Research Article

Diagnosis of thyroid cartilage invasion by laryngeal and hypopharyngeal cancers based on CT with deep learning

Yuki Takano, MD[1,2]; Noriyuki Fujima, MD, PhD[2,3]; Junichi Nakagawa, MD, PhD[4]; Hiroki Dobashi, MD[2]; Yukie Shimizu, MD, PhD[2,3]; Motoma Kanaya, MD[1,2]; Satoshi Kano, MD, PhD[5]; Akihiro Homma, MD, PhD[5]; Kohsuke Kudo, MD, PhD[1,2,6,7].

[1] Department of Diagnostic Imaging, Graduate School of Medicine, Hokkaido University, N15-W7, Kita-Ku, Sapporo, Hokkaido 060-8638, Japan

[2] Department of Diagnostic and Interventional Radiology, Hokkaido University Hospital, N14-W5, Kita-Ku, Sapporo, Hokkaido 060-8648, Japan

[3] Global Center for Biomedical Science and Engineering, Faculty of Medicine, Hokkaido University, N14-W5, Kita-Ku, Sapporo, Hokkaido 060-8638, Japan

[4] Department of Radiology, Sapporo-Kosei General Hospital, N3-E8-5, Chuo-Ku, Sapporo, Hokkaido 060-0033, Japan

[5] Department of Otolaryngology-Head and Neck Surgery, Faculty of Medicine and Graduate School of Medicine, Hokkaido University, N15-W7, Kita ku, Sapporo 060-8638, Japan

[6] Department of Nuclear Medicine, Hokkaido University Hospital, N14-W5, Kita-Ku, Sapporo, Hokkaido 060-8648, Japan

[7] Medical AI Human Research and Development Center, Hokkaido University Hospital, N14-W5, Kita-Ku, Sapporo, Hokkaido 060-8648, Japan

Corresponding Author

Noriyuki Fujima

Department of Diagnostic and Interventional Radiology, Hokkaido University Hospital

N14-W5, Kita-Ku, Sapporo, Hokkaido 060-8648, Japan

Phone: +81-11-706-7779, Fax: +81-11-706-7408

E-mail: fujima@med.hokudai.ac.jp

Authors' contributions

YT conceived the study's design, performed the data acquisition, analysis and interpretation of data for the work, and drafted the manuscript. NF conceived the study's design, performed the interpretation of data for the work, the statistical analyses and manuscript editing. JN performed the interpretation of data for the work, methodology development, software development, and manuscript revising. HD, YS, and MK performed the interpretation of data for the work and manuscript revising. AH and SK carried out the data acquisition and manuscript revisions. KK participated in the study design and coordination and helped draft the manuscript. All authors read and approved the final manuscript.

Research Article

Diagnosis of thyroid cartilage invasion by laryngeal and hypopharyngeal cancers based on CT with deep learning

ABSTRACT

Objectives: To develop a convolutional neural network (CNN) model to diagnose thyroid cartilage invasion by laryngeal and hypopharyngeal cancers observed on computed tomography (CT) images and evaluate the model's diagnostic performance.

Methods: We retrospectively analyzed 91 cases of laryngeal or hypopharyngeal cancer treated surgically at our hospital during the period April 2010 through May 2023, and we divided the cases into datasets for training (n=61) and testing (n=30). We reviewed the CT images and pathological diagnoses in all cases to determine the invasion positive- or negative-status as a ground truth. We trained the new CNN model to classify thyroid cartilage invasion-positive or -negative status from the pre-treatment axial CT images by transfer learning from Residual Network 101 (ResNet101), using the training dataset. We then used the test dataset to evaluate the model's performance. Two radiologists, one with extensive head and neck imaging experience (senior reader) and the other with less experience (junior reader) reviewed the CT images of the test dataset to determine whether thyroid cartilage invasion was present.

Results: The following were obtained by the CNN model with the test dataset: area under the curve (AUC), 0.82; 90% accuracy, 80% sensitivity, and 95% specificity. The CNN model showed a significant difference in AUCs compared to the junior reader (p=0.035) but not the senior reader (p=0.61).

Conclusions: The CNN-based diagnostic model can be a useful supportive tool for the assessment of thyroid cartilage invasion in patients with laryngeal or hypopharyngeal cancer.

Keywords

head and neck, laryngeal and hypopharyngeal cancer, thyroid cartilage invasion, deep learning, convolutional neural network.

Introduction

For patients with laryngeal or hypopharyngeal cancer, preoperative staging plays a critical role in determining appropriate treatment plans and predicting the prognosis. In the staging of laryngeal and hypopharyngeal cancers, the extent of thyroid cartilage invasion is an important criterion for classifying tumors as stage T3 or T4a [1, 2]. For localized tumors at stage T3 or below, larynx-preserving treatments may be considered. However, if the cancer is diagnosed as T4a, more invasive treatments (e.g., total laryngectomy) are required [3–7]. An incorrect assessment of thyroid cartilage invasion can have serious consequences for both the quality of life and prognosis of the patient [5–9]. An accurate evaluation of thyroid cartilage invasion is thus necessary in cases of laryngeal or hypopharyngeal cancer.

Computed tomography (CT) and magnetic resonance imaging (MRI) have been widely used for diagnosing thyroid cartilage invasion [10–14]. Computed tomography has high spatial and temporal resolution with a short scan time, but non-calcified cartilage and tumors can have similar attenuation values, resulting in diagnostic accuracy that depends on the radiologist's expertise. MRI can detect conspicuous signal changes within the thyroid cartilage, but it is susceptible to motion artifacts. Although the utility of dual-energy CT (one of the more recent techniques) for diagnosing thyroid cartilage invasion has been described, the availability of this modality is somewhat limited because it is advanced rather than conventional. 'Radiomics' allows high-dimensional quantitative analyses of CT images [15–17], but the use of radiomics usually involves a manual process and a variety of parameter calculations including parameter extraction, selection and integration; its practical application thus remains challenging.

Deep learning (DL), an artificial intelligence method, has been applied to various types of imaging analysis, including in the field of head and neck radiology (e.g., tumor-progression evaluations and diagnoses) [18–20]. Deep learning differs from conventional machine learning approaches, including radiomics, in that it automatically extracts features from training data [18], and it is thus a useful method for classifying and quantifying images with complex and unknown features. The DL approach has the potential to detect thyroid cartilage invasion on CT images with high diagnostic performance. We conducted the present study to develop and evaluate a DL model that we designed to detect thyroid cartilage invasion with the use of neck CT images.

Materials and Methods

This retrospective study was approved by the institutional review board in our hospital, and the requirement for patients' written informed consent was waived.

Database and patient population

Based on their medical records, we selected the cases of 355 patients who were diagnosed as having laryngeal or hypopharyngeal cancer and treated at our institution during the period from April 2010 to May 2023. These 355 patients were further selected with the following inclusion criteria: (*i*) the patient had undergone a surgical resection of the cancer, with no presurgical treatment, (*ii*) a pathological diagnosis regarding whether thyroid cartilage invasion by the primary site was present/absent had been obtained, and (*iii*) the patient's pretreatment CT images reconstructed with a soft tissue kernel including the entire tumor site in its scanning range with a slice thickness <4 mm were available. Ultimately, 91 cases were deemed eligible for this study (Fig. 1). Furthermore, between June 2023 and February 2025, patients who met the previously described inclusion criteria and had accessible CT scans from outside facilities were included in the external validation dataset as a supplemental analysis; a total of ten patients were included.

