Automatic Recognition of Laryngoscopic Images Using a Deep-Learning Technique

Ren, Jianjun; Jing, Xueping; Wang, Jing; Ren, Xue; Xu, Yang; Yang, Qiuyun; Ma, Lanzhi; Sun, Yi; Xu, Wei; Yang, Ning

Published in:
Laryngoscope

DOI:
10.1002/lary.28539

Publication date:
2020

Citation for published version (APA):
Ren, J., Jing, X., Wang, J., Ren, X., Xu, Y., Yang, Q., Ma, L., Sun, Y., Xu, W., Yang, N., Zou, J., Zheng, Y., Chen, M., Gan, W., Xiang, T., An, J., Liu, R., Lv, C., Lin, K., ... Zhao, Y. (2020). Automatic Recognition of Laryngoscopic Images Using a Deep-Learning Technique. *Laryngoscope*, *130*(11), E686-E693. https://doi.org/10.1002/lary.28539

The Laryngoscope
© 2020 The American Laryngological,
Rhinological and Otological Society, Inc.

Automatic Recognition of Laryngoscopic Images Using a Deep-Learning Technique

Jianjun Ren, PhD[†]; Xueping Jing, PhD[†]; Jing Wang, MD; Xue Ren, PhD; Yang Xu, PhD; Qiuyun Yang, MS; Lanzhi Ma, MS; Yi Sun, MS; Wei Xu, PhD; Ning Yang, PhD; Jian Zou, PhD; Yongbo Zheng, PhD; Min Chen, BSMed; Weigang Gan, MD; Ting Xiang, MS; Junnan An, MS; Ruiqing Liu, PhD; Cao Lv, MS; Ken Lin, MS; Xianfeng Zheng, BSMed; Fan Lou, MS; Yufang Rao, MS; Hui Yang, PhD; Kai Liu, PhD; Geoffrey Liu, MD, MSc; Tao Lu, PhD[‡]; Xiujuan Zheng, PhD[‡]; Yu Zhao, MD, PhD[‡]

Objectives/Hypothesis: To develop a deep-learning–based computer-aided diagnosis system for distinguishing laryngeal neoplasms (benign, precancerous lesions, and cancer) and improve the clinician-based accuracy of diagnostic assessments of laryngoscopy findings.

Study Design: Retrospective study.

Methods: A total of 24,667 laryngoscopy images (normal, vocal nodule, polyps, leukoplakia and malignancy) were collected to develop and test a convolutional neural network (CNN)-based classifier. A comparison between the proposed CNN-based classifier and the clinical visual assessments (CVAs) by 12 otolaryngologists was conducted.

Results: In the independent testing dataset, an overall accuracy of 96.24% was achieved; for leukoplakia, benign, malignancy, normal, and vocal nodule, the sensitivity and specificity were 92.8% vs. 98.9%, 97% vs. 99.7%, 89% vs. 99.3%, 99.0% vs. 99.4%, and 97.2% vs. 99.1%, respectively. Furthermore, when compared with CVAs on the randomly selected test dataset, the CNN-based classifier outperformed physicians for most laryngeal conditions, with striking improvements in the ability to distinguish nodules (98% vs. 45%, $P < .001$), polyps (91% vs. 86%, $P < .001$), leukoplakia (91% vs. 65%, $P < .001$), and malignancy (90% vs. 54%, $P < .001$).

Conclusions: The CNN-based classifier can provide a valuable reference for the diagnosis of laryngeal neoplasms during laryngoscopy, especially for distinguishing benign, precancerous, and cancer lesions.

Key Words: Deep learning, laryngoscopic image, artificial intelligence, convolutional neural networks, clinical visual assessment..

Level of Evidence: NA

***Laryngoscope*, 00:1–8, 2020**

From the Department of Otorhinolaryngology (J.R., J.W., Y.X., J.Z., Y.Z., M.C., W.G., T.X., J.A., X.Z., Y.R., H.Y., Y.Z.), West China Hospital, West China Medical School, Sichuan University, Chengdu, China; Medical Oncology and Medical Biophysics (J.R., G.L.), Princess Margaret Cancer Centre, Toronto, Ontario, Canada; Department of Automation (X.J., K.L., X.Z.), College of Electrical Engineering and Information Technology, Sichuan University, Chengdu, China; Department of Radiation Oncology (X.J.), University Medical Center Groningen, Groningen, The Netherlands; Department of Economic Statistics (X.R.), School of Statistics and Management, Shanghai University of Finance and Economics, Shanghai, China; Department of Forensics (Q.Y.), West China School of Preclinical and Forensic Medicine, Sichuan University, Chengdu, China; Department of Preclinical Medicine (L.M., Y.S.), West China School of Preclinical and Forensic Medicine, Sichuan University, Chengdu, China; Department of Biostatistics (W.X.), Princess Margaret Cancer Centre, Toronto, Ontario, Canada; College of Computer Science, Sichuan University (N.Y.), Chengdu, China; Department of Otorhinolaryngology (R.L.), Kunming City Women and Children Hospital, Kunming, China; Department of Otorhinolaryngology (C.L.), The Second Affiliated Hospital of Kunming Medical University, Kunming, China; Department of Otorhinolaryngology (K.L., F.L.), The Affiliated Children's Hospital of Kunming Medical University, Kunming, China; Medicine and Epidemiology, Dalla Lana School of Public Health (G.L.), University of Toronto, Toronto, Ontario, Canada; and the Department of Otolaryngology and Head Neck Surgery (T.L.), The First Affiliated Hospital of Kunming Medical University, Kunming, China.

Additional supporting information may be found in the online version of this article.

Editor's Note: This Manuscript was accepted for publication on December 30, 2019.

[†] J.R. and X.J. contributed equally to this work.

[‡] X.Z., T.L., and Y.Z. contributed equally to this work.

All authors participated in the design of study. J.R., T.L., J.W., Y.X., Q.Y., L.M., and Y.S. carried out the medical imaging collection. J.R., T.L., J.W., Y. Z., and H.Y. did the classification. Y.Z., M.C., J.Z., W.G., T.X., J.A., R.L., C.L., K.L., X.Z., F.L., and Y.R. did the clinician visual assessments to compete with the convolutional neural network. J.R., X.J., X.Z., N.Y., K.L., W.X., G.L., Y.Z., and X.R. conducted the medical image analysis. J.R., X.J., X.Z., G.L., W.X., Y.Z., and T.L. drafted and revised the manuscript. All authors read and approved the final manuscript.

