Title Page

Title

A multimodal deep learning model for laryngeal and hypopharyngeal lesions diagnosis: a multicenter retrospective study

Authors

Jiahong Zhang, MS[a, g]; Kewei Liang, PhD[b, g]; Qingxiu Yao, PhD[c, g]; Yifan Zhang, BS[b]; Qun Mo, PhD[b]; Xiaojing Chen, BM[d]; Siyuan Qu, MM[e]; Yunzhen Luo, MM[f]; Hengchao Chen, PhD[c, *]; Libo Dai, PhD[c, **]; Shuihong Zhou, PhD[c, ***];

[a]Polytechnic Institute, Zhejiang University, Hangzhou, China
[b]School of Mathematical Sciences, Zhejiang University, Hangzhou, China
[c]Department of Otolaryngology, The First Affiliated Hospital, College of Medicine, Zhejiang University, Hangzhou, China
[d]Department of Otolaryngology, The First Affiliated Hospital of Wenzhou Medical University, Wenzhou, China
[e]Department of Otorhinolaryngology-Head and Neck Surgery, The Affiliated Lihuili Hospital, Ningbo University, Ningbo, China
[f]Department of Otolaryngology, The Second Affiliated Hospital of Jiaxing University, Jiaxing, China

[g]These authors contributed equally to this work.
[*]Corresponding Author
[**]Corresponding Author
[***]Corresponding Author

Corresponding Authors:
Shuihong Zhou, PhD, Department of Otolaryngology, The First Affiliated Hospital, College of Medicine, Zhejiang University, 79 Qingchun Road, Hangzhou 310003, China (1190051@zju.edu.cn).
Libo Dai, PhD, Department of Otolaryngology, The First Affiliated Hospital, College of Medicine, Zhejiang University, 79 Qingchun Road, Hangzhou 310003, China (dailibo@163.com).
Hengchao Chen, PhD, Department of Otolaryngology, The First Affiliated Hospital, College of Medicine, Zhejiang University, 79 Qingchun Road, Hangzhou 310003, China (chenhengchao@126.com).

Summary

Background

In authentic clinical settings, patient-related data, including clinical characteristics and outpatient history, contribute to the assessment of laryngeal and hypopharyngeal (LHP) disorders in laryngoscopic images. However, these data have never been utilized in current machine learning models. Therefore, we aim to develop and validate a multimodal deep learning network architecture for diagnosing LHP lesion.

Methods

Our study used an internal cohort consisting of 1,664 valid cases enrolled from one main medical center between April 2021 to August 2024, and three external validation cohorts consisting of 93, 120, and 167 valid cases enrolled from three different centers respectively. We propose a multimodal diagnostic algorithm framework that integrates laryngoscopy images, patient complaints, and clinical characteristics as multimodal input data. The framework comprises a multimodal backbone network, along with components for modal fusion and interaction. We evaluated the performance of our model on an internal validation cohort and three external validation cohorts. The performance metrics include accuracy, precision, recall, and F1-score. Additionally, we compared the performance of our model with laryngologists of different qualifications (experts, seniors and juniors) based on the aforementioned metrics.

Findings

On the internal validation set, our model achieves an accuracy of 91·67%, a precision of 91·67%, a recall of 91·87% and an F1-score of 91·82%, which surpasses the performance of expert laryngologists. On the three external validation sets, our model achieves superior performance over senior laryngologists.

Interpretation

The multiclassification diagnostic model architecture proposed in this study is capable of effectively predicting LHP disorders and it has great potential for future clinical applications.

Funding

This study was sponsored by the Key R&D Program of Zhejiang Province (No.2023C03066).

Keywords

Deep Learning, Laryngeal and hypopharyngeal lesions diagnosis, Multimodal fusion, Medical artificial intelligence

Research in context

Evidence before this study

We searched PubMed on January 22, 2025 using the term ("Deep learning" OR "Convolutional neural network" OR "Artificial intelligence" OR "Dataset" OR "Computer-aided") AND ("image" OR "imaging" OR "endoscopy") AND ("Laryngeal" OR "Larynx" OR "Pharyngeal" OR "Pharynx" OR "vocal cord" OR "vocal fold"), without language restriction. A total of 31 studies have been identified that have employed artificial intelligence (AI) methodologies, with the use of endoscopy or publicly available datasets for the purpose of predicting the diagnosis of LHP lesions. Of these, three studies were conducted by multimodal methods, which integrate different formats of dataset from the same individual, and had enhanced accuracy. Nonetheless, the three studies were binary classifications that distinguished between malignant and non-malignant cases. Moreover, only one of the studies employed a data format other than images, namely voice. Notably, there is a lack of literature addressing the use of outpatient data integrated into laryngoscopic images for predicting LHP lesions in multiclass operations.

Added value of this study

In this study, a multiclassification diagnostic model architecture based on multimodal information fusion of outpatient history, clinical characteristics (age and sex) and laryngoscopic images was developed to predict LHP disorders. The study was validated in multicenter datasets and compared with laryngologists of different qualifications. The results showed that the integrated outpatient data, which were readily available, enabled the model to predict LHP disorders more accurately than using images alone, and outperformed senior clinicians with 3-10 years' experience.

Implications of all the available evidence

Our study demonstrated that the proposed multiclassification diagnostic model enabled early prediction of LHP lesions using outpatient data and laryngoscopic images. This provides a foundation for clinical decision-making and minimizing unnecessary biopsy. In addition, similar to a real outpatient diagnostic process, higher accuracy can be expected as more data formats besides images are integrated.

Introduction

Laryngeal cancer ranks as the second most prevalent head and neck malignancy globally,[1] while hypopharyngeal cancer constitutes approximately 3% of such cases.[2] In 2022, global new cases reached 188,960 and 86,276 respectively for these laryngeal and hypopharyngeal (LHP) cancers.[3] Despite declining incidence and improved treatments, survival rates for LHP cancers have shown minimal improvement, particularly for hypopharyngeal cancer which maintains a poor 30-35% 5-year survival rate.[4,5] Early detection remains critical for preserving vital functions and survival outcomes.

Laryngoscopy serves as the primary diagnostic tool,[6] yet endoscopist training requires significant time investment. Deep learning shows promise in addressing this challenge through lesion classification and malignancy prediction.[7–9] Current AI applications demonstrate 0.806-0.997 accuracy in laryngeal lesion assessment, with 0.91 sensitivity and 0.94 specificity for benign/malignant differentiation.[13] However, existing models face limitations: reduced efficacy in multiclass differentiation and exclusive reliance on image-based analysis, neglecting clinically available patient data like demographics and medical history.

This study developed a multimodal deep learning model integrating laryngoscopic images with clinical features and medical history to reduce misdiagnosis in LHP lesions, which often arise from overlapping symptoms. The model was validated through external testing across diverse clinical settings and comparative analysis with clinicians, demonstrating its potential to improve diagnostic accuracy and promote equitable healthcare delivery in resource-limited regions.

Methods

Ethnics

This retrospective study was jointly conducted by four hospitals in China: the First Affiliated Hospital of Zhejiang University School of Medicine (Center 1) in Hangzhou, the Second Affiliated Hospital of Jiaxing University (Center 2) in Jiaxing, the Ningbo Medical Center Lihuili Hospital (Center 3) in Ningbo, and the First Affiliated Hospital of Wenzhou Medical University (Center 4) in Wenzhou. The study was conducted in collaboration with the School of Mathematical Sciences at Zhejiang University. Since the data extraction was part of routine treatment, no dedicated informed consent was required. The study adhered to the STARD-2015 guidelines and received approval from the Ethics Committee of Center 1 in 2024 (approval number: [2024B] IIT Ethics Approval No. 1292), from Center 2 in 2025 (approval number: The Second Hospital of Jiaxing Ethics 2025 Study No. 078), from Center 3 (approval number: 2025SL097), and from Center 4 (approval number: 2025-R095).

Data collection

We established a multimodal diagnostic model for LHP lesions by integrating laryngoscopy, outpatient history data, and clinicopathological findings. The raw laryngoscopic images were obtained under white light mode from clinical cases at Center 1 between April 2021 and August 2024. Inclusion criteria for malignant tumor cases required: patients with confirmed pathological diagnosis through surgery or biopsy at Center 1, and available preoperative laryngoscopic images from Center 1. Exclusion criteria specified: patients who had previously undergone radical surgery, chemoradiation, or targeted therapy for laryngeal tumors. Non-cancerous lesions (including vocal nodules, vocal polyps, and vocal cord leukoplakia) and normal control cases were randomly selected from contemporaneous clinical records without specific inclusion/exclusion criteria. Imaging was acquired using fiberoptic endoscopes (ENF-VH2 and ENF-VT3, Olympus Medical Systems Corp.) and 70° rigid laryngoscopes (D0104, Zhejiang Tiansong Medical Instrument Co. Ltd).

This study categorized patients into three groups: cancer (CA), normal (NORM), and non-cancerous lesions (NCL). NORM comprised clinically diagnosed normal individuals or non-space-occupying conditions (e.g., chronic

pharyngitis/laryngitis). CA included pathologically confirmed LHP cancers, sarcomas, or lymphomas. NCL was further classified per the 5th WHO guidelines[14] into benign lesions (BL; e.g., acute epiglottitis, vocal nodules, cysts, or polyps), low-grade squamous intraepithelial lesions (LG SILs; encompassing hyperplastic epithelium with or without mild atypia), and high-grade squamous intraepithelial lesions (HG SILs; moderate/severe atypia or carcinoma in situ). Case allocation adhered to clinical diagnosis for NORM and select BL subgroups (where biopsies were unnecessary), while CA, HG, LG, and remaining BL subgroups were confirmed by histopathology. One representative unprocessed image per subject was analyzed.

The perceptual modalities involved in the model of this work are threefold: patient features, patient complaints, and laryngoscopic images. The patient features involve two numerical data types: patient gender and age. For samples with different degrees of deterioration, we assign different target risk scores respectively. An example of multimodal samples is shown in **Fig. 1**.

Datasets setting

A total of 1708 samples were collected from Center 1 in this study. After removing samples with the exclusion criteria mentioned above, we obtained 1664 valid samples. These samples were further divided into a training set with 1424 samples and an internal validation set with 240 samples. To better evaluate the model's performance on the three major categories, the internal validation set was randomly sampled with 80 data points for NORM, NCL and CA each.

For the construction of the external validation set, we obtained 93 samples from Center 2, 120 samples from Center 3, and 167 samples from Center 4. **Fig. 2** shows the process of data collection and partitioning.

In summary, we enrolled 1708 samples in this study as internal set and 380 samples as external set. We used these datasets to construct and evaluate our model.

Model framework

This study utilizes a text and images processing backbone network for laryngoscopy images and patient complaints inputs. On this basis, a patient feature processing network is added to map clinical features into a space consistent with the dimensionality of image and text features. Finally, the three types of features are fused to generate a multimodal comprehensive vector for the patient, which is fed into a scoring network to predict the patient's disease risk score with a number between 0(NORM) and 2(CA). The complete model details and architecture are depicted in Supplementary A1, Supplementary A2 and **Fig. 3**.

We assigned predefined baseline scores (NORM: 0·0; NCL: 1·0; CA: 2·0) and adopted SIL classification thresholds (BL: 1·0; LG: 1·1; HG: 1·3) after exploratory analysis (Supplementary A2). Final classifications were determined by minimal distance to these baseline scores.

Moreover, this study utilizes a combined contrastive learning objective function.[15,16] This loss function incorporates class information into the construction of the embedding feature space, allowing the model to obtain diverse forms of information from the multimodal and multi-granularity signals, thereby enhancing its performance.

Clinician-machine comparison

The diagnostic ability of our model was compared with that of ten different qualified laryngologists using the internal validation set and three external validation sets. The ten qualified laryngologists were divided into three groups: experts,

seniors and juniors. The experts group consists of four laryngologists, each with over 10 years of clinical experience; The seniors group consists of three laryngologists, each with over 3 years of clinical experience and the juniors group consists of three laryngologists, each with over 1 year of clinical experience.

These ten laryngologists were not involved in the screening and cleaning of the data set and were unaware of all the laryngoscopic and pathological findings. The laryngoscope images, complaints, age and sex were assigned to each laryngologist for independent testing. Laryngologists need to first choose one of the three groups of normal, non-cancerous, and cancerous lesions, and then score each patient with a number between 0 and 2. A high score indicates a worse degree of the lesion. Those who tend to be normal can be scored 0·0 points, those who tend to be benign lesions (such as polyps, nodules, cysts, etc., without epithelial hyperplasia or atypia) can be scored 1·0 points, and those who tend to be LG SILs can be scored 1·1 points. Those who tend to be HG SILs can be scored 1·3 points, and those who tend to be various malignant tumors (excluding in-situ cancer) can be scored 2·0 points. At the same time, the four validation sets were input to our model for testing. Finally, the results provided by the model were compared with those given by the laryngologists to evaluate the diagnostic performance of the model.

