IRBM 44 (2023) 100799

Contents lists available at ScienceDirect

IRBM

journal homepage: www.elsevier.com/locate/irbm

Elsevier Masson France
EM|consulte
www.em-consulte.com

Original Article

Identifying Laryngeal Neoplasms in Laryngoscope Images via Deep Learning Based Object Detection: A Case Study on an Extremely Small Data Set

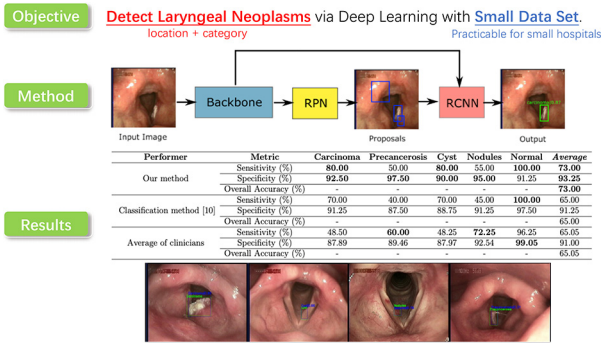
Shijie Fang [a,1], Jia Fu [c,1], Chen Du [b], Tong Lin [a,*], Yan Yan [b,*]

[a] National Key Lab of General Artificial Intelligence, School of Intelligence Science and Technology, Peking University, Beijing, 100871, China
[b] Department of Otolaryngology, Peking University Third Hospital, Beijing, 100191, China
[c] The First Affiliated Hospital of Nanchang University, Nanchang, 330006, China

HIGHLIGHTS

- An extremely small data set with 279 images is built as a more realistic benchmark.
- A small-data learning method is proposed to turn the problem into detection task.
- Experimental results demonstrate the effectiveness of our method (73.00% accuracy).

GRAPHICAL ABSTRACT



Performer	Metric	Carcinoma	Precancerosis	Cyst	Nodules	Normal	Average
Our method	Sensitivity (%)	**80.00**	50.00	**80.00**	55.00	100.00	**73.00**
	Specificity (%)	**92.50**	**97.50**	**90.00**	**95.00**	91.25	**93.25**
	Overall Accuracy (%)	-	-	-	-	-	**73.00**
Classification method [10]	Sensitivity (%)	70.00	40.00	70.00	45.00	**100.00**	65.00
	Specificity (%)	91.25	87.50	88.75	91.25	97.50	91.25
	Overall Accuracy (%)	-	-	-	-	-	65.00
Average of clinicians	Sensitivity (%)	48.50	**60.00**	48.25	**72.25**	96.25	65.05
	Specificity (%)	87.89	89.46	87.97	92.54	**99.05**	91.00
	Overall Accuracy (%)	-	-	-	-	-	65.05

ARTICLE INFO

Article history:
Received 25 April 2023
Received in revised form 28 August 2023
Accepted 30 August 2023
Available online 7 September 2023

Keywords:
Larynx
Artificial intelligence

ABSTRACT

Objectives: Laryngoscopy is a medical procedure for obtaining a view of the human larynx. It is challenging for clinicians to distinguish laryngeal neoplasms by human visual observation. Recent deep learning methods can assist clinicians in improving the accuracy of distinguishing. However, existed methods are often trained on large-scale private datasets, while other researchers and hospitals can neither access these private datasets nor afford to build such large-scale datasets. In this paper, we focus on identifying laryngeal neoplasms under the "small data" regime, which is more important for many small hospitals to investigate deep learning models for diagnosis.
Material and methods: We build an extremely small dataset consisting of 279 laryngoscopic images of different categories. We found that traditional deep learning models for image classification cannot achieve satisfactory performance for small data, due to the great variability of recording laryngoscopic images and the small area of the neoplasms. To address these difficulties, we propose to employ object detection methods for this small data problem. Concretely, a Faster R-CNN is implemented here, which combines the DropBlock regularization technique to alleviate overfitting additionally.
Results: Compared to previous methods, our model is more robust to overfitting and can predict the location and category of detected neoplasms simultaneously. Our method achieves 73.00% overall accuracy, which is higher than the average of clinicians (65.05%) and the recent state-of-the-art classification method (65.00%).

* Corresponding authors.
E-mail addresses: lintong@pku.edu.cn (T. Lin), yanyan_ent@bjmu.edu.cn (Y. Yan).
[1] Shijie Fang and Jia Fu contributed equally to this paper.

https://doi.org/10.1016/j.irbm.2023.100799

Conclusion: The proposed method shows great ability to detect both the category and location of neoplasms and can be served as a screening tool to help the final decisions of clinicians.

1. Introduction

In the clinical practice of Otolaryngology, the laryngoscope is the most common and important piece of equipment for the examination the larynx structure. Through a laryngoscope, rigid or flexible, lesion of larynx could be found by ENT (Ear, Nose and Throat) practitioner. Doctors' judgment of laryngoscope images depends on clinical experience, and doctors of different seniority from different levels of hospitals may draw different conclusions from the same laryngoscope images. If there exists a reliable automated system for screening laryngoscopic images, it can give preliminary recommendations that can assist doctors in their interpretation of such images.

