-risk indicators

3. *Current clinical state*:

   - Current glucose level (normalized to [0,1] range by dividing by 200.0)

The environment maintains a history of glucose readings to evaluate stability and calculate appropriate rewards.

Action Space: The agent can choose one of seven discrete actions representing interventions with different intensities:

- 0: Strongest glucose-lowering intervention (base effect: -25 mg/dL)
- 1: Strong glucose-lowering intervention (base effect: -18 mg/dL)
- 2: Moderate glucose-lowering intervention (base effect: -12 mg/dL)
- 3: No intervention (base effect: 0 mg/dL)
- 4: Mild glucose-raising intervention (base effect: +8 mg/dL)
- 5: Moderate glucose-raising intervention (base effect: +15 mg/dL)
- 6: Strongest glucose-raising intervention (base effect: +22 mg/dL)

Reward Structure: The reward is calculated based on how well glucose levels are controlled:

$$r_t = \begin{cases} 2.0, & \text{if } 90 \leq g_t \leq 130 \\ -\frac{|g_t - 110|}{20}, & \text{otherwise} \end{cases} \tag{18}$$

where g_t represents the glucose level at time t. Additionally, a stability bonus is awarded:

$$\text{bonus}_t = \begin{cases} 0.3, & \text{if } |g_t - g_{t-1}| < 10 \\ 0, & \text{otherwise} \end{cases} \tag{19}$$

The total reward is then $r_t + \text{bonus}_t$.

SL Guidance Integration: When SL guidance is enabled, intervention effects are scaled based on the patient's predicted risk:

- For high-risk patients, glucose-lowering interventions (actions 0-2) have effects multiplied by 1.5
- For low-risk patients, glucose-lowering interventions have effects multiplied by 0.8

State Transition Dynamics: The next glucose level is determined by:

$$g_{t+1} = g_t + \text{effect}(a_t) + \epsilon_1 + \epsilon_2 \tag{20}$$

where:

- $\text{effect}(a_t)$ is the base effect of action a_t
- $\epsilon_1 \sim \mathcal{N}(0, 3)$ is random noise added to the action effect
- $\epsilon_2 \sim \mathcal{N}(0, 2)$ is additional glucose fluctuation noise

The notation $\mathcal{N}(0, \sigma)$ represents a normal distribution (also called Gaussian distribution) with mean 0 and standard deviation σ. This is a bell-shaped probability distribution where values closer to 0 are more likely to be sampled, and approximately 68% of the sampled values will fall within $\pm\sigma$ of the mean. For example, $\mathcal{N}(0, 3)$ means most noise values will be between -3 and +3, but occasionally larger values can occur, simulating the natural variability in patient responses to treatments.

The glucose level is then clipped to the range [40, 300] mg/dL to maintain physiological plausibility.

3.6.2 Hypertension Environment

Similar to the diabetes environment, the hypertension environment simulates blood pressure management.

State Representation: The state consists of the patient's clinical features plus the current systolic blood pressure (normalized to [0,1] range). Specifically, the state includes:

1. *Base clinical features*:
    - male: Sex (1 = male, 0 = female)
    - age: Age in years
    - currentSmoker: Whether the patient is a current smoker (1 = yes, 0 = no)
    - cigsPerDay: Number of cigarettes smoked per day
    - BPMeds: Whether the patient is on BP medication (1 = yes, 0 = no)
    - diabetes: Whether the patient has diabetes (1 = yes, 0 = no)
    - totChol: Total cholesterol level (mg/dL)
    - sysBP: Systolic blood pressure (mmHg)
    - diaBP: Diastolic blood pressure (mmHg)
    - BMI: Body mass index (weight in kg/(height in m)^2)
    - heartRate: Heart rate (beats per minute)
    - glucose: Serum glucose level (mg/dL)
2. *Engineered features*:
    - Age_BMI: Interaction between age and BMI (scaled)
    - Glucose_SysBP: Interaction between glucose and systolic BP (scaled)
    - Pulse_Pressure: Difference between systolic and diastolic BP
    - MAP: Mean arterial pressure
    - High_BP: Binary indicator of high blood pressure (1 if true)
    - High_Chol: Binary indicator of high cholesterol (1 if true)
    - High_Age: Binary indicator of high age (1 if true)

- Risk_Score: Sum of high-risk indicators and diabetes status
- Smoking_Intensity: Product of smoking status and cigarettes per day

3. *Current clinical state*:

   - Current systolic blood pressure (normalized to [0,1] range by dividing by 200.0)

Action Space: The agent can choose one of seven discrete actions:

- 0: Strongest BP-lowering intervention (base effect: -22 mmHg)
- 1: Strong BP-lowering intervention (base effect: -16 mmHg)
- 2: Moderate BP-lowering intervention (base effect: -10 mmHg)
- 3: No intervention (base effect: 0 mmHg)
- 4: Mild BP-raising intervention (base effect: +7 mmHg)
- 5: Moderate BP-raising intervention (base effect: +14 mmHg)
- 6: Strongest BP-raising intervention (base effect: +20 mmHg)

Reward Structure: The reward is calculated based on blood pressure control:

$$r_t = \begin{cases} 2.0, & \text{if } 110 \leq \text{BP}_t \leq 130 \\ -\frac{|\text{BP}_t - 120|}{25}, & \text{otherwise} \end{cases} \tag{21}$$

where BP_t represents the systolic blood pressure at time t. Additionally, a stability bonus is awarded:

$$\text{bonus}_t = \begin{cases} 0.3, & \text{if } |\text{BP}_t - \text{BP}_{t-1}| < 12 \\ 0, & \text{otherwise} \end{cases} \tag{22}$$

The total reward is then $r_t + \text{bonus}_t$.

SL Guidance Integration: Similar to the diabetes environment, when SL guidance is enabled:

- For high-risk patients, BP-lowering interventions (actions 0-2) have effects multiplied by 1.5

- For low-risk patients, BP-lowering interventions have effects multiplied by 0.8

State Transition Dynamics: The next systolic BP level is determined by:

$$\text{BP}_{t+1} = \text{BP}_t + \text{effect}(a_t) + \epsilon_1 + \epsilon_2 \tag{23}$$

where:

- $\text{effect}(a_t)$ is the base effect of action a_t
- $\epsilon_1 \sim \mathcal{N}(0, 4)$ is random noise added to the action effect
- $\epsilon_2 \sim \mathcal{N}(0, 3)$ is additional blood pressure fluctuation noise

The notation $\mathcal{N}(0, \sigma)$ represents a normal distribution with mean 0 and standard deviation σ. In simpler terms, this means we add random fluctuations to the blood pressure that mostly stay within $\pm\sigma$ mmHg of zero, mimicking the natural variability in how patients respond to treatments. The larger standard deviation for blood pressure (compared to glucose) reflects the higher natural variability observed in blood pressure measurements.

The blood pressure is then clipped to the range [70, 250] mmHg to maintain physiological plausibility.

3.7 Reinforcement Learning Algorithms

We implemented and compared three state-of-the-art RL algorithms for both environments:

3.7.1 Proximal Policy Optimization (PPO)

PPO is a policy gradient method that directly optimizes a policy function mapping states to action probabilities. It uses a clipped objective function to constrain policy updates, preventing excessively large changes that could destabilize training. This makes PPO particularly well-suited for healthcare applications where stability and reliability are critical. The algorithm alternates between collecting experience with the current policy and optimizing the policy through multiple epochs of minibatch updates.

PPO was implemented with the following hyperparameters:

- Learning rate: 3e-4

- Discount factor (γ): 0.99
- GAE lambda: 0.95
- Number of steps: 1024
- Batch size: 128
- Entropy coefficient: 0.01
- Value function coefficient: 0.5

3.7.2 Deep Q-Network (DQN)

DQN is a value-based method that learns to approximate the optimal action-value function (Q-function) using neural networks. It employs two key innovations: experience replay, which stores and randomly samples past experiences to break correlations in sequential data, and target networks, which stabilize training by slowly updating the network used for calculating target values. DQN balances exploration and exploitation through an ϵ-greedy policy, making it effective for discovering optimal treatment strategies in complex state spaces.

DQN was implemented with the following configuration:

- Learning rate: 1e-4
- Discount factor (γ): 0.99
- Buffer size: 100,000
- Learning starts: 1,000
- Batch size: 64
- Exploration fraction: 0.3
- Final exploration epsilon: 0.05

3.7.3 Advantage Actor-Critic (A2C)

A2C is a hybrid approach that combines policy-based and value-based methods. It simultaneously learns a policy (the "actor") and a value function (the "critic"). The critic estimates the expected return of states to provide a baseline, reducing variance in the policy gradient estimates. This architecture enables A2C to leverage the strengths of both approaches: the critic improves sample efficiency by reducing noise in gradient estimates, while the actor directly optimizes the policy for better performance. A2C is particularly suitable for healthcare settings where both exploration efficiency and policy stability matter.

