MEDINFO 2023 — The Future Is Accessible
J. Bichel-Findlay et al. (Eds.)

doi:10.3233/SHTI231104

Vision-Based Assistance for Vocal Fold Identification in Laryngoscopy with Knowledge Distillation

Thao Thi Phuong DAO[a,b,c,d], Minh-Khoi PHAM[e], Mai-Khiem TRAN[a,b,c], Chanh Cong Ha[f], Boi Ngoc VAN[g], Bich Anh TRAN[h], Minh-Triet TRAN[a,b,c,1]

[a] *University of Science, VNU-HCMC, Ho Chi Minh City, Vietnam*
[b] *John von Neumann Institute, VNU-HCMC, Ho Chi Minh City, Vietnam*
[c] *Vietnam National University, Ho Chi Minh City, Vietnam*
[d] *Otorhinolaryngology Department, Thong Nhat Hospital, Ho Chi Minh City, Vietnam*
[e] *Dublin City University, Dublin, Ireland*
[f] *Otorhinolaryngology Department, 7A Military Hospital, Ho Chi Minh City, Vietnam*
[g] *Otorhinolaryngology Department, Vinmec Central Park International Hospital, Ho Chi Minh City, Vietnam*
[h] *Otorhinolaryngology Department, Cho Ray Hospital, Ho Chi Minh City, Vietnam*

Abstract: Laryngoscopy images play a vital role in merging computer vision and otorhinolaryngology research. However, limited studies offer laryngeal datasets for comparative evaluation. Hence, this study introduces a novel dataset focusing on vocal fold images. Additionally, we propose a lightweight network utilizing knowledge distillation, with our student model achieving around 98.4% accuracy-comparable to the original EfficientNetB1 while reducing model weights by up to 88%. We also present an AI-assisted smartphone solution, enabling a portable and intelligent laryngoscopy system that aids laryngoscopists in efficiently targeting vocal fold areas for observation and diagnosis. To sum up, our contribution includes a laryngeal image dataset and a compressed version of the efficient model, suitable for handheld laryngoscopy devices.

Keywords: Laryngoscopy, vocal folds, knowledge distillation, vision-based assistance

1. Introduction

Deep learning is increasingly pivotal in medical applications, predicting outcomes or detecting anomalies [1]. This is particularly relevant for mobile healthcare devices, which are compact, portable for bedside use, and budget-friendly [2]. Yet, the scalability of these tools faces challenges due to computing resource constraints, especially with neural networks requiring specialized processors for acceleration [3]. Handheld devices like mobile phones further amplify these limitations, constrained by computational power, storage, and battery life [4].

[1] Corresponding author: Minh-Triet Tran, email: tmtriet@fit.hcmus.edu.vn.

In laryngology, smartphone-linked scopes offer portability for diagnoses and treatment planning based on larynx images [5]. Yet, limited research exists on deep learning models for these devices to classify vocal fold images. Thus, a vital need exists for a lightweight deep learning network tailored to handheld laryngoscopy.

Knowledge distillation (KD) transfers wisdom from a potent teacher model to a nimble student model, heightening performance without compromising efficiency [6]. This imparts better performance to a student model than another non-KD student. Resultantly, we form a notably compact, specialized model via KD, derived from a broader, standard model, ideal for portable laryngoscopy devices. In this paper, our major contributions are as follows:

- We present a novel dataset for classifying vocal fold images. After that, we experiment and evaluate state-of-the-art backbones on our vocal fold image dataset.
- We propose a newly simple yet efficient architecture using KD to compress model to match the performance of baseline and use minimal computing resource.
- We develop a smartphone application with our AI model to assist laryngoscopists locate vocal fold area quickly.

2. Methods

2.1. *Dataset collection*

Retrospectively collected from 4,624 images of 876 patients at Cho Ray Hospital, Vietnam (Jan 2020 - Nov 2021), the dataset underwent manual classification by two doctors with cross-validation. Another experienced doctor further refined the ground truth. The classification created two classes: visible and non-visible vocal folds. Approved by Cho Ray Hospital's ethics committee (No.1280/GCN-HDDD) and aligned with Declaration of Helsinki, the study waived patient consent due to its retrospective nature not impacting clinical care or workflow.

