ORIGINAL RESEARCH

Laryngoscope
Investigative Otolaryngology

Automatic classification of informative laryngoscopic images using deep learning

Peter Yao BSE [ORCID] | Dan Witte BA | Hortense Gimonet BA |
Alexander German BS | Katerina Andreadis MSc [ORCID] | Michael Cheng MD |
Lucian Sulica MD [ORCID] | Olivier Elemento PhD | Josue Barnes BS |
Anaïs Rameau MD, MPhil [ORCID]

Department of Otolaryngology-Head and Neck Surgery, Sean Parker Institute for the Voice, Weill Cornell Medicine, New York, New York, USA

Correspondence
Anaïs Rameau, Department of Otolaryngology—Head and Neck Surgery, Sean Parker Institute for the Voice, Weill Cornell Medicine, 240 East 59th St, New York, NY 10022, USA.
Email: anr2783@med.cornell.edu

Abstract

Objective: This study aims to develop and validate a convolutional neural network (CNN)-based algorithm for automatic selection of informative frames in flexible laryngoscopic videos. The classifier has the potential to aid in the development of computer-aided diagnosis systems and reduce data processing time for clinician-computer scientist teams.

Methods: A dataset of 22,132 laryngoscopic frames was extracted from 137 flexible laryngostroboscopic videos from 115 patients. 55 videos were from healthy patients with no laryngeal pathology and 82 videos were from patients with vocal fold polyps. The extracted frames were manually labeled as informative or uninformative by two independent reviewers based on vocal fold visibility, lighting, focus, and camera distance, resulting in 18,114 informative frames and 4018 uninformative frames. The dataset was split into training and test sets. A pre-trained ResNet-18 model was trained using transfer learning to classify frames as informative or uninformative. Hyperparameters were set using cross-validation. The primary outcome was precision for the informative class and secondary outcomes were precision, recall, and F1-score for all classes. The processing rate for frames between the model and a human annotator were compared.

Results: The automated classifier achieved an informative frame precision, recall, and F1-score of 94.4%, 90.2%, and 92.3%, respectively, when evaluated on a hold-out test set of 4438 frames. The model processed frames 16 times faster than a human annotator.

Conclusion: The CNN-based classifier demonstrates high precision for classifying informative frames in flexible laryngostroboscopic videos. This model has the

potential to aid researchers with dataset creation for computer-aided diagnosis systems by automatically extracting relevant frames from laryngoscopic videos.

KEYWORDS

artificial intelligence, computer vision, computer-aided diagnosis, laryngology, machine learning, vocal fold polyp

1 | INTRODUCTION

Machine learning is the process of developing algorithms that have the ability to learn from example data and make predictions on new data.[1] Deep learning is a subfield of machine learning that uses artificial neural networks, algorithms that learn patterns from large amounts of data and make predictions based on those inputs.[2] The applications of deep learning have accelerated in the past 10 years thanks to advances in computing power and new large datasets.[3] In particular, deep learning algorithms have demonstrated substantial success in predicting diagnoses from medical images in colonoscopy,[4] radiology,[5] dermatology,[6] ophthalmology,[7] and pathology.[8]

Flexible laryngoscopy is one of the most common procedures performed by otolaryngologists to evaluate the nasal, nasopharyngeal, oropharyngeal, laryngeal, and hypopharyngeal structures and provides a rich source for images for deep learning models. Given the insufficient number of otolaryngologists to meet healthcare needs both in the USA and globally,[9–12] there is a large potential for machine learning systems to assist in the diagnosis of otolaryngologic disease, especially in areas where specialists are difficult to access. Focusing on laryngology initially, we aim to bridge this gap in otolaryngologic care by building a deep learning model for automatic informative frame selection, a key component of an end-to-end framework for the automated classification of laryngeal lesions on video-laryngoscopy.

