


Article

Gender, Smoking History, and Age Prediction from Laryngeal Images

Tianxiao Zhang ¹, Andrés M. Bur ², Shannon Kraft ², Hannah Kavookjian ², Bryan Renslo ², Xiangyu Chen ¹,
Bo Luo ¹ and Guanghui Wang ^{3,*}

¹ Department of Electrical Engineering and Computer Science, University of Kansas, Lawrence, KS 66045, USA

² Department of Otolaryngology—Head and Neck Surgery, University of Kansas Medical Center, Kansas City, KS 66160, USA

³ Department of Computer Science, Toronto Metropolitan University, Toronto, ON M5B 2K3, Canada

* Correspondence: wangcs@torontomu.ca

Abstract: Flexible laryngoscopy is commonly performed by otolaryngologists to detect laryngeal diseases and to recognize potentially malignant lesions. Recently, researchers have introduced machine learning techniques to facilitate automated diagnosis using laryngeal images and achieved promising results. The diagnostic performance can be improved when patients' demographic information is incorporated into models. However, the manual entry of patient data is time-consuming for clinicians. In this study, we made the first endeavor to employ deep learning models to predict patient demographic information to improve the detector model's performance. The overall accuracy for gender, smoking history, and age was 85.5%, 65.2%, and 75.9%, respectively. We also created a new laryngoscopic image set for the machine learning study and benchmarked the performance of eight classical deep learning models based on CNNs and Transformers. The results can be integrated into current learning models to improve their performance by incorporating the patient's demographic information.

Keywords: laryngeal images; CAM; gender; smoking history; demographic information



Citation: Zhang, T.; Bur, A.M.; Kraft, S.; Kavookjian, H.; Renslo, B.; Chen, X.; Luo, B.; Wang, G. Gender, Smoking History, and Age Prediction from Laryngeal Images. *J. Imaging* **2023**, *9*, 109. <https://doi.org/10.3390/jimaging9060109>

Academic Editor: Carmelo Militello

Received: 5 May 2023

Revised: 22 May 2023

Accepted: 25 May 2023

Published: 29 May 2023



Copyright: © 2023 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (<https://creativecommons.org/licenses/by/4.0/>).

1. Introduction

Flexible laryngoscopy is a commonly used diagnostic tool to visually identify diseases of the larynx [1,2]. While it has advantages over other diagnostic methods given its ease of use and lack of ionizing radiation exposure, discerning between benign and malignant lesions on laryngoscopy requires expert interpretation. Previously, computer vision techniques utilizing deep learning, including Convolutional Neural Networks (CNNs) and Transformers, have been implemented to determine the pathologic diagnosis based on laryngoscopic medical images or videos [3–7]. Such models have been shown to be sufficiently accurate in the diagnosis of laryngeal cancer with only a limited training set [3–8].

The majority of prior studies that utilized machine learning for medical image analysis focused on lesion or polyp detection, segmentation, and classification [9–13]. To date, no studies have attempted to automatically incorporate patient characteristics into lesion detection models by predicting them using laryngeal images. Even for well-trained experts, identifying the age, gender, or smoking status of patients based on laryngoscopy alone is virtually impossible. Fortunately, this is never necessary because this information is readily available to clinicians performing laryngoscopy. However, the incorporation of patient characteristics into deep learning models for medical image analysis typically requires manual entry.

In this study, we have demonstrated the capability of deep learning models, such as CNNs and Transformers, to extract discernible features from laryngeal images, allowing the identification of patients' demographic characteristics. This has the potential to enhance clinical diagnoses by automatically integrating demographic information into intelligent

learning models. For instance, we can automate multi-model learning to improve the detection of laryngeal cancers by considering factors such as the patient's smoking status and age during decision making. Additionally, our research contributes to the field of explainable machine learning (XAI), which emphasizes the provision of clear and interpretable explanations for the decisions and predictions of models. By enhancing transparency and trust, XAI plays a crucial role in medical contexts, where healthcare decisions carry significant importance [14–17]. Analyzing patients' demographic characteristics, especially activation saliency maps, can deepen our understanding of the underlying workings of deep learning models.

This study is the first endeavor to predict the patient's gender, smoking history, and age directly from laryngeal images. We implemented and compared the performance of the following classical CNN-based and Transformer-based deep learning models: ResNet-18 [18], ResNet-50 [18], ResNet-101 [18], DenseNet-121 [19], MobileNetv2 [20], ShuffleNetv2 [21], and ViT [22]. The major contributions of this paper are as follows:

- We performed the first study on predicting the gender, age, and smoking status of the patient purely based on laryngeal images from laryngoscopy.
- We created a dataset of 33,906 laryngeal image frames captured from 398 patients. The dataset is annotated with the clinical diagnosis, the pathologic diagnosis for the lesion, and the patient's demographic information. This is the first large laryngoscopic image set for machine learning studies.
- We implemented and benchmarked the performance of eight classical deep learning models and achieved very promising results.
- We employed the Classification Activation Map (CAM) to visualize and analyze the regions of interest in the image. This approach contributes to the explainability of the learning models by providing insights into which specific areas of the image influenced the decision-making process.

The labeled dataset and developed learning models are available to the research community upon request.

2. Materials and Methods

2.1. Dataset

Data from flexible video stroboscopic exams performed during patient care in the Department of Otolaryngology-Head and Neck Surgery at the University of Kansas Medical Center (KUMC) were collected over a one-year period. Digital videos were collected in MPEG-4 format at 30 frames per second (fps) with a resolution of 720×486 pixels. Each video was labeled with a clinical diagnosis (structurally normal larynx, polyp, papilloma, leukoplakia, or malignant neoplasm) and a pathologic diagnosis for lesions that were biopsied. Additional patient demographic information was captured, including age, sex, and history of tobacco use.