This study was supported by the Fundamental Research Funds for the Central Universities (grant no. 2012017yjsy118), Key Research and Development Support Programs of Chengdu Science and Technology Bureau (grant no. 2018-YFYF-00123-SN).

The authors have no other funding, financial relationships, or conflicts of interest to disclose.

Send correspondence to Yu Zhao, MD, Department of Otorhinolaryngology, West China Hospital, West China Medical School, Sichuan University, No. 37 Guo Xue Alley, Chengdu, Sichuan 610041, China. E-mail: yutzhao@163.com

DOI: 10.1002/lary.28539

INTRODUCTION

The first attempts to examine the human larynx dates back to more than 150 years ago.[1] Imaging technologies in laryngeal diagnostics have since advanced enormously. Laryngology practices have changed over the past several decades as access, visualization, and manipulation of the larynx has become easier, better, and safer. Visualization of the larynx helps physicians better observe the detailed morphology of the glottal structures, so as to make an accurate diagnosis and the best management strategy.

Laryngoscopy is used routinely to diagnose laryngeal conditions, such as vocal fold cysts, nodules, polyps, and especially laryngeal neoplasms. For laryngeal cancer screening, endoscopy detection and obtaining suspicious precancerous/cancer tissue for biopsy are essential. Early detection for laryngeal precancer or cancer by laryngoscopy could result in early therapeutic interventions, benefiting patients' survival rate and prognosis. However, not all physicians have enough training, experience, and equipment necessary to fully visualize the larynx.[2] Human visual observation of laryngeal lesions varies, especially for the early-stage cancers whose diagnoses depend on the pathologic results of biopsies. There is additional waiting time for pathology reports. In clinics lacking good pathology support and narrow band imaging equipment, faster and more accurate image-based diagnosis is in demand.

As a popular technique of deep-learning algorithms, the convolutional neural network (CNN) has demonstrated its impressive power in natural image classification.[3,4] The emergence of the transfer learning technique swept away the barriers in exploiting advanced CNN-based algorithms, and it has therefore gained considerable popularity in medical imaging analysis for various clinical applications.[5–8] Moreover, recent reports showed that CNN-based algorithms could parallel or outperform human experts in visual tasks, such as the recognition of skin cancers, retinal diseases, gastrointestinal disease, and lymph node metastases in breast cancer, to name a few.[9–13]

The application of CNN-based algorithms in the field of laryngoscopy has not been well described. To the best of our knowledge, this study is the first to develop a fully automated system to identify and distinguish laryngeal benign, precancerous, and cancer lesions, and compare its performance to clinical visual assessments (CVAs) delivered by otolaryngologists.

MATERIALS AND METHODS

This study was approved by the Research Ethics Committee of West China Hospital (No. 2018-64). All methods were performed in accordance with the relevant guidelines and regulations, and informed consent was waived by the ethics committee due to the retrospective nature of this study.

Image Datasets

A total of 24,667 independent and clear consecutive laryngoscopy images from 9,231 patients were reviewed and obtained from the database of West China Hospital, Sichuan University, between 2012 and 2017. All participants received laryngoscopy performed by either one of two experienced endoscopists using a flexible 4.9-mm laryngoscope (Olympus Medical Systems, Tokyo, Japan). There were no exclusion criteria based on age, gender, or race. The collected laryngeal images were taken under routine lighting conditions and with wide ranges of zooming and optical magnification. The glottic area was the site of interest, with all sizes of opening of the vocal cords. All duplicate images, unclear images with low resolution, and images without vocal cords were exclude from this study by seven of the authors (J.R., T.L., J.W., Y.X., Q.Y., Y.S., L.M.). A total of 19,433 (80%) images from 7,521 patients were randomly divided into two independent subsets, 14,340 images from 5,250 patients for training and 5,093 images from 2,271 patients for validation (to tune hyperparameters and avoid overfitting). The remaining 20% (5,234 images from 1,710 patients) served as a testing dataset (to verify the generalization ability of the algorithm). Finally, 500 individual laryngoscopy images (each diagnostic class contained 100 images) were randomly chosen from the 5,234 images in the testing dataset; these images were used to compare the performance of the CNN-based algorithm to the CVAs by the otolaryngologists, and formed the performance assessment dataset (PAD). The development of the dataset is described in Figure 1 and Table I.

Image Classification Gold Standard

The laryngoscopic images were classified using two reference standard strategies: otolaryngologist-based and pathology-based labeling. The gold-standard classification of benign lesions, malignant neoplasms, and vocal leukoplakia were based on pathological diagnosis. Each individual's electronic chart was reviewed retrospectively by eight of the authors (J.R., T.L., J.W., Y.W., Y.S., L.M., Q.Y., W.Y.). A total of 8,645 images were labeled as benign lesions (polyps, 2,995 images), leukoplakia (2,120 images), and malignancies (3,530 images) according to their pathological records.

Due to the lack of biopsies for normal and small vocal nodules, these datasets (16,022 images) were manually sorted and labeled by a panel of experts (authors J.R., T.L., J.W., Y.Z., H.Y., Y.W., H.W.) according to the appearance of the vocal fold. Normal vocal folds were defined as smooth and straight edges of vocal folds, without visible blood vessels, erythema, edema, ventricular obliteration, postcricoid hyperplasia, mesh vascularization, or pseudosulcus. Vocal nodules were defined according to their typical shape features, of which nodules were bilateral, symmetrical, and broad based (Fig. 2). If there were disagreements in the labeling for each image, a consensus was reached among all of the experts. The image quantity for each class in each dataset is shown in Table I.

CNN-Based Diagnostic Classification Algorithm

Previous related work—CNN and ResNet

TRANSFER LEARNING. Transfer learning strategy was conducted in this study.[15] The ResNet-101 model,[14] which has been pretrained and showed excellent performance (with 6.4% top-five error and 22.6% top-one error) on the ImageNet dataset (contains 1.2 million images with 1,000 categories),[4] was adopted for the task of classifying laryngoscopy images (see Supporting Information, Methods).

