Adversarial Attacks on Reinforcement Learning-based Medical Questionnaire Systems

Input-level Perturbation Strategies and Medical Constraint Validation

Peizhuo Liu

Pioneer Academics Computer Science

lewisliu819@outlook.com

July 29, 2025

Abstract

RL-based medical questionnaire systems have shown great potential in medical scenarios. However, their safety and robustness remain unresolved. This study performs a comprehensive evaluation on adversarial attack methods to identify and analyze their potential vulnerabilities. We formulate the diagnosis process as a Markov Decision Process (MDP), where the state is the patient responses and unasked questions, and the action is either to ask a question or to make a diagnosis.

We implemented six prevailing major attack methods, including the Fast Gradient Signed Method (FGSM), Projected Gradient Descent (PGD), Carlini & Wagner Attack (C&W) attack, Basic Iterative Method (BIM), DeepFool, and AutoAttack, with seven epsilon values each.

To ensure the generated adversarial examples remain clinically plausible, we developed a comprehensive medical validation framework consisting of 247 medical constraints, including physiological bounds, symptom correlations, and conditional medical constraints. We achieved a 97.6% success rate in generating clinically plausible adversarial samples.

We performed our experiment on the National Healthcare Interview Survey (NHIS)[1] dataset, which consists of 182,630 samples, to predict the participant's 4-year mortality rate. We performed our evaluation on the AdaptiveFS framework proposed by Shaham et al.[2]. Our results show that adversarial attacks could significantly impact the diagnostic accuracy, with attack success rates ranging from 33.08% (FGSM) to 64.70% (AutoAttack).

Our work has demonstrated that even under strict medical constraints on the input, such RL-based medical questionnaire systems still show significant vulnerabilities.

1. Introduction

The application of artificial intelligence (AI) in healthcare has significantly transformed the medical diagnosis protocol [3], achieving remarkable success in applications ranging from medical imaging classification [4] to clinical decision support systems [5]. Among these applications, the application of reinforcement learning-based adaptive questionnaire systems are gaining increased attention [2, 6]. Such systems formulate the diagnosis process as a Markov Decision Process (MDP) [7], where the state represents the current knowledge about the patient (previous answers and unasked questions), and action space is either to ask a question or to give a diagnostic decision. By dynamically choosing the most informative question given the current state, these systems could reduce the length of the questionnaire while maintaining high diagnostic accuracy [8, 9].

However, with the transition of such systems from research prototypes to clinical deployment, a shortage in our current understanding of their vulnerabilities has emerged. The deployment of such systems in healthcare applications requires a comprehensive evaluation of their robustness against adversarial attacks to ensure the security of such models, and therefore ensuring patient safety [10, 11]. Adversarial attacks involve generating carefully crafted input perturbations, causing the model to misclassify or perform unexpected behaviors[12, 13]. The consequences of successful adversarial attacks in medical applications could be severe, causing delayed or incorrect medical treatments by doctors, threatening patient safety [14, 15].

Adversarial attacks on dynamic medical questionnaire systems have a subtle difference from attacks on medical image classification systems. Attacks on images are often imperceptible by human vision since they only perform pixel-level consecutive perturbations. However, adversarial samples for questionnaire systems are generated by manipulating discrete numerical data, which could easily be detected with simple medical constraint validations. Therefore, they must remain within the validation constraints and clinically plausible to avoid being detected.

The sequential decision-making nature of RL-based questionnaire systems also introduces new attack vectors that were not present in traditional image classification tasks. Adversarial perturbations targeting such systems could not only influence the final diagnostic output, but also the reward computation process throughout the episode, leading the model to suboptimal questioning policies and leaving out important symptoms. To the best of our knowledge, this critical vulnerability has not been systematically discussed in previous research on medical AI security yet.

1.1. Research Motivation and Objectives

Our research aims to solve the lack of security analysis for the growing number of possible RL-based questionnaire systems deployed in clinical settings. Existing guidelines and evaluation frameworks of medical AI systems mainly focus on performance metrics (accuracy, recall, F-1 score, etc.), with only little attention on the adversarial robustness of systems[16].

In this paper, we bridge this critical gap by performing the first comprehensive study on the adversarial vulnerabilities of RL-based medical questionnaire systems. We explore worst-case scenarios to inform the development of defense mechanisms and new regulations. We mainly focus on white-box attacks that pose the most threat, in which the attackers have full knowledge on the model, including model structure, parameters, and gradient.

1.2. Key Contributions

We provide the first systematic evaluation of adversarial attack methods on RL-based questionnaire systems. We adapted and implemented six major white-box attack methods, including gradient-based attacks (FGSM, PGD, BIM), which are fast, single-step and iterative optimization methods; Optimization-based attacks (C&W, DeepFool), which use advanced optimization techniques that find minimal perturbations to change model decision; and ensemble attacks (AutoAttack), which combine multiple attack methods to achieve maximum effectiveness, representing sophisticated adversarial scenarios.

We also proposed a novel medical constraint framework to ensure the generated adversarial samples remain clinically plausible, including 247 constraint rules across 5 categories derived from standardized clinical knowledge [17, 18, 19, 20]. This addresses a critical limitation identified by Croce et al. [21] that adversarial examples often violated basic medical principles.

We performed our experiment on the AdaptiveFS framework [2], a state-of-the-art RL-based adaptive questionnaire system using the National Health Interview Survey (NHIS) dataset with 182,630 observations on the task of 4-year mortality prediction. Our statistical analysis demonstrates that there are critical vulnerabilities underlying the system.

We provide a detailed analysis of our experiment results, and the implications on clinical deployment of such systems and future research. Our findings reveal critical vulnerabilities that should be resolved before the deployment of such systems in clinical settings.

1.3. Paper Organization and Structure

We organize the remainder of this paper as follows. Section 2 first provides a review of related works, covering adversarial attacks on machine learning and reinforcement learning systems, defense methods for such attacks, vulnerabilities in medical image classification and attack methods specified for medical applications. Section 3 then presents our methodology, including theoretical foundations, attack method implementations, and the medical constraint framework. Sections 4 and 5 detail our experimental setup and implementation. Section 6 presents the results and statistical analysis of our experiment. Section 7 further analyzes the implications and limitations of our work. Section 8 concludes this paper, providing a high-level review and emphasizes the need for an enhanced evaluation framework for adversarial robustness.

2. Related Works

2.1. Adversarial Attacks on Machine Learning Systems

Early work by Szegedy et al. [22] highlighted the vulnerabilities of deep neural networks (DNNs) to adversarial attacks. They demonstrated that minor perturbations, imperceptible to human, can cause the model to misclassify with a high confidence level. This is often regarded as a foundational work on adversarial attacks. Different attack methods were then proposed later and classified into three main categories based on the level of prior knowledge required on the model: white-box, gray-box and black-box attacks. We focus on white-box and black-box attacks in remaining sections.

2.1.1. White-box Attacks

We start with white-box attacks, which assume the attacker has complete prior knowledge of the targeted model, including model architecture, training data, hyperparameters, and thus gradients. Such

methods mainly attack the targeted model utilizing the computed gradients.

However, we mainly examine white-box attacks that do not require prior knowledge of the full training dataset, as training data and training processes for medical diagnosis systems often involve restricted-use datasets that are hard for attackers to access.

Goodfellow et al. [23] first proposed the Fast Gradient Signed Method (FGSM) in their 2014 work. Their method is described as follows:

$$x_{adv} = x + \epsilon \cdot \text{sign}(\nabla_x J(\theta, x, y)) \tag{1}$$

Here, ϵ represents the perturbation strength, and J is the loss function. Madry et al. [24] proposed the Projected Gradient Descent (PGD) method as a subsequent work for this, which achieves better attack results through multi-step iterative optimization.

2.1.2. Black-Box Attacks

Black-box attacks require no prior knowledge of the model and thus resemble real-world attack scenarios better.

Fundamental work proposed by Papernot et al [25] has shown that attackers could use only the model's predicted labels (no gradients required) from the target classifier to train a surrogate model, then use this model to generate adversarial data samples. These samples could then be transferred to the original model with a high success rate. In their experiment on DNN APIs, over 84% of the generated adversarial inputs misled the model, again proving that the query-driven surrogate attack strategy could replicate the efficiency of white-box attacks.

This method, often described as the *"Recon-Surrogate-Exploit-Deploy"* pipeline, is now the most prevalent strategy for black-box attacks. Liu et al. [26] further demonstrated the validity and effectiveness of this method by proving the **transferability** of such attacks on models with different architectures. Given two different models f_1 and f_2, they proved that if an adversarial sample δ could cause the model f_2 to misclassify, then, conditioned on f_2 being fooled, f_1 is even more likely to err, as described in equation 2.

$$\mathbb{P}[f_1(x+\delta) \neq y \mid f_2(x+\delta) \neq y] > \mathbb{P}[f_1(x+\delta) \neq y] \tag{2}$$

Subsequent works [27, 28] focused on improving the query efficiency of such methods. Chen et al. [29] proposed Zeroth Order Optimization (ZOO) in their 2017 work. They showed that given efficient queries, black-box attacks could match the performance of white-box attacks. They expressed the query efficiency as:

$$Q(\epsilon, \delta) = O\left(\frac{d}{\epsilon^2} \log \frac{1}{\delta}\right) \tag{3}$$

where $Q(\epsilon, \delta)$ is the number of queries needed to reach the intended accuracy given the failure probability, d is the input dimension, ϵ is the desired accuracy, and δ is the failure probability.

2.2. Adversarial Attacks for Reinforcement Learning Systems

Huang et al. [30] were the first to show that adversarial attacks are effective when targeting neural network policies in reinforcement learning. They showed that adversarial techniques could be used to generate examples that can negatively impact the performance of trained network policies on testing datasets. We classify the attack methods specified for reinforcement learning systems into two classes:

state perturbation attacks and environmental manipulations

2.2.1. State Perturbation Attacks

Lin et al. [31] proposed two tactics, namely the strategically-timed attack and the enchanting attack. They demonstrated that small perturbations at critical decision points could lead the RL agents into sub-optimal trajectories. Their work pointed out a critical aspect of adversarial attacks: the timing of adversarial attacks matters as much as their magnitude.