Artificial Intelligence (AI) has its technical advantages in image recognition, which has been shown in imaging diagnosis of skin cancer [1]. With the updating of technology, objective analysis of laryngoscope images is in full swing. Research of Du et al. [2] applies Artificial Neural Network (ANN) on laryngoscope images to determine the validity of color and texture abnormalities in LPR (Laryngopharyngeal Reflux) system. The disadvantage with this method is that using only two features could omit more substantial information contained in the image.

In recent years, a great number of publications apply computer vision techniques to medical imagery such as radiology, pathology, ophthalmology and dermatology, which is benefited by the growing availability of highly structured images [3]. For example, EchoNet [4] is proposed to recognize cardiac structures and predict the systemic phenotypes which are difficult for human interpretation. For skin images with pathology, DermGAN [5] leveraged Generative Adversarial Nets (GANs) [6] for synthesizing clinical images with skin conditions as data augmentation. Convolutional Neural Network (CNN) is employed in [7] for automated detection of diabetic retinopathy and diabetic macular edema in retinal fundus photographs.

Nevertheless, these works are often supported by extensive medical image collections which are easily accessible. For example, CT images will be automatically preserved in digit form, hence researchers don't have to consciously collect and build the dataset. On the other hand, there are some large-scale public datasets such as MURA of bone X-rays [8] and LUNA16 [9] of lung nodules which can be efficiently used for training and evaluation.

While for laryngoscope images, to the best of our knowledge, no public dataset is available. It's challenging to build a dataset from scratch. Firstly, the number of patients with throat disease is not as large as other diseases, and only a small number of patients are willing to take laryngoscopy images. Secondly, laryngoscope images are not saved in digit form for many hospitals in previous years, which makes it hard to collect enough data. Thirdly, privacy policy further limits the amount of related metadata. Fourthly, the annotations can only be made by senior physicians to ensure the correctness of the label, which increases the cost of annotation.

There're few works toward larynx image recognition. For example, Yao et al. [10] employ a CNN for selecting informative frames from laryngoscopic videos. However, their method is not able to predict the category of neoplasms. The most related work to ours is [11], where a widely-used ResNet-101 model [12] is pretrained on the ImageNet natural image dataset and transferred to classify laryngoscope images. To be more detailed, they built a dataset of 24677 consecutive laryngoscope images in total from 9231 patients. This dataset is further divided into three parts: a training set of 14340 images, a validation set of 5093 images, and a test set of 5234 images. There are five categories in their dataset: Normal, Vocal nodule, Leukoplakia, Benign and Malignancy. Using the ResNet-101 model, they achieved an overall accuracy of 96.24%, a sensitivity of 99.02% and a specificity of 99.36% on the test set.

However, we argue there exist several limitations in the method of [11]. First, their high accuracy was achieved on a large training set (14340 images from 5250 patients) collected over six years (from 2012 to 2017). Currently, their dataset cannot be made publicly available. Hence, other researchers must build their own datasets from scratch, possibly with a small number of laryngoscope images. It would be almost impossible for many small hospitals to build dataset of such large-scale from scratch. It is quite questionable whether the transferred ResNet-101 model can achieve satisfactory performance on these small datasets. Even if they released the model, it cannot be directly employed by others since the different laryngoscope scanning tools can cause huge differences in the domain of images. Second, their image classification model can only report the category of the whole image, lacking interpretability when an otolaryngologist attempts to examine the results of the computer-aided diagnosis system.

To address these problems, we propose a diagnosis method for laryngoscope images under small-scale data. Recently learning on small data [13] has attracted much attentions due to the expensive cost of annotation and training, the wide applicability of real scenarios, and the attempt to pursue Artificial General Intelligence (AGI). For this small data problem, we implement an object detection model rather than conventional image classification methods. We argue that image classification can not focus on the small area of the interested region in a laryngoscopic image, particularly in the setting of small data. In this way, we can simultaneously predict the category and detect the region of neoplasms from the input RGB images. To learn from small data, we propose to use DropBlock as the regularization to prevent overfitting. To demonstrate the effectiveness of our method, we built a dataset of 279 images from 279 patients (one image per patient), which is much smaller than [11]. As shown in Fig. 1, our dataset contains five categories that are different from [11]: normal, cyst, nodules, laryngeal carcinoma, and precancerous lesions. We use a rectangular bounding box as the location annotation of the neoplasm (if exists) and the type of neoplasm as the category label.

In summary, our contributions are as follows:

1. A extremely small dataset with 279 images from 279 patients is built as a more realistic benchmark for laryngology practices.
2. A method based on object detection is proposed. By outputting both the category label and the location of pathology, our results are more interpretable than other black-box image classification methods which offer only the category.
3. Experimental results demonstrate the effectiveness of our method. Even on a small-scale dataset, our method achieves 73.00% overall accuracy, which outperforms clinical visual assessments (CVAs) and state-of-the-art automated method.

