Deep Neural Networks for Automatic Detection and Classification of Laryngeal Pathologies in Endoscopic Imaging

Jan Nowak
Adam Mickiewicz University; Poznan Supercomputing and Networking Center
Poznan, Poland *jnowak@man.poznna.pl*

Mikolaj Buchwald
Poznan Supercomputing and Networking Center
Poznan, Poland *mbuchwald@man.poznan.pl*

Szymon Kupinski
Poznan Supercomputing and Networking Center
Poznan, Poland *szymonk@man.poznan.pl*

Juliusz Pukacki
Poznan Supercomputing and Networking Center
Poznan, Poland *pukacki@man.poznan.pl*

Hanna Klimza
Regional Specialist Hospital Wroclaw Research & Development Centre
Wroclaw, Poland *haniaklimza@gmail.com*

Piotr Nogal
Poznań University of Medical Sciences
Poznan, Poland *pionogch@gmail.com*

Joanna Jackowska
Poznań University of Medical Sciences
Poznan, Poland *drjackowska@gmail.com*

Małgorzata Wierzbicka
Wroclaw University of Science and Technology
Wroclaw, Poland *wierzbicka.otolaryngology@gmail.com*

Krzysztof Dyczkowski
Adam Mickiewicz University
Poznan, Poland *chris@amu.edu.pl*

Abstract

This study explores the application of artificial intelligence (AI) methods for the automated detection and classification of laryngeal pathologies in fiberoptic laryngoscopy videos. From recordings of 292 patients, a total of 885 informative image frames were automatically extracted, and subsequently segmented manually by experienced clinicians. Seven distinct pathology categories were examined using two deep learning models, Mask R-CNN, designed for classification, object detection, and segmentation tasks; and EfficientNet V2L, solely for classification. For the classification task, an across-class average imbalance-resistant F1-score was higher for Mask R-CNN model, 0.95 (confidence interval, CI: 0.90–0.98), than for EfficientNet V2L 0.74 (CI: 0.66-0.81; McNemar's test $p<0.001$). In object detection, a mean average precision of 0.36 (CI: 0.35-0.37) was achieved at an intersection over union threshold of 50%. However, segmentation models reached lower performance, average precision 0.29 (0.28-0.30). In sum, for the larynx pathology analysis, DNNs show more potential for classification than segmentation tasks, with an advantage of Mask R-CNN over EfficientNet architecture.

Keywords: deep neural networks, laryngology, pathology classification, medical image segmentation, endoscopic imaging

1. Introduction

Laryngeal pathologies such as tumors, polyps, or cysts can contribute to the worsening of the patient's condition, including the disruption of phonation function of the larynx [9]. Early diagnosis of these lesions improves clinical outcomes through timely intervention [5].

As one of the means of evaluating the pathologies, an endoscopic camera introduced through the nasal cavity of the patient is used to assess the larynx, and the vocal folds in particular [1], [4]. During the examination, the physician may ask the patient to say a certain vowel, for a more accurate diagnosis. In addition, the physician can adjust the parameters of the endoscopic camera, e.g., by changing the range of light the camera is displaying/recording. Two of the most commonly used spectra of light are white light (WL), and narrow-band imaging (NBI).

The current study aims to evaluate the potential of artificial intelligence (AI) in enhancing medical diagnostics in laryngology [13]. Utilizing deep neural networks (DNNs) during the endoscopic evaluation of the respiratory tract can significantly accelerate patient diagnosis processes [12]. Here, we evaluated the performance of two widely used architectures—EfficientNet V2L and Mask R-CNN for full-image classification, as well as for object detection and segmentation (Mark R-CNN only). Our aim was to determine whether current DNN solutions can support and/or automate assessment of the pathologies while identifying current limitations of DNN technology in this regard.

2. MATERIALS AND METHODS

2.1. Dataset

Data were collected from 292 patient cases at the Poznan University of Medical Sciences (PUMS) and anonymized, yielding 885 high-quality expert-annotated laryngeal endoscopic frames. Segmentation was performed by experienced otolaryngologists using Label Studio [3], based on 7 pathology classes: squamous cell carcinoma (SCC) – 207 samples, cyst – 61 samples, dysplasia/carcinoma in situ (DCS) – 69 sample, keratosis without atypia – 36 samples, recurrent respiratory papillomatosis (RRP) – 675 samples, polyp – 67 samples, and Reinke's edema – 21 samples. The images were either from white light (WL) setting, or narrow-band imaging (NBI) spectrum.

