Iowa Initiative for Artificial Intelligence

Final Report

Project title:	Can AI identify occult nodal metastasis in oral squamous cell carcinoma?
Principal Investigator:	Daniella Karassawa Zanoni
Prepared by (IIAI):	Honghai Zhang, Zhi Chen
Other investigators:	Danielli Matsuura, Bruno Policeni
Date:	10/01/2025

Were specific aims fulfilled:	Y
Readiness for extramural proposal?	Y
If yes … Planned submission date	Nov 2025 / Feb 2026
Funding agency	The Medical University of South Carolina
Grant mechanism	MUSC's intramural Idea award
If no … Why not? What went wrong?	N/A

Brief summary of accomplished results:

Trained and cross-validated on 75 3D CT images with 2638 lymph nodes identified, a nnUNet based approach was used to detect lymph nodes and produced >85% precision and recall on nodes with ≥3 mm radius and >90% precision and recall on nodes with ≥4 mm radius.

Trained and cross-validated on 17 3D CT image with 1226 lymph nodes identified and cervical level labeled, a XGBoost model was used to predict cervical levels based on the location and size of the nodes. The model achieved accurate predictions on 57% of the nodes and acceptable (close-enough) predictions on 93% of the nodes.

Research report:

Aims (provided by PI):

Head and neck lymph node metastasis is an important prognostic factor for head and neck squamous cell carcinoma disease and a predictor of tumor recurrence and decreased survival. Patients with positive nodal disease may be treated with neck dissection; however, the procedure carries morbidity and decreases the overall patient's quality of life. The ability to predict head and neck nodal involvement provides adequate treatment while avoiding the morbidity of neck dissection for patients without neck disease.

Aim 1: Automatically detect lymph nodes in 3D CT images.

Aim 2: Automatically classify lymph nodes into the cervical levels.

Data:

265 3D CT images of 265 patients diagnosed with oral squamous cell carcinomas and primarily treated with surgery, including a neck dissection, were collected. The in-plane resolution is 0.3-

0.5 mm. The slice distances are 0.3-0.5 mm (on 55% of the images), 0.7-1.0 mm (25%), and 2-5 mm (20%).

On 75 3D images, 2638 lymph nodes were located by the experts (DKZ, DM) and segmented using LOGISMOS.

On additional 17 3D images, 1226 lymph nodes were located, and their 16 cervical levels were assigned by the experts (DKZ, DM). These nodes were also segmented by LOGISMOS.

AI/ML Approach:

For Aim 1, nnUNet was utilized and its capabilities were tested on the 75 images containing 2638 lymph nodes (no cervical level labels) using 5-fold cross validation.

For Aim 2, ML approach, XGBoost, was utilized to achieve cervical level classification based on the location and size of the nodes.

Experimental methods, validation approach:

Data Preparation:

Upon IRB approval, the 2D DICOM images of patients were uploaded to XNAT after proper anonymization and were accessed by IIAI researchers. To achieve easy management and visualization, the converted 3D images (NIfTI format) were analyzed.

GUI Tool – LogismosByTemplate:

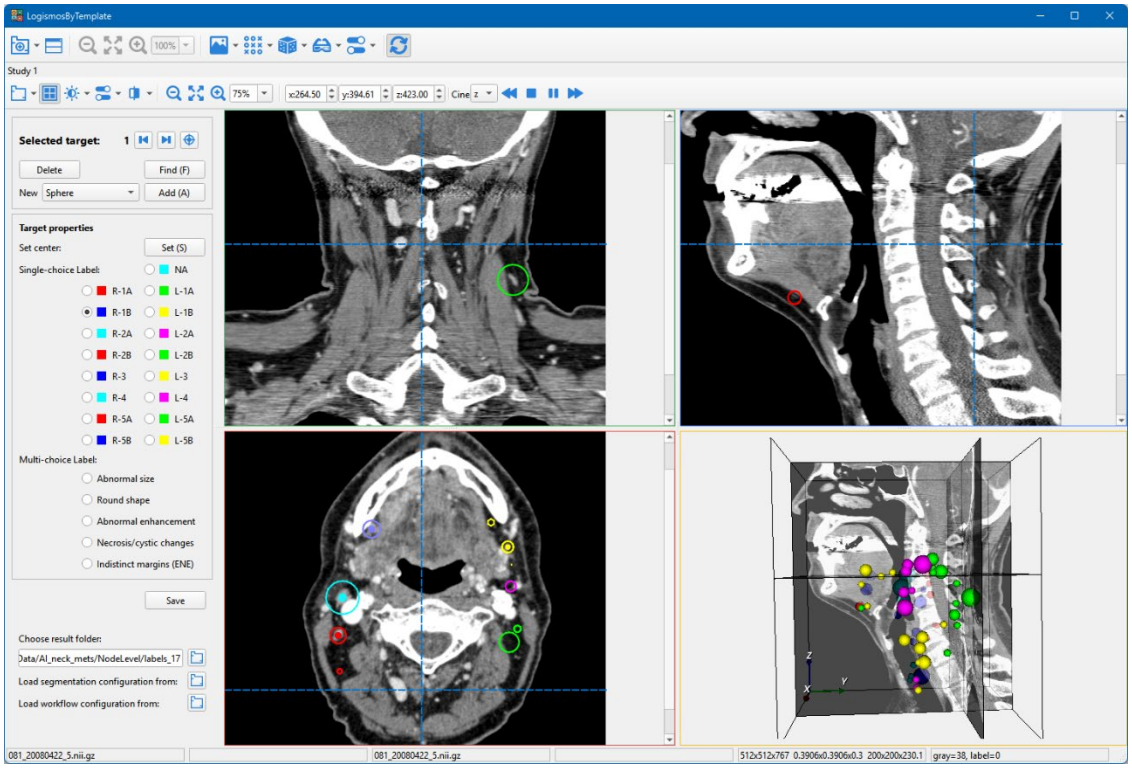

Figure 1. Screenshot of LogismosByTemplate being used to locate lymph nodes and assign cervical level labels.

A previously developed general-purpose application, LogismosByTemplate, was customized for this project. As shown in Fig. 1, it allows the clinical experts to visualize the 3D image, mark location of the lymph nodes (center of 3D spheres in Fig. 1), and assign cervical labels (e.g., R-1A) to the nodes. To speedup this manual process, the clinical experts (DKZ, DM) placed centers of fixed-size spheres inside the identified nodes. Later, IIAI researcher (HZ) used the same application (in a different operating mode) to segment the nodes after adjusting the size of the spheres to fully enclose the nodes in 3D if necessary.

nnUNet Node Segmentation:

We employed the nnUNet framework to segment lymph nodes in 75 neck CT scans with 5-fold cross validation. Preprocessing included voxel resampling and CT-specific intensity normalization. Three 3D models were configured: a low-resolution network trained on down-sampled (larger voxel size) volumes to capture global anatomical context, a full-resolution network trained at original resolution to preserve fine details, and a cascade model in which the low-resolution output guided a full-resolution refinement. Predictions from both models were ensembled and post-processed to refine the final outputs. This coarse-to-fine approach allowed the cascade to better detect small and subtle lesions that might be missed by single-stage models.

In this segmentation approach, the ground truth and AI result are represented by binary images such that a voxel is labeled as either background or lymph node without additional information to differentiate individual nodes. Traditional voxel-wise segmentation evaluation metrics such as Dice coefficient cannot fully evaluate the capability of node detection. The binary images were converted multi-label images such that each lymph node has its own unique label. Object-wise metrics – precision and recall – were derived from the multi-label images.

XGBoost Level Classification:

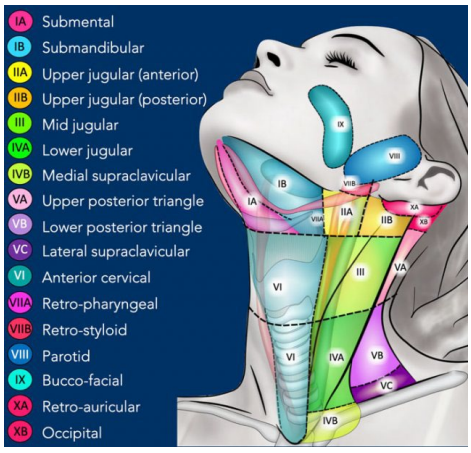

Figure 1. Cervical Lymph Node Map (https://radiologyassistant.nl/head-neck/cervical-node-mapping/cervical-node-map).

We trained an XGBoost classifier to predict lymph node levels from spatial features (node center coordinates and radius) derived from CT data. To avoid patient-level data leakage, we applied 5-fold cross-validation grouped by patient, using stratified grouping. Out-of-fold predictions were collected to compute performance metrics, including overall accuracy and per-class sensitivity, specificity, and accuracy.

A total of 16 lymph node levels were used: R-1A, L-1A, R-1B, L-1B, R-2A, L-2A, R-2B, L-2B, R-3, L-3, R-4, L-4, R-5A, L-5A, R-5B, L-5B. Their locations are shown in Figure 2 and R and L in the labels indicate the right and left sides of the neck.

The 3D coordinate of the lymph node centers and sizes (radii), both in millimeter physical space, were computed from the multi-label ground truth image. Note that the node radius is an estimation based on the volume of the node using $v = \frac{4}{3}\pi r^3$.