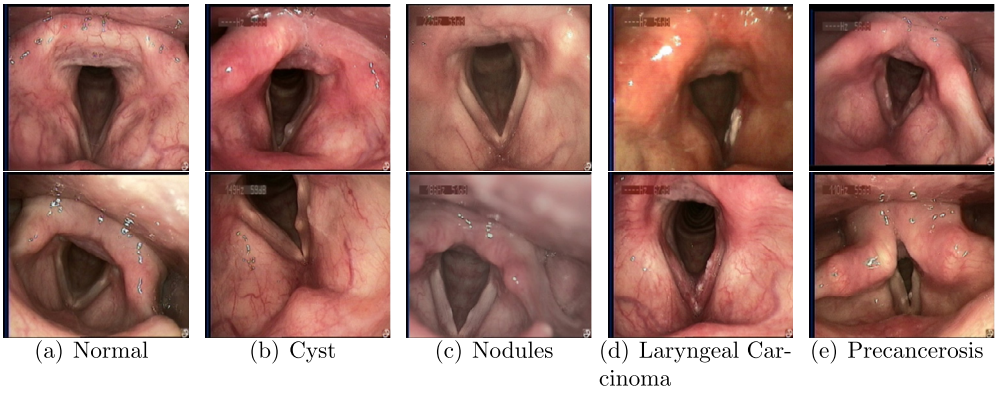


Fig. 1. Some examples of our dataset.

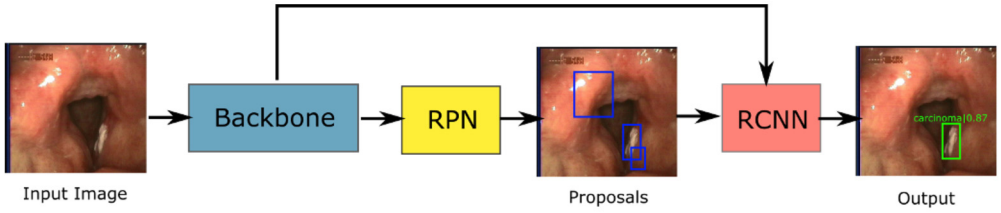


Fig. 2. The pipeline of our method. The input is an RGB image, which is first fed into a backbone CNN to extract image features. Then, a Region Proposal Net (RPN) generates a set of region proposals, which are the potential regions that may contain interested objects. Finally, an R-CNN head takes both the image features and region proposals to give a category prediction (with a confidence score) and the location of a single bounding box.

2. Materials and methods

2.1. Neoplasm detection based on Faster R-CNN

Our method is mainly based on Faster R-CNN [14], one of the most widely used two-stage models for object detection. Compared to classification model [11], Faster R-CNN is capable of predicting the location and category of neoplasms simultaneously. Besides, Faster R-CNN is more lightweight than state-of-the-art detection methods [15,16] and can prevent overfitting to small-scale data. Fig. 2 illustrates the three parts of Faster R-CNN: a backbone module, a Region Proposal Network (RPN), and an R-CNN detector module.

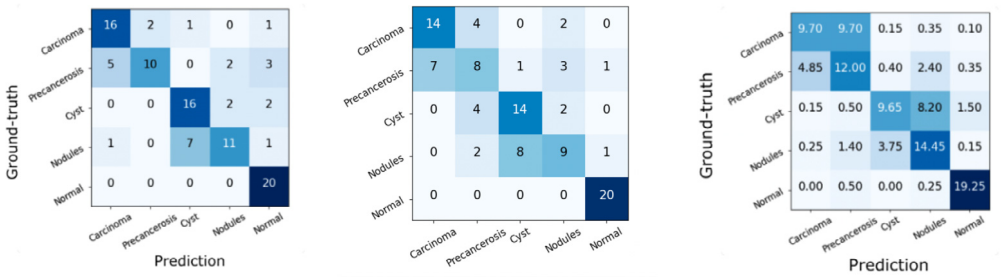
The backbone module is used to extract visual features from the input images. Modern detectors choose various backbones for different purposes. For example, large models like ResNeST [17] are designed for higher performance at the expense of heavy computation. For this task, we choose ResNet-50 (with 50 layers in the model) pre-trained on the ImageNet [18] dataset as the backbone to reduce computation overhead and alleviate overfitting. Since the amounts of training data are extremely small, we use the DropBlock [19] regularization technique to alleviate the overfitting problem. Given the *block_size* and γ as a hyperparameter, DropBlock first generates a sample mask M (with size *block_size* $\times$ *block_size*) at each pixel (i, j) with a Bernoulli distribution, i.e. $M_{(i,j)} \in Bernoulli(\gamma)$. Then, for each zero position $M_{(i,j)}$, we create a spatial mask of a square with the center being (i, j) and shape being *block_size* $\times$ *block_size*, and setting all the values of the feature map A in the square as zero. In this way, multiple results of the feature map can be obtained by sampling multiple times, which can prevent the network from overfitting to some specified patterns.

