의학석사 학위논문

내시경 영상을 활용한 중이 질환 진단에 있어서 딥러닝 모델과 이비인후과 의사의 비교 분석

Analysis of Deep Learning Model and Otolaryngologists in Diagnosing Middle Ear Diseases Using Endoscopic Images

울산대학교 대학원

의 학 과

이 세 은

내시경 영상을 활용한 중이 질환 진단에 있어서 딥러닝 모델과 이비인후과 의사의 비교 분석

지도교수 안 중 호

이 논문을 의학석사 학위 논문으로 제출함.

2024 년 2 월

울산대학교 대학원

의 학 과

이 세 은

이세은의 의학석사 학위 논문을 인준함.

심사위원 안 중 호 (인)

심사위원 박 홍 주 (인)

심사위원 권 지 훈 (인)

울 산 대 학 교 대 학 원

2024 년 2 월

국문 요약

머신 러닝을 기반으로 한 딥 러닝은 수십 년 동안 전문가의 이미지 해석과 자동 분석 사이의 간극을 좁히는 데 급속히 발전해 왔습니다. 우리는 EfficientNet-B4 딥 러닝 모델의 진단 성능을 이비인후과 전문의들과 100 개의 내시경 이미지에서 비교하고, 또한 딥 러닝 결과를 알고 나서 이비인후과 전문가들의 진단 결론 변화를 비교했습니다. 이 모델은 주요 클래스(중이염, 만성 중이염, 없음)와 보조 클래스(고막 윗부에 있는 진주종, 고막염, 귀 진균증, 통기관)를 예측했습니다. 세 명의 이비인후과 교수, 다섯 명의 고년차 레지던트, 그리고 다섯 명의 저년차 레지던트가 동일한 내시경 이미지에서 주요 및 보조 클래스의 선택을 수행했습니다. 딥 러닝 결과를 알고 난 후 다시 선택을 수행했습니다. 주요 클래스의 예측에서 딥 러닝 모델의 정확도는 95.0%였습니다. 딥 러닝 결과를 알기 전에 교수, 고년차 레지던트 및 저년차 레지던트의 정확도는 각각 78.7%, 65.0%, 54.4%였습니다. 딥 러닝 모델과 세 그룹 간에는 유의한 차이가 있었습니다($p<0.001$, 각각). 딥 러닝 결과를 알고 나서 교수, 고년차 레지던트 및 저년차 레지던트의 정확도는 각각 89.7%, 93.8%, 86.6%였습니다. 딥 러닝 결과를 알기 전보다 모든 그룹에서 정확도가 통계적으로 증가했습니다($p<0.001$, 각각). 모든 그룹, 특히 레지던트 그룹의 진단 성능이 딥 러닝 결과를 알고 나서 향상된 것으로 보입니다. 이는 딥 러닝이 중이 질환을 진단하는 데 경험이 적은 의사들뿐만 아니라 교육적 측면에서 레지던트에게도 도움이 될 수 있음을 시사합니다.

차 례

Table Contents

Figure Contents

INTRODUCTION

Middle ear diseases, encompassing conditions such as Acute Otitis Media (AOM), Otitis Media with Effusion (OME), and Chronic Otitis Media (COM), are prevalent not only in developing countries like Southeast Asia, the Western Pacific region, and Africa but also in developed countries. [1] OME is a very common disease with a prevalence of 80-90% in children, and early diagnosis is important due to its potential impact on speech and language development when associated with hearing loss. [2, 3] Otoscopy is essential for diagnosing middle ear diseases including AOM, OME, and COM. [4] However, 64% of African countries have less than one otolaryngologist per million people. [5]. Therefore, non-otolaryngologists have to evaluate and diagnose middle ear diseases when otolaryngologists are in short supply. [5-7] Even skilled otolaryngologists encounter challenges in conducting precise otoscopy in moving children. These cases sometimes lead to misdiagnosis even when otoscopy is performed accurately.

