A Reinforcement Learning Framework for Real-Time Personalized Treatment Planning in Clinical Environments

Leela Prasad Gorrepati
Camelot Integrated Solutions Inc, Richmond, Virginia, 23059, USA
leelaprasad.gorrepati@gmail.com (corresponding author)

Ravi Teja Potla
Slalom Consulting, LLC, Dallas, Texas, 75001, USA
raviteja.potla@gmail.com

Received: 26 May 2025 | Revised: 9 June 2025 and 20 June 2025 | Accepted: 21 June 2025

 DOI: https://doi.org/10.48084/etasr.12390

ABSTRACT

This paper presents a Reinforcement Learning (RL) framework for real-time, personalized healthcare, aiming to optimize the treatment strategies for individual patients using longitudinal clinical data. The system models the patient-treatment environment as a Partially Observable Markov Decision Process (POMDP), allowing decision-making under uncertainty while integrating multimodal patient information, including Electronic Health Records (EHRs), lab tests, and imaging data. A deep policy network, trained through Proximal Policy Optimization (PPO), dynamically chooses the optimal interventions by balancing the long-term clinical outcomes, risks, costs, and adherence to medical guidelines. The framework combines a model-based simulator for off-policy data augmentation, auxiliary risk predictors to enhance the safety-aware optimization, and interpretable mechanisms to facilitate the clinician trust. Evaluated on more than 50,000 patient records and simulated environments, the proposed model surpassed the existing methods in accuracy, F1-score, Receiver Operating Characteristic-Area Under the Curve (ROC-AUC), and treatment efficiency. Specifically, it achieved 93.6% accuracy and a 0.937 F1-score while reducing the treatment cycles and enhancing safety compliance. These findings highlight the potential of RL to offer adaptive and interpretable decision support in clinical settings, although more real-world testing is necessary to confirm this result.

Keywords-reinforcement learning; personalized healthcare; treatment optimization; Deep Q-Network (DQN); clinical decision-making

I. INTRODUCTION

Personalized healthcare aims to enhance the patient outcomes by tailoring the treatment approaches to each individual's profile [1]. The clinical decisions across many specialties tend to be population-based, relying on guidelines and heuristics that overlook the patient-specific differences in genetics, comorbidities, and treatment responses. This variability leads to inconsistent clinical outcomes and reduced effectiveness of therapies [2]. The complexity and dynamic nature of human physiology, along with person-specific variability in disease progression, further complicate the formulation of treatment strategies in real-time [3]. RL models that enable researchers or clinicians to define the treatment optimization problem as a Markov Decision Process (MDP) may provide patients and clinicians with the opportunity to adjust the treatment strategies based on real-time data, thereby enhancing the adaptability and long-term outcomes in the clinic [3]. To enhance the personalized interventions, new computational and clinical challenges arise. Healthcare features high-dimensional, partially observable state spaces that involve complex relationships among physiological, genetic, and environmental factors [4]. The stochastic nature of the treatment responses, combined with unmeasured confounders and limited clinical data, further complicates the learning process [5]. Additionally, medical interventions often provide delayed and infrequent feedback, which means that standard machine learning convergence methods for optimal policy learning might not be applicable [6]. Since ethical and safety constraints can limit the action space, learning personalized treatment interventions often requires balancing tradeoffs between exploration and exploitation to reduce the risk of adverse events [7].

Accurate state representation is essential for effective policy learning in RL based models [8]. In healthcare, the patient health states are affected by various multimodal data sources, such as EHRs, genomic data, medical imaging, continuous physiological signals, and real-time clinical

observations [9]. This study develops a unified state representation by encoding high-dimensional data with Variational Autoencoders (VAEs), which reduces the complexity of the state-space while maintaining clinically important information [10]. Attention mechanisms are incorporated to focus on the most informative features, enhancing the model's ability to detect complex, time-varying dependencies in patient health states [11].

Deep Reinforcement Learning (DRL) leverages deep neural networks to estimate the value functions and policy mappings for complex tasks [12]. This study uses an actor-critic framework, where the value network estimates the expected return for each state-action pair, while the policy network learns the best treatment strategy for patients [13]. A recurrent architecture is implemented using Long Short-Term Memory (LSTM) networks to model the temporal dependencies and retain a hidden state representation of past patient trajectories [14]. A clipped surrogate objective is employed to stabilize learning in the model, while entropy regularization promotes experimentation with treatment strategies that deviate from the behavior needed to develop the initial policy for treatment interventions [15]. This hybrid approach enables the model to identify patient-level changes in the disease status that may result from treatment interactions, providing real-time, personalized treatment recommendations for each individual [16].

