International Journal of Radiation Oncology biology • physics

www.redjournal.org

Physics Contribution

# Attention Guided Lymph Node Malignancy Prediction in Head and Neck Cancer

**Liyuan Chen, PhD,* Michael Dohopolski, MD,* Zhiguo Zhou, PhD,† Kai Wang, MS,* Rongfang Wang, PhD,* David Sher, MD,* and Jing Wang, PhD***

**Medical Artificial Intelligence and Automation (MAIA) Laboratory, Department of Radiation Oncology, University of Texas Southwestern Medical Center, Dallas, Texas; †School of Computer Science and Mathematics, University of Central Missouri, Warrensburg, Missouri*

Received Jun 30, 2020. Accepted for publication Feb 1, 2021.

**Purpose:** Accurate lymph node (LN) malignancy classification is essential for treatment target identification in head and neck cancer (HNC) radiation therapy. Given the constraints imposed by relatively small sample sizes in real-world medical applications, to classify LN malignancy status accurately, we proposed an attention-guided classification (AGC) scheme that (1) incorporates human knowledge (ie, LN contours) into model training to guide model's "learning" direction, alleviating the critical requirement of large training samples by deep learning approaches; and (2) does not require accurate delineation of LNs in the inference stage but can highlight the discriminative region nearby the LN, which is important for malignancy determination.

**Methods and Materials:** In the proposed AGC scheme, there is an attention-guided convolutional neural network (agCNN) module, followed by a classification convolutional neural network (cCNN) module. The input of the proposed AGC scheme is a region of interest (ROI) containing the LN and its surrounding tissues. The agCNN is designed to find the discriminative region in the ROI, which outputs an activation map whose voxel values indicate the importance of the voxels in malignancy prediction. Through multiplying the activation map with the ROI, we obtain the input for the cCNN, which finally outputs the LN malignancy probability. To demonstrate the effectiveness of the proposed scheme, we performed experimental studies using positron emission tomography and contrast-enhanced computed tomography from 129 surgical HNC patients, including 791 LNs, with pathologic ground truth of malignancy status. To evaluate the performance, 5-folder cross validation was used.

**Results:** The sensitivity, specificity, accuracy, and area under the receiver operating characteristic (ROC) curve values obtained by the proposed AGC scheme were 0.91, 0.93, 0.92, and 0.98, respectively, significantly outperforming conventional convolutional neural network and radiomics approaches at a significance level of .05 under a paired ROC comparison statistical test.

**Conclusions:** We developed an AGC scheme that can highlight the discriminative region in an image for LN malignancy prediction, outperforming a conventional radiomics method that requires accurate segmentation and a standard convolutional neural network model without involving segmentation. 

Corresponding author: Jing Wang, PhD; E-mail: Jing.Wang@UTSouthwestern.edu

Disclosures: none.
Research data are not available at this time.

Int J Radiation Oncol Biol Phys, Vol. 110, No. 4, pp. 1171–1179, 2021
0360-3016/$ - see front matter 
https://doi.org/10.1016/j.ijrobp.2021.02.004

## Introduction

Head and neck (HN) cancer is the sixth most common cancer in the United States, with an estimated 65,630 new patients diagnosed in 2020.[1] For patients with HN cancer, the presence of metastatic disease in the cervical lymph nodes (LNs) increases the risk of distant metastases and death by 50% compared with patients with disease limited to the primary site.[2,3] Accurate identification of malignant cervical LNs is critical because this information helps dictate the area needing to be covered by radiation treatment fields. Malignant nodes not appropriately covered by these radiation fields may inevitably lead to locoregional recurrences, whereas misidentifying benign LNs as malignant can lead to larger treatment fields, which may cause unnecessary toxicity.[4]

Physicians use information obtained from physical examinations, contrast-enhanced computed tomography (CT) imaging, and Fluorine-18 fluorodeoxyglucose (FDG) positron emission tomography (PET) imaging to determine whether an LN is likely malignant. Identification is easier when LNs on CT and PET scans are large, necrotic, and highly FDG avid. Unfortunately, not all involved LNs are clinically obvious. Classifying LNs as benign or malignant can be difficult when they are smaller or less FDG avid. A meta-analysis of 32 studies showed that the sensitivity and specificity of PET in detecting LN metastases in patients with head and neck squamous cell carcinoma were 79% and 86%, respectively.[5,6] Failure to treat a malignant node inevitably leads to regional recurrence, whereas overestimating the risk of malignancy will cause unnecessary toxicity. If a misidentified LN target is next to a critical structure, it is more difficult to appropriately spare the organ at risk and increases the chance of subsequent side effects.[7] The accuracy of LN metastasis identification strongly depends on the physician's experience. An automatic and objective strategy with high sensitivity and specificity is needed to differentiate between malignant and benign LNs so that only nodes with a high probability of metastasis are included in the gross tumor volume.

Imaging-based prediction approaches can generally be summarized into 2 categories: handcrafted-feature-based radiomics methods and self-learned-feature-based convolutional neural network (CNN) approaches. Radiomics methods make predictions through extracting and analyzing a large number of predefined quantitative features of the segmented target, which requires accurate delineation of the target.[8] The HN region contains a rich and elaborate lymphatic network of more than 300 nodes; thus, conventional radiomics models requiring accurate delineation of each LN would be labor intensive and time consuming even for experts with the relevant domain knowledge.[9] CNN-based approaches using a region of interest (ROI) containing both the target LN and its surrounding tissues as the input do not require detailed and accurate contours of the target but instead are able to highlight a discriminative region that is important for the final prediction. Yet, the performance of CNN-based approaches heavily relies on a large amount of training data, which is often lacking in many real-world medical applications.