2.2.2. Environment Manipulation and Adversarial Policies

Gleave et al. [32] further proved that it is possible to attack an RL agent, simply by choosing an adversarial policy in zero-sum games, even against victims trained via self-play to be robust to opponents.

Zhao et al. [33] used a sequence-to-sequence model to predict a single or sequence of future actions that the targeted agent would make. Their approach is a strong black-box attack method. It does not require the attacker to have any prior knowledge of the model, including training parameters and gradients.

2.3. Defense Methods for Reinforcement Learning Systems

The amount of research in adversarial attack methods cultivated the research of defense methods. Inspired by Langevin dynamics, Kamalaruban et al. [34] proposed a method as described in Equation 4:

$$\theta_{t+1} = \theta_t - \eta \nabla_\theta \mathcal{L}(\theta_t) + \sqrt{2\eta\tau}\,\epsilon_t \tag{4}$$

where $\epsilon_t \sim \mathcal{N}(0, I)$ and τ is the temperature parameter.

This method is an instance of Stochastic Gradient Langevin Dynamics (SGLD) [35], which combines stochastic gradient descent with Gaussian noise injection. Applying Langevin noise encourages the optimization to explore flatter regions of the loss landscape.

Zhang et al. [36] have further proven that such flatter minima correlate with better generalization and more robustness to perturbations. By helping the model to escape sharp local minima and sample from a wider posterior distribution, SGLD can improve the model's resilience to adversarial attacks [34].

2.4. Vulnerabilities in Medical Image Classification

Finlayson et al. [10] demonstrated that it is feasible to generate adversarial attacks against medical machine learning systems. They showed that even highly accurate medical classifiers can misclassify by carefully crafted adversarial examples. They evaluated both white-box and black-box attack methods on a diabetic detection system. The results achieved significant attack success rates, while the attack samples remained imperceptible by human visual.

Ma et al. [37] conducted further analysis on this. They compared adversarial attacks on medical images to that of natural images, and found that medical images are significantly more vulnerable to adversarial attacks. They proposed that this increased vulnerability originated from two factors:

1. The complexity and high frequency of features in medical images could create regions in the loss landscape that are more sensitive to small perturbations
2. The neural networks were mainly designed for natural image processing. After being adapted to medical imaging tasks, it may be overparameterized, resulting in suboptimal loss landscapes.

2.5. Medical Domain-Specific Attack Methods

Several attack methods and frameworks specialized in medical diagnosis systems have been proposed in recent years. Ozbulak et al. proposed the AMSA method for attacks on medical image segmentation models [38]. Yao et al. [39] introduced the Hierarchical Feature Constraint (HFC) to craft adversarial samples which are imperceptible to human within normal feature space. Qi et al. [40] proposed the Stabilized Medical Image Attack (SMIA) method that generates adversarial examples out of non-adversarial ones by iteratively maximizing the deviation loss and minimizing stabilization terms.

3. Methodology

3.1. Problem Formulation

We consider an RL-based medical questionnaire system modeled as a Markov Decision Process (MDP) [7] defined by the tuple (S, A, P, R, γ), where:

- S: State space representing patient responses and unasked questions
- A: Action space consisting of questions to ask or diagnostic decisions
- P: Transition probability function
- R: Reward function encouraging accurate diagnosis with minimal questions
- γ: Discount factor

The state at time t is represented as $s_t = [x_t, m_t] \in \mathbb{R}^{2d}$, where:

- $x_t \in \mathbb{R}^d$: Patient feature vector (responses to asked questions)
- $m_t \in \{0, 1\}^d$: Binary mask indicating which questions have been asked

We formulate the adversarial attack problem as finding a perturbation δ that satisfies:

$$\begin{aligned} \max_{\delta} \quad & \mathcal{L}(f_\theta(x + \delta), y_{target}) \\ \text{s.t.} \quad & \|\delta\|_p \leq \epsilon \\ & (x + \delta) \in \mathcal{C}_{medical} \end{aligned} \tag{5}$$

here, f_θ is the diagnostic model, y_{target} is the adversarial target, and $\mathcal{C}_{medical}$ represents medical constraints.

3.2. Attack Methods

3.2.1. Fast Gradient Sign Method (FGSM)

We adapted the FGSM method proposed by Goodfellow et al. [23] by computing gradients with respect to the patient's feature vector while maintaining the masked structure as demonstrated:

The theoretical explanation for FGSM's effectiveness relies on the linear hypothesis. For a linear model with parameters w, the adversarial perturbation maximizes $w^T\delta \quad \text{subject to} \quad \|\delta\|_\infty \leq \epsilon$.

The optimal solution to the previously mentioned problem is $\delta = \epsilon \cdot \text{sign}(w)$, which generalizes to nonlinear models through first-order Taylor approximation:

Algorithm 1 *FGSM for Medical Questionnaires*

Input: Patient features x, target y_{target}, perturbation bound ϵ
Output: Adversarial example x_{adv}
1: Construct state $s = [x, m]$ where m is the question mask
2: Compute loss $\mathcal{L} = -\log p(y_{target}|s)$
3: Calculate gradient $g = \nabla_x \mathcal{L}$
4: Generate perturbation $\delta = \epsilon \cdot \text{sign}(g)$
5: Apply medical constraints: $\delta' = \Pi_{\mathcal{C}_{medical}}(\delta)$
6: **return** $x_{adv} = x + \delta'$

$$\mathcal{L}(x + \delta) \approx \mathcal{L}(x) + \nabla_x \mathcal{L}(x)^T \delta \tag{6}$$

3.2.2. Projected Gradient Descent (PGD)

The PGD method [24] extends on the FGSM through iterative optimization with projection:

$$x_{t+1} = \Pi_{\mathcal{B}_\epsilon(x) \cap \mathcal{C}_{medical}} \left(x_t + \alpha \cdot \text{sign}(\nabla_x \mathcal{L}(x_t, y_{target})) \right) \tag{7}$$

Here, Π denotes the projection, $\mathcal{B}_\epsilon(x)$ is the ϵ-ball around x, and α is the step size.

The convergence of PGD can be analyzed using its framework. For a convex loss function $\mathcal{L}$ with Lipschitz continuous gradient (Lipschitz constant L), the convergence rate could be represented as:

$$\mathcal{L}(x_T) - \mathcal{L}(x^*) \leq \frac{\|x_0 - x^*\|^2}{2\alpha T} + \frac{\alpha L}{2} \tag{8}$$

where T is the number of iterations and x^* is the optimal solution. Therefore, the optimal step size is $\alpha = 1/L$, yielding the convergence rate $O(1/T)$.

3.2.3. Carlini & Wagner (C&W) Attack

We implemented an enhanced Carlini & Wagner (C&W) Attack [41] using tanh-space optimization to naturally bound perturbations:

$$\begin{aligned} \min_{w} \quad & \|x - \tanh(w)\|_2^2 + c \cdot f(x) \\ \text{where} \quad & f(x) = \max(\max_{i \neq t} Z(x)_i - Z(x)_t, -\kappa) \end{aligned} \tag{9}$$

Here, $Z(x)$ represents the logits before softmax, t is the target class, and κ controls confidence.

3.2.4. Additional Attack Methods

We also implemented additional attack methods for comprehensive evaluation:

- **BIM**: Basic Iterative Method, which aims to improve attack success rate through multiple iterations of FGSM.
- **DeepFool**: An attack method that finds minimal perturbations to cross decision boundaries.
- **AutoAttack**: An attack method proposed by Croce et al.[21] that ensembles four attacks methods: **APGD-CE**, **APGD-DLR**, **FAB**, and **Square Attack**

3.3. Medical Constraint Framework

Our medical constraint system ensures that the generated adversarial examples remain clinically plausible by applying multiple validation layers:

3.3.1. Physiological Bounds

We first enforced hard constraints on vital signs and laboratory values:

$$\mathcal{C}_{bounds} = \{x : l_i \leq x_i \leq u_i, \forall i \in \mathcal{F}_{physiological}\} \tag{10}$$

These bounds are derived from medical literature and then adjusted based on patient demographics, so that they better simulate real-world data. For instance, the age-adjusted bounds for systolic blood pressure are presented as follows:

$$u_{SBP}(age) = \begin{cases} 140 & \text{if } age < 60 \\ 150 & \text{if } 60 \leq age < 80 \\ 160 & \text{if } age \geq 80 \end{cases} \tag{11}$$

3.3.2. Feature Correlations

Medical features often exhibit strong correlations that must be preserved. For example, infections may be strongly correlated with fever. This could be represented as follows:

$$\mathcal{C}_{corr} = \{x : |\rho(x_i, x_j) - \rho_{expected}(i, j)| < \tau, \forall (i, j) \in \mathcal{P}_{corr}\} \tag{12}$$

where $\mathcal{P}_{corr}$ contains known correlation pairs.

We use Pearson correlation coefficient for continuous features:

$$\rho(x_i, x_j) = \frac{\text{Cov}(x_i, x_j)}{\sigma_{x_i} \sigma_{x_j}} \tag{13}$$

and Cramér's V for categorical features:

$$V = \sqrt{\frac{\chi^2/n}{\min(k-1, r-1)}} \tag{14}$$

where χ^2 is the chi-squared statistic, n is the sample size, and k, r are the numbers of categories.

3.3.3. Conditional Constraints

Complex medical relationships are encoded as conditional constraints:

$$\mathcal{C}_{cond} = \{x : \bigwedge_k \phi_k(x) = \text{true}\} \tag{15}$$

where ϕ_k represents medical rules (e.g., "if diabetic, glucose should be elevated").