Since object detection is a rather difficult problem, Faster R-CNN partitions the whole optimization process into two stages. The first stage is a Region Proposal Net (RPN), which takes the feature maps and generates thousands of "region proposals". These proposals represent the most likely areas that may contain objects of interest. The second stage is an R-CNN Head, which uses the feature maps to predict the category label and to yield more accurate locations on the basis of proposals. For optimization of the RPN, the shapes of region proposals are predefined to reduce the search space, which is also called "anchors". In [14], the scale of anchors is set as 8 to handle the case that some objects may occupy the whole canvas in nature object detection. Since the sizes of neoplasms in laryngoscopic images are rather small, we set the scale of anchors to 3 in order to achieve better performance. The loss function of RPN is composed of two parts: binary classification loss and anchor regression loss, where binary cross entropy and L1 loss are utilized respectively.

The R-CNN head is composed of two fully convolution branches: a classification branch and a regression branch. With the region proposals generated by RPN, the R-CNN Head first uses RoI (Region of Interest) Align operation to registrate the feature map of the whole image onto the region of proposals. Then for each region proposal, the classification branch yields the probability (or confidence) of four categories (except the normal) with a softmax operator, while the regression branch creates a bounding box. Only one bounding box with the highest confidence is kept for each image, assuming that there is only one of RoI in the examined laryngoscope image. Finally, if its confidence is lower than a threshold β, the bounding box will be discarded and the image is classified as the normal category (without any bounding box). In our experiments, the threshold β is empirically set as 0.3. For training R-CNN head, we use cross entropy for the classification branch and L1 loss for regression branch.

2.2. Implementation details

Due to the limitation of this extremely small data set, we use 5-fold cross-validation rather than a fixed partition of the training set and test set. Since the original images are of different sizes in pixels, we first resize them to the fixed 640×640 pixel resolution, which is widely used in object detection tasks. Then, we use a horizontal flip with a probability of 0.5 for data augmentation.



(a) Our method (b) Classification method [11] (c) Average of clinicians

Fig. 3. Confusion matrix of (a) out method; (b) the classification model in [11]; (c) average of clinicians under Setting A.

Table 1
Amount of examples in each category under different settings.

Category	Setting A	Setting B
Carcinoma	49	20
Precancerosis	40	20
Cyst	62	20
Nodules	63	20
Normal	65	20

We tune hyper-parameters using the cross-validation method. The batch size is 2 and the total number of epochs in training is set as 36. The initial learning rate is set as 0.02, which decays by a factor of 0.1 at the 24-th and 33-th epochs. As for the optimizer, we chose stochastic gradient descent (SGD) with a momentum of 0.9 and weight decay of 0.0001. All code is implemented based on the PyTorch deep learning framework and the model is trained and tested with a single NVIDIA 2080Ti GPU.

3. Results

Since the dataset and training code of [11] are not open-source, for demonstrating the effectiveness of our method, we chose 100 samples from the entire dataset and collected the classification results from 20 clinicians of Peking University Third Hospital, Beijing Xiyuan Hospital, and Inner Mongolia Forestry General Hospital. All doctors have completed 3 years of residency training program. One-third of them are doctors under 10 years' experience, one-third of them are non-laryngeal specialty doctors for more than 10 years, and the remainings are laryngeal specialists with more than 10 years' experience. As above, the dataset is composed of 179 images with annotation ($D_x = \{x_1, x_2, \cdots, x_{179}\}$) and 100 images with both annotation and diagnostic results from 20 clinicians ($D_y = \{y_1, y_2, \cdots, y_{100}\}$). We designed two different training settings:

- Setting A: 5-fold cross-validation over D_y. For example, at the first round, $D_x + \{y_{21}, y_{22}, \cdots, y_{100}\}$ are used for training, $\{y_1, y_2, \cdots, y_{20}\}$ are used for testing.
- Setting B: 5-fold cross-validation over the entire dataset. For example, at the first round, $\{x_{37}, x_{38}, \cdots, x_{179}\} + \{y_{21}, y_{22}, \cdots, y_{100}\}$ are used for training, the remainings are used for testing. The category distribution is shown as Table 1.

Setting A is designed for comparing our method with clinicians, while Setting B is mainly for comparing with [11]. Unless specified, Setting A is the default setting of experiments.

The specificity, sensitivity of each category and the overall accuracy over the whole dataset are employed as metrics, which are defined as follows:

$$specifity_i = \frac{TN_i}{E - E_i}, \tag{1}$$

$$sensitivity_i = \frac{TP_i}{E_i}, \tag{2}$$

$$overall\_accuracy == \frac{\sum_{i=1}^{m} TP_i}{E}, \tag{3}$$

where i is the category index and $m = 5$ is the total number of categories. The variables TP_i and TN_i refer to the number of true positive and true negative samples, respectively, for each category i. E_i is the number of samples for each category and $E = \sum_{i=1}^{m} E_i$ is the total number of samples in the test set.