In this study, we propose an attention-guided classification (AGC) scheme that incorporates human knowledge (ie, LN contours) to guide our model's "learning" given the constraints imposed by a relatively limited training data set. This method was motivated by human experience that voxels within and surrounding an LN contour are important for appropriately classifying an LN as malignant or not, especially for LNs with ill-defined, irregular margins.[10] Similar to commonly used CNN approaches, the proposed AGC scheme would use a ROI composed of the LN and its surrounding tissue as an input for the LN malignancy prediction. Instead of a direct prediction through a classification CNN (cCNN) architecture (which commonly includes several convolutional, down-sampling, and fully connected layers) after the ROI input, we added an attention-guided CNN (agCNN) module in front of the cCNN. This attention module was designed to generate an activation map from human knowledge (ie, LN contours given by the physicians), which would be used together with the original ROI in the subsequent cCNN for the final malignancy prediction. This activation map can be similar to the LN contour but is not necessarily the same because it can provide additional information "contained" within the surrounding tissues, while also accommodating data quality imperfection. That is, the activation map learned from the agCNN module provides guidance on the learning focus for malignancy predictions. Hence, the additional information gained by the agCNN module can be reflected in the decreased amount of data needed in the training process of the subsequent cCNN module (ie, the number of LNs with known malignant/benign status). Note that the proposed method needs LN contours only during the training stage, not during the inference or testing stage. Once the model is well trained, the ROIs containing the LNs alone are sufficient for final malignancy prediction in the inference stage, without requiring accurate contours of LNs.

## Methods and Materials

### Patient data set

This data set included 129 patients with oropharyngeal squamous cell carcinoma (OPSCC) who had preoperative PET and contrast-enhanced CT imaging and underwent surgical neck dissection. Contrast-enhanced CT images were scanned on a Philips CT scanner with 3 mm slice thickness and 1.1719 mm pixel spacing. PET images were scanned on a Siemens PET scanner with 3 mm slice thickness and 4.0728 mm pixel spacing. PET images were resampled to have the same resolution as CT images. A total of 791 LNs were contoured by a radiation oncologist on contrast-enhanced CT guided by PET: 620 were

benign and 171 were malignant based on subsequent pathology reports obtained from the neck dissections. We randomly partitioned the whole patient data set into 5 fold, labeled F1, F2, F3, F4, and F5, which included 25, 26, 26, 26, and 26 patients' data, respectively. Detailed distributions of benign and malignant nodes in each folder are listed in Table 1. To evaluate the performance of the proposed AGC scheme, 5-folder cross validation was performed. That is, each time, we used 3 folders' data to train the model, whose hyperparameters were tuned based on 1 folder's validation data; we then used the remaining folder's data to test the model's performance. In total, we performed 5 iterations, such that every folder was used for testing.

## AGC scheme

The proposed AGC scheme is shown in Figure 1. There are 2 modules in the scheme: (1) the agCNN module and (2) the cCNN module. The inputs of the proposed AGC scheme are 2 ROI patches of size $64 \times 64 \times 48$, which are separately extracted from contrast-enhanced CT and PET images. Here, PET and contrast-enhanced CT images need to be aligned with each other. In this study, we used the commercial Velocity software package (Varian Medical Systems) to perform the rigid registration between PET and contrast-enhanced CT images. The ROI patches of each LN contain the LN and its surrounding tissues. Surrounding tissues here refer to the tissues within a 10-voxel expansion in each dimension (x, y, and z) from the bounding box of the LN. The size $64 \times 64 \times 48$ was chosen to cover the largest LN. The agCNN module was designed to find the discriminative region in the ROI patches, outputting an activation map whose voxel values indicate the importance of the voxels for the final malignancy prediction. Through multiplying the activation map with the ROI patches, we obtained the inputs for the cCNN module, which outputs the final LN malignancy prediction probabilities. During the training stage, agCNN's loss function minimized the differences between the activation map and reference node contour, and the cCNN's loss function minimized the differences between the predicted probabilities and the node malignancy status. The agCNN and cCNN modules were trained simultaneously through minimizing the combination of agCNN's loss and cCNN's loss functions. That is, for each module, minimizing the other module's loss function was a regularizer to guide its learning direction. For the agCNN module, the activation map was not only trained to be close to the LN contour but also to be useful for the LN malignancy prediction. For the cCNN module, the malignancy prediction was trained to be made mainly based on the characteristics of the LN, with the constraint that the activation map should be similar to the LN contour. In the inference stage, for each LN, only ROI patches extracted from contrast-enhanced CT and PET images containing this LN and surrounding tissues are needed for the malignancy status prediction.

**Table 1** Distributions of benign and malignant nodes in 5 folders

| | Benign, no. | Malignant, no. | Total, no. |
|---|---|---|---|
| F1 | 109 | 43 | 152 |
| F2 | 94 | 31 | 125 |
| F3 | 119 | 38 | 157 |
| F4 | 142 | 28 | 170 |
| F5 | 156 | 31 | 187 |
| Total | 620 | 171 | 791 |

*Abbreviation:* F = folder.