Informative frames are keyframes, or a set of salient images for diagnostic accuracy extracted from the underlying video source.[13] Due to the widespread use of laryngoscopy, extracting informative frames from laryngoscopy videos is an effective method for creating datasets for laryngeal lesion classification. However, not all frames in a laryngoscopy video are useful. For the purpose of laryngeal lesion recognition and characterization, only frames that provide a full view of the vocal folds and are of sufficient quality to appreciate the vocal folds are informative. There are various methods for extracting informative frames. Frames may be extracted manually, but the process is labor intensive, time consuming, and subject to reviewer bias. The simplest automated methods are uniform or random sampling of frames from video, however these methods may incorporate uninformative or redundant frames into the dataset.

In laryngology, informative frame selection via deep learning has previously been applied to narrow-band imaging (NBI) video-laryngoscopy, albeit on a very small dataset (720 frames extracted from 18 laryngoscopy videos) prone to overfitting, which is when a model fits the training data well but may perform unreliably on new data.[14–16] We aim to expand on this approach by using a dataset that is an order of magnitude larger than previous studies to develop an informative frame classifier for use with images obtained from flexible laryngostroboscopy, a more commonly used imaging modality than NBI video-laryngoscopy. An informative frame classifier has several uses. First, it can assist laryngologist-computer scientist teams with efficiently creating datasets for machine learning models of disease classification. Similarly, it can also facilitate the assembly of large-scale public datasets for use in validating and testing other models. Furthermore, it can reduce the time laryngologists or trained researchers need to review and label endoscopic videos by automatically tagging informative sections.

Our hypothesis is that automated informative frame selection from video-laryngoscopy is possible and can accelerate the application of computer vision to laryngoscopy. Our objective is to develop and validate a deep learning model for informative frame classification in flexible laryngoscopy images, with informative class precision as the primary outcome measure (Figure 1). Our long-term goal is to use automated informative frame selection to facilitate the development of deep learning models for automated diagnosis that improve the delivery of otolaryngologic care in communities lacking local specialists.

2 | MATERIALS AND METHODS

2.1 | Data

This project was approved by the Weill Cornell Medicine Institutional Review Board (protocol #19-05020151). From our institutional database of flexible distal chip strobolaryngoscopic videos, we identified 55 videos of healthy vocal folds and 82 videos with unilateral vocal fold polyps from 115 patients. Multiple videos from the same patient were included but there was no patient overlap between the healthy and polyp groups. Diagnoses were assigned by the physician at the time of examination and identified using the Sean Parker Institute for the Voice database. Exams were performed by three laryngologists on adult patients in a clinic setting with a Pentax VNL-1590STi naso-pharyngo-laryngoscope. The dates of the videos range from July 13, 2017 to September 10, 2020.

2.2 | Data labels

For the purpose of laryngeal lesion classification, we defined informative frames as images with a full view of the vocal folds and of sufficient quality to appreciate the vocal folds. Moccia et al. described
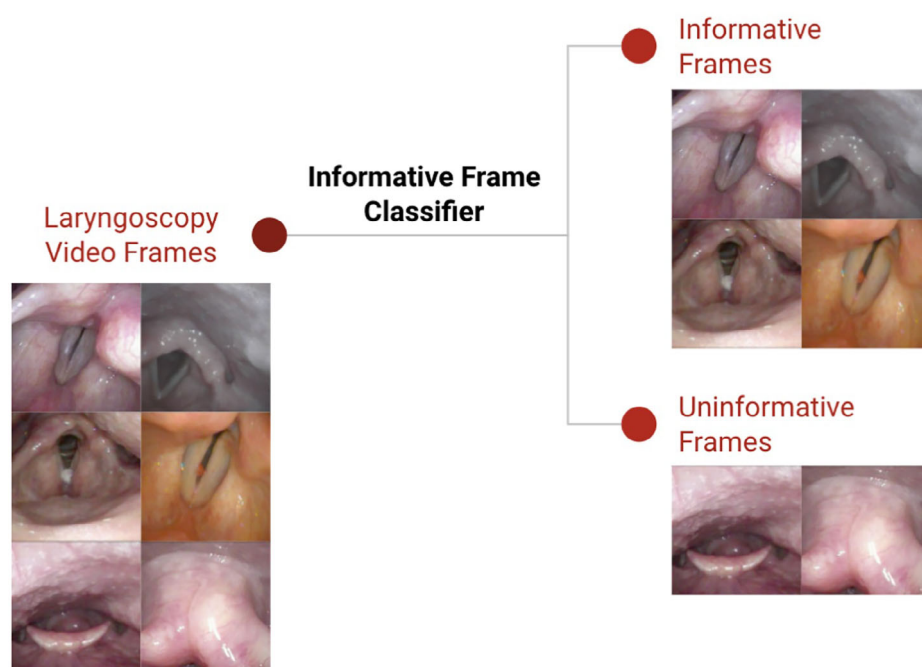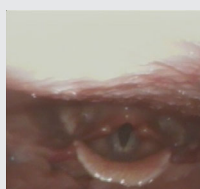
TABLE 1 Selection criteria for informative frames

