Radiology

ORIGINAL RESEARCH • HEAD AND NECK IMAGING

Pretreatment CT Identification of Extranodal Extension in Laryngeal and Hypopharyngeal Cancers Using Deep Learning

Na Shen, MD[*1,2] • Yirui Wang, MS*[*3] • Cheng Yan, BS[4] • Jian Wang, MD[4] • Dandan Zheng, MD[5] • Xuewei Wang, MD[1] • Dazhou Guo, PhD[3] • Haoshen Li, MS[3] • Qinji Yu, MS[3] • Zi Li, MS[3,6] • Yuzhen Chen, BS[7] • Ke Yan, PhD[3,6] • Le Lu, PhD[3] • Xianghua Ye, MD[5] • Mengsu Zeng, MD[4] • Xinsheng Huang, MD[1] • Tsung-Ying Ho, MD[8] • Fang Zhang, MD**[9] • Dakai Jin, PhD**[3]*

* N.S. and Y.W. contributed equally to this work.

** F.Z. and D.J. are co–senior authors.

Author affiliations, funding, and conflicts of interest are listed at the end of this article.

Radiology 2026; 318(1):e250332 • https://doi.org/10.1148/radiol.250332 • Content codes: HN CT AI

Background: Accurate preoperative identification of pathologic extranodal extension (ENE) at CT is essential for precise treatment decisions in laryngeal and hypopharyngeal squamous cell cancer (LHSCC). However, human interpretation of ENE is neither reliable nor reproducible.

Purpose: To develop and evaluate the diagnostic performance of a new deep learning tool, DeepENE, in detecting metastatic and ENE lymph nodes on preoperative CT scans in patients with LHSCC in a multicenter cohort.

Materials and Methods: In this retrospective study, patients with LHSCC from Zhongshan Hospital, Fudan University (April 2011–August 2022), were included in training, validation, and internal test sets to develop DeepENE. For the reference standard, lymph nodes were segmented on CT scans and labeled for metastasis and ENE status based on pathologic findings. DeepENE was tested using three external cohorts of patients with LHSCC (external test sets 1–3) and one external cohort of patients with oral squamous cell carcinoma. The primary diagnostic metric was the area under the receiver operating characteristic curve (AUC). The performance of DeepENE was compared with that of five board-certified head and neck cancer specialists using the DeLong method.

Results: Overall, 289 patients with LHSCC with 1954 pathologically confirmed lymph nodes were evaluated. DeepENE achieved an AUC of 0.93 for ENE diagnosis in the internal test set under fivefold cross-validation, and AUCs of 0.96, 0.87, and 0.90 in external test sets 1, 2, and 3, respectively. DeepENE outperformed the five experts, especially in early-stage ENE detection in external test set 2 (AUC of 0.87 for DeepENE vs mean AUC of 0.66 for readers; $P < .001$). In external test set 1, DeepENE maintained a high sensitivity of 97% at specificity of 90%, compared with experts' mean sensitivity of 77% ($P = .003$). In external test sets 2 and 3, DeepENE had sensitivity of 78% and 80%, compared with experts' mean sensitivity of 36% ($P < .001$) and 46% ($P < .001$), respectively.

Conclusion: DeepENE accurately detected ENE on preoperative CT scans in patients with LHSCC and outperformed head and neck cancer specialists.

Supplemental material is available for this article.

Laryngeal and hypopharyngeal squamous cell carcinoma (LHSCC) are prevalent cancers that occur spatially close to each other (1–4). Cancer in the hypopharynx, located on the medial wall of the pyriform fossa, often spreads to the ipsilateral laryngeal structures. Additionally, laryngeal cancer can extend into the paraglottic region of the hypopharynx. According to cancer statistics, there were over 250 000 new LHSCC cases and over 130 000 LHSCC cancer deaths worldwide in 2022 (5,6). Nodal staging is crucial for the management and prognosis of LHSCC. Studies have shown that positive lymph nodes are associated with an increased risk of death in LHSCC, with the presence of up to five metastatic nodes associated with a 19% increase in mortality risk (7).

Additionally, lymph node extranodal extension (ENE) (ie, tumor cells invading beyond the lymph node capsule into surrounding tissue) is one of the most negative prognostic factors (8). For example, 5-year overall survival for patients with laryngeal cancer is 32.9% with ENE compared with 56.7% without ENE (9), and 3-year overall survival for patients with hypopharyngeal cancer is 20% lower in those with ENE than in those without ENE (10). Accurate preoperative diagnosis of metastatic and ENE lymph nodes is key to precise treatment decisions. Identifying ENE in patients with locally advanced LHSCC would also facilitate and optimize clinical trials, allowing delivery of more advanced treatment (11,12).

In current clinical practice, ENE can be definitively diagnosed only through postoperative pathologic examination because imaging findings can be subtle and inconsistent at preoperative diagnostic imaging. With contrast-enhanced CT, physician readers exhibit poor diagnostic performance for identifying ENE, with high interobserver variability (13). Area under the receiver operating characteristic curve (AUC) values for physician readers are reported to be below 0.70, with sensitivities and specificities varying from less than 45% to 90% (13,14). These wide-ranging results suggest that human interpretation of "ill-defined nodal borders" (the imaging feature of ENE) (15,16) is neither reliable nor reproducible for accurate ENE detection.

Deep learning has demonstrated promising results in medical imaging. Recent studies have shown that well-developed deep learning models can match or surpass experts in various tasks

Abbreviations

AUC = area under the receiver operating characteristic curve, ENE = extranodal extension, LHSCC = laryngeal and hypopharyngeal squamous cell cancer

Summary

A deep learning diagnostic tool, DeepENE, accurately detected extranodal extension on preoperative CT scans in patients with laryngeal and hypopharyngeal squamous cell cancer, outperforming head and neck cancer specialists.

Key Results

- In this multicenter retrospective study of 289 patients, a deep learning model, DeepENE, identified extranodal extension (ENE) on preoperative CT scans in laryngeal and hypopharyngeal squamous cell cancer and achieved area under the receiver operating characteristic curve (AUC) values of 0.96, 0.87, and 0.90 in three external test sets (total $n = 117$).
- DeepENE outperformed five head and neck specialists in predicting ENE, especially in early-stage ENE detection (AUC of 0.87 for DeepENE vs mean AUC of 0.66 for physician readers; $P < .001$).

related to cancer screening, diagnosis, and treatment planning (17–20). For ENE detection, Kann et al (13,14) showed that using dual deep networks could increase the preoperative diagnostic AUC for positive cervical lymph nodes to 84%–90% at CT imaging. However, these studies focused mainly on patients with oral cancer or human papilloma virus–associated oropharynx cancer, for which the lymph node metastasis region (neck stations I–II) differs from that of LHSCC (neck stations II–IV). Moreover, the architectures of these dual deep networks were not jointly optimized to capture comprehensive ENE imaging features, and their three-dimensional convolutional kernel is not suitable for common diagnostic CT scans with 3–5-mm section thickness.

The aim of this retrospective study was to develop a new deep learning tool, DeepENE, for detecting metastatic lymph nodes with and without ENE on preoperative CT scans in patients with LHSCC, and to evaluate its diagnostic performance in a multicenter cohort.

Materials and Methods

Patients

This retrospective multicenter study was reviewed and approved by the institutional review board of Zhongshan Hospital, Fudan University (B2022-170R). The institutional review board waived the requirement for informed consent from patients due to the retrospective nature of the study and because all procedures performed were part of standard care. This article follows TRIPOD (Transparent Reporting of a Multivariable Prediction Model for Individual Prognosis or Diagnosis) Statement guidelines (21).

Five retrospective cohorts were included in this study. The internal dataset consisted of patients with pathologically confirmed LHSCC treated between April 2011 and August 2022 at Zhongshan Hospital, Fudan University. This dataset comprised the training, validation, and internal test sets that were used to develop lymph node metastasis and ENE prediction models. Patients were included if they underwent tumor and neck lymph node dissection, underwent preoperative contrast-enhanced CT (Appendix S4) within 2 months before surgery, and had complete postoperative pathology reports that recorded the number, location, and size of metastatic lymph nodes and whether there was extracapsular invasion. Patients with prior history of neck surgery, chemotherapy, or radiation therapy or who did not undergo neck dissection were excluded.

Deep learning prediction models were trained and internally examined in the internal dataset using standard nested fivefold cross-validation. Specifically, the data were randomly partitioned into five folds (ie, nonoverlapping subsets). For each of the five subsets in turn, the subset was reserved as the internal test set, and the remaining four subsets were combined for model training and hyperparameter tuning. Looping this process through all five subsets provided more comprehensive and statistically reliable internal evaluation results.

To evaluate the generalizability of the diagnostic model, four patient cohorts across institutions were collected to form four external test sets: *(a)* patients with LHSCC treated at the Eye & ENT Hospital of Fudan University between October 2018 and September 2024 (external test set 1), *(b)* patients with LHSCC treated at Chang Gung Memorial Hospital between February 2018 and July 2023 (external test set 2), *(c)* patients with LHSCC selected from The Cancer Genome Atlas Head-Neck Squamous Cell Carcinoma Collection (The Cancer Imaging Archive) between May 1989 and July 2003 (external test set 3), and *(d)* patients with oral squamous cell carcinoma treated at Chang Gung Memorial Hospital between March 2018 and October 2019 (external test set 4). The inclusion and exclusion criteria for the three external LHSCC cohorts were the same as those for the internal cohort. For external test set 4, the inclusion and exclusion criteria were the same as those for the internal cohort, except that patients with oral squamous cell carcinoma were included. The patient flowchart is presented in Figure 1.

Lymph Node Metastasis Status Annotation

A physician-in-the-loop procedure was developed for lymph node labeling on CT scans. It involved semiautomatic lymph node segmentation using a recently developed transformer-based model, LN-DETR (22), and manual refinement by physicians (Table S1). A deep segmentation model (23) was developed to automatically delineate three-dimensional lymph node stations (I–V) in the neck region (Table S2). Two head and neck cancer surgeons (N.S. and X.H., with over 15 and over 25 years of experience, respectively) labeled the metastasis and ENE status of each lymph node based on refined lymph node masks, autosegmented stations, and pathology reports. Annotations were first made independently, and a review board resolved inconsistent annotations, if consensus could not be reached after discussion. Detailed labeling procedures and criteria are provided in Appendix S1.