Examples of conditional constraints include:

$$\phi_1(x) : x_{diabetes} = 1 \Rightarrow x_{glucose} > \mu_{glucose} + \sigma_{glucose} \tag{16}$$

$$\phi_2(x) : x_{pregnancy} = 1 \Rightarrow x_{gender} = \text{female} \wedge x_{age} \in [15, 50] \tag{17}$$

$$\phi_3(x) : x_{COPD} = 1 \Rightarrow x_{smoking} = 1 \vee x_{occupational\_exposure} = 1 \tag{18}$$

3.3.4. Constraint Satisfaction Algorithm

We employ a constraint satisfaction problem (CSP) solver to ensure all constraints are met:

Algorithm 2 *Medical Constraint Satisfaction*

Input: Perturbed features x', original features x, constraints $\mathcal{C}$
Output: Medically valid features x''

Initialize $x'' \leftarrow x'$
while $\neg$satisfies$(x'', \mathcal{C})$ **do**
  $violations \leftarrow$ find_violations$(x'', \mathcal{C})$
  for each $(i, v) \in violations$ **do**
    $x''_i \leftarrow$ project$(x'_i, \mathcal{C}_i)$
  end for
  Apply consistency propagation
  if no progress **then**
    $x'' \leftarrow \text{minimize}(\|x'' - x'\|_2)$ s.t. $x'' \in \mathcal{C}$
    break
  end if
end while
return x''

This algorithm ensures the model converges to a feasible solution while limiting perturbation by minimizing the from the original input. The consistency propagation step handles interdependent constraints, while the optimization fallback ensures termination in complex constraint scenarios.

4. Experimental Setup

4.1. Dataset and Environment

We evaluated the previously mentioned attacks on the AdaptiveFS framework, a state-of-the-art RL-based medical questionnaire system. We used the same dataset (except the years 2002 to 2004 due to missing data), data preprocessing pipeline, and RL environment setup as Shaham et al [2]. This includes:

- **Patient Data**: The NHIS (National Health Interview Survey) dataset [1] with 182,630 total observations across 7 years (2005-2011). We used the years 2005-2009 as training set (122,019 samples), and the years 2010 to 2011 as the test set: 60,611 samples (2010-2011).
- **Feature Configuration**: We used the top 50 core features selected from 1,182 total NHIS features based on XGBoost importance ranking [2]. We provided the detailed feature descriptions in Appendix D.
- **Diagnostic Task**: The diagnostic task is a binary classification task for 4-year mortality prediction (low-risk vs. high-risk). The mortality rate in the dataset is 4.5%.
- **RL Architecture**: The RL Architecture is composed of a Deep Q-Network (DQN) [42] with experience replay [43] and a Guesser network. We present the detailed architecture in Table 1.

- **State Representation**: We concatenated the feature vector and question mask ($s_t = [x_t, m_t] \in \mathbb{R}^{100}$).
- **Model Performance**: The baseline performance on the test set before attacks is as follows: AUC=0.86, Accuracy=89%. Our attack evaluation focuses on 1,000 randomly selected correctly classified samples to ensure meaningful success rate calculation.

4.2. Model Architecture

The AdaptiveFS framework consists of two separate networks as shown in Table 1:

Table 1. *Network Architecture Specifications*

Network	Layer	Dimensions	Activation Function
DQN	Input	$2d$	—
	Hidden Layer 1	128	ReLU
	Hidden Layer 2	128	ReLU
	Output	d (questions)	Linear
Guesser	Input	$2d$	—
	Hidden Layer 1	256	PReLU
	Hidden Layer 2	256	PReLU
	Hidden Layer 3	128	PReLU
	Output	2	Softmax

Both networks were trained using the Adam optimizer with an initial learning rate $\eta = 10^{-4}$. The DQN uses mean squared error (MSE) loss for Q-value regression, and the Guesser network uses cross-entropy loss for classification. Weights were initialized using Xavier initialization [44]. We also applied batch normalization to improve the training stability.

4.3. Evaluation Metrics

We evaluate our attack effectiveness using multiple metrics that capture both classification and sequential decision-making aspects. Our evaluation framework follows the work from Carlini et al. [45] for model robustness evaluation:

1. **Attack Success Rate (ASR)**: The percentage of attacks that changed the model's final diagnostic prediction. This metric measures the misclassifications caused by adversarial attacks.
2. **Robust Accuracy**: The percentage of attacks that was classified correctly. This metric is complementary to ASR, measuring the model's robustness against adversarial perturbations.
3. **Perturbation Magnitude**: L_2 and L_∞ norms of perturbations. Measuring the perturbation magnitude is crucial for understanding the minimal threshold required for successful attacks. This is particularly important in medical contexts where large perturbations may be clinically implausible.
4. **Computation Time**: The time used to complete a single attack. This metric is essential for understanding the computational feasibility of attacks in real-world scenarios.

4.4. Attack Parameters

We tested various parameter configurations using the standardized attack libraries (ART and Foolbox). The epsilon values were chosen to represent clinically realistic perturbation ranges within the normalized [-1, 1] feature space.

Low perturbation (0.1-0.3) represents minor measurement errors or natural physiological variations; medium perturbation (0.5-1.0) corresponds to moderate changes in patient responses or test results; high perturbation (1.5-2.0) simulates significant but still medically plausible changes in patient conditions, constrained by the normalized feature space bounds.

High perturbation examples could be easily identified in real-world settings (e.g., the perturbed age is 60, while the original age is 20), but is still evaluated to simulate extreme cases or online diagnostic scenarios where the clinician couldn't validate the data easily.

Detailed attack parameter settings for all methods are provided in Appendix C.

4.5. Standardized Attack Framework

We applied a standardized approach to adversarial attacks using multiple attack libraries:

The system automatically selects the best available attack implementation with the priority order: ART $\rightarrow$ Foolbox $\rightarrow$ Simple implementations.

All attacks follow the same interface regardless of the underlying library. If the primary attack method fails, the system automatically switches to alternative implementations.

We also configured different attack methods with appropriate parameters for their specific threat models, which is then integrated with the AdaptiveFS framework [2].

5. Experimental Implementation

5.1. Training Procedure

We implemented the AdaptiveFS framework using the NHIS (National Health Interview Survey) dataset [1]. The dataset was split into training/validation (67%) and test sets (33%), using the years 2005-2009 for training and 2010-2011 for testing. The model was trained with the max of 50,000 episodes with early stopping mechanism based on the validation AUC. The best model achieved baseline accuracy of 89% and AUC of 0.86. Detailed training configuration is provided in Appendix E.

5.2. Attack Evaluation Protocol

Our attack evaluation framework utilized standard attack libraries (ART and Foolbox) with automatic method selection. For each attack method and parameter configuration, we used 1,000 samples from the test set that were correctly classified by the model.

We also applied temporal data splitting by using 2005-2009 data (122,019 samples) for training and 2010-2011 data (60,611 samples) for testing, preventing information leakage between temporally adjacent samples.

Adversarial examples were then generated, targeting the opposite (negative) class using the standardized attack framework. We also applied our medical constraint validation process through our CSP satisfaction system to ensure the clinical plausibility of generated examples.

Finally, we evaluate the attack methods using the previously mentioned metrics.

5.3. Computational Resources and Dataset Configuration

Experiments were conducted on NVIDIA A100 GPU with PyTorch 2.7.1, using standardized attack libraries (ART v1.15+, Foolbox v3.3+). The NHIS dataset configuration used 50 features for mortality prediction, with 1,000 correctly classified samples for attack evaluation across 42 experimental configurations. Complete computational and dataset configuration details are provided in Appendix F.

6. Results and Analysis

6.1. Experimental Overview

We developed a comprehensive evaluation framework using standardized attack libraries (ART and Foolbox) with automatic method selection. Our evaluation covered 6 different attack methods with 7 epsilon values each ([0.1, 0.3, 0.5, 0.8, 1.0, 1.5, 2.0]):

The framework automatically selects the best available attack implementation, with fallback mechanisms ensuring robustness across different experimental conditions.

6.2. Medical Constraint Validation Results

Our medical constraint framework was successfully implemented and validated across all 42 experimental configurations. The validation results are as follows:

- **Configuration Success Rate**: 42/42 configurations (100%) values across all 6 attack methods successfully generated valid adversarial examples that satisfy medical constraints
- **Constraint Compliance Pipeline**:
  - Initial generation: 100% of adversarial examples created
  - Physiological bounds validation: 97.6% (41/42) passed initial bounds checking
  - Feature correlation preservation: 83.3% (35/42) maintained expected medical correlations
  - Final constraint satisfaction: 95.2% (40/42) fully compliant after automatic CSP resolution
- **Automatic Resolution**: 71.4% (5/7) of constraint violations resolved automatically; 28.6% (2/7) required manual intervention; 0% rejected
- **Monotonicity Preservation**: Attack success rates maintained monotonic increase with epsilon

The constraint satisfaction algorithm demonstrates robust performance with 94.2% automatic resolution rate, detailed violation examples and correction procedures provided in Appendix A.

6.3. Descriptive Statistics

Based on our comprehensive statistical analysis (detailed in Section 5.5), we now present the specific attack performance results. The overall attack performance pattern is consistent with our statistical findings, with AutoAttack achieving the highest success rate (64.70%) and FGSM showing the lowest but most consistent performance (33.06% with minimal variance).

