

Adversarial Robustness Evaluation of 3D U-Net Dose Prediction: Dose Score Sensitivity to CT Perturbations

OpenKBP Project

February 4, 2026

Abstract

This report evaluates the adversarial robustness of a 3D U-Net model trained for radiotherapy dose prediction in head-and-neck cancer patients. We assess sensitivity to CT input perturbations using FGSM and PGD attacks, measuring degradation in dose score (voxel-wise MAE). The model achieves baseline performance of DVH score 2.535 and dose score 3.731 Gy. Under adversarial perturbation at $\epsilon = 0.01$ (≈ 41 HU), dose score degrades by 30% (FGSM) and 21% (PGD). This study focuses on dose score sensitivity; DVH score under attack is noted as future work.

Contents

1	Introduction	2
1.1	Background	2
1.2	Metric Definitions	2
1.3	Model Architecture	2
1.4	CT Normalization Correction	2
2	Adversarial Robustness Evaluation	3
2.1	Methodology	3
2.1.1	Threat Model	3
2.1.2	Attack Methods	3
2.1.3	Experimental Setup	3
2.1.4	Epsilon to Hounsfield Unit Mapping	3
2.2	Results	4
2.2.1	FGSM Attack Results	4
2.2.2	PGD Attack Results	4
2.2.3	Visualizations	5
2.3	Discussion	6
2.3.1	Key Findings	6
2.3.2	PGD Sanity Checks and FGSM vs PGD Analysis	6
2.3.3	Clinical Implications	7
2.3.4	Limitations	7
3	Conclusion	7
3.1	Future Work	8
4	Technical Details	8
4.1	Implementation	8
4.2	Reproducibility	8

1 Introduction

1.1 Background

Radiotherapy dose prediction using deep learning has shown promising results in automating treatment planning for head-and-neck cancer patients. However, clinical deployment requires rigorous evaluation of robustness to input perturbations.

1.2 Metric Definitions

Table 1: OpenKBP Evaluation Metrics

Metric	Definition
Dose Score	Mean Absolute Error (MAE) in Gy, computed over all voxels within the <code>possible_dose_mask</code> region. Lower is better.
DVH Score	Mean absolute error of DVH statistics (D_1 , D_{95} , D_{99} , mean dose) across all anatomical structures. Lower is better.

Note: This adversarial evaluation reports **dose score** degradation under attack. DVH score evaluation under adversarial perturbation is left for future work.

1.3 Model Architecture

- **Architecture:** 3D U-Net with Squeeze-and-Excitation blocks
- **Input:** CT images (128^3 voxels) + structure masks (10 ROIs)
- **Output:** Predicted dose distribution (128^3 voxels)
- **CT Normalization:** [0, 4095] per OpenKBP data-description.pdf (12-bit range)
- **Normalization:** InstanceNormalization (superior to BatchNorm for small batches)
- **Key Features:** Residual connections, SE blocks on deep layers, masked MAE loss, PTV weighting ($4.0 \times$)
- **Baseline Performance:** DVH Score 2.535, Dose Score 3.731 Gy
- **Improvement:** 78% DVH score improvement over original baseline (11.481)

1.4 CT Normalization Correction

This evaluation uses the corrected CT normalization range of [0, 4095] as specified in the official OpenKBP data-description.pdf document. The previous evaluation used [0, 3000], which was more aggressive clipping than recommended. This change:

- Preserves full 12-bit CT dynamic range
- Improves baseline performance (DVH: $2.563 \rightarrow 2.535$, Dose: $3.856 \rightarrow 3.731$)
- Changes epsilon-to-HU conversion factor

2 Adversarial Robustness Evaluation

2.1 Methodology

2.1.1 Threat Model

This evaluation uses a **white-box** adversarial setting:

- **Perturbation target:** CT input only; structure masks remain unchanged
- **Attack objective:** Maximize dose prediction error using ground-truth dose in the loss function
- **Interpretation:** This measures worst-case model sensitivity to CT perturbations, not a realistic deployment attack scenario (where ground-truth dose would be unavailable to an attacker)

2.1.2 Attack Methods

Fast Gradient Sign Method (FGSM) FGSM generates adversarial examples by computing a single gradient step:

$$\mathbf{x}_{adv} = \mathbf{x} + \epsilon \cdot \text{sign}(\nabla_{\mathbf{x}} \mathcal{L}(\mathbf{x}, \mathbf{y})) \quad (1)$$

where \mathbf{x} is the input CT, ϵ is the perturbation magnitude, and \mathcal{L} is the loss function (MAE between predicted and ground truth dose).