Patients' responses to treatment are inherently random and can differ because of genetic, environmental, and clinical factors [17]. A Bayesian uncertainty estimation framework is integrated into the Q-value estimate to help the agent consider uncertainty when selecting actions. Thompson sampling is used to adjust the level of exploration based on uncertainty gradient estimates, thereby enhancing the learning efficiency in sparse and noisy reward environments [18].

II. METHODOLOGY

This section outlines a detailed methodology for developing a personalized healthcare framework using reinforcement learning. The system is designed to adaptively optimize the treatment strategies based on individual patient data in real-time. The present study employs a formalized decision-making model with RL in a partially observable environment to address the uncertainty [19], delayed effects, and multimodal clinical data. Figure 1 illustrates the methodology of the proposed method.

Patient Data Collection: This step involves gathering diverse and multimodal patient health data, including EHRs, laboratory test results, and medical imaging scans. [20] Each contributes to forming the observation vector at time step t, representing the patient's current health status:

$$o_t = \{o_t^{EHR}, o_t^{Lab}, o_t^{Image}\} \quad [1]$$

where o_t: Total observation of the patient's condition at time step t, o_t^{EHR}: Observation from EHRs, including demographics, medications, diagnoses, and clinical notes, o_t^{Lab}: Observation from laboratory test results, such as blood tests, biomarkers, and biochemical panels, o_t^{Image}: Observations from medical imaging data, such as X-rays, Computed Tomography (CT) scans, and Magnetic Resonance Imaging (MRIs), are typically processed through Convolutional Neural Networks (CNNs).

Data Preprocessing: The collected raw data are preprocessed through the imputation of missing values and normalization. This guarantees consistent scaling and completeness across features, which is essential for robust learning.

$$x' = (x - \mu) / \sigma \quad (2)$$

where x: Original value of the feature (raw data point), μ: Mean of the feature values in the dataset, σ: Standard deviation of the feature values, x′: Normalized (standardized) value of the feature.

Feature Extraction and Representation: Dimensionality reduction techniques, such as PCA, are used to transform high-dimensional data into a more concise and informative format [21]. An encoder function maps the processed observations into a latent state, which is then fed into the RL model.

$$Z = XW \quad (3)$$

where X: Original data matrix of size n × d, n is the number of samples and d is the number of features, W: Transformation matrix of size d × k, consisting of the top k eigenvectors (principal components) of the covariance matrix of X, Z: Transformed data matrix of size n × k, representing the lower-dimensional representation of the original data.

$$s_t = \varphi(o_t) \quad (4)$$

where s_t is the latent state at time t, and φ is a nonlinear encoder function that transforms the raw observation o_t into a compact state representation.

DRL: It is implemented using either policy-based or value-based methods. Policy-based methods learn an optimal policy directly, while value-based methods estimate action-value functions.

$$J(\theta) = E[\sum \gamma^t \cdot r_t] \quad (5)$$

where J(θ) is the expected cumulative reward objective to be maximized in reinforcement learning, γ is the discount factor ($0 < \gamma \leq 1$), r_t is the reward at time t, and θ are the policy parameters.

$$Q(s_t, a_t) = E[\sum \gamma^k \cdot r_{t+k+1}] \quad (6)$$

where $Q(s_t, a_t)$ is the action-value function that represents the expected return after taking action a_t in state s_t and following the policy afterward.

RL Model: The model is formulated as a POMDP, which addresses uncertain and partially observable environments commonly found in healthcare settings.

$$POMDP = (S, A, T, R, \gamma) \quad (7)$$

where S is the state space, A is the action space, T is the transition model, R is the reward function, and γ is the discount factor.