THE CNN-BASED CLASSIFIER. The CNN-based classifier utilized a single ResNet-101 model to classify laryngoscopic images into five conditions. We modified the number of output classes of the last layer from 1,000 to five, and fine-tuned parameters across all the layers using laryngoscopic image training data.

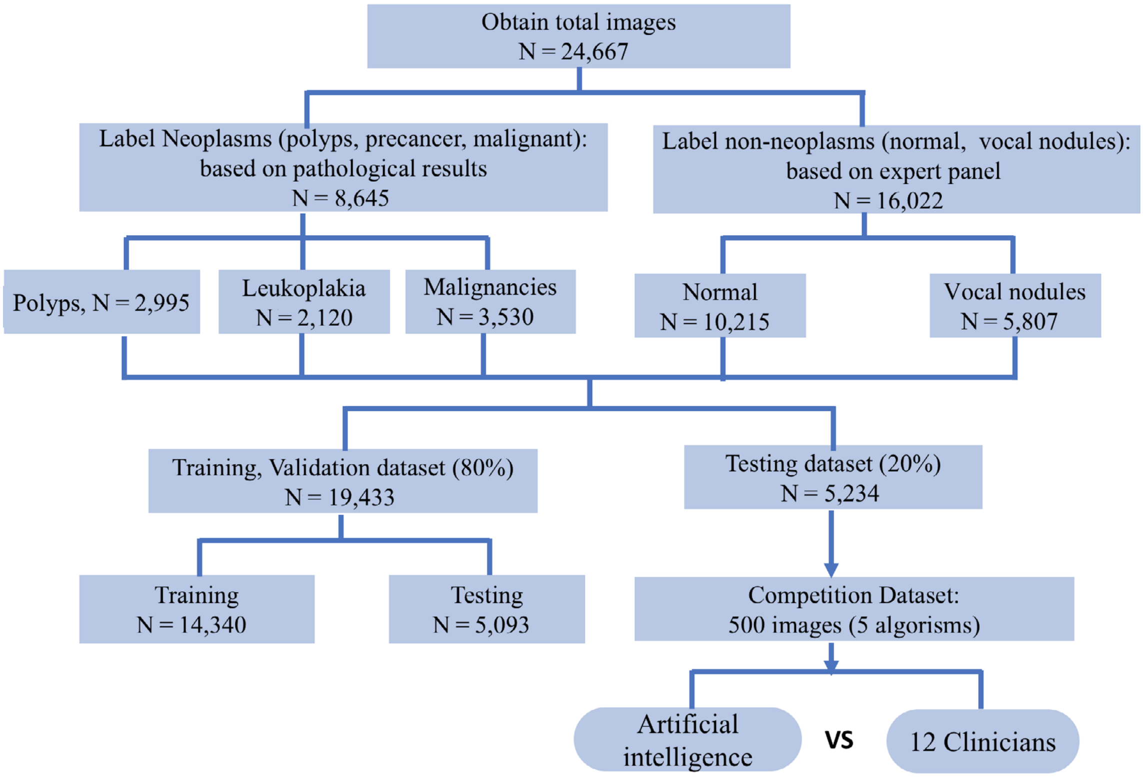


Fig. 1. The flow diagram of the dataset creation. [Color figure can be viewed in the online issue, which is available at www.laryngoscope.com.]

Training, validation, and testing

TRAINING. To match the original input dimensions of the ResNet-101 model, we resized the images to 240 × 240 pixels, and then cropped them to 224 × 224 pixels randomly in the process of training. We changed the brightness and contrast of the input images randomly, flipping them horizontally, with a probability of 0.5, and rotating them randomly between −30° and 30° for data augmentation.

Weighted cross-entropy loss function was used to rescale the weight of each class to deal with the imbalanced dataset during training.[16] The loss weights for leukoplakia, polyps, malignancy, normal, and vocal nodules in the direct-ResNet classifier were set as 5.62, 3.53, 2.90, 1, and 1.86 according to the amount of training images.

Stochastic gradient descent optimizers, with a momentum of 0.9 and weight decay of 0.000001, were used in the training procedures. The initial learning rate was set to 0.0001. The learning rates were a decade every 10 epochs, with a factor of 10 for the convergence of the models (Supporting Figure S1A, B).

VALIDATION AND TESTING. The validation dataset was used to find the best hyperparameters and to avoid overfitting during training, and the test dataset was used to evaluate the performances of different classifiers at the end.

Pytorch, a popular deep-learning framework (https://github.com/pytorch), was used to establish the CNN-based classifier. The proposed classifier was trained on an Ubuntu 16.04 computer with one Intel (Santa Clara, CA) Core i7-5930K CPU, two NVIDIA (Santa Clara, CA) GTX 1080 8 GB GPUs, and 16 GB RAM memory.

The t-distributed stochastic neighbor embedding method was used to explore the internal features learned by the CNN-based classifier.[17] Gradient-weighted class activation mapping (Grad-CAM) was also used to generate the activation map of predicted classes.[18] Grad-CAM is a technique that helps explain the decision-making process of the CNN model by producing a

TABLE I.
The Number of Different Diagnostic Images for Each Condition by Dataset.

Subsets	Training Dataset	Validation Dataset	Testing Dataset	Total	Performance Assessment Dataset (Selected From the Testing Dataset)
Normal	6,129	2,043	2,043	10,215	100
Vocal nodules	3,279	1,264	1,264	5,807	100
Leukoplakia	1,089	366	665	2,120	100
Benign	1,734	705	556	2,995	100
Malignancy	2,109	715	706	3,530	100
Total	14,340 (58.1%)	5,093 (20.6%)	5,234 (21.2%)	24,667 (100%)	500

The percentages in parentheses are the proportions of each subset.