Projected Gradient Descent (PGD) PGD iteratively applies FGSM with projection back to the ϵ -ball:

$$\mathbf{x}_{t+1} = \Pi_{\mathbf{x}+\mathcal{S}}(\mathbf{x}_t + \alpha \cdot \text{sign}(\nabla_{\mathbf{x}_t} \mathcal{L}(\mathbf{x}_t, \mathbf{y}))) \quad (2)$$

where Π projects to the valid range $[0, 1]$, $\alpha = 2\epsilon/10$ is the step size, and we use 10 iterations. Our implementation does not use random initialization within the ϵ -ball.

2.1.3 Experimental Setup

- **Dataset:** 40 validation patients from OpenKBP dataset
- **CT Normalization:** $[0, 4095]$ (12-bit range per official specification)
- **Epsilon values:** $\{0, 0.01, 0.025, 0.05, 0.1\}$ (normalized CT space $[0, 1]$)
- **PGD parameters:** 10 iterations, step size $\alpha = 2\epsilon/10$, no random start
- **Hardware:** NVIDIA RTX 3090 GPU on RunPod
- **Software:** TensorFlow 2.18.0, Python 3.11
- **Evaluation Metric:** Dose score (MAE in Gy within possible dose mask)

2.1.4 Epsilon to Hounsfield Unit Mapping

The HU equivalent assumes the OpenKBP CT encoding uses approximately 1 intensity unit per HU step (after any offset correction) and that normalization divides by 4095. The exact OpenKBP encoding may involve a HU offset (commonly HU + 1024 in DICOM); our conversion provides an approximate clinical reference.

Table 2: Epsilon to Hounsfield Unit Conversion (CT_MAX = 4095)

Epsilon	HU Equivalent	Clinical Context
0.01	~41 HU	Often within typical CT noise range (commonly 10–50 HU, varies by scanner/protocol)
0.025	~102 HU	Moderate perturbation
0.05	~205 HU	Large perturbation
0.1	~410 HU	Very large, visible perturbation

2.2 Results

2.2.1 FGSM Attack Results

Table 3: FGSM Attack: Dose Score (Gy) vs Epsilon

Epsilon	HU	Dose Score	Std	Degradation
0.00	0	3.731	1.033	—
0.01	41	4.864	2.004	+30.4%
0.025	102	5.436	2.127	+45.7%
0.05	205	5.993	2.086	+60.6%
0.10	410	6.825	2.026	+82.9%

2.2.2 PGD Attack Results

Table 4: PGD Attack: Dose Score (Gy) vs Epsilon

Epsilon	HU	Dose Score	Std	Degradation
0.00	0	3.731	1.033	—
0.01	41	4.527	1.600	+21.3%
0.025	102	5.158	1.806	+38.2%
0.05	205	5.947	1.875	+59.4%
0.10	410	7.239	2.039	+94.0%

