THE LANCET
Digital Health

Supplementary appendix

Supplement to: Kann BH, Likitlersuang J, Bontempi D, et al. Screening for extranodal extension in HPV-associated oropharyngeal carcinoma: evaluation of a CT-based deep learning algorithm in patient data from a multicentre, randomised de-escalation trial. *Lancet Digit Health* 2023; published online April 20. https://doi.org/10.1016/S2589-7500(23)00046-8.

SUPPLEMENTARY MATERIAL

Screening for extranodal extension with deep learning: evaluation in ECOG-ACRIN E3311, a randomized trial for HPV-associated oropharynx carcinoma

TABLE OF CONTENTS

S1. Supplemental Methods

E3311 Lymph Node Segmentation and Labeling

CT scans were reviewed in conjunction with the LND pathology reports and individual lymph nodes were manually segmented and labeled as negative, nodal metastasis (NM), or nodal metastasis with extranodal extension (ENE) per a standard operating procedure described in detail previously.[1] In summary, a lymph node was only classified into the above categories if it could be deduced from correlative review of the pathology report that, (1) ENE and/or NM was confirmed present, and (2) the CT-identified LN matched in location, anatomic level, and size as described in the pathology report. For this study, the lymph node with largest short-axis diameter was segmented in every study, so long as a certain pathologic correlation could be made. The node with longest SAD radiographically was annotated by comparison with the documented size of the corresponding pathologic node at the corresponding lymph node level.

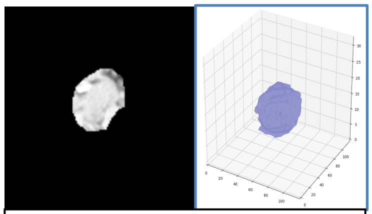
E3311 pathology reports generally delineate the nodal level and the size of the lymph node where ENE is present (i.e. "extranodal extension is present at level IIA, node measuring 2.5 cm in diameter"), which could be correlated back to the CT scan. For the purposes of annotation, the largest node was identified via measuring short-axis diameter and then was compared to the pathology report – if there was corresponding lymph node on pathology report that matched in anatomic level and size, then labels (benign vs positive and no ENE vs +ENE) was assigned with certainty. This process was then iterated for an additional 1-2 lymph nodes of varying smaller sizes that had certain pathologic correlations. In patients where no ENE was present in the entire specimen, "no ENE" could also be ascribed to the segmented nodes with certainty. For the smaller 1-2 lymph nodes, in rare cases when pathologic lymph node size was not documented, but there was only one node in a specified station with documented ENE, and that node corresponded to a radiographic node with SAD ≥10 mm, then a certain annotation could be made. If there was ambiguity in the pathology report, or stations with multiple positive nodes and no clear way to distinguish them (via size or lymph node level), then the node would not be segmented or annotated. An example of an ambiguous case would be one with two positive nodes in the same nodal level, one with ENE, one without, and the node with ENE was either a) not specified by size or b) had equivalent size as the other positive node. In addition to presence or absence of ENE, the extent of ENE was recorded (in millimeters) when available. The initial segmentation was performed by B.H.K. and then each segmentation was reviewed for accuracy by each radiologist in the study, and necessary changes were made prior to model testing. Segmentations and reader review were performed using the open-source software 3D Slicer v4.10 (Boston, MA).[2] Scans were reviewed in axial, sagittal, and coronal planes. In all, 313 E3311 lymph nodes were segmented and annotated, and all of these were included in the primary study analysis.

Image preprocessing and deep learning algorithm training pipeline protocol

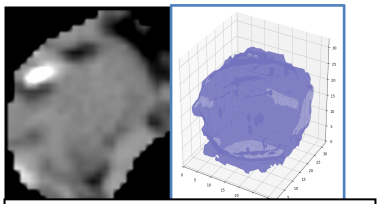
We utilized the *DualNet* architecture, training heuristics, and hyperparameters as described in detail in prior work.[1] All preprocessing and training were performed using Python v3.8.5. Neural network training was done with Tensorflow v2.4 utilizing Keras v2.2. The algorithm pipeline was the same as used in prior work[1], with the following modifications that were made *a priori* for this study: to prevent the reliance on commercial software for lymph node segmentation and reader evaluation (as done in prior work with OsirixMD [Geneva, Switzerland]), we utilized the open-source, well-established 3D Slicer software (Boston, MA; slicer.org) for lymph node segmentation, converting losslessly from DICOM to Nearly Raw Rasterized Data (nrrd) format during curation. Additionally, given heterogeneity in slice thickness within the E3311 dataset (given the numerous participation institutions), we elected to resample all lymph node data to 1 x 1 x 3 mm voxel spacing.

The steps for the protocol are summarized as follows for reference and reproducibility:

1. Raw DICOM image conversion to NRRD utilizing Pydicom package and Simple ITK
2. Lymph node segmentation in 3D Slicer saved to NRRD
3. Interpolation to 1 (x) x 1 (y) x 3 (z) mm spacing with 10 mm dilation to encompass surrounding lymph node tissue environment
4. Centering and cropping to a 118 x 118 x 32 voxel array (Box Input)
5. Hounsfield Unit centering and normalization to unit variance with soft tissue windowing
6. Derivation of second input array (Small Input) via cropping to the edge of the voxel region of interest and rescaling the Box Input to 32 x 32 x32



Box Input (example)



Small Input (example)

7. Data splitting randomly 80% / 20% for training and tuning, stratified by node category (benign, positive without ENE, or ENE)
8. Random upsampling of training and tuning data for the ENE (minority) class 2:1 to improve tuning class balance
9. Passing dual input arrays from training set into a custom data generator for augmentation
10. Training and validation data are passed into *DualNet* (**Figure S5**) for training. Training was performed on a single RTX Titan 24 GB Graphical Processing Unit (Nvidia; Santa Clara, CA).

Training Details

To harness the hypothesized benefit of increasing training data, we combined the three datasets (Yale, Sinai, TCGA-TCIA), described in detail in our prior work[1,3] **(Table S2)**. Following curation of a combined dataset, and converting segmentations and images to 3D NRRD format, there were a total of 797 valid lymph nodes with annotated ground truth available for the development set: 605 from the Yale dataset, 138 from the Mount Sinai dataset, and 54 from TCGA-TCIA. There was some attrition from the Yale and TCGA-TCIA datasets from our prior work, given inconsistent metadata that corrupted the conversion to NRRD when shifting to the 3D Slicer-pipeline with Z-axis resampling. Therefore, these nodes were excluded for practical reasons. The remaining data was split into training (80%, n=637) and internal validation (20%, n=160) sets of nodes, stratified by nodal category (benign, metastatic without ENE, or ENE). Data augmentation for the present study was performed with a real-time data generator utilizing a series of affine transformations, including random rotation, horizontal and vertical flipping. An additional augmentation feature was developed to mimic variability in human contouring of lymph nodes by introducing random erosions and dilations across the lymph node region of interest on the order of several millimeters. We hypothesized that this would increase generalizability of the algorithm at test time. Together, real-time augmentation would be expected to generate >60x individual node representations over the course of training (n >47,820). The model was specified to train for a maximum of 300 epochs with early stopping once validation loss (i.e. binary cross-entropy loss) ceased to improve for 20 epochs. Learning rate began at 0.001 and was halved if loss plateaued for 10 epochs. The network was trained on an in-house RTX TITAN GPU (Nvidia, Santa Clara, CA) using Tensorflow v2.4. Testing was performed in a Python v3.6.4 environment on both a desktop 2.6 GHz Intel Core i5 computer processing unit (CPU) (Apple; Cupertino, CA) and a Tesla V100 GPU (Nvidia; Santa Clara, CA). Upon execution of training and validation, the DLA obtained minimized loss at epoch 130 (Figure S6) and was locked at this point for further testing. AUC on the internal validation set at this point was 0.93.