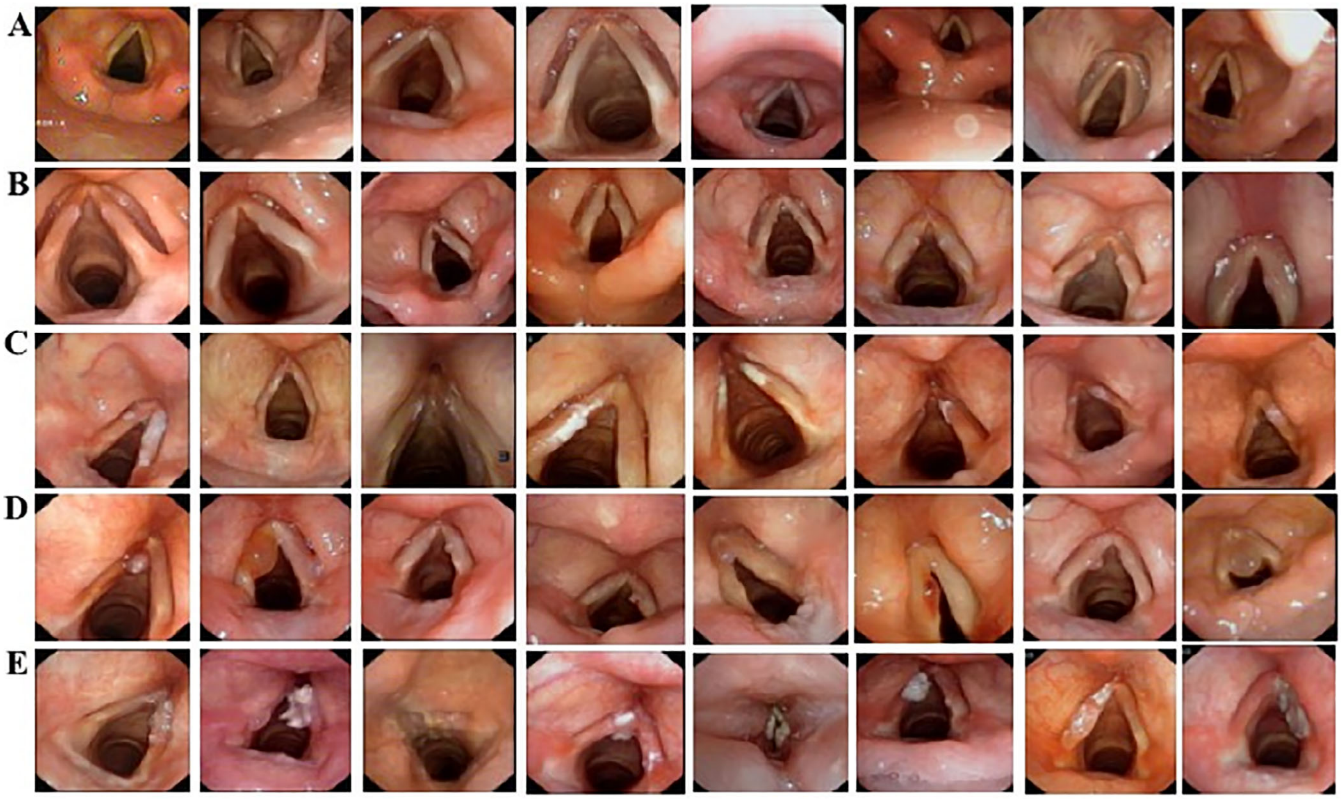


Fig. 2. Representative laryngoscopic images for the reference standard (A) normal, (B) vocal nodule, (C) leukoplakia, (D) benign, (E) malignancy. [Color figure can be viewed in the online issue, which is available at www.laryngoscope.com.]

coarse localization map highlighting important regions in the image, and the produced heatmap can also be used to guide the physicians during examination of the clinical image.[19]

Comparison between the deep-learning–based algorithm and clinician visual assessments. The PAD (500 images) mentioned previously was used to compare the performance between the CNN-based algorithm and CVAs. A panel of 12 otolaryngologists (J.Z., T.X., W.G., R.L., C.L., K.L., X.Z., Y.Z. J.A., Y.R., F.L., M.C.) from the three tertiary hospitals participated in this comparison. An otolaryngologist would earn one point if his/her judgement for one image was correct. The distribution of professional titles of the 12 doctors was: 41.7% (five individuals: clinicians 2, 7, 9, 10, and 12) were residents; 33.3% (four individuals: clinicians 3, 5, 6, and 8) were attending doctors; and 25% (three individuals: clinicians 1, 4, and 11) were vice or chief physicians. All physicians performed blinded assessments according to the appearance of images without time constraint. Receiver operating characteristics (ROCs) were generated to compare the recognition ability for each class between the proposed algorithm and that of the 12 physicians.

Statistical Analysis

The Student *t* test was applied to analyze the accuracy rate (score). ROC curves and dichotomized tables were used to determine sensitivity and specificity comparing either CVA or the CNN-based algorithm with the gold standard. The correlation between time consumption and accuracy rate of physicians was analyzed by using the Pearson correlation test. Analysis of variance was conducted to test whether there were differences between the physicians' professional titles and accuracy. All statistical procedures were performed using SPSS version 17.0 (IBM, Armonk, NY); significance was set at an α of .05.