A total of 398 video sequences were included for analysis and randomly separated into training ($n = 319$, 80%) and testing ($n = 79$, 20%) cohorts. Every 10th video frame was extracted from each video sequence, creating a dataset of 33,906 laryngeal images in total, with 26,424 for training and 7,482 for testing. All classification models were pretrained on the ImageNet benchmark [23]. Transfer learning was then used to fine-tune the learning models using the collected training set. Finally, the testing set was employed to evaluate the performance of the final classification models.

2.2. Deep Learning Models

The following classical deep learning models were implemented and compared: ResNet-18 [18], ResNet-50 [18], ResNet-101 [18], DenseNet-121 [19], MobileNetv2 [20], ShuffleNetv2 [21], and Vision Transformer [22]. The general process of deep-learning-based classification is shown in Figure 1. Given an input of a laryngeal image frame, the trained network can predict the gender, smoking history, or age of the patient purely based on the features in the input image. Below is a brief introduction the implemented learning models.

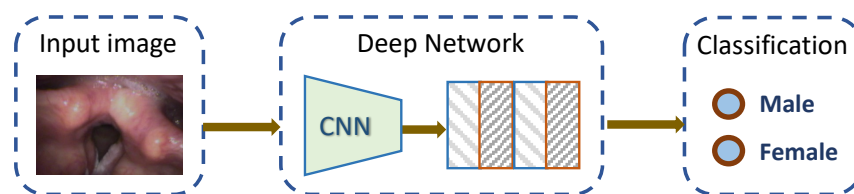


Figure 1. Illustration of using deep learning models for laryngeal image classification. The deep learning models were pretrained on ImageNet and then fine-tuned on the laryngeal dataset using transfer learning. The output prediction could be gender, smoking history, or age.

ResNet: ResNet [18] designs a residual connection to facilitate the training of deep neural networks. The gradients can be easily back-propagated via short connections so that the deep neural networks can be optimized more easily and have better performance than their shallow counterparts. Since its introduction, ResNet has become the benchmark for almost all computer vision tasks and has achieved state-of-the-art performance in almost all tasks. Additionally, shortcut connections can be applied to other classic models, such as Transformers, to achieve state-of-the-art performance in both natural language processing and computer vision applications.

DenseNet: In DenseNet [19], the layers are connected with each other directly so that the gradient can flow smoothly, preventing information flow from vanishing, which is a common difficulty in deep neural network training. The features from different layers are combined by concatenation instead of summation.

MobileNetv2: MobileNetv2 [20] is based on MobileNetv1 [24], which separates the convolutions into depthwise separable convolutions and pointwise convolutions with fewer parameters and computations. MobileNetv2 introduces an inverted residual block that projects the feature maps to a high dimension and then back to a low dimension. The proposed inverted module reduces memory access and accelerates inference speed.

ShuffleNetv2: ShuffleNetv2 [21] was developed from ShuffleNetv1 [25] to empirically design high-efficiency mobile-level networks. Practical guidelines were incorporated for higher efficiency and a more lightweight network, including equal channel width, group convolution cost, less network fragmentation, and fewer element-wise operations.

Vision Transformers: Transformers [26] were initially designed for natural language processing for global connections between long-range tokens. Transformers have since been applied to computer vision tasks and have achieved state-of-the-art performance in classification [27,28] and object detection [29,30]. For image classification, the images are split into patches of the same size, which are embedded into tokens and fed into the Transformer blocks. Usually, there is an extra class token that interacts with all other tokens and produces the ultimate class prediction. Due to the lack of inductive bias, Vision Transformers [22] normally require more data and much longer training epochs to converge.

2.3. Training Settings

Given that the current dataset was relatively small compared to other benchmark datasets in computer vision, transfer learning was employed, and all deep learning models were pretrained on ImageNet [23]. For each learning model, the same structure and hyperparameters as reported in the original paper were utilized. The batch size was set to 16, and the initial learning rate was 0.00005 (reduced by 0.2 each epoch) with a total of 5 epochs. The optimizer utilized was Adam [31], and all code was written with PyTorch [32].

2.4. The Metrics for Evaluation

We evaluated the performance of our deep learning models on the laryngeal dataset using four commonly used metrics: precision, recall, F1 score, and overall accuracy. The definitions of these metrics can be found in [13].

Precision assesses the accuracy of positive predictions by measuring the proportion of correctly classified positive instances out of all instances predicted as positive. It indicates how well the model identifies positive instances and has a low false-positive rate. Recall, also known as sensitivity or the true-positive rate, measures the proportion of correctly classified positive instances out of all actual positive instances. It focuses on the model's ability to detect all positive instances and has a low false-negative rate. The F1 score combines precision and recall into a single value, providing a balanced measure. It is calculated as the harmonic mean of precision and recall, considering both false positives and false negatives. The F1 score serves as an overall performance metric, providing a single evaluation measure. Overall accuracy measures the proportion of correctly classified instances, including both positive and negative, out of all instances.

These metrics offer insights into different aspects of the model's classification abilities. When evaluating a medical image classification model, it is crucial to consider the specific requirements and priorities of the application. The importance of each metric may vary depending on the context. Additionally, it is important to interpret these metrics alongside domain-specific considerations, such as the severity of misclassifications and their potential impact on patient outcomes.