The performance of the classification was evaluated based on the true-vs-predicted confusion matrix. Based on the locations of the levels, the prediction errors are further categorized as follows.

- L-R errors: The prediction is on the wrong side of the neck.
- Non-neighbor errors: The predicted level is NOT spatially connected to the true level.
- Neighbor errors: The predicted level is spatially connected (neighbor) to the true level.

Results:

Aim 1:

Table 1. Performance of AI node detection.

	Ground Truth (G)				AI Detection (A)			
Node radius (mm)	# of nodes	True_G Positive	False Negative	Recall	# of nodes	True_A Positive	False Positive	Precision
all	2638	2012	626	0.76	3048	1983	1065	0.65
≥2	1942	1644	298	0.85	1904	1471	430	0.77
≥3	855	762	93	0.89	832	711	121	0.85
≥4	340	310	30	0.91	339	311	28	0.92
≥5	127	116	11	0.91	133	124	9	0.93
≥6	42	39	3	0.93	47	45	2	0.96
≥7	22	21	1	0.95	24	22	2	0.92

The performance of the AI node detection is listed in Table 1.

- Recall = True_G Positive / (True_G Positive + False Negative)
- Precision = True_A Positive / (True_A Positive + False Positive)

The AI node detection shows very good performance on lymph nodes with ≥3 mm radius. Although small nodes (<3 mm radius) constitute the majority (~70%) of the nodes in the training data, detecting those is difficult. The most probable cause is lack of information in the voxel

space between slices: 20% of the images have 2-5 mm slice distance and 25% with 0.7-1 mm slice distance.

The discrepancy between the highlighted 'True_G Positive' and 'True_A Positive' columns is caused by several factors. When a ground truth node is overlapped with any AI detected node, it is marked as belong to `True_G Positive` and vice versa. A G-node can overlap with more than one A-nodes and vice versa. In addition, the node radii are computed from the multi-label images of the 'G' and 'A' separately and thus could affect radius-based grouping. This discrepancy could be resolved by post-processing of AI results including visual inspection, object merging, and LOGISMOS segmentation.

Ami 2:

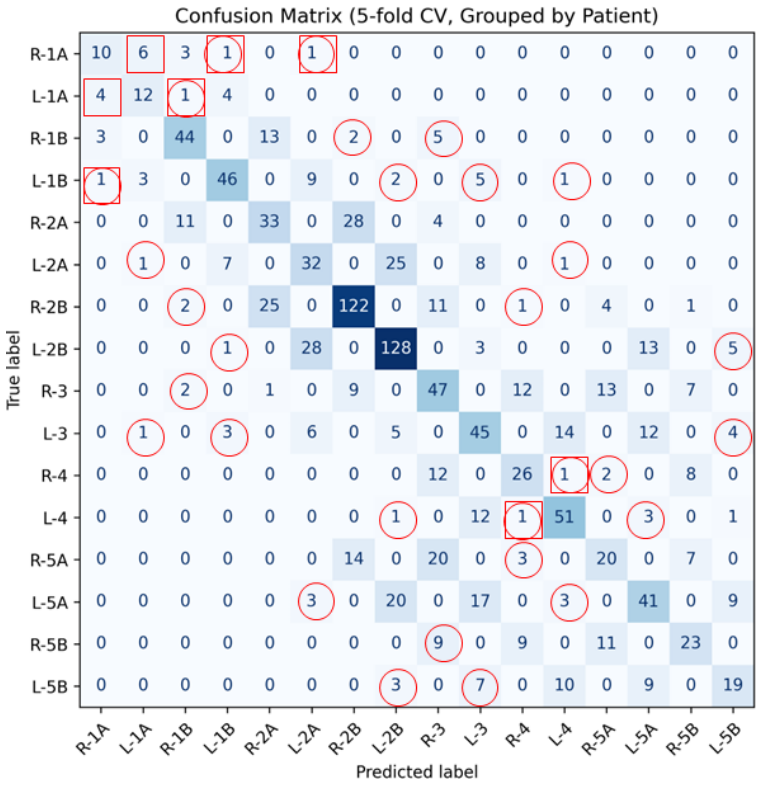

Figure 3. Confusion matrix of node level prediction. L-R errors are marked by red squares and non-neighbor errors are marked by red circles.

The confusion matrix of node level prediction is shown in Fig. 3. The correct level predictions are on the diagonal elements of the matrix. The L-R and non-neighbor errors are also marked. Unmarked non-diagonal elements with non-zero values indicate neighbor errors.

Out of all 1226 lymph nodes.