Criterium	Description	Informative	Uninformative
Vocal fold visibility	If abducted, bilateral vocal folds are visible from vocal process to anterior commissure. (If adducted, 80% visibility is acceptable)		
Lighting and exposure	Vocal folds are well exposed by light such that the vocal folds are clearly distinguished from surrounding structures.		
Focus	Vocal folds are in focus with minimal blurring.		
Camera distance	Camera is appropriate distance from vocal folds such that details on vocal folds are discernable.		

three categories of uninformative frames in laryngoscopic videos: (1) blurred frames due to motion, (2) frames with specular reflections due to secretions, and (3) underexposed frames due to varying illumination conditions.[14] In conjunction with a laryngologist [AR], we developed criteria based on vocal fold visibility, lighting, focus, and camera distance to distinguish informative and uninformative frames (Table 1). We extracted 22,132 laryngoscopic frames from 137 videos in the database. The extracted frames were manually labeled as

informative or uninformative by two independent reviewers resulting in 18,114 informative frames and 4018 uninformative frames. The two annotators [PY, AG] were given informative and uninformative examples for reference. Any conflicts in labeling were resolved by consensus with input from the laryngologist.

2.3 | Classifier

Our labeled dataset was split into training and test sets by patient in an 80:20 ratio. A ResNet-18 model, a type of convolutional neural network (CNN) architecture, pre-trained on ImageNet was trained using transfer learning to classify frames as informative or uninformative (Figure S1).[17] In machine learning, a loss function maps decisions to their associated costs, an optimizer tunes the model in response to the output of the loss function, and weight decay is a regularization technique used to avoid overfitting. In our model, cross-entropy loss was used as the loss function and Adam[18] was used as the model optimizer with a weight decay of 1E-5. Hyperparameters are values used to control the learning process. Hyperparameters tuned using cross-validation included the learning rate (how fast the model learns), number of epochs (how long the model trains), and number of frozen layers (the degree to which we keep the model weights from transfer learning). During cross-validation the training data was split into five partitions, keeping frames from the same video in the same partition to avoid overfitting.

	Predicted	
	Negative	Positive
Negative	True Negative (TN)	False Positive (FP)
Positive	False Negative (FN)	True Positive (TP)

(with **Actual** label spanning the Negative/Positive rows)

$$Recall = \frac{TP}{TP + FN} \qquad Precision = \frac{TP}{TP + FP} \qquad F1 - Score = \frac{2TP}{2TP + FP + FN}$$

FIGURE 2 Definition of quantitative performance metrics: recall, precision, and F1-score

2.4 | Gradient-weighted class activation mapping

Gradient-weighted class activation mapping (Grad-CAM) is a technique for improving model transparency by producing visual explanations for decisions from CNN-based models.[19] Grad-CAM was applied to identify important regions in the image that the informative frame classifier used for predictions.