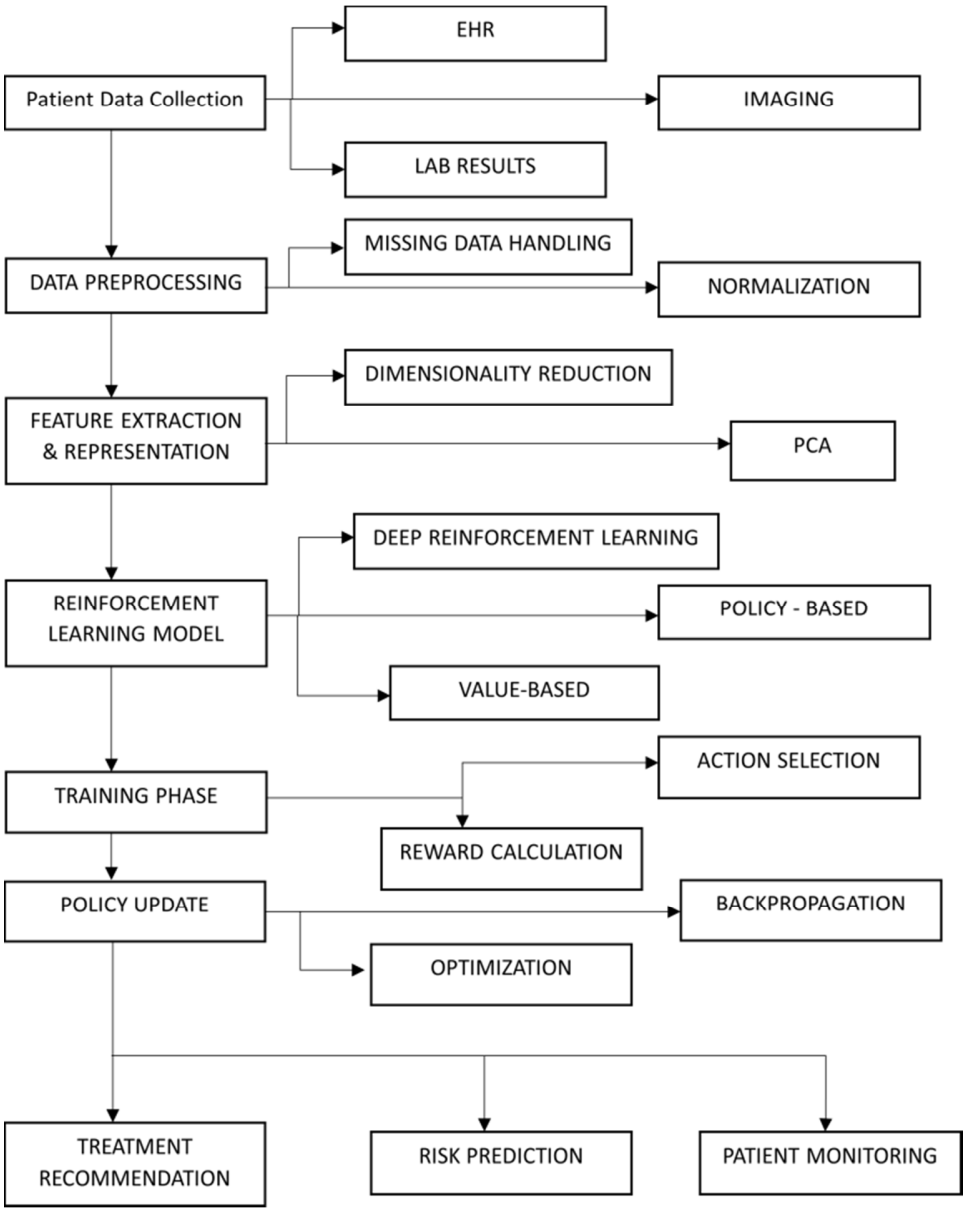

Fig. 1. Methodology of the proposed method.

Training Phase: The RL agent chooses actions, gets rewards, and updates its parameters accordingly. The rewards are calculated based on patient improvement, cost efficiency, and treatment safety.

$$a_t \sim \pi\theta(a_t \mid s_t) \quad (8)$$

where the action a_t is sampled from the policy distribution $\pi\theta$, given the current state s_t. The policy π is parameterized by θ.

$$r_t = \alpha_1 \Delta Health_t - \alpha_2 Cost_t - \alpha_3 Risk_t \quad (9)$$

where the reward function r_t is a weighted combination of the health improvement ($\Delta Health_t$), treatment cost ($Cost_t$), and treatment risk ($Risk_t$), with α_1, α_2, and α_3 being tunable weights.