3.1. Comparing to the classification network and clinicians

We implemented their method and trained it on our dataset to compare with them. We also tried to directly use the trained parameters of [11] to yield predictions on our dataset. It obtained an average accuracy of 56.67% for carcinoma, nodules and normal (the precancerosis and cyst were not involved in their dataset). The result was worse than expected, which may be contributed to the domain difference of our dataset. For a more fair comparison, we train and evaluate their model from scratch on our dataset.

The evaluation results of our method, the ResNet-101 model used in [11], and human clinics are shown in Table 2 and Table 3. Under setting A, our method achieves an overall accuracy of 73.00% on the proposed dataset, which is better than the method of [11] (65.00%) and clinicians (65.05%). Besides, our method has the highest average specificity and sensitivity. The confusion matrix and ROC curve are given by Fig. 3 and Fig. 4 respectively.

For human experts, it tends to be challenging to distinguish carcinoma (our AUC=0.8962) from precancerosis (our AUC=0.8413) (see Fig. 3(c)) for their high visual similarity. For our method, the sensitivity and specificity are 31.50% and 4.61% higher than the clinicians because the deep convolution network is able to capture the minor differences between them that a human cannot. Our method also achieves the best performance in the cyst category (our AUC=0.8556). However, the network does not outperform in the nodules category (our AUC=0.7819), which may be attributed to the limited data. Nodules tend to be tiny lesion and the network may require more samples to learn the difference. The specificity of normal images (our AUC=0.9787) is lower than experts while the sensitivity is higher, which is highly related to the hyper-parameter β. If we use a bigger value for β, the precision will be increased and recall will be reduced. The used $\beta = 0.3$ is empirically set. Since the baseline method cannot predict the regions of lesions, direct numerical comparison on the performance of detection is not available. To illustrate the effectiveness of our method, we give some illustrations in Fig. 5.

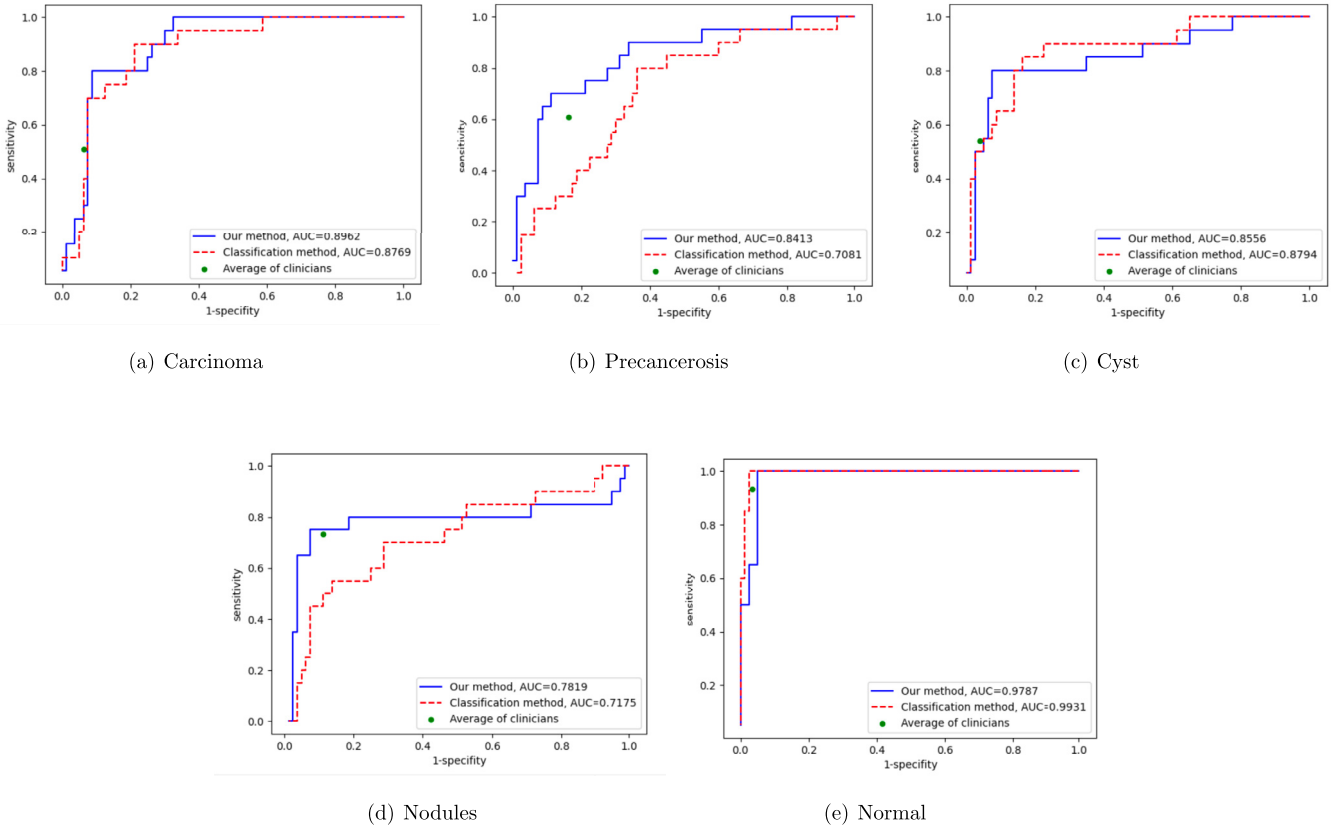
Our method also achieves higher average sensitivity and specificity compared to [11] under both setting A and setting B. As for nodules, the sensitivity of our method is significantly lower than [11], since our method contains both detection (RPN) and classification (R-CNN Head). Tiny lesion like nodules is difficult to