TABLE 2 Precision, recall, and F1-score for informative and uninformative classes

	Precision	Recall	F1-Score
Informative	0.94	0.90	0.92
Uninformative	0.63	0.76	0.69

FIGURE 4 Confusion matrix of the performance of the informative frame classifier on the test set. The individual table cell values represent the number of images in each category

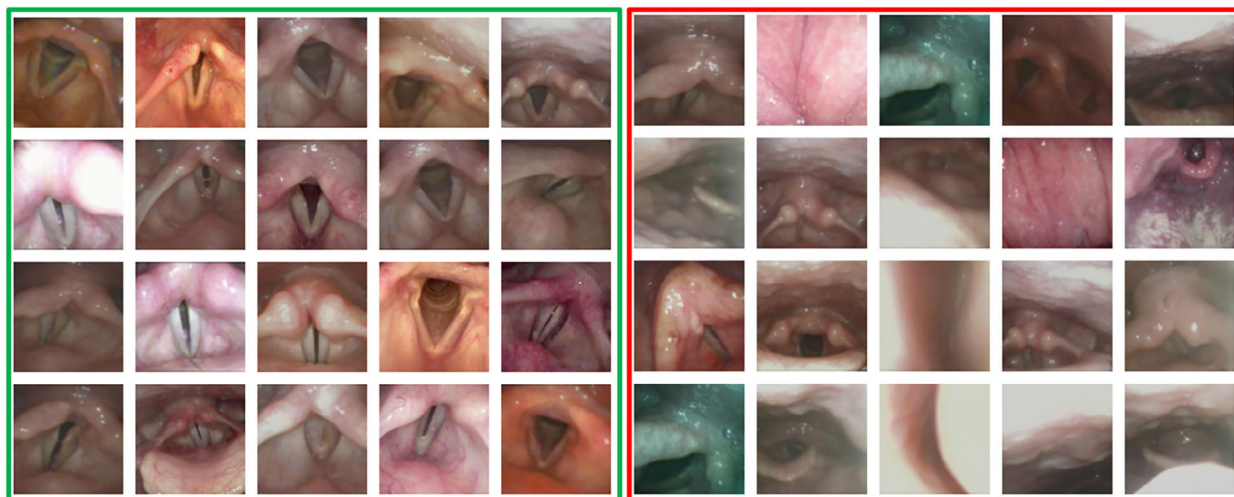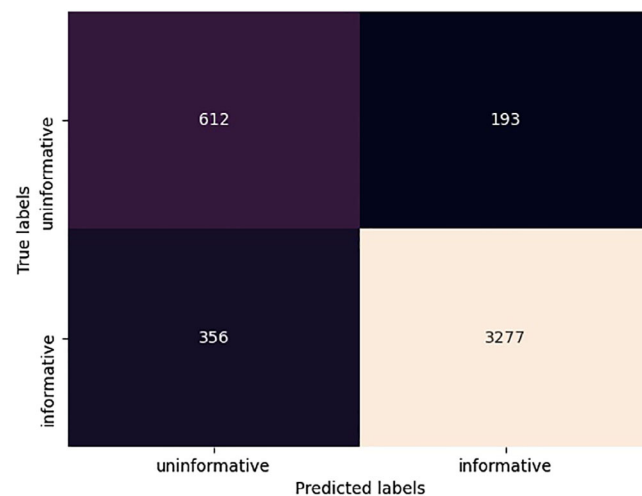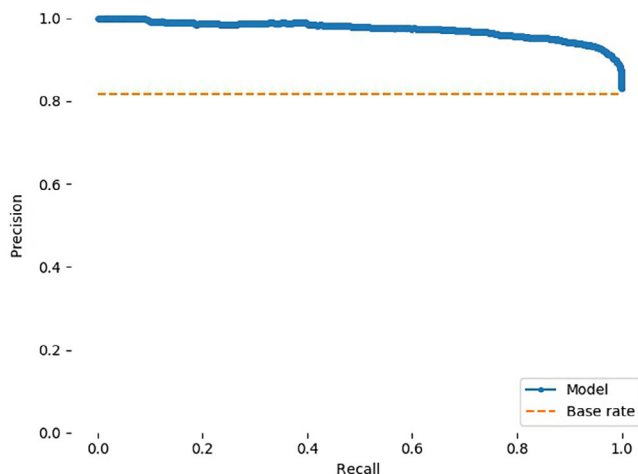**FIGURE 3** A representative sample of frames correctly classified as informative (green border) and uninformative (red border)