RESULTS

Classification Results of the Deep-Learning–Based Algorithm

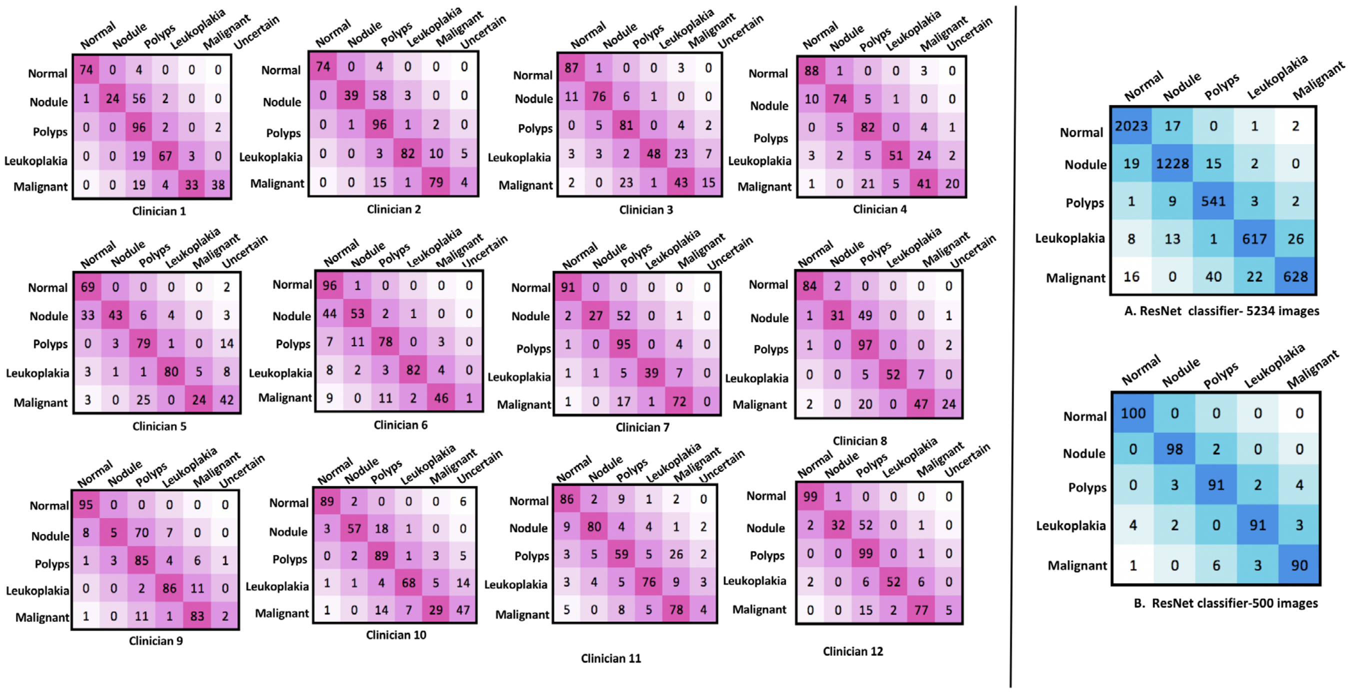
The CNN-based classifier achieved an overall accuracy of 96.24% on the testing dataset, with high sensitivities versus specificities of 92.78% versus 98.95% for leukoplakia (the area under the ROC curve [AUC] = 0.9975), 97.30% versus 99.67% for polyps (AUC = 0.9972), 88.95% versus 98.29% for malignancy (AUC = 0.9956), 99.02% versus 99.36% for normal (AUC = 0.9991), and 97.15% versus 99.09% for vocal nodules (AUC = 0.9976), respectively (Table II). The performance details in the classification for laryngeal conditions by CNN are shown in Figure 3. We visualized the image representations at the last hide layer of the CNN-based classifier by forward propagation of test images (Supporting Fig. S1C, D). A predictive classification demo is available at https://github.com/xpjing-SCU/Laryngoscope-image-classification.

TABLE II.
The Performance of CNN Classifier and 12 Clinicians.

		Clinician Visual Assessment Scores						
Performer	Subsets	Mean ± SD	*P* Value	Sensitivity	*P* Value	Specificity	*P* Value	Accuracy
CNN (testing dataset, 5,234 images)	Normal	—	—	99.02%	—	99.36%	—	96.24%
	Vocal nodule	—	—	97.15%	—	99.09%	—	
	Polyps	—	—	97.30%	—	99.67%	—	
	Leukoplakia	—	—	92.78%	—	98.95%	—	
	Malignancy	—	—	89%	—	99.33%	—	
CNN (performance assessment, dataset, 500 images)	Normal	—	—	100%	—	98.75%	—	94%
	Vocal nodule	—	—	98%	—	98.7%	—	
	Polyps	—	—	91%	—	99.5%	—	
	Leukoplakia	—	—	91%	—	98.7%	—	
	Malignancy	—	—	90%	—	98.25%	—	
Clinicians (performance assessment, dataset,500 images)	Normal	86 ± 2.71	.000*	96.7% (94.5-98.9)	.014	99% (98.6-99.7)	.013*	86%
	Vocal nodules	45 ± 6.75	.000*	50.1% (36.2-64)	.000*	89% (86.1-91.4)	.000*	45%
	Polyps	86 ± 3.33	.000*	90.1% (83.8-96.3)	.776	97% (95.6-98.6)	.431	86%
	Leukoplakia	65 ± 4.67	.000*	78.3% (72.8-83.7)	.001*	95% (94-96.3)	.001*	65%
	Malignant	54 ± 6.31	.000*	69.4% (62.1-76.5)	.000*	94% (93.4-95)	.000*	54%
	Total	62 ± 7.93	.015*	77.8% (67.1-88.6)	.971	94% (90.8-98.1)	.98	62%

*Significant difference.

The specificity and sensitivity for clinicians were presented with 95% confidence interval. The scores (mean ± SD) reflect the numbers of correct diagnoses, where each point reflects one right diagnosis. *P* value: clinician visual assessment compared with the CNN classifier. Sensitivity: true positive/(true positive + false positive). Specificity: true negative/(true negative + false positive). CNN = convolutional neural network; SD = standard deviation.



	Normal	Nodule	Polyps	Leukoplakia	Malignant	Uncertain
Normal	74	0	4	0	0	0
Nodule	1	24	56	2	0	0
Polyps	0	0	96	2	0	2
Leukoplakia	0	0	19	67	3	0
Malignant	0	0	19	4	33	38

Clinician 1

	Normal	Nodule	Polyps	Leukoplakia	Malignant	Uncertain
Normal	74	0	4	0	0	0
Nodule	0	39	58	3	0	0
Polyps	0	1	96	1	2	0
Leukoplakia	0	0	3	82	10	5
Malignant	0	0	15	1	79	4

Clinician 2

	Normal	Nodule	Polyps	Leukoplakia	Malignant	Uncertain
Normal	87	1	0	0	3	0
Nodule	11	76	6	1	0	0
Polyps	0	5	81	0	4	2
Leukoplakia	3	3	2	48	23	7
Malignant	2	0	23	1	43	15

Clinician 3

	Normal	Nodule	Polyps	Leukoplakia	Malignant	Uncertain
Normal	88	1	0	0	3	0
Nodule	10	74	5	1	0	0
Polyps	0	5	82	0	4	1
Leukoplakia	3	2	5	51	24	2
Malignant	1	0	21	5	41	20

Clinician 4

	Normal	Nodule	Polyps	Leukoplakia	Malignant	Uncertain
Normal	69	0	0	0	0	2
Nodule	33	43	6	4	0	3
Polyps	0	3	79	1	0	14
Leukoplakia	3	1	1	80	5	8
Malignant	3	0	25	0	24	42