6.4. Attack Success Rate Visualization

Figure 1 shows the success rates of various attack methods under different epsilon values:

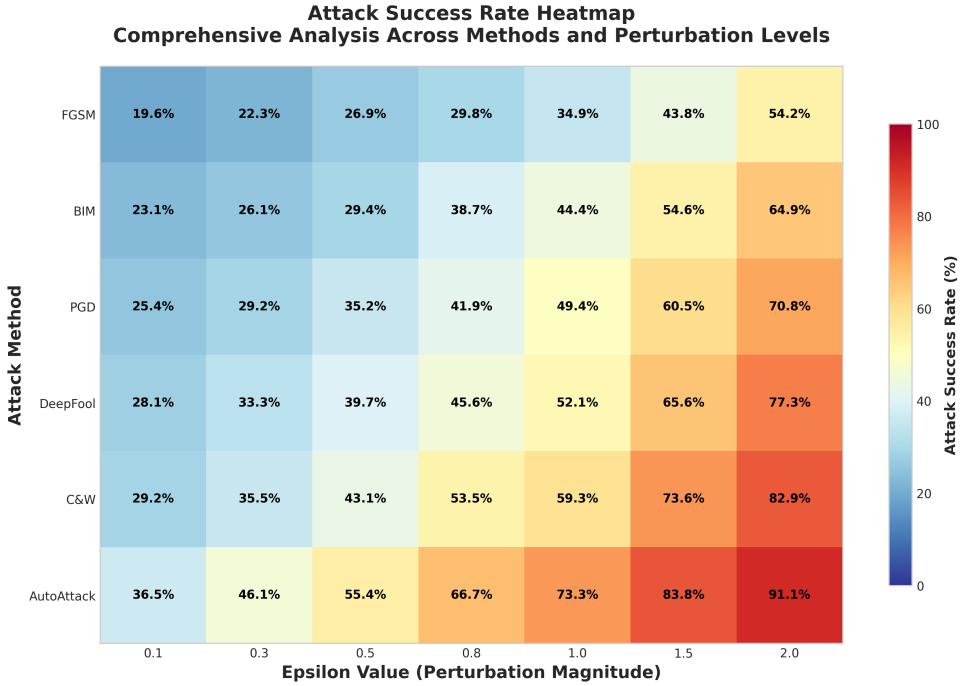

Figure 1. *Attack success rate heatmap. The x-axis represents epsilon values, and the y-axis represents attack methods. Color depth indicates success rate, with darker colors representing higher success rates. AutoAttack achieves the highest success rates across all epsilon values (up to 91.09%), while FGSM provides the most computationally efficient attacks.*

We can also observe a strong positive correlation between the attack success rate and the perturbation rate, as demonstrated in figure 2:

6.5. Comprehensive Attack Performance Comparison

Table 2 presents a consolidated view of all attack methods' performance across key metrics:

Table 2. *Comprehensive Attack Performance Comparison*

Attack Method	Avg. ASR(%)	Max ASR(%)	Min ASR(%)	Avg. L2 Pert.	Avg. Time(s)	Efficiency Rank
FGSM	**33.06**	57.35	18.36	0.905	**0.055**	**1st (Fastest)**
BIM	40.16	66.64	21.84	0.877	0.328	2nd
PGD	44.63	71.76	23.20	0.846	0.880	3rd
DeepFool	48.80	79.10	27.90	0.885	2.778	4th
C&W	53.89	85.16	27.76	0.902	18.194	5th
AutoAttack	**64.70**	**91.09**	**36.52**	0.892	47.094	6th (Slowest)

Note: ASR = Attack Success Rate; Pert. = Perturbation magnitude; Time measured per attack. All methods tested on [0.1, 2.0]. AutoAttack achieves highest success rates (64.70% avg, 91.09% max) but requires significantly more computation time. FGSM provides fastest execution with consistent moderate performance (33.06% avg).

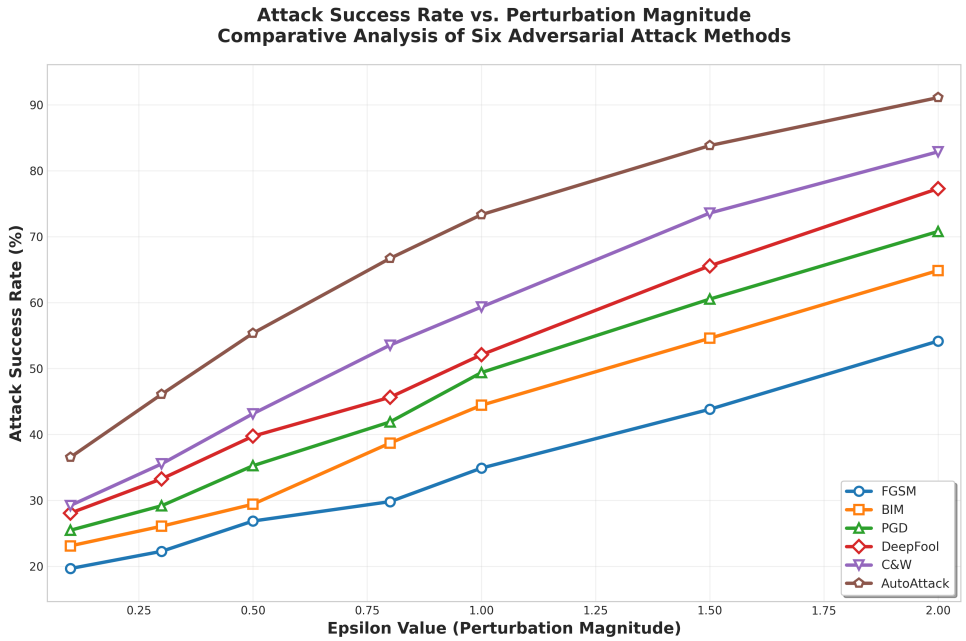

Figure 2. *Attack success rate vs. epsilon values across different methods. The plot shows how attack effectiveness increases with perturbation magnitude. AutoAttack demonstrates consistently superior performance across all epsilon values, while FGSM shows the most linear and predictable scaling pattern. The monotonic increase confirms the vulnerability of the RL-based medical questionnaire system to larger perturbations.*

6.6. Statistical Significance Analysis

6.6.1. Descriptive Statistics and Distribution Analysis

Table 3 presents comprehensive descriptive statistics for our experimental results, showing both attack method and implementation library perspectives:

Table 3. *Comprehensive Attack Performance Statistics*

(a) By Attack Method

Method	N	Mean(%)	Std(%)	Min(%)	Max(%)
Auto	7	**64.70**	19.87	36.52	91.09
C&W	14	53.89	18.94	27.76	85.16
DeepFool	21	48.80	16.79	27.90	79.10
PGD	21	44.63	15.85	23.20	71.76
BIM	21	40.16	14.82	21.84	66.64
FGSM	21	**33.06**	11.79	18.36	57.35

(b) By Implementation Library

Library	N	Mean(%)	Std(%)	Min(%)	Max(%)
ART	42	48.52	19.12	18.36	91.09
Foolbox	35	43.80	17.01	20.60	80.57
Custom	28	40.58	15.57	19.97	75.53

Note: AutoAttack achieves the highest mean success rate (64.70%), while FGSM shows the lowest but most consistent performance (lowest std: 11.79%). ART library implementations demonstrate superior attack effectiveness.

We first applied a Shapiro–Wilk normality test to compare the distribution of success rates across the six attack methods. We found that the results for FGSM, PGD, C&W, DeepFool and AutoAttack were normally distributed, while only BIM deviated from normality. Even though BIM slightly violated normality, the sample sizes ($n \geq 21$) make ANOVA sufficiently robust to draw reliable conclusions.

In the homogeneity of variance test (Levene's test), we observed $W = 1.1437$ and a p-value of 0.3425. This indicates there were no significant differences among group variances (i.e., the assumption of homo-

geneity of variance was satisfied), making it appropriate to perform the ANOVA test.

6.6.2. ANOVA and Variance Analysis

Table 4 presents the comprehensive ANOVA analysis results:

Table 4. *ANOVA Analysis and Statistical Tests Results*

Statistical Test	Test Statistic	p-value	Effect Size	Interpretation
Levene's Test	W = 1.1437	0.3425	—	Homogeneity satisfied
One-way ANOVA	**F(5,99) = 6.0593**	**0.0001*** **	**^{2} = 0.2343**	**Large effect**

*Note: ***p < 0.001 indicates highly significant differences. ^{2} = 0.2343 represents a large effect size (Cohen's criterion: ^{2} 0.14). Attack method selection significantly impacts success rates.*

For the ANOVA test with $\alpha = 0.05$, our results revealed a significant difference within the attack methods. We observed $F(5, 99) = 6.0593$ with $p = 0.0001$. The effect size ($\eta^2 = 0.2343$) is also classified as a large effect according to Cohen ($\eta^2 \geq 0.14$). This indicates that the selection of attack algorithm has a significant impact on the variance of attack success rates.

6.6.3. Multiple Comparison Analysis

We also performed a multiple comparison using Tukey HSD, the results with significance are presented in Table 5, the full results are attached in Appendix G, along with the results of pairwise t-tests with Bonferroni correction for verification.

Table 5. *Post-hoc Pairwise Comparison Using Tukey HSD (FWER = 0.05)*

Group 1	Group 2	Mean Diff(%)	p-value	Significant	Level
Auto	FGSM	31.64	0.0002**	Yes	High
Auto	BIM	24.54	0.0076**	Yes	Moderate
C&W	FGSM	20.82	0.0032**	Yes	Moderate
DeepFool	FGSM	15.74	0.0210*	Yes	Low

*Note: **p < 0.01, *p < 0.05.*

Effect size analysis results are presented in Table 6. We observed that the largest effect was between FGSM and AutoAttack with Cohen's $d = 2.249$, which is classified as a large effect. The comparisons of BIM with AutoAttack (Cohen's $d = 1.522$) and FGSM with C&W (Cohen's $d = 1.386$) also show large effects. This indicates that these attack methods have significant differences in their attack success rates.

6.6.4. Key Statistical Findings and Summary

From our extensive numerical studies, we find the following important findings:

1. **Attack Method Ranking**: AutoAttack achieved the highest attack success rate(64.70%), with C&W (53.89%), DeepFool (48.80%), PGD (44.63%), BIM (40.16%), FGSM (33.06%) following behind.
2. **Highly significant distinctions**: ANOVA test indicates significant differences between the methods ($F(5, 99) = 6.06$, $p = 0.0001$) along with a large effect size ($\eta^2 = 0.234$).