In the medical field, deep learning (DL) based on machine learning (ML) has rapidly advanced over decades to bridge the gap between expert image interpretation and automated analysis. [8-11] One study trained a deep learning model with dermoscopic images of skin lesions, and the overall classification accuracy of the trained deep learning model was 76.5%. [9] In some studies, deep learning was conducted through images such as computed tomography (CT), ultrasound (US), and simple radiography.[12-14] One study evaluated the accuracy of two deep learning models for detecting femoral neck fractures on radiographs which was 88.1% and 94.4%, respectively. [15]

In otolaryngology field, one study divided otoscope images into balanced and imbalanced sets based on prevalence and compared the diagnostic performance of deep learning and specialist. As a result, the deep learning model showed a bias for prevalence compared to specialists. [8]

In our previous study, we developed a deep learning model capable of simultaneously diagnosing two or more diseases using an endoscopy image database that included various diseases. We customized the EfficientNet-B4 architecture to include shared and task-specific layers for multi-task learning. A total of 6630 RGB images were reformatted into 256×256×3 using circular cropping and utilized for deep learning contruction. This deep learning model can predict two non-coexisting diseases (OME and COM) along with four concurrently detectable diseases (attic cholesteatoma, myringitis, otomycosis, and ventilating tube). The accuracy of this deep learning model ranged from 72% to 98%. [16]

Many studies show that deep learning's image classification performance equals or exceeds that of medical specialists.[9-13, 15] However, in actual practice, there are not many studies on whether deep learning is helpful to doctors who make diagnoses by referring to deep learning.

In this study, we compared the diagnostic performance of the deep learning model and the otolaryngologist using a dataset of 100 endoscopic images. Additionally, we analyzed how referring to deep learning results influenced the diagnostic conclusions of otolaryngologists.

MATERIALS AND METHODS

1. Diagnosis model construction by deep learning network

The deep learning network architecture proposed by [16], which is based on Efficientnet-B4 [17], was utilized. It has shared layers and task-specific layers that enable more accurate diagnosis of not only the primary classes of None, OME, and COM, but also other diseases that can be detected such as Attic cholesteatoma, Myringitis, Otomycosis, and Ventilation tube. (Figure 1)

RGB images reformatted into 256×256×3 with circular cropping were fed to the deep networks as inputs. The pre-trained weights from ImageNet were leveraged for transfer learning. Categorical cross-entropy loss was employed to train the models for multi-class classification. To prevent overfitting, data augmentation methods that randomly altered (−90° to 90°), shifted (0–20% of image size in horizontal and vertical axes), magnified (0–20%), mirrored horizontally, modified brightness (0–20%) and reduced resolution (0–50%) of the images were used. The models were trained for 200 epochs with a differential learning rate, which was initially set as 10-3 and decreased by half when the validation loss stagnated for 50 epochs. To assess the model performance more objectively and robustly, 5-fold cross-validation was performed.

2. Data description

Endoscopic images of the tympanic membrane (TM) were collected in Asan Medical Center from Jan 2018 to Dec 2020. Endoscopic examination was mainly performed for diagnosis, and captured

images were stored in the hospital server without patient identification information. The collected images were classified into one primary class with only one selection and four secondary classes with multiple selections possible. Two experienced otologists blindly annotated each collected image. In the primary class with only one selection, the collected images were annotated as either 'Otitis media with effusion' (OME), 'Chronic otitis media' (COM), or 'None' which meant the absence of OME and COM. OME was defined as the presence of effusion in the middle ear cavity, which was presented by the color change of the TM, such as amber, or the air-fluid level in the middle ear cavity. COM was defined as visible perforation of TM. In the second class with multiple selections possible, the collected images were annotated as ‘Attic cholesteatoma’, ‘Myringitis’, ‘Otomycosis’, and ‘Ventilation tube’. Attic cholesteatoma was defined as the retracted TM or bony destruction in the attic. Myringitis was defined as any inflammation of the TM, including acute otitis media. Otomycosis was defined as the whitish debris or visible pore of fungus in the TM or external auditory canal. Ventilating tube was defined as the tube inserted through the TM. For example, when COM with otomycosis was identified in the image, the primary class was ’COM’ and the secondary classes were ’True’ for otomycosis, and ‘False’ for attic cholesteatoma, myringitis, and ventilating tube. The present study is in compliance with the Declaration of Helsinki and research approval was granted from the Institutional Review Board of the Asan Medical Center with a waiver of research consent (IRB no. 2021–0837).