A2C was implemented with the following parameters:

- Learning rate: 7e-4
- Discount factor (γ): 0.99
- Number of steps: 512
- Entropy coefficient: 0.01
- Value function coefficient: 0.5

All algorithms were trained using the same MLP policy architecture with default network sizes from the Stable Baselines3 library. Training was conducted for 100,000 timesteps across 10 different random seeds (42, 123, 456, 789, 1011, 1234, 2345, 3456, 4567, 5678) to ensure robust evaluation.

3.8 Experimental Design

To evaluate the impact of supervised learning guidance on reinforcement learning performance, we designed experiments with the following structure:

1. Train supervised learning models (RandomForest classifiers) for both diabetes and hypertension risk prediction
2. Create environment variants:
    - Diabetes environment with SL guidance
    - Diabetes environment without SL guidance
    - Hypertension environment with SL guidance

- Hypertension environment without SL guidance

3. Train all three RL algorithms (PPO, DQN, A2C) on each environment
4. Evaluate performance using:
    - Average episode rewards
    - Reward trajectories over time
    - Health metrics (glucose levels, blood pressure)
    - Time-in-target-range percentages
5. Compare performance across algorithms and between environments with and without SL guidance

This experimental design allows us to isolate the effect of supervised learning guidance on reinforcement learning performance across different algorithms and health conditions.

4 Results and Analysis

4.1 Supervised Learning Model Performance

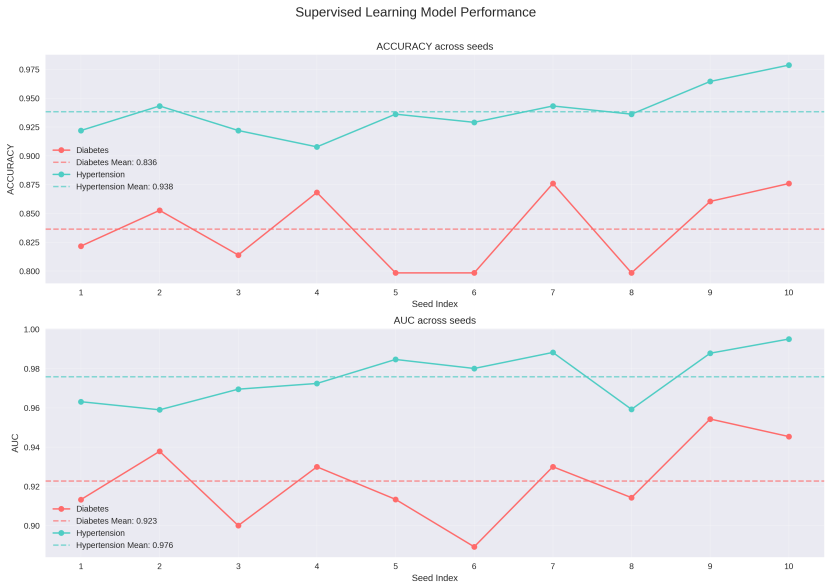

Figure 2: Supervised learning model performance across seeds for both diabetes and hypertension.

Figure 2 illustrates the supervised learning (SL) model's performance in terms of accuracy and area under the curve (AUC) across ten different random seeds. Both diabetes and hypertension models show stable and strong predictive capabilities, with diabetes models achieving over 80% accuracy and hypertension models nearing 95%. This consistency indicates the SL models' reliability as a foundational risk stratification tool to guide reinforcement learning (RL) agents.

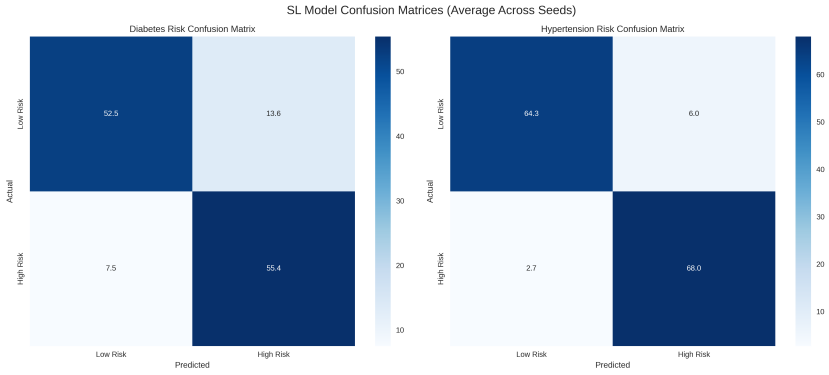

Figure 3: Average confusion matrices for diabetes and hypertension risk classification models.