FIGURE 5 Precision-recall curve which summarizes the trade-off between the true positive rate and the positive predictive value for our model using different probability thresholds
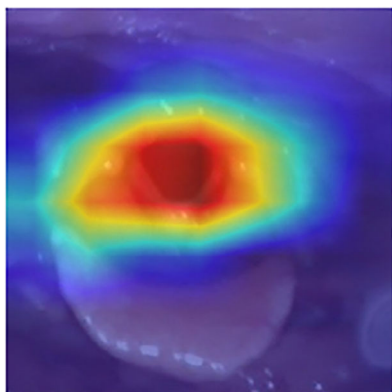

FIGURE 6 An attention map generated by Grad-CAM overlaid on an image correctly predicted to be informative. The model's focus, as illustrated by the warm colors, is on relevant key structures, particularly the glottis and vocal folds

2.5 | Outcomes

The primary outcome measure was precision (also called positive predictive value) for the informative class. We chose this as our primary metric to minimize the number of false positives (uninformative frames classified as informative), which are detrimental to the model's diagnostic accuracy. The secondary outcome metrics were recall (also called the true positive rate), precision, and F1-score (a combination measure of precision and recall) for all classes, which are defined in Figure 2. The time required for the model to evaluate 10,000 frames was also measured compared with the speed of a human annotator. The model was implemented using the Pytorch library in the Python programming language and run on a server with a NVIDIA Tesla V100 32GB graphics card and 64GB RAM.

3 | RESULTS

The labeling of informative and uninformative frames by two trained annotators demonstrated substantial inter-rater reliability with a Cohen's kappa of 0.69. The model was trained using a learning rate of 1E-4 for 5 epochs with three layer groups frozen. The classifier achieved an average informative frame precision of 92.7% on the left-out partition during cross-validation and 94.4% on the test set. The model was able to classify 31.5 frames per second, which is approximately 16 times faster than a human annotator who we measured to have an output of 2 frames per second. A representative sample of correctly classified frames are displayed in Figure 3. Outcomes metrics for the informative and uninformative classes are detailed in Table 2. The confusion matrix is reported in Figure 4. The precision-recall curve is reported in Figure 5. A representative heatmap generated using Grad-CAM for an image correctly predicted as informative by the classifier is shown in Figure 6.

4 | DISCUSSION

We developed and validated a deep learning classifier capable of automatically identifying informative frames in flexible laryngoscopy videos with 94.4% precision. This classifier addresses a critical need for an efficient and accurate method to automatically select informative frames for inclusion in laryngoscopic deep learning datasets.

Flexible laryngoscopy yields a rich set of images ideal for deep learning applications. Yet, a major challenge is the considerable time and labor necessary to assemble datasets of laryngoscopy images to train diagnostic machine learning models. Previous models for laryngeal lesion recognition have relied on manual extraction of informative frames to build training datasets. For example, Ren et al. required seven reviewers examining over 24,000 images to build a dataset to train a classifier to recognize five classes of laryngeal pathology.[20] In another study of a classifier for laryngeal cancer, one endoscopist manually reviewed over 14,000 images for the dataset.[21]

Our classifier provides an automated and accurate method of selecting informative frames without human input. Automatic identification of informative frames in laryngoscopy is underdeveloped in comparison to other endoscopic modalities. For example, machine learning has been successfully applied to identify informative and uninformative frames in colonoscopy videos as an important first step toward performing automated pathology detection.[22–24]