3. Results

A total of 398 video sequences were utilized in our analysis, which were further divided into two cohorts: a training cohort consisting of 319 sequences and a testing cohort comprising 79 sequences. The models were trained using the training cohort, taking into account the ground-truth information regarding the patient's gender, smoking history, and age. Subsequently, the classification performance of the models was assessed using the independent testing set. In this section, we begin by evaluating the model's performance at the image level and subsequently present the results at the patient (sequence) level. This approach allows us to examine both the individual image classification accuracy and the overall performance across the entire video sequence.

3.1. The Performance of Deep Learning Models at Image Level

Figure 2 depicts the loss curves of the different models employed for age, gender, and smoking history predictions. The loss measure utilized in the analysis is the average loss calculated across all previous loss values, resulting in a smoothed loss curve. During the training process, the loss curves converged quickly, and the training was terminated after five epochs. It is observed that MobileNetv2, ShuffleNetv2, and ViT-B exhibit higher loss values compared to the other models at convergence.

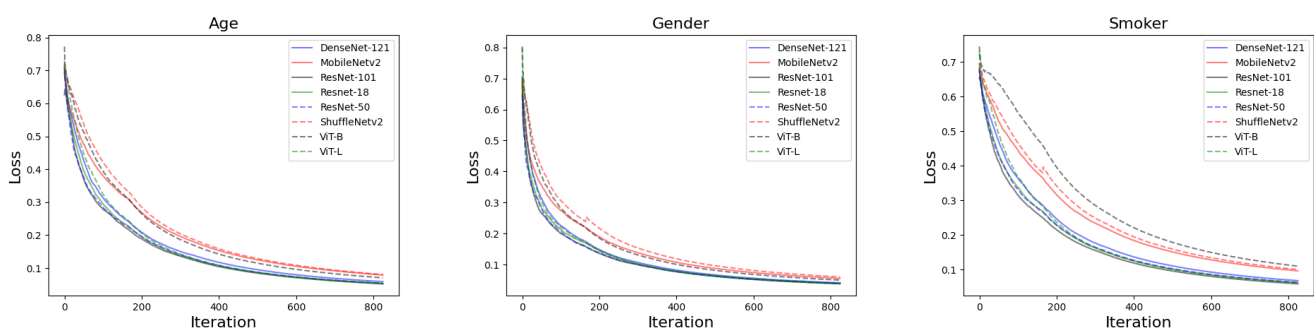


Figure 2. The average loss curve for training the models to predict age, gender, and smoking history. The left, middle, and right graphs are the loss curves for predicting age, gender, and smoking history, respectively.

The evaluation metrics, namely, precision, recall, F1 score, and overall accuracy, are presented in Tables 1–4. These metrics assess the performance of the models in predicting three target categories: gender (male or female), smoking history (smoker or non-smoker), and age (<50 or \geq 50). Each category represents a binary classification problem. In the case

of smoking history, a non-smoker is defined as a patient who has never smoked, while a smoker refers to a patient with any smoking history. For age prediction, patients are divided into two groups: young (<50) and senior (≥ 50), creating a binary classification scenario. The reported precision, recall, F1 score, and overall accuracy provide a comprehensive assessment of the models' performance across these classification tasks.

Table 1 provides the precision values calculated for each class, along with the mean and standard deviation computed across all deep learning models. Overall, the models exhibited consistent performance across the experiments, although the standard deviations for predicting female gender and age < 50 were relatively large. Among all deep learning models, ResNet-50 achieved the highest precision for predicting male gender, with a value of 94.7%. In contrast, the precision for age < 50 was significantly lower at only 44% compared to the other categories. The high precision for predicting male gender can be attributed to the distinguishable features between male and female patients, as well as the clear visual differences between male and female images. In contrast, discerning features related to age becomes more challenging, particularly for patients near the age threshold.

Table 1. The precision for predicting gender, smoking history, and age on the larynx dataset.

DL Models	Gender		Smoking History		Age	
	Male	Female	Smoker	Non-Smoker	<50	≥ 50
ResNet-18	93.9	73.4	65.3	63.3	41.6	88.6
ResNet-50	94.7	70.9	64.2	62.8	44.5	87.8
ResNet-101	92.4	70.9	66.7	64.0	45.7	89.7
DenseNet-121	94.6	72.3	62.0	60.0	44.3	89.5
MobileNetv2	93.6	66.1	64.1	61.5	43.3	89.5
ShuffleNetv2	92.0	66.8	64.6	63.1	38.5	88.0
ViT-L	93.5	66.4	66.6	63.1	47.7	89.1
ViT-B	94.3	69.4	64.7	63.6	46.6	88.4
Mean	93.6	69.5	64.8	62.7	44.0	88.8
Std	0.99	2.81	1.50	1.30	2.93	0.73

The standard deviation of accuracy among the models, as shown in Table 1, indicates that there is relatively low variation in performance across the different models. Notably, the lightweight deep learning models, such as ResNet-18, outperform the more complex models with a larger number of parameters (e.g., ResNet-18 achieves the highest precision for predicting "female"). This observation suggests that the limited size of the dataset may favor simpler models, as they are less prone to overfitting. The dataset's relatively small size may also contribute. Complex models such as Vision Transformers typically require a larger amount of data to achieve optimal performance. While precision provides valuable insights into the models' performance, it is important to consider other metrics as well. The following sections will present the models' performance based on additional evaluation metrics.