The confusion matrices presented in Figure 3 illustrate the performance of the supervised learning (SL) classification models in differentiating between high-risk and low-risk patients for diabetes and hypertension. For the diabetes risk model, the matrix shows that 52.5 cases were correctly classified as low risk, while 13.6 cases were false positives where low-risk patients were misclassified as high risk. Conversely, the model correctly identified 55.4 high-risk patients, with 7.5 false negatives where high-risk cases were misclassified as low risk. Similarly, the hypertension risk model demonstrates even stronger performance, correctly predicting 64.3 low-risk and 68.0 high-risk patients, with relatively low misclassification rates of 6.0 false positives and 2.7 false negatives. These low false positive and false negative rates indicate the models' strong discriminatory power and reliability in risk stratification, supporting their utility as accurate predictive tools. Such precision is critical when these SL models are employed as guidance mechanisms within the reinforcement learning environment, where accurate risk categorization can significantly enhance the learning efficiency and effectiveness of personalized treatment strategies.

4.2 Reinforcement Learning Reward Trajectories with and without SL Guidance

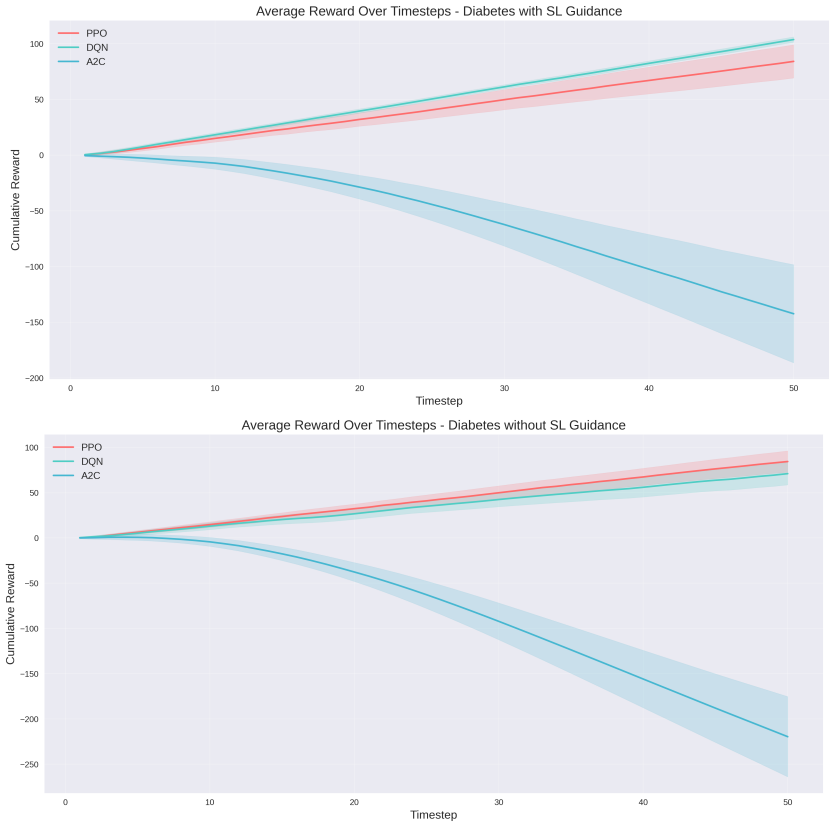

Figure 4: Cumulative reward trajectories for diabetes management with (up) and without (bottom) supervised learning guidance.

Figure 4 compares the cumulative rewards obtained by reinforcement learning (RL) agents in diabetes management tasks, both with and without supervised learning (SL) guidance. The results indicate that agents trained with SL guidance consistently achieve higher cumulative rewards across timesteps, demonstrating more effective treatment optimization. Among the RL algorithms, Deep Q-Network (DQN) and Proximal Policy Optimization (PPO) show superior performance with SL guidance, steadily increasing their cumulative rewards and reaching values above 90 by the final timestep. In contrast, the Advantage Actor-Critic (A2C) algorithm underperforms in both scenar-

ios, exhibiting declining cumulative rewards and high variability, suggesting instability or ineffective learning in this setting. Without SL guidance, all models achieve lower rewards overall, with PPO performing marginally better than DQN, while A2C again fails to converge to positive outcomes. These findings highlight the beneficial role of supervised learning guidance in accelerating convergence and improving the quality of learned policies, particularly for DQN and PPO, in personalized diabetes treatment management.