Related studies have used machine learning to identify informative frames in NBI laryngoscopy videos, demonstrating the potential for automated informative frame selection to be applied to other laryngoscopic imaging modalities.[14–16] Moccia et al. achieved a classification recall of 91% for informative frames using support vector machines on a dataset of 720 frames from 18 NBI laryngoscopic videos.[14] Patrini et al. and Galdran et al. expanded on this work by using a CNN-based classifier to achieve a classification recall of 98% and 100%, respectively, for informative frames on the

same dataset.[15,16] Our work adds to previous approaches in several ways. First, we apply informative frame classification to flexible laryngoscopy broadly. NBI is a specialized optical technique that enhances the diagnostic capability of endoscopy in characterizing precancerous and cancer laryngeal lesions.[25] However, limited access to endoscopes equipped with specialized filters and additional training time limits the clinical applications of NBI endoscopy, especially in surgical deserts and low-income countries.[21] On the other hand, white-light laryngoscopy is a more commonly used imaging modality that is broadly accessible and has a wider range of clinical and research applications. Second, previous models were trained and validated on a small dataset of 720 frames from 18 videos. Small datasets are prone to overfitting especially when used with complex models such as CNNs. Furthermore, the small dataset is unlikely to capture the full range of viewpoints, lighting, background, and scale present in real-world informative and uninformative frames. We addressed this shortcoming by creating a larger dataset of 22,132 frames from 137 videos, an order of magnitude larger than the dataset used by Moccia et al. Additionally, we applied Grad-CAM to enhance the transparency of our CNN model and investigate the region of the image that contributed most to decisions by the classifier. As intended, Grad-CAM (Figure 6) demonstrates that the classifier uses the region of the image encompassing the vocal folds to predict informativeness. Lastly, previous datasets only contained images from patients with laryngeal squamous cell carcinoma. Our model extends informative frame selection to images from healthy patients and patients with vocal fold polyps.

Automated informative frame selection will play a critical role in accelerating the application of computer vision and machine learning to laryngoscopy. Primarily, it can assist with efficiently creating datasets for machine learning with limited human intervention. To test the application of the informative frame classifier to disease classification models in future work, we intend to use the automatically extracted informative frames to build a vocal fold lesion classifier. This future work would demonstrate the integration of the informative frame classifier into an end-to-end system for laryngeal disease classification.

Our study has several limitations. First, despite our best efforts to develop criteria to define the two classes, informativeness is a spectrum with no clear boundary and some frames exist in the gray area between informative and uninformative. Therefore, in future work it may be helpful to predict informativeness as a continuous rather than categorical variable. We also recognize that informativeness is subjective and may vary by laryngologist and application of interest. Our definition of informative frames was specifically tailored toward extracting frames useful for laryngeal lesion recognition. However, researchers can customize our work as needed by retraining the model on frames labeled by criteria for informativeness that best suit their application. In addition, data augmentation, a technique that aims to increase the size and diversity of the training dataset by applying transformations to the input images, was not applied to our dataset because we found that data augmentation did not meaningfully improve the performance of the model. Furthermore, the informative frame classifier was developed using laryngoscopy data from a single institution and may not generalize to other laryngoscopy videos

broadly. However, we believe this limitation is mitigated by use of a large dataset relative to previous publications on this topic and adherence to pre-defined criteria for labeling frames that were developed in collaboration with a laryngologist. Also, to increase dataset size and diversity of frames, we included frames from multiple exam videos for some patients. Only 16.5% of patients (19 out of 115 patients) contributed multiple exams, whereas 83.5% of patients only contributed one exam to the dataset. To avoid data leakage, or accidental sharing of information between training and test sets with the potential to skew performance scores, we were careful to split the data by patient when assigning frames to training and test sets so that frames from the same patient would not appear in both the training and test sets. When sampling frames from the same exam, we made sure to sample frames across the full length of videos to include representative frames throughout the laryngoscopic exam, rather than sampling consecutive frames which may be similar to each other. Lastly, the model was trained on frames of healthy vocal folds and vocal folds with polyps due to our intent to use the frames for vocal fold polyp classification in future work. The informative frame selection may not generalize to other pathologies. Nonetheless, our informative frame labeling criteria were pathology agnostic and our classifier can be easily retrained on labeled frames containing other pathologies.