3. Comparison of diagnostic performance between deep learning network and otolaryngologists

Diagnostic performance was evaluated using 100 randomly selected images that were not used in the deep learning construction. First, to compare diagnostic performance, three otology professors, five senior residents, and five junior residents performed selections of the primary class and the secondary class in the endoscopic images (Version 1.0) for 1 minute each. In the same method, deep learning models performed selections of the primary class and the secondary class in Version 1.0. The accuracy and F1-score of each class in Version 1.0 were calculated and compared between deep learning network and otolaryngologists. Second, to evaluate whether deep learning model was helpful for diagnosis, same otolaryngologists performed selections of the primary class and the second class in the same endoscopic images (Version 2.0) by referring to deep learning selection. The accuracy and F1-score of Version 2.0 were calculated to compare the deep learning network and otolaryngology diagnosis performance. Also, the accuracy and F1 score of version 1.0 and version 2.0 were compared.

4. Comparison of diagnostic performance between otolaryngologists referring to deep learning results

The accuracy, and F1-score of three otology professors (PF), five senior (SR), and five junior residents (JR) were compared in Version 1.0 and Version 2.0, respectively. To compare the changes in the diagnostic conclusions of otolaryngologists by referring to the deep learning results, 1) Cases with inconsistency between the first selection (version 1.0) and the second selection (version 2.0) in

the total data were selected. 2) In the case of inconsistency, Version 2.0 was divided into a group that selected an answer same with the deep learning model and a group that selected an answer different from the deep learning model. 3) Compared to the gold standard, the accuracy was calculated for each primary and secondary class of the two groups, and the difference in accuracy between the three otolaryngologist groups was compared.

5. Statistical analysis

All statistical analyses were performed by using SAS software (version 9.4; SAS Institute, Cary, NC). P values of accuracy were calculated by use of generalized estimating equations (GEE) method to account for patient clustering effect. P values of F1-score were calculated by use of bootstrap method. A p-value less than 0.05 was considered statistically significant.

RESULTS

1. Demographic characteristics

A total of 11,970 endoscopic images were enrolled in this study. In primary class, 1,657 images, 1,707 images, and 3,971 images were annotated as 'OME', 'COM', and 'None', respectively. In the second classes, 1,333 images and 1,152 images were annotated as ‘Attic cholesteatoma’ and ‘Myringitis’, respectively. And, 452 images and 1,698 images were annotated as ‘Otomycosis’ and ‘Ventilation tube’, respectively.

2. Diagnostic performance of deep learning network

Table 1 shows the diagnostic performance of deep learning network. In the prediction of the primary class, the accuracy and the F1 score were 95.0% and 94.9%, respectively. Both sensitivity and specificity were 95.2%. The accuracy and F1 score of attic cholesteatoma class were 98.0% and 94.1%, respectively. In myringitis class, the accuracy was 94.0% and the F1-score was 88.0%. The accuracy and F1 score of otomycosis class were 97.0% and 84.2%, respectively. The accuracy of the ventilation tube class was 99.0%, and the F1 score was 97.9%, showing the highest accuracy and F1 score among the four second classes. In the second classes, the myringitis class showed the lowest accuracy at 94.0%, and the otomycosis class showed the lowest F1 score at 84.2%.

3. Comparison of diagnostic performance between deep learning network and otolaryngologists

Table 2 shows the otolaryngologist's diagnosis performance of version 1.0 before knowing the deep learning results. In the professor group, the accuracy and F1 score of the primary class were 78.7% and 79.2%, respectively. Furthermore, the average accuracy and F1 score of four second classes were 94.8% and 83.1%, respectively. In senior resident group, the accuracy and F1 score of the primary class were 65.0% and 65.5%, respectively. And, the average accuracy and F1 score of four second classes were 95.1% and 84.6%, respectively. In junior resident group, the accuracy and F1 score of the primary class were 54.4% and 53.6%, respectively. And, the average accuracy and F1 score of four second classes were 91.2% and 72.1%, respectively. In both primary class and second classes, the accuracy and F1 scores increased in the order of senior residents, senior residents, and professors. All three groups had statistically significantly lower accuracy in the primary class than deep learning model ($p=0.001$, $p<0.001$ and $p<0.001$, respectively) (Table 3)

Table 2 shows the otolaryngologist's diagnosis performance of version 2.0 after knowing the deep learning results. In the professor group, the accuracy and F1 score of the primary class were 89.7% and 89.9%, respectively. And, the average accuracy and F1 score of four second classes were 96.3% and 88.5%, respectively. In senior resident group, the accuracy and F1 score of the primary class were 93.8% and 93.8%, respectively. And, the average accuracy and F1 score of four second classes were 97.1% and 92.0%, respectively. In junior resident group, the accuracy and F1 score of the primary class were 86.6% and 86.6%, respectively. And, the average accuracy and F1 score of four second classes were 95.3% and 85.5%, respectively.