## agCNN module

The architecture of the proposed agCNN module adopts the U-net structure,[11,12] which is commonly used for medical image segmentation. The agCNN module consists of an encoder and a decoder and several parallel skip concatenations between the encoder and decoder. The inputs of the agCNN module are 2 ROI patches extracted from contrast-enhanced CT and PET, respectively. The encoder used convolution and max-pooling layers to obtain the input's global features, while the decoder, which was associated with parallel skip connection, used convolution transpose, concatenation, and convolution layers to up-sample the output image's size and then incorporate detailed local features (eg, edge information) to generate an activation map subsequently used in the following cCNN module.

The reference output of the agCNN module is the LN contour given by the physicians, denoted as $GT_{contour}$. The discriminative region loss—that is, the agCNN loss, measured by the dice similarity coefficient index (DSC) between the activation maps ($M_{activation}$) and $GT_{contour}$—is formulated as follows:

$$Loss_{agCNN} = 1 - DSC(M_{activation}, GT_{contour}) = 1 - \frac{2|M_{activation}, \cap GT_{contour}|}{|M_{activation}, | + |GT_{contour}|} \quad (1)$$

## Classification-CNN module

Motivated by the promising results in a previous work[13] using AlexNet-based CNN[14] to predict LN metastases as belonging to 1 of 3 categories (benign, suspicious, or malignant) aided by contrast-enhanced CT and PET, we adopted the same CNN architecture in the proposed cCNN module to classify LN malignancy status into 2 classes: benign and malignant. Note that any advanced-classification CNN architecture[15,16] can be incorporated in this cCNN module.

The reference output of the proposed cCNN module is the nodal malignancy status obtained from the pathology reports (benign, 0; malignant, 1). The malignancy prediction loss—that is, the cCNN loss, measured by the binary

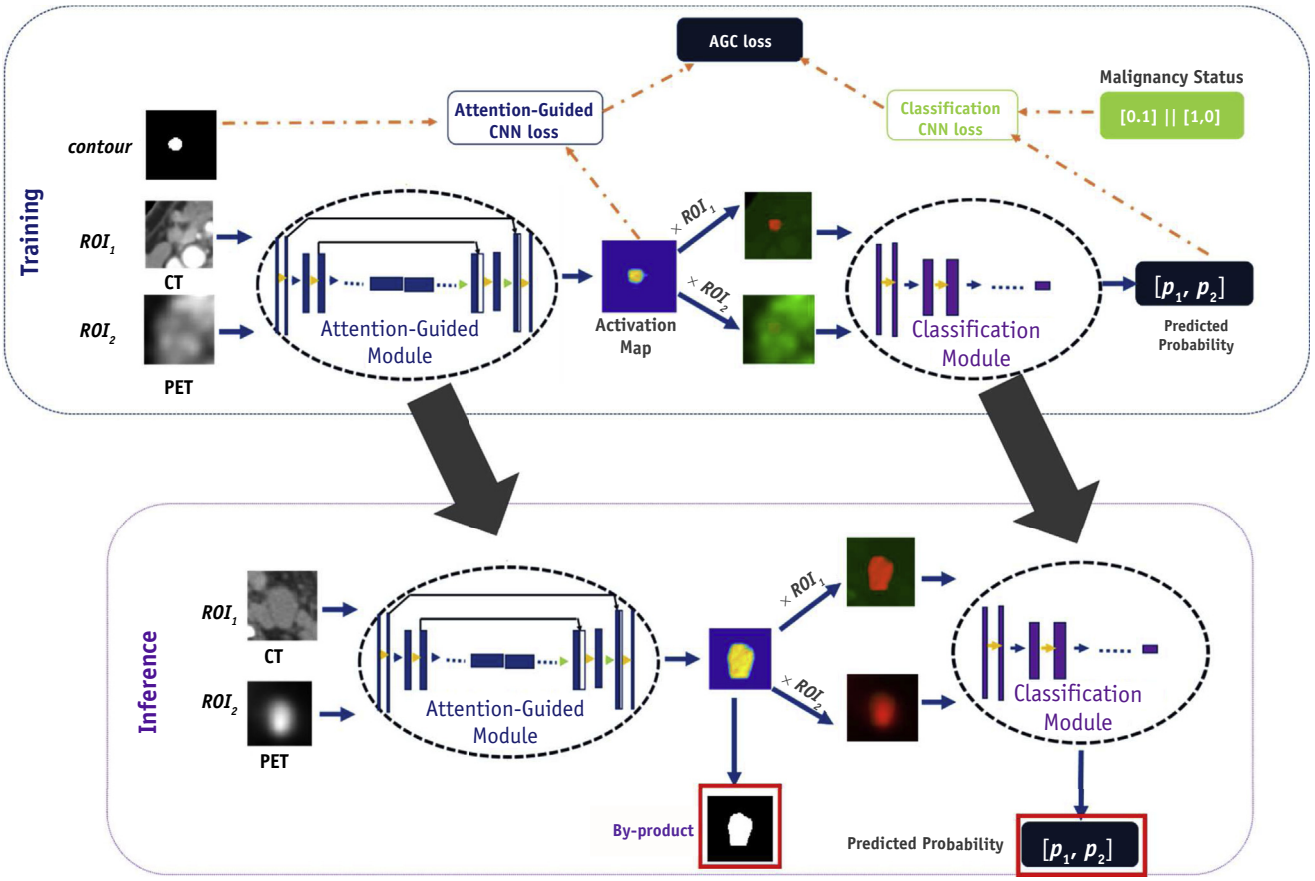


**Fig. 1.** The workflow of the proposed AGC scheme for lymph node malignancy prediction. The upper blue box contains the training stage, and the lower purple box contains the inference stage. In the inference stage, only the PET and CT ROI patches are needed as inputs. These 2 ROI patches first go through the attention-guided module to generate an activation map. Multiplying the activation map with the original ROI patches provides inputs for the subsequent classification module, and hence the final malignancy prediction results. *Abbreviations:* AGC = attention-guided classification; CNN = convolutional neural network; CT = computed tomography; PET = positron emission tomography; ROI = region of interest.