2.2. Model development

The balanced ratio for all training and validation dataset was 80% to 20%, meaning, e.g., that for squamous cell carcinoma there were 207 samples in the training dataset, and 44 in the validation dataset, etc. The models were trained on the 80% of the available dataset, and then tested on a 20% validation subset.

Two architectures of DNNs were used to perform the research, Mask R-CNN implemented on the Detectron2 platform with PyTorch (X_101_32x8d_FPN_3x) [2], [16], and EfficientNet V2L architecture [8], [11], [14] implemented with the Keras framework and TensorFlow.

2.3. Preprocessing and Data Augmentation

Underexposed, smeared, overexposed and reflection frames were removed, according to the methodology described elsewhere [4], and only informative frames were retained.

For datasets trained on the Mask R-CNN model running on the Detectron2 framework, no data augmentation was applied. For datasets used in the TensorFlow framework for the EfficientNet V2L model, images were transformed with normalization, random rotation with a maximum of 20 degrees, approximation with a maximum degree of 0.2, and random horizontal flipping.

2.4. Training Parameters

Training parameters for the Mask R-CNN model used in the study were the following: batch size=2, Learning Rate=0.0125, number of epochs=30, number of classes=7, Backbone freeze at 2. Parameters for the EfficientNet V2L model were: batch size=[**8**,16,32], learning rate=[0.00001, **0.00002**, 0.00005], number of epochs=1860, dropout rate=[**0.1**, 0.2, 0.3, 0.4, 0.5], and L2=[**0.001**, 0.002, 0.005]. This include the initial set of parameters, and the best set selected with Keras-Tuner [8] (with **bolded** font), which was used for the final training. The number of epochs was selected based on previous experience of the research team and could potentially be increased in further studies.

2.5. Model evaluation

To evaluate the performance of the trained models, the following metrics were used: (1) for classification: precision, recall, F1 score; (2) for object detection and lesion segmentation: average precision (AP), average recall (AR), and intersection under union (IoU) [10]. For AP and AR, IoU was set to IoU ≥ 0.5 with a cap of 0.05 to IoU ≤ 0.95, which means that for a detection to be considered correct, the common part of the object detected by the model and the real object is a minimum of 50% of the total detection area and the real object together, and so the common part for each IoU threshold is summed and then divided by the number of thresholds, making them bet the average AP [15]. The difference in the performance of the two analyzed models was evaluated with the McNemar's test.

3. Model Results

3.1. Classification results of Mask R-CNN model X_101_32x8d_FPN_3x vs EfficientNet V2L

For the classification task, an cross-class average imbalance-resistant F1-score for Mask R-CNN model was 0.95 (confidence interval, CI: 0.90–0.98). The complete results for the Mask R-CNN model are presented in the left panel of the Table 1, by class and across all classes (averaged).

The class-averaged F1-score for the EfficientNet V2L was 0.74 (CI: 0.66-0.81). Table 1, right panel, shows all the results for the EfficientNet V2L model.

The analyzed models differed in terms of performance (χ^2 = 26.7, p<0.001).

3.2. Mask R-CNN model object detection results X_101_32x8d_FPN_3x

In object detection, a mean average precision of 0.36 (CI: 0.35-0.37) was achieved for an intersection over union threshold of 50%.

Table 1. Comparison of classification results for Mask R-CNN and EfficientNet V2L on the laryngeal pathology dataset. Metrics shown: precision, recall, and F1-score with 95% confidence intervals.