Algorithm Calibration

Given the increasing recognition that neural networks outputs can be prone to miscalibration, we applied temperature scaling to the internal validation predicted probabilities.[4] Temperature scaling, an extension of Platt Scaling[5], is a widely used method whereby uncalibrated predictions are used as features for a logistic regression model, which was, importantly, trained on the ***internal validation set*** (n=160) to return the calibrated probabilities (without actually retraining any model parameters), using the following transform, which was optimized with respect to the negative log likelihood:

$$\hat{q}_i = \max_k \, \sigma_{\mathrm{SM}}(\mathbf{z}_i/T)^{(k)}.$$

This transform was then applied to the E3311 test set predicted probabilities and calibration plots were plotted for the original predictions and the calibrated predictions. Temperature scaling, by definition, does not affect model accuracy or discriminatory performance.

Expected Calibration Error (ECE) was calculated via the following equation:

$$\mathrm{ECE} = \sum_{m=1}^{M} \frac{|B_m|}{n} \left| \mathrm{acc}(B_m) - \mathrm{conf}(B_m) \right|,$$

Where *n* is the number of samples, and the difference between the *acc* and *conf* for a given bin represents the calibration gap. Temperature scaling was implemented via the Python package, NetCal (https://pypi.org/project/netcal/).

Contour Variance

The source code for the contour variance custom script can be found in the "contour_variance.py" file at https://github.com/bhkann/DualNet-ENE

Ancillary Analyses

To determine if DLA prediction on the scan's largest node could be a surrogate for patient-level ENE, we evaluated DLA prediction on the largest node to predict patient-level pathologic diagnosis of ENE (at any node) and compared these results to reader predictions of patient-level ENE following review of the entire scan. To determine if the DLA might augment radiologist performance, we conducted a simulated experiment whereby DLA predictions were used to augment uncertain reader Likert scores of 2 and 3 by substituting the DLA prediction (from the model with optimized YI threshold) in place of the reader prediction in those instances.

1. Kann BH, Aneja S, Loganadane GV, et al. Pretreatment Identification of Head and Neck Cancer Nodal Metastasis and Extranodal Extension Using Deep Learning Neural Networks. *Scientific Reports*. 2018;8(1):14036. doi:10.1038/s41598-018-32441-y

2. Fedorov A, Beichel R, Kalpathy-Cramer J, et al. 3D Slicer as an image computing platform for the Quantitative Imaging Network. *Magn Reson Imaging*. 2012;30(9):1323-1341. doi:10.1016/j.mri.2012.05.001

3. Kann BH, Hicks DF, Payabvash S, et al. Multi-Institutional Validation of Deep Learning for Pretreatment Identification of Extranodal Extension in Head and Neck Squamous Cell Carcinoma. *Journal of Clinical Oncology*. 2020;38(12):1304-1311. doi:10.1200/jco.19.02031

4. Guo C, Pleiss G, Sun Y, Weinberger KQ. On Calibration of Modern Neural Networks. *arXiv:170604599 [cs]*. Published online August 3, 2017. Accessed February 5, 2020. http://arxiv.org/abs/1706.04599

5. Platt JC. Probabilistic Outputs for Support Vector Machines and Comparisons to Regularized Likelihood Methods. In: *Advances in Large Margin Classifiers*. MIT Press; 1999:61-74.

6. Maxwell JH, Rath TJ, Byrd JK, et al. Accuracy of computed tomography to predict extracapsular spread in p16-positive squamous cell carcinoma. *Laryngoscope*. 2015;125(7):1613-1618. doi:10.1002/lary.25140

7. Chai RL, Rath TJ, Johnson JT, et al. Accuracy of computed tomography in the prediction of extracapsular spread of lymph node metastases in squamous cell carcinoma of the head and neck. *JAMA Otolaryngol Head Neck Surg*. 2013;139(11):1187-1194. doi:10.1001/jamaoto.2013.4491

S2. Model Development and Training Dataset from Combined Model Dataset: Study Patient and Lymph Node Characteristics

	Model Development Combined Dataset					
	Yale		Mount Sinai		TCIA-TCGA	
Patient Cohort	Patients (N=270)	Lymph Nodes (n=653)	Patients (N=82)	Lymph Nodes (n=130)	Patients (N=62)	Lymph Nodes (n=70)
Primary Cancer Site	N (%)	n (%)	N (%)	n (%)	N (%)	n (%)
Oropharynx	72 (26.7)	178 (27.3)	41 (50.0)	71 (54.6)	1 (1.6)	1 (1.4)
Oral Cavity	106 (39.3)	251 (38.4)	32 (39.0)	44 (33.8)	51 (82.3)	59 (84.3)
Larynx/Hypopharynx/ Nasopharynx	48 (17.8)	126 (19.3)	9 (11.0)	15 (11.6)	10 (16.1)	10 (14.3)
Salivary Gland	18 (6.7)	36 (5.5)	0 (0)	0 (0)	0 (0)	0 (0)
Unknown/Other	26 (9.6)	62 (9.5)	0 (0)	0 (0)	0 (0)	0 (0)
Pathologic T-stage						
T0	5 (1.9)	17 (2.6)	0 (0)	0 (0)	0 (0)	0 (0)
T1	36 (13.3)	91 (13.9)	25 (30.5)	38 (29.2)	5 (8.1)	5 (7.1)
T2	72 (26.7)	172 (26.3)	32 (39.0)	53 (40.8)	16 (25.8)	18 (25.7)
T3	37 (13.7)	94 (14.4)	9 (11.0)	15 (11.5)	19 (30.6)	22 (31.4)
T4	44 (16.3)	107 (16.4)	16 (19.5)	24 (18.5)	21 (33.9)	24 (34.4)
Unknown	76 (28.2)	172 (26.3)	0 (0)	0 (0)	1(1.6)	1 (1.4)
Pathologic N-stage						
N0	83 (30.7)	185 (28.3)	17 (20.7)	20 (15.4)	24 (38.7)	28 (40.0)
N1	38 (14.1)	82 (12.6)	11 (13.4)	17 (13.1)	12 (19.4)	14 (20.0)
N2	76 (28.2)	209 (32.0)	53 (64.6)	92 (72.7)	24 (38.7)	26 (37.1)
N3	9 (3.3)	33 (5.1)	1 (1.2)	1 (0.8)	0 (0)	0 (0)
Unknown	64 (23.7)	144 (22.0)	0 (0)	0 (0)	2 (3.2)	2 (2.9)
HPV/p16 Status						
Negative	188 (69.6)	454 (69.5)	44 (53.7)	67 (51.5)	6 (9.7)	7 (10.0)
Positive	76 (28.2)	185 (28.3)	38 (46.3)	63 (48.5)	0 (0)	0 (0)
Unknown	6 (2.2)	14 (2.2)	0 (0)	0 (0)	56 (90.3)	53 (90.0)
Lymph Node Pathology						
Negative		380 (58.2)	-	55 (42.3)	-	29 (41.4)
Nodal Metastasis, ENE(-)		153 (23.4)	-	54 (41.5)	-	46 (34.3)
Node Metastasis, ENE(+)		120 (18.4)	-	21 (16.2)	-	17 (24.3)

Abbreviations: ENE = extranodal extension

For the purposes of this study, the *DualNet* model was retrained after adding the Mount Sinai and TCIA-TCGA (The Cancer Genome Atlas Head-Neck Squamous Cell Carcinoma (TCGA-HNSC)) datasets to the original Yale training data. Patients with non-oropharyngeal carcinoma who did not undergo HPV or p16 testing were coded as negative, given the very low incidence of HPV/p16 positive tumors in these disease sites. Abbreviations: ENE = extranodal extension. For further detail, see “Methods” sections from 1) Kann BH, Aneja S, Loganadane GV, et al. Pretreatment Identification of Head and Neck Cancer Nodal Metastasis and Extranodal Extension Using Deep Learning Neural Networks. Scientific Reports. 2018;8(1):14036. doi:10.1038/s41598-018-32441-y ; and 2) Kann BH, Hicks DF, Payabvash S, Mahajan A, Du J, Gupta V, Park HS, Yu JB, Yarbrough WG, Burtness BA, Husain ZA, Aneja S. Multi-Institutional Validation of Deep Learning for Pretreatment Identification of Extranodal Extension in Head and Neck Squamous Cell Carcinoma. J Clin Oncol. 2020 Apr 20;38(12):1304-1311. doi: 10.1200/JCO.19.02031.