- Level prediction on 699 nodes is correct with an accuracy of 57%.
- There are only 16 (1%) L-R errors that concentrates (13 occurrences) on level 1A.

- The number of non-neighbor errors is 76 corresponding to an error rate of 6%.
- The rate of neighbor errors is 36%.

If we deem neighbor errors to be acceptable, the overall accuracy of the acceptable prediction is 93%, which is excellent considering the facts that only a small number of images are labeled, and the numbers of nodes associated with different labels are highly unbalanced.

To gain more insight into the cause of errors, the variation of image appearances was accessed as shown in Fig. 4. The very low L-R error rate indicates that the image-agnostic XGBoost model can identify the left and right sides of the neck. However, the variations in the orientation and curve of the neck are probably the main cause of the neighbor errors.

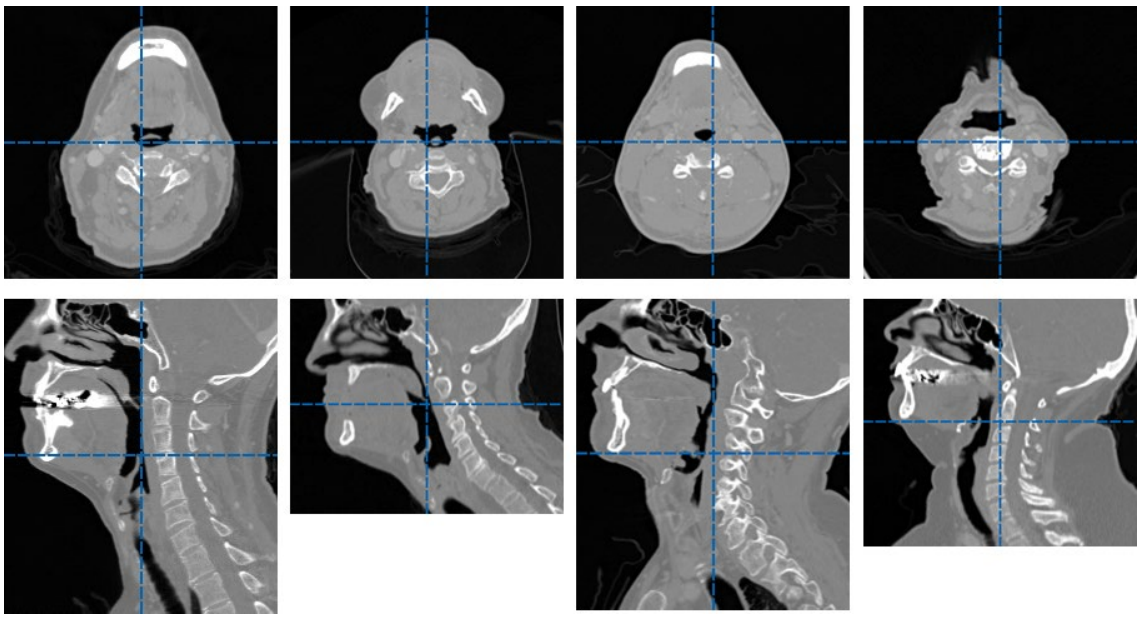

Figure 4. Examples of mid-transverse (top row) and mid-sagittal (bottom) row 2D slices.

Ideas/aims for future extramural project:

The main deterministic factor of the performance of any AI/ML method is the size of training data with high-quality labels. Acquiring 3D medical images and associated labels can be time-consuming especially for labels which requires human efforts from trained experts. In fact, most efforts of this project were spent in data labeling.

Once a base AI/ML model with reasonable performance is available, it can be utilized to reduce efforts required to label new data to increase the size of the training data for the next-iteration model. The results of this project show that: i) the base models for node detection and level classification produced expected performance, ii) baseline software tool and data processing workflow are established and ready to be used for new data.

In the ideal case, if large training data is available, the task of node detection and level classification can be achieved by a single nnUNet based model. We explored this possibility using the 16 3D images with labeled levels, but the results showed that 16 is far from enough.

Beyond enlarging training set, potential further improvements of the level classification model include adding anatomical and image awareness to the model. For example, locations of anatomical structures that are often used in practice to identify node levels can be added to the model. The data for these locations can come from manually place landmarks and/or outputs of AI model trained on images.

Due to limited available data and human effort, we did not test the feasibility of using AI/ML to produce diagnostic labels of the lymph nodes. The LogismosByTemplate software already supports adding additional labels to nodes. In another previous project, we utilized ML trained on various shape, intensity, and texture features of tumors to predict the tumor progression and produced promising results. The same approach should be applicable to lymph nodes in the future.

Publications resulting from project:

None yet

Grant proposal to be re-submitted in Feb 2026.