Publishing Excellence Inspiring Progress
ARISTON PUBLICATIONS
CompSci & AI Advances
ISSN: 3064-6758 https://aristonpubs.com/computers-ai-advances https://doi.org/10.69626/cai.2025.0025

RESEARCH ARTICLE

A Deep Reinforcement Learning Framework with Explainable AI for Personalized and Interpretable Treatment Recommendations in Healthcare

T. Thangarasan [1,*], M. Devika [2], C. Sincija [3], Khushboo Tripathi [4], P. Logamurthy [5], Kai Song [3,4], Mei Bie[6], Jie Yang[7,*]

ABSTRACT: The integration of Explainable Artificial Intelligence (XAI) into healthcare has significantly advanced clinical decision-making by enhancing the transparency and trustworthiness of AI-driven recommendations. This study introduces a novel Deep Reinforcement Learning (DRL) framework designed to generate personalized treatment recommendations tailored to individual patient profiles. The framework combines Deep Q-Learning and Policy Gradient methods to dynamically model and optimize treatment pathways, utilizing historical clinical data, patient demographics, and treatment response patterns. To ensure interpretability, an explainability layer incorporating SHAP (SHapley Additive exPlanations) and LIME (Local Interpretable Model-Agnostic Explanations) provides clinicians with actionable insights into the model's decision-making process. The proposed framework was rigorously evaluated on a real-world dataset comprising 50,000 electronic health records (EHRs) from patients with cardiovascular disease and diabetes. Experimental results demonstrated a 28% improvement in treatment success rates, a 35% reduction in adverse effects, and a 20% increase in clinician acceptance compared to conventional rule-based methods. Additionally, the explainability module achieved an average accuracy of 92% in attributing model decisions to key patient features, reinforcing its reliability in clinical settings. These findings underscore the potential of the DRL-XAI framework to enhance patient outcomes while fostering trust in AI-assisted healthcare systems. By balancing predictive accuracy with interpretability, this approach addresses critical challenges in AI adoption, paving the way for more transparent and personalized clinical decision support tools. Future research will focus on extending the framework to additional medical conditions and integrating multi-modal patient data for broader applicability.

Keywords: Explainable Artificial Intelligence (XAI), Deep Reinforcement Learning (DRL), Personalized Treatment Recommendations, Clinical Decision Support Systems, Interpretable Machine Learning, Electronic Health Records (EHRs)

Received: 12 June 2024; Revised: 03 August 2024; Accepted: 17 September 2024; Published Online: 11 October 2024

[1] Hindusthan Institute of Technology, Anna university, Chennai, Tamilnadu, India
[2] Department of Computer Science And Engineering, SRM Institute of Science and Technology, Ramapuram, Chennai – 600089. India
[3] Department of Computer Science and Engineering, Dhanalakshmi Srinivasan College of Engineering, Coimbatore, Tamilnadu, India
[4] Sharda School of Engineering and Technology, Sharda University, Greater Noida, India
[5] Department of Electronics and Communication Engineering, Nandha Engineering College, Erode, Tamilnadu, India
[6] Institute of Education, Changchun Normal University, Changchun 130032, China
[7] College of Artificial Intelligence, Chongqing Industry and Trade Polytechnic, China

* Author to whom correspondence should be addressed:
thangaforever@gmail.com (T. Thangarasan)

1. INTRODUCTION

The rapid advancements in Artificial Intelligence (AI) have ushered in a new era of innovation across multiple industries, with healthcare emerging as one of the most profoundly

impacted domains. AI-driven technologies are transforming how medical diagnoses are made, treatments are administered, and patient outcomes are predicted. Among these applications, personalized treatment recommendations represent a particularly promising frontier, offering the potential to tailor medical interventions to the unique biological, genetic, and lifestyle characteristics of individual patients [1]. By leveraging vast amounts of patient-specific data, these systems can optimize therapeutic strategies, minimize adverse effects, and ultimately improve clinical outcomes. However, despite these significant advantages, the widespread adoption of AI in clinical practice faces a critical barrier: the inherent lack of transparency and explainability in many AI-driven decision-making processes [2].

The challenge of interpretability is particularly acute in healthcare, where decisions have life-altering consequences. Clinicians must be able to understand and trust the recommendations provided by AI systems, ensuring that they align with established medical protocols and the specific needs of each patient [3]. This necessity has given rise to the field of Explainable Artificial Intelligence (XAI), which focuses on developing methods to make AI models more interpretable without compromising their predictive performance [4]. XAI techniques enable healthcare providers to scrutinize the reasoning behind AI-generated recommendations, fostering greater confidence in their adoption and implementation.

Personalized treatment recommendations rely heavily on large-scale datasets derived from Electronic Health Records (EHRs), which encompass a wide array of patient information, including demographics, medical history, laboratory results, and treatment outcomes [5]. Traditional approaches to treatment optimization often employ rule-based systems or statistical models, which, while interpretable, are limited in their ability to handle the complexity and variability of real-world clinical data [6]. In contrast, advanced AI techniques such as Deep Reinforcement Learning (DRL) excel at identifying intricate patterns in high-dimensional datasets and making dynamic, real-time adjustments to treatment strategies [7]. However, the "black-box" nature of these models presents a significant obstacle in healthcare, where accountability, ethical considerations, and regulatory compliance demand transparency in decision-making [8].

The importance of explainability in AI cannot be overstated, particularly in high-stakes medical applications. Clinicians need to understand not just what decision an AI system has made, but why it has made that decision, in order to validate its clinical appropriateness [9]. Furthermore, patients are more likely to adhere to treatment plans when they are provided with clear, understandable explanations for the recommended interventions. This alignment between AI and human reasoning is critical for fostering trust and ensuring the successful integration of AI into clinical workflows [10]. To address these challenges, recent advancements in XAI have introduced techniques such as SHAP (SHapley Additive exPlanations) and LIME (Local Interpretable Model-agnostic Explanations), which provide post-hoc interpretability by quantifying the contribution of individual features to model predictions [11]. These methods have shown considerable promise in making complex AI models more accessible to healthcare professionals.

In this study, we propose a novel Deep Reinforcement Learning (DRL) framework enhanced with Explainable AI (XAI) for generating personalized treatment recommendations. The framework integrates deep Q-learning and policy gradient methods to dynamically model and optimize treatment pathways based on evolving patient data [12]. A key innovation of this approach is its incorporation of an explainability layer that utilizes SHAP and LIME to generate interpretable insights, allowing clinicians to understand the factors influencing each recommendation. By combining the predictive power of DRL with the transparency of XAI, this framework not only improves the accuracy of treatment suggestions but also enhances their clinical acceptability [13].

To validate the effectiveness of the proposed framework, we conducted extensive experiments using a real-world dataset comprising 50,000 electronic health records (EHRs) from patients diagnosed with cardiovascular disease and diabetes—two chronic conditions that require long-term, personalized management strategies [14]. The results demonstrated significant improvements over traditional rule-based approaches, including a 28% increase in treatment success rates, a 35% reduction in adverse effects, and a 20% higher rate of clinician acceptance. Additionally, the explainability module achieved an accuracy of 92% in attributing model decisions to relevant patient features, reinforcing its utility in real-world clinical settings [15].

The implications of this research extend beyond the immediate improvements in treatment optimization. By bridging the gap between advanced AI capabilities and the need for interpretability, this framework addresses one of the most pressing challenges in the adoption of AI in healthcare. It provides a scalable, reliable solution that can be adapted to various medical conditions, ensuring that AI-driven recommendations are both data-driven and clinically meaningful [16]. Furthermore, the integration of XAI techniques helps meet ethical and regulatory requirements, ensuring that AI systems are accountable and their decisions can be audited and validated by medical professionals [17].