S3. CT Scans and Image Acquisition

All scans, pathology reports, and EHR data were de-identified in accordance with Health Insurance Portability and Accountability Act prior to transfer to the study investigators.

CT Images were obtained in their entirety as de-identified, decompressed Digital Imaging and Communications in Medicine (DICOM) files. E3311 Protocol stipulated that diagnostic imaging must have been performed <30 days prior to trial registration, and surgery was to be performed <4 weeks after registration. Patients were specifically excluded if there was evidence of extensive or "matted/fixed" pathologic adenopathy on preoperative imaging. Of 251 study scans initially obtained, 187 were uncorrupted CT scans with contrast enhancement with linked pathology reports, and for 178 a certain correlation was able to be made in regards to ENE. Of 178 scans included in the study, 46 institutions were represented, with a median of 3 patients per institution (range: 1 – 17). Scans were performed on 22 different CT models.

S3A. E3311 CT Scanner Characteristics

CT Scanner Manufacturers and Models Used for Study Patients

Manufacturer	Model	Patients (n=178) (n)	 (%)
GE Medical Systems	LightSpeed	45	25.3%
	Discovery	31	17.4%
	BrightSpeed	2	1.1%
	Optima	7	3.9%
	Revolution	8	4.5%
	SafeCT	1	0.6%
	HiSpeed	1	0.6%
	Other	1	0.6%
Siemens	SOMATOM	31	17.4%
	Biograph	2	1.1%
	Perspective	2	1.1%
	Sensation	11	6.2%
	Other	6	3.4%
Philips	Brilliance	12	6.7%
	Gemini	3	1.7%
	iCT	1	0.6%
	Ingenuity	2	1.1%
	Intellispace	1	0.6%
NMS	NeuViz	1	0.6%
Toshiba	Aquilion	8	4.5%
Velocity Medical Solutions	VelocityAI	1	0.6%
Other		1	0.6%
	Total	178	100%

CT Scanner Specifications and Deviation Tables

a) Scan Characteristic Deviation Table

Scan Characteristic	Mean	Median	Mode	Range (min – max)	Standard Deviation
Pixel Size (cm)	0.52	0.49	0.49	(0.35 –1.15)	0.13
Slice Thickness (cm)	2.57	2.5	2.5	(.625 – 5.0)	0.65
Tube Voltage (kVp)	117.6	120.0	120.0	(80.0 – 140.0)	10.42

b) Scan Characteristic Distribution

Slice Thickness (cm)	n = 82	(%)
5.0	2	1.1%
4.0	3	1.7%
3.75	3	1.7%
3.0	58	32.6%
2.5	81	45.5%
2.0	16	9.0%
1.25	8	4.5%
1.0	2	1.1%
0.75	2	1.1%
0.625	3	1.7%
Axial Spatial Resolution (pixels)		
512 x 512	178	100%
IV Contrast Bolus		
Optiray 300	5	2.8%
Omnipaque	78	43.8%
Isovue 370	16	9.0%
Visipaque	1	0.56%
Unknown	78	43.8%

S4A-B. Lymph Node Characteristics for E3311 Dataset

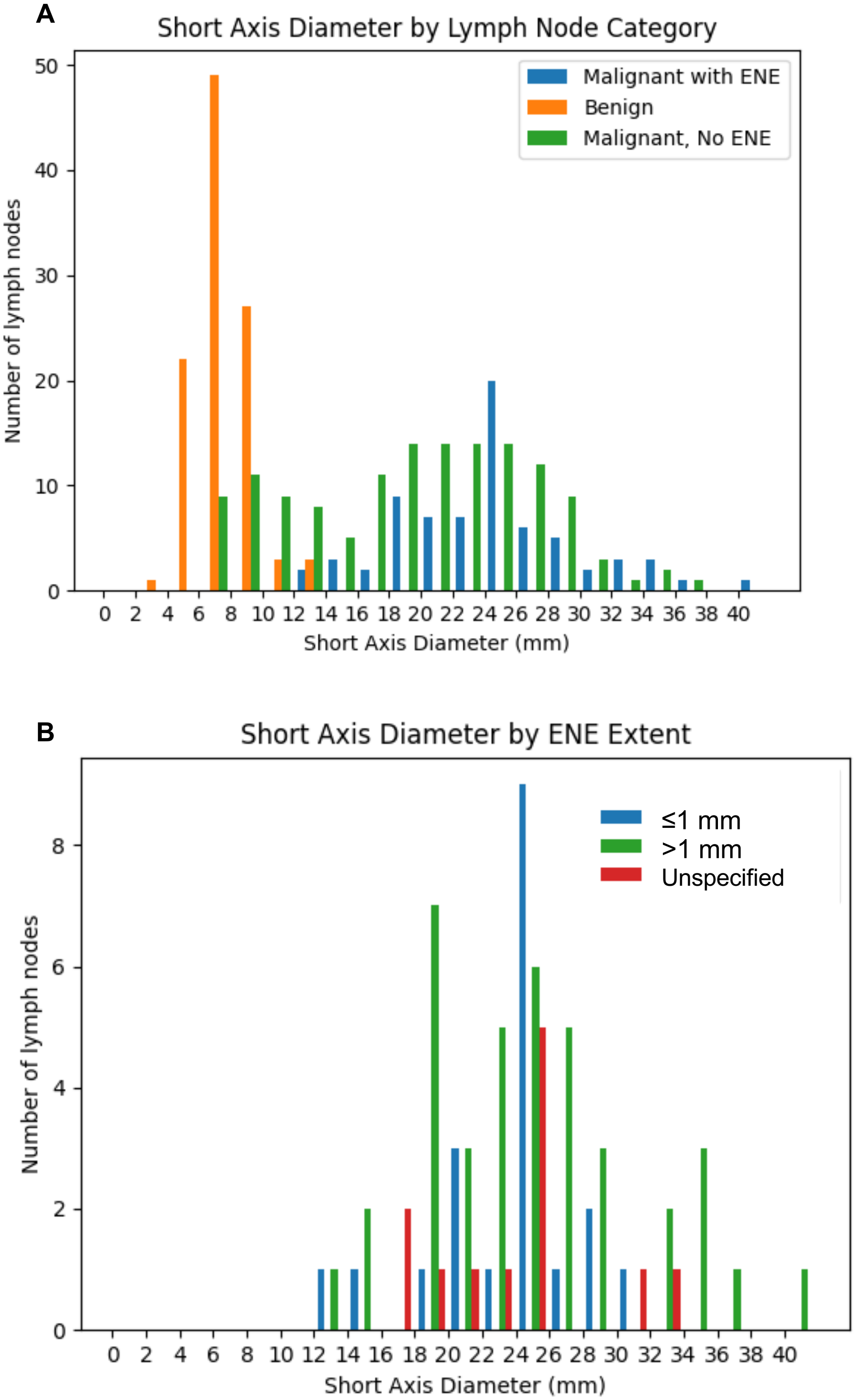
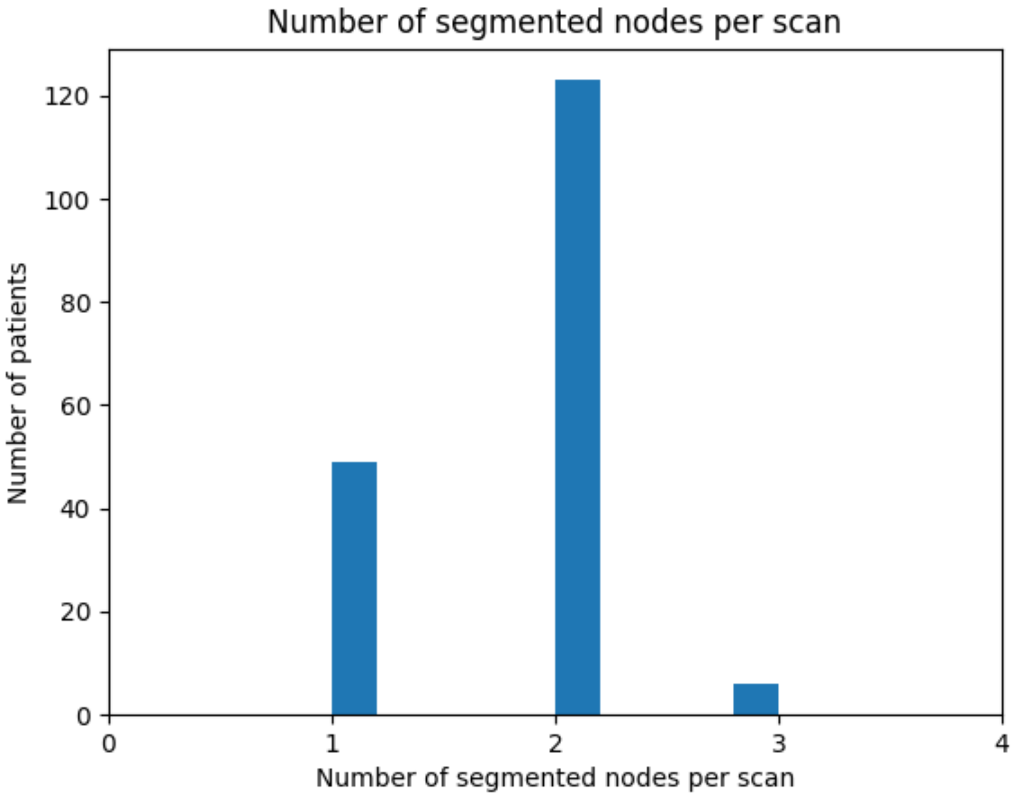