In summary, we developed a deep learning model for automated selection of informative frames from flexible laryngoscopy videos, demonstrating high precision. This model has the potential to aid AI researchers with dataset creation for computer-aided diagnosis systems by automatically extracting relevant frames from laryngoscopic videos.

CONFLICT OF INTEREST

Anaïs Rameau is medical advisor for Perceptron Health, Inc. Dan Witte is co-founder of Perceptron Health, Inc.

AUTHOR CONTRIBUTIONS

Peter Yao: conceptualization, methodology, validation, resources, data curation, writing—original draft, writing—review & editing, visualization, project administration, funding acquisition. Dan Witte: conceptualization, methodology, software, validation, resources, writing—review & editing, visualization. Hortense Gimonet: conceptualization, methodology, software, validation, resources, writing—review & editing, visualization. Alexander German: data curation, writing—review & editing. Katerina Andreadis: resources, project administration. Michael Cheng: conceptualization, resources, data curation. Lucian Sulica: conceptualization, resources, writing—review & editing, supervision. Olivier Elemento: conceptualization, methodology, software, writing—review & editing, supervision, funding acquisition. Josue Barnes: conceptualization, methodology, software, supervision, project administration. Anaïs Rameau: conceptualization, methodology, resources, writing—review & editing, supervision, project administration, funding acquisition.

ORCID

Peter Yao https://orcid.org/0000-0002-1127-9030
Katerina Andreadis https://orcid.org/0000-0001-8586-450X

Lucian Sulica https://orcid.org/0000-0001-8989-2073
Anaïs Rameau https://orcid.org/0000-0003-1543-2634

REFERENCES

1. Liu Y, Chen PHC, Krause J, Peng L. How to read articles that use machine learning: users' guides to the medical literature. *JAMA*. 2019; 322(18):1806. doi:10.1001/jama.2019.16489

2. Hinton G. Deep learning: a technology with the potential to transform health care. *Jama*. 2018;320(11):1101. doi:10.1001/jama.2018.11100

3. Esteva A, Robicquet A, Ramsundar B, et al. A guide to deep learning in healthcare. *Nat Med*. 2019;25(1):24-29. doi:10.1038/s41591-018-0316-z

4. Urban G, Tripathi P, Alkayali T, et al. Deep learning localizes and identifies polyps in real time with 96% accuracy in screening colonoscopy. *Gastroenterology*. 2018;155(4):1069-1078.e8. doi:10.1053/j.gastro.2018.06.037

5. Cheng JZ, Ni D, Chou YH, et al. Computer-aided diagnosis with deep learning architecture: applications to breast lesions in US images and pulmonary nodules in CT scans. *Sci Rep*. 2016;6:24454. doi:10.1038/srep24454

6. Esteva A, Kuprel B, Novoa RA, et al. Dermatologist-level classification of skin cancer with deep neural networks. *Nature*. 2017;542(7639):115-118. doi:10.1038/nature21056

7. 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(22):2402. doi:10.1001/jama.2016.17216

8. Cireşan DC, Giusti A, Gambardella LM, Schmidhuber J. Mitosis detection in breast cancer histology images with deep neural networks. In: Mori K, Sakuma I, Sato Y, Barillot C, Navab N, eds. *Medical Image Computing and Computer-Assisted Intervention: MICCAI 2013. Lecture Notes in Computer Science*. Springer; 2013:411-418. doi:10.1007/978-3-642-40763-5_51

9. Hughes CA, McMenamin P, Mehta V, Pillsbury H, Kennedy D. Otolaryngology workforce analysis. *Laryngoscope*. 2016;126(Suppl 9):S5-S11. doi:10.1002/lary.26238

10. Vickery TW, Weterings R, Cabrera-Muffly C. Geographic distribution of otolaryngologists in the United States. *Ear Nose Throat J*. 2016;95(6):218-223. doi:10.1177/014556131609500607

11. Fagan JJ, Jacobs M. Survey of ENT services in Africa: need for a comprehensive intervention. *Glob Health Action*. 2009;2(1):1932. doi:10.3402/gha.v2i0.1932.