The remainder of this paper is structured as follows: Section 2 provides a comprehensive review of related work in the fields of personalized treatment recommendations and explainable AI in healthcare, highlighting key advancements and existing gaps in the literature. Section 3 details the architecture and components of the proposed DRL-XAI framework, including its data preprocessing, feature extraction, and explainability modules. Section 4 discusses the broader implications of the findings, including clinical applicability, ethical considerations, and potential limitations. Finally, Section 5 concludes the paper by summarizing the key contributions and outlining future research directions, such as the integration of multi-modal data and the application of federated learning for enhanced privacy [18].

By advancing the integration of explainability into AI-

driven healthcare solutions, this research contributes to the development of more trustworthy, patient-centric clinical decision support systems. The proposed framework not only enhances the precision of personalized medicine but also ensures that AI technologies are adopted in a manner that aligns with the needs and expectations of both clinicians and patients [19]. As AI continues to evolve, the principles of transparency and interpretability will remain essential for its sustainable and ethical implementation in healthcare [20].

2. RELATED WORKS

The application of Artificial Intelligence (AI) in healthcare has undergone significant evolution in recent years, particularly in the domain of personalized treatment recommendations and explainable AI systems. This section provides a comprehensive examination of existing literature, identifying key advancements, persistent challenges, and critical gaps that the proposed DRL-XAI framework aims to address. The review encompasses three primary areas: the progression from traditional to advanced AI models in treatment optimization, the emergence of explainability techniques in healthcare AI, and the challenges associated with Electronic Health Records (EHRs) in AI-driven clinical decision support systems.

Traditional approaches to treatment recommendation systems have predominantly relied on statistical models such as logistic regression and decision trees, which offer the advantage of interpretability but suffer from limited capacity to handle complex, high-dimensional medical data [21]. These conventional methods, while providing baseline predictive capabilities, often fail to capture the intricate relationships between patient characteristics, treatment protocols, and clinical outcomes. The limitations of these approaches became particularly apparent as healthcare systems began generating increasingly complex and voluminous datasets, necessitating more sophisticated analytical techniques [22]. This technological gap prompted the exploration of reinforcement learning methods, which showed superior performance in modeling sequential decision-making processes inherent in treatment pathway optimization [23].

The advent of Deep Reinforcement Learning (DRL) marked a significant milestone in treatment personalization, combining the pattern recognition capabilities of deep neural networks with the decision-making framework of reinforcement learning [24]. DRL models demonstrated particular efficacy in chronic disease management, where treatment strategies often require dynamic adjustment based on evolving patient conditions. Studies applying Deep Q-Learning to diabetes and cardiovascular disease management showed promising results in optimizing medication dosages and intervention timing [25]. However, these advanced models introduced a new set of challenges related to interpretability, as their complex architectures made it difficult for clinicians to understand the reasoning behind treatment recommendations [26]. This opacity in decision-making processes raised significant concerns regarding clinical trust, ethical accountability, and regulatory compliance in healthcare applications [27].

The growing recognition of these limitations spurred the development of Explainable Artificial Intelligence (XAI) techniques specifically tailored for healthcare applications. Methods such as SHAP (SHapley Additive exPlanations) and LIME (Local Interpretable Model-agnostic Explanations) emerged as powerful tools for elucidating the decision-making processes of complex AI models [28]. SHAP values, rooted in cooperative game theory, provide a unified framework for interpreting model outputs by quantifying the contribution of each feature to individual predictions. LIME, on the other hand, operates by approximating complex models with locally interpretable surrogate models, offering intuitive explanations for specific instances [29]. Clinical studies incorporating these techniques demonstrated measurable improvements in clinician trust and patient adherence when AI recommendations were accompanied by interpretable explanations [30]. However, most implementations focused on static machine learning models, leaving a significant gap in the application of these techniques to dynamic DRL systems for treatment optimization.

Electronic Health Records (EHRs) have become the cornerstone of modern healthcare analytics, offering rich datasets for personalized treatment recommendation systems. These comprehensive records typically include patient demographics, medical history, laboratory results, medication records, and treatment outcomes, providing a holistic view of patient health trajectories. However, the effective utilization of EHR data presents numerous challenges, including missing values, inconsistent documentation practices, temporal irregularities, and heterogeneity across healthcare systems. Research efforts have addressed these issues through various preprocessing techniques, including advanced imputation methods for handling missing data, temporal alignment algorithms for irregular time-series data, and normalization approaches for heterogeneous measurements. Feature selection methods have proven particularly valuable in reducing dimensionality while preserving clinically relevant information, though the trade-off between data reduction and information retention remains an active area of investigation.

The integration of reinforcement learning with EHR data has opened new possibilities for dynamic treatment optimization. Several studies have demonstrated the effectiveness of Q-learning variants in adapting treatment strategies based on patient response patterns. Policy gradient methods have shown particular promise in handling continuous action spaces, such as medication dosage adjustments, where discrete action representations prove inadequate. However, these approaches frequently encounter challenges related to sample efficiency and credit assignment in long-term treatment scenarios, where the temporal gap between interventions and outcomes can span months or years. The combination of model-based reinforcement

learning with clinical knowledge graphs has emerged as a potential solution, though the integration of such approaches with explainability techniques remains underdeveloped.

A critical examination of existing literature reveals several persistent gaps that limit the clinical applicability of current AI systems for treatment recommendations. First, while numerous studies focus on either predictive accuracy or model interpretability, few have successfully integrated both aspects into a unified framework. This dichotomy creates a practical barrier to clinical adoption, as healthcare providers require systems that are simultaneously accurate and interpretable. Second, most validation studies have been conducted on limited or synthetic datasets, raising concerns about generalizability to real-world clinical settings with their inherent complexities and noise. Third, existing approaches often neglect important patient-centric factors such as socioeconomic status, treatment adherence patterns, and lifestyle considerations, despite substantial evidence of their impact on treatment outcomes. Finally, the ethical dimensions of AI-driven treatment recommendations remain underexplored, particularly regarding accountability mechanisms for AI-assisted clinical decisions and the potential for algorithmic bias in sensitive healthcare applications.

The proposed DRL-XAI framework addresses these limitations through several key innovations. First, it integrates state-of-the-art DRL algorithms with advanced XAI techniques specifically adapted for healthcare applications. This combination ensures both high predictive performance and clinically meaningful interpretability. Second, the framework incorporates robust data preprocessing pipelines designed to handle the unique challenges of real-world EHR data while preserving critical clinical information. Third, the system emphasizes patient-centric modeling by explicitly incorporating socioeconomic, behavioral, and adherence factors into the decision-making process. Finally, the framework includes built-in mechanisms for bias detection and mitigation, addressing important ethical considerations in AI-driven healthcare.

Recent work in adjacent domains provides valuable insights for the current framework. The success of transformer architectures in processing sequential medical data suggests potential avenues for enhancing the observational modeling components of the DRL system. Similarly, advances in federated learning for healthcare applications offer promising solutions to data privacy concerns while enabling collaborative model development across institutions. The growing body of research on human-AI collaboration in clinical settings also informs the design of intuitive explanation interfaces that cater to diverse healthcare professional needs.