Clinician 5

	Normal	Nodule	Polyps	Leukoplakia	Malignant	Uncertain
Normal	96	1	0	0	0	0
Nodule	44	53	2	1	0	0
Polyps	7	11	78	0	3	0
Leukoplakia	8	2	3	82	4	0
Malignant	9	0	11	2	46	1

Clinician 6

	Normal	Nodule	Polyps	Leukoplakia	Malignant	Uncertain
Normal	91	0	0	0	0	0
Nodule	2	27	52	0	1	0
Polyps	1	0	95	0	4	0
Leukoplakia	1	1	5	39	7	0
Malignant	1	0	17	1	72	0

Clinician 7

	Normal	Nodule	Polyps	Leukoplakia	Malignant	Uncertain
Normal	84	2	0	0	0	0
Nodule	1	31	49	0	0	1
Polyps	1	0	97	0	0	2
Leukoplakia	0	0	5	52	7	0
Malignant	2	0	20	0	47	24

Clinician 8

	Normal	Nodule	Polyps	Leukoplakia	Malignant	Uncertain
Normal	95	0	0	0	0	0
Nodule	8	5	70	7	0	0
Polyps	1	3	85	4	6	1
Leukoplakia	0	0	2	86	11	0
Malignant	1	0	11	1	83	2

Clinician 9

	Normal	Nodule	Polyps	Leukoplakia	Malignant	Uncertain
Normal	89	2	0	0	0	6
Nodule	3	57	18	1	0	0
Polyps	0	2	89	1	3	5
Leukoplakia	1	1	4	68	5	14
Malignant	1	0	14	7	29	47

Clinician 10

	Normal	Nodule	Polyps	Leukoplakia	Malignant	Uncertain
Normal	86	2	9	1	2	0
Nodule	9	80	4	4	1	2
Polyps	3	5	59	5	26	2
Leukoplakia	3	4	5	76	9	3
Malignant	5	0	8	5	78	4

Clinician 11

	Normal	Nodule	Polyps	Leukoplakia	Malignant	Uncertain
Normal	99	1	0	0	0	0
Nodule	2	32	52	0	1	0
Polyps	0	0	99	0	1	0
Leukoplakia	2	0	6	52	6	0
Malignant	0	0	15	2	77	5

Clinician 12

	Normal	Nodule	Polyps	Leukoplakia	Malignant
Normal	2023	17	0	1	2
Nodule	19	1228	15	2	0
Polyps	1	9	541	3	2
Leukoplakia	8	13	1	617	26
Malignant	16	0	40	22	628

A. ResNet classifier- 5234 images

	Normal	Nodule	Polyps	Leukoplakia	Malignant
Normal	100	0	0	0	0
Nodule	0	98	2	0	0
Polyps	0	3	91	2	4
Leukoplakia	4	2	0	91	3
Malignant	1	0	6	3	90

B. ResNet classifier-500 images

Fig. 3. Confusion matrix of the performance of clinician visual assessments (CVAs) for each of the 12 physicians and the convolutional neural network (CNN) classifier on different datasets. The purple tables present the performance details of the individual clinician. The blue tables describe the performance of the classification model on two datasets: (A) the confusion matrix of the CNN classifier on 5,234 test images, and (B) the confusion matrix of the CNN classifier on the 500 performance assessments dataset that had been selected to compare the performances of the CNN-based algorithm and CVAs. The individual table cell values represent the number of images in each category. For each confusion matrix, vertical labels represent the gold-standard results (true labels), whereas the horizontal labels represent the decisions made by the classifiers (predicted labels). [Color figure can be viewed in the online issue, which is available at www.laryngoscope.com.]