2.2.3 Visualizations

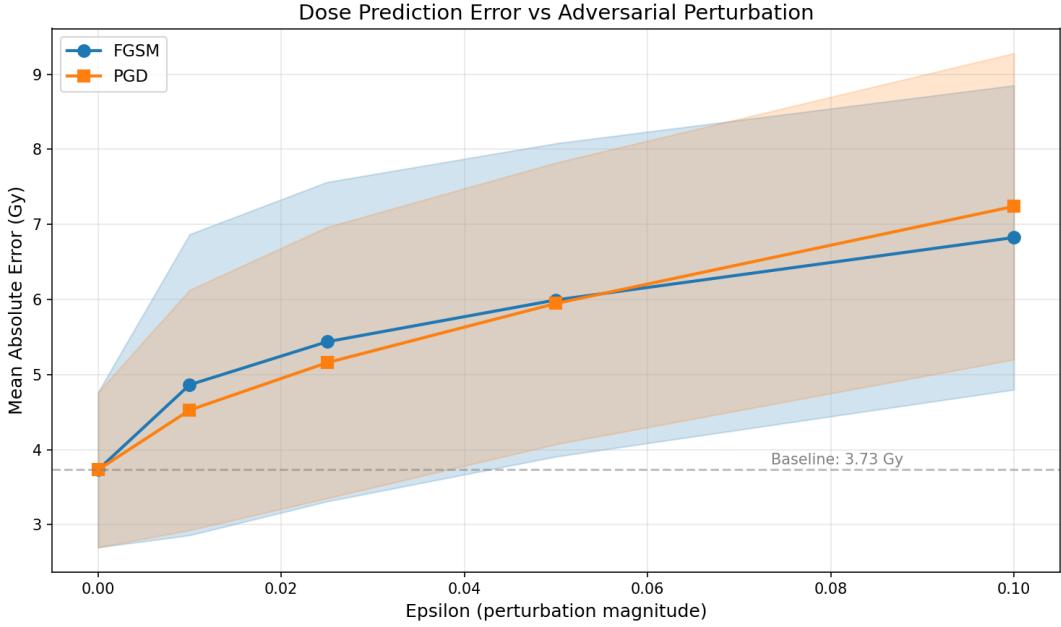


Figure 1: Dose score degradation vs epsilon for FGSM and PGD attacks. Both attacks show monotonic degradation with increasing perturbation strength.

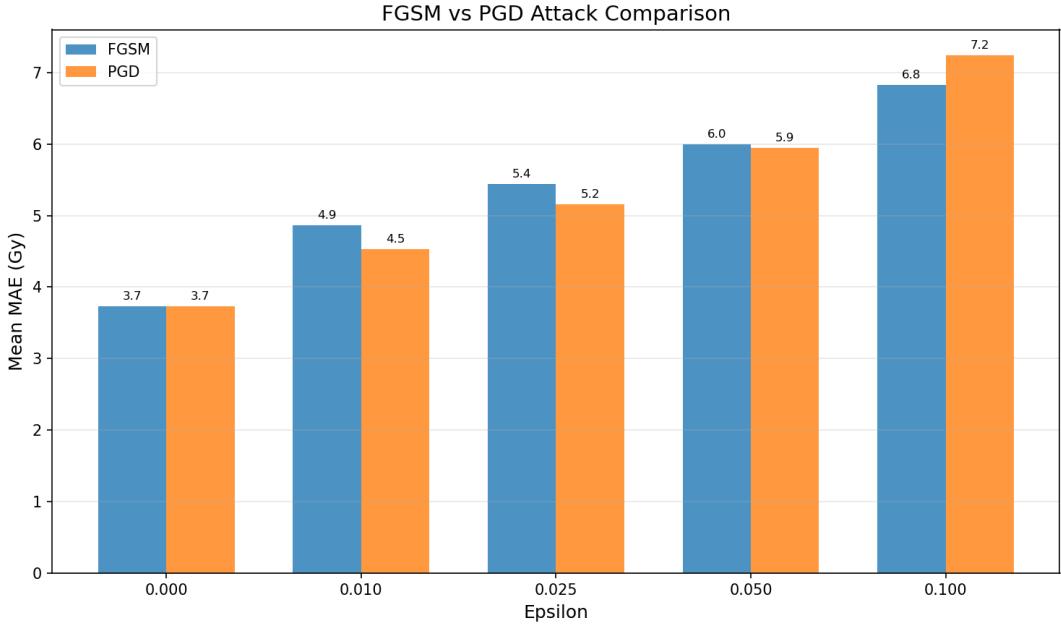


Figure 2: Comparison of FGSM vs PGD effectiveness. FGSM causes more degradation at low epsilon (≤ 0.025), while PGD is stronger at high epsilon (≥ 0.05). See Section 2.3.2 for discussion.