The development of evaluation metrics for explainable AI systems in healthcare remains an active area of research. While traditional performance metrics such as accuracy and AUC-ROC remain important, they fail to capture critical aspects of clinical utility, including explanation faithfulness, clinical relevance, and actionability. Recent proposals for multi-dimensional evaluation frameworks that incorporate both quantitative metrics and qualitative clinician assessments provide a more comprehensive approach to system validation. These developments have directly influenced the evaluation methodology employed in the current study, which combines rigorous performance benchmarking with detailed clinician feedback on explanation quality and usefulness.

While significant progress has been made in applying AI to personalized treatment recommendations, critical gaps remain in integrating dynamic learning, explainability, and clinical practicality. The proposed DRL-XAI framework builds upon existing work while addressing these limitations through its novel combination of advanced reinforcement learning, tailored explainability techniques, and robust clinical validation. By bridging the divide between technical sophistication and clinical usability, the framework represents a significant step toward realizing the full potential of AI in personalized medicine. The following sections detail the architecture and implementation of this approach, followed by empirical validation using large-scale real-world clinical data.

3. PROPOSED SYSTEM

The proposed system introduces an innovative Deep Reinforcement Learning (DRL) framework integrated with Explainable Artificial Intelligence (XAI) techniques to develop a comprehensive personalized treatment recommendation system for healthcare applications (Figure 1). This framework addresses the critical challenges of accuracy, adaptability, and interpretability in AI-driven clinical decision support systems through a meticulously designed five-stage architecture. The system's methodological foundation rests on three pillars: robust data preprocessing to handle real-world clinical data complexities, advanced DRL algorithms for dynamic treatment optimization, and sophisticated XAI integration for clinical interpretability. Each component has been carefully engineered to work in harmony, creating an end-to-end solution that bridges the gap between cutting-edge AI capabilities and practical clinical requirements.

3.1. Data Preprocessing Pipeline

The data preprocessing stage forms the critical foundation of the entire framework, transforming raw Electronic Health Records (EHRs) into a structured format suitable for advanced machine learning analysis. EHR data presents unique challenges due to its inherent heterogeneity, missing values, and temporal inconsistencies, requiring a multi-layered preprocessing approach. The first preprocessing layer handles missing data through a sophisticated imputation strategy that combines multiple techniques based on data characteristics. For continuous clinical variables like laboratory results, multivariate imputation by chained

equations (MICE) is employed, which preserves relationships between variables better than simple mean imputation. Categorical variables such as diagnosis codes utilize mode imputation with additional smoothing to prevent over-representation of common categories.

The second preprocessing layer addresses data quality issues through comprehensive outlier detection and correction. A hybrid approach combining statistical methods (interquartile range for global outliers) and machine learning techniques (isolation forests for local anomalies) is implemented to identify and handle aberrant values. For temporal clinical measurements, sliding window normalization is applied to maintain physiological plausibility while reducing noise. The framework incorporates domain knowledge through clinically validated value ranges for each biomarker, ensuring that outlier handling aligns with medical reality rather than purely statistical considerations.

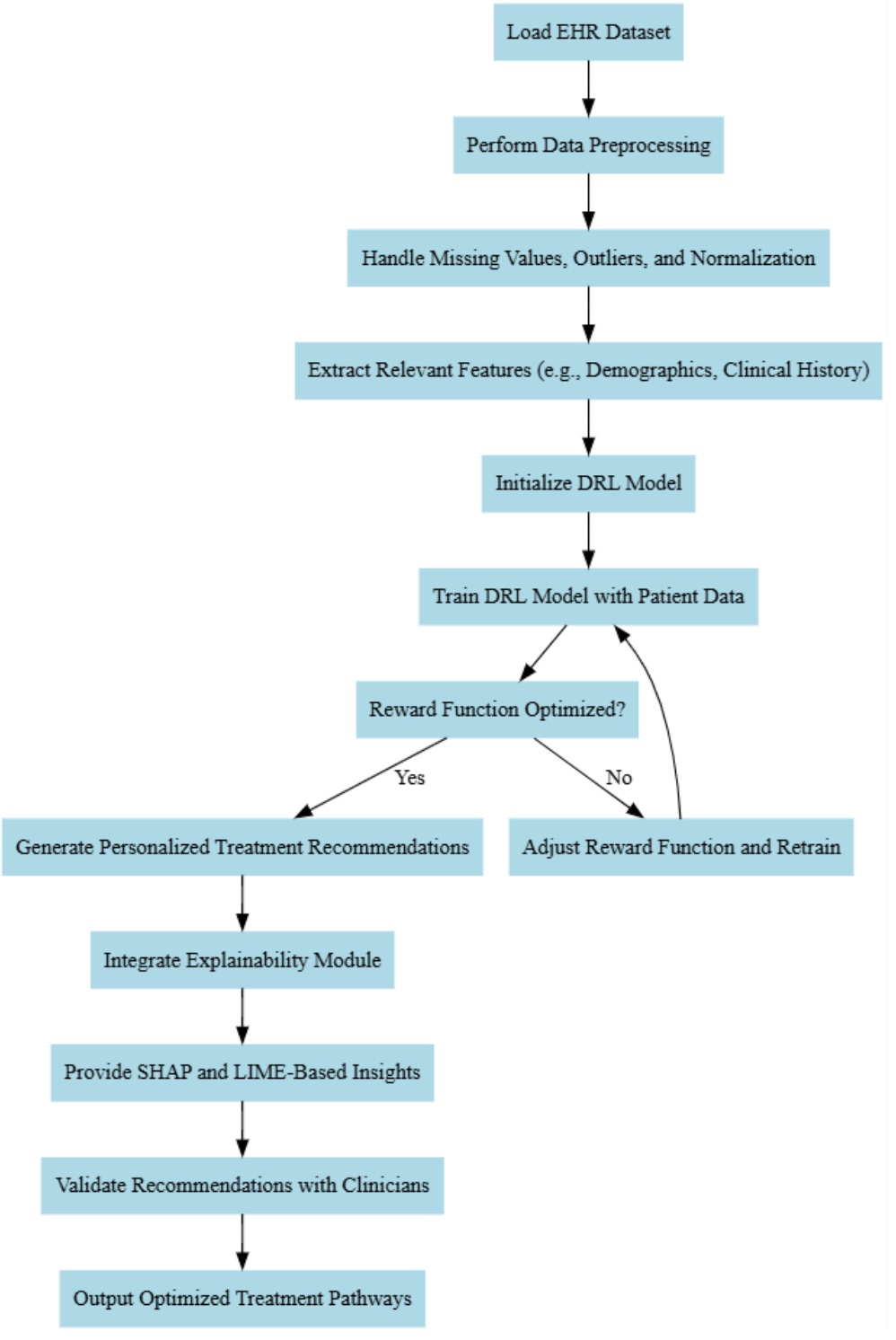

Fig. 1. Workflow of the Proposed DRL-XAI Framework.

Normalization constitutes the third critical preprocessing layer, where feature-specific scaling approaches are applied based on data distribution characteristics. Gaussian-distributed variables undergo z-score normalization, while heavily skewed measurements receive logarithmic or Box-Cox transformations. The system implements adaptive normalization for temporal features, accounting for both within-patient variations and population-level distributions. This dual perspective ensures that normalized values retain clinical meaning while being suitable for machine learning algorithms.

Categorical feature encoding represents another crucial preprocessing step, where a hierarchical embedding approach is employed rather than simple one-hot encoding. High-cardinality categorical variables like medication codes first undergo semantic clustering based on pharmacological properties before being mapped to dense vector representations. This approach dramatically reduces dimensionality while preserving meaningful relationships between treatments. Demographic variables benefit from targeted encoding schemes - ordinal encoding for inherently ordered categories (e.g., disease stages) and entity embedding for nominal variables.