DeepENE Model Development

The overall model development workflow is shown in Figure 2. The developed model, DeepENE, is a two-stream 2.5-dimensional multiscale deep feature fusion network aiming to effectively fuse local and global lymph node characteristics to differentiate negative, metastasis-positive, and ENE-positive nodal status. Details on DeepENE model architecture, preprocessing, and network training can be found in Appendixes S2 and S3, Figure S1, and Table S3. The model was trained and internally evaluated using nested

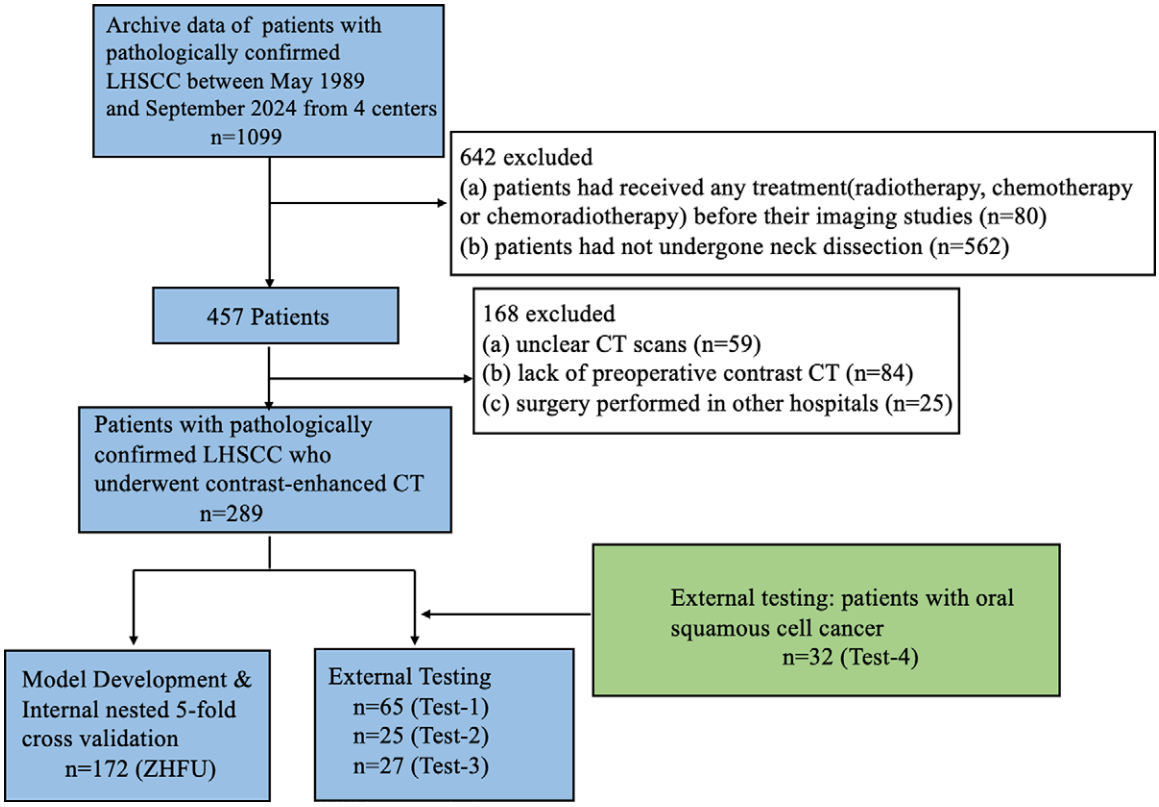

Figure 1: Patient flowchart for this study. External test sets 1–3 contained patients with laryngeal and hypopharyngeal squamous cell cancer (LHSCC): external test set 1 (Test-1) from Eye & ENT Hospital of Fudan University, external test set 2 (Test-2) from Chang Gung Memorial Hospital, and external test set 3 (Test-3) from The Cancer Genome Atlas Head-Neck Squamous Cell Carcinoma Collection (The Cancer Imaging Archive). External test set 4 contained patients with oral squamous cell carcinoma from Chang Gung Memorial Hospital. ZHFU = Zhongshan Hospital, Fudan University.

fivefold cross-validation in the internal dataset. In external evaluation, bootstrapping was used to perform the model ensemble, with the model predictions averaged across five individual folds, each with five nested runs. The codes used for model inference are provided at *https://github.com/ENThuang/rtnet*.

Reader Study

The reader study involved five board-certified physicians: two radiologists (both with over 15 years of experience, one of whom was a specialized head and neck radiologist), two radiation oncologists (with over 5 and 10 years of experience, respectively), and one nuclear medicine physician (with over 15 years of experience). For all of these readers, clinical duties involved detection of lymph node metastasis. No readers were involved in annotation of the training set or algorithm development. Readers were provided with CT scans with preidentified and segmented masks of lymph nodes and asked to make independent judgments on the metastasis and ENE status of each node based on their professional experience using the preoperative clinical references 15 and 24 (Appendix S5). For each node, readers chose one of three categories: benign, metastasis without ENE, or metastasis with ENE. In the reader study, no artificial intelligence predictions were provided to the readers.

Statistical Analysis

Statistical analysis was performed by one author (Y.W.) using R version 4.2.2 (R Foundation for Statistical Computing) and Stata version 17.0 (StataCorp). The primary metric used to evaluate the performance of DeepENE was the AUC. The 95% CIs were calculated, and the AUC of DeepENE was compared with that of each reader using the DeLong method (25), with $P < .05$ considered to indicate a statistically significant difference. The specific metrics of DeepENE were determined using various probability thresholds: the Youden index (optimizing the combined sensitivity and specificity) and thresholds that yielded clinically meaningful false-positive rates of 5%, 10%, 15%, 20%, and/or 30%. Categorical data were analyzed with the Pearson χ^2 test or Fisher exact test, and continuous data were analyzed with the Student *t* test or Mann-Whitney *U* test. Fleiss κ was used to evaluate interobserver agreement.

Results

Patient Characteristics

A total of 289 patients with LHSCC and 32 patients with oral squamous cell carcinoma (mean age, 61 years ± 9 [SD]; age range, 33–89 years; 307 men) were included in the final study sample (internal dataset and all external test sets). From an initial 1099 patients with LHSCC treated at the participating hospitals or selected from The Cancer Genome Atlas Head-Neck Squamous Cell Carcinoma Collection, patients were excluded for the following reasons: received treatments before CT imaging ($n = 80$), lacked preoperative contrast-enhanced neck CT ($n = 84$), poor-quality CT scans ($n = 59$), surgical dissection performed in other hospitals ($n = 25$), or neck dissection not performed ($n = 562$). The dataset used for model training and internal testing included 172 patients from Zhongshan Hospital, Fudan University. The

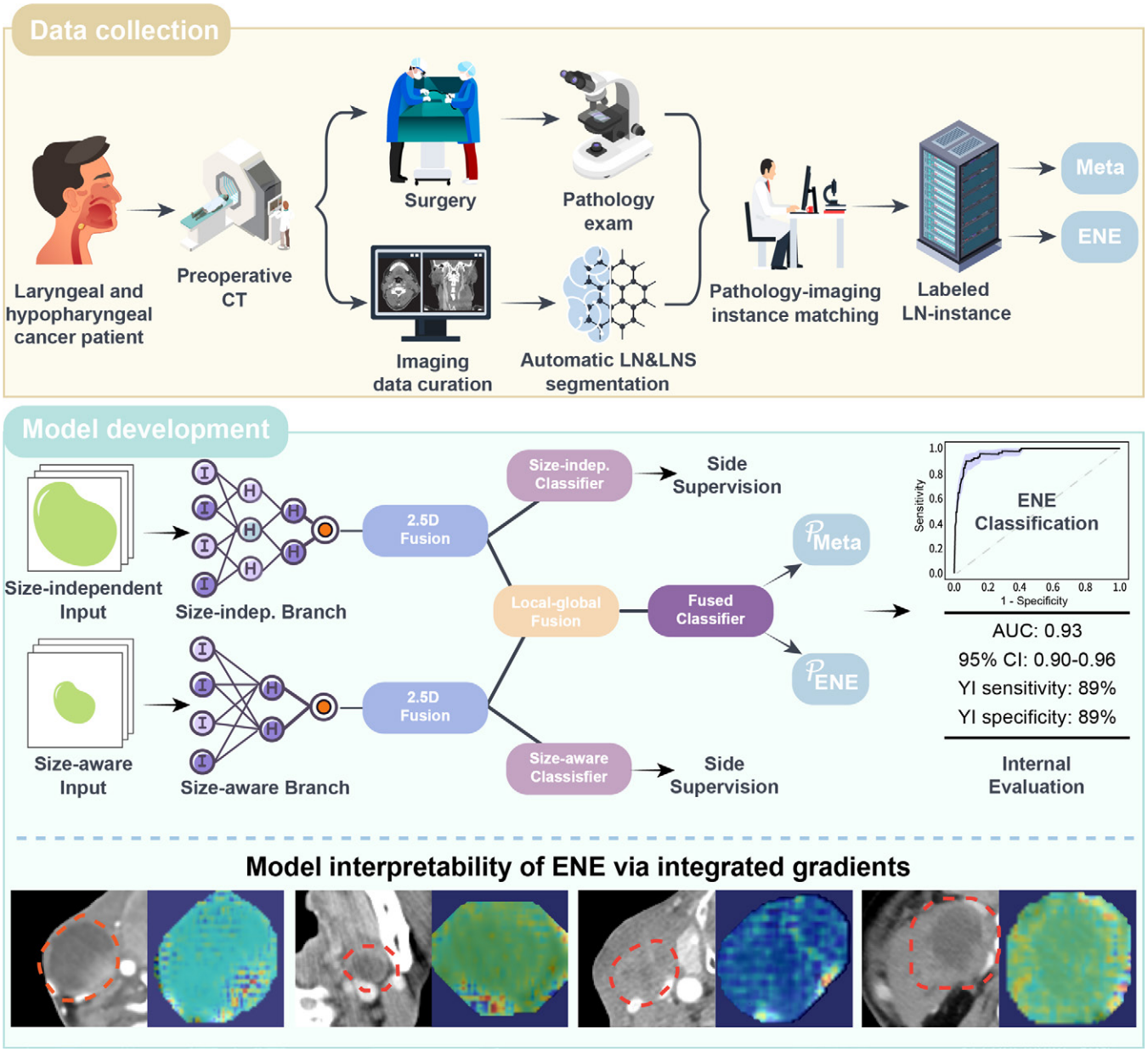

Figure 2: The development workflow for the deep learning model consisted of two main stages: data collection and model development. Data were collected from patients with laryngeal and hypopharyngeal cancer who underwent contrast-enhanced preoperative CT. Pathology reports were collected after surgery to provide the reference standard for lymph node (LN) metastasis (Meta) status labeling. A human-in-the-loop lymph node mask delineation procedure was used, involving a lymph node station (LNS) autosegmentation deep network and a lymph node detection deep network, to produce initial lymph node candidates and instance-wise mask delineation. Two senior physicians with over 20 years of experience reviewed and refined the machine-generated masks. Then, matching between the pathology report and CT scan was conducted to label instance-wise metastasis status. For model development, a size-independent (size-indep.) and size-aware two-stream network architecture was adopted to simultaneously encode unaltered lymph node features and zoomed-in features that better accounted for subtle localized changes in textures and boundaries. Fivefold cross-validation was used to evaluate the developed model on the internal set. To examine model interpretability, integrated gradients were used to identify the contribution of the raw input pixels to extranodal extension (ENE)–positive predictions. Results showed that the model correctly focused on the boundary areas where ENE characteristics were present, as shown in the example CT images at the bottom of the figure. The dashed outlines on the CT images indicate the boundaries of the lymph nodes. To the right of each CT image is a magnified image of the lymph node showing the most salient region used to classify ENE. AUC = area under the receiver operating characteristic curve, 2.5D = 2.5-dimensional, YI = Youden index.

four external test sets consisted of 149 patients total, including 65, 25, and 27 patients with LHSCC in external test sets 1, 2, and 3, respectively, and 32 patients with oral squamous cell carcinoma in external test set 4. Detailed patient characteristics are shown in Tables 1 and 2.