Statistics

In this article, categorical variables were expressed as counts (percentages) and analyzed using the Chi-square test, while continuous variables were expressed as mean ± standard deviation and analyzed using the student's t-test, Mann–Whitney U test, or the Kruskal-Wallis test, as appropriate. We considered $p < 0·05$ as statistically significant. The performance evaluation of the model as well as the laryngologists utilized accuracy (Acc.) as well as precision (Pre.), recall (Rec.), and F1-score (F1.) with the Macro Average aggregation method. Bootstrapping was used to compute all 95% CIs. We also calculated the Mean Squared Error (MSE) of both the model and the laryngologists relative to the true labels to compare their performance. All statistical analyses were performed using Python (version 3.12.1) and R software (version 4.3.2).

Role of funding source

The funder of this study was involved in the study design, data collection, data analysis, and manuscript preparation. All authors had access to the dataset.

Result

Summary of clinical characteristics

The detailed characteristics of participants in different datasets, including age, sex, and the distribution of the labels, are presented in **Table 1**. It can be seen from the table that except for the sex characteristic, significant differences were observed in all characteristics between the five different datasets. This indicates that we need a model with a strong generalization ability in order to achieve good predictive performance across different datasets.

Model Performance

In this section, we discuss the model's performance in the three-class classification task (NORM, NCL, CA) on the internal validation set. We first compared the model developed in this study with other typical artificial intelligence algorithms on the internal validation set, including ResNet[17], Vit[18], Swin Transformer[19] , and CLIP[20]. For models that cannot accept text input directly, we utilized BERT[21] to incorporate textual modality information. The experimental results can be found in **Table 2**. Our multimodal interaction approach achieves an accuracy of 91.67%, a precision of 91.67%, a recall of 91·87% and a f1-score of 91·82% in overall performance, surpassing other models.

The architecture proposed in this study is based on the regression task of sample risk score. As the quality of the model's predictions for sample risk estimation is crucial, from a qualitative perspective, we plotted the histogram of the predicted risk estimations on the internal validation set (refer to **Fig. 4**) and calculated the mean and variance of the predicted risk estimations for each class (refer to **Table 3**).

To assess model performance from a quantitative perspective, we conducted Wilcoxon rank-sum test and t-tests between every two adjacent risk classes. The experimental results for these tests are presented in **Table 4**.

It can be observed that for the cancerous and normal classes, the predicted risk scores of most model samples are concentrated around the class baselines. Moreover, the statistical tests indicate significant separation between adjacent classes around normal and cancer classifications, demonstrating a relatively high level of regression accuracy.

Clinician-machine comparison

In order to assess the diagnostic performance of our model in clinical applications, we compared the performance of our model with that of 10 laryngologists of varying qualifications on internal and external validation sets.

First, for the three-class classification task (NORM, NCL, CA), after plotting the confusion matrices of the model on each validation set (**Fig. 5**), we computed our model's accuracy, precision, recall and F1-score against the laryngologists. The results are shown in **Table 5**. The model surpasses that of the expert group across all datasets (0·20%-5·44% higher) on the precision metric. Although the model performs slightly worse on one external validation set (Center 2), its performance on the other two external validation sets (Center 3 and Center 4) surpasses that of senior laryngologists.

Finally, we evaluated our model's performance in the comprehensive five-category classification task (NORM, BL, LG, HG, and CA). The risk baseline scores for these categories were set at 0.0 (NORM), 1.0 (BL), 1.1 (LG), 1.3 (HG), and 2.0 (CA), respectively. To assess accuracy, we first adjusted both the model's output and the laryngologists' risk scores to their closest predefined baseline values. Subsequently, we calculated the mean squared error (MSE) between the adjusted risk scores and the ground truth labels. The results are presented in Table S6. The overall performance of the model significantly surpasses that of juniors (0·101-0·204 lower), reaching an overall level comparable to that of seniors, who have over three years of experience.

Overall, the model shows marginally better performance than senior clinicians in three-class classification, while demonstrating comparable capability to senior clinicians in more comprehensive five-class classification tasks.

Model study

To assess the extent to which information from different modalities can provide diagnostic references and to identify which modality is most effective in diagnosis, we conducted ablation experiments on multimodal information, with detailed findings presented in Supplementary A3. Overall, in the context of LHP lesions, images contain the most critical information, while patient complaint texts and characteristics serve as secondary supportive information that enhances diagnostic accuracy.

Discussion

The diagnosis of LHP cancer has long been a challenging problem in the academic community. A review of past literature indicates that most diagnostic algorithms have primarily focused on laryngoscopic image analysis (see Table S7). This study represents a pioneering effort to integrate outpatient history and clinical features onto laryngoscopic

images devoid of lesion annotation, with the objective of developing a multiclassification prediction model for LHP diseases. The diagnostic efficacy of this approach was validated by multicenter data and compared with different levels of specialists. Compared to other methods, our algorithm achieves commendable diagnostic accuracy (see **Table 2**).

Recently, multimodal methods have been employed for laryngoscopic image prediction[22,23]. However, both methods have only integrated two distinct image modalities, namely white light and narrow-band imaging (NBI), with the requirement of annotated information on the lesion sites. In contrast, we have fused image, text and feature information in order to more realistically reproduce the process of patient consultation. This additional information has been demonstrated to be beneficial in ablation experiments (see Table S5), which is a significant factor contributing to our ability to achieve high diagnostic efficacy on the three-classification task with only two thousand cases.

Although NBI has demonstrated efficacy in detecting precancerous lesions and delineating surgical margins, and prior machine learning studies have shown promising results,[22–24] its high equipment costs and specialized training requirements limit accessibility, particularly in resource-constrained settings. In contrast, our multimodal model using white light examination offers cost-effective generalizability, overcoming these barriers. Furthermore, unlike previous studies focused solely on vocal fold lesions[25,26], our investigation included supraglottic and hypopharyngeal lesions, expanding clinical applicability to anatomically complex regions. Despite adapting to this more challenging diagnostic environment, the model maintained robust accuracy across three external datasets with varying image qualities and achieved performance that surpasses clinicians with 3-10 years of experience. This underscores its practical utility in diverse clinical workflows, particularly where advanced imaging technologies or subspecialty expertise are unavailable.

There is a distinction between benign and varying grades of SILs within NCL; however, this distinction is relatively minor in comparison to the differences between themselves, CA and NORM groups (see **Fig. 4** and **Table 4**). To more effectively leverage this nonlinear differentiation in the multiclassification task, we originally devised the scoring method. In comparison to the conventional classification approach, the scoring method enhances the accuracy of the model (Scoring = 88·91% vs. Classifier = 87·32%, see Table S4).

Moreover, the question of whether squamous epithelial hyperplasia without atypia constitutes a precancerous lesion remains a point of controversy within the context of the WHO classification criteria.[14] Without considering the prognosis, we demonstrate that squamous epithelial hyperplasia without atypia is more similar to low-grade precancerous lesions from a novel perspective of machine learning (accuracy of SILs = 88·03% vs. Dysplasia = 83·72%, see Table S1).

To delve deeper into the features attended to by the model during prediction, we have generated attention heatmaps illustrating the features extracted from both image and text modalities. The experimental results are depicted in **Fig. 6**.

In the normal group, the attention was more dispersed, with a tendency to focus on specific anatomical structures, including the rima glottidis, arytenoid cartilage, ventricular cord, and tubercle of the epiglottis. Concurrently, some pharyngeal secretions that were not directly pertinent to the diagnosis but frequently observed during examination were successfully disregarded (see **Fig. 6**, sample (a)). It seems that deep learning is more effective than humans at discerning the nature of a lesion from a relatively blurry picture (see **Fig. 6**, sample (b)). Such images might be excluded in other studies[22,27]. Moreover, the model is less likely to be misled by prior precancerous pathology or biopsy results documented in the patient's outpatient history (see **Fig. 6**, sample (c)). In clinical practice, physicians might assume a recurrent lesion remains non-cancerous based on past reports, whereas in this case, our model correctly predicted malignancy despite the patient's history suggesting a precancerous condition. This indicates that the model developed in this study exhibits a

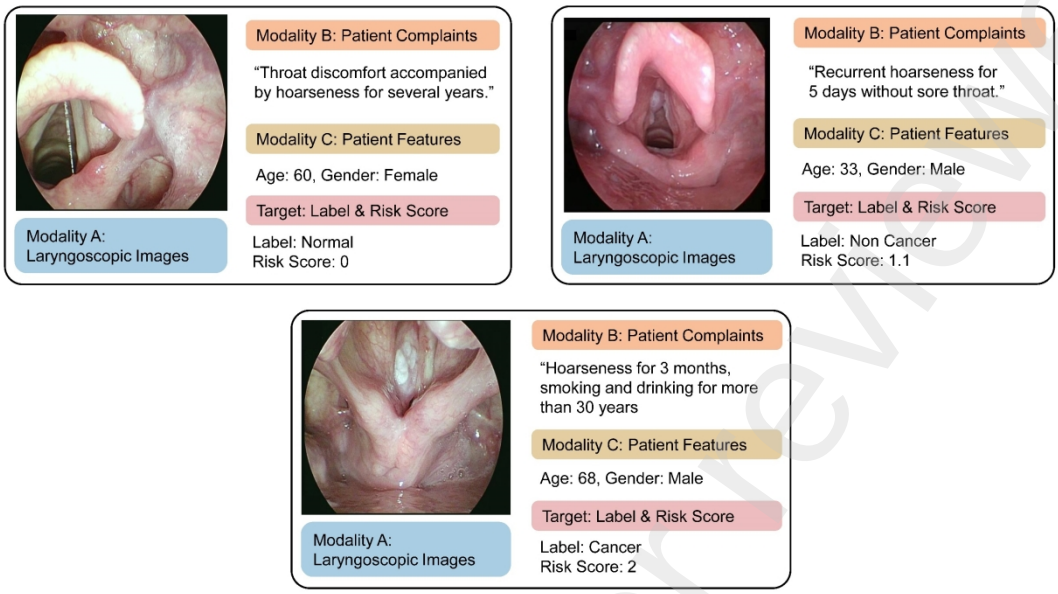
certain level of robustness against interference.

Our study has several limitations. First, while multi-centered, the predominantly Asian cohort may limit generalizability to other ethnicities or regions with differing disease patterns. Second, insufficient sample sizes across disease stages—e.g., pharyngeal reflux-associated laryngeal granuloma (see **Fig. 6**, sample (d)), frequently seen clinically but underrepresented here due to low surgical rates[28]—restricted the model's access to comprehensive data, risking misclassification. Additionally, variability in patient-reported data formats and imaging angles across centers (e.g., lesions partially outside the field of view) constrained predictive accuracy, notably in external validation sets (see **Fig. 6**, sample (e-f)), potentially explaining their weaker performance.

In summary, this study pioneered the integration of outpatient history and clinical features onto laryngoscopic images, enhancing diagnostic comprehensiveness by synthesizing multimodal clinical data, enabling the model to achieve superior performance over clinicians with 3-10 years of experience in multiclass classification of LHP disorders. This advancement addresses the critical need for accessible tools in settings lacking subspecialty expertise, reducing diagnostic uncertainty in anatomically complex regions. Future studies will standardize the data while expanding the sample to ensure equitable applicability across diverse patient demographics, enhance generalizability, and optimize the model's performance for seamless integration into routine clinical workflows.

Tables and Figures

Fig. 1: Multimodal patient sample example. Three modalities are involved, which are patient features, patient complaints, and laryngoscopic images.



[a] Three modalities are involved, which are patient features, patient complaints, and laryngoscopic images.

Fig. 2: Data collection and partitioning flowchart.

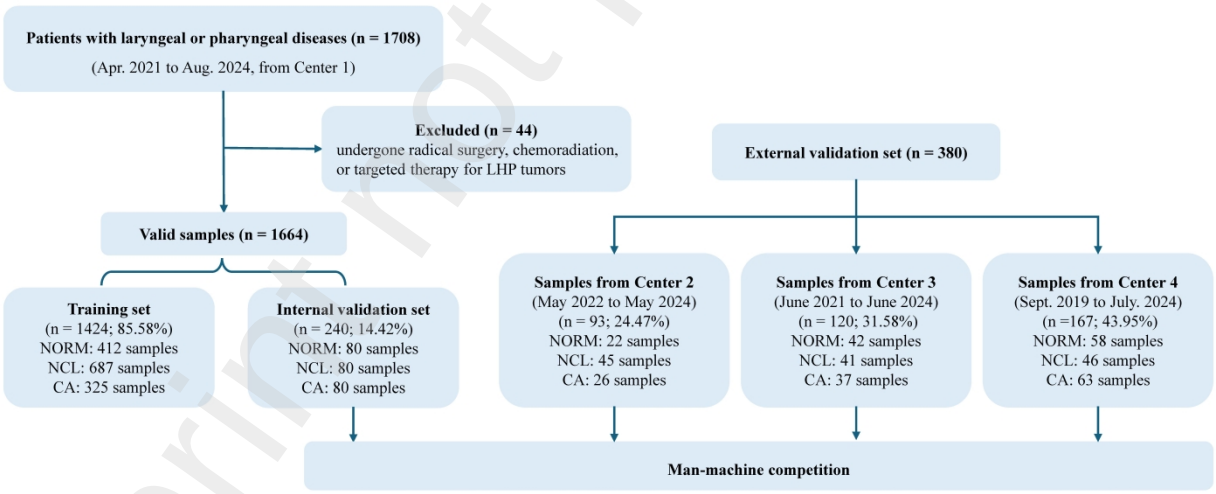


Fig. 3: Multimodal patient information interaction model structure.