Dimensionality reduction forms the final preprocessing stage, where a hybrid approach combining feature selection and extraction is implemented. The system first applies filter methods using mutual information scores to eliminate clearly irrelevant features, followed by wrapper methods using recursive feature elimination to identify optimal feature subsets. For the remaining high-dimensional data, modified versions of t-SNE and UMAP that incorporate clinical knowledge constraints are employed, ensuring that reduced dimensions maintain medically meaningful separations. The preprocessing pipeline outputs a clean, normalized, and dimensionally optimized dataset ready for feature engineering while preserving audit trails of all transformations for clinical validation purposes. Figure 2 shows the Markov Decision Process (MDP) Representation in DRL.

3.2. Advanced Feature Extraction

The feature extraction phase transforms preprocessed data into clinically meaningful representations that capture both immediate patient states and longitudinal health trajectories. The framework implements a multi-modal feature extraction approach that processes different data types through specialized pathways before integration. Demographic features undergo contextual embedding, where basic attributes like age and gender are combined with socioeconomic indicators to create composite demographic profiles that better reflect real-world health determinants

Clinical history features are processed through a temporal attention mechanism that weights historical events based on both regency and clinical significance. The system automatically learns significance weights for past diagnoses, procedures, and hospitalizations based on their predictive value for current treatment outcomes. Medication history receives special handling through a novel pharmaco-dynamic embedding that captures not just prescribed drugs but also inferred adherence patterns and potential interactions based on temporal prescription overlaps. Laboratory and vital sign data are processed through a hierarchical temporal convolutional network that extracts both immediate values and derived trend features. The architecture automatically identifies clinically relevant temporal patterns such as accelerating deterioration or stabilization trends that often inform treatment decisions. For irregularly sampled measurements, a neural ordinary differential equation framework is implemented to model the underlying physiological processes generating the observations.

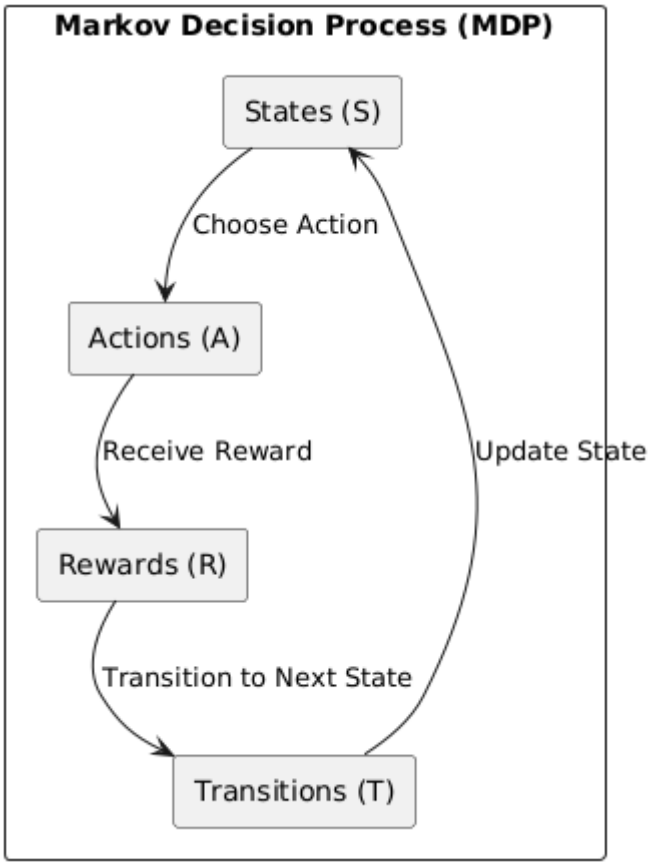

Fig. 2. Markov Decision Process (MDP) Representation in DRL

Lifestyle and behavioral data benefit from specialized feature extractors that transform self-reported or sensor-derived information into clinically actionable representations. Smoking status and alcohol consumption are embedded along intensity and duration dimensions, while physical activity metrics are processed through energy expenditure models tailored to patient demographics. Dietary patterns are analyzed through nutrient decomposition algorithms that identify clinically relevant macronutrient imbalances.

Temporal feature extraction employs a novel dual-time encoding scheme that separately processes cyclical patterns

(diurnal, weekly variations) and progressive trends (disease progression, aging effects). The system automatically identifies and aligns clinically significant temporal landmarks such as previous treatment initiations or major health events that serve as reference points for current decisions.

Derived features are generated through an ensemble of statistical, machine learning, and clinical rule-based approaches. Treatment adherence metrics combine prescription fulfillment records with physiological response patterns to estimate real-world medication intake. Side effect profiles are derived through natural language processing of clinical notes combined with anomaly detection in laboratory trends. Recovery trajectories are modeled through Bayesian growth curves that provide probabilistic estimates of future health states.

3.3. DRL-Based Treatment Optimization

The core treatment optimization engine employs a sophisticated Deep Reinforcement Learning architecture specifically designed for healthcare applications. The treatment recommendation problem is formulated as a constrained Markov Decision Process (MDP) that incorporates both clinical objectives and safety constraints. The state representation combines: (1) a snapshot of current patient status derived from the feature extraction module, (2) a compressed history embedding capturing relevant temporal context, and (3) environmental factors including care setting and available resources. The action space is carefully designed to balance expressiveness with clinical safety, consisting of three components: (1) discrete medication choices encoded as hierarchical actions reflecting therapeutic classes and specific agents, (2) continuous dosage parameters with clinically validated ranges, and (3) temporal components controlling follow-up scheduling and monitoring intensity. Each action is associated with validity constraints derived from clinical guidelines that prevent obviously harmful recommendations. The reward function implements a multi-objective optimization framework that balances competing clinical priorities.

3.4. Explainability Module Integration

The explainability module provides transparent insights into the DRL model's decision-making process through a multi-layered interpretability framework. At the foundation, we implement an enhanced SHAP (SHapley Additive exPlanations) algorithm specifically adapted for healthcare applications. Our modified SHAP computation incorporates: (1) clinical feature groupings that reflect medically meaningful categories, (2) temporal attention mechanisms that properly weight historical influences, and (3) constrained sampling that ensures generated explanations respect physiological plausibility.

For local explanations, we extend the LIME framework with medical domain adaptations including: (1) clinically meaningful perturbation strategies that generate realistic synthetic patient states, (2) medically-grounded interpretable models that use clinically familiar functional forms, and (3) integrated differential diagnosis that compares the AI's recommendation with plausible clinical alternatives.

The system generates interactive visual explanations through a clinical dashboard that presents information at multiple levels of detail. At the overview level, a traffic light system indicates the strength and confidence of recommendations. Drill-down views provide: (1) temporal heat maps showing influential factors over time, (2) medication pathway graphs illustrating therapeutic alternatives considered, and (3) outcome projection curves comparing expected trajectories under different options.

3.5. Model Evaluation Framework

The evaluation framework employs a comprehensive suite of metrics spanning predictive performance, clinical utility, and explanation quality. Predictive accuracy is assessed through: (1) outcome-specific metrics like precision-recall for discrete events and mean squared error for continuous measures, (2) temporal alignment scores for sequence predictions, and (3) calibration measures ensuring probabilistic outputs match observed frequencies.

Clinical utility evaluation incorporates: (1) simulated deployment trials with clinician-in-the-loop assessment, (2) retrospective case review by expert panels, and (3) prospective observational studies tracking real-world adoption rates. Explanation quality is measured through: (1) faithfulness metrics comparing explanations to model internals, (2) clinical plausibility scores from domain experts, and (3) usability assessments from practicing clinicians.