4. Comparison of diagnostic performance between otolaryngologists referring to deep learning results

When the accuracy of version 1 and version 2 were compared, the accuracy of version 2 increased statistically significantly in all three groups in the primary class. ($p=001$, $p<0.001$, and $p<0.001$, respectively) (Table 4) As a result of referring to the deep learning results, 22.7% (n=68) of the professor group (n=300), 33.6% (n=168) of the senior resident group (n=500), and 42.40% (n=212) of the junior resident group (n=500) changed their selection in primary class. In the professor group, 73.5% (n=157) of them changed their selection with the same result as deep learning. In the resident groups, 93.5% (n = 157) of senior resident and 88.7% (n = 188) of junior resident changed their selection with the same results as deep learning. In this case, the accuracy of the primary class was 96.0%, 96.82%, and 95.21%, respectively, in the order of professor, senior resident, and junior resident. However, in the case of changing the selection to a different result from deep learning, the accuracy of the primary class was 0%, 27.27%, and 8.33%, respectively, in the order of professor, senior resident, and junior resident. Accuracy was significantly higher in all three groups when the selection was changed with the same result compared to when the selection was changed with a different result from deep learning. ($p<0.001$, $p<0.001$, and $p<0.001$ respectively) (Figure 2)

DISCUSSION

In this study, a deep learning model with high accuracy and F1 score of 94% to 98% was constructed using 11,970 endoscopic images. Compared with deep learning, three groups of otolaryngologists showed lower accuracy and F1 scores. Also, in the first selection performed before knowing deep learning results, more experienced otolaryngologists had higher accuracy and F1 scores in both the primary and second classes. However, in the second selection performed after knowing deep learning results, three groups of otolaryngologists showed higher accuracy and F1 scores than the first selection. Impressively, in the second selection, both the accuracy and F1 scores of the senior and junior groups increased to the same level as the professor group. As a result of referring to the deep learning results, the resident groups changed their choice more than the professor group, and the accuracy was higher when changing to the same selection as the deep learning result than when changing to a different selection from deep learning.

In one particular study, a neural network was developed using 267 intraoperative endoscopic images, successfully predicting the presence of middle ear effusion with an accuracy of 83.8% in pediatric cases. [18] Khan et al. classified a total of 2,484 endoscopic images into three categories: normal, perforation, and middle ear effusion, and constructed a deep learning model with an accuracy of 95%. [19] Similarly, Zeng et al. classified a total of 20,542 endoscopic images into eight categories: normal, middle ear cholesteatoma, chronic suppurative otitis media, external auditory canal bleeding, impacted cerumen, external otomycosis, secretory otitis media, and tympanic membrane calcification.

They trained their model using these images and achieved an overall accuracy of 95.59%. [20] In previous studies, deep learning models were constructed to diagnose a single disease. However, in actual clinical practice, many cases involve the coexistence of multiple diseases. Therefore, we developed a deep learning model in our previous study that can simultaneously predict two non-coexisting diseases (OME and COM) along with four concurrently detectable diseases (attic cholesteatoma, myringitis, otomycosis, and ventilating tube). This model performed predictions within an accuracy range of 72% to 98%. [16] By augmenting the number of images used in this deep learning process and performing additional training, the performance improved to 94-98%. This performance level is comparable to or even surpasses that of other studies. The reason for the improved performance is likely the result of training with a sufficient amount of high-quality databases.