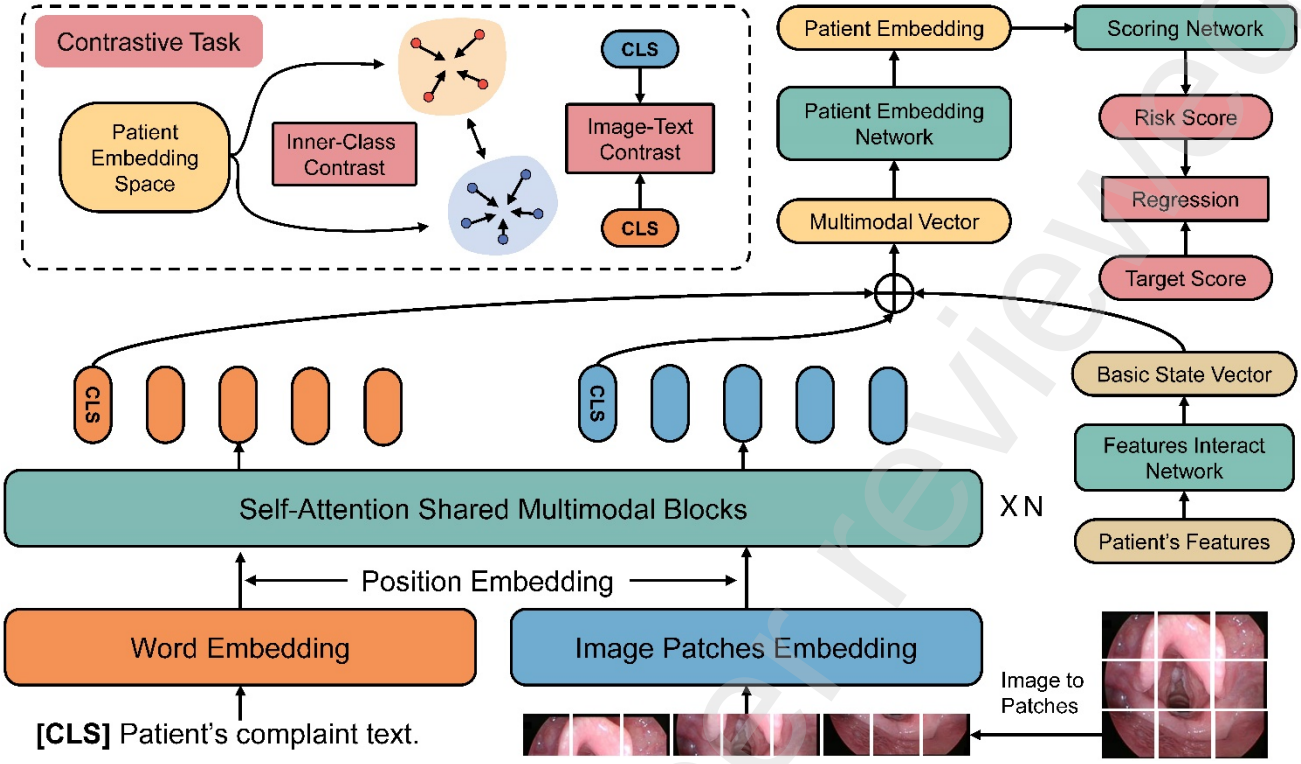


Fig. 4: Distribution of predicted risk scores for each category of the model on the internal validation set.

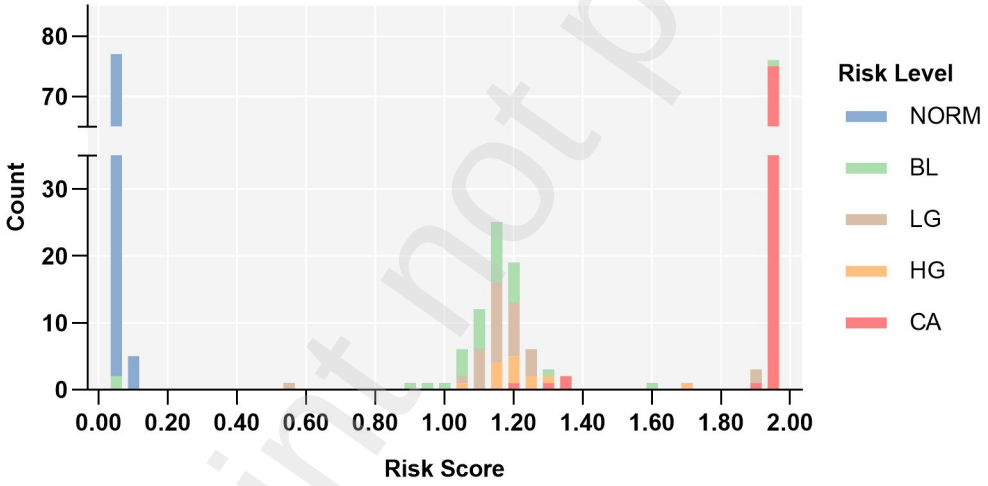


Fig. 5: Confusion matrices of the model on the internal validation set and three external validation sets.

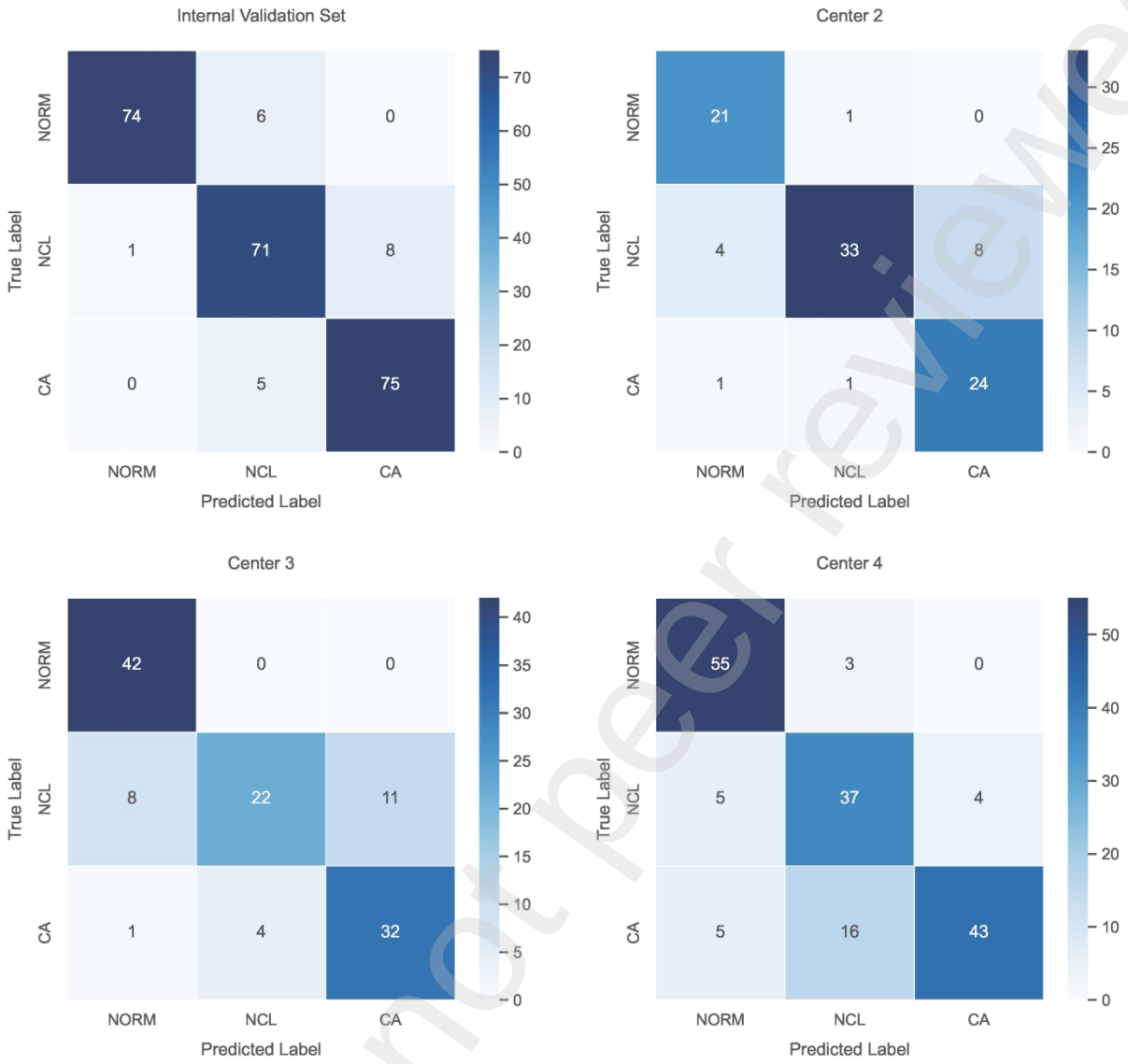
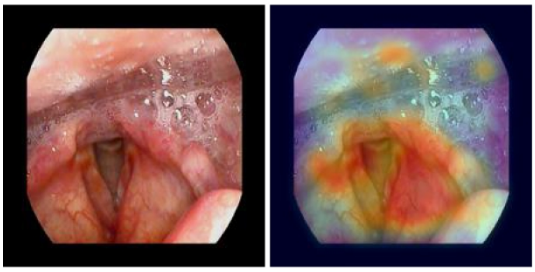


Fig. 6: Attention visualization of feature extraction in multimodal samples.



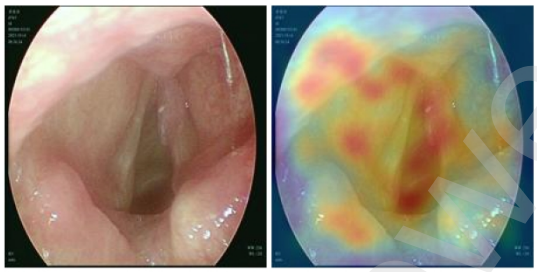
(a) **Center:** LH, **Label:** NCL, **Sex:** Male, **Age:** 50 yr.

Complaint Text:

Due to enlarged lymph node, patient was admitted to local hospital . Laryngoscopy showed leukoplakia of the right vocal cord . Patient had a history of smoking : 1 pack of cigarettes per day , for more than 30 years .

Risk Score Evaluations:

True score: 1.1
Model prediction score: 1.21
Expert group predictions - A: 1.1, B: 1.1, C: 2.0, D: 2.0
Senior group predictions - A: 1.1, B: 1.3, C: 1.1
Junior group predictions - A: 1.1, B: 1.0, C: 1.3



(b) **Center:** FAHZU, **Label:** CA, **Sex:** Male, **Age:** 77 yr.

Complaint Text:

Hoarseness for several months , no pain , no itching , little coughing , eating well , spitting a little white sputum . Foreign body sensation .

Risk Score Evaluations:

True Score: 2.0
Model prediction score: 1.97
Expert group predictions - A: 2.0, B: 1.3, C: 1.3, D: 2.0
Senior group predictions - A: 1.1, B: 1.1, C: 1.3
Junior group predictions - A: 1.3, B: 1.3, C: 1.3

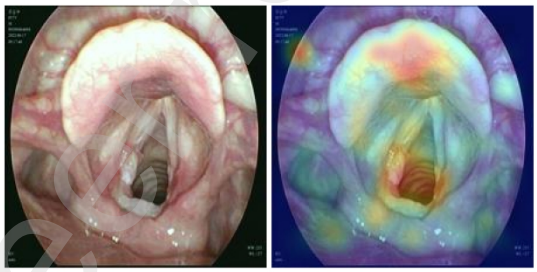
(c) **Center:** LH, **Label:** CA, **Sex:** Male, **Age:** 69 yr.

Complaint Text:

The patient had hoarseness due to excessive use of voice 4 years ago , and had undergone vocal cord tumor resection , with postoperative pathology showing leukoplakia . The hoarseness was relieved after the operation . 1 month ago, the patient had hoarseness without obvious cause , with no sore throat , fever , cough, sputum , blood in sputum , and difficulty in breathing and swallowing , and denied any smoking and drinking history.

Risk Score Evaluations:

True score: 2.0
Model prediction score: 1.97
Expert group predictions - A: 2.0, B: 2.0, C: 1.3, D: 2.0
Senior group predictions - A: 1.3, B: 2.0, C: 2.0
Junior group predictions - A: 1.3, B: 1.3, C: 1.3



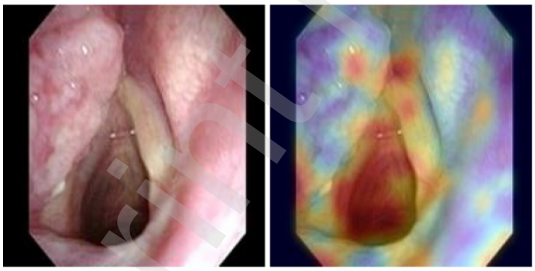
(d) **Center:** FAHZU, **Label:** NCL, **Sex:** Male, **Age:** 58 yr.