CT images

The CT images of the total patient population had been obtained by various scanners from five vendors. Of the 91 cases, three cases had only non-contrast enhanced images, 53 cases had only contrast-enhanced images, and the other 35 cases had both non-contrast-enhanced and contrast-enhanced images. Of the 10 external validation cases, one had only non-contrast-enhanced images, and nine had only contrast-enhanced images. We used the axial reconstructed CT images for the evaluation. Other image parameters were as follows: acquisition slice thickness of 0.50–1.00 mm (median: 0.50 mm), reconstruction slice thickness of 1.00–3.75 mm (median: 2.00 mm), tilt of 0 degrees (except for one case at -5 degrees), field of view (FOV) of 240–500 mm (median: 400 mm), acquisition and reconstruction matrix of 512×512, and use of a soft tissue reconstruction kernel. The window level was set at 60 Hounsfield units (HU) with a width of 300 HU.

Ground truth determination for thyroid cartilage invasion

The reference standard was determined based on pathological reports and tumor specimens. Three radiologists with 2, 8, and 17 years of experience in head and neck radiology reviewed the pathological reports and specimens to correlate the tumor location between the CT and pathological findings in the cases that involved thyroid cartilage invasion. All available medical records were used as additional supportive information. We also determined whether the quality of the CT images was appropriate for interpretation, by using a Digital Imaging and Communication in Medicine (DICOM) viewer (XTREK, J-Mac System, Sapporo). The range of evaluation on axial sections was set from the inferior edge of the hyoid bone to the lower end of the thyroid cartilage. All CT images in the range of evaluation were judged as invasion-positive or -negative slices. Thereafter, each patient was classified as having invasion-positive or -negative status according to slice-based diagnosis. In addition, we further selected specific cases for the supplemental subgroup analysis as follows: (1) invasion-positive cases that showed invasion into nonossified cartilage, and (2) invasion-negative cases that had nonossified cartilage adjacent to the tumor extension.

Configuration of datasets for the training and test

We classified the 91 cases into four groups according to whether or not there was thyroid cartilage invasion from the laryngeal or hypopharyngeal cancer. We randomly split each group's cases into the training and test datasets at an approx. 7:3 ratio. Sixty-one cases (which included 18 invasion-positive cases) comprised the training dataset used to create the diagnostic model, and the other 30 cases (which included 10 invasion-positive cases) comprised the test dataset used to evaluate the performance of the established model (Fig. 1).

Data configuration

We first performed segmentation on all axial CT images. A rectangular region of interest (ROI) was manually drawn to encompass the thyroid cartilage. The superior-inferior extent of each ROI ranged from the inferior tip of the hyoid bone to the inferior tip of the thyroid cartilage, anterior-posteriorly from the skin to the anterior vertebral bodies and laterally to include both internal carotid arteries. Masking outside the thyroid cartilage of each ROI was performed based on the aspect ratios of the image's width and height. Each masking of lateral triangles was set with a base three-tenths of the rectangle's width and a height six-tenths of the rectangle's height, and masking of central triangle was set with an equal height and a width that was half of the rectangle.

For all datasets, we prepared two types of images: those with masking and those without masking. All CT images were adjusted to the window level of 60 HU and the window width 300 HU. Each processed image was then exported as a Joint Photographic Experts Group (JPEG) document.

The training dataset for the invasion-negative group included all extracted CT slices and consisted of 1,251 images from 43 cases (non-contrast only: n=2, contrast only: n=25, both non-contrast and contrast: n=16). In contrast, the training dataset for the invasion-positive group included only those specific slices that were identified as invasion-positive as described in the Ground Truth determination for thyroid cartilage invasion described above. This training dataset consisted of 153 images from 18 cases (contrast only: n=10, both non-contrast and contrast: n=8).

Before the training session, data augmentation techniques such as horizontal flipping, random rotation, and vertical and/or horizontal shifting were applied to each image to improve the model's robustness, generating five additional images for each invasion-negative original image and 25 additional images for each invasion-positive original image. As a result, the training data set consisted of 7,566 images for the invasion-negative group and 3,978 images for the invasion-positive group.

We used all of the extracted CT images of each case in both the invasion-positive and -negative groups for the test dataset, which was composed of 547 images from 20 invasion-negative cases (contrast only: n=12, both non-contrast and contrast: n=8) and

269 images (including 49 invasion-positive slices) from 10 invasion-positive cases (non-contrast only: 1, contrast only: n=6, both non-contrast and contrast: n=3). The test dataset was evaluated without the use of any data augmentation procedures. The external validation dataset was processed in the same manner as the test dataset described above.

Computer configuration

The model was established using the 64-bit version of the Windows operating system (Windows 10 Pro ver. 22H2) with an Intel Core i9 10940X 14core/28thread 3.30-GHz central processing unit (CPU), NVIDIA Quadro M2000 graphics processing unit (GPU) cards, and 256-GB (32GB×8) DDR4-2666 octa-channel memory for training and validation. The time required for each epoch was approx. 130 min. All image analyses were performed using MATLAB (R2023a, MathWorks, Natick, MA, USA) and Metavol software (https://www.metavol.org) [21].

Model development

We developed a DL model to classify the invasion-positive status and invasion-negative status of the axial CT images by using transfer learning from ResNet101 [22], which is a pre-trained convolutional neural network (CNN) algorithm devoted to image classification. ResNet101 consists of residual blocks in which the input image passes through convolution and activation functions toward the output. The input node has a bypass to the output node in each residual block. The residual block helps to avoid the degradation problem [22]. In the present study, the output of the last layer was connected to a fully connected layer with a sigmoid function to make the prediction. The main hyperparameters for the DL model were set as follows: the optimizer was adaptive moment (Adam), the learning rate was 0.00001, the number of epochs was 12, and the batch size was 64. We used 30% of the training data as the internal validation during the training.

Model test

We assessed the diagnostic performance of the DL model on the test dataset. First, we used the newly developed CNN model to determine whether each CT slice in the test dataset was classified as having thyroid cartilage invasion or not. Next, for each patient, we grouped the slices consecutively judged as invasion-positive by the CNN model and calculated the number of slices in each group as the number of consecutive invasion-positive slices. If a patient had multiple groups of invasion-positive slices, the group with the highest number of slices was considered the patient's consecutive invasion-positive slice count. If a patient had both non-contrast enhanced and contrast-enhanced images, the number of consecutive slices showing invasion-positive was compared between two series, and the larger number was selected as the output data for the case. Finally, we assessed the performance of the CNN model by adjusting the threshold for the number of consecutive positive slices in order to classify each patient as either having thyroid cartilage invasion or not having it. The steps of image processing and the DL analysis are depicted in Figure 2. The external validation dataset was also processed in a similar fashion.

Visual evaluation by radiologists

Two radiologists, i.e., a senior radiologist (SR) with extensive experience in head and neck imaging (15 years in radiology, board-certified) and a junior radiologist (JR) with 1 year of experience in radiology reviewed the CT images from the test dataset to determine whether there was thyroid cartilage invasion by the laryngeal or hypopharyngeal cancer. They evaluated all axial CT slices of the neck, including the thyroid cartilage, without segmented images.

Visual representation with gradient-weighted class activation mapping (Grad-CAM)

We used gradient-weighted class activation mapping Grad-CAM [23] to generate saliency maps that reflect the ROIs in the output from the test dataset. Grad-CAM creates a saliency map by using the gradient information flowing into the last convolutional layer of the trained CNN model. The saliency map visualizes the regions that are crucial for predicting the target concept, providing insights into which parts of the image played a significant role in the CNN model's decision-making process.