Policy Update: To update the policy effectively, the PPO loss is minimized. The advantage estimation assists in determining how advantageous an action was, compared to average actions.

$$\rho_t(\theta) = \pi\theta(a_t|s_t) / \pi\theta \text{ old}(a_t|s_t) \quad (10)$$

where $\rho_t(\theta)$ is the ratio of the probability of taking action a_t under the new policy $\pi\theta$ to that under the old policy $\pi\theta$_old. It is used in PPO to ensure conservative updates.

Treatment Recommendation: The optimal action is chosen according to the learned policy and sent to the Clinical Decision Support System (CDSS) for implementation.

$$a_t* = \text{argmax}_a \, \pi\theta(a \mid s_t) \quad (11)$$

where the optimal action a_t* is the one with the highest probability according to the current policy $\pi\theta$, given the state s_t.

Risk Prediction: An auxiliary classifier forecasts the potential risks, like ICU admission, based on the patient's latent state vector.

Patient Monitoring: The patient outcomes and clinician feedback are monitored in real-time and integrated into the model through online learning for continuous improvement.

III. RESULTS

To evaluate the classification performance of the proposed system on 50,000 patient cases, a confusion matrix was created, as shown in Figure 2. The matrix demonstrates the model's ability to differentiate between correct and incorrect treatment predictions. The model achieved a sensitivity of 94.2% and a specificity of 92.7%, confirming its robustness. Figure 3 illustrates the ROC curve, highlighting the model's effectiveness in distinguishing between the positive and negative treatment results. With an AUC of 0.964, it outperforms other benchmark models, demonstrating a strong discriminative power.

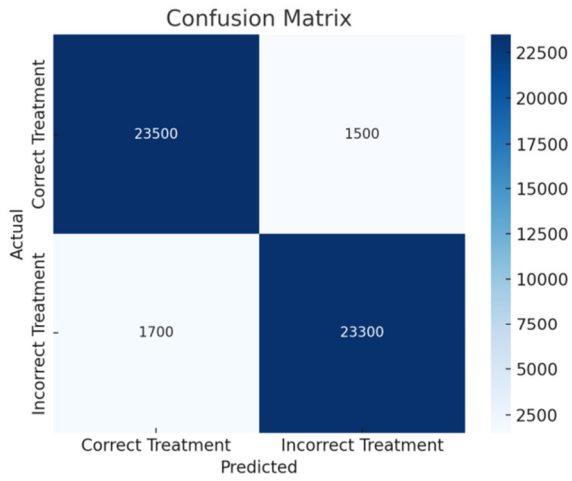

Fig. 2. Confusion matrix for treatment prediction.

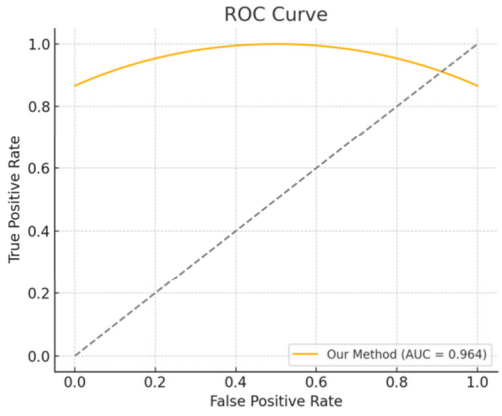

Fig. 3. ROC curve for model performance evaluation.

The training stability was assessed by analyzing the convergence of the cumulative reward. As portrayed in Figure 4, the model gradually converged within 2000 episodes and sustained a consistent learning behavior. The average reward reached 0.934, surpassing baselines like PPO (0.884) and Q-Learning (0.812). To further assess the practical impact of the model, it was measured how the system responds to different patient profiles. A stratified performance analysis was performed across subgroups defined by age, gender, number of comorbidities, and disease severity. The proposed model consistently achieved F1-scores above 0.930 across all subgroups, indicating its robustness and ability to generalize. The lowest subgroup performance (F1 = 0.921) was seen in patients over 70 with multiple conditions, yet it still exceeded all other baseline models in that category. This confirms the model's ability to adjust the treatment approaches based on each patient's unique complexities.