Complaint Text:

Recurrent hoarseness , acid reflux , never smoked , drank 2 catties of rice wine per day for more than 20 years . Local laryngoscope found neoplasm in vocal cords . History of atrial fibrillation.

Risk Score Evaluations:

True score: 1.1
Model prediction score: 1.96
Expert group predictions - A: 1.1, B: 1.1, C: 1.3, D: 1.1
Senior group predictions - A: 1.3, B: 1.3, C: 1.0
Junior group predictions - A: 1.1, B: 1.4, C: 1.3



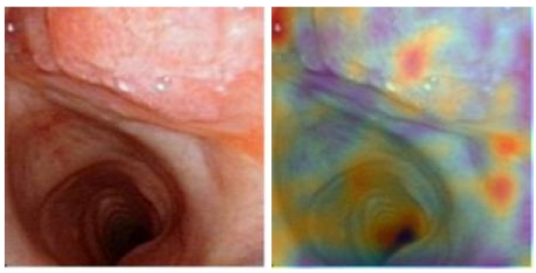
(e) **Center:** FAHWMU, **Label:** CA, **Sex:** Male, **Age:** 61 yr.

Complaint Text:

A left- sided neck mass was found 1 month ago , and is now present with mild sore throat and no hemoptysis.

Risk Score Evaluations:

True score: 2.0
Model prediction score: 0.04
Expert group predictions - A: 2.0, B: 2.0, C: 2.0, D: 2.0
Senior group predictions - A: 2.0, B: 2.0, C: 2.0
Junior group predictions - A: 2.0, B: 2.0, C: 2.0



(f) **Center:** JSH, **Label:** CA, **Sex:** Male, **Age:** 64 yr.

Complaint Text:

A right sided neck mass was found for more than 2 months.

Risk Score Evaluations:

True score: 2.0
Model prediction score: 0.04
Expert group predictions - A: 2.0, B: 2.0, C: 2.0, D: 2.0
Senior group predictions - A: 2.0, B: 0.0, C: 2.0
Junior group predictions - A: 0.0, B: 1.1, C: 1.5

[a] The text combined with the two images above it, forms a sample group. The image on the left is the original image, while the one on the right is the model's attention heatmap. The colors in the illustration represent the degree of attention,

with a higher degree of attention indicated by a hotter (redder) color. The letters (A, B, C, D) in risk score evaluations represent different laryngologists within each group. The scores highlighted in red, provided by the model or the laryngologists, indicate classification errors made by the model or the laryngologists in the three-class classification task.

Table 1: Basic Characteristics of Participants

Characteristic	Training Set (n=1424)	Internal Validation Set (n= 240)	Center 2 (n=93)	Center 3 (n=120)	Center 4 (n=167)	P Value
Age	54·94±13·74	57·73±14·30	56·89 ±13·85	56·44±13·85	57·66 ±12·52	0·0040
Sex						0·19
Male	1020(71·63%)	178(74·17%)	72(77·42%)	94(78·33%)	130(77·84%)	
Female	404(28·37%)	62(25·83%)	21(22·58%)	26(21·67%)	37(22·16%)	
Label						<0·001
NORM	412(28·93%)	80(33·33%)	22(23·66%)	42(35·00%)	58(34·73%)	
NCL	687(48·24%)	80(33·33%)	45(48·39%)	41(34·17%)	46(27·54%)	
CA	325(22·82%)	80(33·33%)	26(27·96%)	37(30·83%)	63(37·72%)	

NORM = Normal group, NCL = Non-cancerous lesions, CA = Cancer group

Table 2: Comparison of performance of models of different modalities and scales on data sets

Model	Train Set				Internal Validation Set			
	Acc.	Pre.	Rec.	F1.	Acc.	Pre.	Rec.	F1.
Vision Only								
Resnet50	96·86	95·95	97·23	96·55	79·17	80·38	79·17	79·05
Resnet101	95·28	94·25	95·86	94·96	81·25	82·67	81·25	81·48
ViT-Base	96·02	95·32	96·00	95·60	81·34	82·45	81·25	81·52
ViT-Large	96·12	95·30	95·93	95·59	82·50	83·05	82·50	82·65
SwinT	95·60	94·84	95·33	95·06	84·17	86·27	84·17	84·32
CLIP-Vision	95·39	94·86	94·65	94·75	84·48	84·46	84·37	84·12
Vision-Language								
Resnet50 + BERT	96·65	95·89	97·02	96·42	81·27	84·96	81·25	81·49
Resnet101 + BERT	99·79	99·76	99·76	99·77	82·08	83·32	82·08	82·21
ViT-Large + BERT	98·32	97·79	98·48	98·13	84·58	86·49	84·58	84·85
SwinT + BERT	98·89	96·35	98·43	98·02	85·83	86·26	85·83	85·96
CLIP	99·81	99·22	99·58	98·89	86·67	86·91	86·67	86·74
Visual-Textual interaction and Patient information								
MFDN (ours)	99·90	99·94	99·87	99·90	**91·67**	**91·67**	**91·87**	**91·82**

CLIP = Contrastive language-image pretraining, Acc. = Accuracy, Pre. = Precision, Rec, = Recall, F1. = F1-score, MFDN = Multimodal Fusion Deep Network

Table 3: The average score and standard deviation of each type of model prediction risk score on the internal validation set.

	NORM	BL	LG	HG	CA
Score	0.11±0.29	1.17±0.36	1.22±0.22	1.26±0.23	1.91±0.36

NORM = Normal group, BL = Benign lesion, LG = Low grade, HG = High grade, CA = Cancer group

Table 4: Statistical tests of model predicted risk scores on the internal validation set.

Alternative Hypothesis	Sum Rank		T-Test	
	U-value	P-Value	T-Value	P-Value
NORM< BLL	-7.78	<0.0001	-14.88	<0.0001
BLL<LG	-2.21	0.01	-0.64	0.26
LG<HG	-0.81	0.21	-0.46	0.33
HG<CA	-5.32	<0.0001	-9.62	<0.0001

NORM = Normal group, BL = Benign lesion, LG = Low grade, HG = High grade, CA = Cancer group

Table 5: Comparison of performance of our model against laryngologists on the three-class classification task.

Datasets	Metrics	Expert	Senior	Junior	MFDN (ours) [a]
Internal Validation Set	Acc.	86·24 [81·67, 90·79]	83·06 [78·33, 87·50]	77·44 [74·17, 84·58]	**91·67** [88·33, 95·00]
	Pre.	86·23 [81·71, 90·74]	83·05 [78·69, 87·34]	79·44 [74·37, 84·15]	**91·67** [88·15, 94·98]
	Rec.	87·49 [83·28, 91·34]	84·61 [80·12, 88·61]	83·04 [78·57, 87·28]	**91·87** [88·31, 95·15]
	F1.	86·50 [81·94, 90·84]	82·92 [78·14, 87·29]	79·46 [74·12, 84·41]	**91·82** [88·08, 95·03]
Center 2	Acc.	86·22 [79·35, 92·47]	85·30 [77·42, 92·47]	80·65 [72·04, 88·20]	*83·87* [76·34, 91·40]
	Pre.	85·14 [76·90, 92·38]	84·49 [76·88, 91·80]	77·82 [69·20, 85·76]	**87·03** [80·27, 93·21]
	Rec.	88·43 [82·12, 94·17]	86·90 [79·20, 93·74]	85·05 [76·69, 92·60]	83·35 [75·19, 90·77]
	F1.	86·55 [78·79, 92·80]	85·28 [76·97, 92·18]	79·60 [70·36, 87·60]	*82·45* [76·17, 91·25]
Center 3	Acc.	80·42 [73·33, 87·50]	77·50 [70·00, 85·00]	73·61 [65·83, 80·83]	<u>80·00</u> [72·50, 86·67]
	Pre.	79·85 [72·83, 86·63]	76·68 [69·40, 83·74]	72·21 [65·66, 78·64]	**80·05** [73·70, 85·65]
	Rec.	82·74 [75·99, 88·77]	80·90 [73·68, 87·85]	81·78 [75·10, 87·40]	80·46 [72·95, 87·46]
	F1.	80·28 [72·86, 86·96]	76·86 [68·88, 84·10]	70·82 [63·02, 78·49]	<u>78·66</u> [70·73, 85·18]

Center 4	Acc.	77·53 [70·83, 83·93]	74·60 [67·86, 80·97]	69·84 [62·50, 76·19]	**80·36** [74·39, 86·31]
	Pre.	78·07 [71·31, 84·25]	75·91 [69·65, 81·87]	71·85 [65·80, 77·67]	**80·82** [75·00, 86·37]
	Rec.	79·58 [73·74, 85·10]	77·71 [71·41, 83·27]	75·69 [69·59, 81·28]	**80·73** [74·98, 86·06]
	F1.	77·58 [70·95, 83·63]	74·29 [67·41, 80·70]	69·48 [62·47, 76·28]	**79·82** [73·49, 85·44]

Data in brackets are 95% CIs. Acc. = Accuracy, Pre. = Precision, Rec, = Recall, F1. = F1-score, CI = Confidence Interval

[a] **Bold**, underline and *italics* respectively indicate that the model's performance surpasses that of expert, senior, junior laryngologists.

References

1 Echanique KA, Evans LK, Han AY, Chhetri DK, John MAS. Cancer of the Larynx and Hypopharynx. *Hematology/Oncology Clinics* 2021;**35**:933–47.

2 Zhu J, Li B, Hu JJ, et al. Undifferentiated Small Round Cell Sarcoma of the Postcricoid Region of the Hypopharynx: A Rare Case Report and Review of the Literature. *OTT* 2021;**Volume 14**:4537–44.

3 Bray F, Laversanne M, Sung H, et al. Global cancer statistics 2022: GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries. *CA A Cancer J Clinicians* 2024;**74**:229–63.

4 Newman JR, Connolly TM, Illing EA, et al. Survival trends in Hypopharyngeal cancer: A population-based review: Survival Trends in Hypopharyngeal Cancer. *The Laryngoscope* 2015;**125**:624–9.

5 Hall SF, Groome PA, Irish J, O'Sullivan B. The Natural History of Patients With Squamous Cell Carcinoma of the Hypopharynx. *The Laryngoscope* 2008;**118**:1362–71.

6 Wang S, Chen Y, Chen S, Zhong Q, Zhang K. Hierarchical dynamic convolutional neural network for laryngeal disease classification. *Scientific Reports* 2022;**12**:13914.

7 Haroz EE, Rebman P, Goklish N, et al. Performance of Machine Learning Suicide Risk Models in an American Indian Population. *JAMA Netw Open* 2024;**7**:e2439269.

8 Shu Q, Pang J, Liu Z, et al. Artificial Intelligence for Early Detection of Pediatric Eye Diseases Using Mobile Photos. *JAMA Network Open* 2024;**7**:e2425124–e2425124.

9 Park H, Yun J, Lee SM, et al. Deep Learning–based Approach to Predict Pulmonary Function at Chest CT. *Radiology* 2023;**307**:e221488.

10 Du S, Guo J, Huang D, et al. Diagnostic accuracy of deep learning-based algorithms in laryngoscopy: a systematic review and meta-analysis. *Eur Arch Otorhinolaryngol* 2024.

11 Kang Y, Yang L, Hu Y, et al. Self-Attention Mechanisms-Based Laryngoscopy Image Classification Technique for Laryngeal Cancer Detection. *Head & Neck* 2024:hed.27999.

12 Baldini C, Azam MA, Sampieri C, et al. An automated approach for real-time informative frames classification in laryngeal endoscopy using deep learning. *Eur Arch Otorhinolaryngol* 2024;**281**:4255–64.

13 Żurek M, Jasak K, Niemczyk K, Rzepakowska A. Artificial Intelligence in Laryngeal Endoscopy: Systematic Review and Meta-Analysis. *JCM* 2022;**11**:2752.

14 Zidar N, Gale N. Update from the 5th Edition of the World Health Organization Classification of Head and Neck Tumors: Hypopharynx, Larynx, Trachea and Parapharyngeal Space. *Head and Neck Pathol* 2022;**16**:31–9.

15 Zhang Y, Jiang H, Miura Y, Manning CD, Langlotz CP. Contrastive Learning of Medical Visual Representations from Paired Images and Text. Proceedings of the 7th Machine Learning for Healthcare Conference, PMLR; 2022, p. 2–25.

16 Gunel B, Du J, Conneau A, Stoyanov V. Supervised Contrastive Learning for Pre-trained Language Model Fine-tuning 2021.

17 He K, Zhang X, Ren S, Sun J. Deep residual learning for image recognition. Proceedings of the IEEE conference on computer vision and pattern recognition, 2016, p. 770–8.