Class	**Mask R-CNN**			**EfficientNet V2L**		
	Precision	Recall	F1-score	Precision	Recall	F1-score
Cyst	1.00 (1.00–1.00)	0.91 (0.70–1.00)	0.95 (0.82–1.00)	0.86 (0.50–1.00)	0.46 (0.20–0.75)	0.60 (0.31–0.80)
Dysplasia/carcinoma in situ	1.00 (1.00–1.00)	0.75 (0.50–1.00)	0.86 (0.67–1.00)	1.00 (1.00–1.00)	0.46 (0.20–0.77)	0.63 (0.33–0.87)
Keratosis without atypia	1.00 (1.00–1.00)	0.67 (0.25–1.00)	0.80 (0.40–1.00)	0.67 (0.00–1.00)	0.33 (0.00–0.80)	0.44 (0.00–0.80)
Papillomatosis (RRP)	0.95 (0.88–0.99)	1.00 (1.00–1.00)	0.97 (0.94–0.99)	0.77 (0.67–0.86)	0.89 (0.81–0.96)	0.83 (0.75–0.89)
Polyp	1.00 (1.00–1.00)	0.92 (0.71–1.00)	0.96 (0.83–1.00)	0.59 (0.40–0.80)	0.93 (0.77–1.00)	0.72 (0.55–0.88)
Reinke's edema	1.00 (1.00–1.00)	1.00 (1.00–1.00)	1.00 (1.00–1.00)	1.00 (0.00–1.00)	0.67 (0.00–1.00)	0.80 (0.00–1.00)
Squamous cell carcinoma	0.92 (0.83–1.00)	0.98 (0.93–1.00)	0.95 (0.89–0.99)	0.72 (0.59–0.85)	0.69 (0.56–0.81)	0.71 (0.60–0.80)
Macro avg	0.98 (0.96–0.99)	0.89 (0.80–0.96)	0.93 (0.85–0.97)	0.80 (0.61–0.88)	0.63 (0.51–0.75)	0.68 (0.52–0.77)
Weighted avg	0.95 (0.92–0.98)	0.95 (0.91–0.98)	**0.95 (0.90–0.98)**	0.77 (0.71–0.83)	0.75 (0.68–0.82)	**0.74 (0.66–0.81)**

Unbalanced accuracy results across all classes (i.e., all correctly classified samples over all samples) were the following: for Mask R-CNN: 0.95 (0.91–0.98); for EfficientNet V2L: 0.75 (0.68–0.82). In parentheses, 95% confidence intervals are reported. Abbreviations: DCS – dysplasia/carcinoma in situ, RRP – papillomatosis, SCC – squamous cell carcinoma.

3.3. Mask R-CNN model segmentation results X_101_32x8d_FPN_3x

The Mask R-CNN achieved a mean average precision (mAP@[0.50:0.95]) of 0.29 (CI: 0.28-0.30) and an average recall of 0.36 (CI: 0.29-0.43).

4. Conclusion

In this study we demonstrated the capabilities and compared two deep neural networks architectures for the automatic detection and classification of laryngeal pathologies in a real-life medical dataset.

The classification task was the easiest for DNN, as compared to object detection and segmentation. Mask R-CNN proved better than EfficientNet V2L in this task, suggesting that the use of faster R-CNN as the classification backbone within the Mask R-CNN architecture may offer better suitability or performance for medical image classification tasks. In a similar work where EfficientNetV2L-LGBM model was used for classification, validation accuracy of 0.97 was obtained, compared to 0.75 in our study, indicating overfitting in the first case [6].

The classification of objects from bounding box areas yielded moderate object detection performance, indicating that while the model can assist in lesion pre-identification, careful clinician verification remains essential to ensure diagnostic accuracy [13]. The segmentation results indicate even more limited model capabilities, as for most pathology categories the results were unsatisfactory [9].

The presented results have undergone rigorous verification, also in cooperation with medical experts. Of the analyzed use-cases, classification, especially with Masked R-CNN yielded the most promising results, in terms of potential future applications such as diagnosis support or treatment monitoring [7].

References

[1] Azam, M. A., Sampieri, C., Ioppi, A., Africano, S., Vallin, A., Mocellin, D., Fragale, M., Guastini, L., Moccia, S., Piazza, C., Mattos, L. S., and Peretti, G.: Deep Learning Applied to White Light and Narrow Band Imaging Videolaryngoscopy: Toward Real-Time Laryngeal Cancer Detection. In: *The Laryngoscope* 132.9 (2022), pp. 1798–1806.

[2] He, K., Gkioxari, G., Dollár, P., and Girshick, R.: *Mask R-CNN*. 2018. arXiv: `1703.06870 [cs.CV]`. URL: `https://arxiv.org/abs/1703.06870`.

[3] *Label Studio*. `https://labelstud.io/`. Accessed July 01, 2025. 2020.