Lymph Node Characteristics

In the 172 patients in the internal set, 1011 lymph nodes were semiautomatically segmented and labeled on CT scans. Of these, 226 (22.4%) were malignant nodes without ENE, and 94 (9.3%) were malignant nodes with ENE (Table 1). Median short-axis diameter was 2.2 cm (range, 0.5–3.7 cm) for malignant nodes with ENE (Fig 3A), 1.0 cm (range, 0.4–2.7 cm) for malignant nodes without ENE, and 0.6 cm (range, 0.5–3.0 cm) for benign nodes.

In external test set 1, 471 lymph nodes were labeled on CT scans from 65 patients. Of segmented nodes, 92 (19.5%) were malignant without ENE, and 31 (6.6%) were malignant with ENE (Table 2). Median short-axis diameter was 2.2 cm (range,

Table 1: Patient and Lymph Node Characteristics in All Datasets and the Internal Dataset

	All Datasets		Internal Dataset	
Characteristic	Patients (n = 321)	Lymph Nodes (n = 2363)	Patients (n = 172)	Lymph Nodes (n = 1011)
Mean age (y)*	61 ± 9	NA	61 ± 8	NA
Sex				
Male	307	NA	170	NA
Female	14	NA	2	NA
Primary cancer site				
Larynx	170 (53.0)	1091 (46.2)	108 (62.8)	589 (58.3)
Hypopharynx	119 (37.1)	863 (36.5)	64 (37.2)	422 (41.7)
Oral cavity	32 (10.0)	409 (17.3)	NA	NA
Pathologic T stage				
T1	17 (5.3)	172 (7.3)	5 (2.9)	30 (3.0)
T2	94 (29.3)	640 (27.1)	54 (31.4)	306 (30.3)
T3	114 (35.5)	748 (31.7)	73 (42.4)	427 (42.2)
T4	96 (29.9)	803 (34.0)	40 (23.3)	248 (24.5)
Pathologic N stage				
N0	75 (23.4)	450 (19.0)	38 (22.1)	184 (18.2)
N1	33 (10.3)	289 (12.2)	14 (8.1)	67 (6.6)
N2	83 (25.9)	614 (26.0)	45 (26.2)	243 (24.0)
N3	130 (40.5)	1010 (42.7)	75 (43.6)	517 (51.1)
Lymph node pathologic finding				
Benign	87 (27.1)	1780 (75.3)	49 (28.5)	691 (68.3)
Metastasis without ENE	112 (34.9)	402 (17.0)	55 (32.0)	226 (22.4)
Metastasis with ENE	122 (38.0)	181 (7.7)	68 (39.5)	94 (9.3)

Note.—Unless otherwise indicated, data are numbers of patients or lymph nodes, with percentages in parentheses. ENE = extranodal extension, NA = not applicable.

* Data are means ± SDs.

1.1–4.2 cm) for malignant nodes with ENE (Fig 3A), 1.2 cm (range, 0.6–2.3 cm) for malignant nodes without ENE, and 0.7 cm (range, 0.5–1.4 cm) for benign nodes.

In external test set 2, 194 lymph nodes were labeled on CT scans from 25 patients. Among the segmented nodes, 10 (5.2%) were malignant nodes without ENN, and 23 (11.9%) were malignant nodes with ENE (Table 2). The median short-axis diameter was 1.1 cm (range, 0.5–3.8 cm) for malignant nodes with ENE (Fig 3A), 0.8 cm (range, 0.6–1.5 cm) for malignant nodes without ENE, and 0.6 cm (range, 0.5–1.9 cm) for benign nodes. Note that the size distribution of malignant nodes with ENE in external test set 2 was substantially smaller than that in the internal dataset and external test set 1 (Fig 3A), and the percentage of ENE lymph nodes at early pathologic stages in external test set 2 was also much higher than that in external test set 1 (Table S4). This indicated a more challenging diagnostic task in patients in external test set 2.

For external test set 3, 278 lymph nodes were labeled on CT scans from 27 patients. Of the segmented nodes, 35 (12.6%) were malignant nodes without ENE, and 17 (6.1%) were malignant nodes with ENE (Table 2). The median short-axis diameter was 1.3 cm for malignant nodes with ENE (range, 0.5–6.1 cm) (Fig 3A), 1.0 cm (range, 0.5–2.6 cm) for malignant nodes without ENE, and 0.6 cm (range, 0.5–1.3 cm) for benign nodes.

For patients with oral squamous cell carcinoma in external test set 4, 409 lymph nodes were labeled on CT scans from 32 patients. Among the segmented nodes, 39 (9.5%) were malignant nodes without ENE, and 16 (3.9%) were malignant nodes with ENE (Table 2). The median short-axis diameter was 0.8 cm (range, 0.5–1.3 cm) for malignant nodes with ENE, 0.7 cm (range, 0.4–2.0 cm) for malignant nodes without ENE, and 0.5 cm (range, 0.2–1.8 cm) for benign nodes.

DeepENE Performance for ENE Prediction

DeepENE achieved an AUC of 0.93 (95% CI: 0.90, 0.96) for ENE identification at the lesion level in the internal dataset under fivefold cross-validation (Table 3). Sensitivity, specificity, and accuracy were all 89% when calculated using the Youden index threshold.

DeepENE achieved an overall AUC of 0.91 (95% CI: 0.86, 0.95) in all external testing data (Fig 3B), substantially higher than that of the five readers (mean AUC, 0.75 [95% CI: 0.68, 0.82]; P < .001). Specifically in external test set 1 (Fig 3C, Table 4), DeepENE had an AUC of 0.96 (95% CI: 0.94, 0.98), markedly higher than that of the five readers (mean AUC, 0.85 [95% CI: 0.76, 0.93]; P < .001). There was large variation in readers' sensitivity (mean sensitivity, 77%; range, 55%–90%), while they normally exhibited good specificity (mean specificity, 93%; range, 86%–97%). Interobserver agreement was moderate (Fleiss κ = 0.49). In comparison, DeepENE achieved a sensitivity of 97%, significantly higher than the mean sensitivity of human readers (P = .003), at a specificity of 80%–90%.

In external test set 2 (Fig 3D, Table 4), physician readers encountered substantial difficulty in identifying nodal ENE and had lower sensitivities and AUCs compared with their performance in external test set 1. Physician readers exhibited a

Table 2: Patient and Lymph Node Characteristics in the External Datasets

Characteristic	External Test Set 1		External Test Set 2		External Test Set 3		External Test Set 4	
	Patients (*n* = 65)	Lymph Nodes (*n* = 471)	Patients (*n* = 25)	Lymph Nodes (*n* = 194)	Patients (*n* = 27)	Lymph Nodes (*n* = 278)	Patients (*n* = 32)	Lymph Nodes (*n* = 409)
Mean age (y)*	62 ± 9	NA	60 ± 10	NA	61 ± 8	NA	55 ± 11	NA
Sex								
Male	62	NA	24	NA	20	NA	31	NA
Female	3	NA	1	NA	7	NA	1	NA
Primary cancer site								
Larynx	33 (51)	244 (51.8)	11 (44)	75 (38.7)	18 (67)	183 (65.8)	NA	NA
Hypopharynx	32 (49)	227 (48.2)	14 (56)	119 (61.3)	9 (33)	95 (34.2)	NA	NA
Oral cavity	NA	NA	NA	NA	NA	NA	32 (100)	409 (100)
Pathologic T stage								
T1	1 (2)	10 (2.1)	1 (4)	10 (5.2)	0 (0)	0 (0.0)	10 (31)	122 (29.8)
T2	21 (32)	135 (28.7)	3 (12)	24 (12.4)	3 (11)	15 (5.4)	13 (41)	160 (39.1)
T3	28 (43)	218 (46.3)	7 (28)	47 (24.2)	5 (19)	47 (16.9)	1 (3)	9 (2.2)
T4	15 (23)	108 (22.9)	14 (56)	113 (58.2)	19 (70)	216 (77.7)	8 (25)	118 (28.9)
Pathologic N stage								
N0	12 (18)	48 (10.2)	12 (48)	81 (41.8)	12 (44)	130 (46.8)	1 (3)	7 (1.7)
N1	0 (0)	0 (0.0)	2 (8)	36 (18.6)	2 (7)	14 (5.0)	15 (47)	172 (42.1)
N2	27 (42)	222 (47.1)	0 (0)	0 (0.0)	5 (19)	64 (23.0)	6 (19)	85 (20.8)
N3	26 (40)	201 (42.7)	11 (44)	77 (39.7)	8 (30)	70 (25.2)	10 (31)	145 (35.5)
Lymph node pathologic finding								
Benign	13 (20)	348 (73.9)	12 (48)	161 (83.0)	12 (44)	226 (81.3)	1 (3)	354 (86.6)
Metastasis without ENE	27 (42)	92 (19.5)	2 (8)	10 (5.2)	7 (26)	35 (12.6)	21 (66)	39 (9.5)
Metastasis with ENE	25 (38)	31 (6.6)	11 (44)	23 (11.9)	8 (30)	17 (6.1)	10 (31)	16 (3.9)

Note.—Unless otherwise indicated, data are numbers of patients or lymph nodes, with percentages in parentheses. External test set 1 was from the Eye & ENT Hospital of Fudan University. External test set 2 was from Chang Gung Memorial Hospital. External test set 3 was from The Cancer Genome Atlas Head-Neck Squamous Cell Carcinoma Collection. External test set 4 was from Chang Gung Memorial Hospital. ENE = extranodal extension, NA = not applicable.