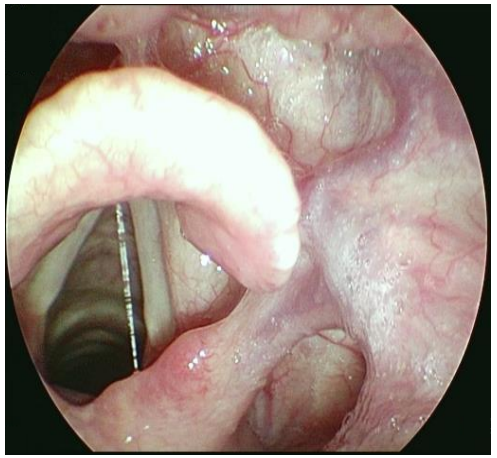
18 Alexey D. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv Preprint arXiv: 201011929* 2020.

19 Liu Z, Lin Y, Cao Y, et al. Swin transformer: Hierarchical vision transformer using shifted windows. Proceedings of the IEEE/CVF international conference on computer vision, 2021, p. 10012–22.

20 Radford A, Kim JW, Hallacy C, et al. Learning transferable visual models from natural language supervision. International conference on machine learning, PMLR; 2021, p. 8748–63.

21 Kenton JDM-WC, Toutanova LK. Bert: Pre-training of deep bidirectional transformers for language understanding. Proceedings of naacL-HLT, vol. 1, Minneapolis, Minnesota; 2019, p. 2.

22 Tie C, Li D, Zhu J, et al. Multi-Instance Learning for Vocal Fold Leukoplakia Diagnosis Using White Light and Narrow-Band Imaging: A Multicenter Study. *The Laryngoscope* 2024;**134**:4321–8.
23 Li Y, Gu W, Yue H, et al. Real-time detection of laryngopharyngeal cancer using an artificial intelligence-assisted system with multimodal data. *J Transl Med* 2023;**21**:698.
24 You Z, Han B, Shi Z, et al. Vocal cord leukoplakia classification using deep learning models in white light and narrow band imaging endoscopy images. *Head & Neck* 2023;**45**:3129–45.
25 Cho WK, Lee YJ, Joo HA, et al. Diagnostic Accuracies of Laryngeal Diseases Using a Convolutional Neural Network-Based Image Classification System. *The Laryngoscope* 2021;**131**:2558–66.
26 He Y, Cheng Y, Huang Z, et al. A deep convolutional neural network-based method for laryngeal squamous cell carcinoma diagnosis. *Annals of Translational Medicine* 2021;**9**.
27 Tran BA, Dao TTP, Dung HDQ, et al. Support of deep learning to classify vocal fold images in flexible laryngoscopy. *American Journal of Otolaryngology* 2023;**44**:103800.
28 Karkos PD, George M, Van Der Veen J, et al. Vocal Process Granulomas: A Systematic Review of Treatment. *Ann Otol Rhinol Laryngol* 2014;**123**:314–20.



Modality A:
Laryngoscopic Images

Modality B: Patient Complaints

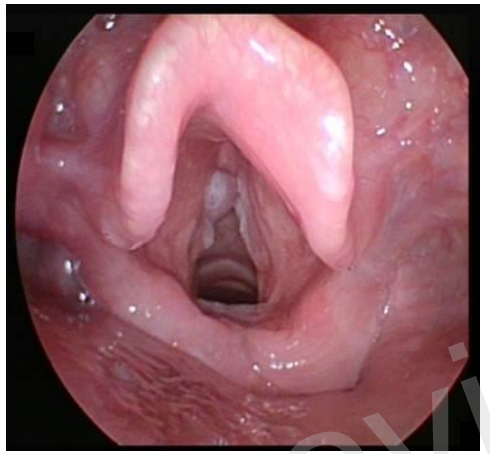
"Throat discomfort accompanied by hoarseness for several years."

Modality C: Patient Features

Age: 60, Gender: Female

Target: Label & Risk Score

Label: Normal
Risk Score: 0



Modality A:
Laryngoscopic Images

Modality B: Patient Complaints

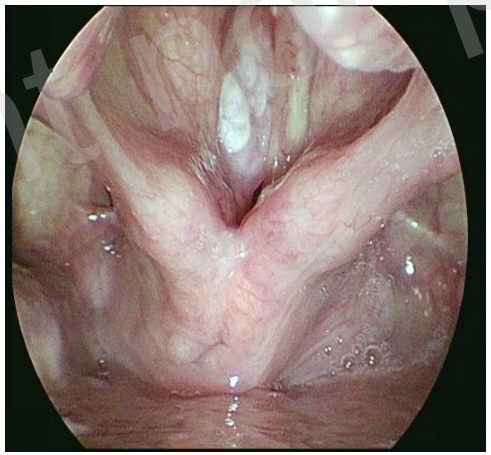
"Recurrent hoarseness for 5 days without sore throat."

Modality C: Patient Features

Age: 33, Gender: Male

Target: Label & Risk Score

Label: Non Cancer
Risk Score: 1.1



Modality A:
Laryngoscopic Images

Modality B: Patient Complaints

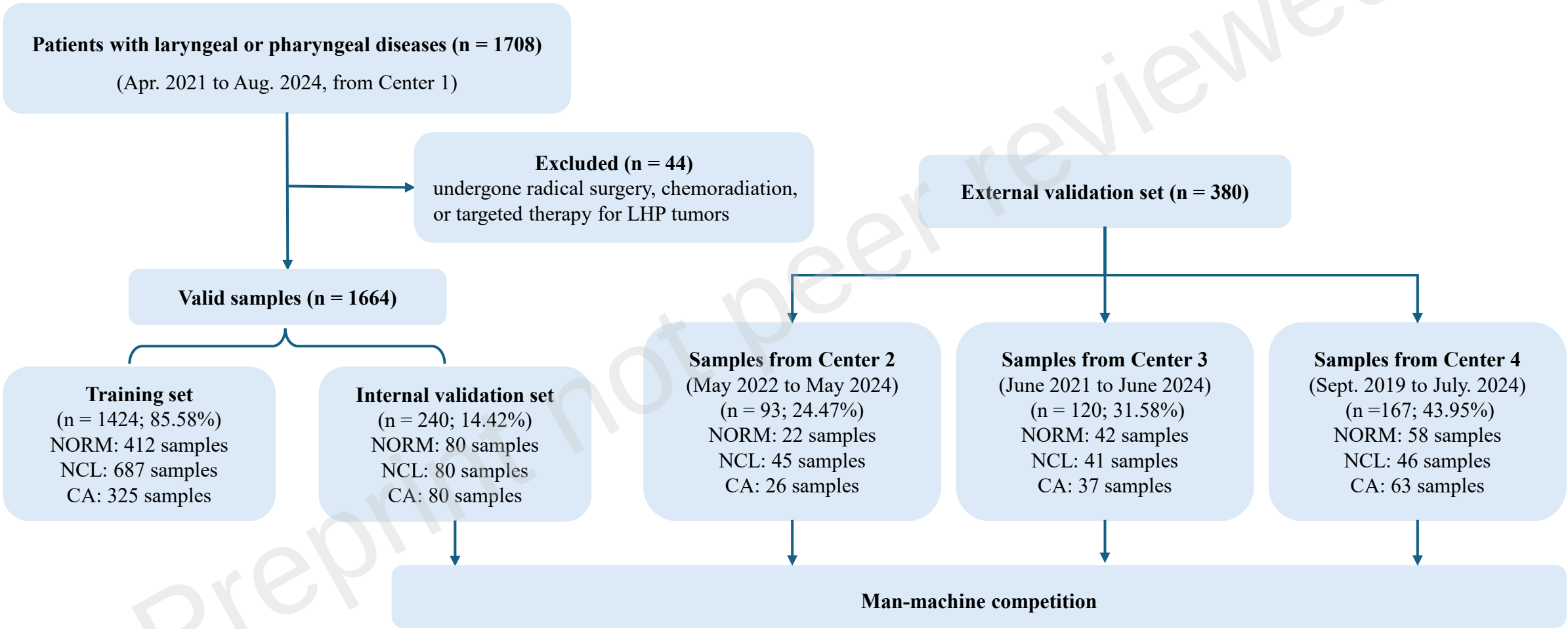
"Hoarseness for 3 months, smoking and drinking for more than 30 years