Statistical analyses

To evaluate both the developed CNN model and the accuracy of the two radiologists, we assessed the area under the receiver operating characteristic curve (AUC), sensitivity, specificity, accuracy, positive predictive value (PPV), and negative predictive value (NPV) with the test dataset. In the analysis of the CNN model, the cutoff value for a receiver operating characteristic (ROC) curve, derived from the number of consecutive positive slices, was set at the maximum Youden index. The comparison of AUCs and the calculation of 95% confidence intervals (CIs) for the AUCs were performed using the DeLong method [24]. We calculated the 95%CIs (except for those of the AUCs) by the Clopper-Pearson method. Group comparison (except for those of the AUCs) was performed by the Mann-Whitney U test or Fisher's exact test. The external validation dataset was also evaluated in the same way. In addition, as subgroup analyses, the performance of the developed CNN model on specific cases in the test dataset were assessed as follows: 1) only cases in which the target thyroid cartilage was identified as nonossified cartilage, 2) only cases with both contrast-enhanced and non-contrast-enhanced images, evaluating diagnostic performance using each type separately. All analyses were performed using BellCurve for Excel ver. 4.07 (https://bellcurve.jp/ex/) or R ver. 4.2.2. Statistical significance was defined as a p-value <0.05.

Results

The detailed demographics of the 91 patients are summarized in Table 1. All of the cases were classified as squamous cell carcinoma. In the following text, the CNN model trained using masking for each dataset will be referred to as the 'the CNN model with masking,' while the model without masking will be referred to simply as the 'the CNN model without masking.' The diagnostic performance of the two CNN models and the two radiologists was as follows: AUC 0.82 (95%CI: 0.63–1.00), accuracy 90% (27/30, 73%–98%), sensitivity 80% (8/10, 43%–98%), specificity 95% (19/20, 75%–100%), PPV 89% (8/9, 52%–100%), NPV 90% (19/21, 70%–99%) for the CNN model with masking when the number of consecutive positive slices for the threshold was four; AUC 0.72 (0.51–0.93), accuracy 73% (22/30, 54%–88%), sensitivity 80% (8/10, 43%–98%), specificity 70% (14/20, 46%–88%), PPV 57% (8/14, 29%–82%), NPV 88% (14/16, 62%–98%) for the CNN model without masking when the number of consecutive positive slices for the threshold was three; AUC 0.75 (0.58–0.72), accuracy 77% (23/30, 57%–90%), sensitivity 60% (6/10, 26%–88%), specificity 85% (17/20, 62%–97%), PPV 67% (6/9, 30%–93%), NPV 81% (17/21, 58%–95%) for the SR; and AUC 0.53 (0.33–0.72), accuracy 50% (15/30, 31%–69%), sensitivity 40% (4/10, 12%–74%), specificity 55% (11/20, 32%–77%), PPV 31% (4/13, 9%–61%), and NPV 65% (11/17, 38%–86%) for the JR. The AUCs are listed in Table 2 and illustrated in Figure 3. The diagnostic performance of the CNN models and the radiologists is summarized in Table 2.

The AUC of the CNN model with masking for the detection of thyroid cartilage invasion was significantly larger than those of the CNN model without masking (p=0.019) and the JR (p=0.035), but there was no significant difference between the CNN model with masking and the SR (p=0.61). The AUC of the CNN model without masking did not show a significant difference compared to that of the SR (p=0.84) or JR (p=0.19).

We generated saliency maps for CT images in the test dataset. Representative saliency maps for the models with and without masking are shown in Figure 4. Compared to the CNN model without masking, the CNN model with masking showed greater saliency in the region of the thyroid cartilage.

In the analysis of the external validation datasets, the diagnostic accuracy of the CNN model with masking was 80% (8/10, 48–100%). All other information regarding the external validation datasets is provided in Supplemental Tables 1 and 2.

As results of a subgroup analysis targeting nonossified thyroid cartilage, a total of five cases in the test dataset were included, consisting of three invasion-positive cases which showed invasion into nonossified cartilage, and two invasion-negative cases which showed nonossified cartilage adjacent to the tumor extension. The overall diagnostic accuracy was 100% (5/5, 48–100%) for the CNN models (both with and without masking), 40% (2/5, 5–85%) for SR, and 60% (3/5, 15–95%) for JR. As results of the other subgroup analysis, 11 cases from the test dataset with both non-contrast enhanced and contrast-enhanced images available were included. The overall diagnostic accuracy was 91% (10/11, 58–100%) for the CNN model with masking, using either contrast-enhanced or non-contrast-enhanced images, with both image types demonstrating identical performance. Details of the subgroup analyses are presented in Supplemental Tables 3 and 4.

Discussion

We developed and evaluated a CNN model to predict tumor invasion of the thyroid cartilage that uses axial neck CT images of patients with laryngeal or hypopharyngeal cancer. To the best of our knowledge, this is the first study to create and validate a CNN-based model for determining the presence of thyroid cartilage invasion. The best-performing model achieved an AUC of 0.82 and 90% accuracy. For our performance evaluation using ROC curves, the threshold was set based on the number of consecutive positive slices to convert individual CT image assessments into case-level diagnoses. Similar methods have been used to evaluate a diagnostic performance against radiologists' readings [25, 26]. In the reading process, it is considered reasonable to estimate thyroid cartilage invasion not only by changes in the thyroid cartilage (mainly osteolytic changes) but also by considering the surrounding tumor conditions and changes across contiguous slices. In the present study, the saliency map showed strong saliency over the thyroid cartilage. In particular, when images with masking were used, the saliency map indicated a reduction in features other than the thyroid cartilage, which contributed to improved model performance. The entire image contains a large amount of extraneous information, and thus masking the necessary parts is considered crucial for efficiently building a highly accurate model. We used a relatively simple linear method for the masking in this study. Other investigations have employed a CNN model for tissue segmentation in the neck [27–29], which could potentially change the performance evaluation if a more accurate extraction of only the thyroid cartilage is achieved. Furthermore, a CNN model based on three-dimensional (3D) images could improve diagnostic performance by preserving volumetric information, however the amount of calculation needed would be large, making it difficult to use in practice. In contrast, 2.5-dimensional (2.5D) approaches, which combine consecutive slices to simulate 3D features, are more practical and could improve diagnostic accuracy [30,31]. The 2.5D method could also help automate the evaluation process based on the number of consecutive positive slices, as used in the current study. Additionally, we analyzed images using a fixed window level and width. Mostly similar settings have been employed in the previous deep learning study assessing tumor invasion into bone tissue [25], likely because they enable visualization of both soft tissue and bone to some extent. Therefore, we believe the settings used here are acceptable. Ideally, future studies should use native DICOM files to avoid information loss due to the conversion from DICOM to JPEG format, although using native DICOM files would increase data volume. In the subgroup analyses of the current study, the CNN model demonstrated high diagnostic performance even in nonossified cartilage and achieved high accuracy with both non-contrast-enhanced and contrast-enhanced images. These findings may suggest the model's high applicability; however, the number of cases included in these subgroup analyses was very limited, and further validation is needed.

The novel CNN model described herein offers several benefits. It demonstrated higher diagnostic accuracy than that of a junior radiologist and comparable accuracy to that of a highly experienced radiologist, which suggests that this CNN model could enable high-precision diagnoses of thyroid cartilage invasion in CT images, even in hospitals without experienced radiologists. Although CT has high spatial and temporal resolution with short scan times, non-calcified cartilage and tumors can have similar attenuation values, which results in relatively low sensitivity (67%) for detecting thyroid cartilage invasion [14]; this might cause the difficulty in achieving an accurate diagnosis except by an expert who is highly specialized in head and neck radiology.