Table 2
The results of our method (**Setting A**, single-run), classification method [11] (**Setting A**, single-run) and 20 different clinicians. Bold indicates the best result of each category. Since the number of samples for each category is the same, the overall accuracy is numerically equal to average sensitivity.

Performer	Metric	Carcinoma	Precancerosis	Cyst	Nodules	Normal	*Average*
Our method	Sensitivity (%)	**80.00**	50.00	**80.00**	55.00	**100.00**	**73.00**
	Specificity (%)	**92.50**	**97.50**	**90.00**	**95.00**	91.25	**93.25**
	Overall Accuracy (%)	-	-	-	-	-	**73.00**
Classification method [11]	Sensitivity (%)	70.00	40.00	70.00	45.00	**100.00**	65.00
	Specificity (%)	91.25	87.50	88.75	91.25	97.50	91.25
	Overall Accuracy (%)	-	-	-	-	-	65.00
Average of clinicians	Sensitivity (%)	48.50	**60.00**	48.25	**72.25**	96.25	65.05
	Specificity (%)	87.89	89.46	87.97	92.54	**99.05**	91.00
	Overall Accuracy (%)	-	-	-	-	-	65.05

Table 3
The results of our method (**Setting B**, single-run), classification method [11] (**Setting B**, single-run). Bold indicates the best result of each category.

Performer	Metric	Carcinoma	Precancerosis	Cyst	Nodules	Normal	*Average*
Our method	Sensitivity (%)	67.35	**40.00**	**69.35**	71.43	**95.38**	**68.70**
	Specificity (%)	**97.83**	95.40	**92.17**	**91.20**	86.92	**92.70**
	Overall Accuracy (%)	-	-	-	-	-	**71.32**
Classification method [11]	Sensitivity (%)	**69.39**	12.50	66.13	**80.95**	90.77	63.95
	Specificity (%)	91.30	**98.33**	**92.17**	85.19	**92.52**	91.90
	Overall Accuracy (%)	-	-	-	-	-	68.10



(a) Carcinoma (b) Precancerosis (c) Cyst

(d) Nodules (e) Normal

Fig. 4. The Receiver operating characteristics (ROC) curves of each category. The threshold of the output for our method and [11] are varied in the interval 0 to 1 to generate each threshold point. AUC is the area under the ROC curve.

detection state. However, our method obtains significantly higher sensitivity on precancerosis, cyst, and normal categories. It is additionally worth mentioning that our method is able to predict the exact location of neoplasm while the method of [11] cannot.

3.2. Comparing to the other detectors

To demonstrate the effectiveness of our method, we compare the results of Faster R-CNN and other state-of-the-art models of object detection – Cascade R-CNN [16], ResNeSt [17], NAS-FPN [20], and Double Head [21]. Cascade R-CNN consists of a sequence of detectors trained with increasing Intersection over Union (IoU) thresholds, to be sequentially more selective against close false positives. ResNeSt is the extension of ResNet, which applies channel-wise attention on different network branches to leverage their success in capturing cross-feature interactions and learning diverse representations. NAS-FPN was discovered using Neural Ar-

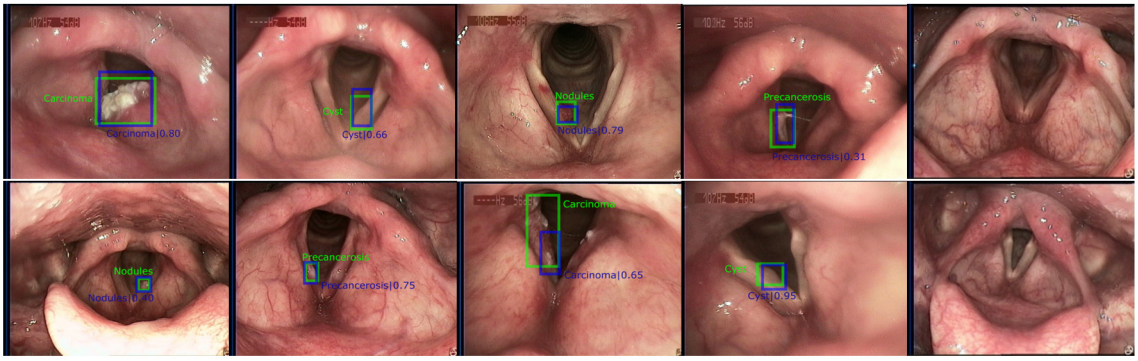


Fig. 5. Some examples of the results with each category shown in a column. The green bounding boxes are the ground truth and the blue bounding boxes are the prediction of our model. For clarity, only the bounding box with the highest confidence score is shown in each image. The two images in the last column are normal, so no ground truth is annotated. Our model succeeds to filter all proposals and no prediction is generated for the normal category.