cross entropy (BCE) between the reference malignancy label ($GT_{label}$) and the predicted probability ($p$)—is formulated as follows:

$$Loss_{cCNN} = BCE(p, GT_{label})$$
$$= -(GT_{label}\log(p) + (1 - GT_{label})\log(1 - p)) \quad (2)$$

## Training objective function for the AGC scheme

The training objective function for the proposed AGC scheme is a combination of (1) discriminative region loss ($Loss_{agCNN}$) of the agCNN module and (2) malignancy prediction loss ($Loss_{cCNN}$) of the cCNN module. The mathematical formula of the proposed training objective function is as follows:

$$Loss = (1 - \alpha) \cdot Loss_{agCNN} + \alpha \cdot Loss_{cCNN} \quad (3)$$

Here, $\alpha$ is a trade-off parameter to balance the attention and prediction loss that was obtained empirically. Under such a loss function, the activation map generated from the agCNN module is close to the LN contour, but not necessarily limited to be exactly the same as the LN contour, which allows for contributions from surrounding tissues for the final malignancy prediction and accommodates imperfection in the data quality.

The model was developed using the Tensorflow Library in Python. All the experiments were performed using an NVIDIA ROV100 GPU equipped with 32G memory. We used an Adam optimization algorithm to train the proposed AGC scheme. The learning rate, $\beta_1$, and $\beta_2$ for the Adam algorithm are set as $1 \times 10^{-4}$, 0.9, and 0.999, respectively. The trade-off parameter $\alpha$ in the loss function of Equation 3 is set as 0.25. The average training time for each folder was $\sim$3 hours. All these parameters were tuned via a grid search (learning rate range is $[10^{-5}, 10^{-3}]$, and trade-off parameter range is [0, 1]) and selected based on the performance of a validation data set.

## Comparison methods and evaluation criteria

To evaluate the classification performance of the proposed AGC scheme, we compared it with a conventional

radiomics model that requires accurate LN contours, the cCNN module using the ROI patch as input (cCNN−ROI), and the cCNN module using the segmented nodes as input (cCNN−NodeOnly). Based on 1 of our previously published works,[13] 257 features, including intensity, geometry, and texture features, were first extracted from the contoured LNs in PET and CT images for the radiomics model. Intensity features include the minimum, mean, median, maximum, standard deviation, skewness, kurtosis, sum, and variance values in the LNs. Geometric features contain the eccentricity, elongation orientation, major diameter, minor diameter, volume, perimeter, and bounding box volume of the LNs. Texture features were generated based on a 3D gray-level co-occurrence matrix consisting of homogeneity, texture variance, max-probability, contrast, energy, sum-mean, cluster prominence, inertia, entropy, correlation, cluster shade, and inverse variance. A support vector machine was used to build the predictive model, where feature selection and model parameter training were performed simultaneously. Four evaluation criteria were used to quantitatively evaluate the classification performance of four different methods: sensitivity, specificity, accuracy (ACC), and area under the receiver operating characteristic curve (AUC).

## Results

Figure 2 illustrates loss (total loss, attention loss, and classification loss) and evaluation criteria (M) along the training Epochs. Overall, the proposed AGC scheme achieved 0.91 sensitivity, 0.93 specificity, 0.92 ACC, and 0.98 AUC values. The detailed quantitative results of the other 3 models are listed in Table 2. The ACC, sensitivity, and specificity listed in Table 2 were obtained using a threshold of 0.5. As shown, the proposed AGC scheme, which does not require precise LN contours, can generate impressive sensitivity and AUC values. Even though the radiomics model slightly outperformed the proposed AGC in ACC (0.93 vs. 0.92, respectively), the sensitivity, a measure of significant clinical interest, was improved substantially in the proposed AGC, from 0.76 to 0.91. The receiver operating characteristic curves (ROCs) of the 4 different methods are also plotted in Figure 3, which illustrates that the proposed AGC scheme can better differentiate between benign and malignant LNs. We also investigated the statistical significance of the difference between the proposed AGC scheme and the other 3 methods (Table 3) using a paired empirical (nonparametric) ROC comparison test developed by Number Cruncher Statistical System.[17] That is, a z-test was used to compare the AUCs of 2 prediction tests in a paired design. Based on the *P* values shown in Table 3, the proposed AGC scheme significantly outperforms the other methods at a significance level of .05. Note that the quantitative results listed in Tables 2 and 3 represent the total testing results over the 5 folders.