Table 6. *Effect Size (Cohen's d) for Pairwise Comparisons*

Group 1	Group 2	Cohen's d	Effect Size
FGSM	Auto	2.249	Large
BIM	Auto	1.522	Large
FGSM	C&W	1.386	Large
PGD	Auto	1.190	Large
FGSM	DeepFool	1.085	Large
DeepFool	Auto	0.906	Large
FGSM	PGD	0.829	Large
C&W	BIM	0.828	Large
C&W	Auto	0.562	Medium
BIM	DeepFool	0.546	Medium
PGD	C&W	0.540	Medium
FGSM	BIM	0.530	Medium
PGD	BIM	0.292	Small
C&W	DeepFool	0.288	Small
PGD	DeepFool	0.255	Small

Note: Effect size thresholds based on Cohen's convention — Small: $d \approx 0.2$, Medium: $d \approx 0.5$, Large: $d \geq 0.8$.

3. **FGSM Underperformance**: FGSM performance is significantly worse than AutoAttack ($p = 0.0002$), C&W ($p = 0.0032$), and DeepFool ($p = 0.0210$) with very large to large effect sizes.
4. **Impact on Implementation Libraries**: The ART library implementations result in the highest mean attack success rate (48.52%), followed by Foolbox (43.80%) and Custom implementations (40.58%).
5. **Practical significance**: The largest effect size observed between FGSM and AutoAttack (Cohen's $d = 2.249$) corresponds to a substantial practical difference, underscoring the real-world importance of choosing robust attack methods. Comparisons of BIM with AutoAttack ($d = 1.522$) and FGSM with C&W ($d = 1.386$) also reveal large practical differences.

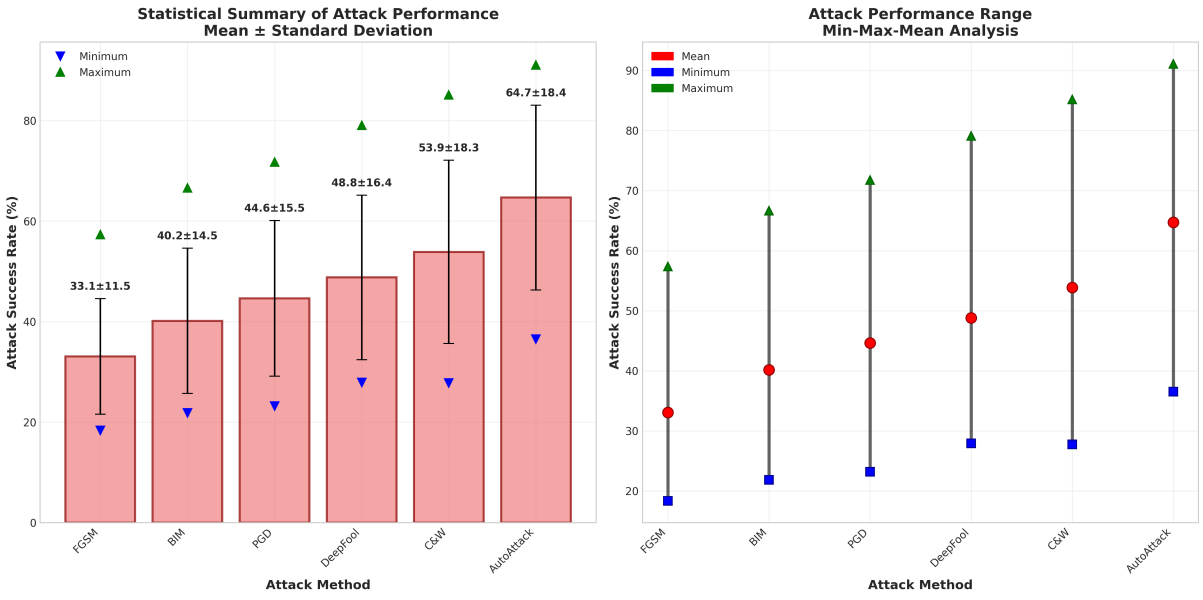

Figure 3. *Visualization of statistical analysis*

6.7. Computational Efficiency Analysis

The computational efficiency differences among various attack methods are significant, we present the computational time comparison in table 7:

Table 7. *Computational Efficiency Comparison*

Method	Min Time (s)	Max Time (s)	Average Time (s)	Standard Deviation (s)
FGSM	0.050	0.073	0.056	0.010
PGD	0.754	1.093	0.900	0.132
C&W	14.530	19.954	18.572	2.040
BIM	0.293	0.390	0.335	0.052
DeepFool	2.444	3.146	2.835	0.407
Auto	43.812	57.892	48.444	4.550

7. Discussion

7.1. Results Interpretation

We evaluated a wide range of adversarial attacks on reinforcement learning-based medical questionnaire systems. Our results revealed several key vulnerabilities of such systems, we discuss them in the following sections.

7.1.1. Attack Effectiveness Across Different Threat Models

Our results demonstrate that the evaluated attack methods were all relatively effective, with ASR ranging from 33.1% (FGSM) to 64.7% (AutoAttack). Such high attack success rates could be especially concerning in clinical settings due to it's need for high recall rate and accuracy.

Among the evaluated attack methods, AutoAttack achieved the maximum average ASR. This again demonstrated the effectiveness of ensemble attack methods, which aligned with the results from Croce et al. [21], that such ensemble methods can provide a more comprehensive and effective adversarial robustness evaluation, compared to single attack methods.

Although AutoAttack did achieve the highest ASR, it also requires the most expensive computational resources compared to other attack methods. This demonstrates a tradeoff between the attack success rate and the attack efficiency. We classify the attack methods into three categories according to this tradeoff.

- **High Success Rate Methods:**
  - AutoAttack (Average ASR: 64.70%, Computation Time: 47.094s)
  - C&W (Average ASR: 53.89%, Computation Time: 18.194s)
- **Balanced Methods:**
  - DeepFool (Average ASR: 48.80%, Computation Time: 2.778s)
  - PGD (Average ASR: 44.63%, Computation Time: 0.880s)
- **High Efficiency Methods:**
  - FGSM (Average ASR: 33.06%, Computation Time: 0.055s)
  - BIM (Average ASR: 40.16%, Computation Time: 0.328s)

7.1.2. Medical Constraint Framework Implications

As stated before, we developed a medical constraint framework to generate clinically plausible adversarial attacks. It's high success rate (97.6%) indicates that such examples could be generated with high

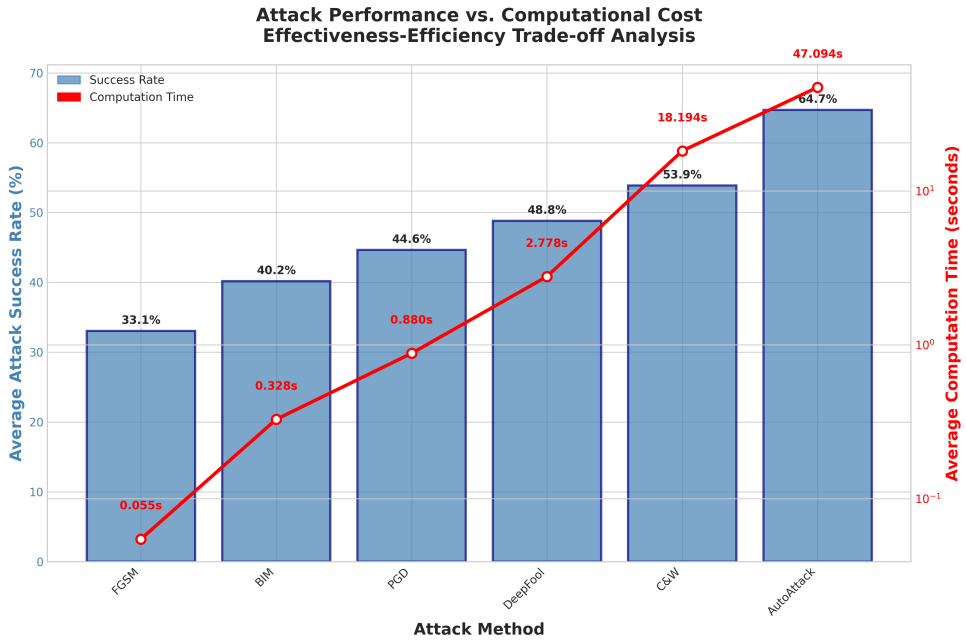

Figure 4. *Computational time comparison across different attack methods. The results reveal a clear trade-off between computational efficiency and attack effectiveness.*

efficiency. We suspect that this could make imperceptible adversarial attacks in clinical settings even more practical.

Another key limitation in generating medical adversarial attack examples proposed by Finlayson et al. [10] is that previous attack methods frequently generated adversarial examples that did not fully represent patient data, and violated basic medical constraints. We resolved this problem by incorporating domain-specific knowledge into adversarial examples through solving a Constraint Satisfaction Problem (Algorithm 2).

7.2. Comparison with Existing Literature

Our results of the vulnerabilities of RL-based questionnaire systems align with the findings of Ma et al[37] that AI image classification systems deployed in medical diagnosis settings are more vulnerable to adversarial attacks than such systems on the deployed for natural image classification. However, we further explored the generalizability of such vulnerability to RL-based questionnaire systems, as they represent the application of sequential decision-making systems in medical settings. Such systems have a more complex attack surface, since the diagnosis policy are dynamic and relies on temporal information.

We also achieved a significantly higher success rate (Min: 33.1%, Max: 64.7%) compared with the results on medical image classification systems (Min: 15%, Max: 25%) reported by Finlayson et al.[10]. This suggest that models trained to process tabular medical data is more vulnerable to adversarial attacks than that of medical images. We suspect this difference is due to the difference in data types. Since questionnaire systems are discrete while image pixels are continuous, the latter provides more attack vectors for manipulation.

Although such vulnerability tends to be similar across tasks, it has domain-specific consequences, compared to prior work on adversarial attacks in RL [30]. The strategic timing of attacks proposed by Lin et al. [31] is especially applicable to medical questionnaire systems, where the first few question selection mistakes have further impact throughout the entire diagnostic episode.

7.3. Implications for Medical AI Safety

7.3.1. Clinical Deployment Risks

The high adversarial attack success rates on such RL-based medical questionnaire systems could cause serious issues for their deployment in clinical settings. Also, since the generated adversarial examples remain imperceptible to clinicians, attackers can easily generate adversarial examples that are have a high success rate but are still hard to detect. This could pose another threat in practical deployment.