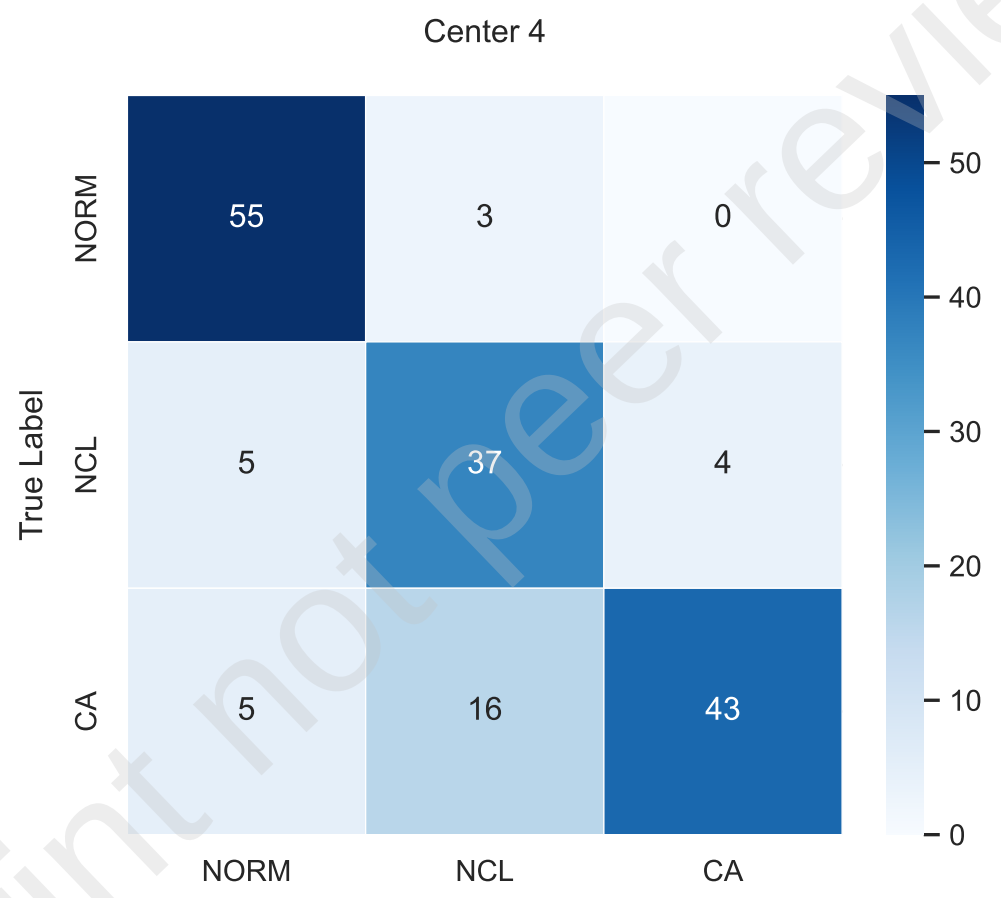
Modality C: Patient Features

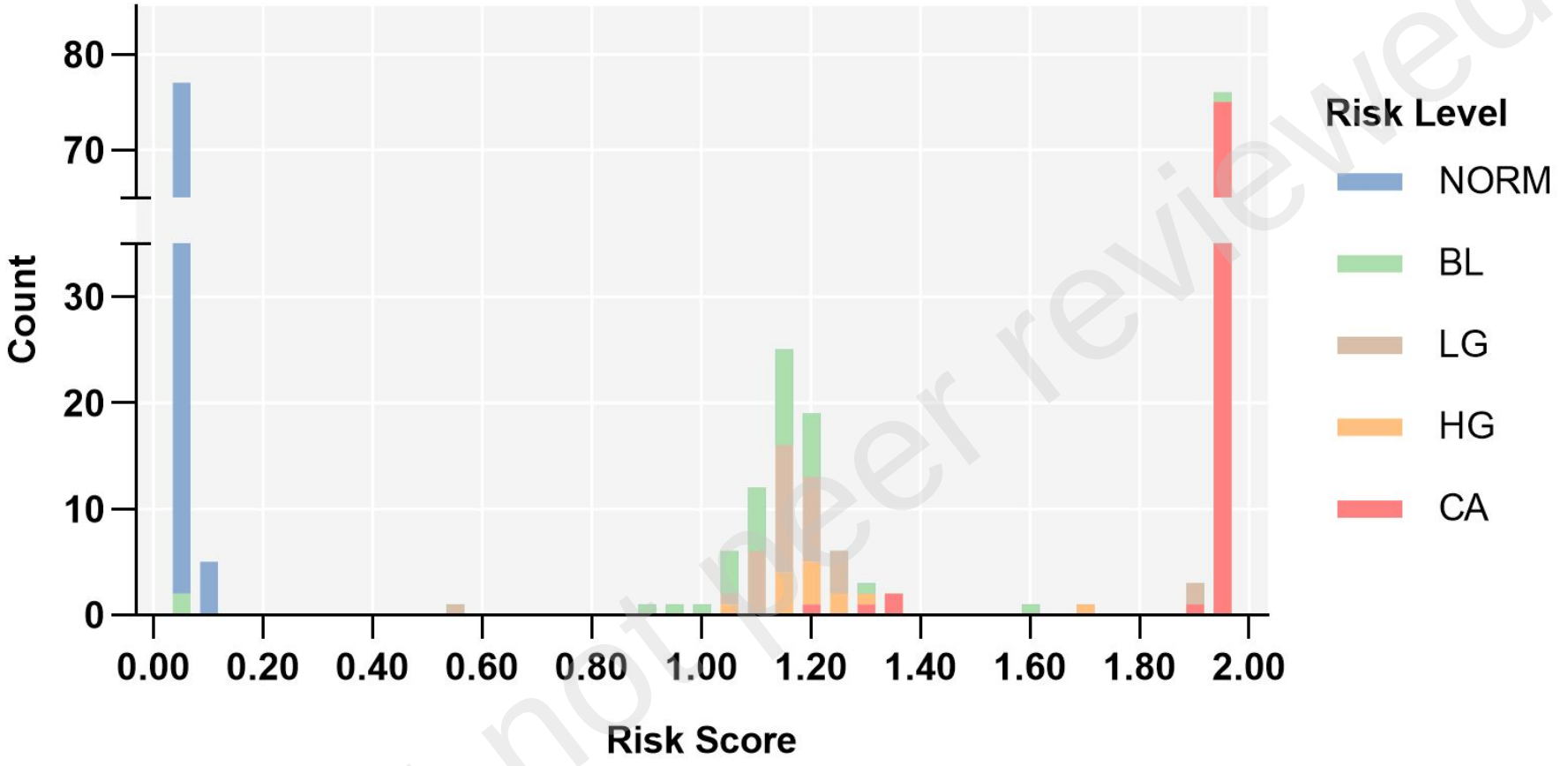
Age: 68, Gender: Male

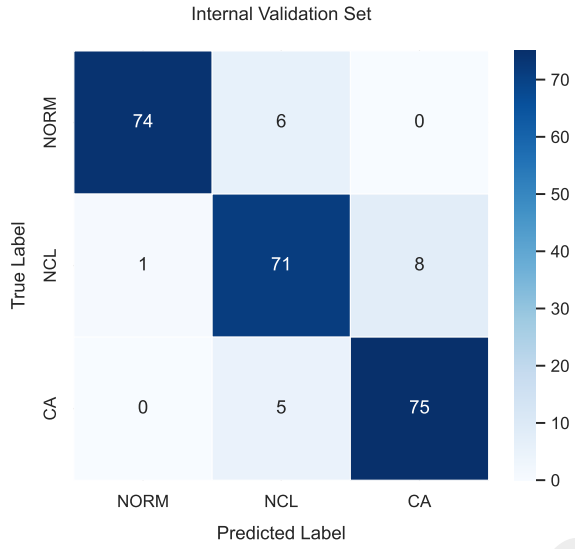
Target: Label & Risk Score

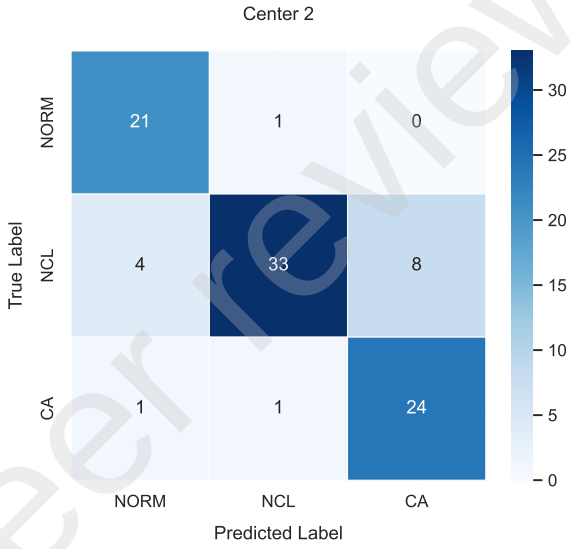
Label: Cancer
Risk Score: 2

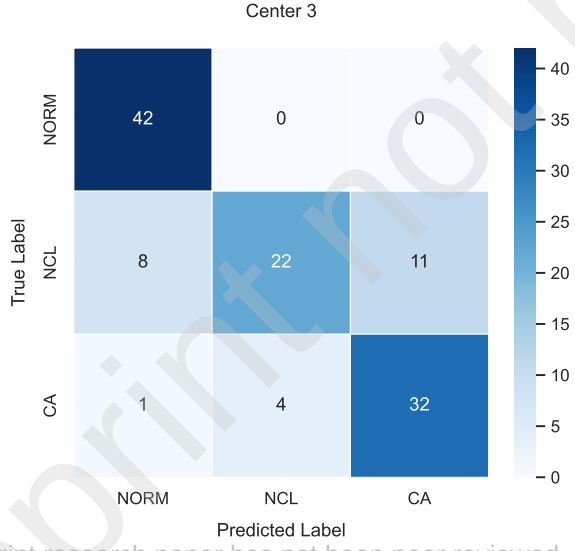


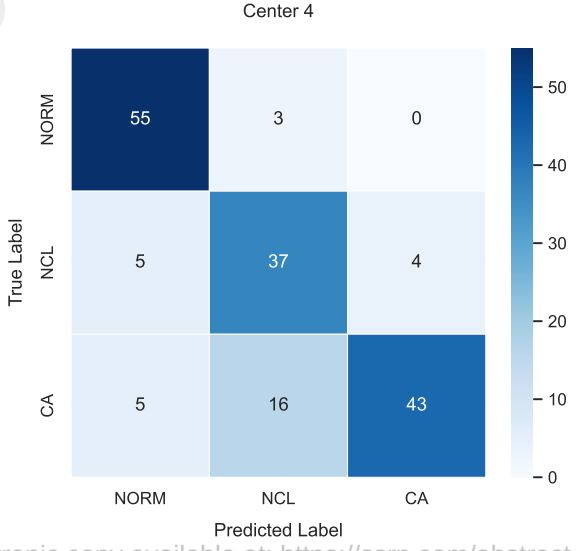


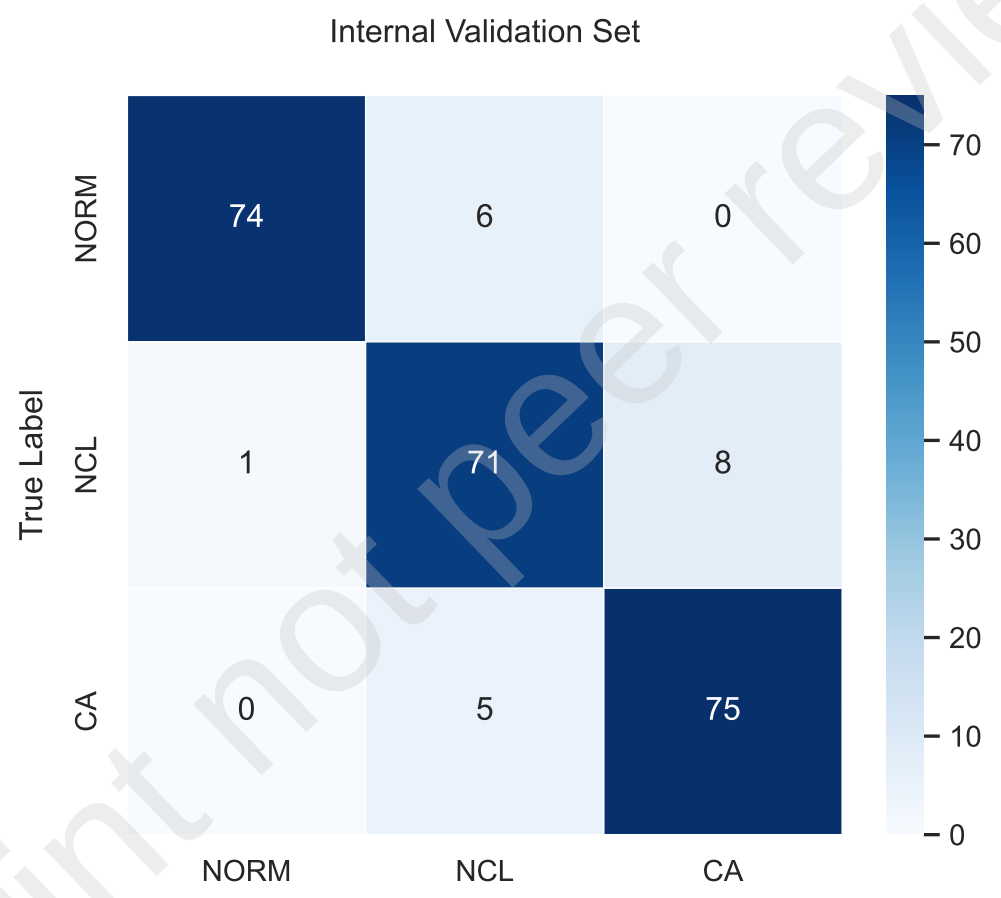


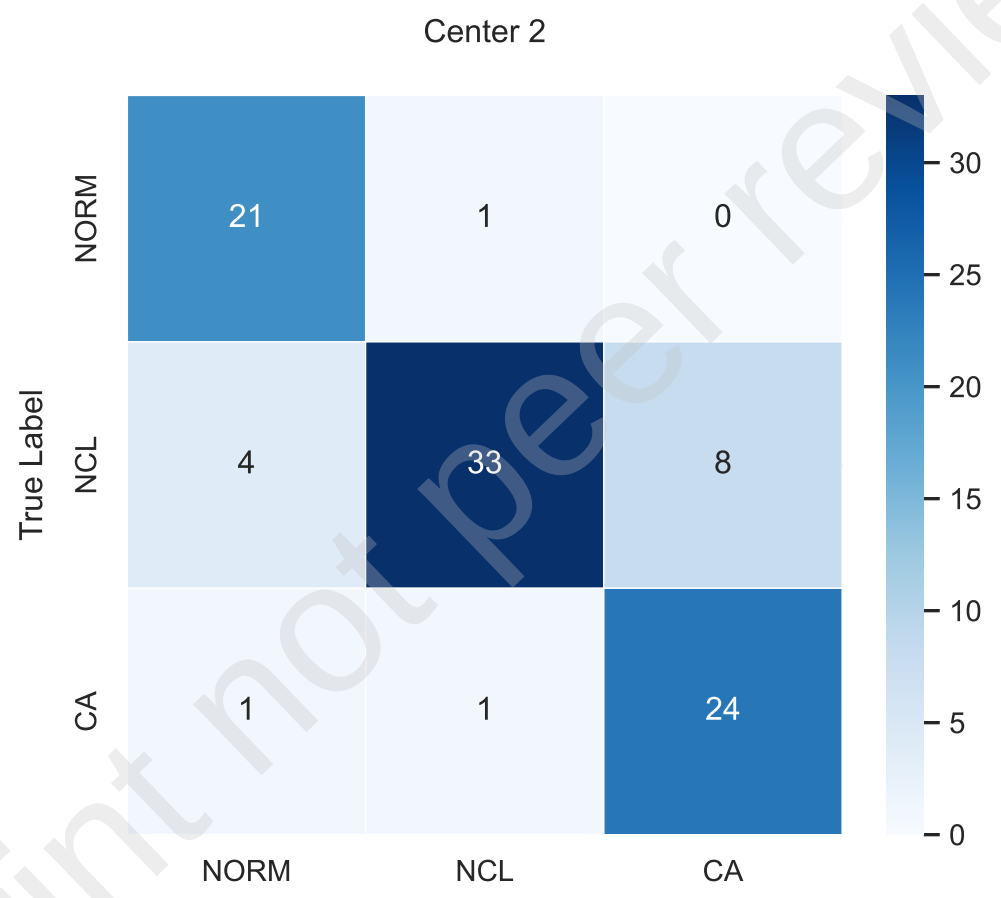


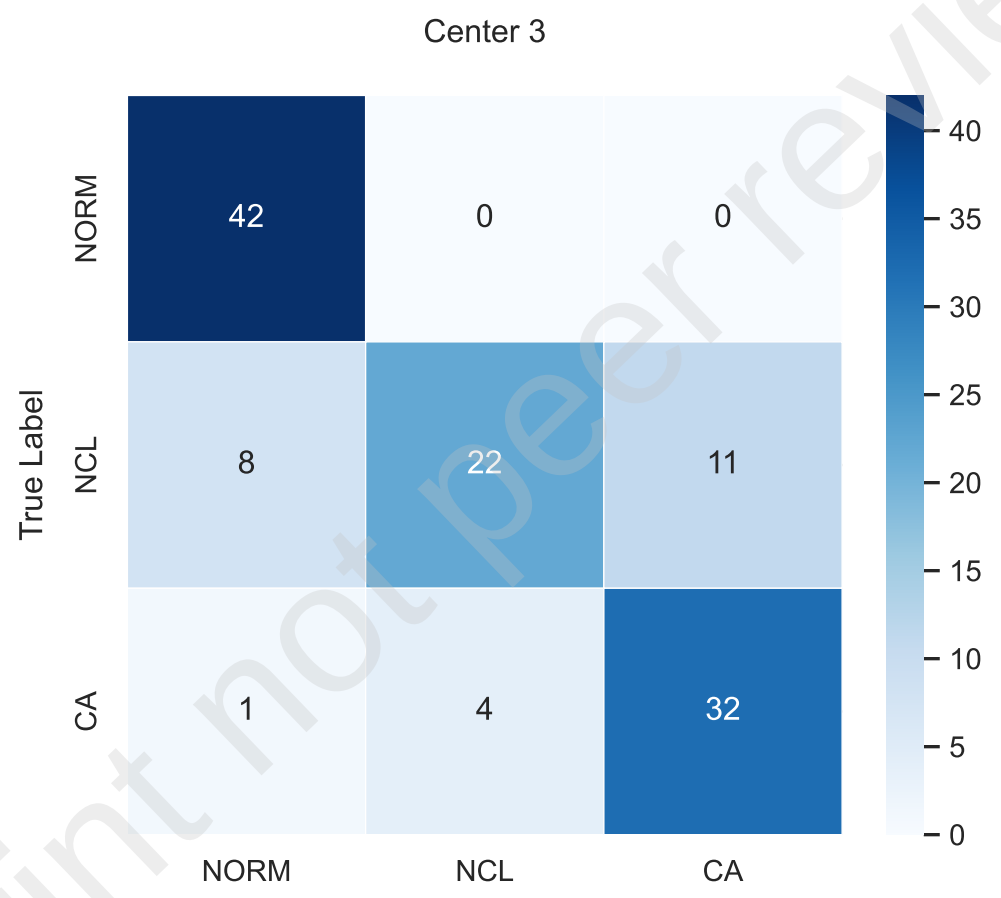


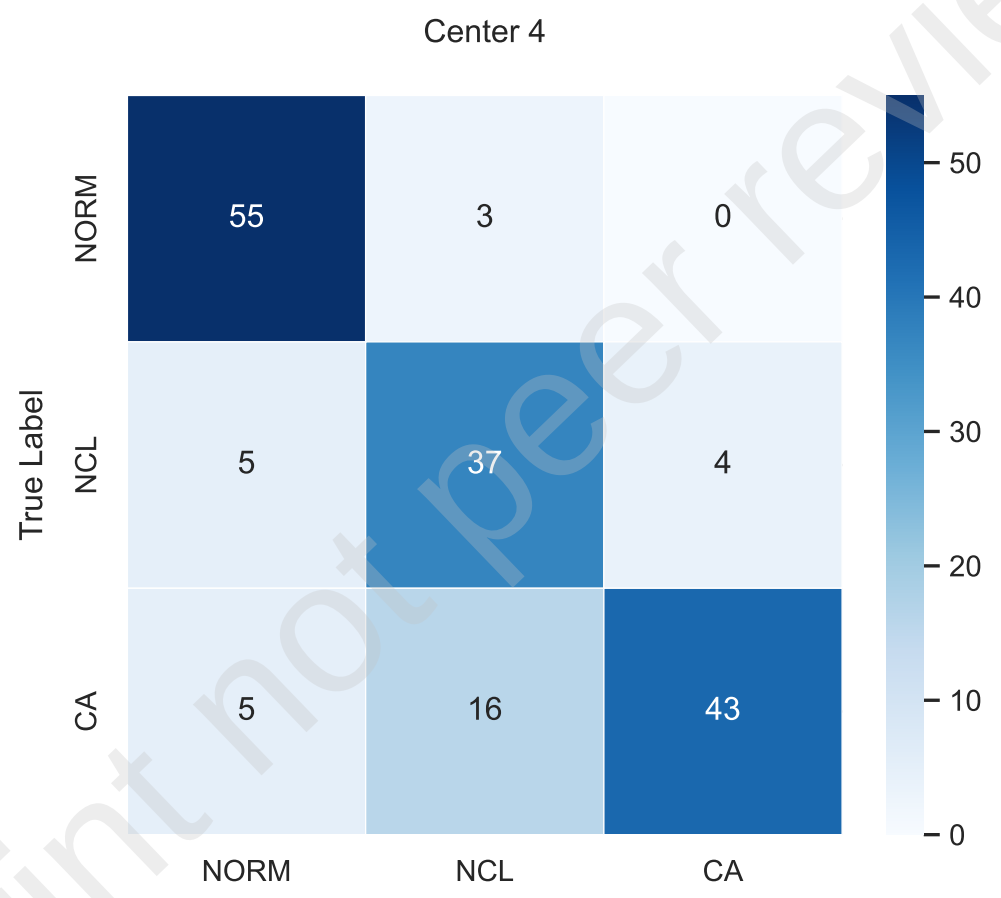


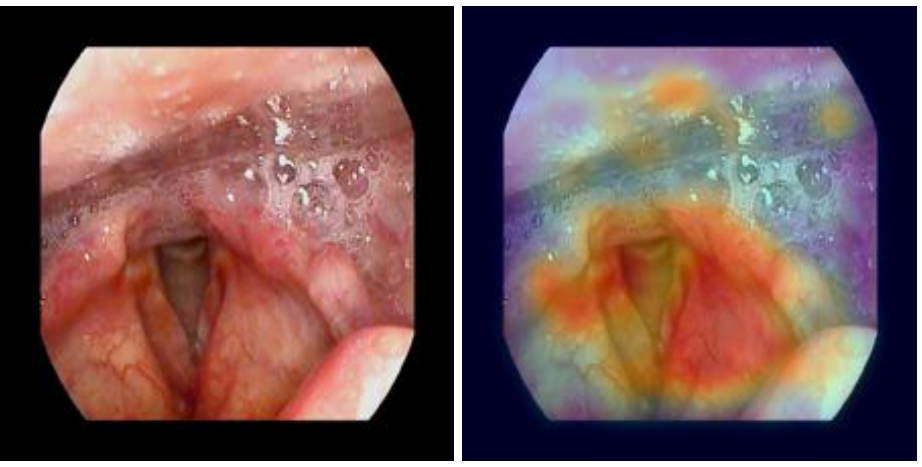












(a) **Center:** LH, **Label:** NCL, **Sex:** Male, **Age:** 50 yr.

Complaint Text:

Due to enlarged lymph node, patient was admitted to local hospital . Laryngoscopy showed leukoplakia of the right vocal cord . Patient had a history of smoking : 1 pack of cigarettes per day , for more than 30 years .

Risk Score Evaluations:

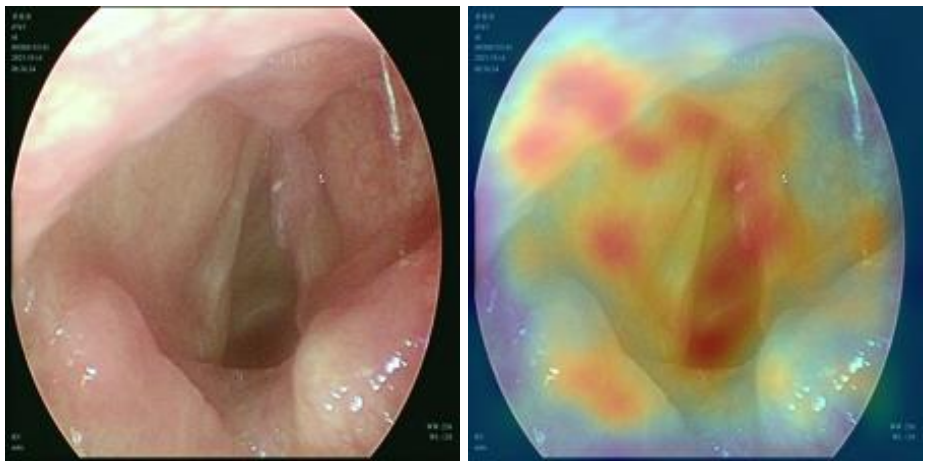
True score: 1.1

Model prediction score: 1.21

Expert group predictions - A: 1.1, B: 1.1, C: 2.0, D: 2.0

Senior group predictions - A: 1.1, B: 1.3, C: 1.1

Junior group predictions - A: 1.1, B: 1.0, C: 1.3



(b) **Center:** FAHZU, **Label:** CA, **Sex:** Male, **Age:** 77 yr.

Complaint Text:

Hoarseness for several months , no pain , no itching , little coughing , eating well , spitting a little white sputum . Foreign body sensation .