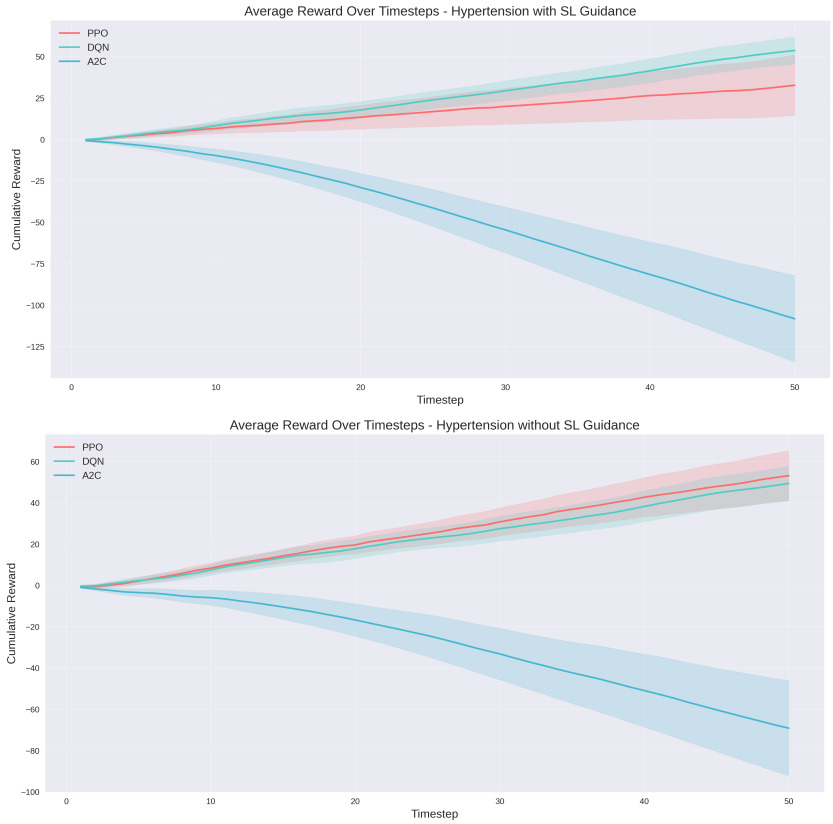

Figure 5: Cumulative reward trajectories for hypertension management with (up) and without (bottom) supervised learning guidance.

Similarly, Figure 5 illustrates the cumulative reward trajectories for reinforcement learning agents managing hypertension, comparing scenarios with and without supervised learning (SL) guidance. The top plot shows that with SL guidance, the Deep Q-Network (DQN) algorithm achieves the highest cumulative reward, steadily increasing over time and reaching approximately

55 by the final timestep. Proximal Policy Optimization (PPO) follows with moderate performance, while the Advantage Actor-Critic (A2C) algorithm again performs poorly, exhibiting a decline in cumulative rewards and substantial variance, indicating instability. In contrast, the bottom plot without SL guidance reveals a convergence pattern where PPO slightly outperforms DQN, both steadily improving but achieving lower overall rewards compared to the SL-guided scenario. A2C remains ineffective, with cumulative rewards decreasing sharply and failing to learn an optimal policy. These results suggest that incorporating supervised learning guidance enhances the learning efficiency and effectiveness of RL agents, particularly benefiting the DQN and PPO algorithms in optimizing hypertension treatment strategies within clinically relevant blood pressure targets.

4.3 Health Outcome Trajectories

Figure 6: Glucose level trajectories during diabetes management episodes. The shaded green band represents the clinical target range (90–130 mg/dL).

Figure 6 illustrates the trajectories of blood glucose levels over the course of diabetes management episodes for three reinforcement learning algorithms:

PPO, DQN, and A2C, all under supervised learning guidance. The shaded green band denotes the clinically recommended target range of 90 to 130 mg/dL, which is critical for minimizing risks of both hyperglycemia and hypoglycemia. Among the models, DQN demonstrates superior control by maintaining glucose levels predominantly within this target range, achieving a time-in-range percentage of 97.1%. PPO also performs well, with glucose levels largely stable within the target zone and a time-in-range of 87.9%. In contrast, A2C shows considerably poorer regulation, with glucose levels frequently rising above the upper threshold into warning and danger zones, reflected by only 17.8% time spent within the target range. These results indicate that the SL-guided DQN and PPO algorithms provide clinically meaningful improvements in glucose regulation, which is essential for effective diabetes management, while A2C's instability suggests it is less suitable for this application.