The sensitivity for detecting thyroid cartilage invasion using MRI is reported to be 75%–100% [10, 12, 32, 33], which is higher than that of conventional CT. However, the laryngeal or hypopharyngeal region is significantly affected by motion artifacts from breathing and swallowing. Particularly MRI techniques are considered more sensitive to motion artifacts than CT due to its long acquisition time; this tendency was reported prominent in locally advanced cases involving invasion of the thyroid cartilage [32]. Moreover, edema or inflammatory changes in cartilage can present similarly of MR imaging findings to local tumor invasion, resulting in relatively low specificity to detect thyroid cartilage invasion [10, 12, 14, 32, 33]. In contrast, recent advancements in MRI for the head and neck region, especially the deep learning reconstruction technique, aim to reduce scan time and improve image contrast by the effective denoising [34, 35]. However, these technologies still require time before they can be widely used in clinical practice. In the meantime, conventional CT remains the primary imaging modality for staging laryngeal and hypopharyngeal cancers, due to its accessibility and widespread availability. Although the sensitivity afforded by dual-energy CT (DECT) as a representative advanced imaging technique has been reported as 86%–89% [32, 36], the availability of DECT for daily routine clinical practice is limited at many institutions. In this situation, conventional CT remains a widely used and highly accessible technique for staging laryngeal and hypopharyngeal cancers [13, 37], and it enables diagnoses with the use of commonly available CT images without requiring dedicated imaging modality or techniques such as MRI or DECT.

Radiomics enables the quantitative analysis of images by clinicians without a high level of experience in image interpretation of the head and neck, with 80% sensitivity reported with the use of a CT-based radiomics signature [15]. However, the application of radiomics may encounter difficulties in extracting and selecting features when dealing with numerous and complex features, because of the differences among radiomics analytical methodologies [18–20]. The extra task of manually setting the ROI on the lesion also often arises in radiomics analyses. Unlike such radiomics analyses, the CNN-based method eliminates the need for complex lesion ROI settings, extensive feature extraction, and integration with complicated calculation processes to achieve high diagnostic accuracy. Considering these advantages, a CNN-based analysis may offer greater utility compared to other analytical methodologies.

This study has several limitations. Due to its retrospective design and a patient population drawn from a single institution, the sample size is small (n=91), although we addressed it as much as possible through data augmentation. Accordingly, the results presented herein should be regarded as preliminary: further validation with large-scale, prospectively acquired test datasets is needed to address this limitation. The need to manually set rectangular ROIs could introduce observer-related bias in clinical settings. Although a head/neck investigation reported that the diagnostic performance is not significantly affected by ROI delineation [26], a deep learning-based automated cartilage segmentation could further improve the robustness and reproducibility of the pipeline, given the recent advances in auto-segmentation technology [27–29]. In addition, for cases with both contrast-enhanced and non-contrast CTs, the scan with the highest number of consecutive positive slices was selected for case-based diagnosis. This selection strategy may introduce potential bias that leads to overestimation of diagnostic performance. In future studies, employing ensemble approaches or fusion methods that integrate multiple modalities could contribute to a more balanced and accurate analysis. Moreover, although we used the pathological diagnosis as the reference standard, there could have been a time gap between the timing of the patients' CT scans and the period when the surgeries were performed, which means that thyroid cartilage invasion might not have been present at the time of CT imaging even if it was pathologically confirmed later. Finally, the CT slice thickness was not strictly standardized. The CT slice thickness was limited to <4 mm, and adjustments may be needed to apply this model to patients imaged with 5-mm slices. In addition, our dataset included varying CT slice thicknesses (1.0–3.8 mm). Slice thickness is known to influence diagnostic performance in the assessment of thyroid cartilage invasion [13]. Further analysis of the relationship between model performance and slice thickness variation will be necessary.

Conclusions

We developed a CNN model that classifies whether thyroid cartilage invasion is present or absent, based on features present in neck CT images of laryngeal and hypopharyngeal cancers. This artificial intelligence model has the potential to be an effective diagnostic support tool, particularly in medical institutions that lack specialized expertise in head and neck imaging.

Abbreviations

2.5D: 2.5-dimensional

3D: three-dimensional

AUC: area under the curve

CNN: convolutional neural network

CT: computed tomography

DECT: dual-energy computed tomography

DICOM: Digital Imaging and Communication in Medicine

DL: deep learning

Grad-CAM: gradient-weighted class activation map

HU: Hounsfield units

JPEG: Joint Photographic Experts Group

JR: junior reader

MRI: magnetic resonance imaging

NPV: negative predictive value

PPV: positive predictive value

ROC: receiver operating characteristic

ROI: region of interest

SR: senior reader

Table and Figure captions

Table 1. Characteristics of the patients with laryngeal or hypopharyngeal cancer.

Table 2. The AUC and model performance for the test dataset.

Fig. 1. The flowchart of patient selection and dataset configuration.

Fig. 2. The workflow of the development and evaluation of the CNN models. Conv: convolution, CNN: convolutional neural network, ROC: receiver operating characteristic curve.

Fig. 3. The receiver operating characteristic curves for the test dataset.

Fig. 4. The saliency maps for the test dataset. *Left column:* The original images from the test dataset, with *arrows* pointing to the regions where thyroid cartilage invasion was detected. *Middle and right columns:* The saliency maps, both of which show stronger saliency in the thyroid cartilage when using the CNN model with masking.

References

1. A.A. Forastiere, N. Ismaila, J.S. Lewin, C.A. Nathan, D.J. Adelstein, A. Eisbruch, G. Fass, S.G. Fisher, S.A. Laurie, Q.T. Le, B. O'Malley, W.M. Mendenhall, S. Patel, D.G. Pfister, A.F. Provenzano, R. Weber, G.S. Weinstein, G.T. Wolf, Use of Larynx-Preservation Strategies in the Treatment of Laryngeal Cancer: American Society of Clinical Oncology Clinical Practice Guideline Update, J. Clin. Oncol. 36 (2018) 1143–1169. https://doi.org/10.1200/jco.2017.75.7385
2. R. Hermans, Staging of laryngeal and hypopharyngeal cancer: Value of imaging studies, Eur. Radiol. 16 (2006) 2386–2400. https://doi.org/10.1007/s00330-006-0301-7
3. E. Katsoulakis, N. Riaz, M. Hu, L. Morris, E. Sherman, S. McBride, N. Lee, Hypopharyngeal squamous cell carcinoma: Three-dimensional or intensity-modulated radiotherapy? A single institution's experience, Laryngoscope. 126 (2016) 620–626. https://doi.org/10.1002/lary.25509
4. T. Jin, X. Li, D. Lei, D. Liu, Q. Yang, G. Li, X. Pan, Preservation of laryngeal function improves outcomes of patients with hypopharyngeal carcinoma, Eur. Arch. Otorhinolaryngol. 272 (2015) 1785–1791. https://doi.org/10.1007/s00405-014-3115-2
5. A.A. Forastiere, H. Goepfert, M. Maor, T.F. Pajak, R. Weber, W. Morrison, B. Glisson, A. Trotti, J.A. Ridge, C. Chao, G. Peters, D.J. Lee, A. Leaf, J. Ensley, J. Cooper, Concurrent chemotherapy and radiotherapy for organ preservation in advanced laryngeal cancer, N. Engl. J. Med. 349 (2003) 2091–2098. https://doi.org/10.1056/nejmoa031317
6. M.Y. Lee, M. Belfiglio, J. Zeng, C.W. Fleming, S. Koyfman, N.P. Joshi, E. Lamarre,

B. Prendes, J. Scharpf, R.R. Lorenz, N.M. Woody, D.J. Adelstein, J.L. Geiger, D.J. Chute, J.A. Ku, Primary total laryngectomy versus organ preservation for locally advanced T3/T4a laryngeal cancer, Laryngoscope. 133 (2023) 1122–1131. https://doi.org/10.1002/lary.30254

7. Y.S. Choi, S.G. Park, E.K. Song, S.H. Cho, M.R. Park, K.U Park, K.H. Lee, I.C. Song, H.J. Lee, D.Y. Jo, S. Kim, H.J. Yun, Comparison of the therapeutic effects of total laryngectomy and a larynx-preservation approach in patients with T4a laryngeal cancer and thyroid cartilage invasion: A multicenter retrospective review, Head Neck. 38 (2016) 1271–1277. https://doi.org/10.1002/hed.24438