* Data are means ± SDs.

mean AUC of 0.85 (95% CI: 0.76, 0.93) in external test set 1 versus 0.66 (95% CI: 0.54, 0.79) in external test set 2 (mean decrease, 0.19; $P < .001$) and a mean sensitivity of 77% (24 of 31) in external test set 1 versus 36% (eight of 23) in external test set 2 (mean decrease, 41%; $P < .001$). Interobserver agreement was poor in external test set 2 (Fleiss $\kappa = 0.33$). DeepENE had higher performance than physician readers in external test set 2 (AUC of 0.87 [95% CI: 0.76, 0.97] for DeepENE vs mean AUC of 0.66 [95% CI: 0.54, 0.79] for physician readers; $P < .001$). Using the Youden index threshold, the sensitivity of DeepENE was 78% with a specificity of 90%. When the threshold was set to allow no more than a 5% false-positive rate, DeepENE still yielded a sensitivity of 70%, which is much higher than the sensitivity achieved by physician readers (mean sensitivity, 36%; sensitivity for individual readers: 17%, 35%, 39%, 39%, 52%; all $P < .001$). Examples of correct and incorrect predictions made by DeepENE and physician readers are shown in Figure 4.

In external test set 3 (Fig 3E, Table 4), DeepENE yielded an AUC of 0.90 (95% CI: 0.82, 0.98), substantially higher than that of the five readers (mean AUC, 0.71; AUC range, 0.58–0.90; $P < .001$). Physician readers generally had low sensitivity (mean, 46%; range, 18%–88%), but they exhibited high specificity (mean, 96%; range, 93%–99%). Interobserver agreement was fair (Fleiss $\kappa = 0.41$). In comparison, DeepENE achieved a balanced performance with a sensitivity of 87% at a specificity of 87% at the Youden index threshold, significantly higher than the mean sensitivity of the five readers (46%; $P < .001$). Confusion matrices for DeepENE and the five human readers in the three external test sets are provided in Figure S2, further demonstrating the performance consistency of DeepENE across different centers.

DeepENE Performance for Lymph Node Metastasis Prediction

The performance of DeepENE and physician readers in distinguishing metastatic and benign lymph nodes was evaluated in the three external LHSCC test sets (Fig 5, Table 5). DeepENE achieved AUCs of 0.91 (95% CI: 0.87, 0.94), 0.83 (95% CI: 0.75, 0.92), and 0.83 (95% CI: 0.76, 0.89) in external test sets 1, 2, and 3 respectively, markedly higher than those of the five readers (external test set 1: mean AUC, 0.74 [95% CI: 0.69, 0.80], $P < .001$; external test set 2: mean AUC, 0.74 [95% CI: 0.63, 0.84], $P = .02$; external test set 3: mean AUC, 0.74 [95% CI: 0.65, 0.82], $P = .003$). Physician readers exhibited large performance variation across datasets. For example, reader 1 had high sensitivity (83%) in external test set 1 but lower sensitivity in external test sets 2 and 3 (67% and 60%).

Quantitative results were also calculated for the performance of DeepENE and physician readers in distinguishing between metastatic nodes without ENE and those with ENE in the three external LHSCC test sets (Fig 6, Table 6). In this

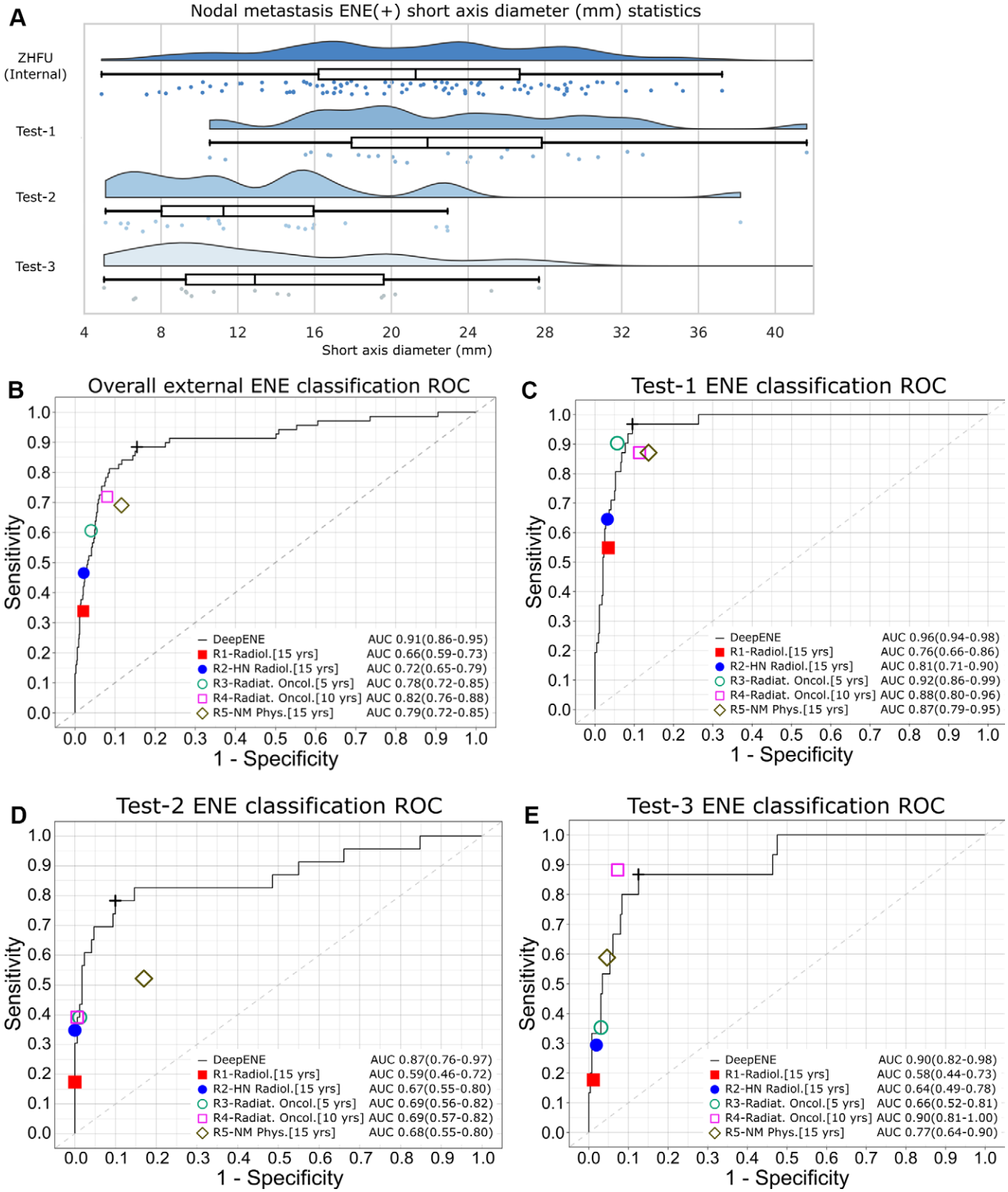

Figure 3: Extranodal extension (ENE) classification performance of DeepENE and physician readers. **(A)** Distribution of short-axis diameters of lymph nodes with ENE in the internal dataset (training, validation, internal test sets from Zhongshan Hospital, Fudan University [ZHFU]) and three external test sets of patients with laryngeal and hypopharyngeal squamous cell carcinoma: external test set 1 (Test-1) from the Eye & ENT Hospital of Fudan University, external test set 2 (Test-2) from Chang Gung Memorial Hospital, and external test set 3 (Test-3) from The Cancer Genome Atlas Head-Neck Squamous Cell Carcinoma Collection. In addition to the smoothed distribution, individual short-axis values are shown as dots, and box and whisker plots along the distribution are shown (box, IQR; line, median; whiskers, 1.5 × IQR). Samples from external test sets 2 and 3 had overall smaller lymph nodes with ENE compared with the internal set and external test set 1, which suggests that external test sets 2 and 3 represent more challenging diagnostic cohorts. **(B–E)** Receiver operating characteristic (ROC) curves of DeepENE performance and comparison with five physician readers (R1, R2, R3, R4, R5) on **(B)** all data from external test sets 1–3, **(C)** external test set 1, **(D)** external test set 2, and **(E)** external test set 3. Note that DeepENE surpassed all five physician readers on external test set 2 (all $P < .001$). Years of experience for readers are given in brackets. AUC = area under the ROC curve, HN = head and neck, NM Phys. = nuclear medicine physician, Radiat. Oncol. = radiation oncologist, Radiol. = radiologist.

Table 3: Quantitative Performance of DeepENE for Lymph Node Metastasis and ENE Classification in the Internal Cohort under Fivefold Cross-validation

	Metastasis Identification				ENE Identification			
Analysis	AUC*	Sensitivity (%)	Specificity (%)	Accuracy (%)	AUC*	Sensitivity (%)	Specificity (%)	Accuracy (%)
DeepENE	0.84 (0.82, 0.87)				0.93 (0.90, 0.96)			
Probability threshold								
Youden index		70	86	80		89	89	89
FPR 30%		81	70	74		95	70	72
FPR 20%		75	80	78		91	80	81
FPR 10%		60	90	78		87	90	90

Note.—The numerators and denominators for sensitivity, specificity, and accuracy in the table are shown in Table S7. AUC = area under the receiver operating characteristic curve, ENE = extranodal extension, FPR = false-positive rate.

* Data in parentheses are 95% CIs.

task, DeepENE outperformed the five readers in external test sets 1 and 2 (by AUC margins of 0.11 and 0.15, respectively; both $P < .05$) but not in external test set 3 ($P = .06$). Physician readers showed significant performance variation across datasets. For instance, the mean sensitivity of physician readers was 77%, 36%, and 46% in external test sets 1, 2, and 3, respectively, while the mean specificity was 72%, 94%, and 79%, respectively.