Figure S4A-B. Short Axis Diameter (SAD) by ENE status among all nodes (n=313). (A) Median SAD was 7 mm (4-14 mm) for benign, 20 mm (6-37 mm) for malignant without ENE, and 24 mm (12-42 mm) for ENE (p<0.001). (B) Of ENE nodes (n=71), SAD for ENE ≤1 mm, >1 mm, and unspecified ENE was each 25, 25, and 24 mm, respectively (p=.60).

S4C. Number of nodes segmented and annotated per E3311 scan (n=178 scans, 313 lymph nodes)



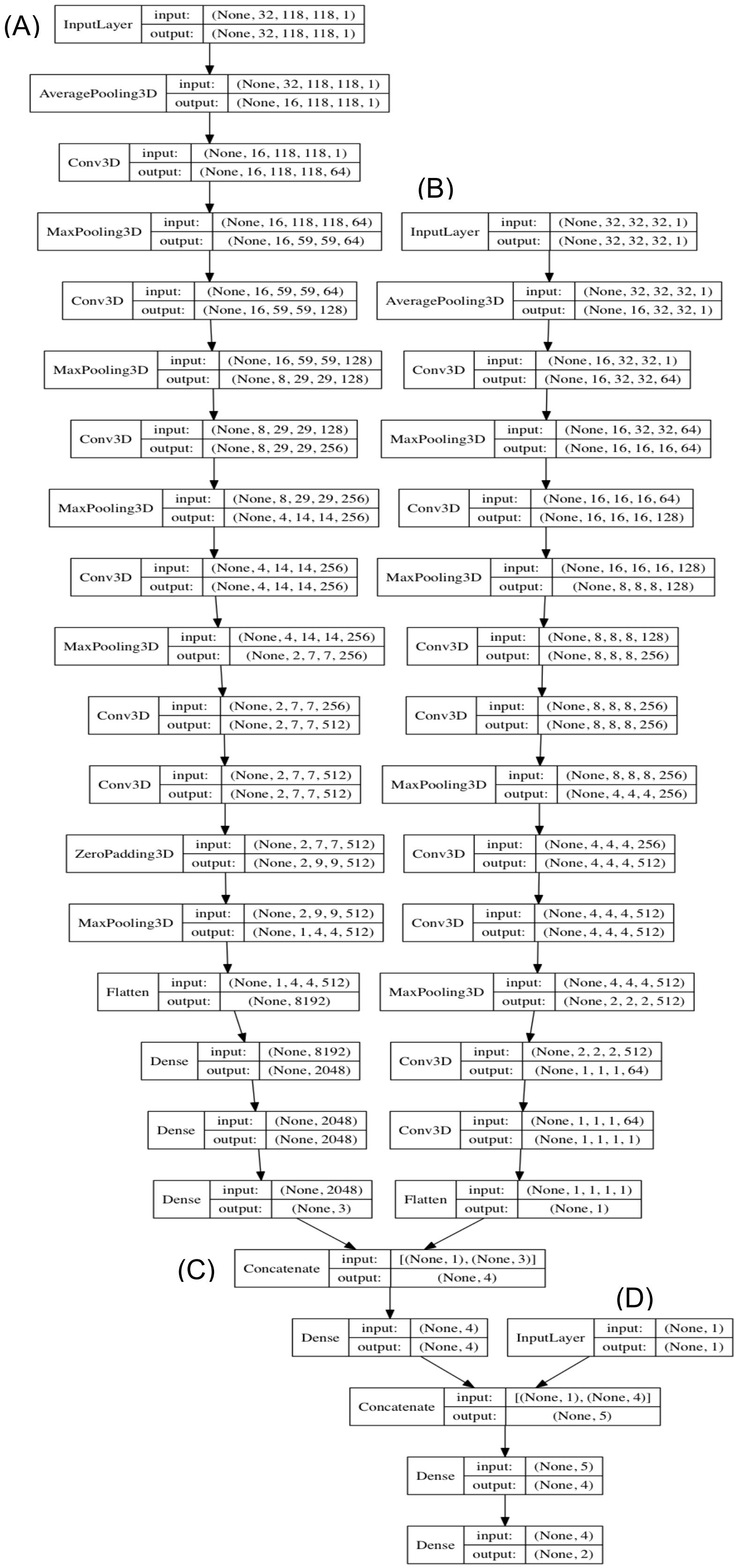
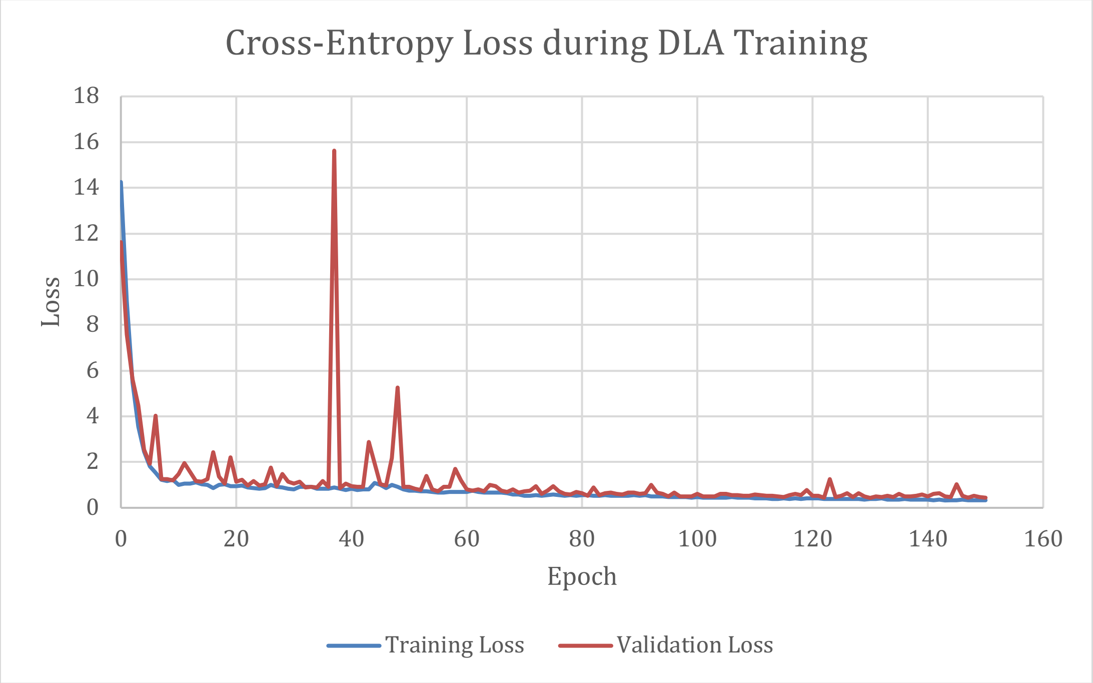