8. A. Tawfik, A.M.A. El-Fattah, A. Hassan, F.A. Helal, H.A. Ebada, Discrepancy between clinical and pathological staging of laryngeal carcinoma: A dilemma to be solved, Eur. Arch. Otorhinolaryngol. 281 (2024) 2507–2513. https://doi.org/10.1007/s00405-024-08506-2

9. B. Li, M. Bobinski, R. Gandour-Edwards, D.G. Farwell, A.M. Chen, Overstaging of cartilage invasion by multidetector CT scan for laryngeal cancer and its potential effect on the use of organ preservation with chemoradiation, Br. J. Radiol. 84 (2011) 64–69. https://doi.org/10.1259/bjr/66700901

10. M. Becker, P. Zbären, J.W. Casselman, R. Kohler, P. Dulguerov, C.D. Becker, Neoplastic invasion of laryngeal cartilage: Reassessment of criteria for diagnosis at MR imaging, Radiology. 249 (2008) 551–559. https://doi.org/10.1148/radiol.2492072183

11. M. Becker, P. Zbären, J. Delavelle, A.M. Kurt, C. Egger, D.A. Rüfenacht, F. Terrier, Neoplastic invasion of the laryngeal cartilage: Reassessment of criteria for diagnosis at CT, Radiology. 203 (1997) 521–532.

https://doi.org/10.1148/radiology.203.2.9114116

12. M. Becker, P. Zbären, H. Laeng, C. Stoupis, B. Porcellini, P. Vock, Neoplastic invasion of the laryngeal cartilage: Comparison of MR imaging and CT with histopathologic correlation, Radiology. 194 (1995) 661–669. https://doi.org/10.1148/radiology.194.3.7862960
13. E. Chau, K.S. Traylor, B.F. Branstetter, CT of the larynx: Is an additional high-resolution acquisition necessary for diagnostic accuracy?, AJNR Am. J. Neuroradiol. 44 (2023) 1421–1424. https://doi.org/10.3174/ajnr.a8048
14. S.J. Cho, J.H. Lee, C.H. Suh, J.Y. Kim, D. Kim, J.B. Lee, M.K. Lee, S.R. Chung, Y.J. Choi, J.H. Baek, Comparison of diagnostic performance between CT and MRI for detection of cartilage invasion for primary tumor staging in patients with laryngo-hypopharyngeal cancer: A systematic review and meta-analysis, Eur. Radiol. 30 (2020) 3803–3812. https://doi.org/10.1007/s00330-020-06718-8
15. R. Guo, J. Guo, L. Zhang, X. Qu, S. Dai, R. Peng, V.F.H. Chong, J. Xian, CT-based radiomics features in the prediction of thyroid cartilage invasion from laryngeal and hypopharyngeal squamous cell carcinoma, Cancer Imaging. 20 (2020) 81. https://doi.org/10.1186/s40644-020-00359-2
16. Q. Liu, S. Liu, Y. Mao, X. Kang, M. Yu, G. Chen, Machine learning model to preoperatively predict T2/T3 staging of laryngeal and hypopharyngeal cancer based on the CT radiomic signature, Eur. Radiol. 34 (2024) 5349–5359. https://doi.org/10.1007/s00330-023-10557-8
17. C.M. Chiesa-Estomba, M. Mayo-Yanez, O. Guntinas-Lichius, V. Vander-Poorten, R.P. Takes, R. de Bree, G.B. Halmos, N.F. Saba, S. Nuyts, A. Ferlito, Radiomics in hypopharyngeal cancer management: A state-of-the-art review, Biomedicines. 11

(2023) 805. https://doi.org/10.3390/biomedicines11030805

18. X. Chen, Q. Yu, J. Peng, Z. He, Q. Li, Y. Ning, J. Gu, F. Lv, H. Jiang, K. Xie, A combined model integrating radiomics and deep learning based on contrast-enhanced CT for preoperative staging of laryngeal carcinoma, Acad. Radiol. 30 (2023) 3022–3031. https://doi.org/10.1016/j.acra.2023.06.029

19. H. Tomita, T. Kobayashi, E. Takaya, S. Mishiro, D. Hirahara, A. Fujikawa, Y. Kurihara, H. Mimura, Y. Kobayashi, Deep learning approach of diffusion-weighted imaging as an outcome predictor in laryngeal and hypopharyngeal cancer patients with radiotherapy-related curative treatment: A preliminary study, Eur. Radiol. 32 (2022) 5353–5361. https://doi.org/10.1007/s00330-022-08630-9

20. X. Xu, L. Xi, L. Wei, L. Wu, Y. Xu, B. Liu, B. Li, K. Liu, G. Hou, H. Lin, Z. Shao, K. Su, Z. Shang, Deep learning assisted contrast-enhanced CT-based diagnosis of cervical lymph node metastasis of oral cancer: A retrospective study of 1466 cases, Eur. Radiol. 33 (2023) 4303–4312. https://doi.org/10.1007/s00330-022-09355-5

21. K. Hirata, K. Kobayashi, K.P. Wong, O. Manabe, A. Surmak, N. Tamaki, S.C. Huang, 2014. A semi-automated technique determining the liver standardized uptake value reference for tumor delineation in FDG PET-CT. PLoS. One. 9, e105682. https://doi.org/10.1371/journal.pone.0105682

22. H. Kaiming, Z. Xiangyu, R. Shaoqing, S. Jian, 2016. Deep residual learning for image recognition. 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR) 1:770. Las Vegas, NV, USA. https://doi.org/10.1109/CVPR.2016.90

23. R.R. Selvaraju, M. Cogswell, A. Das, R. Vedantam, D. Parikh, D. Batra, 2017. Grad-CAM: Visual explanations from deep networks via gradient-based localization.

2017 IEEE International Conference on Computer Vision (ICCV), 618–626. https://doi.org/10.1109/ICCV.2017.74

24. E.R. DeLong, D.M. DeLong, D.L. Clarke-Pearson, Comparing the areas under two or more correlated receiver operating characteristic curves: A nonparametric approach, Biometrics. 44 (1988) 837–845. https://doi.org/10.2307/2531595

25. J. Nakagawa, N. Fujima, K. Hirata, M. Tang, S. Tsuneta, J. Suzuki, T. Harada, Y. Ikebe, A. Homma, S. Kano, K. Minowa, K. Kudo, Utility of the deep learning technique for the diagnosis of orbital invasion on CT in patients with a nasal or sinonasal tumor, Cancer Imaging. 22 (2022) 52. https://doi.org/10.1186/s40644-022-00492-0

26. J. Nakagawa, N. Fujima, K. Hirata, T. Harada, N. Wakabayashi, Y. Takano, A. Homma, S. Kano, K. Minowa, K. Kudo, Diagnosis of skull-base invasion by nasopharyngeal tumors on CT with a deep-learning approach, Jpn. J. Radiol. 42 (2024) 450–459. https://doi.org/10.1007/s11604-023-01527-7

27. D. Rao, P. K, R. Singh, V. J, Automated segmentation of the larynx on computed tomography images: A review. Biomed. Eng. Lett. 12 (2022) 175–183. https://doi.org/10.1007/s13534-022-00221-3

28. Y.C. Lin, G. Lin, S. Pandey, C.H. Yeh, J.J. Wang, C.Y. Lin, T.Y. Ho, S.F. Ko, S.H. Ng, Fully automated segmentation and radiomics feature extraction of hypopharyngeal cancer on MRI using deep learning, Eur. Radiol. 33 (2023) 6548–6556. https://doi.org/10.1007/s00330-023-09827-2