The performance of DeepENE was compared with that of a recently developed deep learning–based ENE classification tool, DualNet, that was previously evaluated in patients with human papilloma virus–associated oropharyngeal cancer (14). In the three external LHSCC test sets, DeepENE had numerically higher AUCs than DualNet in most settings. For example, DeepENE had a numerically higher AUC than DualNet in external test set 2 for ENE prediction (0.87 vs 0.82; $P = .05$), metastasis prediction (0.83 vs 0.80; $P = .08$), and distinguishing between metastatic nodes with and without ENE (0.80 vs 0.73; $P = .07$) (Table S5). In the diagnostically relatively easy external test set 1, DeepENE and DualNet had almost the same performance.

DeepENE Performance for Metastasis and ENE Prediction in Oral Cavity Cancer

To assess whether lymph node metastasis regions affected the nodal differentiation ability of the model, the performance of DeepENE was examined in external test set 4. Nodal metastasis regions in oral squamous cell carcinoma are mostly focused in neck stations I–II, unlike LHSCC, for which most metastatic lymph nodes are observed in neck stations II–IV. DeepENE yielded an AUC of 0.85 (95% CI: 0.80, 0.91) for metastasis and 0.82 (95% CI: 0.73, 0.90) for ENE prediction (Table S6). The AUC for metastasis identification in external test set 4 (0.85 [95% CI: 0.80, 0.91]) was comparable to that in the LHSCC cohorts in external test set 2 (0.83 [95% CI: 0.75, 0.92]; $P = .67$) and external test set 3 (0.83 [95% CI: 0.76, 0.89]; $P = .76$), but ENE classification performance in external test set 4 (AUC, 0.82 [95% CI: 0.73, 0.90]) was numerically decreased compared with performance in the LHSCC cohorts in external test set 2 (AUC, 0.87 [95% CI: 0.76, 0.97]; $P = .23$) and external test set 3 (AUC, 0.90 [95% CI: 0.82, 0.98]; $P = .08$).

Discussion

Previous studies on applying deep learning models in extranodal extension (ENE) identification have focused mainly on oropharyngeal cancer (14), oral cancer (26,27), or a mix of the two (13). In this work, we used multicenter datasets to develop and evaluate the diagnostic performance of a deep learning model, DeepENE, in predicting lymph node metastasis and ENE in patients with laryngeal and hypopharyngeal squamous cell cancer (LHSCC). External testing showed that DeepENE could identify lymph node metastasis and ENE with high accuracy in patients with LHSCC in different cohorts with different levels of ENE severity. Furthermore, the diagnostic performance of DeepENE for identifying ENE surpassed that of five physicians, including board-certified radiologists, oncologists, and a nuclear medicine physician, in three external test sets (external test set 1: AUC of 0.96 for DeepENE vs mean AUC of 0.85 for readers, $P < .001$; external test set 2: 0.87 vs 0.66, $P < .001$; external test set 3: 0.90 vs 0.71, $P < .001$). DeepENE had higher sensitivity (97%, 78%, and 87% in external test sets 1, 2, and 3, respectively) compared with the mean sensitivity for physician readers (external test set 1: 77% [range, 55%–90%], $P = .003$; external test set 2: 36% [range, 17%–52%], $P < .001$; external test set 3: 46% [range, 18%–88%], $P < .001$).

In addition to consistent improvements in diagnostic performance, DeepENE offers the advantages of objectivity, reproducibility, and flexibility in adjusting the probability threshold to fit different clinical scenarios. In external test set 2, experts encountered substantial difficulty in identifying nodal ENE accurately, as reflected by lower sensitivities and AUCs compared with their performance in external test set 1 (a mean decrease of 40% in sensitivity and 0.19 in AUC). They also exhibited poor interobserver agreement, with a Fleiss κ of 0.33. We believe this is because the majority of ENE lymph nodes in external test set 1 (81%, 25 of 31) belonged to the most severe ENE grade (grade 3), where the tumor had invaded beyond perinodal fat to encase surrounding structures, thus making ENE easier to identify. In contrast, lymph nodes in external test set 2 had early-stage ENE and showed more subtle CT findings, which proved more difficult for human observers (15,28). DeepENE detected most cases of early-stage ENE in the difficult external test set 2 (AUC, 0.87; sensitivity, 78%; specificity, 90%).

Table 4: Quantitative Results of ENE Classification in Three External Test Sets of Patients with Laryngeal and Hypopharyngeal Squamous Cell Cancer

Analysis	AUC*	*P* Value†	Sensitivity	Specificity	Accuracy	Fleiss κ‡
External test set 1						0.49
DeepENE	0.96 (0.94, 0.98)					
Probability threshold						
Youden index			97	90	91	
FPR 20%			97	80	81	
FPR 10%			97	90	91	
FPR 5%			74	95	94	
Reader R1	0.79 (0.68, 0.90)	<.001	55	97	94	
Reader R2	0.81 (0.71, 0.90)	<.001	65	97	95	
Reader R3	0.92 (0.86, 0.99)	<.001	90	94	94	
Reader R4	0.88 (0.80, 0.96)	<.001	87	89	89	
Reader R5	0.87 (0.79, 0.95)	<.001	87	86	86	
Reader mean	0.85 (0.76, 0.93)	<.001	77	93	92	
External test set 2						0.33
DeepENE	0.87 (0.76, 0.97)					
Probability threshold						
Youden index			78	90	89	
FPR 20%			83	80	80	
FPR 10%			78	90	89	
FPR 5%			70	95	92	
Reader R1	0.59 (0.46, 0.72)	<.001	17	100	90	
Reader R2	0.67 (0.55, 0.80)	<.001	35	100	92	
Reader R3	0.69 (0.56, 0.82)	<.001	39	99	92	
Reader R4	0.69 (0.57, 0.82)	<.001	39	99	92	
Reader R5	0.68 (0.55, 0.80)	<.001	52	83	79	
Reader mean	0.66 (0.54, 0.79)	<.001	36	96	89	
External test set 3						0.41
DeepENE	0.90 (0.82, 0.98)					
Probability threshold						
Youden index			87	87	87	
FPR 20%			87	80	81	
FPR 10%			80	90	89	
FPR 5%			53	95	93	
Reader R1	0.58 (0.44, 0.73)	<.001	18	99	94	
Reader R2	0.64 (0.49, 0.78)	<.001	29	98	94	
Reader R3	0.66 (0.52, 0.81)	<.001	35	97	93	
Reader R4	0.90 (0.81, 1.00)	.05	88	93	92	
Reader R5	0.77 (0.64, 0.90)	<.001	59	95	93	
Reader mean	0.71 (0.56, 0.86)	<.001	46	96	93	

Note.—External test set 1 (n = 65 patients) was from the Eye & ENT Hospital of Fudan University. External test set 2 (n = 25 patients) was from Chang Gung Memorial Hospital. External test set 3 (n = 27 patients) was from The Cancer Genome Atlas Head-Neck Squamous Cell Carcinoma Collection. Reader R1 was a radiologist with over 15 years of experience. Reader R2 was a radiologist specialized in head and neck cancer with over 15 years of experience. Reader R3 was a radiation oncologist with over 5 years of experience. Reader R4 was a radiation oncologist with over 10 years of experience. Reader R5 was a nuclear medicine physician with over 15 years of experience. The numerators and denominators for sensitivity, specificity, and accuracy in the table are shown in Table S8. AUC = area under the receiver operating characteristic curve, ENE = extranodal extension, FPR = false positive rate.

* Data in parentheses are 95% CIs.

† *P* value for comparison of performance (AUC) against DeepENE.

‡ Fleiss κ for interobserver agreement across readers.

The superior performance of DeepENE may be attributed to the following factors. First, the training data were from a high-volume tertiary hospital, encompassing diverse nodal ENE conditions in patients with LHSCC of different severity. Second, our two-stream multiscale deep feature fusion network model effectively captured the intensity and texture patterns of nodal ENE status at both local and global perspectives. Third, human eyes may not generally be as sensitive as computerized models to subtle grayscale CT intensity differences, which is especially important for detecting early ENE signals. Lastly, there has been