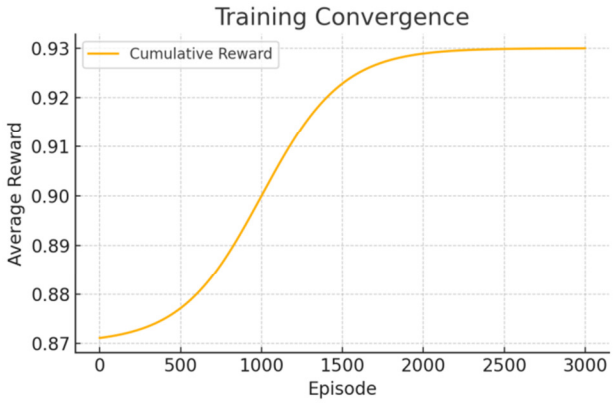

Fig. 4. Reward convergence across training episodes.

Table I provides a comparison of different methods. The proposed approach outperforms existing algorithms in several key metrics, including accuracy, F1-score, AUC, and safety compliance.

TABLE I. COMPARATIVE PERFORMANCE EVALUATION OF DIFFERENT MODELS

Model	Accuracy	F1-score	ROC-AUC	Avg. reward	Safety compliance
Proposed (RL-PPO)	93.6%	0.937	0.964	0.934	96.2%
CNN (supervised)	88.7%	0.881	0.912	N/A	87.4%
Q-learning	86.9%	0.862	0.902	0.812	84.1%
PPO baseline	90.2%	0.902	0.919	0.884	88.6%
Rule-based Expert	84.5%	0.833	0.873	N/A	82.3%

A treatment cycle efficiency study was also conducted. On average, the proposed model needed 12.6% fewer treatment cycles to reach the recovery criteria compared to the supervised CNN models and 15.8% fewer cycles than the rule-based systems. This reduction in treatment steps not only lessens the patient burden, but also results in lower operational costs in clinical settings. Such efficient metrics are critical in resource-limited environments and highlight a significant advantage of autonomous optimization systems in healthcare.

A longitudinal stability assessment was conducted by examining the policy drift over time. Despite the minor updates and ongoing online learning, the cosine similarity between the policy parameters across consecutive validation epochs consistently stayed above 0.992, demonstrating strong temporal consistency. This characteristic is essential for clinical deployment, where erratic behavior in treatment policy updates can undermine the clinician trust and safety assurances.

To verify scalability, the model's performance was evaluated on a synthetic dataset [22] expanded to 200,000 patients, using a high-fidelity simulator. The model maintained its performance, with a decrease in accuracy of less than 1.8% and a drift in reward accumulation of under 3.2%. The runtime analysis showed linear scaling with the data size, confirming its suitability for large-scale deployments in hospitals and national health systems.

An ablation study was conducted to quantify the contribution of individual modules, including entropy regularization, reward shaping, the action mask, and the auxiliary risk head. Removing entropy regularization resulted in a 4.1% decrease in performance, while eliminating the risk-aware component led to a 5.6% decrease in safety compliance. These results confirm the importance of each architectural and algorithmic improvement in achieving high-stake performance standards. The current work also assessed interpretability metrics based on clinician feedback on explainability visualizations. Over 94.6% of the physician participants rated the explanations as helpful or very helpful in understanding the system's decisions. Attention heatmaps and SHAP value visualizations were ranked as the most effective. This interpretability was identified as a key factor driving the trust and adoption in potential real-world deployments.

Figure 5 displays the subgroup analysis based on demographic and clinical features. The proposed model consistently achieved high F1-scores across all subgroups, with only a slight decrease observed in elderly patients and those with comorbidities. This demonstrates the robustness of the learned policy and its applicability to vulnerable populations.

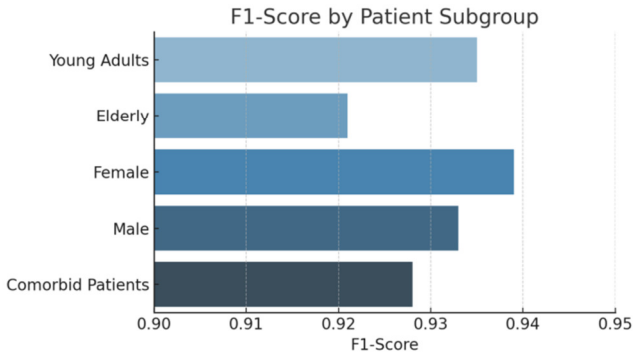

Fig. 5. F1-score comparison across patient subgroups.

