2026 JULY 19–22 VANCOUVER, BC
JOINT AAPM | COMP MEETING

Adversarial Robustness Evaluation of AI Based Dose Prediction Models Using Clinical DVH Metrics

Rahim Chowdhury[1], Neil Tripathi[2], Lei Ren[2], Amit Sawant[2], Weixian Liao[1], Birjoo Vaishnav[2]
[1] Veterans Health Administration, Baltimore MD
[2] University of Maryland School of Medicine, Baltimore MD

ABSTRACT

Deep learning–based dose prediction models are increasingly used in radiotherapy, yet their robustness to input uncertainty remains poorly understood. We developed a framework to evaluate 3D head-and-neck dose prediction models by applying controlled adversarial perturbations to CT images and assessing changes in clinically relevant dose-volume metrics. Results demonstrated measurable degradation in target coverage and organ-at-risk sparing, even for small perturbations, with effects increasing systematically as perturbation magnitude increased. These findings highlight the importance of incorporating robustness testing and uncertainty assessment into the development and clinical evaluation of AI-based dose prediction models

CONTACT

Birjoo Vaishnav, PhD, DABR

University of Maryland School of Medicine,
Baltimore MD

bvaishnav@som.umaryland.edu

410-427-2039(o)

INTRODUCTION

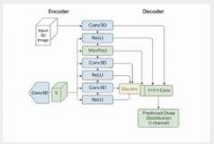
Deep learning-based treatment planning models are increasingly being deployed clinically, yet their evaluation is typically limited to prediction accuracy under standard test conditions. Consequently, model robustness to input variability and uncertainty remains poorly understood. Accuracy-based assessments may overlook vulnerabilities that can affect clinically relevant outputs.

To stress-test model performance, we apply FGSM and PGD adversarial perturbations to CT inputs. These methods generate small, targeted input changes designed to maximize output deviations, enabling systematic evaluation of robustness and model sensitivity.

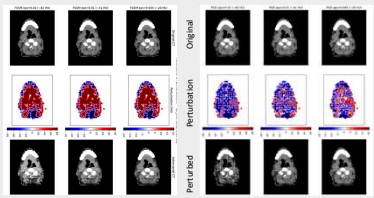
METHODS AND MATERIALS

Data set
Our framework is created using the
Open Knowledge-Based Planning (OpenKBP) dataset, a publicly available benchmark for radiotherapy dose prediction.
340 Head and Neck Patients
Anonymized CT Images with PTV & OAR Contours
70/63/56 Gy Prescriptions
Organs at Risk contoured:
Brainstem, Spinal Cord, Parotids, Larynx, Mandible, Esophagus

Dose Prediction Model
3D U-Net architecture
Data split: 200 Training, 40 validation, 100 test plans
Encoder-decoder network with skip connections
Inputs: CT image + structure masks
Output: voxel-wise 3D dose distribution
Trained using mean squared error loss between predicted and clinical dose maps



Adversarial perturbation model :
Adversarial perturbations in test input data are optimized to maximally impact model output(3).
Below: Perturbations for epsilon 0.02,0.01, 0.005
Fast Gradient Sign Method Post Gradient Descent



RESULTS

- Measurable dosimetric changes were observed even at the smallest perturbation level (ε = 0.001), with decreases in PTV coverage metrics and increases in hotspot metrics.
- Increasing perturbation magnitude produced progressive degradation in target coverage, with D95 reductions reaching 6-9% at ε = 0.01 and exceeding 15% at ε = 0.05.
- OAR endpoints demonstrated similar sensitivity, with notable increases in D0.1cc, V30, and V50, indicating reduced sparing robustness at higher perturbation levels.
- Both FGSM and PGD attacks resulted in increasing prediction error and performance degradation as ε increased, with PGD consistently producing larger effects than FGSM.
- The greatest degradation occurred at the highest perturbation magnitudes, demonstrating that deep learning dose prediction models are vulnerable to structured input perturbations.