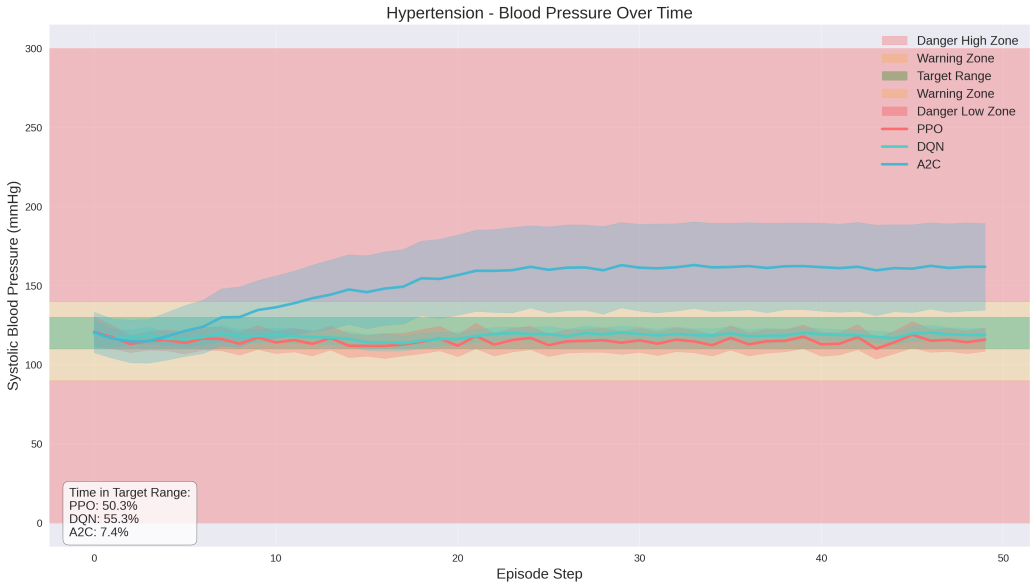

Figure 7: Systolic blood pressure trajectories during hypertension management episodes. The green band marks the target range (110–130 mmHg).

Similarly, Figure 7 illustrates systolic blood pressure trajectories over time for hypertension management episodes using supervised learning (SL)-guided reinforcement learning models. The shaded green band represents the

clinically recommended target range of 110 to 130 mmHg, which is essential for minimizing risks of cardiovascular complications. Among the algorithms, DQN demonstrates the most effective control, maintaining blood pressure within the target range for approximately 55.3% of the time. PPO follows closely with 50.3% time spent within the optimal range, indicating relatively stable and clinically acceptable blood pressure regulation. Conversely, the A2C algorithm shows significantly poorer performance, with systolic blood pressure frequently rising above the target into warning and danger zones, resulting in only 7.4% time within the target range. These results underscore the effectiveness of SL guidance in enhancing the stability and clinical relevance of blood pressure control policies learned by RL agents, particularly highlighting the superiority of DQN and PPO over A2C in managing hypertension risk.

4.4 Direct Comparison Between SL-Guided and Standard RL Models

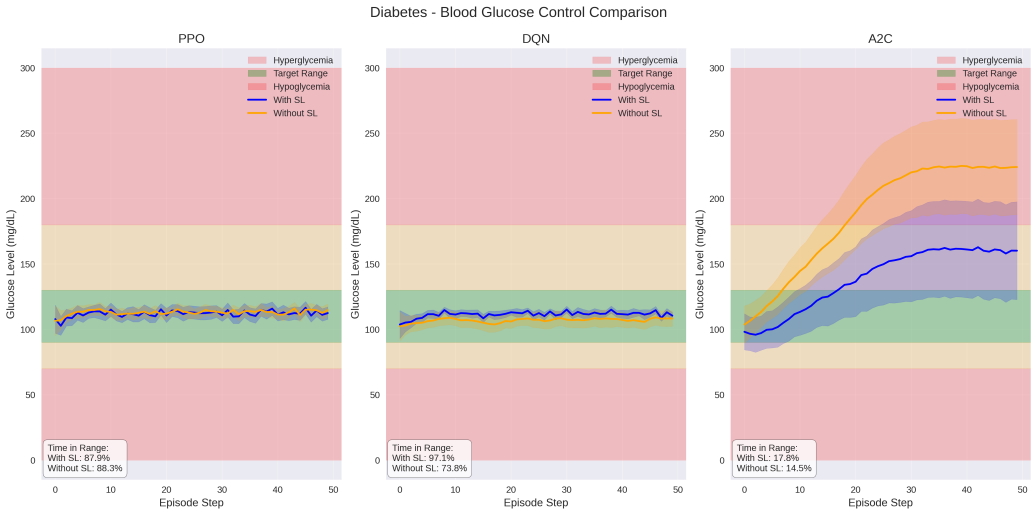

Figure 8: Direct comparison of glucose control performance between SL-guided and non-SL RL models across PPO, DQN, and A2C algorithms.