Many studies showed that deep learning's image classification performance equals or exceeds that of medical specialists [9, 10, 12, 13, 15], one study divided 7500 otoscope images into balanced and imbalanced sets based on prevalence and compared. The machine learning model showed an accuracy of 77% on the balanced test set and 82% on the imbalanced test set, which was similar to the performance levels of 71% and 72% achieved by the otolaryngologists, respectively. [8] In this study, deep learning achieved an accuracy of 94%, better than the performance of the previous study. On the other hand, the performance of otolaryngologists was 54.7% to 87.8%, similar to previous studies. This suggested that deep learning showed higher accuracy compared to otolaryngologists when

trained with a sufficient amount of data. Therefore, for uncommon diseases, enhancing accuracy through data augmentation could enable the deep learning model to classify various middle ear diseases in the future. Deep learning could be helpful in the diagnostic process when endoscopic findings are nonspecific for diagnosing uncommon diseases. Therefore, we suggested that deep learning can be used for diagnosing not only common diseases but also uncommon diseases in the future.

As a result of referring to the deep learning results, the resident groups changed their choice more than the professor group, which showed a higher reliance on deep learning due to lack of experience. Surprisingly, in the second selection, the accuracy of the resident group increased to a level similar to that of the professor group, suggesting that deep learning can be helpful for diagnosis in residents. Additionally, one study reported that otoscope simulation improved confidence in diagnosing otological disease, with 71% of learners agreeing or strongly agreeing, compared to conventional training. They suggested that this simulation in a safe environment could expose learners to a variety of common and rare diseases in a short amount of time, providing a new addition to existing methods. [4] In the same vein, deep learning can also be used as an aid to educating residents in a safe environment.

There were some limitations on our study. First, despite the large enough database of over 10,000 images, the deep learning data set was collected from a single center. Second limitation of our study is the absence of physicians who were non-otolaryngologists. When comparing reliance on deep

learning, it may be helpful to compare accuracy between non-otolaryngologists and otolaryngologists. However, we addressed the potential variance in dependence on deep learning based on experience by categorizing the resident group into two subgroups according to their level of expertise.

CONCLUSION

We developed a deep learning model that accurately diagnosed various diseases from a single endoscopic picture. Accuracy increased when training with high quality and sufficient quantity of images. Therefore, in the future, we proposed to expand the use of deep learning by enhancing accuracy through data augmentation for the diagnosis of not only common diseases but also rare diseases. When there were discrepancies between the diagnostic results of deep learning and doctors, less experienced doctors tended to adjust the diagnosis more based on the results of deep learning. As a result, diagnostic performance was improved, reaching the level of expertise comparable to that of professors. Our research findings provide insights into the potential integration of deep learning into clinical practice for diagnosing middle ear diseases, as well as the possibility of education for residents through deep learning.

Table 1. Diagnostic performance of deep learning model

Variables	Accuracy	F1-score	Sensetivity	Specificity
Primary class (P)	0.950	0.949	0.952	0.952
None	0.950	0.952	0.980	0.918
OME	0.980	0.952	0.909	1.000
COM	0.970	0.943	0.926	0.986
Attic cholesteatoma (S1)	0.980	0.941	0.941	0.988
Myringitis (S2)	0.940	0.880	0.786	1.000
Otomycosis (S3)	0.970	0.842	0.889	0.978
Ventilatation tube (S4)	0.990	0.979	1.000	0.987

OME, Otits media with effusion; COM, chronic otitis media.

Table 2. Diagnostic performance of otolaryngologists in Version 1.0 (before knowing the results of deep learning) and Version 2.0 (after knowing the results of deep learning)

	Professor group (n=3)		Senior resident group (n=5)		Junior resident group (n=5)	
	Accuracy	**F1-score**	**Accuracy**	**F1-score**	**Accuracy**	**F1-score**
Version 1.0						
Primary class	0.897	0.899	0.938	0.938	0.866	0.866
None	0.897	0.889	0.938	0.939	0.872	0.870
OME	0.957	0.910	0.974	0.940	0.948	0.875
COM	0.940	0.898	0.964	0.936	0.912	0.853
Attic cholesteatoma	0.973	0.917	0.990	0.970	0.984	0.952
Myringitis	0.917	0.832	0.914	0.826	0.884	0.743
Otomycosis	0.963	0.800	0.982	0.892	0.958	0.759
Ventilation tube	0.997	0.993	0.996	0.991	0.984	0.965
Version 2.0						
Primary class	0.897	0.899	0.938	0.938	0.866	0.866
None	0.897	0.889	0.938	0.939	0.872	0.870
OME	0.957	0.910	0.974	0.940	0.948	0.875
COM	0.940	0.898	0.964	0.936	0.912	0.853
Attic cholesteatoma	0.973	0.917	0.990	0.970	0.984	0.952
Myringitis	0.917	0.832	0.914	0.826	0.884	0.743
Otomycosis	0.963	0.800	0.982	0.892	0.958	0.759
Ventilation tube	0.997	0.993	0.996	0.991	0.984	0.965

OME, Otits media with effusion; COM, chronic otitis media.