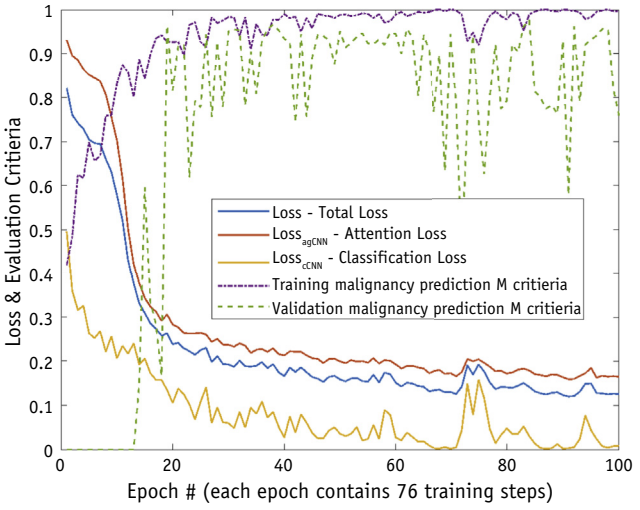


**Fig. 2.** Loss (total loss, attention loss, and classification loss) and evaluation criteria (M) along the training epochs. Here, total loss is calculated from attention loss and classification loss, based on Equation 3. M is the evaluation criterion used to select an optimal model for further testing and is a composite value of sensitivity and specificity whose weighting parameters are 0.6 and 0.4, respectively, with a criteria threshold of 0.75. $(M = 0.6 * (sensitivity \geq th) + 0.4 * (specificity \geq th), th = 0.75.)$ *Abbreviations:* agCNN = attention-guided convolutional neural network; cCNN = classification convolutional neural network.

We also investigated the training sample size required by the proposed AGC scheme to get similar prediction results, compared with the conventional cCNN−ROI model (Table 4). Only 60% of training samples were needed by AGC to generate prediction results similar to those obtained by the cCNN−ROI model using 100% of the training sample. We also investigated the statistical significance of the difference within the proposed AGC scheme at different percentages of training samples (Table 5) using a paired empirical (nonparametric) ROC comparison test developed by NCSS.[17] The *P* value between AGC (100%) and AGC (60%) is ∼.07. At the 5% significance level, for

**Table 2** The 4 evaluation criteria, obtained by 4 different approaches

| | Sensitivity | Specificity | ACC | AUC |
|---|---|---|---|---|
| cCNN−NodeOnly | 0.88 | 0.92 | 0.91 | 0.96 |
| cCNN−ROI | 0.87 | 0.95 | 0.93 | 0.96 |
| Radiomics | 0.76 | 0.97 | 0.93 | 0.95 |
| AGC | 0.91 | 0.93 | 0.92 | 0.98 |

Values in Table 2 summarize results over 5 folders.
*Abbreviations:* ACC = accuracy; AGC = attention-guided classification; AUC = area under the receiver operating characteristic curve; cCNN = classification convolutional neural network; ROI = region of interest.

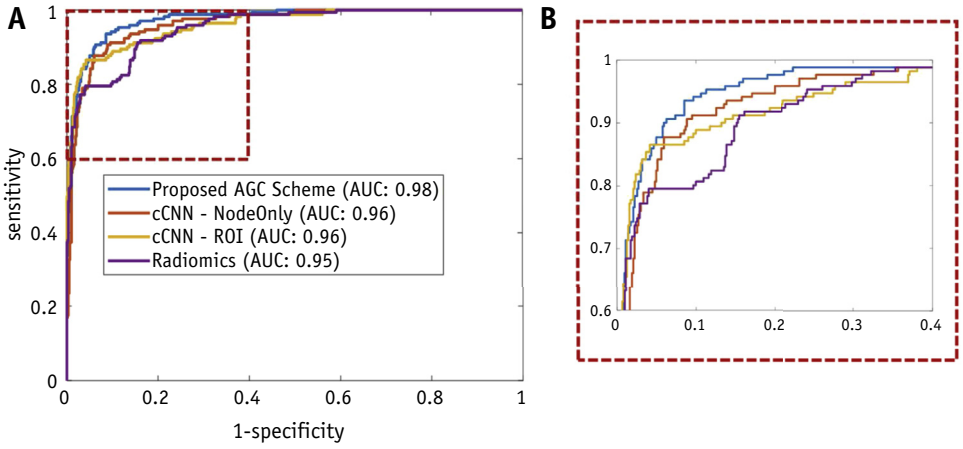


**Fig. 3.** (A) The ROC curves of the 4 different approaches. (B) Zoomed-in area in the rectangular of A. *Abbreviations:* AGC = attention-guided classification; AUC = area under the receiver operating characteristic curve; cCNN = classification convolutional neural network; ROI = region of interest.

the proposed AGC scheme, using 60% of the training sample was not much different than using the original sample size. In addition, the statistical significance of the difference between the proposed AGC scheme using 60% of training sample and cCNN−ROI (100%) was also investigated (Table 5). A $P$ value of ∼.69 further demonstrated that only 60% of samples were needed by AGC to generate prediction results similar to those using 100% of the sample in the conventional cCNN−ROI model.

Aside from achieving accurate malignancy prediction, the proposed AGC scheme can highlight the discriminative region in the ROI. We illustrated the activation maps obtained by the proposed AGC scheme and the reference contours (r-contour) given by the physicians of 3 LNs in Figure 4. Inhomogeneity of intensity values existed in the activation maps, which illustrates the importance of different spatial locations/textures in the node or surrounding the node. By using a threshold of .5, the activation map can be converted into a binary mask (ie, predicted contour), which was found to match well with the r-contour. Dice similarity coefficients between the reference contour and the predicted contour given by activation maps under a threshold of .5 were measured for 5 folders, with a mean value of 0.834 and standard deviation of 0.105.