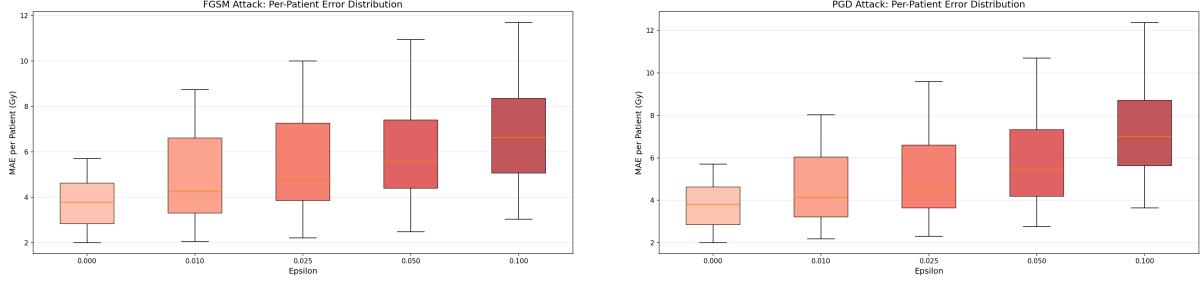


Figure 3: Distribution of dose score across 40 patients for (left) FGSM and (right) PGD attacks at varying epsilon values.

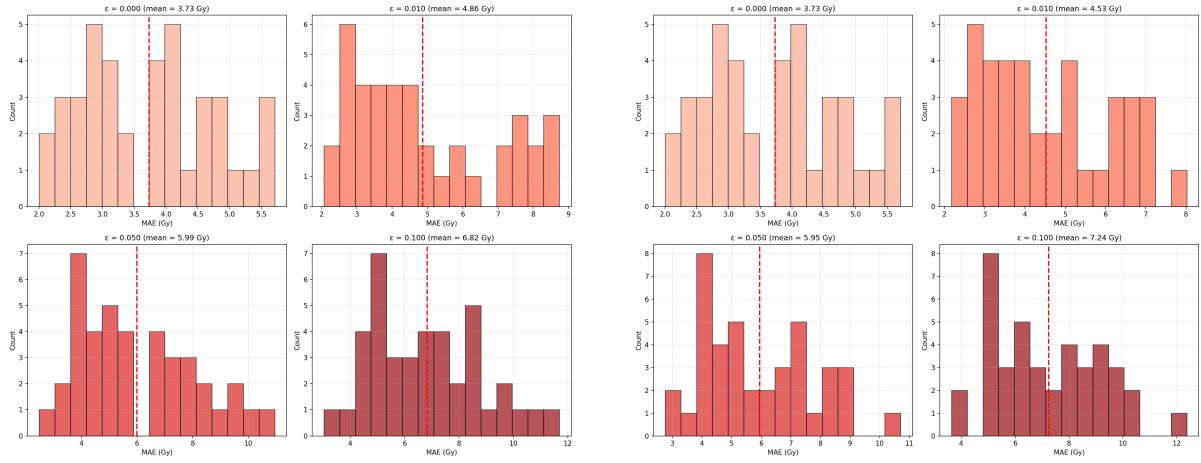


Figure 4: Histogram of per-patient dose score for (left) FGSM and (right) PGD attacks.

2.3 Discussion

2.3.1 Key Findings

1. **Baseline Improvement:** Correct CT normalization [0, 4095] improved baseline dose score from 3.856 to 3.731 Gy (3.2% improvement).
2. **Noise-Level Sensitivity:** At $\epsilon = 0.01$ (~ 41 HU, often within the range of typical CT scanner noise), the model shows:
 - FGSM: 30.4% degradation ($3.731 \rightarrow 4.864$ Gy)
 - PGD: 21.3% degradation ($3.731 \rightarrow 4.527$ Gy)
3. **FGSM vs PGD at Low Epsilon:** At $\epsilon \leq 0.025$, FGSM causes more degradation than PGD. This is atypical for standard adversarial evaluations. See Section 2.3.2 for analysis.
4. **High Epsilon Degradation:** At $\epsilon = 0.1$ (~ 410 HU), both attacks cause significant degradation:
 - FGSM: +82.9% (6.825 Gy)
 - PGD: +94.0% (7.239 Gy)