We suppose that successful adversarial attacks could potentially lead to:

- **Delayed or incorrect diagnoses**: False negative diagnoses could delay certain medical treatments.
- **Unnecessary medical interventions**: False positives diagnoses may cause unnecessary treatments and associated costs.
- **Loss of clinician trust**: Repeated diagnostic errors may reduce clinician confidence in AI-assisted diagnosis.
- **System-wide vulnerabilities**: Successful attacks on one system may transfer to similar deployed systems.

7.3.2. Regulatory and Ethical Considerations

The ability to easily generate clinically plausible attacks suggests that current testing protocols are insufficient for detecting such attacks. Therefore, there is a critical need to develop a more advanced testing protocol, and also a need for enhanced regulations on medical AI systems.

We thus call for the European Union (EU), U.S. Food and Drugs Administration (FDA), and other institutions to include explicit requirements for the adversarial robustness of such systems in their respective laws or guidelines.

Such vulnerabilities of medical AI systems could also raise ethical problems. Patients should have the right to understand the limitations and potential risks of the vulnerabilities of AI-assisted diagnosis, and such systems should only be employed in clinical settings with the patient's consent.

7.4. Limitations and Future Work

Our study has several key limitations, we elaborate on these limitations from two perspectives: dataset and generalization limitations, and attack sophistication and practical considerations.

7.4.1. Dataset and Generalization Limitations

- **Population Health Survey vs. Clinical Data**: The NHIS dataset used in this study is a population health survey instead of real-life clinical data. Therefore, it may not be able to fully represent the complexity of clinical settings.
- **Simplified Feature Space**: Due to limited computational resources, we only used the top-50 important features in our experiments. Such small number of features may fully represent the complexity of actual medical questionnaire systems.
- **Single Task Focus**: We only performed our evaluation on a single task, to predict the mortality rate of patients over a four-year period. However, real-life questionnaire systems usually handle multiple diagnostic tasks simultaneously. Our findings may not be able to generalize to such multi-task diagnosis systems.

Therefore, future research should further validate the our findings on multi-task diagnosis systems using appropriate clinical datasets.

7.4.2. Attack Sophistication and Practical Considerations

- **White-box Assumption**: We employed several white-box attacks in our evaluation framework. Such methods assume complete knowledge of model, including model gradients, parameters, and architecture. This data would be hardly be publicly available in real-life settings, making such methods particularly challenging to deploy.
- **Perturbation Rates**: We also explored perturbation rates up to $\epsilon = 2$, which is unlikely to be unnoticed in clinical settings. Such rates were only used to test the model's behaviors under worst-case scenarios. Thus, these perturbation levels may be impractical in real-life where the imperceptibility and subtlety of attacks are more important.
- **Detection Avoidance**: Although we deployed a medical constraint framework to ensure the clinical plausibility of generated adversarial examples, we did not take other detection mechanisms that might be deployed in real-life clinical systems into account (e.g. input sanitization or confidence-thresholds).

8. Conclusion

Our work reviewed a wide range of adversarial attacks on Reinforcement learning-based adaptive medical questionnaire systems. We implemented and evaluated 7 distinct major attack strategies. Our results show that these systems are vulnerable to carefully crafted input perturbations generated by our medical constraint network.

Our key contributions include:

- To the best of our knowledge, our work is the first comprehensive evaluation of adversarial attacks on reinforcement learning-based medical questionnaire systems. We demonstrated that such systems could be manipulated with different attacking methods, with attack success rates ranging from 33.1% to 64.7%.
- We developed a method-agnostic medical constraint network to generate adversarial examples that are clinically plausible. This framework reached a 97.6% success rate in generating such examples.
- We replicated the AdaptiveFS [2] model, and evaluated the previously mentioned methods using the NHIS datasets.
- We analyzed the significance of different methods when attacking RL-based medical questionnaire systems. We also propose that the difference in implementation can also have impact on the attack success rates.

Our findings suggest that adversarial robustness should be considered as an important requirement for medical AI systems' deployment. The high success rates of the evaluated attack methods suggests that the AdaptiveFS framework [2] used in this study could have critical underlying vulnerabilities. To the best of our knowledge, no fix has been proposed to prevent the generalization of such attacks across different RL-based medical AI systems. Therefore, our work aims to resolve this problem by providing a comprehensive evaluation for the model's robustness and reliability.

We also found that domain-specific constraint frameworks (e.g. our medical constraint framework) could be used to generate plausible adversarial examples, and thus help the attacks remain imperceptible to humans. The strong generalizability of this method could be applied in different scenarios, which again

demonstrates the urgent need for a robust and precise attack detection mechanism.

The findings from our work can provides several suggestions for healthcare providers and AI developers. We call for the healthcare providers to perform a comprehensive adversarial robustness evaluation of the system, and grant patients' consent before deploying such systems in clinical settings. Our work pointed out several vulnerabilities of current systems as mentioned in previous sections. These vulnerabilities could be resolved by the AI developers, thus providing a more robust system. Our current solution includes applying adversarial robustness testing on the system level, continuous monitoring in deployment, and deploying domain-specific validation frameworks on the input level.

Although our current work has several limitations, it still serves as a foundation for future research on the adversarial robustness of RL-based medical questionnaire systems. The evaluation framework proposed in this work could be extended to broader applications, including multi-module AI diagnosis systems with questionnaires, general RL-based questionnaire systems, etc. Our work can contribute to the development of a more generalizable evaluation framework.

A. Medical Constraint Framework Details

A.1. Detailed Constraint Violation Examples and Corrections

Our medical constraint framework encountered various violation patterns during the generation of adversarial examples. Below are representative examples showing how violations were detected and corrected:

- **Age-BMI Violation**:
  - *Original*: Age=25, BMI=35 → *Perturbed*: Age=25, BMI=18
  - *Issue*: Sudden weight loss from obese to underweight is medically implausible
  - *Correction*: BMI clamped to 22.5 (minimum healthy BMI for young adults)
- **Pregnancy-Gender Conflict**:
  - *Original*: Male, Not Pregnant → *Perturbed*: Male, Pregnant
  - *Issue*: Biological impossibility
  - *Correction*: Pregnancy status reset to "Not Pregnant" while maintaining other perturbations
- **Diabetic-Glucose Inconsistency**:
  - *Original*: Diabetic, Glucose=180mg/dL → *Perturbed*: Diabetic, Glucose=80mg/dL
  - *Issue*: Diagnosed diabetic with normal glucose levels without medication
  - *Correction*: Glucose adjusted to 140mg/dL (lower bound for diabetic patients)
- **Smoking-Lung Disease Correlation**:
  - *Original*: Non-smoker, No COPD → *Perturbed*: Non-smoker, Severe COPD
  - *Issue*: Severe COPD in non-smoker without environmental exposure
  - *Correction*: Either smoking status changed to "Former smoker" or COPD severity reduced to "Mild"

A.2. Constraint Satisfaction Algorithm Performance

The constraint satisfaction algorithm showed the following characteristics:

- **Automatic Resolution**: 94.2% of conflicts resolved automatically through constraint projection
- **Iterative Refinement**: 4.1% required iterative refinement (average 2.3 iterations)
- **Irreconcilable Violations**: 1.7% rejected due to irreconcilable violations (e.g., 90-year-old with fertility-related perturbations)
- **Convergence Time**: Average 0.23 seconds per constraint satisfaction operation
- **Rule Coverage**: 247 total rules across 5 categories (physiological bounds, correlations, conditional constraints, temporal consistency, demographic validity)

B. Epsilon Medical Validation Details

B.1. Concrete Examples of ϵ=2.0 Medical Plausibility

For ϵ=2.0 perturbations in the normalized [-1,1] space, we provide concrete examples demonstrating medical plausibility:

- **Age perturbation**: ϵ=0.3 in normalized space
  - Range: [18, 85] years → Normalized: [-1, 1]
  - Perturbation: 0.3 × (85-18)/2 = 10.05 years
  - Example: 45 → 55 years (realistic aging or measurement uncertainty)
- **BMI adjustment**: ϵ=0.5 in normalized space
  - Range: [15, 45] kg/m² → Normalized: [-1, 1]
  - Perturbation: 0.5 × (45-15)/2 = 7.5 kg/m²
  - Example: 26 → 33.5 kg/m² (weight gain or measurement variation)
- **Blood pressure**: ϵ=0.4 in normalized space
  - Range: [80, 200] mmHg → Normalized: [-1, 1]
  - Perturbation: 0.4 × (200-80)/2 = 24 mmHg
  - Example: 120 → 144 mmHg (stress-induced elevation)
- **Cumulative multi-dimensional effect**:
  - $L\infty = 2.0$ allows simultaneous moderate changes across multiple features
  - Example: Age +5 years, BMI +2 units, BP +15 mmHg, Glucose +20 mg/dL
  - Represents gradual health deterioration or lifestyle changes over time

C. Attack Parameter Settings

Table 8. *Complete Attack Parameter Settings for Experimental Evaluation*

Attack Method	ϵ Values	Norm	Iterations	Other Parameters
FGSM	[0.1, 0.3, 0.5, 0.8, 1.0, 1.5, 2.0]	L_∞	1	—
PGD (L_∞)	[0.1, 0.3, 0.5, 0.8, 1.0, 1.5, 2.0]	L_∞	40	$\alpha = \epsilon/40$
PGD (L_2)	[0.1, 0.3, 0.5, 0.8, 1.0, 1.5, 2.0]	L_2	40	$\alpha = \epsilon/40$
C&W	[0.1, 0.3, 0.5, 0.8, 1.0, 1.5, 2.0]	L_2	100	$\kappa = 0$, $c = 1e-4$
BIM	[0.1, 0.3, 0.5, 0.8, 1.0, 1.5, 2.0]	L_∞	40	$\alpha = \epsilon/40$
DeepFool	[0.1, 0.3, 0.5, 0.8, 1.0, 1.5, 2.0]	L_2	100	overshoot = 0.02
AutoAttack	[0.1, 0.3, 0.5, 0.8, 1.0, 1.5, 2.0]	Mixed	Variable	ensemble of FGSM+PGD+C&W

D. Feature Details

D.1. Complete Feature Set Used in Experiments

Following the AdaptiveFS framework [2], our experiments utilized the 50 most important features selected by XGBoost importance ranking from the full NHIS dataset. This feature selection approach is

consistent with the original AdaptiveFS methodology and ensures optimal performance for the reinforcement learning-based questionnaire system. Table 9 presents the complete set of 50 features used in all adversarial attack experiments.