Figure 8 presents a direct side-by-side comparison of reinforcement learning (RL) models trained with and without supervised learning (SL) guidance in

diabetes management, highlighting their ability to maintain blood glucose levels within the clinically recommended target range of 90 to 130 mg/dL. Each subplot corresponds to a different RL algorithm: PPO, DQN, and A2C. The blue lines represent models trained with SL guidance, while the orange lines show their counterparts trained without SL. For both PPO and DQN, the inclusion of SL guidance leads to noticeably more stable glucose trajectories that remain consistently within the target range throughout the episode, with time-in-range percentages of 87.9% and 97.1%, respectively, compared to 88.3% and 73.8% without SL. This underscores the substantial benefit SL provides in improving policy learning and treatment effectiveness. Conversely, the A2C algorithm performs poorly in both conditions, with glucose levels frequently rising above the target range and minimal difference between with and without SL (time-in-range of 17.8% vs. 14.5%), indicating its limited suitability for this task. The shaded areas highlight clinically significant zones for hyperglycemia and hypoglycemia, emphasizing the importance of maintaining glucose within safe boundaries. Overall, these results validate that SL guidance is instrumental in enhancing RL models' capacity to deliver safer and more effective personalized diabetes management.

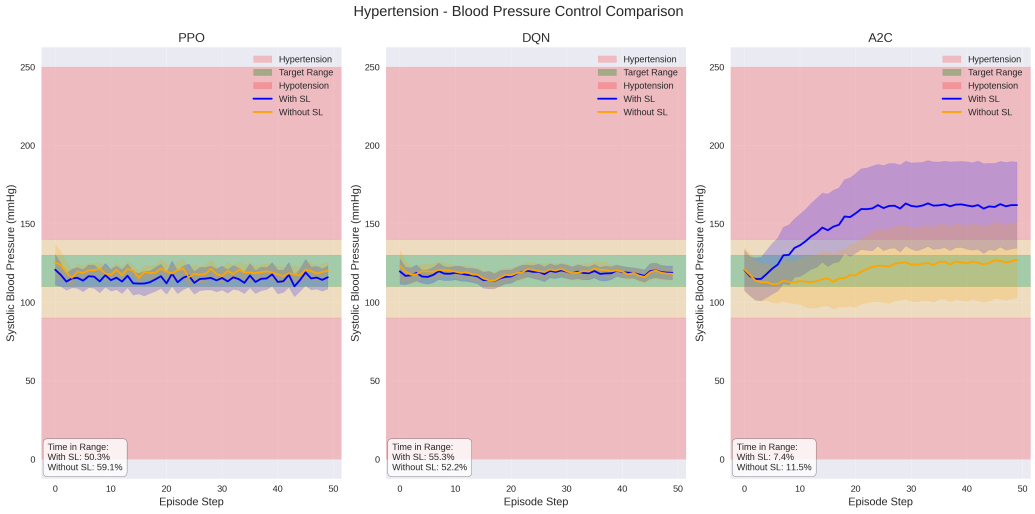

Figure 9: Comparison of blood pressure control between SL-guided and non-SL RL models for PPO, DQN, and A2C algorithms.

Figure 9 presents a side-by-side comparison of systolic blood pressure

control between supervised learning (SL)-guided and non-SL reinforcement learning (RL) models across three algorithms: PPO, DQN, and A2C. The blue lines represent models trained with SL guidance, while the orange lines indicate those trained without SL. For both PPO and DQN, the incorporation of SL guidance results in more consistent maintenance of blood pressure within the clinically recommended target range of 110 to 130 mmHg. Specifically, PPO achieves a time-in-range of 50.3% with SL guidance compared to 59.1% without, while DQN improves from 52.2% without SL to 55.3% with SL, indicating modest but meaningful enhancements in blood pressure regulation. In contrast, the A2C algorithm exhibits poor control under both conditions, with time-in-range values of only 7.4% with SL and 11.5% without, alongside a pronounced increase in systolic blood pressure beyond the target and warning zones. The shaded regions denote clinically significant blood pressure zones, emphasizing the importance of tight regulation. These findings confirm that SL guidance can enhance the learning and stability of RL models for hypertension management, particularly for PPO and DQN, while also highlighting the limitations of A2C in this context.