29. Y. Choi, J. Bang, S.Y. Kim, M. Seo, J. Jang, Deep learning-based multimodal segmentation of oropharyngeal squamous cell carcinoma on CT and MRI using self-configuring nnU-Net, Eur. Radiol. 34 (2024) 5389–5400.

https://doi.org/10.1007/s00330-024-10585-y

30. J. Zhu, L. Zou, X. Xie, R. Xu, Y. Tian, B. Zhang, 2.5D deep learning based on multi-parameter MRI to differentiate primary lung cancer pathological subtypes in patients with brain metastases, Eur. J. Radiol. 180 (2024) 111712. https://doi.org/10.1016/j.ejrad.2024.111712

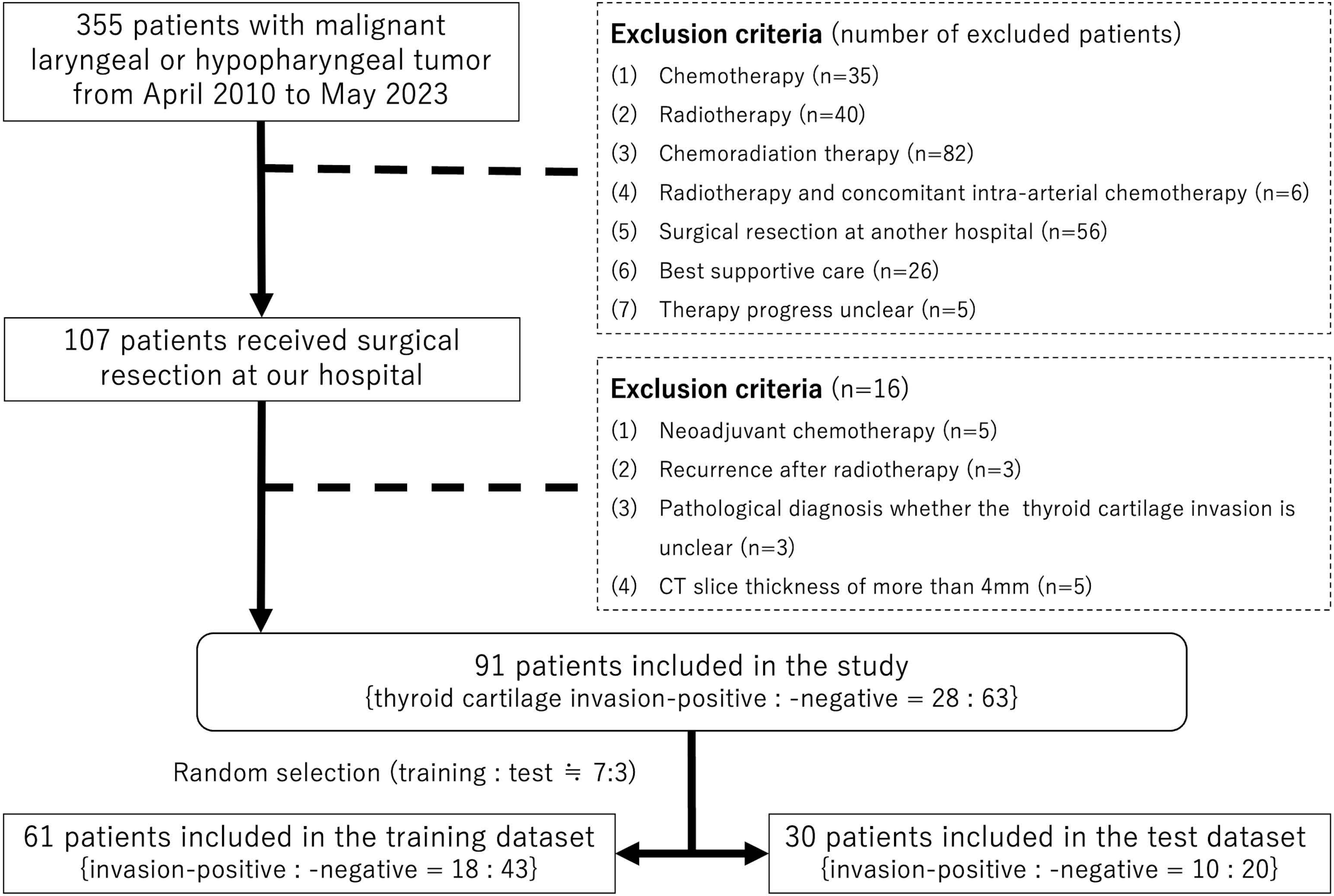
31. A.L.G. Saint-Esteven, M.E. Bogowicz, E. Konukoglu, O. Riesterer, P. Balermpas, M. Guckenberger, S. Tanadini-Lang, J.E. van Timmeren, A 2.5D convolutional neural network for HPV prediction in advanced oropharyngeal cancer. Comput. Biol. Med. 142 (2022) 105215. https://doi.org/10.1016/j.compbiomed.2022.105215

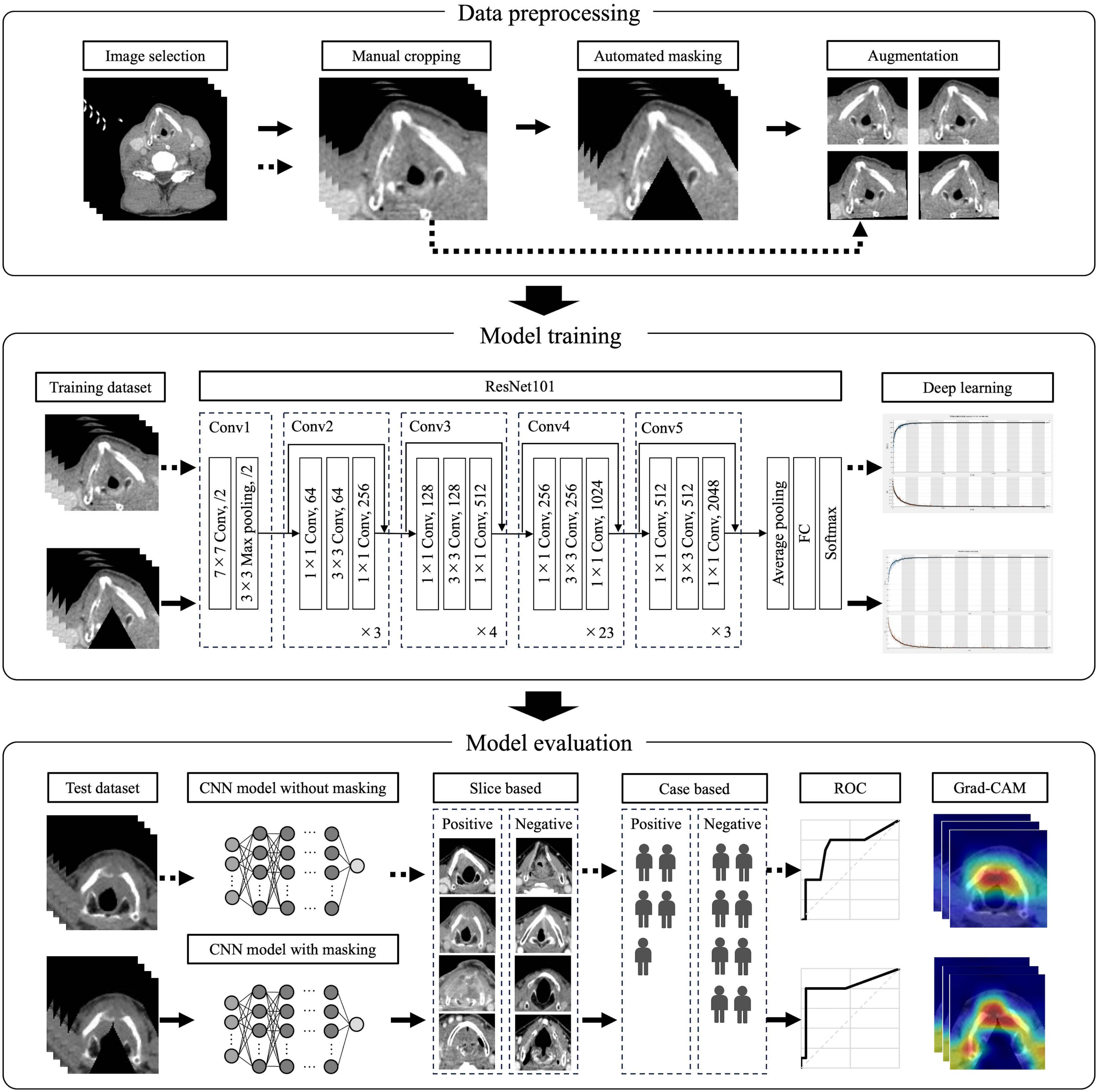
32. H. Kuno, K. Sakamaki, S. Fujii, K. Sekiya, K. Otani, R. Hayashi, T. Yamanaka, O. Sakai, M. Kusumoto, Comparison of MR imaging and dual-energy CT for the evaluation of cartilage invasion by laryngeal and hypopharyngeal squamous cell carcinoma, AJNR Am. J. Neuroradiol. 39 (2018) 524–531. https://doi.org/10.3174/ajnr.a5530