[4] Moccia, S., Vanone, G. O., Momi, E. D., Laborai, A., Guastini, L., Peretti, G., and Mattos, L. S.: Learning-based classification of informative laryngoscopic frames. In: *Computer Methods and Programs in Biomedicine* 158 (2018), pp. 21–30. URL: https://www.sciencedirect.com/science/article/pii/S0169260717312130.

[5] Montenegro, C., Mattavelli, D., Lancini, D., Paderno, A., Marazzi, E., Rampinelli, V., Tomasoni, M., and Piazza, C.: Treatment and outcomes of minor salivary gland cancers of the larynx and trachea: a systematic review. In: *ACTA Otorhinolaryngologica Italica* 43.6 (2023), p. 365.

[6] Nobel, S. M. N., Swapno, S. M. M. R., Islam, M. R., Safran, M., Alfarhood, S., and Mridha, M. F.: A machine learning approach for vocal fold segmentation and disorder classification based on ensemble method. In: *Scientific Reports* 14 (2024). URL: https://doi.org/10.1038/s41598-024-64987-5.

[7] Nogal, P., Buchwald, M., Staśkiewicz, M., Kupiński, S., Pukacki, J., Mazurek, C., Jackowska, J., and Wierzbicka, M.: Endoluminal larynx anatomy model – towards facilitating deep learning and defining standards for medical images evaluation with artificial intelligence algorithms. ENG. In: *Polish Journal of Otolaryngology* 76.5 (2022), pp. 37–45.

[8] O'Malley, T., Bursztein, E., Long, J., Chollet, F., Jin, H., Invernizzi, L., et al.: *KerasTuner*. https://github.com/keras-team/keras-tuner. 2019.

[9] Paderno, A., Piazza, C., Del Bon, F., Lancini, D., Tanagli, S., Deganello, A., Peretti, G., De Momi, E., Patrini, I., Ruperti, M., Mattos, L. S., and Moccia, S.: Deep Learning for Automatic Segmentation of Oral and Oropharyngeal Cancer Using Narrow Band Imaging: Preliminary Experience in a Clinical Perspective. In: *Frontiers in Oncology* Volume 11 - 2021 (2021). URL: https://www.frontiersin.org/journals/oncology/articles/10.3389/fonc.2021.626602.

[10] Ren, S., He, K., Girshick, R., and Sun, J.: *Faster R-CNN: Towards Real-Time Object Detection with Region Proposal Networks*. 2016. arXiv: 1506.01497 [cs.CV]. URL: https://arxiv.org/abs/1506.01497.

[11] Russakovsky, O., Deng, J., Su, H., Krause, J., Satheesh, S., Ma, S., Huang, Z., Karpathy, A., Khosla, A., Bernstein, M., Berg, A. C., and Fei-Fei, L.: *ImageNet Large Scale Visual Recognition Challenge*. 2015. arXiv: 1409.0575 [cs.CV]. URL: https://arxiv.org/abs/1409.0575.

[12] Sampieri, C., Azam, M. A., Ioppi, A., Baldini, C., Moccia, S., Kim, D., Tirrito, A., Paderno, A., Piazza, C., Mattos, L. S., and Peretti, G.: Real-time laryngeal cancer boundaries delineation on white light and narrow-band imaging laryngoscopy with deep learning. In: *The Laryngoscope* 134.6 (2024), pp. 2826–2834. URL: https://doi.org/10.1002/lary.30710.

[13] Sampieri, C., Baldini, C., Azam, M. A., Moccia, S., Mattos, L. S., Vilaseca, I., Peretti, G., and Ioppi, A.: Artificial Intelligence for Upper Aerodigestive Tract Endoscopy and Laryngoscopy: A Guide for Physicians and State-of-the-Art Review. In: *Otolaryngology–Head and Neck Surgery* 169.4 (2023), pp. 811–829.

[14] Tan, M. and Le, Q. V.: *EfficientNetV2: Smaller Models and Faster Training*. 2021. arXiv: 2104.00298 [cs.CV]. URL: https://arxiv.org/abs/2104.00298.

[15] Tanimoto, T.: An Elementary Mathematical Theory of Classification and Prediction. International Business Machines Corporation, 1958. URL: https://books.google.pl/books?id=yp34HAAACAAJ.

[16] Wu, Y., Kirillov, A., Massa, F., Lo, W.-Y., and Girshick, R.: *Detectron2*. https://github.com/facebookresearch/detectron2. 2019.