(A)
InputLayer input: (None, 32, 118, 118, 1) output: (None, 32, 118, 118, 1)
AveragePooling3D input: (None, 32, 118, 118, 1) output: (None, 16, 118, 118, 1)
Conv3D input: (None, 16, 118, 118, 1) output: (None, 16, 118, 118, 64)
MaxPooling3D input: (None, 16, 118, 118, 64) output: (None, 16, 59, 59, 64)
Conv3D input: (None, 16, 59, 59, 64) output: (None, 16, 59, 59, 128)
MaxPooling3D input: (None, 16, 59, 59, 128) output: (None, 8, 29, 29, 128)
Conv3D input: (None, 8, 29, 29, 128) output: (None, 8, 29, 29, 256)
MaxPooling3D input: (None, 8, 29, 29, 256) output: (None, 4, 14, 14, 256)
Conv3D input: (None, 4, 14, 14, 256) output: (None, 4, 14, 14, 256)
MaxPooling3D input: (None, 4, 14, 14, 256) output: (None, 2, 7, 7, 256)
Conv3D input: (None, 2, 7, 7, 256) output: (None, 2, 7, 7, 512)
Conv3D input: (None, 2, 7, 7, 512) output: (None, 2, 7, 7, 512)
ZeroPadding3D input: (None, 2, 7, 7, 512) output: (None, 2, 9, 9, 512)
MaxPooling3D input: (None, 2, 9, 9, 512) output: (None, 1, 4, 4, 512)
Flatten input: (None, 1, 4, 4, 512) output: (None, 8192)
Dense input: (None, 8192) output: (None, 2048)
Dense input: (None, 2048) output: (None, 2048)
Dense input: (None, 2048) output: (None, 3)
(B)
InputLayer input: (None, 32, 32, 32, 1) output: (None, 32, 32, 32, 1)
AveragePooling3D input: (None, 32, 32, 32, 1) output: (None, 16, 32, 32, 1)
Conv3D input: (None, 16, 32, 32, 1) output: (None, 16, 32, 32, 64)
MaxPooling3D input: (None, 16, 32, 32, 64) output: (None, 16, 16, 16, 64)
Conv3D input: (None, 16, 16, 16, 64) output: (None, 16, 16, 16, 128)
MaxPooling3D input: (None, 16, 16, 16, 128) output: (None, 8, 8, 8, 128)
Conv3D input: (None, 8, 8, 8, 128) output: (None, 8, 8, 8, 256)
Conv3D input: (None, 8, 8, 8, 256) output: (None, 8, 8, 8, 256)
MaxPooling3D input: (None, 8, 8, 8, 256) output: (None, 4, 4, 4, 256)
Conv3D input: (None, 4, 4, 4, 256) output: (None, 4, 4, 4, 512)
Conv3D input: (None, 4, 4, 4, 512) output: (None, 4, 4, 4, 512)
MaxPooling3D input: (None, 4, 4, 4, 512) output: (None, 2, 2, 2, 512)
Conv3D input: (None, 2, 2, 2, 512) output: (None, 1, 1, 1, 64)
Conv3D input: (None, 1, 1, 1, 64) output: (None, 1, 1, 1, 1)
Flatten input: (None, 1, 1, 1, 1) output: (None, 1)
(C)
Concatenate input: [(None, 1), (None, 3)] output: (None, 4)
Dense input: (None, 4) output: (None, 4)
(D)
InputLayer input: (None, 1) output: (None, 1)
Concatenate input: [(None, 1), (None, 4)] output: (None, 5)
Dense input: (None, 5) output: (None, 4)
Dense input: (None, 4) output: (None, 2)

Figure S5. Deep Learning Algorithm (*DualNet*) Architecture Schematic Dimension-preserving input "BoxNet" DLA (A) is merged with size-invariant "SmallNet" DLA (B) to form the *DualNet* input (C) with model output. The model has the capability to merge HPV/p16-status (D), though this was found in prior work to not improve performance.

S6. DLA Retraining on Combined Yale-Sinai-TCIA dataset (n=797)



S7. Sample Size Calculation for AUC Comparison

We calculated the necessary sample size of labeled lymph nodes for each dataset to evaluate the study's primary endpoint, area under curve (AUC) of the receiver operating characteristic (ROC) curve, which plots sensitivity versus the false-positive-rate (FPR). We calculated sample size to detect a Type I error of 5% with 80% Power. We assumed a null hypothesis of AUC 0.70 (based on existing literature and prior work),[1,6,7] an alternative hypothesis of AUC 0.85 (based on our prior external validation results),[3] and a ratio of negative to ENE lymph nodes of 4:1 (based on our prior work), yielding a sample of at least 155 lymph nodes needed. Power calculation was performed using MedCalc® v19 (MedCalc Software, Belgium).

1. Kann BH, Aneja S, Loganadane GV, et al. Pretreatment Identification of Head and Neck Cancer Nodal Metastasis and Extranodal Extension Using Deep Learning Neural Networks. *Scientific Reports*. 2018;8(1):14036. doi:10.1038/s41598-018-32441-y

2. Fedorov A, Beichel R, Kalpathy-Cramer J, et al. 3D Slicer as an image computing platform for the Quantitative Imaging Network. *Magn Reson Imaging*. 2012;30(9):1323-1341. doi:10.1016/j.mri.2012.05.001

3. Kann BH, Hicks DF, Payabvash S, et al. Multi-Institutional Validation of Deep Learning for Pretreatment Identification of Extranodal Extension in Head and Neck Squamous Cell Carcinoma. *Journal of Clinical Oncology*. 2020;38(12):1304-1311. doi:10.1200/jco.19.02031

4. Guo C, Pleiss G, Sun Y, Weinberger KQ. On Calibration of Modern Neural Networks. *arXiv:170604599 [cs]*. Published online August 3, 2017. Accessed February 5, 2020. http://arxiv.org/abs/1706.04599

5. Platt JC. Probabilistic Outputs for Support Vector Machines and Comparisons to Regularized Likelihood Methods. In: *Advances in Large Margin Classifiers*. MIT Press; 1999:61-74.

6. Maxwell JH, Rath TJ, Byrd JK, et al. Accuracy of computed tomography to predict extracapsular spread in p16-positive squamous cell carcinoma. *Laryngoscope*. 2015;125(7):1613-1618. doi:10.1002/lary.25140

7. Chai RL, Rath TJ, Johnson JT, et al. Accuracy of computed tomography in the prediction of extracapsular spread of lymph node metastases in squamous cell carcinoma of the head and neck. *JAMA Otolaryngol Head Neck Surg*. 2013;139(11):1187-1194. doi:10.1001/jamaoto.2013.4491

S8. F1 Score, Positive Predictive Value, Negative Predictive Value for DLA and Radiologists

		ENE Overall			ENE >1 mm		
	Threshold	F1 Score	PPV	NPV	F1 Score	PPV	NPV
DLA	Best YI	.62	.48	.96	.53	.37	.98
	FPR<=30%	.62	.47	.96	.50	.35	.96
	FPR<=20%	.60	.51	.91	.51	.40	.93
	FPR<=10%	.53	.58	.86	.49	.49	.90
R1		.48	.50	.85	.44	.39	.90
R2		.53	.45	.88	.48	.36	.93
R3		.49	.33	.97	.40	.25	1.0
R4		.43	.41	.83	.37	.30	.88

S9.

a) ENE identification performance among lymph nodes with short-axis diameter ≥**10 mm (n=204 nodes)**

	AUC (95% CI)	Sensitivity	Specificity	Accuracy
DLA	.74 (.67 - .81)	.63	.72	.66
R1	.61 (.54 - .68)	.46	.76	.66
R2	.62 (.55 - .69)	.63	.61	.62
R3	.55 (.51 - .59)	.96	.14	.43
R4	.55 (.48 - .62)	.45	.65	.58