Table 3. Comparison of accuracy between deep learning and otolaryngologists

	Primary class		**Attic cholesteatoma**		**Myringitis**		**Otomycosis**		**Ventilation tube**	
	Accuracy	P-value	Accuracy	P-value	Accuracy	P-value	Accuracy	P-value	Accuracy	P-value
Deep learning	0.950	ref	0.980	ref	0.940	ref	0.970	ref	0.990	ref
Version 1.0										
Pf group	0.787	0.001	0.963	0.346	0.870	0.013	0.963	0.751	0.993	0.741
SR group	0.650	<.0001	0.966	0.449	0.876	0.026	0.962	0.713	0.998	0.256
JR group	0.544	<.0001	0.942	0.135	0.754	<.0001	0.960	0.613	0.992	0.842

Pf group, Professor group; SR group, Senior resident group; JR group, Junior resident group.

Table 4. Comparison of accuracy of otolaryngologists before and after referencing deep learning results

	Primary class			**Attic cholesteatoma**			**Myringitis**			**Otomycosis**			**Ventilation tube**		
	Ver. 1	Ver. 2	P value	Ver. 1	Ver. 2	P value	Ver. 1	Ver. 2	P value	Ver. 1	Ver. 2	P value	Ver. 1	Ver. 2	P value
Pf	0.787	0.897	<.0001	0.963	0.973	0.248	0.870	0.917	0.010	0.963	0.963	>0.999	0.993	0.997	0.571
SR	0.650	0.938	<.0001	0.966	0.990	0.000	0.876	0.914	0.032	0.962	0.982	0.077	0.998	0.996	0.571
JR	0.544	0.866	<.0001	0.942	0.984	<.0001	0.754	0.884	<.0001	0.960	0.958	0.847	0.992	0.984	0.191

Ver. 1, Version 1.0; Ver. 2, Version 2.0; Pf, Professor group; SR, Senior resident group; JR, Junior resident group.

Figure 1. Schematic diagram of a deep learning network for multi-class classification.