Risk Score Evaluations:

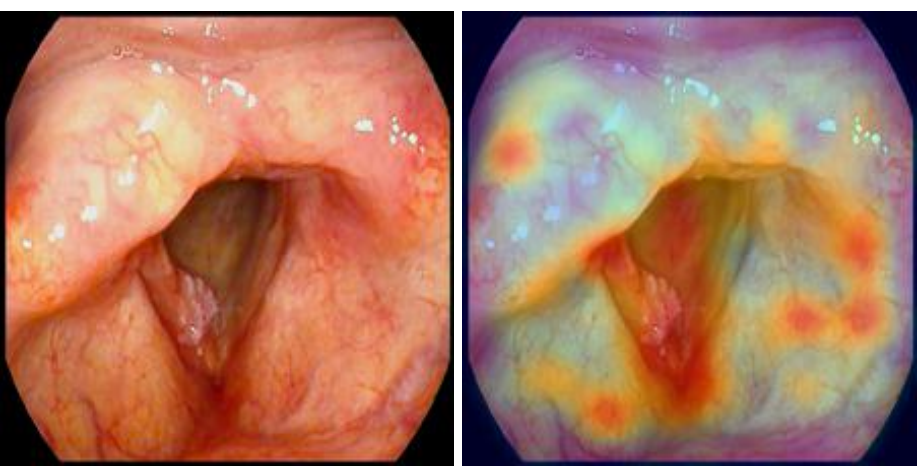
True Score: 2.0

Model prediction score: 1.97

Expert group predictions - A: 2.0, B: 1.3, C: 1.3, D: 2.0

Senior group predictions - A: 1.1, B: 1.1, C: 1.3

Junior group predictions - A: 1.3, B: 1.3, C: 1.3



(c) **Center:** LH, **Label:** CA, **Sex:** Male, **Age:** 69 yr.

Complaint Text:

The patient had hoarseness due to excessive use of voice 4 years ago , and had undergone vocal cord tumor resection , with postoperative pathology showing leukoplakia . The hoarseness was relieved after the operation . 1 month ago, the patient had hoarseness without obvious cause , with no sore throat , fever , cough, sputum , blood in sputum , and difficulty in breathing and swallowing , and denied any smoking and drinking history.

Risk Score Evaluations:

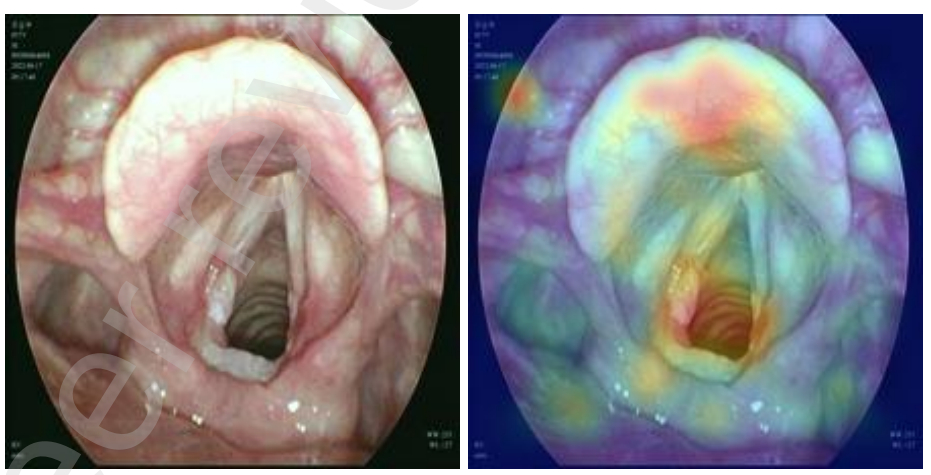
True score: 2.0

Model prediction score: 1.97

Expert group predictions - A: 2.0, B: 2.0, C: 1.3, D: 2.0

Senior group predictions - A: 1.3, B: 2.0, C: 2.0

Junior group predictions - A: 1.3, B: 1.3, C: 1.3



(d) **Center:** FAHZU, **Label:** NCL, **Sex:** Male, **Age:** 58 yr.

Complaint Text:

Recurrent hoarseness , acid reflux , never smoked , drank 2 catties of rice wine per day for more than 20 years . Local laryngoscope found neoplasm in vocal cords . History of atrial fibrillation.