Table 4
The results (5-fold cross validation, single-run) of our method and other state-of-the-art object detectors. Bold indicates the best result. Since the number of samples for each category is the same, the overall accuracy is numerically equal to average sensitivity.

Method	Cascade R-CNN	ResNeSt	NAS-FPN	Double Head	Faster R-CNN (ours)
Average Sensitivity (%)	66.00	59.00	65.00	67.00	**73.00**
Average Specificity (%)	91.50	89.75	91.25	91.75	**93.25**
Overall Accuracy (%)	66.00	59.00	65.00	67.00	**73.00**

chitecture Search, and it consists of a combination of top-down and bottom-up connections to fuse features across scales. Double Head has a fully connected head focusing on classification and a convolution head for bounding box regression in order to address the imbalance problem of a traditional R-CNN head. The experimental results of these methods are given in Table 4. (For simplification, only the average sensitivity and specificity of all categories are listed.) It is clear that our method achieves the best results among these models. The main reason for this may be attributed to the tiny scale of our dataset. Using a complicated model can result in an overfitting problem, especially for a dataset with only 279 samples.

4. Discussion

Experience and expertise can influence doctors' judgment of laryngoscope images. For doctors of different hospitals and seniority, the accuracy of the judgment of laryngoscope images varies. Balanced sensitivity and specificity are characteristic of good screening methods. Hence, the method obtained in this study could serve as a screening tool to help inexperienced doctors diagnose correctly. According to the results of our study, the image processing methods have relatively high sensitivity and specificity to the recognition of laryngoscope images of various vocal cord lesions.

Computer-aided methods for laryngeal disease diagnosis have not been widely used in real deployments due to the limitation of data. Ren et al. [11] proposed a CNN-based classification method and a large-scale dataset that contains 24667 samples in total. Using this dataset, they achieved remarkable results. However, their model can only predict the category of the input image, which gives the results limited applicability. Additionally, it is an intensive task to build such a dataset, which may hinder the usage of a computer-aided method of diagnosis in other hospitals. To address these issues, we propose an effective method based on Faster R-CNN. Our method can not only predict the category of neoplasms but also generate a bounding box to indicate its location. More importantly, the proposed method is a more practicable method for many hospitals that can not afford to build a large-scale dataset like [11]. The results produced on the small-scale dataset demonstrate the effectiveness and robustness of our method.

5. Significance

The significance of this research in the actual clinical setting can be concluded by four points:

1. The advantage of rapid diagnosis. Employing the deep learning method saves time to develop a perioperative plan for diseases such as laryngeal cancer and precancerous lesions, and prepare more fully and rationally.
2. The increased accuracy of identifying benign hypertrophic lesions of the vocal cords helps to accurately remove lesions under laryngeal microsurgery and maximize the preservation of normal mucosa, thereby improving vocalization. Especially for vocal cord cysts, the surgical method is different from the vocal cord polyps, and complete removal of the cyst can effectively reduce postoperative recurrence.
3. Help young doctors or non-laryngology-specialist and primary doctors to improve the accuracy of diagnosis of vocal cord lesions, so as to achieve the homogenization level of diagnosis.
4. It has a positive effect in assisting in the diagnosis of vocal cord lesions in remote consultation. In the condition of the COVID-19 pandemic, telemedicine is an option that benefits both doctors and patients.

6. Conclusion

In this paper, a computer-aided method is proposed for laryngeal disease diagnosis with small data. Previously developed computer-aided diagnosis methods have been rarely used in real settings since they only generate a prediction for the whole image and cannot tell the accurate location of disease. This makes the prediction results unreliable for clinicians. Besides, they rely on large-scale dataset and cannot be easily transferred, which hinder small hospitals to utilize them in diagnosis. To address these issues, we adopt the object detection model rather than image classification model so that both the category label and location of disease can be predicted simultaneously. Besides, we employ DropBlock technical to learn from small data. The clinicians can then easily further validate the accuracy of the prediction. To demonstrate the effectiveness of our model, we collected 279 samples

from 279 different patients to build a small dataset. Our method achieves an overall accuracy of 73.00%, which is better than the method of [11] as well as human experts.

Human and animal rights

The authors declare that the work described has not involved experimentation on humans or animals.

Funding

This work has been supported by: 1. Beijing Academy of Artificial Intelligence (BAAI) 2. National Key R&D Program of China (No. 2018AAA0100300) 3. Sino-Russian Mathematics Center.

Author contributions

All authors attest that they meet the current International Committee of Medical Journal Editors (ICMJE) criteria for Authorship.