Sensitivity, Specificity, and Accuracy scores for DLA reflect use of threshold with the highest Youden Index.

b) ENE identification performance among lymph nodes with **short-axis diameter ≥20 mm (n=131 nodes) and ≥30 mm (n=23 nodes)**

	AUC (95% CI)	
	SAD ≥20 mm	SAD ≥ 30 mm
DLA	.65 (.55 - .74)	.78 (.58 - .97)
R1	.58 (.49 - .66)	.61 (.40 - .82)
R2	.59 (.50 - .67)	.57 (.39 - .75)
R3	.50 (.48 - .53)	.50 [na – na]*
R4	.50 (.42 - .59)	.54 (.34 - .73)

*For R3, all nodes were predicted ENE in this subgroup

S10. Partial AUCs for ENE Prediction

Max False Positive Rate	Partial AUC
.50	.81
.40	.78
.30	.74
.20	.69
.10	.62

S11. Simulated patient-level ENE prediction based on analysis of only the largest positive node (n=178 nodes and patients). Scan-level ENE performance for radiologists (R1-4) was based on the overall impression of likelihood of ENE at the scan-level.

	AUC (95% CI)
DLA	.68 (.60 - .76)
R1	.59 (.51 - .66)
R2	.62 (.55 - .69)
R3	.54 (.50 - .58)
R4	.54 (.46 - .61)

S12. ENE predictive performance on the largest node for each scan (n=178 nodes).

	AUC (95% CI)
DLA	.70 (.62 - .78)
R1	.57 (.50 - .64)
R2	.63 (.55 - .70)
R3	.54 (.50 - .58)
R4	.55 (.47 - .62)

S13. ENE predictive performance for ENE < 1 mm (excluding nodes with unspecified or ≥ 1 mm ENE) (n=262 nodes).

	AUC (95% CI)
DLA	.81 (.73 - .88)
R1	.61 (.50 - .72)
R2	.61 (.50 - .72)
R3	.64 (.56 - .73)
R4	.60 (.49 - .72)

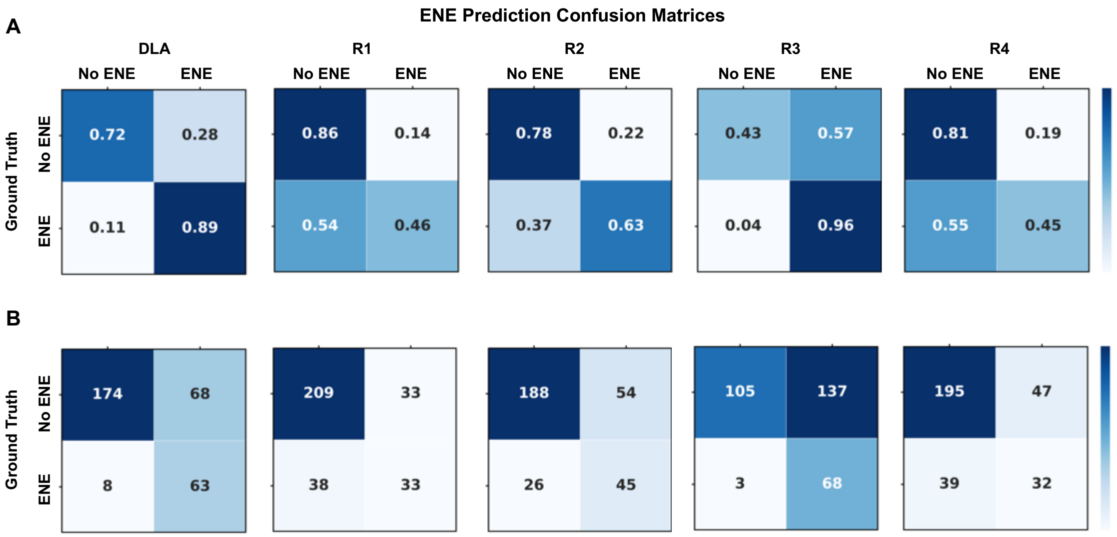


Figure S14. Confusion matrices demonstrating specificity (top left), false positive rate (top right), false negative rate (bottom left), and sensitivity (bottom right) results for ENE prediction for the DLA and R1-4, with percentages (A) and raw results (B).

S15. Radiologist performance for ENE prediction before and after DLA-assistance.

ENE Overall						
	AUC		Sensitivity		Specificity	
	Original	With DLA Assistance	Original	With DLA Assistance	Original	With DLA Assistance
R1	.66	.81	.46	.90 (+.47)	.86	.72 (-.14)
R2	.71	.81	.63	.90 (+.27)	.78	.71 (-.07)
R3	.70	.80	.96	.90 (-.06)	.43	.71 (+.28)
R4	.63	.80	.45	.89 (+.44)	.81	.72 (-.09)

In this simulated experiment, DLA predictions were use in place of radiologist predictions only when nodes were assigned an uncertain Likert score of 2 or 3. Generally, sensitivity was increased (+47%, +27%, -6%, +44% for R1-4, respectively), at the expense of specificity (-14%, -7%, +28%, -9% for R1-4, respectively), with the exception being R3. Kappa inter-rater agreement improved from 0.32 to 0.97.

S16. Sensitivity analysis for >1 mm ENE excluding nodes with uncertain extent of ENE (n=301 nodes).

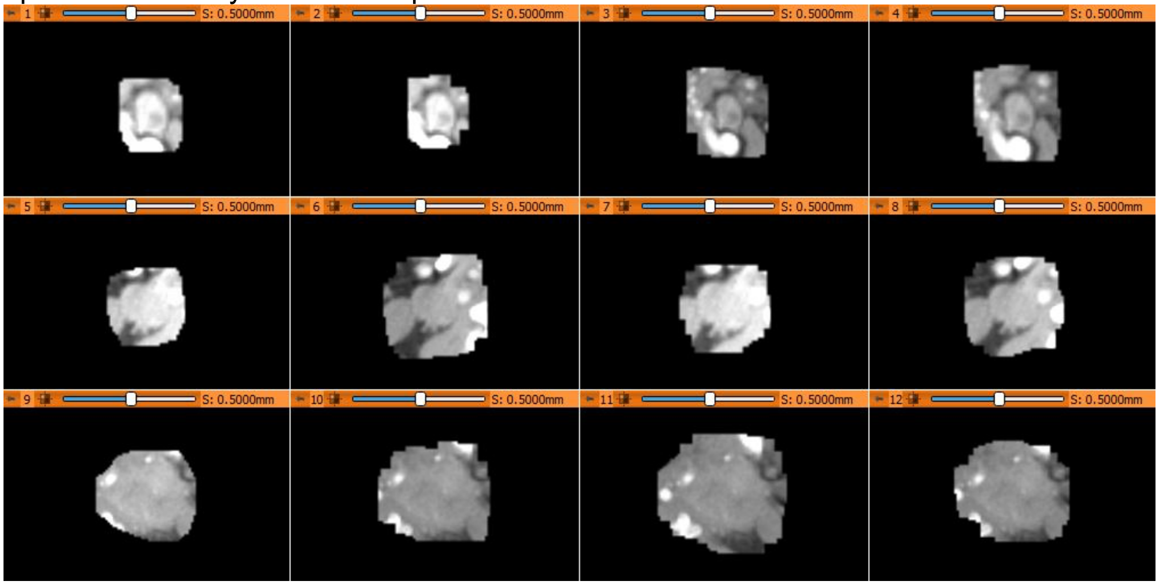
	AUC
DLA	.85 [.80 - .90]