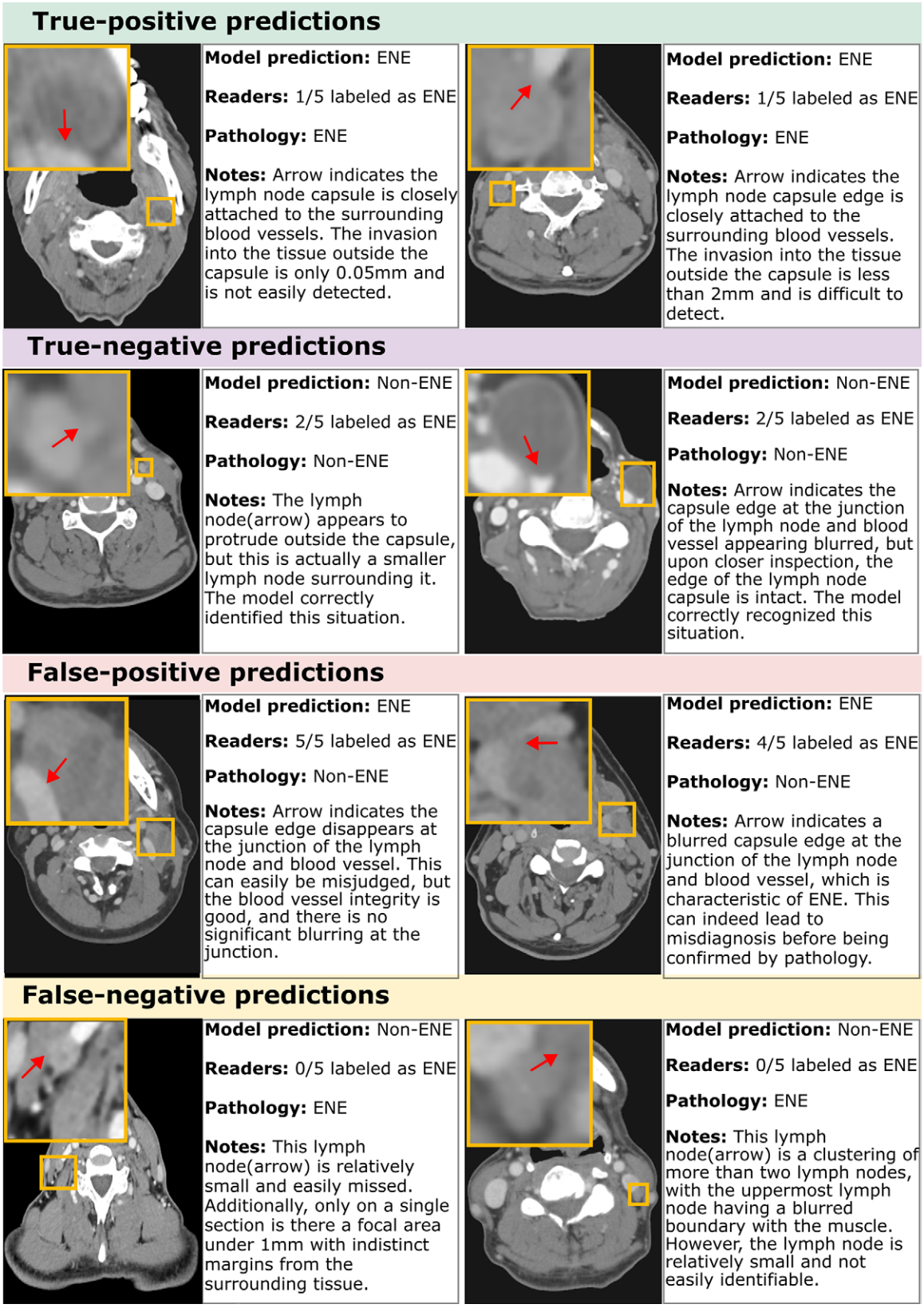

Figure 4: Qualitative analysis of successes and failures of extranodal extension (ENE) classification by DeepENE and experts. All images are axial contrast-enhanced CT images using the soft-tissue window (window level, 50 HU; window width, 350–450 HU). First row, left: image in a 47-year-old male patient with hypopharyngeal cancer who underwent left neck dissection (from external test set 2). First row, right: image in a 60-year-old male patient with hypopharyngeal cancer who underwent bilateral neck dissection (from external test set 2). Second row, left: image in a 64-year-old male with laryngeal cancer who underwent bilateral neck dissection (from external test set 1). Second row, right: image in a 69-year-old male patient with hypopharyngeal cancer who underwent bilateral neck dissection (from external test set 2). Third row, left: image in a 57-year-old male patient with hypopharyngeal cancer who underwent left neck dissection (from external test set 1). Third row, right: image in a 66-year-old male with hypopharyngeal cancer who underwent bilateral neck dissection (from external test set 1). Fourth row, left: image in a 60-year-old male patient with hypopharyngeal cancer who underwent bilateral neck dissection (from external test set 2). Fourth row, right: image in a 50-year-old male patient with laryngeal cancer who underwent left neck dissection (from external test set 2). Insets show enlargements, with arrows indicating relevant features described in each panel. External test set 1 was from the Eye & ENT Hospital of Fudan University. External test set 2 was from Chang Gung Memorial Hospital.

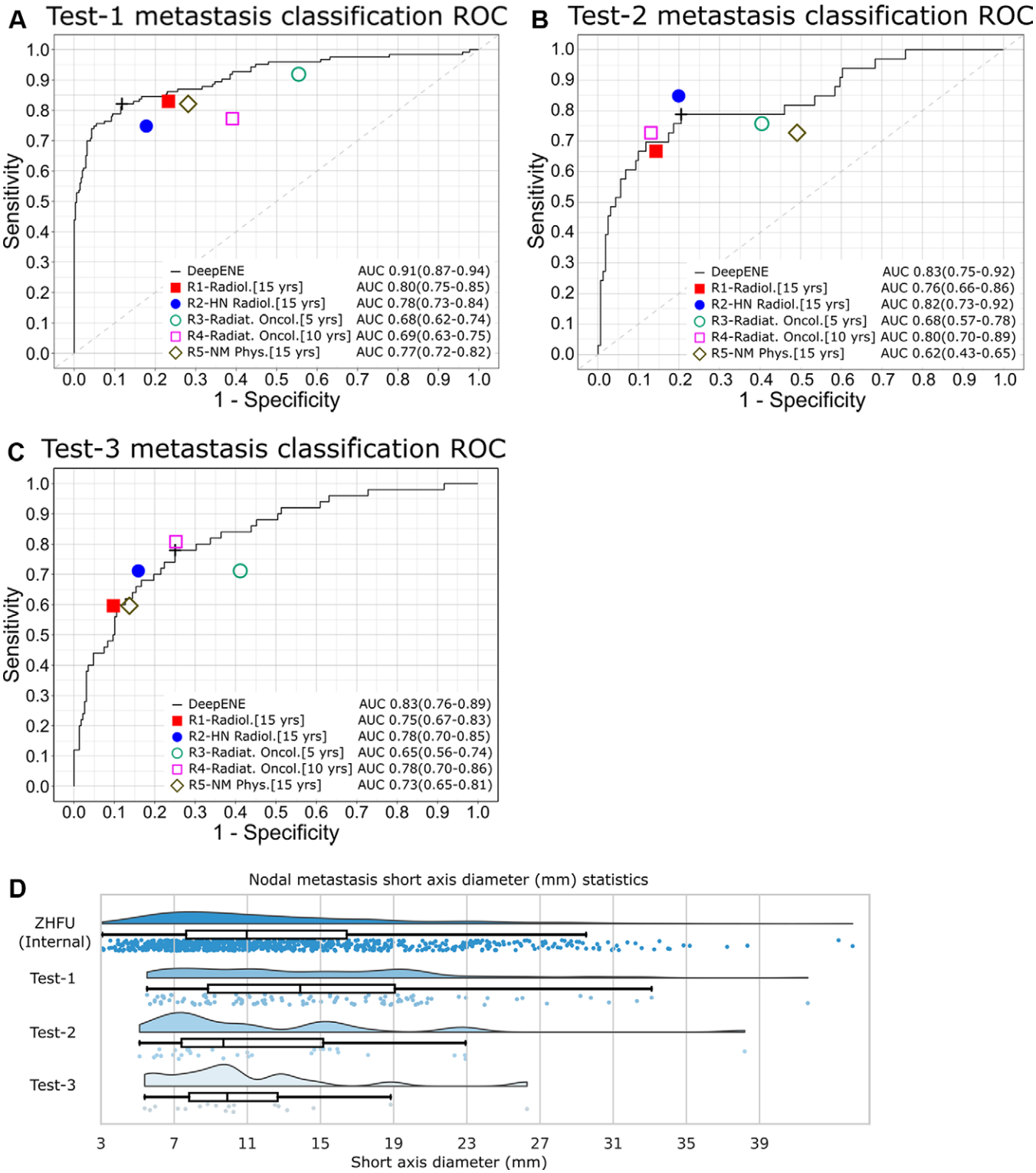

Figure 5: Nodal metastasis classification performance of DeepENE and physician readers. **(A–C)** Receiver operating characteristic (ROC) curves of DeepENE performance and comparison with five physician readers (R1, R2, R3, R4, R5) in three external test sets of patients with laryngeal and hypopharyngeal squamous cell carcinoma: **(A)** external test set 1 (Test-1) from the Eye & ENT Hospital of Fudan University, **(B)** external test set 2 (Test-2) from Chang Gung Memorial Hospital, and **(C)** external test set 3 (Test-3) from The Cancer Genome Atlas Head-Neck Squamous Cell Carcinoma Collection. **(D)** Distribution of short-axis diameters of metastatic lymph nodes in the internal dataset (training, validation, internal test sets from Zhongshan Hospital, Fudan University [ZHFU]) and the three external test sets. In addition to the smoothed distribution, individual short-axis values are shown as dots, and box and whisker plots along the distribution are shown (box, IQR; line, median; whiskers, 1.5 × IQR). DeepENE generally outperformed the physician readers on nodal metastasis classification in the three external test sets. Years of experience for readers are given in brackets. AUC = area under the ROC curve, HN = head and neck, NM Phys. = nuclear medicine physician, Radiat. Oncol. = radiation oncologist, Radiol. = radiologist.

Table 5: Quantitative Results of Lymph Node Metastasis Classification in Three External Test Sets of Patients with Laryngeal and Hypopharyngeal Squamous Cell Cancer

Analysis	AUC*	*P* Value†	Sensitivity	Specificity	Accuracy	Fleiss κ‡
External test set 1						0.41
DeepENE	0.91 (0.87, 0.94)					
Probability threshold						
Youden index			82	88	87	
FPR 30%			87	70	74	
FPR 20%			85	80	81	
FPR 10%			79	90	87	
Reader R1	0.80 (0.75, 0.85)	<.001	83	77	78	
Reader R2	0.78 (0.73, 0.84)	<.001	75	82	80	
Reader R3	0.68 (0.62, 0.74)	<.001	92	45	57	
Reader R4	0.69 (0.63, 0.75)	<.001	77	61	65	
Reader R5	0.77 (0.72, 0.82)	<.001	82	72	75	
Reader mean	0.74 (0.69, 0.80)	<.001	82	67	71	
External test set 2						0.37
DeepENE	0.83 (0.75, 0.92)					
Probability threshold						
Youden index			79	80	79	
FPR 30%			79	70	71	
FPR 20%			76	80	79	
FPR 10%			67	90	86	
Reader R1	0.76 (0.66, 0.86)	.05	67	86	82	
Reader R2	0.82 (0.73, 0.92)	.43	85	80	81	
Reader R3	0.68 (0.57, 0.78)	<.001	76	60	62	
Reader R4	0.80 (0.70, 0.89)	.22	73	87	85	
Reader R5	0.62 (0.43, 0.65)	<.001	73	51	55	
Reader mean	0.74 (0.63, 0.84)	.02	75	73	73	
External test set 3						0.40
DeepENE	0.83 (0.76, 0.89)					
Probability threshold						
Youden index			78	75	76	
FPR 30%			78	70	71	
FPR 20%			70	80	78	
FPR 10%			50	90	83	
Reader R1	0.75 (0.67, 0.83)	<.001	60	90	85	
Reader R2	0.78 (0.70, 0.85)	.07	71	84	82	
Reader R3	0.65 (0.56, 0.74)	<.001	71	59	61	
Reader R4	0.78 (0.70, 0.86)	.07	81	75	76	
Reader R5	0.73 (0.65, 0.81)	.002	60	86	81	
Reader mean	0.74 (0.65, 0.82)	.003	69	79	77	

Note.—External test set 1 (n = 65 patients) was from the Eye & ENT Hospital of Fudan University. External test set 2 (n = 25 patients) was from Chang Gung Memorial Hospital. External test set 3 (n = 27 patients) was from The Cancer Genome Atlas Head-Neck Squamous Cell Carcinoma Collection. Reader R1 was a radiologist with over 15 years of experience. Reader R2 was a radiologist specialized in head and neck cancer with over 15 years of experience. Reader R3 was a radiation oncologist with over 5 years of experience. Reader R4 was a radiation oncologist with over 10 years of experience. Reader R5 was a nuclear medicine physician with over 15 years of experience. The numerators and denominators for sensitivity, specificity, and accuracy in the table are shown in Table S9. AUC = area under the receiver operating characteristic curve, ENE = extranodal extension, FPR = false positive rate.

