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

Multimodal Cross-Attention-Based Predictive Model with Self-Attention Refinement for Lymphedema Risk Following Breast 3D-CRT Using 3D CT and Voxel-Wise Dose Distributions

Jae Hyun Seok[1, 2]; Hojin Kim, PhD[1*]; Jee Suk Chang, MD[1]; Jin Sung Kim, PhD[1, 2, 3]

[1]Department of Radiation Oncology, Yonsei Cancer Center, Heavy Ion Therapy Research Institute, Yonsei University College of Medicine, Seoul, Korea

[2]Medical Physics and Biomedical Engineering Lab (MPBEL), Yonsei University College of Medicine, Seoul, South Korea

[3]Oncosoft Inc., Seoul, South Korea

ABSTRACT

Purpose: To mitigate the loss of spatial information inherent to DVH-based and coarse categorical descriptors used for breast cancer–related lymphedema prediction, we present a multi-modal cross-attention predictive model with self-attention refinement using patient-specific 3D CT anatomy and dose distributions

Material and method: We developed a voxel-level multimodal architecture comprising modality-specific 3D DenseNet encoders for CT and dose, an intra-modality self-attention module for representation refinement, and a bidirectional cross-attention fusion module to explicitly learn anatomical–dosimetric interactions (CT and dose). To isolate architectural contributions, we performed controlled benchmarking of intra-modality self-attention variants (bottleneck attention module (BAM), convolution block attention module (CBAM), multi-head self-attention (MHSA), and a vision transformer using SwinUNETR) under identical fusion settings. Inputs were standardized via axillary lymph node–centered alignment to ensure consistent inclusion of axillary Levels I–III within a fixed-depth volume, producing 512x512x64 3D volumes. The framework was evaluated on 119 right-sided breast cancer patients treated with 3D conformal radiotherapy using five repeated patient-level splits. Ablation analyses quantified the contribution of the best-performing self-attention variant and assessed the necessity of interaction modeling by removing cross-attention.

Results: Across controlled comparisons, BAM provided the most stable and discriminative performance. The full configuration (DenseNet3D+BAM+bidirectional cross-attention) achieved a mean AUC of 0.80 and a mean accuracy of 0.74. Relative to the DenseNet3D baseline with cross-attention (mean AUC 0.61; mean accuracy 0.60), BAM substantially improved discrimination. In contrast, removing cross-attention markedly degraded performance (mean AUC 0.54; mean accuracy 0.56), underscoring the importance of explicit CT–dose interaction learning. Grad-CAM and cross-attention visualizations consistently highlighted axillary lymph node Levels I–III and adjacent soft-tissue regions/dose patterns, supporting clinically plausible spatial focus.

Conclusion: Voxel-level integration of CT anatomy and 3D dose, coupled with systematic benchmarking of intra-modality self-attention and bidirectional cross-attention fusion, enables more accurate and interpretable lymphedema risk prediction beyond DVH-style summaries.

CONTACT

[JaeHyun Seok]
[Yonsei University College of Medicine]
[50-1, Yonsei-ro, Seoul, Republic of Korea]
[jhseok6768@yonsei.ac.kr]
[+82-10-9895-6768]

INTRODUCTION

Breast cancer–related lymphedema is a chronic complication associated with arm swelling, functional impairment, pain, and reduced quality of life [1, 2].

Existing prediction approaches mainly rely on clinical factors, irradiation-field categories, or DVH metrics, which compress complex 3D anatomy and dose distributions into coarse variables [3-5].

Such representations cannot fully capture spatial heterogeneity or voxel-level interactions between axillary anatomy and delivered radiation dose.

This study developed a multimodal framework using 3D CT and dose distributions, with intra-modality attention refinement and bidirectional cross-attention for spatially explicit lymphedema prediction.

METHODS AND MATERIALS

Dataset

- Retrospective cohort of 119 breast cancer patients treated with right-sided 3D CRT
- Lymphedema (LE) cases: 23 patients
- Five repeated patient-level splits:
  - Training set per split: 103 patients (15 LE)
  - Test set per split: 16 patients (8 LE)

Preprocessing

- Each patient had paired 3D CT and dose volumes.
- CT and dose volumes were normalized and resampled to 0.9765 × 0.9765 × 5 mm.
- Cropped to 512 × 512 × 64 centered on the axillary lymph node region

Model Configurations

➢ Multimodal Framework
- Separate encoders extracted modality-specific CT and dose features.
- Bidirectional cross-attention modeled CT-to-dose and dose-to-CT interactions.
- Fusion and classification settings were identical across all configurations.

Four self-attention variants were compared:

1. DenseNet3D + BAM
   - Parallel channel and spatial attention applied after DenseBlocks
2. DenseNet3D + CBAM
   - Sequential channel and spatial attention applied within DenseBlocks
3. DenseNet3D + MHSA
   - DenseNet features converted into spatial tokens
   - Global self-attention captured long-range context.
4. Swin Transformer
   - Hierarchical window-based self-attention
   - Replaced the DenseNet3D encoder

Hyperparameter

- Optimizer: Adam
- Loss: Binary cross-entropy

Evaluation

- AUC, accuracy, recall, precision

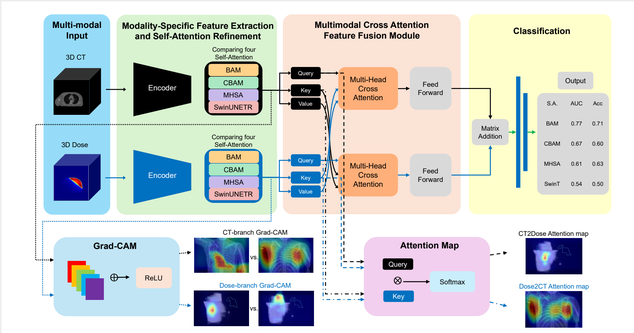

Figure 1. Overall architecture and interpretability workflow of the proposed multimodal attention-based framework.