Risk Score Evaluations:

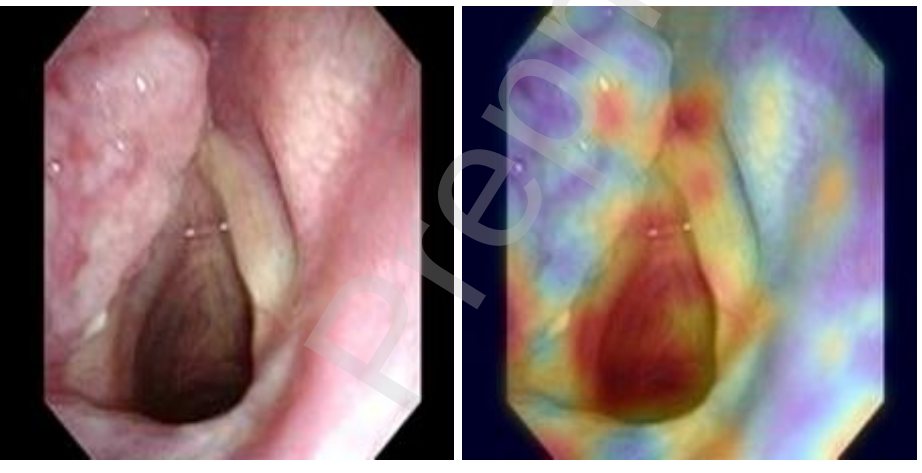
True score: 1.1

Model prediction score: 1.96

Expert group predictions - A: 1.1, B: 1.1, C: 1.3, D: 1.1

Senior group predictions - A: 1.3, B: 1.3, C: 1.0

Junior group predictions - A: 1.1, B: 1.4, C: 1.3



(e) **Center:** FAHWMU, **Label:** CA, **Sex:** Male, **Age:** 61 yr.

Complaint Text:

A left- sided neck mass was found 1 month ago , and is now present with mild sore throat and no hemoptysis.

Risk Score Evaluations:

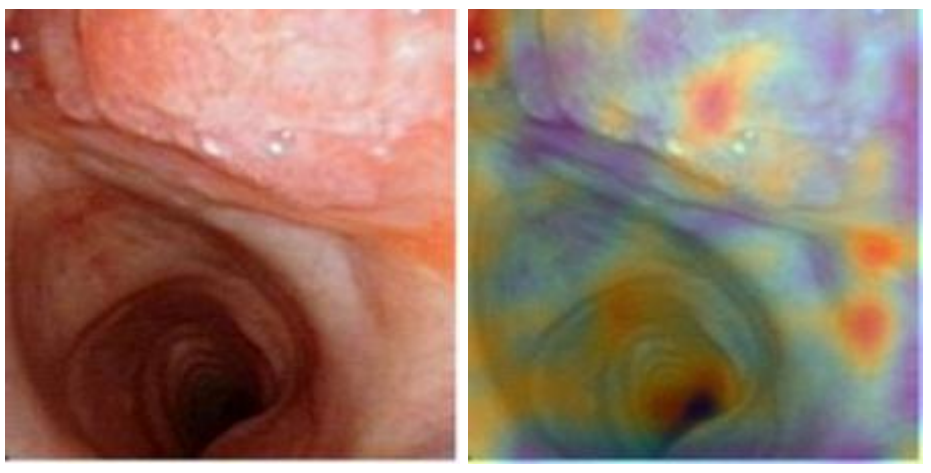
True score: 2.0

Model prediction score: 0.04

Expert group predictions - A: 2.0, B: 2.0, C: 2.0, D: 2.0

Senior group predictions - A: 2.0, B: 2.0, C: 2.0

Junior group predictions - A: 2.0, B: 2.0, C: 2.0



(f) **Center:** JSH, **Label:** CA, **Sex:** Male, **Age:** 64 yr.

Complaint Text:

A right sided neck mass was found for more than 2 months.

Risk Score Evaluations:

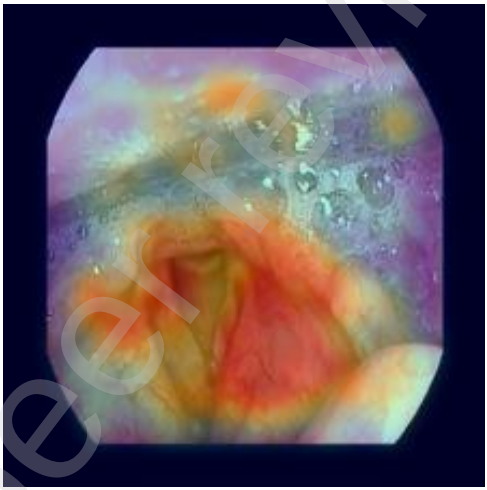
True score: 2.0

Model prediction score: 0.04

Expert group predictions - A: 2.0, B: 2.0, C: 2.0, D: 2.0

Senior group predictions - A: 2.0, B: 0.0, C: 2.0

Junior group predictions - A: 0.0, B: 1.1, C: 1.5

(a) **Center:** LH, **Label:** NCL, **Sex:** Male, **Age:** 50 yr.

Complaint Text:

Due to enlarged lymph node, patient was admitted to local hospital . Laryngoscopy showed leukoplakia of the right vocal cord . Patient had a history of smoking : 1 pack of cigarettes per day , for more than 30 years .

Risk Score Evaluations:

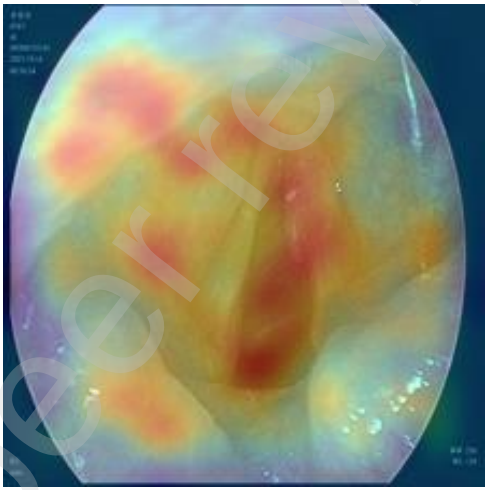
True score: 1.1

Model prediction score: 1.21

Expert group predictions - A: 1.1, B: 1.1, C: 2.0, D: 2.0

Senior group predictions - A: 1.1, B: 1.3, C: 1.1

Junior group predictions - A: 1.1, B: 1.0, C: 1.3

(b) **Center:** FAHZU, **Label:** CA, **Sex:** Male, **Age:** 77 yr.

Complaint Text:

Hoarseness for several months , no pain , no itching , little coughing , eating well , spitting a little white sputum . Foreign body sensation .

Risk Score Evaluations:

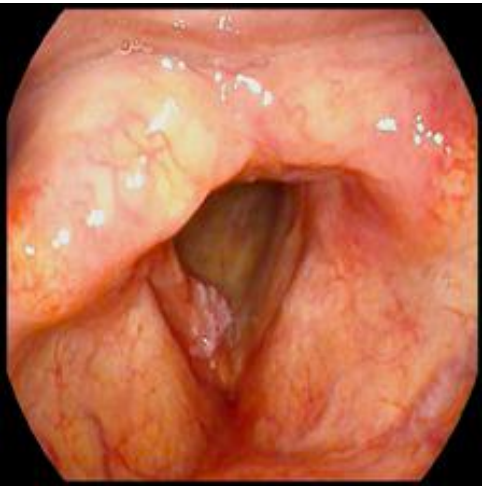
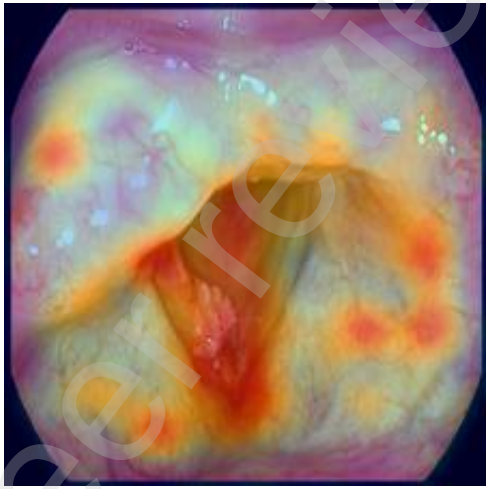
True Score: 2.0

Model prediction score: 1.97

Expert group predictions - A: 2.0, B: 1.3, C: 1.3, D: 2.0

Senior group predictions - A: 1.1, B: 1.1, C: 1.3

Junior group predictions - A: 1.3, B: 1.3, C: 1.3

(c) **Center:** LH, **Label:** CA, **Sex:** Male, **Age:** 69 yr.

Complaint Text:

The patient had hoarseness due to excessive use of voice 4 years ago , and had undergone vocal cord tumor resection , with postoperative pathology showing leukoplakia . The hoarseness was relieved after the operation . 1 month ago, the patient had hoarseness without obvious cause , with no sore throat , fever , cough, sputum , blood in sputum , and difficulty in breathing and swallowing , and denied any smoking and drinking history.

Risk Score Evaluations:

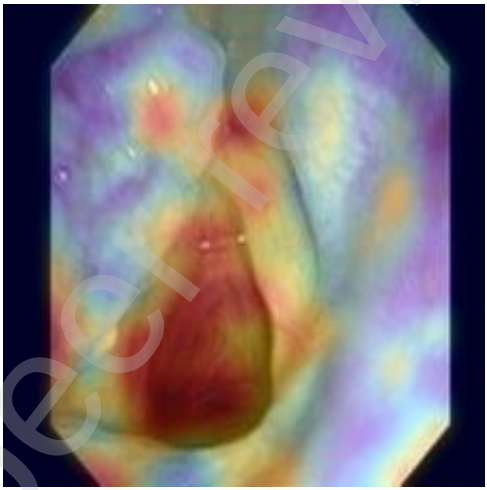
True score: 2.0

Model prediction score: 1.97

Expert group predictions - A: 2.0, B: 2.0, C: 1.3, D: 2.0

Senior group predictions - A: 1.3, B: 2.0, C: 2.0

Junior group predictions - A: 1.3, B: 1.3, C: 1.3

(e) **Center:** FAHWMU, **Label:** CA, **Sex:** Male, **Age:** 61 yr.

Complaint Text:

A left- sided neck mass was found 1 month ago , and is now present with mild sore throat and no hemoptysis.

Risk Score Evaluations:

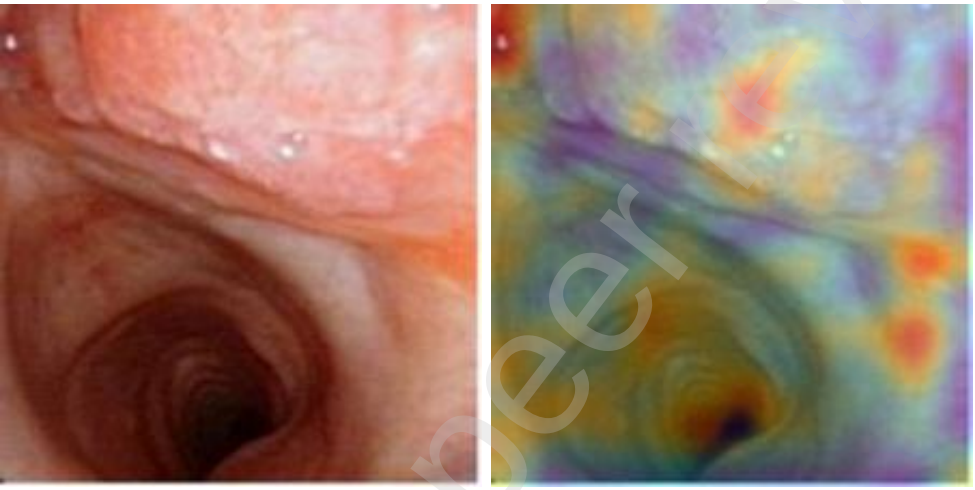
True score: 2.0

Model prediction score: 0.04

Expert group predictions - A: 2.0, B: 2.0, C: 2.0, D: 2.0

Senior group predictions - A: 2.0, B: 2.0, C: 2.0

Junior group predictions - A: 2.0, B: 2.0, C: 2.0



(f) **Center:** JSH, **Label:** CA, **Sex:** Male, **Age:** 64 yr.

Complaint Text:

A right sided neck mass was found for more than 2 months.

Risk Score Evaluations:

True score: 2.0

Model prediction score: 0.04

Expert group predictions - A: 2.0, B: 2.0, C: 2.0, D: 2.0

Senior group predictions - A: 2.0, B: 0.0, C: 2.0

Junior group predictions - A: 0.0, B: 1.1, C: 1.5


(d) **Center:** FAHZU, **Label:** NCL, **Sex:** Male, **Age:** 58 yr.

Complaint Text:

Recurrent hoarseness , acid reflux , never smoked , drank 2 catties of rice wine per day for more than 20 years . Local laryngoscope found neoplasm in vocal cords . History of atrial fibrillation.

Risk Score Evaluations:

True score: 1.1

Model prediction score: 1.96

Expert group predictions - A: 1.1, B: 1.1, C: 1.3, D: 1.1

Senior group predictions - A: 1.3, B: 1.3, C: 1.0

Junior group predictions - A: 1.1, B: 1.4, C: 1.3