* Data in parentheses are 95% CIs.

† *P* value for comparison of performance (AUC) against DeepENE.

‡ Fleiss κ for interobserver agreement across readers.

no consensus on the diagnostic criteria for imaging-detected ENE. Hence, systematic training for experts in this task is lacking. Only in July 2024 did the Head and Neck Cancer International Group publish a consensus recommendation on the diagnostic criteria for imaging-detected ENE (29), which might help improve future performance of physician readers.

Previous studies have indicated the difficulty of preoperative ENE detection. Experts exhibited a wide range of sensitivity (45%–96%) and specificity (43%–96%), with most experts' sensitivity falling between 60% and 80% (15). Regarding deep learning–based approaches, Kann et al (13) reported AUC values of 0.9 and 0.84 for their artificial intelligence model in two external test

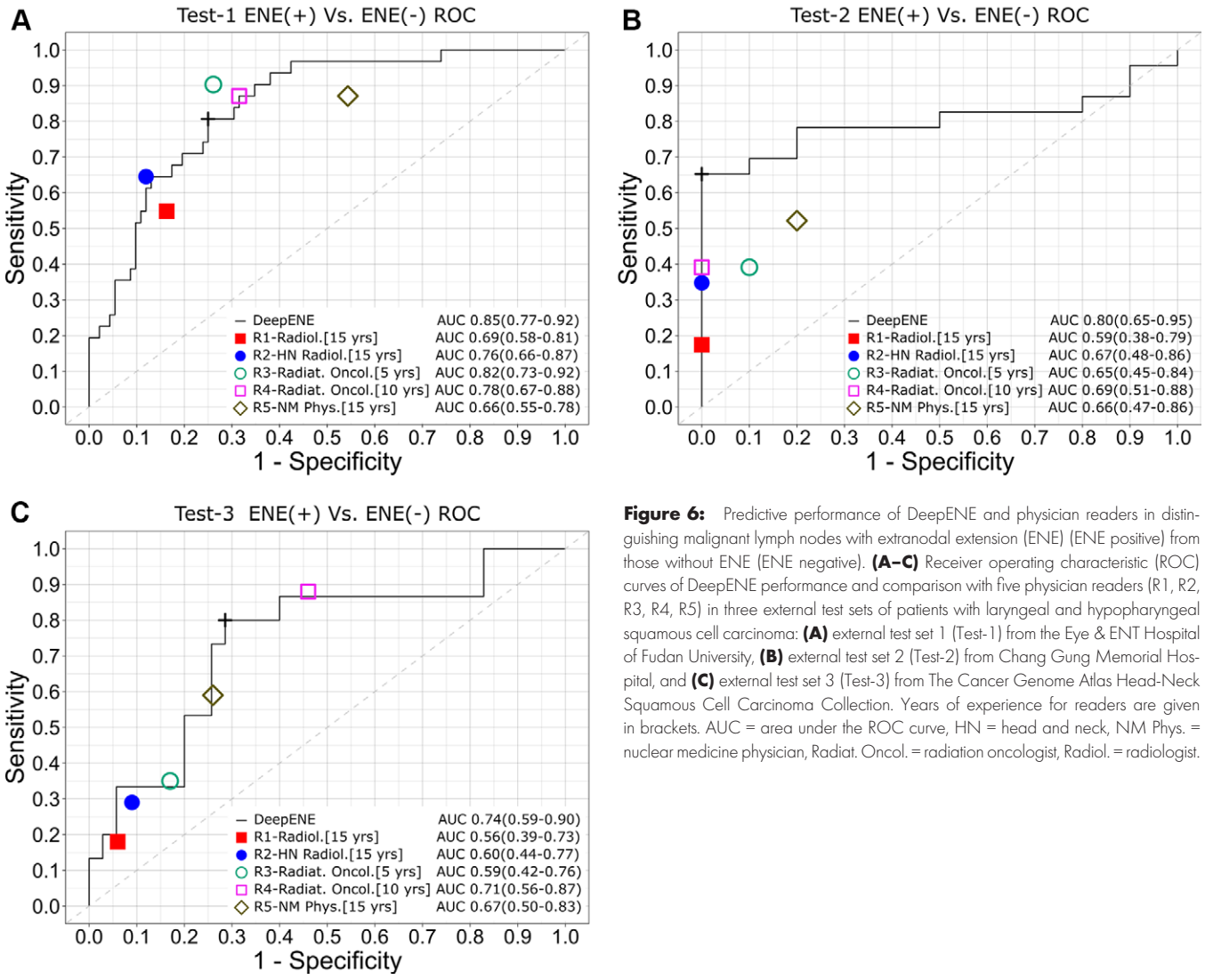

Figure 6: Predictive performance of DeepENE and physician readers in distinguishing malignant lymph nodes with extranodal extension (ENE) (ENE positive) from those without ENE (ENE negative). **(A–C)** Receiver operating characteristic (ROC) curves of DeepENE performance and comparison with five physician readers (R1, R2, R3, R4, R5) in three external test sets of patients with laryngeal and hypopharyngeal squamous cell carcinoma: **(A)** external test set 1 (Test-1) from the Eye & ENT Hospital of Fudan University, **(B)** external test set 2 (Test-2) from Chang Gung Memorial Hospital, and **(C)** external test set 3 (Test-3) from The Cancer Genome Atlas Head-Neck Squamous Cell Carcinoma Collection. Years of experience for readers are given in brackets. AUC = area under the ROC curve, HN = head and neck, NM Phys. = nuclear medicine physician, Radiat. Oncol. = radiation oncologist, Radiol. = radiologist.

sets of mostly oral and oropharyngeal cancer. They also evaluated the model's performance in a multicenter cohort of patients with human papilloma virus–associated oropharyngeal cancer, achieving an AUC of 0.86 (sensitivity of 49% at specificity of 90%, or sensitivity of 72% at specificity of 80%) (14). Yet these studies did not specifically focus on patients with LHSCC. They mainly focused on identifying ENE in nodes of larger sizes, with large lymph nodes selected and annotated in their training and test sets. Hence, the model's performance in identifying smaller nodes is unknown. In contrast, we annotated both large and small lymph nodes and demonstrated consistent high performance of DeepENE in patients with a wide range of lymph node sizes (median short-axis diameter for nodes with ENE was 2.2 and 1.1 cm in external test sets 1 and 2, respectively).

ENE status has been found to be a high-risk prognostic factor in laryngeal and hypopharyngeal cancer (8–10). The latest American Joint Committee on Cancer guidelines for head and neck cancers have updated pathologic N stage with positive ENE to N3b and stage IVB, regardless of the primary tumor status and number or size of lymph node metastases (28). Therefore, the developed model has the potential to help optimize cancer management: When ENE can be identified preoperatively, personalized precision treatment decisions can be made, such as functional neck dissection for N1–N3a or radical neck dissection for ENE if patients qualify for surgical resection. Moreover, DeepENE could contribute to the design of clinical trials by helping researchers accurately recruit patients with and those without ENE (30). Reliable and accurate preoperative diagnosis of ENE will facilitate effective clinical trials for patients with LHSCC (28).

Our study had some limitations. First, although our study involved a large number of patients with LHSCC, testing in more external patient cohorts would further benefit evaluation of the algorithm's performance. Second, our deep learning model should be integrated into the current clinical workflow to allow prospective examination of its role in real-world clinical decision-making processes. We have started preparation for model deployment in the hospital. Third, clinical data could be combined with imaging data to build a multimodal diagnostic model to further improve performance. Finally, contrast-enhanced MRI is a suitable alternative imaging technique for ENE diagnosis, as it provides superior soft-tissue contrast compared with CT. We plan to further explore the multimodal approach and incorporate MRI in future work.

In conclusion, we developed a new deep learning diagnostic tool, DeepENE, that accurately detected extranodal extension

Table 6: Quantitative Performance of DeepENE in Distinguishing Nodal Metastasis with ENE from Metastasis without ENE in Three External Test Sets

Analysis	AUC*	*P* Value†	Sensitivity	Specificity	Accuracy
External test set 1					
DeepENE	0.85 (0.77, 0.92)				
Probability threshold					
Youden index			81	75	76
FPR 30%			81	70	73
FPR 20%			71	80	78
FPR 15%			65	85	80
Reader R1	0.69 (0.58, 0.81)	<.001	55	84	76
Reader R2	0.76 (0.66, 0.87)	.02	65	88	82
Reader R3	0.82 (0.73, 0.92)	.30	90	74	78
Reader R4	0.78 (0.67, 0.88)	.03	87	68	73
Reader R5	0.66 (0.55, 0.78)	<.001	87	46	56
Reader mean	0.74 (0.64, 0.85)	.004	77	72	73
External test set 2					
DeepENE	0.80 (0.65, 0.95)				
Probability threshold					
Youden index			65	100	76
FPR 30%			78	70	76
FPR 20%			78	80	79
FPR 15%			70	85	74
Reader R1	0.59 (0.38, 0.79)	.003	17	100	42
Reader R2	0.67 (0.48, 0.86)	.05	35	100	55
Reader R3	0.65 (0.45, 0.84)	.02	39	90	55
Reader R4	0.69 (0.51, 0.88)	.09	39	100	58
Reader R5	0.66 (0.47, 0.86)	.04	52	80	61
Reader mean	0.65 (0.45, 0.85)	.03	36	94	54
External test set 3					
DeepENE	0.74 (0.59, 0.90)				
Probability threshold					
Youden index			80	71	74
FPR 30%			80	70	73
FPR 20%			53	81	72
FPR 15%			33	85	68
Reader R1	0.56 (0.39, 0.73)	.009	18	94	69
Reader R2	0.60 (0.44, 0.77)	.03	29	91	71
Reader R3	0.59 (0.42, 0.76)	.02	35	83	67
Reader R4	0.71 (0.56, 0.87)	.30	88	54	65
Reader R5	0.67 (0.50, 0.83)	.10	59	74	69
Reader mean	0.63 (0.45, 0.80)	.06	46	79	68

Note.—External test set 1 (n = 65 patients) was from the Eye & ENT Hospital of Fudan University. External test set 2 (n = 25 patients) was from Chang Gung Memorial Hospital. External test set 3 (n = 27 patients) was from The Cancer Genome Atlas Head-Neck Squamous Cell Carcinoma Collection. Reader R1 was a radiologist with over 15 years of experience. Reader R2 was a radiologist specialized in head and neck cancer with over 15 years of experience. Reader R3 was a radiation oncologist with over 5 years of experience. Reader R4 was a radiation oncologist with over 10 years of experience. Reader R5 was a nuclear medicine physician with over 15 years of experience. The numerators and denominators for sensitivity, specificity, and accuracy in the table are shown in Table S10. AUC = area under the receiver operating characteristic curve, ENE = extranodal extension, FPR = false positive rate.