S17. Robustness and Adversarial Experiments

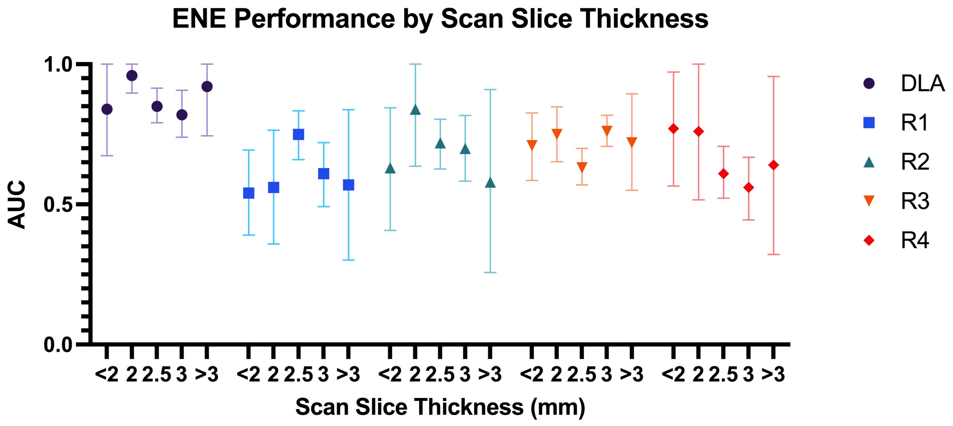
S17A. Robustness Studies. Ten tests were performed iteratively for each adversarial input experiment, contour variation (1) and Gaussian noise input (2). Contours were varied randomly throughout the periphery from 1-10 mm. Guassian noise in a range of +/- 5 HU max was added to each voxel, which had been previously shown to degrade performance in medical imaging deep learning applications.

Test No.	ENE AUC Contour Variance	ENE AUC Gaussian Noise (+/- 5 HU)
1	0.861	0.854
2	0.857	0.852
3	0.856	0.851
4	0.859	0.854
5	0.867	0.852
6	0.864	0.853
7	0.865	0.853
8	0.858	0.853
9	0.860	0.852
10	0.856	0.853
Mean	0.860 +/- .004	0.853
Original Test AUC: 0.857		

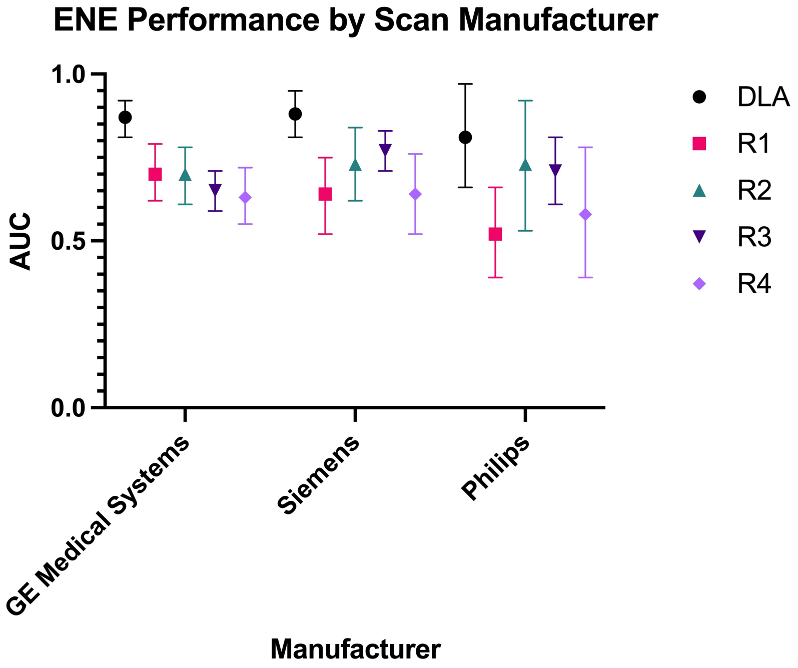
S17B. Representative examples of contour variation. Left column represents the original lymph node region of interest (segmented lymph node plus a uniform dilation of 10 mm). Columns 2-4 represent random contour variance to mimic a real-world scenario of inter-operator variability in contour input.