Table 2 provides the recall rates for each class. The recall rate measures the proportion of positive samples that are correctly identified among all positive samples in each category. While the recall rates exhibit relatively high variations compared to precision, the results remain consistent across all models. Notably, smoking history exhibits the lowest recall rate among all deep learning models. This can be attributed, in part, to the inherent variability in smoking habits among individuals. Some smokers who have a minimal smoking frequency or have quit smoking for an extended period may display fewer visible changes in their larynx, making it challenging to distinguish them from non-smokers solely based on visual cues. As a result, accurately identifying these individuals as smokers becomes more difficult, leading to a lower recall rate for smoking history prediction.

Table 2. The recall for predicting gender, smoking history, and age on the larynx dataset.

DL Models	Gender		Smoking History		Age	
	Male	Female	Smoker	Non-Smoker	<50	≥50
ResNet-18	83.5	89.3	59.6	68.7	68.7	71.7
ResNet-50	81.1	91.0	59.8	67.1	63.6	76.7
ResNet-101	81.9	86.8	59.8	70.6	70.6	75.4
DenseNet-121	82.3	90.8	54.8	66.8	70.5	74.0
MobileNetv2	76.7	89.6	56.3	68.9	71.1	72.7
ShuffleNetv2	78.1	86.6	59.9	67.6	68.4	67.9
ViT-L	76.9	89.5	57.7	71.4	67.3	78.4
ViT-B	79.8	90.5	61.2	67.0	65.0	78.2
Mean	80.0	89.3	58.6	68.5	68.2	74.4
Std	2.58	1.70	2.17	1.73	2.72	3.57

To comprehensively evaluate the performance of the deep learning models on the larynx dataset, it is important to consider metrics that incorporate both precision and recall. The F1 score, as presented in Table 3, computes the harmonic mean of precision and recall, providing a balanced assessment of the models. The F1 scores among the different deep learning models exhibit consistency, as indicated by the small standard deviations. Notably, the performance for predicting male gender, female gender, and age ≥ 50 surpasses that of other classes in terms of the F1 score. This implies that the models achieve a good balance between precision and recall for these categories, resulting in higher overall performance.

Table 3. The F1 score for predicting gender, smoking history, and age on the larynx dataset.

DL Models	Gender		Smoking History		Age	
	Male	Female	Smoker	Non-Smoker	<50	≥50
ResNet-18	88.4	80.6	62.3	65.9	51.8	79.3
ResNet-50	87.3	79.7	61.9	64.9	52.4	81.9
ResNet-101	86.9	78.1	63.1	67.1	55.5	82.0
DenseNet-121	88.0	80.5	58.2	63.2	54.4	81.0
MobileNetv2	84.3	76.1	60.0	65.0	53.8	80.2
ShuffleNetv2	84.5	75.4	62.2	65.2	49.3	76.7
ViT-L	84.4	76.2	61.8	67.0	55.9	83.4
ViT-B	86.4	78.6	62.9	65.3	54.3	83.0
Mean	86.3	78.2	61.6	65.5	53.4	80.9
Std	1.67	2.06	1.65	1.25	2.17	2.19

The overall accuracy of each learning model was assessed by calculating the number of correctly predicted samples divided by the total number of samples. Table 4 presents the results, showing that gender prediction achieved the highest overall accuracy, followed by age and smoking history predictions. Notably, gender prediction exhibited a particularly high mean accuracy among the three tasks, with an average overall predicted accuracy of 83.2%. The impressive accuracy of gender prediction suggests that deep learning models can effectively capture and analyze specific features present in laryngeal images that are indicative of a patient’s gender. These distinguishing features may not be readily discernible to human experts, underscoring the potential of deep learning models in extracting valuable information from medical images.

Gender prediction presents a straightforward binary classification task, whereas age and smoking history classification pose more significant challenges due to their continuous nature. Dividing age into specific thresholds becomes difficult, as the distinguishing features between different age groups may not be readily apparent. Similarly, predicting smoking history is complex due to the wide range of addiction levels among smokers. For instance, the characteristics of a social smoker or someone with a short smoking history may differ significantly from those of a heavy smoker. Consequently, the boundaries between

smokers and non-smokers are not always clearly discernible, despite the existence of distinct boundaries between heavy smokers and non-smokers. These challenges in establishing clear boundaries likely contribute to the relatively low accuracy observed in age prediction compared to gender prediction.

Table 4. The overall accuracy for predicting gender, smoking history, and age on the larynx dataset.

DL Models	Gender	Smoking History	Age
ResNet-18	85.5	64.2	71.0
ResNet-50	84.4	63.5	73.7
ResNet-101	83.6	65.2	74.3
DenseNet-121	85.2	60.9	73.2
MobileNetv2	81.0	62.6	72.3
ShuffleNetv2	81.0	63.8	68.0
ViT-L	81.2	64.6	75.9
ViT-B	83.4	64.1	75.2
Mean	83.2	63.6	73.0
Std	1.87	1.34	2.53

Despite these inherent difficulties, the developed learning models still achieved notable mean accuracies of 73% for age classification and 63.6% for smoking history prediction. These results demonstrate the models' capability to capture meaningful patterns and extract relevant information from laryngeal images, enabling reasonably accurate predictions. Although classifying age and smoking history entails inherent complexities, the achieved accuracies indicate that the models have successfully learned and utilized discriminative features to make informed predictions in these challenging tasks. These findings highlight the potential of machine learning in extracting valuable information from laryngeal images for age and smoking history classification.