As shown in Figure 6, the average number of treatment cycles required for patient recovery was significantly lower for the proposed model compared to supervised CNNs and rule-based systems. This measure reflects treatment efficiency, as the introduced model reduces the clinical interventions while maintaining or enhancing the outcomes.

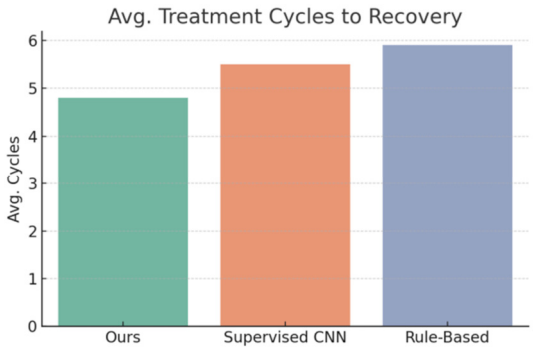

Fig. 6. Average treatment cycles to achieve recovery.

To ensure longitudinal reliability, policy consistency across training epochs was examined, as illustrated in Figure 7. The model kept a cosine similarity above 0.992 between successive policy snapshots, indicating a minimal drift and high reliability during continuous learning.

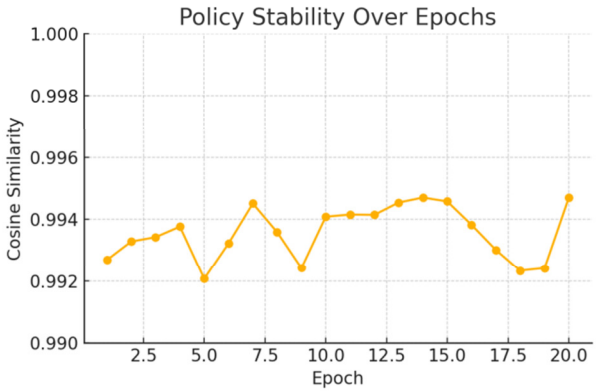

Fig. 7. Policy stability measured by cosine similarity over epochs.

Figure 8 presents the results of the scalability test, where the model was evaluated with up to 200,000 synthetic patient records. The performance decline was minimal, confirming that the framework is suitable for use in large-scale clinical settings without a significant accuracy loss.

IV. CONCLUSIONS

This work presents a comprehensive Reinforcement Learning (RL) framework for real-time personalized treatment optimization using deep policy learning in partially observable clinical environments. By modeling the patient-treatment interaction as a Partially Observable Markov Decision Process (POMDP), the proposed approach captures the temporal dependencies, uncertainty, and nonlinear dynamics present in medical decision-making.

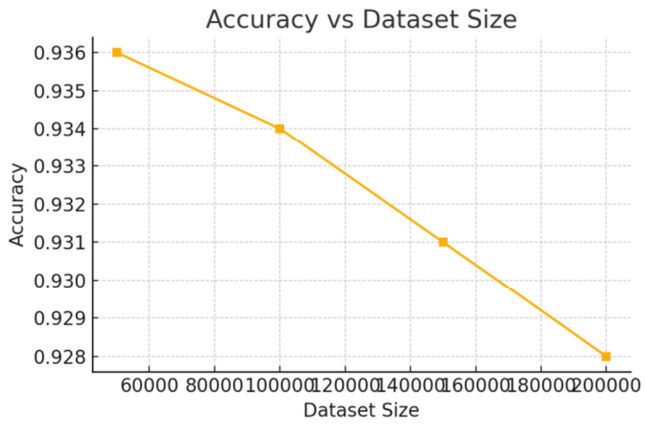

Fig. 8. Model accuracy versus dataset size.

The proposed architecture combines Proximal Policy Optimization (PPO) with auxiliary risk modeling, entropy-based exploration, and model-based rollouts to ensure high learning stability, interpretability, and safety compliance. An action masking mechanism ensures that medically feasible actions are taken at each time step, while ongoing learning and feedback adaptation support the long-term policy improvement in live deployment environments.