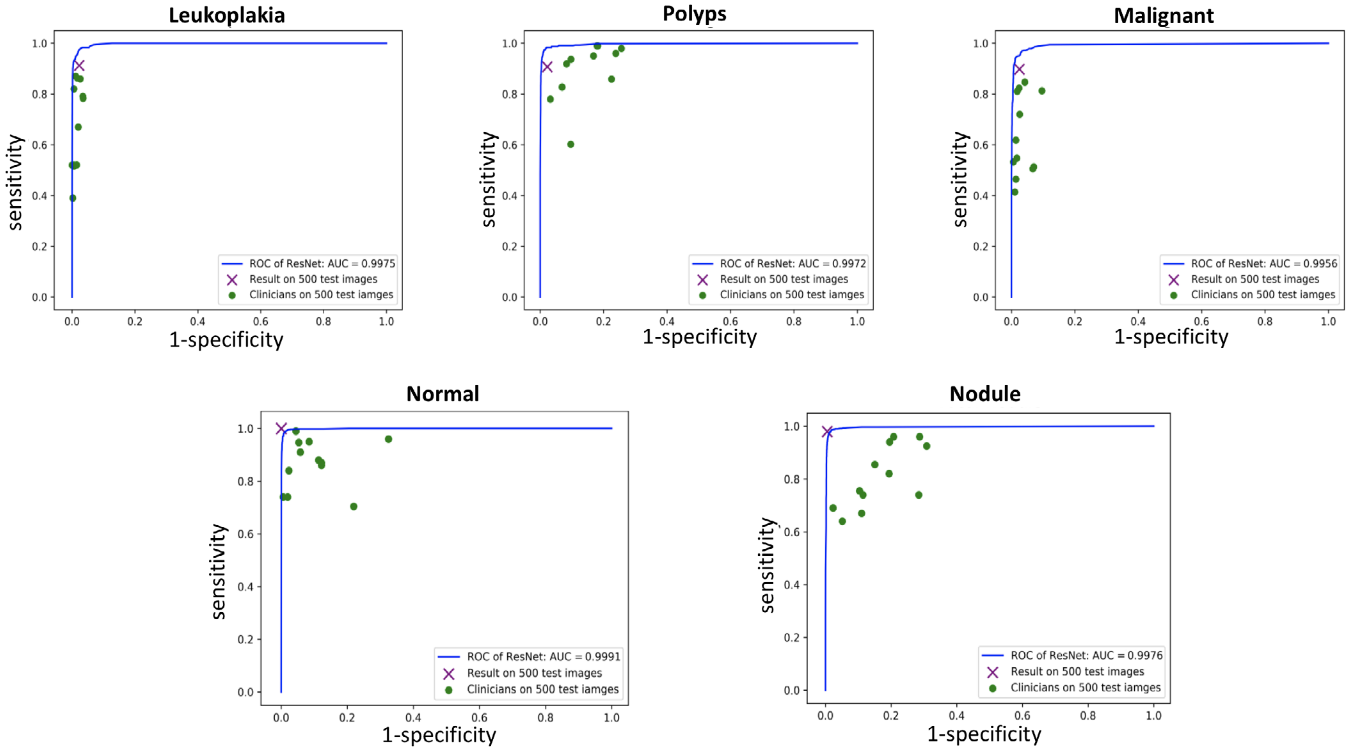


Fig. 4. Receiver operating characteristics (ROC) curves of the convolutional neural network (CNN) classifier comparing with the 12 physicians. Each green point represents the performance of a human clinician. The X represents the performances of the CNN classifier when using the testing sets of the 500 performance assessments dataset. The ROC analyses for each class were performed in a one-versus-all scheme. For each condition (normal, vocal nodule, leukoplakia, benign, and malignant), the threshold of the output in ResNet was varied in the interval 0 to 1 to generate the ROC curves for each threshold point. AUC = the area under the ROC curve. [Color figure can be viewed in the online issue, which is available at www.laryngoscope.com.]

The Decision-Making Process of the CNN Model and Heatmap-Based Error Analysis

The Grad-CAM technique was utilized to analyze the decision-making process of CNN models. Supporting Figure S2 shows those correctly classified test images and their corresponding activation maps. As for those misclassified test images, their Grad-CAM–based heatmaps could help explain how those wrong decisions were made by the CNN model (Supporting Fig. S3).

The Performance of CVA by Otolaryngologists. The panel of 12 physicians obtained an average score of 62% for the PAD. There were no significant differences among the scores of junior, intermediate, and senior physicians. Because there was a certain level of uncertainty in some of the clinician ratings, this was regarded as an error when determining accuracy. The accuracy of physicians, with corresponding specificities and sensitivities for each class, are listed in the Table II. The answers for each test by individuals are presented in the Figure 3. Physicians had the greatest difficulty distinguishing leukoplakia and malignant lesions. The average time physicians spent to assess the PAD was 141.9 ± 83.5 minutes (mean ± standard deviation) (range = 37.5–300 minutes); there was no significant association between the time spent assessing and the accuracy rates.

Comparison Between the CNN-Based Classifier With Human Experts Using CVAs. The CNN-based classifier had been tested with the same PAD as the physicians. It only took the computer 22.7 seconds to diagnose all images, which was on average 500 times more time efficient than CVAs. In this task, the CNN-based classifier achieved better overall accuracy than human experts (94% vs. 62%, $P < .001$) and outperformed physicians in all lesion recognitions, especially in the subsets of nodules (98% vs.45%, $P < .001$), polyps (91% vs. 86%, $P < .001$), leukoplakia (91% vs. 65%, $P < .001$), and malignancy (90% vs. 54%, $P < .001$) (Table II).

In the aspects of identifying sensitivity and specificity, the integrated CNN outperformed CVA significantly in the classes of normal (CNN vs. CVA: sensitivity 100% vs. 96.7%; specificity 99% vs. 99%), vocal nodules (CNN vs. CVA: sensitivity = 98% vs. 50%, $P < .001$; specificity = 99% vs. 89%, $P = .003$), benign (CNN vs. CVA: sensitivity = 91% vs. 90%, $P = .655$; specificity = 99.5% vs. 97%, $P < .001$), leukoplakia (CNN vs. CVA: sensitivity = 91% vs. 78%, $P < .001$; specificity = 98.7% vs. 95%; $P = .72$), and malignant (CNN vs. CVA: sensitivity = 90% vs. 69%, $P = .015$; specificity = 98% vs. 94%, $P = .18$). These comparisons are presented in Figures 3 and 4 and Table II.

To keep our machine learning in a real-world practice approach, routine lighting conditions and wide ranges of zooming and optical magnifications were performed, with the only vocal cords to be visible, of which continuous images from different laryngeal segments could reflect the dynamic process of laryngoscopy to some extent. Supporting Figure S4 illustrates laryngeal conditions in different angles, distances, and brightness in our database. Furthermore, gender and age were not restricted in this study to guarantee the universality of this technique, and the automatic system maintained high overall accuracy (above 95%) even though the anatomic structures among male, female, the elderly, and children have different appearances.

DISCUSSION

Laryngoscopy is the diagnostic procedure for many laryngeal diseases, alongside clinical symptoms and pathological findings. Visualization of the larynx under laryngoscopy is among the first steps to diagnosis, but it is fraught with subjectivity and is dependent greatly on the experience of the examiner. The ideal methods to reduce individual variability efficiently, especially for patients with cancerous or precancerous lesions, will greatly help the early screening/detection of laryngeal cancers and improve patients' prognosis.

Computer-aided diagnosis systems based on deep-learning techniques for laryngeal diseases have been rarely utilized. Witt et al. conducted artificial neural networks on the laryngoscopy images of 62 patients to identify laryngopharyngeal reflux.[20] Verikas et al. categorized 785 vocal fold images by using the kernel-based automated method and achieved a 94% accurate rate.[21] Ilgner et al. developed an automated system to differentiate healthy and diseased laryngeal images based on co-occurrence matrices, with an 81% correct classification rate when testing the system on a very small set of 35 images.[22]

In this study, we developed a CNN-based classifier for the screening and diagnosis of laryngeal disease. This deep-learning technique took a practical approach; we neither utilized artificial light nor zoom or optical amplification restrictions, and no limit or exclusion was performed for age and genders, which maximally kept the natural diversity of different laryngeal appearance of different disease conditions in the real world. Moreover, the automatic intelligent system using a deep-learning algorithm based on large database training turned out to be efficient and accurate for all conditions whether they were male or female and elderly or children.