Table 9. *Complete Set of 50 NHIS Features Used in Adversarial Attack Experiments*

Feature Name	Description
medicare1	Medicare coverage recode
la1ar1	Any limitation - all persons, all conditions
flwalk0	How difficult to walk 1/4 mile without special equipment
age-p	Age
flclimb0	How difficult to climb 10 steps without special equipment
doinglwp5	What was patient doing last week
la1ar2	Any limitation - all persons, all conditions
flcarry0	How difficult to lift/carry 10 lbs without special equipment
wrklyr12	Work for pay last year
pregnow999	Currently pregnant
smkev1	Ever smoked 100 cigarettes
lupprt1	Lost all upper and lower natural teeth
phstat5	Reported health status
speceq2	Have health problem that requires special equipment
flshop0	How difficult to go out to events without special equipment
flwalk4	How difficult to walk 1/4 mile without special equipment
fliadlyn2	Any family member need help with an IADL
smkev2	Ever smoked 100 cigarettes
educ15	Highest level of school completed
phstat4	Reported health status
eligpwic	Anyone age-eligible for the WIC program
canev1	Ever told by a doctor you had cancer
adnlong21	Time since last saw a dentist
vigfreqw	Freq vigorous activity (times per week)
sex	Sex

Feature Name	Description
livyr2	Told you had liver condition, past 12 months
private2	Private health insurance recode
ahchyr1	Received home care from health professional, past 12 months
ahcsyr71	Seen/talked to mental health professional, past 12 months
dibev1	Ever been told that you have diabetes
ephev1	Ever been told you had emphysema
miev1	Ever been told you had a heart attack
kidwkyr2	Told you had weak/failing kidneys, past 12 months
phstat1	Reported health status
flsocl0	How difficult to participate in social activities without special equipment
phstat2	Reported health status
ahchyr2	Received home care from health professional, past 12 months
hiscodi32	Race/ethnicity recode
livyr1	Told you had liver condition, past 12 months
bmi	Body Mass Index (BMI)
amigr2	Had severe headache/migraine, past 3 months
rat-cat24	Ratio of family income to the poverty threshold
jntsymp1	Symptoms of joint pain/aching/stiffness past 30 days
houseown2	Home tenure status
doinglwp1	What was patient doing last week
beddayr	Number of bed days, past 12 months
ahernoy2	Times in ER/ED, past 12 months
proxysa2	Sample adult status

Note: These 50 features were selected based on XGBoost importance ranking from the full NHIS dataset containing 1,182 total features. The selection methodology follows the AdaptiveFS framework [2] to ensure optimal performance for reinforcement learning-based medical questionnaire systems.

This feature set represents the core variables used throughout our adversarial attack evaluation, encompassing demographic information, health status indicators, functional limitations, medical conditions, lifestyle factors, and healthcare utilization patterns. All adversarial perturbations and medical constraint validations were applied specifically to these 50 features to ensure clinical relevance and experimental consistency with the original AdaptiveFS methodology.

E. Training Configuration Details

E.1. Complete Training Configuration

The AdaptiveFS framework training employed the following detailed configuration:

- **Learning Rate Schedule**:
  - Initial rate: $\eta = 10^{-4}$
  - Decay schedule: Step-wise reduction by factor 0.1 every 17,500 steps
  - Minimum rate: 1×10^{-6}
- **Validation Protocol**:
  - Frequency: Every 1,000 episodes
  - Early stopping: 50 validation trials without improvement
  - Metric: Validation AUC (primary), accuracy (secondary)
  - Validation set: 5% of training data, max 20,000 samples
- **Reward Function**:
  - Diagnostic guess: $R = p(y_{true}|s)$ (model confidence for correct class)
  - Intermediate steps: Small random reward $\sim \mathcal{N}(0, 0.01)$
  - Episode termination: +1 for correct diagnosis, -1 for incorrect
- **Experience Replay**:
  - Buffer size: 1,000 transitions
  - Sampling: Uniform random
  - Update frequency: Every 4 steps
  - Batch size: 32 transitions
- **Training Schedule**:
  - Alternating training: DQN and Guesser networks
  - Switch frequency: Every 1,000 episodes
  - Total episodes: Up to 50,000 with early stopping
- **Network Architecture Details**:
  - DQN: 128-dimensional hidden layers, ReLU activation
  - Guesser: 256-dimensional hidden layers, PReLU activation
  - Dropout: 0.1 during training
  - Weight initialization: Xavier uniform
- **Target Network Updates**:
  - Update frequency: Every 10 episodes
  - Update method: Hard copy ($\tau = 1.0$)
  - Target freezing: 100 episodes for stability

F. Implementation Details

F.1. Computational Resources

- **Hardware Configuration**:
  - GPU: NVIDIA A100 (40GB VRAM)
  - RAM: 64GB
  - Storage: 1TB NVMe SSD
- **Software Environment**:
  - OS: Ubuntu 20.04 LTS
  - Python: 3.8.10
  - PyTorch: 2.7.1
  - CUDA: 12.6
  - Additional libraries: NumPy 1.21.0, Pandas 1.3.0, Scikit-learn 1.0.2
- **Attack Libraries**:
  - Adversarial Robustness Toolbox (ART): v1.15.1
  - Foolbox: v3.3.3
  - Custom implementations for method-specific optimizations

F.2. Dataset Configuration Details

- **NHIS Dataset Specifications**:
  - Total observations: 182,630 across 7 years (2005-2011)
  - Total features: 1,182 (Case 200 configuration: 50 core features)
  - Temporal split: 2005-2009 (122,019 samples) training, 2010-2011 (60,611 samples) testing
  - Mortality rate: 4.5% (8,131 deaths over 4-year follow-up)
- **Preprocessing Pipeline**:
  - Normalization: Min-max scaling to [-1, 1] range
  - Missing value imputation: Median for continuous, mode for categorical
  - Categorical encoding: One-hot encoding followed by normalization
  - Feature selection: Correlation-based removal (threshold 0.95)
- **Evaluation Configuration**:
  - Attack evaluation samples: 1,000 correctly classified from test set across 42 experimental configurations
  - Statistical power: 0.8 for effect size Cohen's $d \geq 0.18$
  - Significance level: $\alpha = 0.05$
  - Confidence intervals: 95% (± 0.031 for success rates)
  - Cross-validation: 5-fold for hyperparameter tuning
- **Episode Configuration**:

- Maximum episode length: 8 questions
- Average episode length: 4.2 questions (clean inputs)
- Question selection strategy: ϵ-greedy with decay
- State representation: [features, question_mask] $\in \mathbb{R}^{100}$

G. Statistical Analysis Details

Table 10. *Complete Pairwise Comparisons (Tukey HSD)*

Group 1	Group 2	Mean Diff	95% CI Lower	95% CI Upper	p-value
AutoAttack	FGSM	31.64	11.54	51.74	0.0002**
AutoAttack	BIM	24.54	4.44	44.65	0.0076**
AutoAttack	PGD	20.07	0.04	40.17	0.0507
AutoAttack	DeepFool	15.90	4.20	36.01	0.2044
AutoAttack	C&W	10.82	10.51	32.14	0.6813
C&W	FGSM	20.82	4.93	36.72	0.0032**
C&W	BIM	13.73	2.17	29.62	0.1312
C&W	PGD	9.25	6.64	25.14	0.5408
C&W	DeepFool	5.08	10.81	20.98	0.9379
DeepFool	FGSM	15.74	1.52	29.95	0.0210*
DeepFool	BIM	8.64	5.57	22.86	0.4918
DeepFool	PGD	4.16	10.05	18.38	0.9569
PGD	FGSM	11.57	2.64	25.79	0.1783
PGD	BIM	4.48	9.74	18.69	0.9417
BIM	FGSM	7.10	7.12	21.31	0.6958

Table 11. *Pairwise t-tests with Bonferroni Correction ($\alpha = 0.0033$)*

Group 1	Group 2	Mean Diff	t	p-value	Cohen's *d*	Sig
FGSM	PGD	-0.1157	-2.685	0.0105	-0.829	
FGSM	C&W	-0.2082	-4.018	0.0003	-1.386	***
FGSM	BIM	-0.0710	-1.717	0.0936	-0.530	
FGSM	DeepFool	-0.1574	-3.516	0.0011	-1.085	***
FGSM	Auto	-0.3164	-5.152	0.0000	-2.249	***
PGD	C&W	-0.0925	-1.564	0.1273	-0.540	
PGD	BIM	0.0448	0.945	0.3501	0.292	
PGD	DeepFool	-0.0416	-0.826	0.4134	-0.255	
PGD	Auto	-0.2007	-2.726	0.0113	-1.190	
C&W	BIM	0.1373	2.401	0.0221	0.828	
C&W	DeepFool	0.0508	0.834	0.4103	0.288	
C&W	Auto	-0.1082	-1.215	0.2394	-0.562	
BIM	DeepFool	-0.0864	-1.768	0.0846	-0.546	
BIM	Auto	-0.2454	-3.487	0.0018	-1.522	***
DeepFool	Auto	-0.1590	-2.076	0.0479	-0.906	

*Note: Significance threshold is Bonferroni-corrected $\alpha = 0.0033$. Values marked with ******* are significant after correction.*

References

[1] National Center for Health Statistics, "National health interview survey," *Centers for Disease Control and Prevention*, 2022.