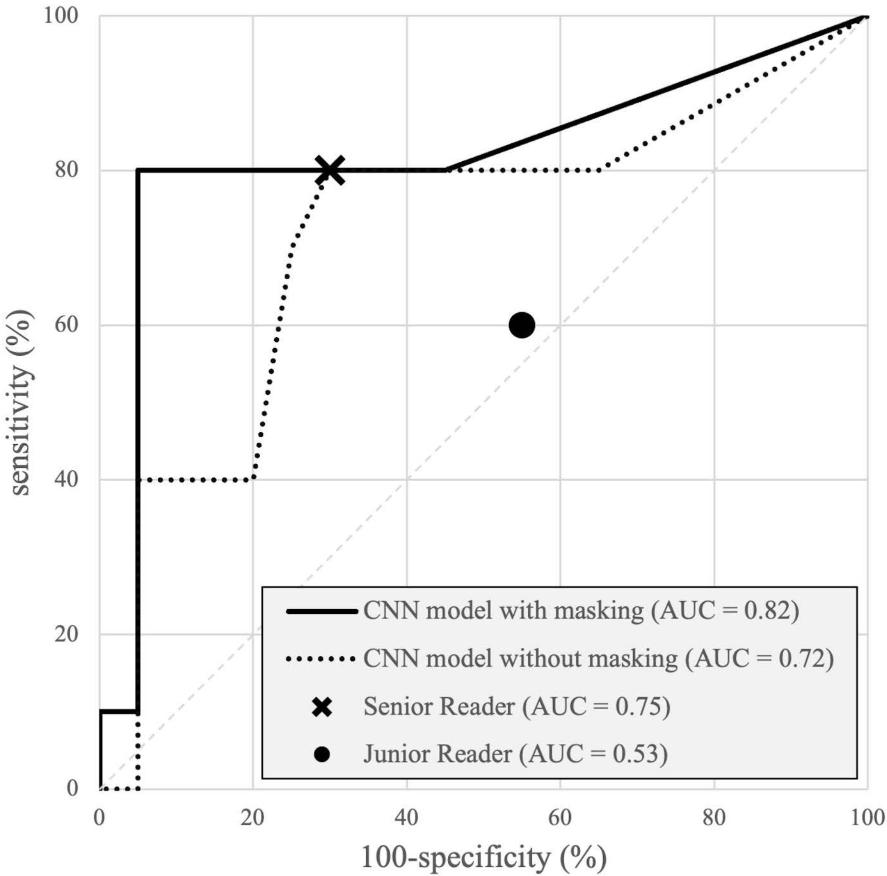
33. C.J. Park, J.H. Kim, S.S. Ahn, S.K. Lee, Y.W. Koh, J. Kim, Preoperative MRI evaluation of thyroid cartilage invasion in patients with laryngohypopharyngeal cancer: Comparison of contrast-enhanced 2D spin-echo and 3D T1-weighted radial gradient recalled-echo techniques, AJNR Am. J. Neuroradiol. 42 (2021) 1690–1694. https://doi.org/10.3174/ajnr.a7213

34. N. Fujima, K. Kamagata, D. Ueda, S. Fujita, Y. Fushimi, M. Yanagawa, R. Ito, T. Tsuboyama, M. Kawamura, T. Nakaura, A. Yamada, T. Nozaki, T. Fujioka, Y. Matsui, K. Hirata, R. Tatsugami, S. Naganawa, Current State of Artificial Intelligence in Clinical Applications for Head and Neck MR Imaging, Magn. Reson. Med. Sci. 22 (2023) 401-414. https://doi.org/10.2463/mrms.rev.2023-0047

35. H. Liu, D. Deng, W. Zeng, Y. Huang, C. Zheng, X. Li, H. Li, C. Xie, H. He, G. Xu, AI-assisted compressed sensing and parallel imaging sequences for MRI of patients with nasopharyngeal carcinoma: comparison of their capabilities in terms of examination time and image quality, Eur. Radiol. 33 (2023) 7686-7696. https://doi.org/10.1007/s00330-023-09742-6
36. H. Kuno, H. Onaya, R. Iwata, T. Kobayashi, S. Fujii, R. Hayashi, K. Otani, H. Ojiri, T. Yamanaka, M. Satake, Evaluation of cartilage invasion by laryngeal and hypopharyngeal squamous cell carcinoma with dual-energy CT, Radiology. 265 (2012) 488–496. https://doi.org/10.1148/radiol.12111719
37. H. Kuno, H. Onaya, S. Fujii, H. Ojiri, K. Otani, M. Satake, Primary staging of laryngeal and hypopharyngeal cancer: CT, MR imaging and dual-energy CT, Eur. J. Radiol. 83 (2014) e23–e35. https://doi.org/10.1016/j.ejrad.2013.10.022


355 patients with malignant laryngeal or hypopharyngeal tumor from April 2010 to May 2023
Exclusion criteria (number of excluded patients)
(1) Chemotherapy (n=35)
(2) Radiotherapy (n=40)
(3) Chemoradiation therapy (n=82)
(4) Radiotherapy and concomitant intra-arterial chemotherapy (n=6)
(5) Surgical resection at another hospital (n=56)
(6) Best supportive care (n=26)
(7) Therapy progress unclear (n=5)
107 patients received surgical resection at our hospital
Exclusion criteria (n=16)
(1) Neoadjuvant chemotherapy (n=5)
(2) Recurrence after radiotherapy (n=3)
(3) Pathological diagnosis whether the thyroid cartilage invasion is unclear (n=3)
(4) CT slice thickness of more than 4mm (n=5)
91 patients included in the study
{thyroid cartilage invasion-positive : -negative = 28 : 63}
Random selection (training : test ≒ 7:3)
61 patients included in the training dataset
{invasion-positive : -negative = 18 : 43}
30 patients included in the test dataset
{invasion-positive : -negative = 10 : 20}


Data preprocessing
Image selection
Manual cropping
Automated masking
Augmentation
Model training
Training dataset
ResNet101
Deep learning
Conv1
7 × 7 Conv, /2
3 × 3 Max pooling, /2
Conv2
1 × 1 Conv, 64
3 × 3 Conv, 64
1 × 1 Conv, 256
× 3
Conv3
1 × 1 Conv, 128
3 × 3 Conv, 128
1 × 1 Conv, 512
× 4
Conv4
1 × 1 Conv, 256
3 × 3 Conv, 256
1 × 1 Conv, 1024
× 23
Conv5
1 × 1 Conv, 512
3 × 3 Conv, 512
1 × 1 Conv, 2048
× 3
Average pooling
FC
Softmax
Model evaluation
Test dataset
CNN model without masking
CNN model with masking
Slice based
Positive
Negative
Case based
Positive
Negative
ROC
Grad-CAM