Physicians have the challenge in distinguishing precancerous lesions and cancer, which are generally difficult clinical entities, often with much heterogeneity of appearance and a continuum of severity (leading to borderline and potentially overlapping diagnostic states). If the appearance of the neoplasm is highly suspected to be cancer, physicians will take a biopsy under the guidance of the laryngoscope; however, misdiagnosis is still not rare if the cancer is in situ, in which case it is hard to distinguish from leukoplakia. Our newly developed laryngeal automatic diagnosing system performed excellently in distinguishing precancerous lesions (leukoplakia) and cancer, with sensitivities and specificities over 90%, as well as in identifying a normal larynx and vocal nodules. Regardless, the performances were significantly better than manual CVA by physicians in almost all comparisons, but never performed significantly worse under any circumstance. Our results thus suggested that deep-learning techniques have value in the setting of clinical laryngoscopy assessments, which will help develop a modern automatic system for diagnosing laryngeal lesions by noninvasive laryngoscopies.

However, this diagnostic system was designed to supplement, but not replace, clinical assessments, which could be useful for the screening of laryngeal disease. Automatic diagnostic classification algorithms could help physicians to make more confident determinations in laryngoscopies, especially in the case of precancers/early cancers. The rapidity of assessment by algorithms ensured that clinical decision making was not delayed, and could raise flags in some cases that would allow for further clinical investigation, such as giving recommendation for biopsy for highly suspicious malignant neoplasms. Computer-aided diagnosis is thus becoming a next-generation tool for the diagnosis of human disease, which can offer promising applicability to daily clinical practice for reducing the burden of endoscopists and the waiting time of patients, while making the cancer-screening procedure become more efficient. Such technology will also benefit patients in remote and rural areas through telemedicine technology,and improve the quality of medical care in developing countries and areas.

The limitations of this study included that we only focused on five major laryngeal conditions, meaning that we did not train or test rare diagnostic entities. Further research is necessary to evaluate additional types of laryngeal diseases such as papilloma and amyloidosis. Another limitation was that for the group of small vocal nodules, of which tissues were quite limited and were not routinely biopsied or no surgery was performed, the criteria for inclusion in the database was mainly according to the panel of experts. However, due to its distinct appearance and consensus by all experts, the accuracy of the database itself could be guaranteed.

CONCLUSION

This newly developed computer-aided system provides a valuable reference for the screening of laryngeal neoplasms during laryngoscopy, especially for distinguishing benign precancerous lesions and cancer in a real-word condition.

BIBLIOGRAPHY

1. Alberti P. The history of laryngology: a centennial celebration. *Otolaryngol Head Neck Surg* 1996;114:345-354.
2. Stachler R, Francis DO, Schwartz SR, et al. Clinical practice guideline: hoarseness (dysphonia) (update). *Otolaryngol Head Neck Surg* 2018;158: S1-S42.
3. Russakovsky O, Deng J, Krause J, et al. ImageNet large scale visual recognition challenge. *Int J Comput Vis* 2015;115:211-252.
4. Deng J, Dong W, Socher R, Li L. ImageNet: a large-scale hierarchical image database. Paper presented at: IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR 2009); June 20–25, 2009; Miami, Florid, USA. IEEE, 2009.
5. Pan S, Yang Q. A survey on transfer learning. *IEEE Trans Knowl Data Eng* 2009;22:1345-1359.

6. Shen D, Wu G, Suk H. Deep learning in medical image analysis. *Ann Rev Biomed Eng* 2017;19:221-248.
7. Litjens G, Kooi T, Bejnordi B, et al. A survey on deep learning in medical image analysis. *Med Image Anal* 2017;42:60-88.
8. Zhang JP, Xia Y, Xie Y, Fulham M, Feng DD. Classification of medical images in the biomedical literature by jointly using deep and handcrafted visual features. *IEEE J Biomed Health Inform* 2017;22:1521-1530.
9. Esteva A, Kuprel B, Novoa R, et al. Dermatologist-level classification of skin cancer with deep neural networks. *Nature* 2017;542:115-118.
10. Gulshan V, Peng L, Coram M, et al. Development and validation of a deep learning algorithm for detection of diabetic retinopathy in retinal fundus photographs. *JAMA* 2016;316:2402-2410.
11. Yu L, Chen H, Dou Q, Qin J, Heng PA. Integrating online and offline three-dimensional deep learning for automated polyp detection in colonoscopy videos. *IEEE J Biomed Health Inform* 2017;21:65-75.
12. Bejnordi BE, Veta M, Johannes van Diest P, et al. Diagnostic assessment of deep learning algorithms for detection of lymph node metastases in women with breast cancer. *JAMA* 2017;318:2199-2210.
13. Hirasawa T, Aoyama K, Tanimoto T, et al. Application of artificial intelligence using a convolutional neural network for detecting gastric cancer in endoscopic images. *Gastric Cancer* 2018;21:653-660.
14. He K, Zhang X, Ren S, Sun J. Deep residual learning for image recognition. Paper presented at: IEEE Conference on Computer Vision and Pattern Recognition (CVPR); June 26–July 1, 2016; Las Vegas, NV.
15. Shin HC, Roth HR, Gao M, et al. Deep convolutional neural networks for computer-aided detection: CNN architectures, dataset characteristics and transfer learning. *IEEE Trans Med Imaging* 2016;35:1285-1298.
16. Xie S, Tu Z. Holistically-nested edge detection. The IEEE International Conference on Computer Vision (ICCV); December 7-13, 2015; Santiago, Chile.
17. van der Maaten L, Hinton G. Visualizing data using t-SNE. *J Mach Learn Res* 2008;9:2579-2605.
18. Ramprasaath RS, Michael C, Abhishek D, et al. Grad-CAM: visual explanations from deep networks via gradient-based localization. *Int J Comput Vis* 2020;128:336-359.
19. Cheng CT, Ho TY, Lee TY, et al. Application of a deep learning algorithm for detection and visualization of hip fractures on plain pelvic radiographs. *Eur Radiol* 2019;29:5469-5477.
20. Witt DR, Chen H, Mielens JD, et al. Detection of chronic laryngitis due to laryngopharyngeal reflux using color and texture analysis of laryngoscopic images. *J Voice* 2014;28:98-105.
21. Verikas A, Gelzinis A, Bacauskiene M, Uloza V. Integrating global and local analysis of color, texture and geometrical information for categorizing laryngeal images. *Int J Pattern Recognit Artifici Intel* 2006;20:1187-1205.
22. Ilgner JF, Palm C, Schutz AG, Spitzer K, Westhofen M, Lehmann TM. Colour texture analysis for quantitative laryngoscopy. *Acta Otolaryngol* 2003; 123:730-734.