In summary, deep learning models demonstrate strong performance in predicting gender, smoking history, and age, with gender prediction being particularly notable. The models surpass human doctors in extracting this information solely from laryngeal images, showcasing their potential in advancing medical image analysis. This finding underscores the promising role of deep learning models in leveraging visual data to enhance diagnostic capabilities in healthcare. By effectively identifying subtle patterns and characteristics, these models can aid healthcare professionals in providing more accurate assessments based on laryngeal images, ultimately improving patient care and outcomes.

3.2. Overall Performance Based On Patients

The experiments described above were evaluated based on individual image frames, but in clinical settings, all frames in a video sequence belong to the same patient. Therefore, it is more meaningful to evaluate the performance of classification at the sequence level. This section reports the overall accuracy of gender, smoking history, and age predictions at the patient level by combining the results of all frames in the same sequence.

Two methods were used to evaluate sequence-level predictions: majority voting and probability voting. In majority voting, the final prediction is based on the majority of the predicted image labels in the sequence. In probability voting, the predicted probabilities for the correct and wrong labels are separately aggregated, and the final prediction is assigned to the one with higher aggregated probabilities. The comparative results are presented in Table 5. It is evident that when using sequence-based prediction, the overall accuracy for predicting gender, smoking history, and age is much higher than that based on individual frames, as shown in Table 4. We also notice that the overall performance of majority-based voting is slightly better than that of the probability-based approach.

Table 5. The overall accuracy for patients.

DL Models	Gender		Smoking History		Age	
	Majority	Prob	Majority	Prob	Majority	Prob
ResNet-18	90.7	88.9	66.9	62.5	77.5	73.1
ResNet-50	88.9	86.5	62.1	59.1	81.0	78.2
ResNet-101	84.5	84.7	64.8	62.3	77.8	74.4
DenseNet-121	88.0	87.0	61.7	58.3	83.6	79.0
MobileNetv2	84.5	81.2	63.2	59.7	77.5	73.5
ShuffleNetv2	83.7	79.8	67.5	63.0	73.5	70.4
ViT-L	84.3	82.2	69.1	66.6	83.5	80.6
ViT-B	91.1	89.1	68.3	66.8	85.7	82.3
Mean	87.0	84.9	65.4	62.3	80.0	76.5
Std	2.87	3.31	2.71	3.00	3.85	3.89

3.3. Visualization

In order to illustrate the response strength within different areas of images that correspond to the prediction results, a Classification Activation Map (CAM) [33] was extracted. Only results obtained by ResNet-50 were utilized for visualization, as similar results were obtained by other learning models. The CAMs for the prediction of gender, smoking history, and age are illustrated in Figure 3, Figure 4, and Figure 5, respectively.

For visualization, CAM maps were overlaid on top of the original laryngeal images with a ratio of 2:5 so that the high-response areas in the original images could be easily recognized. The red color indicates a high response, and the blue color represents a low response. High-response areas correspond to areas that contribute more to the prediction results.

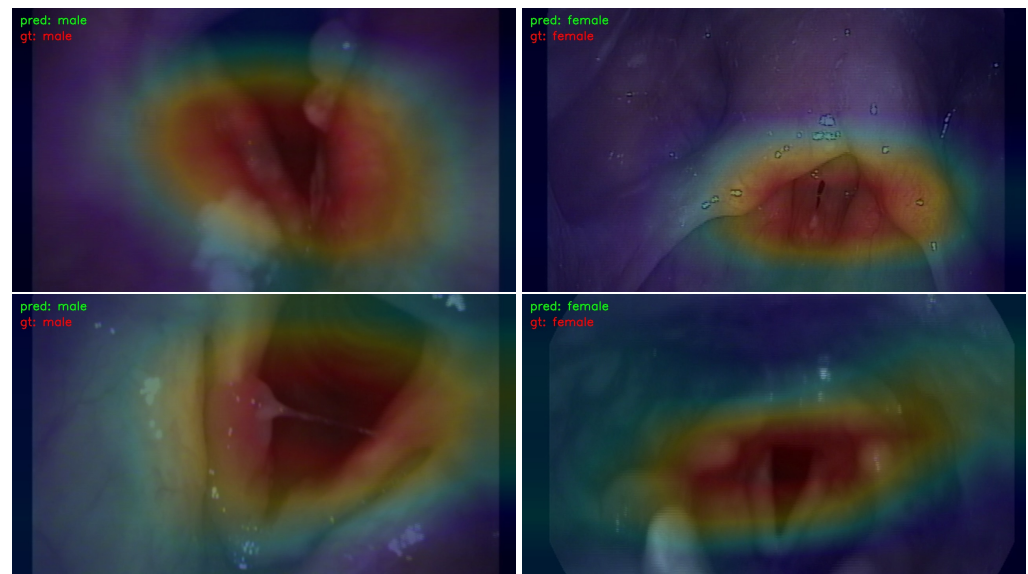


Figure 3. The CAM visualization of gender prediction. “pred” stands for the predicted result, and “gt” represents the ground truth. The left column demonstrates the maps for male patients, and the right column illustrates the maps for female patients. The red color indicates the areas in the image that have a high response for the predicted result, and the blue color means the areas on the image that have a low response for the predicted result.

Figure 3 shows the CAM map for gender prediction. The left images are from male patients, and the right images are from female patients. The high-response areas were similar in both male and female images; they involved the true and false focal folds and, partially, the arytenoids. For smoking history and age prediction, the corresponding CAMs

are illustrated in Figures 4 and 5, respectively. The high-response areas had increased arytenoid involvement and were less focused on the vocal folds.