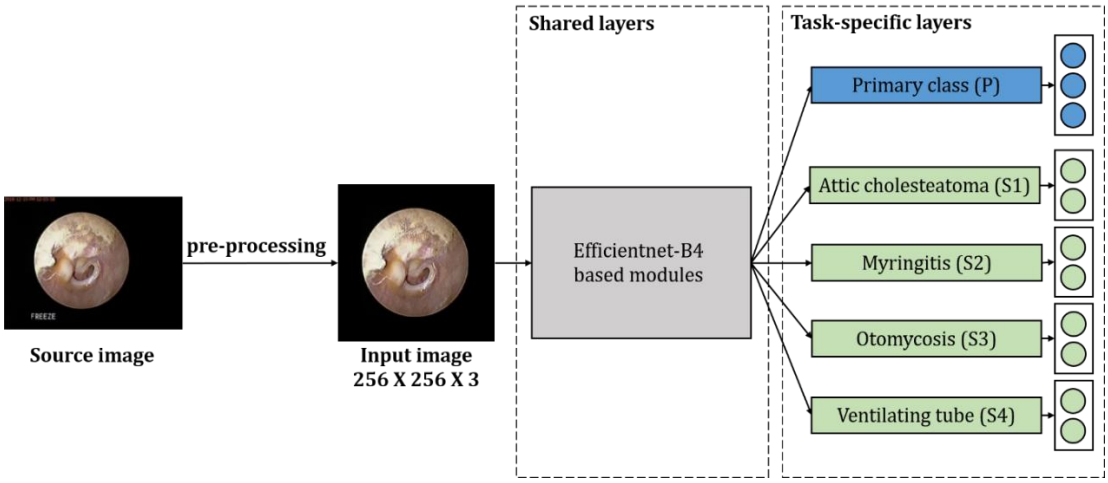
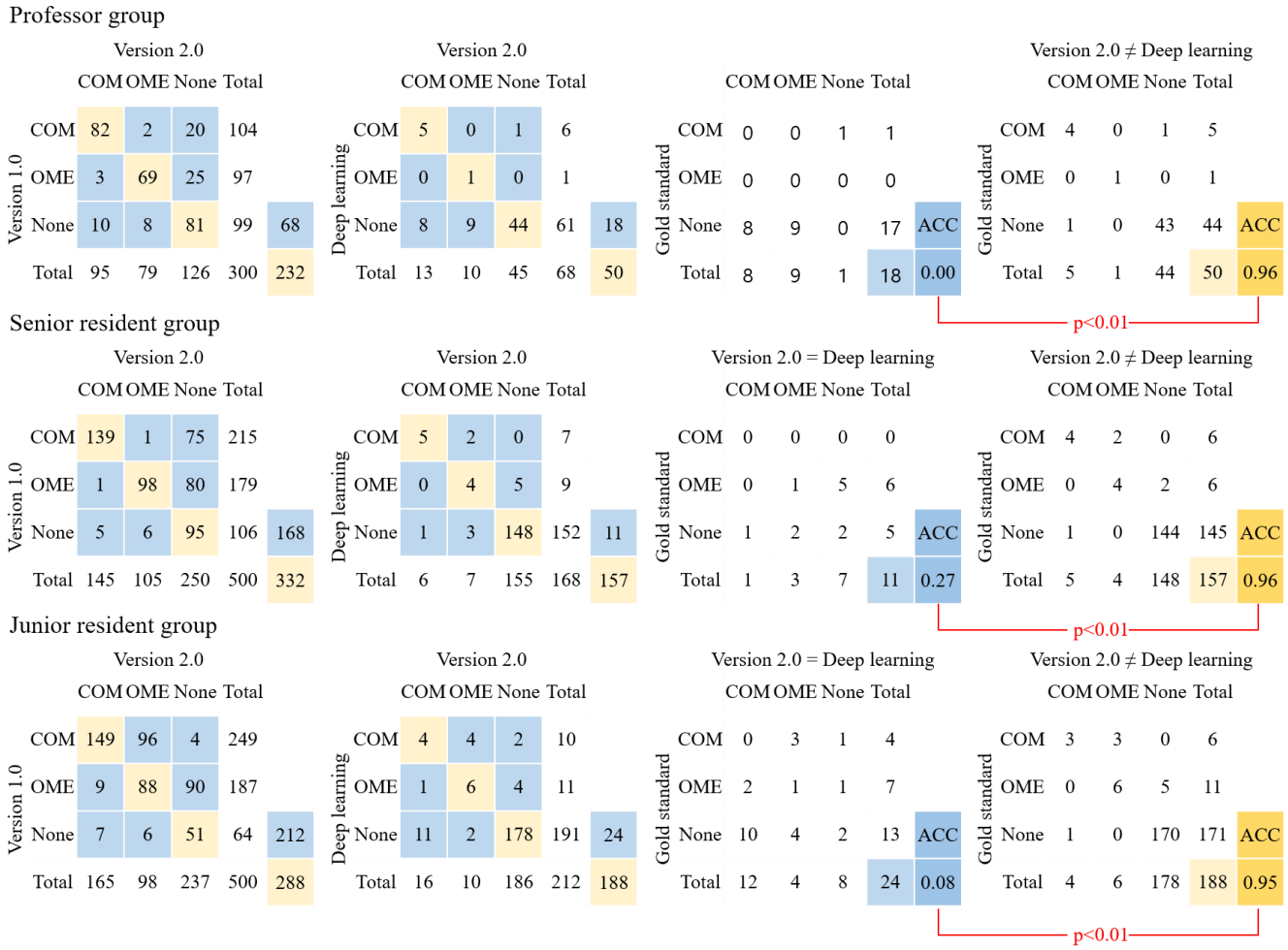


Figure 2. Cross tables comparing the changes in the diagnosis conclusions of the three groups by referring to the deep learning results. ACC, accuracy.