S17C. DLA and Radiologist Performance with Scan Slice Thickness Variation



S17D. DLA and Radiologist Performance by Scanner Manufacturer



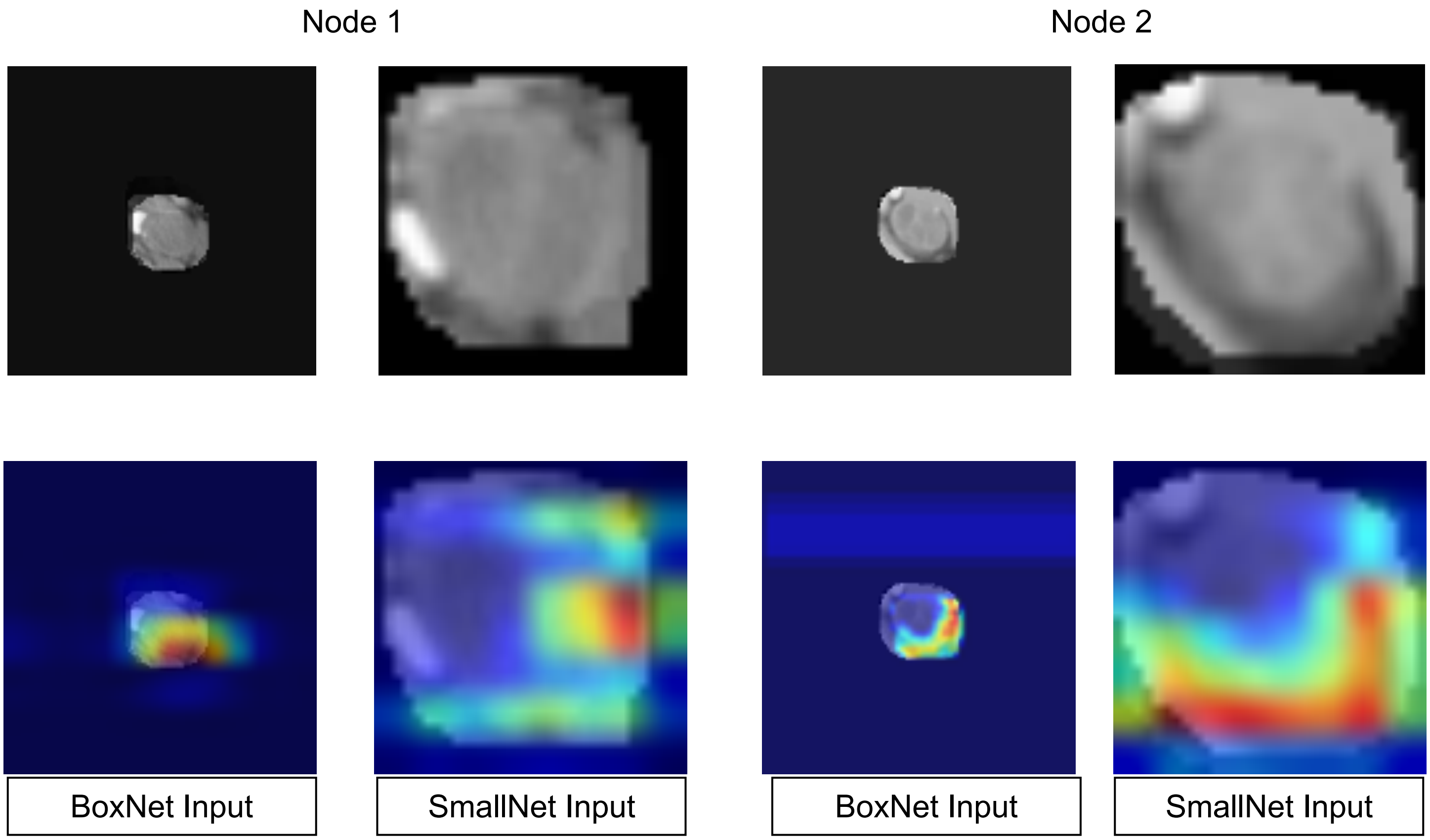
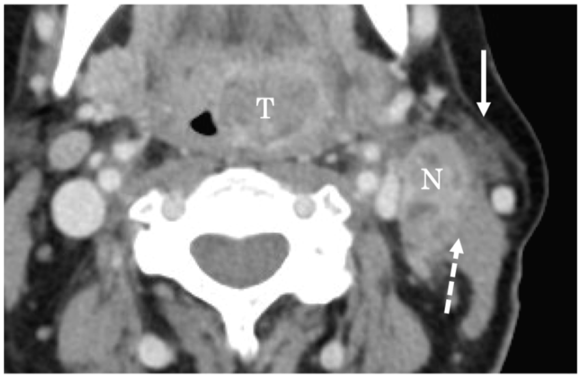
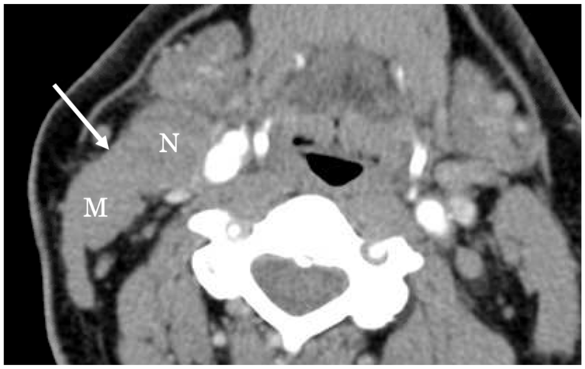
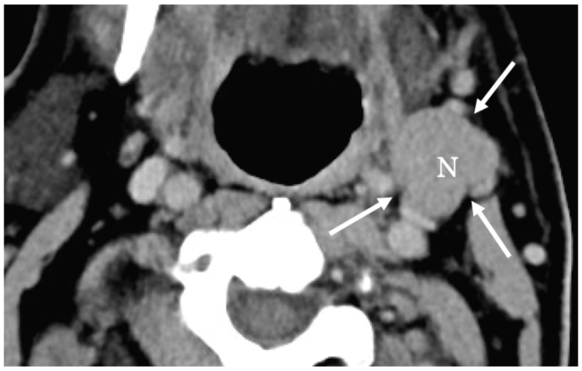
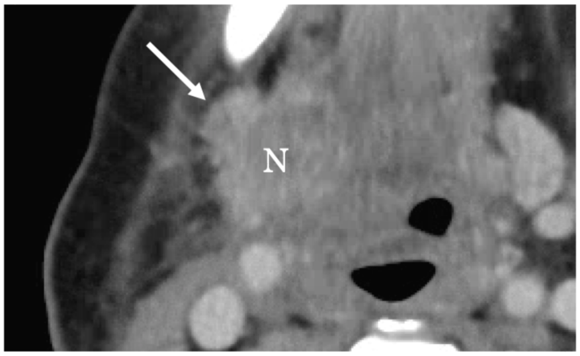
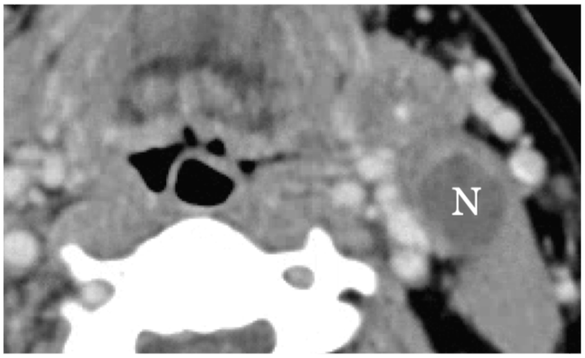
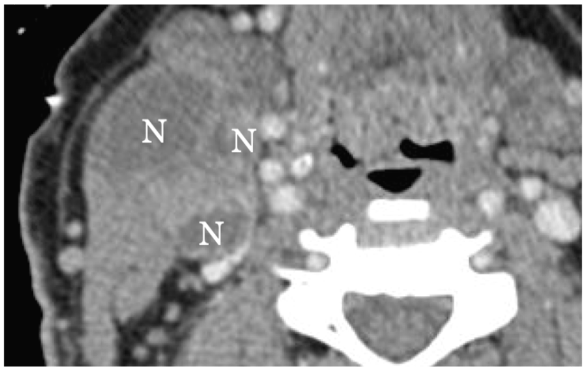
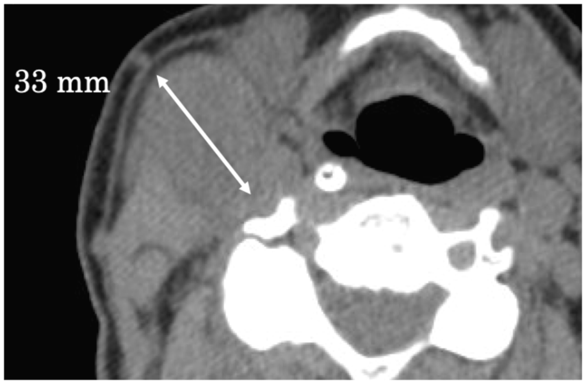


Figure S18. Gradient-weighted Class Activation Maps (GradCAMs) for two representative lymph nodes.GradCAMs generate a heat map, whereby the "hottest" pixels (red: hot; blue: cold) represent regions that are most influential in determining a particular image class - in this case - extranodal extension. These representative cases demonstrate importance of the nodal periphery in DLA ENE classification.

S18. Educational Tool for Radiologists with Visual and Descriptive CT Features Associated with ENE

CT Feature and Description	Example
1. *Perinodal fat stranding* Metastatic lymph node (N) with infiltration into adjacent fat anteriorly (solid arrow) and muscle posteriorly (dashed arrow). Note multiple regions of hypoattenuating central necrosis. A primary tumor (T) at the left base of tongue is also noted.	
2. *Absent perinodal fat planes* Lymph node (N) with absent perinodal fat plane (arrow) to the adjacent muscle (M).	
3. *Lobular contours* Lymph node (N) with lobular contours delineated by indentations at the margins (arrows).	

4. *Irregular nodal margins* Lymph node (N) with margins that are irregular, spiculated and indistinct (arrow).	
5. *Central necrosis* Lymph node (N) with significant intranodal central low attenuation or cystic appearance.	
6. *Matted/ conglomerate nodes* Nodal matting, defined as conglomerate of 3 or more lymph nodes (N) with an absence of internodal fat planes. Note a few also demonstrate central necrosis.	
7. *Size > 30 mm* Measured in greatest axial dimension, from outer margin to outer margin (double arrows).	

S19. Post-Study Radiologist Survey

1. **How many years have you been in practice as a diagnostic radiologist? (including fellowship, but excluding residency):**

   R1: 6
   R2: 4
   R3: 17
   R4: 11

2. **How challenging do you find the task of ENE identification for head and neck cancer patients on diagnostic CT scan overall (i.e. on a routine clinical basis)?** (1- "not challenging at all" to 5- "extremely challenging"):

   R1: 3
   R2: 4
   R3: 3
   R4: 4

3. **How challenging did you find the task of ENE identification on the study patients?** (1- "not challenging at all" to 5- "extremely challenging"):

   R1: 3
   R2: 4
   R3: 3
   R4: 4

4. **Compared to your head and neck radiology peers, do you think you tend to be conservative (*under-calling*) or aggressive (*over-calling*) in your identification of ENE on CT?** (1-"conservative/under-calling" to 5-"aggressive/over-calling"; 3=neutral):

   R1: 1
   R2: 2
   R3: 4
   R4: 2

5. **Do you think the educational tool provided for the study helped improve your ability to predict ENE for the study (i.e. improved your accuracy)?** (1- "not at all" to 5- "extremely helpful"):
   R1: 4
   R2: 2
   R3: 3
   R4: 1

S20. E3311 Trial Arm Assignment for Study Cohort

E3311 Arm	Study Cohort N = 178 patients	E3311 Total Population N = 359
A (low-risk, T1-2N0-1)	2 (1%)	38 (11%)
B (intermediate-risk; de-escalated radiotherapy)	56 (31%)	100 (28%)
C (intermediate-risk; standard radiotherapy)	63 (35%)	108 (30%)
D (high-risk, ≥1 mm ENE, positive margin, or >4 metastatic lymph nodes)	57 (32%)	113 (31%)