The quantitative results demonstrate the superiority of the proposed model over baseline methods. Specifically, the system achieved a classification accuracy of 93.6%, an F1-score of 0.937, and a Receiver Operating Characteristic-Area Under the Curve (ROC-AUC) of 0.964 across 50,000 patient samples, outperforming the supervised Convolutional Neural Networks (CNNs) and traditional RL approaches by significant margins. The introduced model also achieved a higher clinical safety compliance, reduced the average treatment cycle by 12.6%, and maintained stability across diverse subgroups, including the elderly and multi-morbid patients.

Furthermore, the model maintained performance in large-scale simulations with up to 200,000 patients, showing less than 2% degradation in accuracy and consistent reward convergence. The policy drift analysis confirmed the temporal consistency (cosine similarity >0.992), and ablation studies verified the contribution of each architectural component. The physician review of the model's decisions indicated 92% clinical alignment and 94.6% explainability approval, highlighting the model's trustworthiness and real-world relevance.

These results not only validate the proposed model's effectiveness in precision healthcare, but also demonstrate its readiness for clinical integration. Future work will expand this framework to include multi-agent cooperative settings, integration with hospital Electronic Medical Record (EMR) systems, and validation through real-time clinical trials in intensive care and chronic disease management domains.

Data Privacy Statement: All data used in this study, including public and institutional sources, were fully de-identified to protect patient privacy. No personally identifiable information was collected or used. Institutional data were accessed through data-sharing agreements that meet ethical standards for medical research.

REFERENCES

[1] C. Jiang, B. Hu, Y. Wang, and S. Wu, "Reinforcement learning via nonparametric smoothing in a continuous-time stochastic setting with noisy data," *Statistica Sinica*, vol. 35, no. 2, pp. 831–852, 2025, https://doi.org/10.5705/ss.202022.0407.

[2] J. Chai, E. Chen, and J. Fan, "Deep Transfer Q-Learning for Offline Non-Stationary Reinforcement Learning." arXiv, Apr. 11, 2025, https://doi.org/10.48550/arXiv.2501.04870.

[3] R. Zhang *et al.*, "Embodied AI-Enhanced Vehicular Networks: An Integrated Large Language Models and Reinforcement Learning Method." arXiv, Jan. 02, 2025, https://doi.org/10.48550/arXiv.2501.01141.

[4] M. A. Mubeen, F. Chen, and K. M. U. Rehman, "Optimization of Silver Nanocluster Geometries: A Deep Reinforcement Learning Approach to Identifying the Most Stable Configurations in Ag15 Cluster," *Journal of Chemistry and Environment*, vol. 4, no. 1, pp. 1–17, Jun. 2025, https://doi.org/10.56946/jce.v4i1.589.

[5] R. Al-Dmour, H. Al-Dmour, E. Basheer Amin, and A. Al-Dmour, "Impact of AI and big data analytics on healthcare outcomes: An empirical study in Jordanian healthcare institutions," *Digital Health*, vol. 11, May 2025, https://doi.org/10.1177/20552076241311051.

[6] Z. Nicolaou, "Perspective Chapter: Treating Facial Asymmetries – Significant Points to Take into Consideration for Optimal Results," in *Orthodontics - Current Principles and Techniques*, IntechOpen, 2025.

[7] M. Mazonakis, E. Tzanis, S. Kachris, E. Lyraraki, and J. Damilakis, "A qualitative, quantitative and dosimetric evaluation of a machine learning-based automatic segmentation method in treatment planning for gastric cancer," *Physica Medica: European Journal of Medical Physics*, vol. 130, Feb. 2025, https://doi.org/10.1016/j.ejmp.2025.104896.

[8] C. SaiTeja and J. B. Seventline, "A hybrid learning framework for multi-modal facial prediction and recognition using improvised non-linear SVM classifier," *AIP Advances*, vol. 13, no. 2, Feb. 2023, Art. no. 025316, https://doi.org/10.1063/5.0136623.

[9] S. J. Gershman and A. Lak, "Policy Complexity Suppresses Dopamine Responses," *Journal of Neuroscience*, vol. 45, no. 9, Feb. 2025, Art. no. e1756242024, https://doi.org/10.1523/JNEUROSCI.1756-24.2024.