**Table 3** $P$ values in paired ROC comparison tests between the proposed and the other 3 methods

| Methods | $P$ values |
|---|---|
| AGC versus cCNN−NodeOnly | .0167 |
| AGC versus cCNN−ROI | .0094 |
| AGC versus radiomics | .0004 |

*Abbreviations:* AGC = attention-guided classification; cCNN = classification convolutional neural network; ROC = receiver operating characteristic; ROI = region of interest.

## Discussion

In the proposed AGC scheme, 2 ROI patches are extracted from aligned PET and contrast-enhanced CT images and are used as the model inputs. All images are resampled to the same resolution. In this study, we performed 1 rigid registration to align PET and contrast-enhanced CT images. However, the model is not limited to situations in which only 1 registration is performed. Large deformations between contrast-enhanced CT and PET (eg, different neck flexions/extensions), deformable image registration, or multiple registrations can also be used to align PET and contrast-enhanced CT for nodes in different cervical levels. Although the proposed agCNN does not require perfectly aligned PET and CT because they are used as 2 input channels, better registration between PET and CT may facilitate model training. In addition, a 10-voxel expansion from the bounding box of the LN was performed in the ROI patch extraction in this study. The purpose of this surrounding tissue expansion was to ensure the entire LNs were

**Table 4** Evaluation criteria values

| Model, % of sample | Sensitivity | Specificity | AUC |
|---|---|---|---|
| AGC, 100% | 0.91 | 0.93 | 0.98 |
| AGC, 80% | 0.91 | 0.93 | 0.98 |
| AGC, 60% | 0.87 | 0.95 | 0.96 |
| AGC, 40% | 0.84 | 0.92 | 0.92 |
| cCNN−ROI, 100% | 0.87 | 0.95 | 0.96 |
| cCNN−ROI, 80% | 0.78 | 0.95 | 0.95 |
| cCNN−ROI, 60% | 0.78 | 0.95 | 0.95 |
| cCNN−ROI, 40% | 0.74 | 0.94 | 0.93 |

Values were obtained by the proposed AGC scheme and the cCNN−ROI method using different percentages of training samples. Values in Table 4 summarize results over 5 folders.

*Abbreviations:* AGC = attention-guided classification; AUC = area under the receiver operating characteristic curve; cCNN = classification convolutional neural network; ROI = region of interest.

**Table 5** *P* values in paired ROC comparison tests between the proposed AGC and the cCNN − ROI method using different percentages of training samples

| Methods (% of sample) | *P* values |
|---|---|
| AGC (100%) versus AGC (80%) | .0913 |
| AGC (100%) versus AGC (60%) | .0662 |
| AGC (100%) versus AGC (40%) | .0001 |
| AGC (60%) versus cCNN−ROI (100%) | .6852 |
| cCNN−ROI (100%) versus cCNN−ROI (80%) | .2294 |
| cCNN−ROI (100%) versus cCNN−ROI (60%) | .0506 |
| cCNN−ROI (100%) versus cCNN−ROI (40%) | .0002 |

*Abbreviations:* AGC = attention-guided classification; cCNN = classification convolutional neural network; ROI = region of interest.

included in the extracted patch to be used as input for CNN models. In our implementation, we first identified a bounding box that covered an LN based on manual contours. Given the uncertainties of manual contours and our intention to identify relevant surrounding tissues for malignancy prediction in agCNN, we expanded the initially identified LN bounding box to make sure the whole LN and enough surrounding regions were included in the extracted patch for CNN input. Based on the model architecture and GPU memory, 10-voxel expansion was selected in this work. Although we expanded the bounding box isotopically with the potential of including bone or muscle, we did not perform any additional steps to exclude bone or muscle, which can be included in the extracted patch. This is because the proposed agCNN module was designed to identify important and useful regions for malignancy prediction in the ROI patch. Although bone or muscle can be included in the ROI patch, the designed agCNN module can identify which regions are needed by the subsequent cCNN module.

We also investigated the model performance with a decreased number of training samples for the proposed AGC scheme and cCNN-ROI model in this work. Although the model performance as measured by sensitivity gradually decreases with reduced training samples in both the AGC scheme and cCNN-ROI model, the overall performance as measured by AUC does not change dramatically with the reduced training sample. This phenomenon can be attributed to the following 2 reasons. First, the majority of LNs (620/791 = 78.4%) were benign in our data set. Thus, the model performance measured by the AUC was mainly driven by specificity. For example, when the training sample was reduced to 40% in the cCNN-ROI model, although sensitivity was reduced from 0.87 (using 100% of the training sample) to 0.74, the specificity was reduced slightly from 0.95 to 0.94, resulting in a slightly decreased AUC from 0.96 to 0.93. Second, during the training process, the samples in the same validation folder are used for choosing an optimal model. As such, although the training samples are reduced, the chosen model based on the same validation set may have similar properties, leading to similar results on the test set.