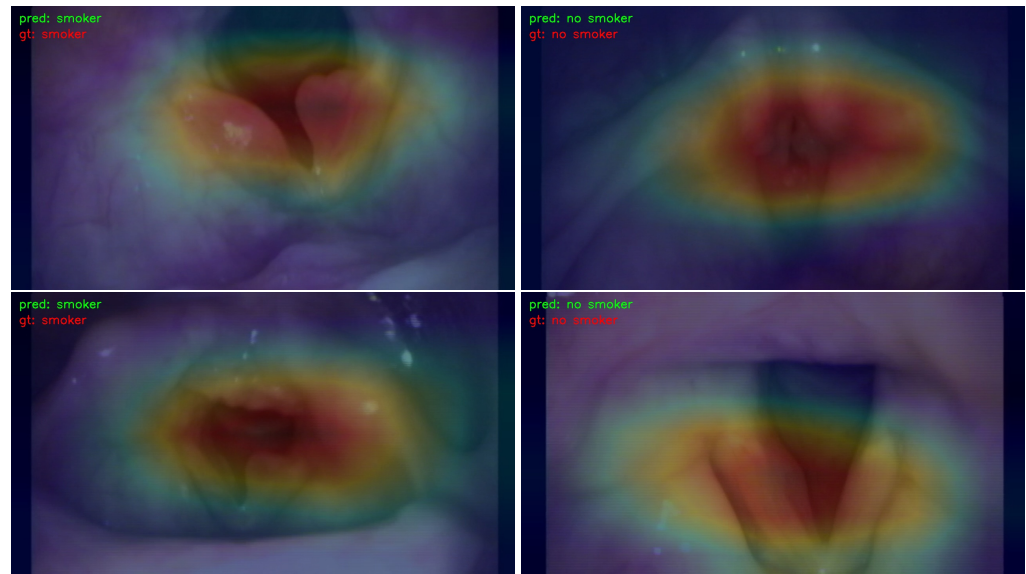


Figure 4. The CAM visualization of smoking history prediction. “pred” stands for the predicted result, and “gt” represents the ground truth. The left column demonstrates the maps for male patients, and the right column illustrates the maps for female patients. The red color indicates the areas in the image have a high response for the predicted result, and the blue color means the areas in the image have a low response for the predicted result.

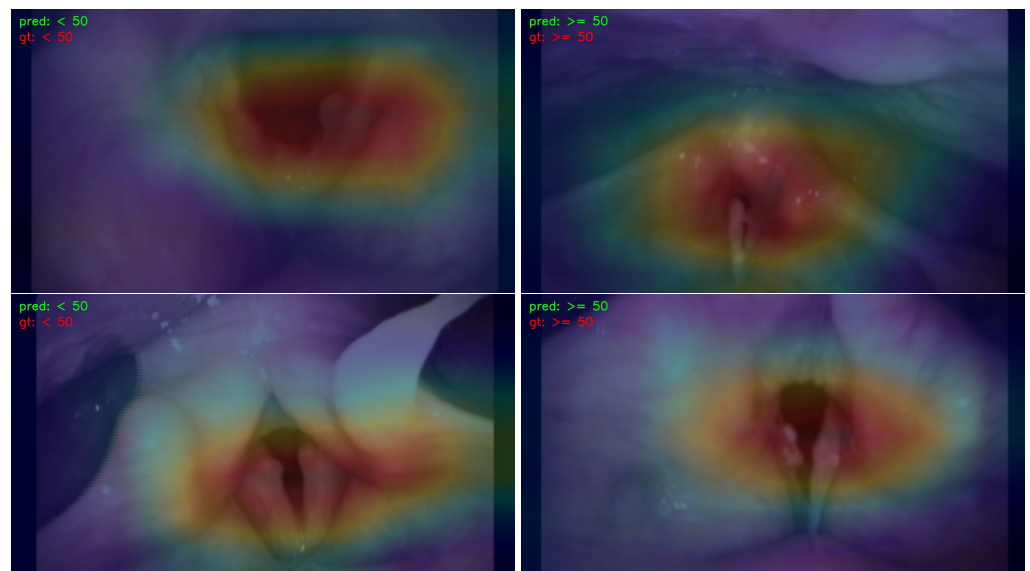


Figure 5. The CAM visualization of age prediction. “pred” stands for the predicted result, and “gt” represents the ground truth. The left column demonstrates the maps for male patients, and the right column illustrates the maps for female patients. The red color indicates the areas in the image that have a high response for the predicted result, and the blue color means the areas in the image that have a low response for the predicted result.

4. Conclusions

This is the first study to employ deep learning models with computer visualization to predict the gender, smoking history, and age of patients from laryngeal images. The deep learning models tested achieved consistent and promising results for these tasks. Visualizing

the CAMs of the laryngeal images revealed that the high-response areas were focused primarily around the true and false vocal folds, which indicates that these areas may exhibit subtle differences among patients of different genders, ages, and smoking statuses.

While we have annotated a laryngoscopic dataset in this study, the trained models may exhibit poor generalizability due to the relatively small scale of the dataset. To mitigate this limitation, it is essential to explore strategies that enhance service continuity in medical image classification. One viable solution is the design of self-organized systems [34] that can dynamically optimize model parameters based on the scalability of the dataset, thereby improving the reliability and adaptability of the system. In our future studies, we will integrate the findings of this study into a comprehensive model for laryngeal disease classification by leveraging multi-modality learning techniques to effectively combine information from various sources, leading to more accurate and reliable diagnostic outcomes. We believe that these advancements will contribute to improved diagnostic capabilities and ultimately benefit patient care.