100
80
60
40
20
0
sensitivity (%)
0
20
40
60
80
100
100-specificity (%)
CNN model with masking (AUC = 0.82)
CNN model without masking (AUC = 0.72)
Senior Reader (AUC = 0.75)
Junior Reader (AUC = 0.53)

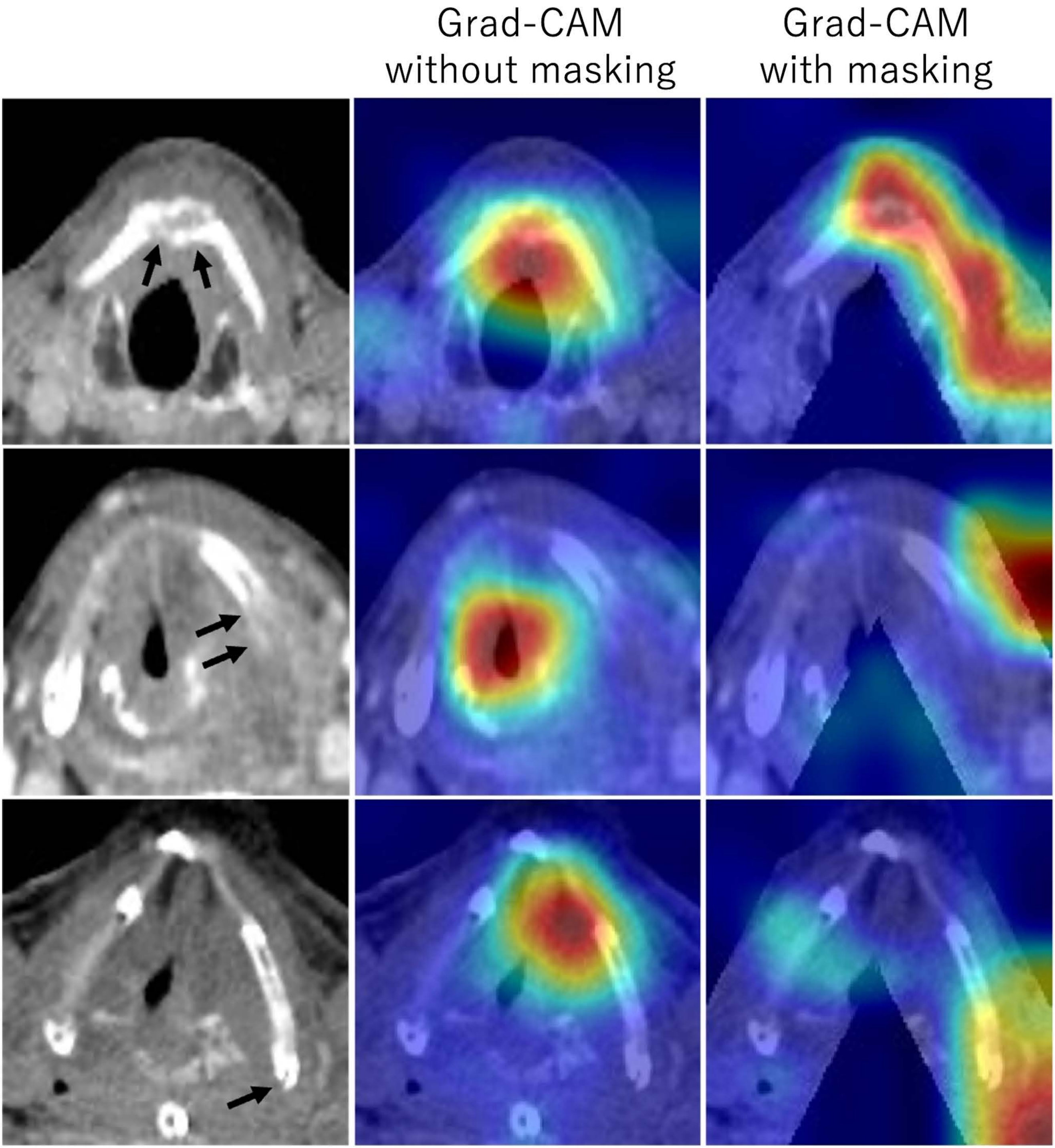

Grad-CAM
without masking
Grad-CAM
with masking

Table 1. Dataset demographics

	Total	Training dataset	Test dataset	p value
Total number of patients	91	61	30	
Sex[a]				1.00(1.00)[b]
Male	85(26)	57(17)	28(9)	
Female	6(2)	4(1)	2(1)	
Age (mean ± SD, years)[a]	71 ± 7	71 ± 7	72 ± 7	0.88(0.83)[c]
Period between operations (mean ± SD, days)[a]	27 ± 21	28 ± 22	24 ± 18	0.47(0.31)[c]
Primary site[a]				0.74(0.72)[b]
Supraglottis	23(4)	14(3)	9(1)	
Glottis	25(12)	18(8)	7(4)	
Subglottis	2(1)	1(0)	1(1)	
Hypopharynx	41(11)	28(7)	13(4)	
Histopathologic diagnosis				0.81[b]
With thyroid cartilage invasion	28	18	10	
Without thyroid cartilage invasion	63	43	20	
Number of CT slices				< 0.001*[b]
With thyroid cartilage invasion	202	153	49	
Without thyroid cartilage invasion	2028	1261	767	
T stage[a]				0.86(1.00)[b]
T1	3(0)	2(0)	1(0)	
T2	15(0)	9(0)	6(0)	
T3	45(7)	32(5)	13(2)	
T4	28(21)	18(13)	10(8)	
N stage[a]				0.82(1.00)[b]
N0	38(12)	25(7)	13(5)	
N1	4(1)	3(1)	1(0)	
N2	32(11)	23(7)	9(4)	
N3	17(4)	10(3)	7(1)	

CT: computed tomography, SD standard deviation.

[a] Numbers in parentheses are with thyroid cartilage invasion.

[b] Fisher's exact test.

[c] Mann-Whitney U test.

* indicates a significant difference

Table 2. Diagnostic performances[a]

	CNN model with masking	CNN model without masking	Senior Reader	Junior Reader
AUC[b]	0.82(0.63-1.00)	0.72(0.51-0.93)	0.75(0.58-0.92)	0.53(0.33-0.72)
Accuracy(%)[c]	90(73-98)	73(54-88)	77(57-90)	50(31-69)
Sensitivity(%)[c]	80(43-98)	80(43-98)	60(26-88)	40(12-74)
Specificity(%)[c]	95(75-100)	70(46-88)	85(62-97)	55(32-77)
PPV(%)[c]	89(52-100)	57(29-82)	67(30-93)	31(9-61)
NPV(%)[c]	90(70-99)	88(62-98)	81(58-95)	65(38-86)
p value[d]		0.019*	0.61	0.035*

AUC: area under the curve, CI: confidence intervals, CNN: convolutional neural network, NPV: negative predictive value, PPV: positive predictive value.

For the CNN model without masking, the value was taken when the cutoff for the number of consecutive slices is set to ≥3. Similarly, for the CNN model with masking, the cutoff was set to ≥4.

[a] Numbers in parentheses are 95% CI.

[b] Numbers of 95% CI are determined with the DeLong method.

[c] Numbers of 95% CI are determined with the Clopper-Pearson method.

[d] DeLong test for differences in AUC compared to CNN model with masking assessment.

* indicates a significant difference.