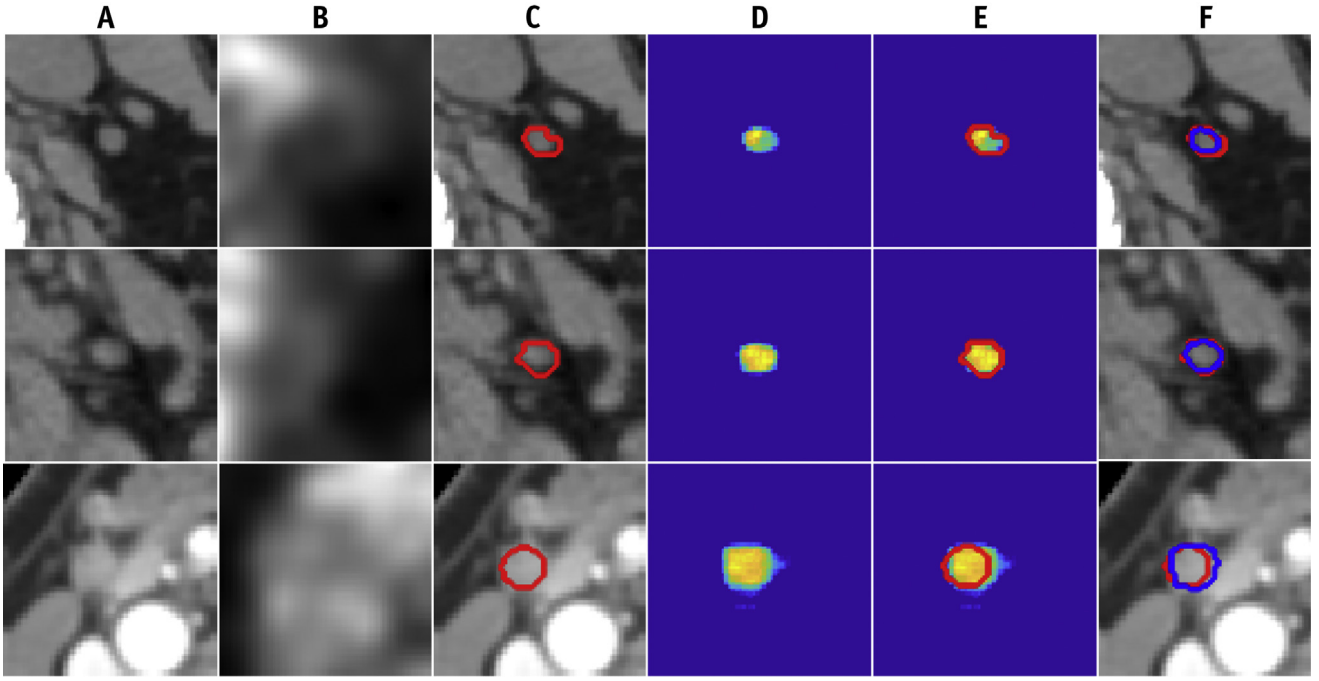


**Fig. 4.** Activation maps obtained by the AGC scheme and the reference contours of 3 lymph nodes. (A) CT; (B) PET; (C) r-contour; (D) activation; (E) overlay of r-contour and activation map; and (F) r-contour and p-contour. *Abbreviations:* AGC = attention-guided classification; CT = computed tomography; p-contour = predicted contour; PET = positron emission tomography; r-contour = reference contours.

In this work, imperfection in data quality was mainly due to uncertainties in manual contours in the training data set. The contours of nodes in the training data set might not be perfect because they were obtained manually by clinicians. As a consequence, the contours are subject to human variations, which can be affected by other human factors, such as the experience level and personal style of the clinician. The uncertainty caused by contour uncertainties could be mitigated by the proposed AGC scheme. In AGC, contours essentially provide guidance on the generation of the activation map, which is allowed to deviate from the contour in the proposed scheme. More specifically, the proposed scheme allows us to identify important regions for malignancy prediction, even though they are not within a node contour. In comparison, regions incorrectly contoured may be identified as less important for prediction, receiving relatively low intensity values in the activation map.

Although the primary goal of the proposed AGC is to classify whether an LN is malignant or benign, the activation map generated by the agCNN module can be used for further applications. As shown in Figure 3, the activation maps can be used to automatically generate contours around LNs by selecting an appropriate threshold. Furthermore, because the agCNN module not only focuses on LN itself but also relevant regions surrounding the LN that contribute to malignancy prediction, the treating physician can focus on this region to confirm there is no alternative explanation for the classification (eg, CT artifact, misregistration).

Compared with results obtained by the standard CNN focusing on segmented LNs only in 1 study[18] (eg, an AUC of 0.91 for HN LN malignance prediction from pretreatment CT), the proposed AGC scheme obtained excellent classification performance, achieving a sensitivity of 0.91 and an AUC of 0.98 on our testing data set. Nevertheless, several aspects need further improvement. In this study, we investigated the model performance on 129 patients with OPSCC from 1 institution. We first partitioned the entire data set into 5 folders. For each iteration, we used 3 folders for training, 1 folder for validation (ie, tuning hyperparameters and selecting the optimal model), and the last folder for testing. Hence, for each iteration, we have an independent testing data set to evaluate the model performance. In total, we performed 5 iterations, such that every folder was used for testing. However, an external data set from other institutions would be needed to further evaluate the generalizability of the model.

The model developed in this work is focused on OPSCC. It is of great interest to study whether the established LN malignancy prediction model can be directly extended to a different site, such as the larynx. However, LNs in the larynx region are generally smaller than those in the oropharyngeal region, which may require further fine-tuning using data from the specific site to adapt the model for optimal performance. This question can be answered once we can construct a data set for the larynx site, which is warranted in a future study.