Author Contributions: Methodology, T.Z.; software, T.Z.; validation, S.K., H.K., B.R. and G.W.; formal analysis, A.M.B.; investigation, B.R.; data curation, X.C.; writing—original draft, T.Z.; supervision, B.L. and G.W.; project administration, A.M.B. and G.W. All authors have read and agreed to the published version of the manuscript.

Funding: This work was supported in part by the Natural Sciences and Engineering Research Council of Canada (NSERC) under grant no. ALLRP 576612-22 and the National Institutes of Health (NIH) under grant no. 1R03CA253212-01.

Institutional Review Board Statement: The study was conducted in accordance with the Declaration of Helsinki, and approved by the KUMC Human Research Protection Program (protocol code STUDY00150076, 10 April 2023).

Informed Consent Statement: It was waived since the study uses anonymized data without personally identifiable information.

Data Availability Statement: Not applicable.

Conflicts of Interest: The authors have declared no conflict of interest.

References

1. Leipzig, B.; Zellmer, J.E.; Klug, D. The Role of Endoscopy in Evaluating Patients With Head and Neck Cancer: A Multi-Institutional Prospective Study. *Arch. Otolaryngol. Neck Surg.* **1985**, *111*, 589–594. [[CrossRef](#)]
2. Ebisumoto, K.; Sakai, A.; Maki, D.; Robinson, K.; Murakami, T.; Iijima, H.; Yamauchi, M.; Saito, K.; Watanabe, T.; Okami, K. Tumor detection with transoral use of flexible endoscopy for unknown primary head and neck cancer. *Laryngoscope Investig. Otolaryngol.* **2021**, *6*, 1037–1043. [[CrossRef](#)] [[PubMed](#)]
3. Xiong, H.; Lin, P.; Yu, J.G.; Ye, J.; Xiao, L.; Tao, Y.; Jiang, Z.; Lin, W.; Liu, M.; Xu, J.; et al. Computer-aided diagnosis of laryngeal cancer via deep learning based on laryngoscopic images. *EBioMedicine* **2019**, *48*, 92–99. [[CrossRef](#)] [[PubMed](#)]
4. Halicek, M.; Lu, G.; Little, J.V.; Wang, X.; Patel, M.; Griffith, C.C.; El-Deiry, M.W.; Chen, A.Y.; Fei, B. Deep convolutional neural networks for classifying head and neck cancer using hyperspectral imaging. *J. Biomed. Opt.* **2017**, *22*, 060503. [[CrossRef](#)] [[PubMed](#)]
5. Azam, M.A.; Sampieri, C.; Ioppi, A.; Africano, S.; Vallin, A.; Mocellin, D.; Fragale, M.; Guastini, L.; Moccia, S.; Piazza, C.; et al. Deep Learning Applied to White Light and Narrow Band Imaging Videolaryngoscopy: Toward Real-Time Laryngeal Cancer Detection. *Laryngoscope* **2022**, *132*, 1798–1806. [[CrossRef](#)] [[PubMed](#)]
6. Takiyama, H.; Ozawa, T.; Ishihara, S.; Fujishiro, M.; Shichijo, S.; Nomura, S.; Miura, M.; Tada, T. Automatic anatomical classification of esophagogastroduodenoscopy images using deep convolutional neural networks. *Sci. Rep.* **2018**, *8*, 7497. [[CrossRef](#)] [[PubMed](#)]
7. Wang, S.; Chen, Y.; Chen, S.; Zhong, Q.; Zhang, K. Hierarchical dynamic convolutional neural network for laryngeal disease classification. *Sci. Rep.* **2022**, *12*, 13914. [[CrossRef](#)]
8. Ren, J.; Jing, X.; Wang, J.; Ren, X.; Xu, Y.; Yang, Q.; Ma, L.; Sun, Y.; Xu, W.; Yang, N.; et al. Automatic recognition of laryngoscopic images using a deep-learning technique. *Laryngoscope* **2020**, *130*, E686–E693. [[CrossRef](#)]
9. Wilson, B.S.; Tucci, D.L.; Moses, D.A.; Chang, E.F.; Young, N.M.; Zeng, F.G.; Lesica, N.A.; Bur, A.M.; Kavookjian, H.; Mussatto, C.; et al. Harnessing the Power of Artificial Intelligence in Otolaryngology and the Communication Sciences. *J. Assoc. Res. Otolaryngol.* **2022**, *23*, 319–349. [[CrossRef](#)]
10. Li, K.; Fathan, M.I.; Patel, K.; Zhang, T.; Zhong, C.; Bansal, A.; Rastogi, A.; Wang, J.S.; Wang, G. Colonoscopy polyp detection and classification: Dataset creation and comparative evaluations. *PLoS ONE* **2021**, *16*, e0255809. [[CrossRef](#)]