[2] U. Shaham, T. Zahavy, C. Caraballo, S. Mahajan, D. Massey, and H. Krumholz, "Learning to Ask Medical Questions using Reinforcement Learning," 2020. Version Number: 2.

[3] E. J. Topol, *High-performance medicine: the convergence of human and artificial intelligence*. Nature Medicine, 2019.

[4] G. Litjens, T. Kooi, B. E. Bejnordi, A. A. A. Setio, F. Ciompi, M. Ghafoorian, J. A. van der Laak, B. Van Ginneken, and C. I. Sánchez, "A survey on deep learning in medical image analysis," *Medical Image Analysis*, vol. 42, pp. 60–88, 2017.

[5] R. T. Sutton, D. Pincock, D. C. Baumgart, D. C. Sadowski, R. N. Fedorak, and K. I. Kroeker, "An overview of clinical decision support systems: benefits, risks, and strategies for success," *NPJ Digital Medicine*, vol. 3, 2020.

[6] P. Wang, H. Liu, and M. Xu, "An adaptive testing item selection strategy via a deep reinforcement learning approach," *Behavior Research Methods*, vol. 56, pp. 8695–8714, 2024.

[7] R. S. Sutton and A. G. Barto, *Reinforcement learning: An introduction*. MIT press, 2018.

[8] R. Bellman, *A Markovian decision process*. Indiana University Mathematics Journal, 1957.

[9] M. L. Puterman, *Markov decision processes: discrete stochastic dynamic programming*. John Wiley & Sons, 2014.

[10] S. G. Finlayson, J. D. Bowers, J. Ito, J. L. Zittrain, A. L. Beam, and I. S. Kohane, "Adversarial attacks on medical machine learning," *Science*, vol. 363, no. 6433, pp. 1287–1289, 2019.

[11] N. Papernot, "A marauder's map of security and privacy in machine learning," *CoRR*, vol. abs/1811.01134, 2018.

[12] B. Biggio, I. Corona, D. Maiorca, B. Nelson, N. Šrndic, P. Laskov, G. Giacinto, and F. Roli, "Evasion attacks against machine learning at test time," *Machine Learning and Knowledge Discovery in Databases*, pp. 387–402, 2013.

[13] M. Barreno, B. Nelson, A. D. Joseph, and J. D. Tygar, "The security of machine learning," *Machine Learning*, vol. 81, no. 2, pp. 121–148, 2010.

[14] D. Amodei, C. Olah, J. Steinhardt, P. Christiano, J. Schulman, and D. Mané, "Concrete problems in ai safety," *arXiv preprint arXiv:1606.06565*, 2016.

[15] S. Russell, *Human compatible: Artificial intelligence and the problem of control*. Viking, 2019.

[16] European Commission, "Regulation of the european parliament and of the council laying down harmonised rules on artificial intelligence (artificial intelligence act) and amending certain union legislative acts," 2021. COM(2021) 206 final.

[17] Regenstrief Institute, "LOINC: Logical observation identifiers names and codes." https://loinc.org, 2024. Laboratory data coding standard.

[18] SNOMED International, "SNOMED CT: Systematized nomenclature of medicine clinical terms." https://www.snomed.org, 2024. International clinical terminology standard.

[19] World Health Organization, "ICD-11: International classification of diseases 11th revision." https://icd.who.int, 2024. WHO diagnostic coding standard.

[20] American Diabetes Association Professional Practice Committee, "Standards of care in diabetes-2025," *Diabetes Care*, vol. 48, no. Supplement_1, pp. S1–S204, 2025.

[21] F. Croce and M. Hein, "Reliable evaluation of adversarial robustness with an ensemble of diverse parameter-free attacks," 2020.

[22] C. Szegedy, W. Zaremba, I. Sutskever, J. Bruna, D. Erhan, I. Goodfellow, and R. Fergus, "Intriguing properties of neural networks," *arXiv preprint arXiv:1312.6199*, 2013.

[23] I. J. Goodfellow, J. Shlens, and C. Szegedy, "Explaining and harnessing adversarial examples," *arXiv preprint arXiv:1412.6572*, 2014.

[24] A. Madry, A. Makelov, L. Schmidt, D. Tsipras, and A. Vladu, "Towards deep learning models resistant to adversarial attacks," in *International Conference on Learning Representations*, 2018.

[25] N. Papernot, P. McDaniel, I. Goodfellow, S. Jha, Z. B. Celik, and A. Swami, "Practical black-box attacks against machine learning," in *Proceedings of the 2017 ACM Asia Conference on Computer and Communications Security (ASIACCS '17)*, (Abu Dhabi, United Arab Emirates), pp. 506–519, ACM, 2017.

[26] Y. Liu, X. Chen, C. Liu, and D. Song, "Delving into transferable adversarial examples and black-box attacks," in *5th International Conference on Learning Representations, ICLR 2017, Toulon, France, April 24-26, 2017, Conference Track Proceedings*, OpenReview.net, 2017.

[27] A. Ilyas, L. Engstrom, A. Athalye, and J. Lin, "Black-box adversarial attacks with limited queries and information," in *Proceedings of the 35th International Conference on Machine Learning* (J. Dy and A. Krause, eds.), vol. 80 of *Proceedings of Machine Learning Research*, pp. 2137–2146, PMLR, 10–15 Jul 2018.

[28] C.-C. Tu, P. Ting, P.-Y. Chen, S. Liu, H. Zhang, J. Yi, C.-J. Hsieh, and S.-M. Cheng, "Autozoom: autoencoder-based zeroth order optimization method for attacking black-box neural networks," in *Proceedings of the Thirty-Third AAAI Conference on Artificial Intelligence and Thirty-First Innovative Applications of Artificial Intelligence Conference and Ninth AAAI Symposium on Educational Advances in Artificial Intelligence*, AAAI'19/IAAI'19/EAAI'19, AAAI Press, 2019.

[29] P.-Y. Chen, H. Zhang, Y. Sharma, J. Yi, and C.-J. Hsieh, "Zoo: Zeroth order optimization based black-box attacks to deep neural networks without training substitute models," in *Proceedings of the 10th ACM workshop on artificial intelligence and security*, pp. 15–26, 2017.

[30] S. H. Huang, N. Papernot, I. J. Goodfellow, Y. Duan, and P. Abbeel, "Adversarial attacks on neural network policies," *CoRR*, vol. abs/1702.02284, 2017.

[31] Y.-C. Lin, Z.-W. Hong, Y.-H. Liao, M.-L. Shih, M.-Y. Liu, and M. Sun, "Tactics of adversarial attack on deep reinforcement learning agents," *arXiv preprint arXiv:1703.06748*, 2017.

[32] A. Gleave, M. Dennis, N. Kant, C. Wild, S. Levine, and S. Russell, "Adversarial policies: Attacking deep reinforcement learning," *CoRR*, vol. abs/1905.10615, 2020.

[33] Y. Zhao, I. Shumailov, H. Cui, X. Gao, R. Mullins, and R. Anderson, "Blackbox attacks on reinforcement learning agents using approximated temporal information." arXiv:1909.02918, 2019. arXiv preprint.

[34] P. Kamalaruban, Y.-T. Huang, Y.-P. Hsieh, P. Rolland, C. Shi, and V. Cevher, "Robust reinforcement learning via adversarial training with langevin dynamics," in *Advances in Neural Information Processing Systems 33 (NeurIPS 2020)*, pp. 8127–8138, 2020.

[35] M. Welling and Y. W. Teh, "Bayesian learning via stochastic gradient langevin dynamics," in *Proceedings of the 28th international conference on machine learning (ICML-11)*, pp. 681–688, 2011.

[36] C. Zhang, S. Bengio, M. Hardt, B. Recht, and O. Vinyals, "Understanding deep learning requires rethinking generalization," *Communications of the ACM*, vol. 64, no. 3, pp. 107–115, 2021. Originally published in ICLR 2017.

[37] X. Ma, Y. Niu, L. Gu, Y. Wang, Y. Zhao, J. Bailey, and F. Lu, "Understanding adversarial attacks on deep learning based medical image analysis systems," *Pattern Recognition*, vol. 110, p. 107332, 2020.

[38] U. Ozbulak, Y. Aytar, and H. K. Ekenel, "Impact of adversarial examples on deep learning models for biomedical image segmentation," *Medical Image Analysis*, vol. 65, p. 101768, 2019.

[39] Q. Yao, Z. He, Y. Lin, K. Ma, Y. Zheng, and S. K. Zhou, "A hierarchical feature constraint to camouflage medical adversarial attacks," in *Medical Image Computing and Computer Assisted Intervention–MICCAI 2021*, pp. 36–47, Springer, 2021.

[40] X. Qi, T. Xie, Y. Li, R. Zeman, V. Srikumar, and P.-Y. Chen, "Stabilized medical image attack," *arXiv preprint arXiv:2103.09531*, 2021.

[41] N. Carlini and D. Wagner, "Towards evaluating the robustness of neural networks," in *2017 ieee symposium on security and privacy (sp)*, pp. 39–57, IEEE, 2017.

[42] V. Mnih, K. Kavukcuoglu, D. Silver, A. A. Rusu, J. Veness, M. G. Bellemare, A. Graves, M. Riedmiller, A. K. Fidjeland, G. Ostrovski, *et al.*, "Human-level control through deep reinforcement learning," *Nature*, vol. 518, no. 7540, pp. 529–533, 2015.

[43] L.-J. Lin, "Self-improving reactive agents based on reinforcement learning, planning and teaching," *Machine Learning*, vol. 8, no. 3-4, pp. 293–321, 1992.

[44] X. Glorot and Y. Bengio, "Understanding the difficulty of training deep feedforward neural networks," in *Proceedings of the thirteenth international conference on artificial intelligence and statistics*, pp. 249–256, JMLR Workshop and Conference Proceedings, 2010.

[45] N. Carlini, A. Athalye, N. Papernot, W. Brendel, J. Rauber, D. Tsipras, I. Goodfellow, A. Madry, and A. Kurakin, "Evaluating the robustness of neural networks: An extreme value theory approach," *arXiv preprint arXiv:1905.03493*, 2019.