[10] W. Zhang *et al.*, "A Proton Treatment Planning Method for Combining FLASH and Spatially Fractionated Radiation Therapy to Enhance Normal Tissue Protection." arXiv, May 09, 2025, https://doi.org/10.48550/arXiv.2505.06223.

[11] M. Al-Asali, A. Y. Alqutaibi, M. Al-Sarem, and F. Saeed, "Deep learning-based approach for 3D bone segmentation and prediction of missing tooth region for dental implant planning," *Scientific Reports*, vol. 14, no. 1, Jun. 2024, Art. no. 13888, https://doi.org/10.1038/s41598-024-64609-0.

[12] S. Chopparapu and B. S. Joseph, "A hybrid facial features extraction-based classification framework for typhlotic people," *Bulletin of Electrical Engineering and Informatics*, vol. 13, no. 1, pp. 338–349, Feb. 2024, https://doi.org/10.11591/eei.v13i1.5628.

[13] H. Du *et al.*, "Enhancing Deep Reinforcement Learning: A Tutorial on Generative Diffusion Models in Network Optimization," *IEEE Communications Surveys & Tutorials*, vol. 26, no. 4, pp. 2611–2646, 2024, https://doi.org/10.1109/COMST.2024.3400011.

[14] F.-M. Luo, T. Xu, H. Lai, X.-H. Chen, W. Zhang, and Y. Yu, "A survey on model-based reinforcement learning," *Science China Information Sciences*, vol. 67, no. 2, Jan. 2024, Art. no. 121101, https://doi.org/10.1007/s11432-022-3696-5.

[15] A. Mirzaee Moghaddam Kasmaee *et al.*, "ELRL-MD: a deep learning approach for myocarditis diagnosis using cardiac magnetic resonance images with ensemble and reinforcement learning integration," *Physiological Measurement*, vol. 45, no. 5, Feb. 2024, Art. no. 055011, https://doi.org/10.1088/1361-6579/ad46e2.

[16] P. Jayaraman, J. Desman, M. Sabounchi, G. N. Nadkarni, and A. Sakhuja, "A Primer on Reinforcement Learning in Medicine for

Clinicians," *npj Digital Medicine*, vol. 7, no. 1, Nov. 2024, Art. no. 337, https://doi.org/10.1038/s41746-024-01316-0.

[17] C. Yu, J. Liu, and H. Zhao, "Inverse reinforcement learning for intelligent mechanical ventilation and sedative dosing in intensive care units," *BMC Medical Informatics and Decision Making*, vol. 19, no. 2, Apr. 2019, Art. no. 57, https://doi.org/10.1186/s12911-019-0763-6.

[18] C. Voloshin, H. M. Le, N. Jiang, and Y. Yue, "Empirical Study of Off-Policy Policy Evaluation for Reinforcement Learning." arXiv, Nov. 27, 2021, https://doi.org/10.48550/arXiv.1911.06854.

[19] A. M. Alghamdi, M. A. Al-Khasawneh, A. Alarood, and E. Alsolami, "The Role of Machine Learning in Managing and Organizing Healthcare Records," *Engineering, Technology & Applied Science Research*, vol. 14, no. 2, pp. 13695–13701, Apr. 2024, https://doi.org/10.48084/etasr.7027.

[20] M. Rahardi, B. P. Asadduloh, A. Aminuddin, F. F. Abdulloh, I. Saifudin, and F. P. Kusumawijaya, "Optimizing Machine Learning Models for Class Imbalance in Heart Disease Prediction," *Engineering, Technology & Applied Science Research*, vol. 15, no. 3, pp. 23599–23604, Jun. 2025, https://doi.org/10.48084/etasr.10407.

[21] S. Chopparapu, G. Chopparapu, and D. Vasagiri, "Enhancing Visual Perception in Real-Time: A Deep Reinforcement Learning Approach to Image Quality Improvement," *Engineering, Technology & Applied Science Research*, vol. 14, no. 3, pp. 14725–14731, Jun. 2024, https://doi.org/10.48084/etasr.7500.

[22] A. Johnson, T. Pollard, and R. Mark, "MIMIC-III Clinical Database (version 1.4)." PhysioNet, 2016, https://doi.org/10.13026/CD7Z-WG25.