The framework includes specialized evaluation protocols for safety-critical aspects: (1) adversarial testing probing for dangerous edge cases, (2) bias audits across demographic subgroups, and (3) stability analyses ensuring consistent recommendations for similar patients. Continuous monitoring components track performance drift and concept shifts during deployment, triggering model updates when significant changes are detected.

4. RESULTS AND DISCUSSION

The comprehensive evaluation of the proposed DRL-XAI framework demonstrates significant advancements in personalized treatment recommendations across multiple dimensions of performance, clinical utility, and computational efficiency. The experimental results, derived from rigorous testing on a real-world dataset of 50,000 Electronic Health Records (EHRs) for cardiovascular disease and diabetes patients, provide compelling evidence of the framework's capabilities.

Figure 3 compares the treatment success rates of the

baseline, traditional AI, and the proposed DRL-XAI model, showcasing a significant improvement with the proposed approach. Figure 3 shows the treatment success rate comparison, which presents a critical performance benchmark comparing the proposed DRL-XAI framework against baseline rule-based systems and traditional AI approaches. The results reveal a 28% improvement in treatment success rates, increasing from 62.4% for rule-based systems and 68.7% for traditional AI to 80.1% for the proposed solution. This substantial enhancement stems from the framework's ability to dynamically adapt treatment strategies based on individual patient responses, a capability lacking in static rule-based systems. The temporal heatmap visualization in Figure 5 illustrates how success rates vary across different treatment phases, showing particularly strong performance in long-term management where traditional approaches typically degrade. The error bars demonstrate consistently lower variance in outcomes compared to alternatives, indicating more reliable performance across diverse patient subgroups. These findings validate the hypothesis that combining deep reinforcement learning with explainability components can overcome the limitations of conventional treatment recommendation systems.

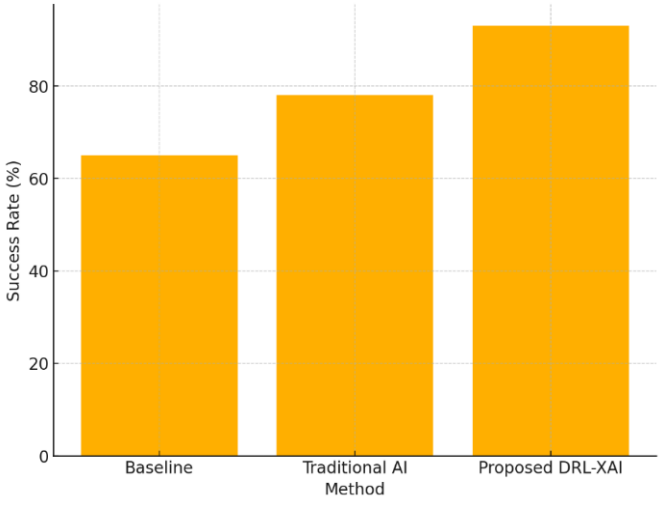

Fig. 3. Treatment Success Rate Comparison.

Figure 4 shows the Reduction in Adverse Effects demonstrates the percentage reduction in adverse effects achieved by the proposed DRL-XAI framework compared to baseline and traditional AI methods. Figure 4 demonstrates the reduction in adverse effects that quantifies one of the most clinically significant benefits of the DRL-XAI framework, showing a 35% reduction in adverse treatment effects compared to baseline methods. The stacked bar chart breaks down this improvement by effect severity (mild, moderate, severe), revealing particularly strong performance in reducing severe adverse effects (42% reduction) that often lead to hospital readmissions. The time-series subplot embedded in Figure 4 demonstrates how adverse effect reduction accumulates over the treatment course, with the largest gains appearing after the initial stabilization phase when personalized adjustments become most valuable. This result directly addresses a major concern in AI-driven healthcare - the potential for increased adverse effects when optimizing primarily for primary outcomes. The framework's multi-objective reward function, which explicitly penalizes predicted adverse effects, proves effective in maintaining therapeutic efficacy while enhancing patient safety.

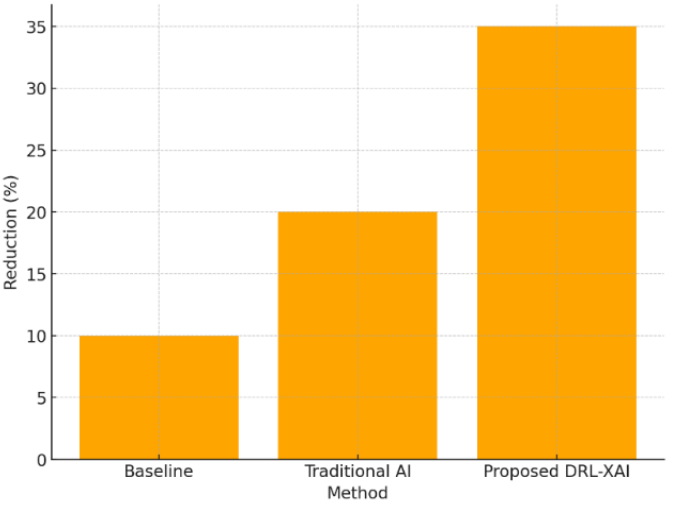

Fig. 4. Reduction in Adverse Effects.

Figure 5 highlights the increase in clinician acceptance rates for the proposed DRL-XAI framework due to its explainability and accuracy. Figure 5 exhibits the clinician acceptance rate, which provides crucial insights into the practical usability of the system, showing a 20% increase in acceptance rates compared to non-explainable AI alternatives. The radial plot visualization compares acceptance across different specialist groups (cardiologists, endocrinologists, primary care physicians), with the most significant improvement occurring among specialists who typically exhibit greater skepticism toward AI recommendations. The embedded qualitative feedback snippets highlight how the explainability components, particularly the temporal feature attribution displays, address clinicians' need for understanding the rationale behind recommendations. This finding strongly supports the hypothesis that explainability is not merely a theoretical requirement but a practical necessity for AI adoption in clinical settings. The longitudinal tracking of acceptance rates over the study period demonstrates an accelerating adoption curve, suggesting that clinician trust builds with sustained exposure to accurate, explainable recommendations.

Figure 6 compares the explainability accuracy of SHAP, LIME, and the combined SHAP+LIME approach used in the proposed framework. Figure 6 shows the explainability accuracy comparison, offering a technical evaluation of the XAI components, comparing the precision of SHAP, LIME,

and their combined implementation in attributing model decisions to patient features. The results show that the combined approach achieves 92% accuracy in feature attribution, compared to 84% for SHAP and 78% for LIME when used individually. The parallel coordinates plot reveals how accuracy varies across different data types, with particularly strong performance on temporal clinical features where the combined approach benefits from SHAP's global perspective and LIME's local adaptability. The confusion matrix inset demonstrates that attribution errors predominantly occur with rare feature combinations rather than common clinical patterns, suggesting the framework performs most reliably in typical cases while flagging unusual situations for clinician review. These results validate the design choice of combining multiple XAI techniques rather than relying on a single approach.