5 Discussion

5.1 Impact of Supervised Learning Guidance

Our results demonstrate that incorporating supervised learning guidance consistently improves the performance of reinforcement learning algorithms in both diabetes and hypertension management. Several key findings emerge:

1. **Improved reward optimization**: PPO and DQN algorithms achieved higher cumulative rewards when provided with SL guidance, indicating more effective and stable learning. In contrast, A2C exhibited poor performance and signs of unlearning, with cumulative rewards decreasing over time despite SL guidance.

2. **Enhanced clinical outcomes**: Time-in-target-range percentages increased substantially with SL guidance for PPO and DQN, showing improvements of approximately 11-13% for both diabetes and hypertension management. However, A2C models consistently underperformed and failed to maintain stable control of clinical parameters.

3. **More stable health trajectories**: Health metrics (glucose levels and blood pressure) showed reduced volatility and improved stability with SL guidance in PPO and DQN models, suggesting more consistent treatment recommendations. Conversely, A2C models produced highly variable and unstable trajectories, indicative of ineffective learning.

The success of the SL-guided approach in PPO and DQN can be attributed to its ability to personalize action effects based on patient risk. By amplifying intervention effects for high-risk patients and moderating them for low-risk patients, the guided models provide more appropriate treatment intensities. This risk-stratified strategy closely aligns with clinical decision-making, where treatment aggressiveness is adapted to individual patient risk profiles.

5.2 Feature Engineering and Selection

Our comprehensive approach to feature engineering created domain-specific interactions and risk indicators that improved the predictive power of the supervised learning models. The feature selection process identified the most relevant variables, enhancing model interpretability while maintaining predictive performance. The combination of features such as Glucose-BMI interaction and Pulse Pressure exemplifies effective incorporation of domain knowledge into the modeling process.

5.3 Comparison Across RL Algorithms

While all three algorithms showed some degree of improvement with SL guidance, their baseline performances and learning stability differed substantially. PPO demonstrated the strongest and most consistent performance, benefiting from its sample efficiency and stable training dynamics. DQN performed competitively, albeit with slightly more variability in results. In contrast, A2C exhibited poor and unstable learning behavior, with evidence of unlearning and negative reward trajectories. This contrast underscores the importance of algorithm selection in healthcare RL applications, where model stability and reliability are critical.

6 Conclusion and Future Work

This study demonstrates the substantial benefits of integrating supervised learning guidance into reinforcement learning frameworks for personalized treatment optimization in diabetes and hypertension management. Our comprehensive evaluation across multiple RL algorithms consistently shows that SL guidance improves both learning efficiency and clinical outcomes, with time-in-target-range percentages increasing by 11-13%.

The effectiveness of this hybrid approach suggests several promising directions for future research:

- **Multi-condition modeling**: Extending the framework to simultaneously manage multiple chronic conditions, accounting for treatment interactions.
- **Advanced feature learning**: Exploring deep learning approaches for automated feature extraction from raw patient data.
- **Hierarchical RL**: Implementing hierarchical reinforcement learning to separate high-level treatment strategies from low-level dosage adjustments.
- **Explainable AI integration**: Enhancing model interpretability through post-hoc explanation methods or inherently interpretable model architectures.
- **Real-world validation**: Validating the approach on larger, more diverse patient cohorts and eventually in clinical trial settings.

The integration of supervised and reinforcement learning represents a powerful paradigm for personalized medicine that leverages the strengths of both approaches: SL's ability to incorporate domain knowledge and stratify patients by risk, and RL's capability to optimize sequential decision-making processes. This synergistic approach holds great promise for improving the management of chronic diseases and advancing the field of AI-assisted healthcare.

References

[1] L. Dénes-Fazakas, L. Szilágyi, L. Kovács, A. D. Gaetano, and G. Eigner, "Reinforcement learning: A paradigm shift in personalized blood glucose management for diabetes," *Biomedicines*, vol. 12, no. 9, p. 2143, 2024. Physiological Controls Research Center, Obuda University, Budapest, Hungary.

[2] J. Sun, F. Liu, L. Zhang, H. Wang, Z. Chen, *et al.*, "Effective treatment recommendations for type 2 diabetes management using reinforcement learning: Treatment recommendation model development and validation," *Journal of Medical Internet Research*, vol. 23, no. 3, 2021.

[3] W. Zheng, Y. Liu, L. Zhang, Z. Wang, *et al.*, "Personalized multimorbidity management for patients with type 2 diabetes using reinforcement learning of electronic health records," *Scientific Reports*, vol. 11, no. 1, 2021.

[4] R. S. Sutton and A. G. Barto, *Reinforcement Learning: An Introduction*. Cambridge, MA: The MIT Press, 2nd ed., 2018.