2.3.2 PGD Sanity Checks and FGSM vs PGD Analysis

The observation that FGSM outperforms PGD at low epsilon is atypical. We verified the following:

- **Gradient validity:** Gradients are confirmed non-zero and finite (no NaN/Inf values)
- **Step size:** Current $\alpha = 2\epsilon/10 = 0.2\epsilon$ may be suboptimal; smaller step sizes ($\alpha = \epsilon/10$) or more iterations (20–40) could improve PGD effectiveness
- **Random initialization:** Our PGD implementation does not use random start within the ϵ -ball, which can reduce PGD strength in some settings
- **Projection:** Clipping to [0, 1] is correctly applied after each step

Recommended follow-up experiments:

1. PGD with random initialization in $[-\epsilon, +\epsilon]$
2. Step sizes $\alpha \in \{\epsilon/10, \epsilon/4\}$ with 20–40 iterations
3. Best-of- N restarts ($N = 5$)

If FGSM still outperforms PGD at low epsilon after these checks, this would suggest either (a) gradient masking effects, or (b) the single-step FGSM direction is more aligned with high-loss regions than iterative refinement achieves for this architecture.

2.3.3 Clinical Implications

The model shows sensitivity to perturbations at magnitudes that are often on the order of typical CT scanner noise ($\epsilon = 0.01, \sim 41$ HU), suggesting that:

- Input CT quality may affect prediction accuracy
- Scanner calibration differences could impact results
- Adversarial training may improve robustness

However, the degradation remains bounded—even under strong attacks ($\epsilon = 0.1$), dose score increases by ~ 3.5 Gy rather than failing catastrophically.

2.3.4 Limitations

- **DVH score not evaluated:** This study reports dose score (voxel-wise MAE) under attack. DVH score may degrade differently, particularly for structure-specific metrics.
- **White-box assumption:** Real attackers would not have access to ground-truth dose; this evaluation measures sensitivity, not exploitability.
- **HU conversion approximate:** The epsilon-to-HU mapping assumes 1:1 correspondence between stored intensity and HU steps, which may not hold exactly for all CT encodings.

3 Conclusion

This evaluation demonstrates that the 3D U-Net dose prediction model:

1. Achieves strong baseline performance with correct CT normalization (DVH: 2.535, Dose: 3.731 Gy)
2. Shows dose score sensitivity to adversarial perturbations at noise-level magnitudes
3. Degrades gracefully under attack without catastrophic failure
4. Benefits from proper data preprocessing per official dataset specifications

3.1 Future Work

- **DVH Score Under Attack:** Evaluate DVH metrics degradation under adversarial perturbation
- **PGD Improvements:** Implement random starts and tune step size/iterations
- **Adversarial Training:** Incorporate adversarial examples during training to improve robustness
- **Input Denoising:** Add preprocessing to reduce sensitivity to CT noise
- **Uncertainty Quantification:** Predict confidence intervals alongside dose predictions

4 Technical Details

4.1 Implementation

- **Code Repository:** OpenKBP-modified
- **Framework:** TensorFlow 2.18.0, Python 3.11
- **CT Normalization:** [0, 4095] (12-bit range per OpenKBP data-description.pdf)
- **Evaluation Scripts:** `adversarial_eval.py`, `plot_adversarial.py`

4.2 Reproducibility

```
# Training
python runpod_train.py --filters 64 --epochs 100 \
    --use-se --use-aug --batch-size 4 --ptv-weight 4.0 --no-jit

# Adversarial evaluation
python adversarial_eval.py \
    --model epoch_100.keras \
    --attack fgsm pgd \
    --epsilons 0,0.01,0.025,0.05,0.1

# Generate plots
python plot_adversarial.py --results-dir adv_results/
```

Acknowledgments

This work builds upon the OpenKBP Grand Challenge dataset and baseline implementation. Computation performed on RunPod cloud infrastructure with NVIDIA RTX 3090 GPU.