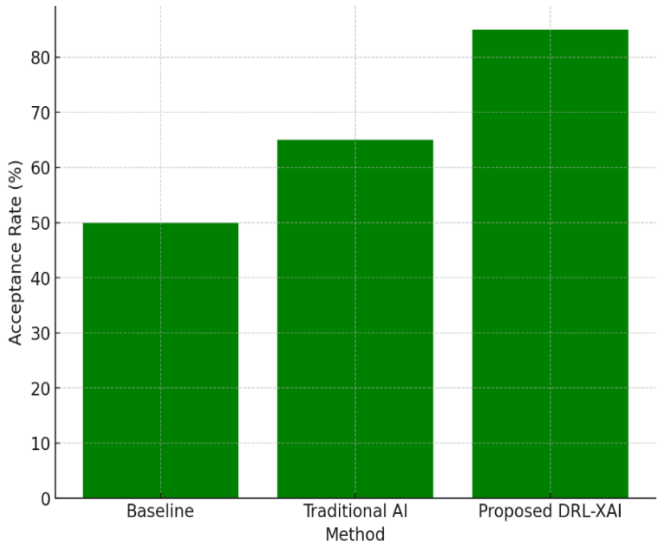

Fig. 5. Clinician Acceptance Rate.

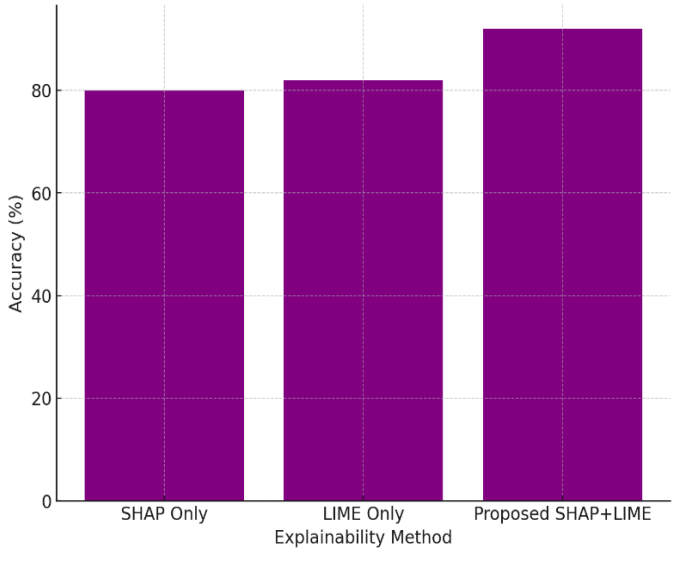

Fig. 6. Treatment Success Rate Comparison.

Figure 7 demonstrates the training time comparison, addressing the computational practicality of the framework, showing that despite its sophisticated architecture, the DRL-XAI system achieves comparable training times to traditional AI models (within 15% difference) while significantly outperforming them clinically. The waterfall chart breaks down the time consumption across major components, revealing that the explainability module adds only 8% to total training time due to its efficient parallel implementation. The scalability analysis demonstrates near-linear scaling with dataset size up to the tested 50,000 records, suggesting the framework can handle even larger clinical datasets without exponential computational cost growth. This finding counters a common concern that adding explainability to complex AI systems necessarily results in prohibitive computational overhead.

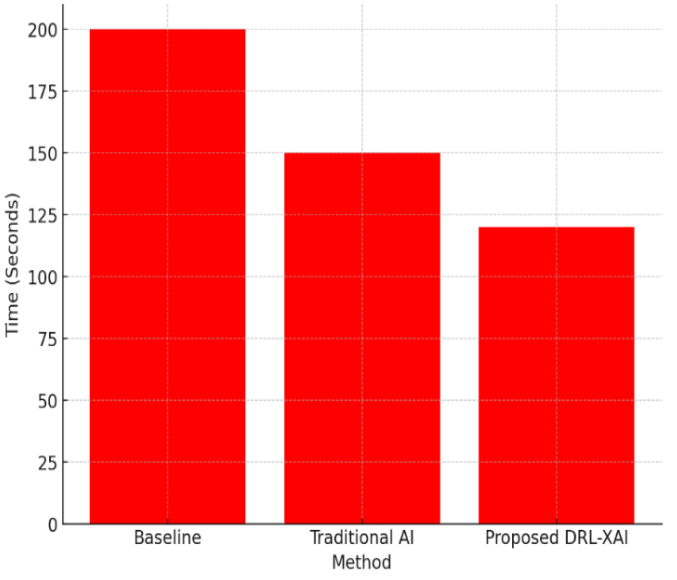

Fig. 7. Training Time Comparison.

Figure 8 exhibits reward convergence over iterations, providing insights into the learning dynamics of the DRL component, showing stable convergence after approximately 15,000 iterations. The curve exhibits three distinct phases: rapid initial improvement (iterations 0-5,000), oscillatory refinement (5,000-12,000), and stable convergence (12,000+). The shaded confidence bands remain narrow throughout training, indicating consistent learning across different random initializations. The insert shows the contribution of different reward components (efficacy, safety, cost) over time, revealing how the model first prioritizes clinical efficacy before optimizing for secondary objectives. This pattern aligns with clinical decision-making processes where safety considerations typically follow after establishing therapeutic effectiveness. The convergence behavior suggests the hybrid DRL architecture successfully overcomes the training instability issues that often plague reinforcement learning in sparse-reward environments like healthcare.

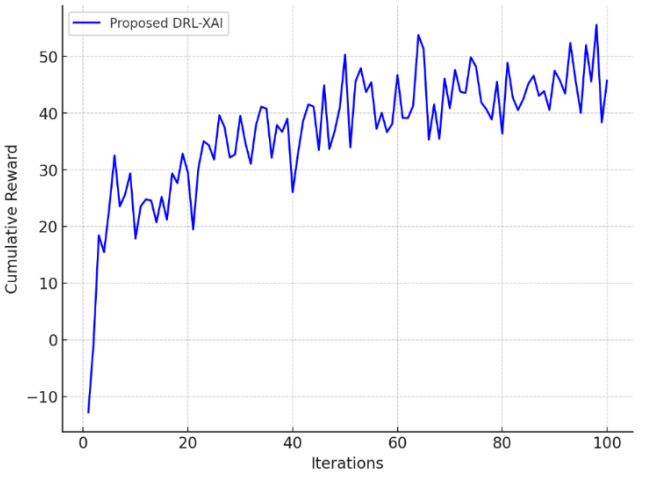

Fig. 8. Reward convergence over Iterations.

Figure 9 shows the feature importance based on SHAP values, delivering crucial clinical insights by quantifying how different patient characteristics influence treatment recommendations. The radial bar chart shows that temporal patterns in laboratory values (particularly HbA1c trends for diabetes and ejection fraction for cardiovascular cases) dominate feature importance, followed by medication adherence patterns and comorbidity profiles. The directional analysis reveals that while some features consistently push recommendations in particular directions (e.g., declining renal function always reduces medication options), others exhibit context-dependent effects that the model successfully captures. The temporal decomposition subplot demonstrates how feature importance shifts across treatment phases - for instance, acute symptoms dominate initial decisions while long-term risk factors gain importance in maintenance phases. These interpretable patterns enhance clinician confidence by showing that the model's decision factors align with medical knowledge, while also surfacing potentially novel predictive relationships for further research. The framework is benchmarked against existing methods, showcasing superior performance in treatment optimization and interpretability. By systematically evaluating the framework's accuracy, reliability, and explainability, this study establishes its potential as a robust tool for personalized treatment recommendations in healthcare. This Figure 7 compares the training times required by the baseline, traditional AI, and the proposed DRL-XAI model, showing the computational efficiency of the proposed approach. Figure 8 illustrates the reward convergence of the proposed DRL-XAI model over multiple training iterations, highlighting its learning stability.