12. Ta N. ENT in the context of global health. *Ann R Coll Surg Engl*. 2019;101(2):93-96. doi:10.1308/rcsann.2018.0138

13. Truong BT, Venkatesh S. Video abstraction: a systematic review and classification. *ACM Trans Multimed Comput Commun Appl*. 2007;3(1):3-es. doi:10.1145/1198302.1198305

14. Moccia S, Vanone GO, Momi ED, et al. Learning-based classification of informative laryngoscopic frames. *Comput Methods Programs Biomed*. 2018;158:21-30. doi:10.1016/j.cmpb.2018.01.030

15. Patrini I, Ruperti M, Moccia S, Mattos LS, Frontoni E, de Momi E. Transfer learning for informative-frame selection in laryngoscopic videos through learned features. *Med Biol Eng Comput*. 2020;58(6):1225-1238. doi:10.1007/s11517-020-02127-7

16. Galdran A, Costa P, Campilho A. Real-time informative laryngoscopic frame classification with pre-trained convolutional neural networks. Paper presented at: 2019 IEEE 16th International Symposium on Biomedical Imaging (ISBI 2019). 2019:87-90. doi:10.1109/ISBI.2019.8759511

17. He K, Zhang X, Ren S, Sun J. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2016:770–778. http://arxiv.org/abs/1512.03385

18. Kingma DP, Ba JL. Adam: A method for stochastic optimization. In *ICLR: International Conference on Learning Representations*. 2015:1–15. http://arxiv.org/abs/1412.6980

19. Selvaraju RR, Cogswell M, Das A, Vedantam R, Parikh D, Batra D. Grad-CAM: visual explanations from deep networks via gradient-based localization. Paper presented at: 2017 IEEE International Conference on Computer Vision (ICCV). 2017:618–626. doi:10.1109/ICCV.2017.74

20. Ren J, Jing X, Wang J, et al. Automatic recognition of laryngoscopic images using a deep-learning technique. *Laryngoscope*. 2020;130(11):E686-E693. doi:10.1002/lary.28539

21. Xiong H, Lin P, Yu JG, et al. Computer-aided diagnosis of laryngeal cancer via deep learning based on laryngoscopic images. *EBioMedicine*. 2019;48:92-99. doi:10.1016/j.ebiom.2019.08.075

22. Ballesteros C, Trujillo M, Mazo C, Chaves D, Hoyos J. Automatic classification of non-informative frames in colonoscopy videos using texture analysis. In: Beltrán-Castañón C, Nyström I, Famili F, eds. *Progress in Pattern Recognition, Image Analysis, Computer Vision, and Applications. Lecture Notes in Computer Science*. Springer International Publishing; 2017:401-408. doi:10.1007/978-3-319-52277-7_49

23. Islam ABMR, Alammari A, Oh J, Tavanapong W, Wong J, de Groen PC. Non-informative frame classification in colonoscopy videos using CNNs. Paper presented at: Proceedings of the 2018 3rd International Conference on Biomedical Imaging, Signal Processing. ICBSP 2018. Association for Computing Machinery; 2018:53-60. doi:10.1145/3288200.3288207

24. Yao H, Stidham RW, Soroushmehr R, Gryak J, Najarian K. Automated detection of non-informative frames for colonoscopy through a combination of deep learning and feature extraction. Paper presented at: 2019 41st Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC). 2019:2402–2406. doi:10.1109/EMBC.2019.8856625

25. Kraft M, Fostiropoulos K, Gürtler N, Arnoux A, Davaris N, Arens C. Value of narrow band imaging in the early diagnosis of laryngeal cancer. *Head Neck*. 2016;38(1):15-20. doi:10.1002/hed.23838

SUPPORTING INFORMATION

Additional supporting information may be found in the online version of the article at the publisher's website.