CRediT authorship contribution statement

Shijie Fang: Coding, Writing. **Jia Fu**: Data collection, Writing. **Chen Du**: Writing. **Tong Lin**: Writing, Reviewing and Editing. **Yan Yan**: Writing, Reviewing and Editing.

Declaration of competing interest

The authors declare that they have no known competing financial or personal relationships that could be viewed as influencing the work reported in this paper.

Acknowledgements

This work was supported by National Key R&D Program of China (No. 2018AAA0100300), NSFC Tianyuan Fund for Mathematics (No. 12026606), Sino-Russian Mathematics Center, Beijing Academy of Artificial Intelligence (BAAI), and Hospital Innovation and Transformation Fund (BYSYZHKC2021102).

References

[1] Esteva A, Kuprel B, Novoa RA, Ko J, Swetter SM, Blau HM, et al. Dermatologist-level classification of skin cancer with deep neural networks. Nature 2017;542(7639):115–8.
[2] Du C, AL-Ramahi J, Liu Q, Yan Y, Jiang J. Validation of the laryngopharyngeal reflux color and texture recognition compared to ph-probe monitoring. Laryngoscope 2017:665–70.
[3] Esteva A, Chou K, Yeung S, Naik N, Madani A, Mottaghi A, et al. Deep learning-enabled medical computer vision. npj Digit Med 2021;4(1):1–9.
[4] Ghorbani A, Ouyang D, Abid A, He B, Chen JH, Harrington RA, et al. Deep learning interpretation of echocardiograms. npj Digit Med 2020;3(1):1–10.
[5] Ghorbani A, Natarajan V, Coz D, Liu Y. DermGAN: synthetic generation of clinical skin images with pathology. In: Machine learning for health workshop. Proceedings of machine learning research, vol. 116. 2019. p. 155–70.
[6] Goodfellow IJ, Pouget-Abadie J, Mirza M, Xu B, Warde-Farley D, Ozair S, et al. Generative adversarial networks. Commun ACM 2020;63(11):139–44.
[7] Voets M, Møllersen K, Bongo LA. Replication study: development and validation of deep learning algorithm for detection of diabetic retinopathy in retinal fundus photographs. CoRR arXiv:1803.04337.
[8] Rajpurkar P, Irvin J, Bagul A, Ding D, Duan T, Mehta H, et al. Mura: large dataset for abnormality detection in musculoskeletal radiographs. arXiv preprint arXiv:1712.06957, 2017.
[9] Setio AAA, Traverso A, et al. Validation, comparison, and combination of algorithms for automatic detection of pulmonary nodules in computed tomography images: The LUNA16 challenge. Med Image Anal 2017;42:1–13.
[10] Yao P, Witte D, Gimonet H, German A, Andreadis K, Cheng M, et al. Automatic classification of informative laryngoscopic images using deep learning. Laryngoscope Invest. Otolaryngol. 2022;7(2):460–6.
[11] Ren J, Jing X, Wang J, Ren X, Xu Y, Yang Q, et al. Automatic recognition of laryngoscopic images using a deep-learning technique. Laryngoscope 2020;130(11):E686–93.
[12] He K, Zhang X, Ren S, Sun J. Deep residual learning for image recognition. In: IEEE conference on computer vision and pattern recognition; 2016. p. 770–8.
[13] Cao X, Bu W, Huang S, Tang Y, Guo Y, Chang Y, et al. A survey of learning on small data. CoRR arXiv:2207.14443.
[14] Ren S, He K, Girshick RB, Sun J. Faster R-CNN: towards real-time object detection with region proposal networks. In: Advances in neural information processing systems; 2015. p. 91–9.
[15] Cao X, Yuan P, Feng B, Niu K. CF-DETR: coarse-to-fine transformers for end-to-end object detection. In: Proceedings of the thirty-sixth AAAI conference on artificial intelligence; 2022. p. 185–93.
[16] Cai Z, Vasconcelos N. Cascade R-CNN: delving into high quality object detection. In: IEEE conference on computer vision and pattern recognition; 2018. p. 6154–62.
[17] Zhang H, Wu C, et al. ResNeSt: split-attention networks. CoRR arXiv:2004.08955 [abs], 2020.
[18] Deng J, Dong W, Socher R, Li L, Li K, Li F. Imagenet: a large-scale hierarchical image database. In: IEEE conference on computer vision and pattern recognition; 2009. p. 248–55.
[19] Ghiasi G, Lin T, Le QV. Dropblock: a regularization method for convolutional networks. In: Advances in neural information processing systems; 2018. p. 10750–60.
[20] Ghiasi G, Lin T, Le QV. NAS-FPN: learning scalable feature pyramid architecture for object detection. In: IEEE conference on computer vision and pattern recognition; 2019. p. 7036–45.
[21] Wu Y, Chen Y, Yuan L, Liu Z, Wang L, Li H, et al. Rethinking classification and localization for object detection. In: IEEE conference on computer vision and pattern recognition; 2020. p. 10183–92.