* Data in parentheses are 95% CIs.

† *P* value for comparison of performance (AUC) against DeepENE.

on preoperative CT scans in patients with laryngeal and hypopharyngeal squamous cell cancer (LHSCC) and outperformed head and neck cancer specialists. Our deep learning model will be prospectively evaluated as a second-opinion assist tool for precise and personalized pretreatment assessment in patients with LHSCC.

Deputy Editor: Yoshimi Anzai
Scientific Editor: Kate Vilas

Author affiliations:

[1] Department of Otolaryngology–Head & Neck Surgery, Zhongshan Hospital, Fudan University, 180 Fenglin Rd, Shanghai 200032, China

[2] Department of Otolaryngology, Zhongshan Hospital, Fudan University, Xiamen, China

[3] DAMO Academy, Alibaba Group, New York, NY

[4] Department of Radiology, Zhongshan Hospital, Fudan University, Shanghai, China

[5] Department of Radiation Oncology, First Affiliated Hospital of Zhejiang University, Hangzhou, China

[6] Hupan Lab, Hangzhou, China

[7] Department of Radiation Oncology, Chang Gung Memorial Hospital, Linkou, Taiwan

[8] Department of Nuclear Medicine, Chang Gung Memorial Hospital, Linkou, Taiwan

[9] Department of Radiology, Eye & ENT Hospital, Fudan University, Shanghai, China

Received February 3, 2025; revision requested April 24; final revision received October 9; accepted November 21.

Address correspondence to: X.H. (email: huang.xinsheng@zs-hospital.sh.cn).

Supplemental material: Supplemental material is available at *Radiology* online.

Funding: This study was supported by grants from the Medical and Engineering Integration Project (XM03241807), Science and Intelligence Special Fund of Fudan University (X24AI059), and Shanghai Natural Science Foundation (25ZR1401055).

Author contributions: Guarantors of integrity of entire study, **N.S., Y.W., J.W., H.L., Q.Y., Y.C., L.L., X.Y., X.H., D.J.**; study concepts/study design or data acquisition or data analysis/interpretation, all authors; manuscript drafting or manuscript revision for important intellectual content, all authors; approval of final version of submitted manuscript, all authors; agrees to ensure any questions related to the work are appropriately resolved, all authors; literature research, **N.S., Y.W., C.Y., J.W., D.Z., X.W., H.L., Q.Y., Y.C., L.L., M.Z., D.J.**; clinical studies, **N.S., J.W., D.Z., X.W., H.L., Q.Y., Z.L., Y.C., M.Z., X.H., F.Z., D.J.**; experimental studies, **Y.W., C.Y., D.Z., D.G., H.L., Q.Y., Z.L., Y.C., K.Y., L.L., X.Y., T.Y.H., F.Z., D.J.**; statistical analysis, **N.S., Y.W., J.W., D.Z., D.G., Q.Y., Y.C., X.Y., X.H., T.Y.H., D.J.**; and manuscript editing, **N.S., J.W., D.Z., Q.Y., Y.C., K.Y., L.L., X.H., F.Z., D.J.**

Data sharing: Data generated or analyzed during the study are available from the corresponding author by request.

Disclosures of conflicts of interest: Please see ICMJE form(s) for author conflicts of interest. These have been provided as supplemental materials.

References

1. Lu JG, Li Y, Li L, Kan X. Overexpression of osteopontin and integrin alphav in laryngeal and hypopharyngeal carcinomas associated with differentiation and metastasis. J Cancer Res Clin Oncol 2011;137(11):1613–1618.
2. Zheng T, Xiao Y, Yang F, Dai G, Wang F, Chen G. The value of dual-layer spectral detector CT in preoperative T staging of laryngeal and hypopharyngeal squamous cell carcinoma. Eur J Radiol 2024;171:111287.
3. Liu Q, Liu S, Mao Y, Kang X, Yu M, Chen G. Machine learning model to preoperatively predict T2/T3 staging of laryngeal and hypopharyngeal cancer based on the CT radiomic signature. Eur Radiol 2024;34(8):5349–5359.
4. Abdeyrim A, He S, Zhang Y, et al. Prognostic value of lymph node ratio in laryngeal and hypopharyngeal squamous cell carcinoma: a systematic review and meta-analysis. J Otolaryngol Head Neck Surg 2020;49(1):31.
5. Xia C, Dong X, Li H, et al. Cancer statistics in China and United States, 2022: profiles, trends, and determinants. Chin Med J (Engl) 2022;135(5):584–590.
6. Siegel RL, Miller KD, Fuchs HE, Jemal A. Cancer statistics, 2022. CA Cancer J Clin 2022;72(1):7–33.
7. Ho AS, Kim S, Tighiouart M, et al. Association of quantitative metastatic lymph node burden with survival in hypopharyngeal and laryngeal cancer. JAMA Oncol 2018;4(7):985–989.
8. Bernier J, Cooper JS, Pajak TF, et al. Defining risk levels in locally advanced head and neck cancers: a comparative analysis of concurrent postoperative radiation plus chemotherapy trials of the EORTC (#22931) and RTOG (#9501). Head Neck 2005;27(10):843–850.
9. Wang Z, Zeng Q, Li Y, Lu T, Liu C, Hu G. Extranodal extension as an independent prognostic factor in laryngeal squamous cell carcinoma patients. J Cancer 2020;11(24):7196–7201.
10. Fan KH, Yeh CH, Hung SP, et al. Prognostic value of radiologic extranodal extension in patients with hypopharyngeal cancer treated with primary chemoradiation. Radiother Oncol 2021;156:217–222.
11. Bar-Ad V, Palmer J, Yang H, et al. Current management of locally advanced head and neck cancer: the combination of chemotherapy with locoregional treatments. Semin Oncol 2014;41(6):798–806.
12. Solimeno LS, Park YM, Lim JY, Koh YW, Kim SH. Treatment outcomes of neoadjuvant chemotherapy and transoral robotic surgery in locoregionally advanced laryngopharyngeal carcinoma. Head Neck 2021;43(11):3429–3436.
13. 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. J Clin Oncol 2020;38(12):1304–1311.
14. 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;5(6):e360–e369.
15. Henson CE, Abou-Foul AK, Morton DJ, et al. Diagnostic challenges and prognostic implications of extranodal extension in head and neck cancer: a state of the art review and gap analysis. Front Oncol 2023;13:1263347.
16. Carvalho P, Baldwin D, Carter R, Parsons C. Accuracy of CT in detecting squamous carcinoma metastases in cervical lymph nodes. Clin Radiol 1991;44(2):79–81.
17. Ardila D, Kiraly AP, Bharadwaj S, et al. End-to-end lung cancer screening with three-dimensional deep learning on low-dose chest computed tomography. Nat Med 2019;25(6):954–961. [Published correction appears in Nat Med 2019;25(8):1319.]
18. McKinney SM, Sieniek M, Godbole V, et al. International evaluation of an AI system for breast cancer screening. Nature 2020;577(7788):89–94. [Published correction appears in Nature 2020;586(7829):E19.]
19. Bulten W, Kartasalo K, Chen PHC, et al; PANDA challenge consortium. Artificial intelligence for diagnosis and Gleason grading of prostate cancer: the PANDA challenge. Nat Med 2022;28(1):154–163.
20. Ye X, Guo D, Ge J, et al. Comprehensive and clinically accurate head and neck cancer organs-at-risk delineation on a multi-institutional study. Nat Commun 2022;13(1):6137.
21. Collins GS, Reitsma JB, Altman DG, Moons KGM. Transparent Reporting of a multivariable prediction model for Individual Prognosis Or Diagnosis (TRIPOD). Ann Intern Med 2015;162(10):735–736.
22. Yu Q, Wang Y, Yan K, et al. Effective lymph nodes detection in CT scans using location debiased query selection and contrastive query representation in transformer. In: Leonardis A, Ricci E, Roth S, Russakovsky O, Sattler T, Varol G, editors. Computer Vision–ECCV 2024 European Conference on Computer Vision, 2024; 180–198.
23. Guo D, Ye X, Ge J, et al. DeepStationing: thoracic lymph node station parsing in CT scans using anatomical context encoding and key organ auto-search. In: de Bruijne M, Cattin PC, Cotin S, et al, editors. Medical Image Computing and Computer Assisted Intervention—MICCAI 2021. Springer, 2021; 3–12.
24. Huang SH, Chernock R, O'Sullivan B, Fakhry C. Assessment criteria and clinical implications of extranodal extension in head and neck cancer. Am Soc Clin Oncol Educ Book 2021;41:265–278.
25. DeLong ER, DeLong DM, Clarke-Pearson DL. Comparing the areas under two or more correlated receiver operating characteristic curves: a nonparametric approach. Biometrics 1988;44(3):837–845.
26. Chen Z, Yu Y, Liu S, et al. A deep learning and radiomics fusion model based on contrast-enhanced computer tomography improves preoperative identification of cervical lymph node metastasis of oral squamous cell carcinoma. Clin Oral Investig 2023;28(1):39.
27. Ariji Y, Sugita Y, Nagao T, et al. CT evaluation of extranodal extension of cervical lymph node metastases in patients with oral squamous cell carcinoma using deep learning classification. Oral Radiol 2020;36(2):148–155.
28. Lydiatt WM, Patel SG, O'Sullivan B, et al. Head and neck cancers—major changes in the American Joint Committee on Cancer eighth edition cancer staging manual. CA Cancer J Clin 2017;67(2):122–137.
29. Henson C, Abou-Foul AK, Yu E, et al. Criteria for the diagnosis of extranodal extension detected on radiological imaging in head and neck cancer: Head and Neck Cancer International Group consensus recommendations. Lancet Oncol 2024;25(7):e297–e307.
30. Faraji F, Aygun N, Coquia SF, et al. Computed tomography performance in predicting extranodal extension in HPV-positive oropharynx cancer. Laryngoscope 2020;130(6):1479–1486.