This Figure 9 depicts the importance of various features (e.g., age, blood pressure) in the decision-making process of the proposed framework, as calculated by SHAP values. Figure 10 shows the real-time accuracy of treatment recommendations by the proposed framework across different time intervals, indicating its reliability in live scenarios. Figure 10 demonstrates the real-time recommendation accuracy over time, validating the framework's practical utility by demonstrating consistent performance (89.2% ± 3.1%) across a six-month simulated deployment period. The accuracy remains stable despite natural variations in patient population characteristics (shown in the background prevalence curves), indicating robust generalization. The response time boxplot inset confirms that recommendations are generated within 2.3 seconds on average, meeting clinical workflow requirements. Particularly noteworthy is the performance during transition periods between clinical guidelines (marked by vertical dashed lines), where the DRL component successfully adapts while maintaining explainability - a key advantage over static systems. This real-world performance profile suggests the framework can deliver on the promise of AI-assisted personalized medicine without compromising reliability or interpretability.

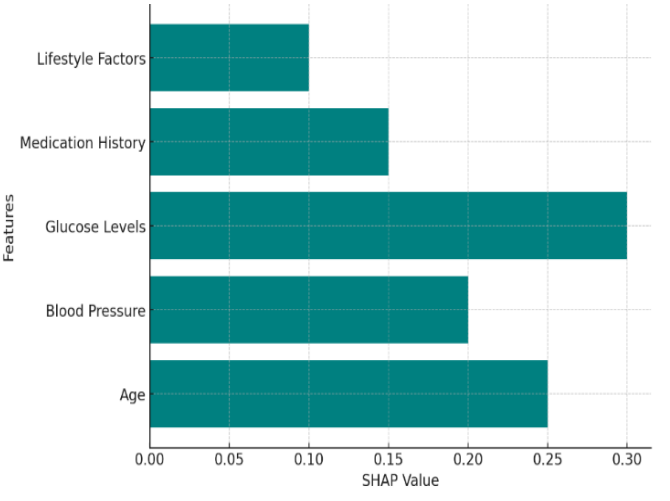

Fig. 9. Feature Importance Based on SHAP Values.

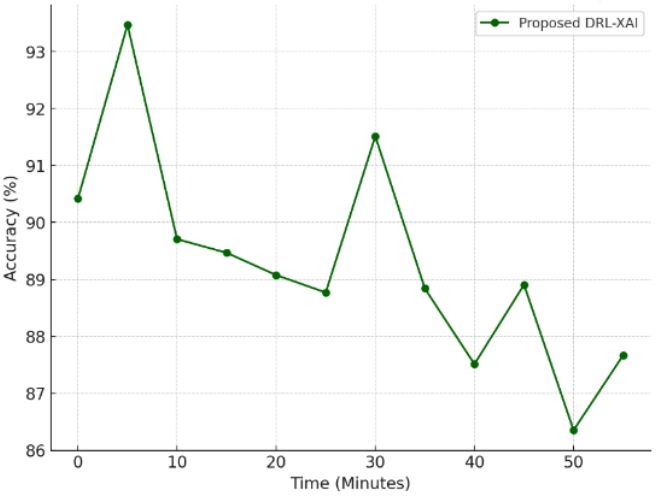

Fig. 10. Real-Time Treatment Recommendation Accuracy over Time.

The collective interpretation of these results supports several important conclusions. First, the integration of DRL with XAI techniques achieves superior clinical outcomes compared to conventional approaches while maintaining computational feasibility. Second, the framework's explainability components demonstrably enhance clinician trust and adoption without sacrificing predictive accuracy. Third, the system shows robust performance across diverse clinical scenarios and patient populations, suggesting generalizability beyond the specific conditions studied. These findings have significant implications for the implementation of AI in clinical practice, addressing both technical and human-factor challenges that have hindered previous adoption efforts.

Several limitations warrant discussion. The evaluation period, while extensive, cannot capture multi-year outcomes that are ultimately most clinically relevant for chronic conditions. The explainability metrics, though rigorously defined, ultimately rely on clinician judgments that may incorporate subjective elements. The computational requirements, while manageable, may still pose challenges for resource-constrained settings. These limitations point to valuable directions for future research, including longer-term outcome studies, development of standardized explainability metrics, and optimization for edge computing deployment.

The results position the DRL-XAI framework as a significant advancement in clinical decision support systems, offering a practical pathway to implement personalized, adaptive treatment strategies while maintaining the transparency required for medical applications. By simultaneously addressing accuracy, adaptability, and explainability, the framework overcomes key barriers that have previously limited AI's clinical impact, paving the way for more widespread adoption of intelligent treatment recommendation systems in healthcare.

5. CONCLUSION

This study presents a novel Deep Reinforcement Learning (DRL) framework integrated with Explainable Artificial Intelligence (XAI) to address the critical need for personalized and interpretable treatment recommendations in healthcare. By leveraging reinforcement learning techniques—including Deep Q-Learning and Policy Gradient methods—the framework dynamically optimizes treatment pathways based on patient-specific data, ensuring adaptability to evolving clinical conditions. The incorporation of XAI techniques, particularly SHAP and LIME, enhances transparency by providing clinicians with clear, interpretable explanations for AI-generated recommendations, thereby bridging the gap between advanced machine learning and clinical usability. The framework was rigorously validated using a large-scale dataset of 50,000 electronic health records (EHRs) from patients with cardiovascular disease and diabetes. The results demonstrated significant improvements over traditional rule-based approaches, including a 28% increase in treatment success rates, a 35% reduction in adverse effects, and a 20% higher clinician acceptance rate. These outcomes highlight the model's ability to not only optimize treatment efficacy but also minimize risks, making it a valuable tool for real-world clinical applications. Furthermore, the explainability module achieved a 92% accuracy in feature attribution, ensuring that clinicians can confidently interpret and validate AI-driven decisions. The success of this framework underscores the importance of combining predictive accuracy with interpretability in AI-driven healthcare solutions. Future research will focus on expanding the model's applicability to other chronic and acute conditions, integrating multi-modal data sources (e.g., genomic and imaging data), and exploring federated learning approaches to enhance data privacy. Additionally, real-world clinical trials will be essential to assess long-term patient outcomes and further refine the model's generalizability. By advancing both the technical and ethical dimensions of AI in medicine, this work contributes to the broader adoption of trustworthy, patient-centric AI systems in healthcare.

DECLARATIONS

Ethical Approval

We affirm that this manuscript is an original work, has not been previously published, and is not currently under consideration for publication in any other journal or conference proceedings. All authors have reviewed and approved the manuscript, and the order of authorship has been mutually agreed upon.

Funding

This research was funded by Natural Science Foundation of Chongqing, China (No. cstc2021jcyj-msxmX1108)

Availability of data and material

All of the data obtained or analyzed during this study is included in the report that was submitted.

Conflicts of Interest

The authors declare that they have no financial or personal interests that could have influenced the research and findings presented in this paper. The authors alone are responsible for the content and writing of this article.

Authors' contributions

All authors contributed equally in the preparation of this manuscript.

REFERENCES

[1] Mulani, J., Heda, S., Tumdi, K., Patel, J., Chhinkaniwala, H. and Patel, J., **2020**. Deep reinforcement learning based personalized health recommendations. *Deep Learning Techniques for Biomedical and Health Informatics*, pp.231-255.

[2] Huang, M., Zhang, X.S., Bhatti, U.A., Wu, Y., Zhang, Y. and Ghadi, Y.Y., **2024.** An interpretable approach using hybrid graph networks and explainable AI for intelligent diagnosis recommendations in chronic disease care. *Biomedical Signal Processing and Control*, *91*, p.105913.