2.2. *Model evaluation*

During the procedure, the model was trained with 80% of the dataset and tested with the remaining 20%. We used various pretrained backbones – VGG19, ResNet50V2, MobileNetV2, DenseNet201, InceptionV3, Xception, and EfficientNetB1 – fine-tuned on Cho Ray Hospital's laryngoscopy dataset. Models underwent 20 epochs, batch size 32, using Adam optimizer. A learning rate policy was applied, decreasing from initial 0.0001 by 0.7 after 5 static epochs. Evaluation used accuracy, recall, and precision (Eqs. 1, 2, 3) to compare backbones' impact, described as follows:

$$\text{Accuracy} = \frac{\text{TP}+\text{TN}}{\text{FP}+\text{TP}+\text{FN}+\text{TN}}, (1)$$

$$\text{Recall} = \frac{\text{TP}}{\text{TP}+\text{FN}}, (2)$$

$$\text{Precision} = \frac{\text{TP}}{\text{TP}+\text{FP}}, (3)$$

where TP: true positive, TN: true negative, FP: false positive, and FN: false negative.

2.3. *Proposed solution based on KD architecture*

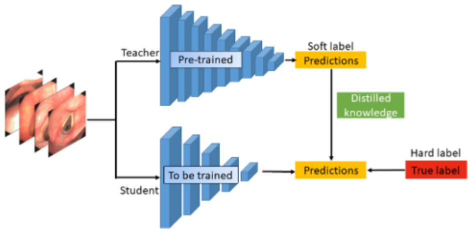


Figure 1. Overview of our KD architecture

Our KD architecture (Figure 1) includes a ResNet-based student network (Figure 2) and a teacher network (EfficientNetB1). The objective function can be described:

$$\alpha * \mathrm{CE}(y_{gt}, y_s) + (1 - \alpha) * \mathrm{KL}(y_t, y_s), \ (4)$$

where α denotes the balance coefficient between supervision and distillation loss. y_{gt} indicates the ground truth. y_s and y_t represent the outputs of the student and the teacher. CE and KL represent standard Cross-Entropy and customized Kullback-Leibler Divergence losses. We make student mimic teacher's predictions (Eq. 5):

$$\mathrm{KL}(y_t, y_s) = \mathrm{F}(y_t) \times \log\left(\frac{F(y_t)}{F(y_s)}\right), \ (5)$$

where F is the softmax normalization with temperature λ. Mathematically, F can be described as Eq. (6):

$$\mathrm{F}(x_i) = \frac{\exp(x_i \mid \lambda)}{\Sigma_j \exp(x_j \mid \lambda)}, \ (6)$$

Additionally, we train a standalone model (α=1, 'scratch'), while student network begins with He-Normalization. Experiments adopt λ=10 and α=0.2-lower α highlights teacher's influence.

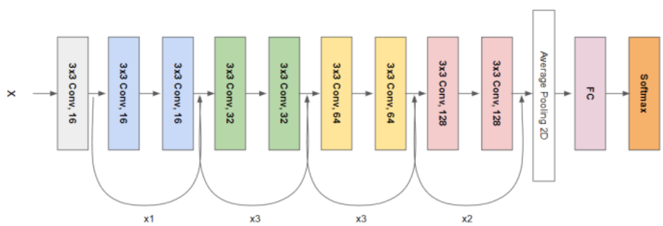


Figure 2. Architecture of our simplified ResNet with shallow residual blocks used as the student network.

3. Results

Among 4,624 samples, 2,147 (46.43%) show existing vocal folds, and 2,480 depict invisible ones. EfficientNetB1 performs best, as in Table 1.

Table 1. Results of state-of-the-art backbones. **Red** values correspond to the best performance.

	VGG 19	ResNet50 V2	MobileNet V2	Inception V3	DenseNet 201	Xception	EfficientNetB1
Accuracy (%)	91.8	98.5	96.1	98.3	98.2	98.2	**98.7**
Recall (%)							
Non vocal fold	94.2	98.8	96.7	98.3	98.3	98.5	**99.2**

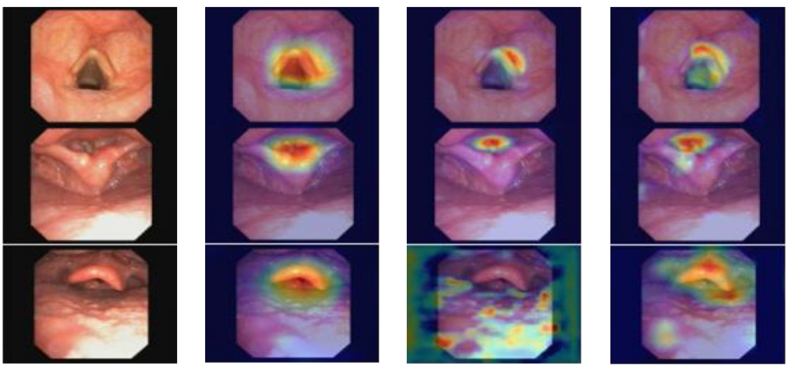
Vocal folds	88.7	98.0	95.3	**98.3**	98.0	97.8	98.0
Precision (%)							
Non vocal fold	91.4	98.5	96.4	**98.6**	98.5	98.3	98.5
Vocal folds	92.3	98.5	95.8	97.8	97.8	98.0	**99.0**

Table 2 shows that distillation aids student in matching teacher's performance, enhancing convergence with minimal resources. Our new network has 0.8M parameters, 88% smaller than EfficientNetB1 and all networks in Table 1, while achieving high accuracy.