RESULTS

Effect of Intra-Modality Attention

- BAM showed the most stable and discriminative performance among the evaluated attention variants.
- CBAM and MHSA showed greater split-to-split variability, while the Swin Transformer–based encoder produced the lowest overall performance.

Contribution of Attention Modules

- Removing either BAM or bidirectional cross-attention reduced predictive performance.
- The larger decrease without cross-attention indicates that explicit CT–dose interaction modeling was the primary contributor to multimodal prediction.
- BAM further improved modality-specific CT and dose representations before fusion.

Explainability Analysis

- Grad-CAM highlighted the ipsilateral axilla, including Levels I–III and adjacent soft tissues.
- Correctly predicted high-risk cases showed concentrated activation along the axillary lymphatic pathway.
- Misclassified cases showed displaced or less anatomically relevant activation.

Table 1. Performance comparison of intra-modality attention configurations across five patient-level splits.

Model	Accuracy	Recall	Precision	AUC	F1
DenseNet + BAM	**0.74**	**0.60**	**0.86**	**0.80**	**0.67**
DenseNet + CBAM	0.54	0.33	0.41	0.63	0.31
DenseNet + MHSA	0.63	0.30	0.83	0.61	0.44
Swin Transformer	0.48	0.20	0.50	0.48	0.28

Table 2. Ablation analysis of BAM refinement and bidirectional cross-attention.

Model	Accuracy	Recall	Precision	AUC	F1
DenseNet + BAM + Cross-attention (proposed)	**0.74**	**0.60**	**0.86**	**0.80**	**0.67**
DenseNet + Cross-attention	0.60	0.23	0.93	0.61	0.35
DenseNet + BAM	0.56	0.23	0.60	0.54	0.31

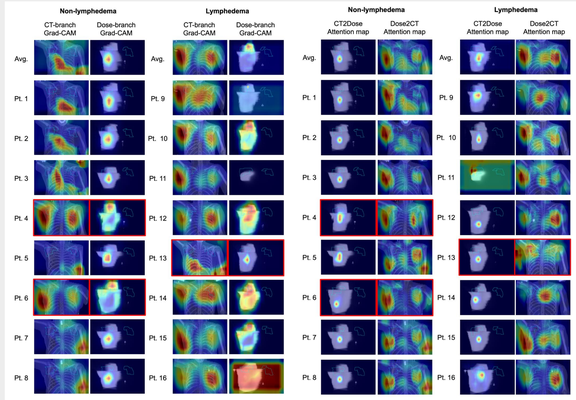

Figure 2. Representative Grad-CAM and bidirectional cross-attention maps from the best-performing DenseNet3D+BAM model. "Avg." denotes the mean map across correctly predicted cases, and "Pt." denotes individual patients. Red boxes indicate misclassified cases.

DISCUSSION

- BAM showed the most stable performance, suggesting that combined channel and spatial refinement was suitable for volumetric anatomy and smooth dose gradients.
- The greater performance loss after removing cross-attention indicates that explicit CT–dose interaction modeling was more important than simple feature aggregation.
- Grad-CAM localization in the ipsilateral axillary region supports the clinical plausibility of the learned representations.
- Further validation is required because of the limited single-institution cohort, class imbalance, restriction to right-sided cases, and lack of external validation.

CONCLUSIONS

- A multimodal framework using 3D CT and voxel-wise dose distributions was developed for lymphedema prediction after breast 3D-CRT.
- DenseNet3D with BAM and bidirectional cross-attention achieved the best overall performance.
- BAM refined modality-specific representations, while cross-attention modeled anatomical–dosimetric interactions beyond simple fusion.
- Grad-CAM localization in the ipsilateral axilla supported the anatomical plausibility of the model.

REFERENCES

1. Taghian NR, et al. Lymphedema following breast cancer treatment and impact on quality of life: a review. Crit Rev Oncol Hematol. 2014;92:227–234.
2. Rockson SG. Lymphedema after breast cancer treatment. N Engl J Med. 2018;379:1937–1944.
3. Byun HK, et al. Risk of lymphedema following contemporary treatment for breast cancer. Ann Surg. 2021;274:170–178.
4. Gross JP, et al. Determining the organ at risk for lymphedema after regional nodal irradiation in breast cancer. Int J Radiat Oncol Biol Phys. 2019;105:649–658.
5. Park YI, et al. Development and validation of a normal tissue complication probability model for lymphedema after radiation therapy in breast cancer. Int J Radiat Oncol Biol Phys. 2023;116:1218–1225.

© Poster Template by Genigraphics® 1.800.790.4001 www.genigraphics.com