Key Finding: Small CT intensity perturbations can lead to clinically meaningful changes in predicted dose distributions despite strong baseline model performance.

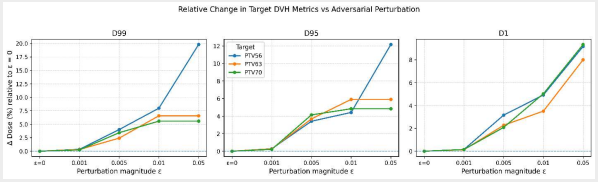


Figure 1 demonstrates that perturbations induce asymmetric responses across regions of the target DVH. Coverage-related metrics exhibit progressive deviation with increasing perturbation strength, while hotspot-related metrics follow a distinct and monotonic escalation pattern. These behaviors highlight that acceptable average performance can coexist with clinically relevant endpoint drift, underscoring the need for robustness-aware evaluation.

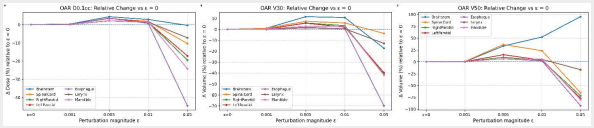


Figure 2 above demonstrates OAR robustness under adversarial perturbations. Relative percentage change in OAR dose and volume metrics with respect to the unperturbed adversarial baseline (ε = 0): (A) D0.1cc, (B) V30, and (C) V50, shown for multiple organs across increasing perturbation magnitudes ε. Volume-based endpoints exhibit greater sensitivity than maximum-dose metrics at higher ε.

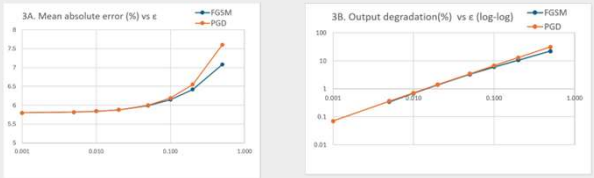


Figure 3. Robustness analysis of the dose prediction model under adversarial perturbations. (A) Mean absolute error (MAE) between predicted and reference dose distributions as a function of perturbation magnitude (ε) for FGSM and PGD attacks. (B) Relative performance degradation as a function of perturbation magnitude shown on a log-log scale. Both metrics increase with perturbation strength, while PGD produces consistently greater performance degradation than FGSM, indicating increased model vulnerability to iterative adversarial perturbations.

REFERENCES

1. Babier A, Mahon R, McNiven A, Diamant A, Chan TCY. Med Phys. 2021;48:4932-4948.
2. Ronneberger O, Fischer P, Brox T. MICCAI. 2015;234-241.
3. Goodfellow IJ, Shlens J, Szegedy C. ICLR. 2015.

DISCUSSION

Deep learning dose prediction models demonstrated increasing sensitivity to CT perturbations as adversarial magnitude increased. While overall prediction accuracy remained acceptable under nominal conditions, clinically relevant DVH metrics showed measurable degradation even at small perturbation levels. PGD attacks produced greater degradation than FGSM attacks, suggesting that iterative perturbations more effectively exploit vulnerabilities in the learned dose mapping. These findings indicate that conventional accuracy metrics alone may not adequately characterize model reliability and that robustness testing can reveal important failure modes not observed during standard validation.

CONCLUSIONS

We adopted adversarial perturbation as a diagnostic tool for robustness of AI models in radiotherapeutic dose prediction. From this study few key points emerge relevant to clinical practice.

Accurate AI models are not necessarily robust AI models.

Small CT perturbations can produce measurable dosimetric changes.

Robustness testing can identify vulnerabilities not captured by standard validation.

Incorporating robustness evaluation may improve confidence in clinical deployment of AI-based planning tools.

FUTURE WORK

The noise models in this work may not represent actual physical noise, impact of which will be presented in ASTRO 2026 as a poster presentation.