Table 2. Comparison between student and teacher performance on our validation set. Our designed network in which KD is applied outperforms MobileNetV2 in both accuracy and light-weight term.

Methods	Accuracy	No. parameters
EfficientNet-B1	98.7%	6.7M
MobileNetV2	96.1%	2.4M
Simple-ResNet-Scratch	96.7%	0.8M
Simple-ResNet-Distilled	98.4%	0.8M

We use Grad-CAM for model transparency (Figure 3), emphasizing informative image regions influencing classifier choices.



(a) Input image (b) EficientNetB1 (c) Simple-ResNet-Scratch (d) Simple-ResNet-Distilled

Figure 3. Grad-CAM visualization on our laryngoscopy image dataset. Noticeably, the student model (d), guided by the knowledge from the teacher (b), can capture the salient features more precisely, while trained-from-scratch model (c) fails to do so in the third case.

4. Discussion

EfficientNetB1 enhances diagnostic speed in flexible laryngoscopy, overcoming reflex and secretion hindrances. It offers high accuracy, aiding vocal fold assessment for swift categorization, especially by inexperienced doctors. In medical contexts, precise diagnosis is vital, and evaluating a converted model's accuracy is key. Our student model maintains 98.4% accuracy like EfficientNetB1, reducing weight by 88%. It's 5% faster than EfficientNetB1 and 2% faster than MobileNetV2 on smartphones, maintaining high accuracy. Moreover, our distilled model excels in feature extraction compared to the scratch model.

As technology advances and outpatient ENT endoscopy demand grows, we developed a smartphone laryngoscopy system. The real-time vocal fold detection system captures and displays endoscopic images on the phone screen, aiding doctors with rapid and precise clinical decisions. This vision-based assistance boasts high accuracy, quick convergence, and minimal memory usage (Figure 4).

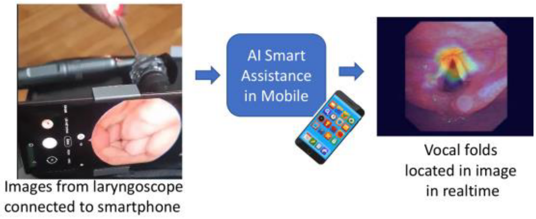


Figure 4. AI smart assistance on smartphone for real time vocal fold detection and localization.

5. Conclusions

Our paper introduces a laryngeal image dataset for vocal fold detection. We created a compact student model using KD, matching EfficientNetB1's accuracy with fewer parameters. We suggested AI-assisted smartphone laryngoscopy for faster, targeted diagnosis, reducing patient examination time. Future plans include expanding the dataset to cover specific vocal fold diseases.

Acknowledgements

This research is funded by Vietnam National University Ho Chi Minh City (VNU-HCM) under grant number DS2020-42-01.

References

[1] Piccialli F, Di Somma V, Giampaolo F, Cuomo S, Fortino G. A survey on deep learning in medicine: Why, how and when?. Inf Fusion. 2021 Feb 1;66:111-37, doi: 10.1016/j.inffus.2020.09.006.

[2] Kim Y, Oh J, Choi SH, Jung A, Lee JG, Lee YS, Kim JK. A portable smartphone-based laryngoscope system for high-speed vocal cord imaging of patients with throat disorders: instrument validation study. JMIR mHealth and uHealth. 2021 Jun 18;9(6):e25816, doi: 10.2196/25816.

[3] Chen C, Zhang P, Zhang H, Dai J, Yi Y, Zhang H, Zhang Y. Deep learning on computational-resource-limited platforms: a survey. Mob Inf Syst. 2020 Mar;2020:1-9, doi: 10.1155/2020/8454327.

[4] Tawalbeh M, Eardley A. Studying the energy consumption in mobile devices. Procedia Comput Sci. 2016 Jan;94:183-9, 10.1016/j.procs.2016.08.028.

[5] Rosen CA, Murry T. Diagnostic laryngeal endoscopy. Otolaryngol Clin North Am. 2000 Aug;33(4):751-7, doi: 10.1016/S0030-6665(05)70241-3.

[6] Zhang L, Song J, Gao A, Chen J, Bao C, Ma K. Be your own teacher: Improve the performance of convolutional neural networks via self distillation. InProceedings of the IEEE/CVF International Conference on Computer Vision; 2019 (pp. 3713-3722).