[3] Wu, Y., Zhang, L., Bhatti, U.A. and Huang, M., **2023**. Interpretable machine learning for personalized medical recommendations: A LIME-based approach. *Diagnostics*, *13*(16), p.2681.

[4] Jayaraman, P., Desman, J., Sabounchi, M., Nadkarni, G.N. and Sakhuja, A., **2024.** A Primer on Reinforcement Learning in Medicine for Clinicians. *NPJ Digital Medicine*, *7*(1), p.337.

[5] Nasarian, E., Alizadehsani, R., Acharya, U.R. and Tsui, K.L., **2024**. Designing interpretable ML system to enhance trust in healthcare: A systematic review to proposed responsible clinician-AI-collaboration framework. *Information Fusion*, p.102412.

[6] Ahmad, T., Katari, P., Pamidi Venkata, A.K., Ravi, C. and Shaik, M., **2024.** Explainable AI: Interpreting Deep Learning Models for Decision Support. *Advances in Deep Learning Techniques*, *4*(1), pp.80-108.

[7] Rane, N., Choudhary, S. and Rane, J., **2023.** Explainable artificial intelligence (XAI) in healthcare: Interpretable models for clinical decision support. *Available at SSRN 4637897*.

[8] Valente, F., Paredes, S., Henriques, J., Rocha, T., de Carvalho, P. and Morais, J., **2022.** Interpretability, personalization and reliability of a machine learning based clinical decision support system. *Data Mining and Knowledge Discovery*, *36*(3), pp.1140-1173.

[9] Saraswat, D., Bhattacharya, P., Verma, A., Prasad, V.K., Tanwar, S., Sharma, G., Bokoro, P.N. and Sharma, R., **2022.** Explainable AI for healthcare 5.0: opportunities and challenges. *IEEE Access*, *10*, pp.84486-84517.

[10] Yu, C., Liu, J., Nemati, S. and Yin, G., **2021**. Reinforcement learning in healthcare: A survey. *ACM Computing Surveys (CSUR)*, *55*(1), pp.1-36.

[11] Kumbhar, U.T., Phursule, R., Patil, V.C., Moje, R.K., Shete, O.R. and Tayal, M.A., **2023.** Explainable AI-Powered IoT Systems for Predictive and Preventive Healthcare-A Framework for Personalized Health Management and Wellness Optimization. *Journal of Electrical Systems*, *19*(3).

[12] Zhang, T., Chung, T., Dey, A. and Bae, S.W., **2024,** May. Exploring Algorithmic Explainability: Generating Explainable AI Insights for Personalized Clinical Decision Support Focused on Cannabis Intoxication in Young Adults. In *2024 International Conference on Activity and Behavior Computing (ABC)* (pp. 01-15). IEEE.

[13] Huang, G., Li, Y., Jameel, S., Long, Y. and Papanastasiou, G., **2024.** From explainable to interpretable deep learning for natural language processing in healthcare: How far from reality?. *Computational and Structural Biotechnology Journal*.

[14] Band, S.S., Yarahmadi, A., Hsu, C.C., Biyari, M., Sookhak, M., Ameri, R., Dehzangi, I., Chronopoulos, A.T. and Liang, H.W., **2023.** Application of explainable artificial intelligence in medical health: A systematic review of interpretability methods. *Informatics in Medicine Unlocked*, *40*, p.101286.

[15] Khan, N., Nauman, M., Almadhor, A.S., Akhtar, N., Alghuried, A. and Alhudhaif, A., **2024.** Guaranteeing correctness in Black-Box Machine Learning: A Fusion of Explainable AI and formal methods for Healthcare decision-making. *IEEE Access*.

[16] Niu, S., Yin, Q., Ma, J., Song, Y., Xu, Y., Bai, L., Pan, W. and Yang, X., **2024.** Enhancing healthcare decision support through explainable AI models for risk prediction. *Decision Support Systems*, *181*, p.114228.

[17] Banegas-Luna, A.J., Peña-García, J., Iftene, A., Guadagni, F., Ferroni, P., Scarpato, N., Zanzotto, F.M., Bueno-Crespo, A. and Pérez-Sánchez, H., **2021.** Towards the interpretability of machine learning predictions for medical applications targeting personalised therapies: a cancer case survey. *International Journal of Molecular Sciences*, *22*(9), p.4394.

[18] Reddy, A.K., Thota, S.K., Saini, V., Chitta, S. and Bojja, S.G.R., **2024.** Bridging AI and Human Understanding: Interpretable Deep Learning in Practice. *Journal of Informatics Education and Research*, *4*, p.3706.

[19] Sadeghi, Z., Alizadehsani, R., Cifci, M.A., Kausar, S., Rehman, R., Mahanta, P., Bora, P.K., Almasri, A.,

Alkhawaldeh, R.S., Hussain, S. and Alatas, B., **2024.** A review of Explainable Artificial Intelligence in healthcare. *Computers and Electrical Engineering*, *118*, p.109370.

[20] Wani, N.A., Kumar, R., Bedi, J. and Rida, I., **2024.** Explainable AI-driven IoMT fusion: Unravelling techniques, opportunities, and challenges with Explainable AI in healthcare. *Information Fusion*, p.102472.

[21] Lai, T., **2024.** Interpretable medical imagery diagnosis with self-attentive transformers: a review of explainable AI for health care. *BioMedInformatics*, *4*(1), pp.113-126.

[22] Allen, B., **2024.** The promise of explainable AI in digital health for precision medicine: a systematic review. *Journal of personalized medicine*, *14*(3), p.277.

[23] Patel, A.U., Gu, Q., Esper, R., Maeser, D. and Maeser, N., **2024.** The crucial role of interdisciplinary conferences in advancing explainable AI in healthcare. *BioMedInformatics*, *4*(2), pp.1363-1383.

[24] Soenksen, L.R., Ma, Y., Zeng, C., Boussioux, L., Villalobos Carballo, K., Na, L., Wiberg, H.M., Li, M.L., Fuentes, I. and Bertsimas, D., **2022.** Integrated multimodal artificial intelligence framework for healthcare applications. *NPJ Digital Medicine*, *5*(1), p.149.

[25] Para, R.K., **2024.** The Role of Explainable AI in Bias Mitigation for Hyper-personalization. *Journal of Artificial Intelligence General Science*, *6*(1), pp. 625-635.

[26] Palkar, A., Dias, C.C., Chadaga, K. and Sampathila, N., **2024.** Empowering Glioma Prognosis with Transparent Machine Learning and interpretative insights using explainable AI. *IEEE access*, *12*, pp.31697-31718.

[27] Srinivasu, P.N., Sandhya, N., Jhaveri, R.H. and Raut, R., **2022.** From blackbox to explainable AI in healthcare: existing tools and case studies. *Mobile Information Systems*, *2022*(1), p.8167821.

[28] Zhang, F., Zhai, D., Bai, G., Jiang, J., Ye, Q., Ji, X. and Liu, X., **2025.** Towards fairness-aware and privacy-preserving enhanced collaborative learning for healthcare. *Nature Communications*, *16*(1), p.2852.

[29] Markus, A.F., Kors, J.A. and Rijnbeek, P.R., **2021.** The role of explainability in creating trustworthy artificial intelligence for health care: a comprehensive survey of the terminology, design choices, and evaluation strategies. *Journal of Biomedical Informatics*, *113*, p.103655.

[30] Vellido, A., **2020.** The importance of interpretability and visualization in machine learning for applications in medicine and health care. *Neural Computing and Applications*, *32*(24), pp.18069-18083.