11. Patel, K.B.; Li, F.; Wang, G. FuzzyNet: A Fuzzy Attention Module for Polyp Segmentation. In Proceedings of the NeurIPS'22 Workshop on All Things Attention: Bridging Different Perspectives on Attention, New Orleans, LO, USA, 2 December 2022.
12. Patel, K.; Bur, A.M.; Wang, G. Enhanced u-net: A feature enhancement network for polyp segmentation. In Proceedings of the 2021 18th Conference on Robots and Vision (CRV), Burnaby, BC, Canada, 26–28 May 2021, pp. 181–188.
13. Patel, K.; Li, K.; Tao, K.; Wang, Q.; Bansal, A.; Rastogi, A.; Wang, G. A comparative study on polyp classification using convolutional neural networks. *PLoS ONE* **2020**, *15*, e0236452. [[CrossRef](#)]
14. Militello, C.; Prinzi, F.; Sollami, G.; Rundo, L.; La Grutta, L.; Vitabile, S. CT radiomic features and clinical biomarkers for predicting coronary artery disease. *Cogn. Comput.* **2023**, *15*, 238–253. [[CrossRef](#)]
15. Gu, R.; Wang, G.; Song, T.; Huang, R.; Aertsen, M.; Deprest, J.; Ourselin, S.; Vercauteren, T.; Zhang, S. CA-Net: Comprehensive attention convolutional neural networks for explainable medical image segmentation. *IEEE Trans. Med. Imaging* **2020**, *40*, 699–711. [[CrossRef](#)]
16. Van der Velden, B.H.; Kuijff, H.J.; Gilhuijs, K.G.; Viergever, M.A. Explainable artificial intelligence (XAI) in deep learning-based medical image analysis. *Med. Image Anal.* **2022**, *79*, 102470. [[CrossRef](#)] [[PubMed](#)]
17. Prinzi, F.; Orlando, A.; Gaglio, S.; Midiri, M.; Vitabile, S. ML-Based Radiomics Analysis for Breast Cancer Classification in DCE-MRI. In Proceedings of the Applied Intelligence and Informatics: Second International Conference, AII 2022, Reggio Calabria, Italy, 1–3 September 2022; Springer: Berlin/Heidelberg, Germany, 2023, pp. 144–158.
18. 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, Las Vegas, NV, USA, 27–30 June 2016; pp. 770–778.
19. Huang, G.; Liu, Z.; Van Der Maaten, L.; Weinberger, K.Q. Densely connected convolutional networks. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Honolulu, HI, USA, 21–26 July 2017; pp. 4700–4708.
20. Sandler, M.; Howard, A.; Zhu, M.; Zhmoginov, A.; Chen, L.C. Mobilenetv2: Inverted residuals and linear bottlenecks. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, 18–22 June 2018; pp. 4510–4520.
21. Ma, N.; Zhang, X.; Zheng, H.T.; Sun, J. Shufflenet v2: Practical guidelines for efficient cnn architecture design. In Proceedings of the European Conference on Computer Vision (ECCV), Munich, Germany, 8–14 September 2018, pp. 116–131.
22. Dosovitskiy, A.; Beyer, L.; Kolesnikov, A.; Weissenborn, D.; Zhai, X.; Unterthiner, T.; Dehghani, M.; Minderer, M.; Heigold, G.; Gelly, S.; et al. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv* **2020**, arXiv:2010.11929.
23. Russakovsky, O.; Deng, J.; Su, H.; Krause, J.; Satheesh, S.; Ma, S.; Huang, Z.; Karpathy, A.; Khosla, A.; Bernstein, M.; et al. Imagenet large scale visual recognition challenge. *Int. J. Comput. Vis.* **2015**, *115*, 211–252. [[CrossRef](#)]
24. Howard, A.G.; Zhu, M.; Chen, B.; Kalenichenko, D.; Wang, W.; Weyand, T.; Andreetto, M.; Adam, H. Mobilenets: Efficient convolutional neural networks for mobile vision applications. *arXiv* **2017**, arXiv:1704.04861.
25. Zhang, X.; Zhou, X.; Lin, M.; Sun, J. Shufflenet: An extremely efficient convolutional neural network for mobile devices. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, 18–22 June 2018; pp. 6848–6856.
26. Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A.N.; Kaiser, L.; Polosukhin, I. Attention is all you need. *arXiv* **2017**, arXiv:1706.03762.
27. Chen, X.; Hu, Q.; Li, K.; Zhong, C.; Wang, G. Accumulated Trivial Attention Matters in Vision Transformers on Small Datasets. In Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision, Waikoloa, HI, USA, 2–7 January 2023; pp. 3984–3992.
28. Liu, Z.; Lin, Y.; Cao, Y.; Hu, H.; Wei, Y.; Zhang, Z.; Lin, S.; Guo, B. Swin transformer: Hierarchical vision transformer using shifted windows. In Proceedings of the IEEE/CVF International Conference on Computer Vision, Montreal, BC, Canada, 11–17 October 2021; pp. 10012–10022.
29. Carion, N.; Massa, F.; Synnaeve, G.; Usunier, N.; Kirillov, A.; Zagoruyko, S. End-to-end object detection with transformers. In Proceedings of the European Conference on Computer Vision, Glasgow, UK, 23–28 August 2020; pp. 213–229.
30. Ma, W.; Zhang, T.; Wang, G. Miti-detr: Object detection based on transformers with mitigatory self-attention convergence. *arXiv* **2022**, arXiv:2112.13310.
31. Kingma, D.P.; Ba, J. Adam: A method for stochastic optimization. *arXiv* **2014**, arXiv:1412.6980.
32. Paszke, A.; Gross, S.; Massa, F.; Lerer, A.; Bradbury, J.; Chanan, G.; Killeen, T.; Lin, Z.; Gimelshein, N.; Antiga, L.; et al. Pytorch: An imperative style, high-performance deep learning library. *arXiv* **2019**, arXiv:1912.01703
33. Zhou, B.; Khosla, A.; Lapedriza, A.; Oliva, A.; Torralba, A. Learning deep features for discriminative localization. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Las Vegas, NV, USA, 27–30 June 2016; pp. 2921–2929.
34. Conti, V.; Militello, C.; Rundo, L.; Vitabile, S. A novel bio-inspired approach for high-performance management in service-oriented networks. *IEEE Trans. Emerg. Top. Comput.* **2020**, *9*, 1709–1722. [[CrossRef](#)]

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.