Additionally, this study does not consider uncertainty of the model output. Incorporating uncertainty into each prediction could potentially improve accuracy because highly uncertain predictions could be further scrutinized by experts. Moreover, having a measure that quantifies the reliability of a prediction could increase physician confidence in adopting machine learning−based models.[19] Prediction uncertainty in relation to cervical LN classifications could be explored in a future study.

## Conclusion

We proposed an AGC scheme to predict cervical LN malignancy status. Compared with a conventional radiomics model[20] requiring accurate delineation of targets to extract features, the proposed AGC scheme does not require accurate delineation of LNs, but can highlight the discriminative region near the LN that is important for determining malignancy status by the model itself, providing further useful information for treatment target delineation andoffering a form of model interpretability. In addition, compared with pure-classification CNN models[18] using whole image/extracted ROIs directly for prediction, the proposed AGC scheme incorporates human knowledge (ie, LN contour) as guidance in the learning process to reduce the critical requirement of sample size for conventional CNNs and hence improve the efficacy of LN malignancy prediction.

## References


1. Siegel RL, Miller KD, Jemal A. Cancer statistics, 2020. *CA Cancer J Clin* 2020;70:7-30.
2. Layland MK, Sessions DG, Lenox J. The influence of lymph node metastasis in the treatment of squamous cell carcinoma of the oral cavity, oropharynx, larynx, and hypopharynx: N0 versus N+. *Laryngoscope* 2005;115:629-639.
3. Popescu B, Patricia E, Bertesteanu SVG, et al. Methods of investigating metastatic lymph nodes in head and neck cancer. *Maedica* 2013;8:384.
4. Nguyen-Tan PF, Zhang Q, Ang KK, et al. Randomized phase III trial to test accelerated versus standard fractionation in combination with concurrent cisplatin for head and neck carcinomas in the Radiation Therapy Oncology Group 0129 trial: Long-term report of efficacy and toxicity. *J Clin Oncol* 2014;32:3858.
5. Finn S, Toner M, Timon C. The node-negative neck: Accuracy of clinical intraoperative lymph node assessment for metastatic disease in head and neck cancer. *Laryngoscope* 2002;112:630-633.
6. Kyzas PA, Evangelou E, Denaxa-Kyza D, et al. 18f-fluorodeoxyglucose positron emission tomography to evaluate cervical node metastases in patients with head and neck squamous cell carcinoma: A meta-analysis. *J Natl Cancer Inst* 2008;100:712-720.
7. Li B, Li D, Lau DH, et al. Clinical-dosimetric analysis of measures of dysphagia including gastrostomy-tube dependence among head and neck cancer patients treated definitively by intensity-modulated radiotherapy with concurrent chemotherapy. *Radiat Oncol* 2009;4:52.
8. Y.-q. Huang, C.-h. Liang, He L, et al. Development and validation of a radiomics nomogram for preoperative prediction of lymph node metastasis in colorectal cancer. *J Clin Oncol* 2016;34:2157-2164.
9. Agarwal AKM. Anatomy, Head and Neck, Lymph Nodes. StatPearls Publishing, Treasure Island. FL; 2020.

10. Hoang JK, Vanka J, Ludwig BJ, et al. Evaluation of cervical lymph nodes in head and neck cancer with CT and MRI: Tips, traps, and a systematic approach. *Am J Roentgenol* 2013;200:W17-W25.
11. Çiçek Ö, Abdulkadir A, Lienkamp SS, Brox T, Ronneberger O. 3D U-Net: learning dense volumetric segmentation from sparse annotation. InInternational conference on medical image computing and computer-assisted intervention. 2016, Springer, Cham, pp. 424-432.
12. Ronneberger O, Fischer P, Brox T. U-net: Convolutional networks for biomedical image segmentation. InInternational Conference on Medical image computing and computer-assisted intervention, 2015, Springer, Cham, pp. 234-241.
13. Chen L, Zhou Z, Sher D, et al. Combining many-objective radiomics and 3d convolutional neural network through evidential reasoning to predict lymph node metastasis in head and neck cancer. *Phys Med Biol* 2019;64:075011.
14. Krizhevsky A, Sutskever I, Hinton GE. Imagenet classification with deep convolutional neural networks. *Adv Neural Inf Process Syst* 2012;25:1097-1105.
15. Iandola FN, Han S, Moskewicz MW, Ashraf K, Dally WJ, Keutzer K. SqueezeNet: AlexNet-level accuracy with 50x fewer parameters and <0.5 MB model size. *arXiv preprint arXiv: 1602.07360* 2016.
16. Ballester P, Araujo RM. On the performance of googlenet and alexnet applied to sketches. In Proceedings of the AAAI Conference on Artificial Intelligence, 2016, Vol. 30.
17. NCSS 2020 Statistical Software. *NCSS, LLC. Kaysville, Utah, USA* 2020. Available at: ncss.com/software/ncss. Accessed March 3, 2021.
18. Kann BH, Aneja S, Loganadane GV, et al. Pretreatment identification of head and neck cancer nodal metastasis and extranodal extension using deep learning neural networks. *Sci Rep* 2018;8:1-11.
19. Leibig C, Allken V, Ayhan MS, et al. Leveraging uncertainty information from deep neural networks for disease detection. *Sci Rep* 2017;7:1-14.
20. Forghani R, Chatterjee A, Reinhold C, et al. Head and neck squamous cell carcinoma: Prediction of cervical lymph node metastasis by dual-energy CT texture analysis with machine learning. *Eur Radiol* 2019; 29:6172-6181.