REFERENCES

1. World Health, O., *Chronic suppurative otitis media : burden of illness and manage ment options*. 2004, Geneva: World Health Organization.
2. Tos, M., *Epidemiology and natural history of secretory otitis.* Otology & Neurotol ogy, 1984. **5**(6): p. 459-462.
3. Shekelle, P., et al., *Diagnosis, natural history, and late effects of otitis media with effusion.* Evid Rep Technol Assess (Summ), 2002(55): p. 1-5.
4. Davies, J., et al., *Otoscopy simulation training in a classroom setting: a novel appr oach to teaching otoscopy to medical students.* Laryngoscope, 2014. **124**(11): p. 25 94-7.
5. World Health, O., *Multi-country assessment of national capacity to provide hearin g care*. 2013, Geneva: World Health Organization.
6. Fagan, J.J. and M. Jacobs, *Survey of ENT services in Africa: need for a comprehen sive intervention.* Glob Health Action, 2009. **2**.
7. Myburgh, H.C., et al., *Otitis Media Diagnosis for Developing Countries Using Ty mpanic Membrane Image-Analysis.* EBioMedicine, 2016. **5**: p. 156-60.
8. Cha, D., et al., *Differential Biases and Variabilities of Deep Learning-Based Artifi cial Intelligence and Human Experts in Clinical Diagnosis: Retrospective Cohort and Survey Study.* JMIR Med Inform, 2021. **9**(12): p. e33049.
9. Fujisawa, Y., et al., *Deep-learning-based, computer-aided classifier developed wit h a small dataset of clinical images surpasses board-certified dermatologists in ski n tumour diagnosis.* Br J Dermatol, 2019. **180**(2): p. 373-381.
10. Haenssle, H.A., et al., *Man against machine: diagnostic performance of a deep lea rning convolutional neural network for dermoscopic melanoma recognition in com parison to 58 dermatologists.* Ann Oncol, 2018. **29**(8): p. 1836-1842.
11. Topol, E.J., *High-performance medicine: the convergence of human and artificial i ntelligence.* Nat Med, 2019. **25**(1): p. 44-56.
12. Yang, Q., et al., *Improving B-mode ultrasound diagnostic performance for focal liv er lesions using deep learning: A multicentre study.* EBioMedicine, 2020. **56**: p. 10

2777.

13. Lee, J.H., et al., *Deep learning-based automated detection algorithm for active pul monary tuberculosis on chest radiographs: diagnostic performance in systematic s creening of asymptomatic individuals.* Eur Radiol, 2021. **31**(2): p. 1069-1080.
14. Ariji, Y., et al., *Contrast-enhanced computed tomography image assessment of cer vical lymph node metastasis in patients with oral cancer by using a deep learning system of artificial intelligence.* Oral Surg Oral Med Oral Pathol Oral Radiol, 201 9. **127**(5): p. 458-463.
15. Adams, M., et al., *Computer vs human: Deep learning versus perceptual training f or the detection of neck of femur fractures.* J Med Imaging Radiat Oncol, 2019. **63** (1): p. 27-32.
16. Choi, Y., et al., *Automated multi-class classification for prediction of tympanic me mbrane changes with deep learning models.* PLoS One, 2022. **17**(10): p. e027584 6.
17. Tan, M. and Q. Le. *Efficientnet: Rethinking model scaling for convolutional neural networks*. in *International conference on machine learning*. 2019. PMLR.
18. Crowson, M.G., et al., *Machine Learning for Accurate Intraoperative Pediatric Mi ddle Ear Effusion Diagnosis.* Pediatrics, 2021. **147**(4).
19. Khan, M.A., et al., *Automatic detection of tympanic membrane and middle ear infe ction from oto-endoscopic images via convolutional neural networks.* Neural Net w, 2020. **126**: p. 384-394.
20. Zeng, X., et al., *Efficient and accurate identification of ear diseases using an ense mble deep learning model.* Sci Rep, 2021. **11**(1): p. 10839.

영문 요약

Deep learning based on machine learning has rapidly advanced over decades to bridge the gap between expert image interpretation and automated analysis. We compared a custom EfficientNet-B4 deep learning model's diagnostic performance with otolaryngologists in 100 endoscopic images, and compared the change in the diagnosis conclusion of the otolaryngologist after knowing the deep learning results. The model predicted primary (otitis media with effusion, chronic otitis media, 'None') and secondary classes (attic cholesteatoma, myringitis, otomycosis, ventilating tube). Three otology professors, five senior residents, and five junior residents performed selection of the primary class and the secondary class from the same endoscopic images. After knowing deep learning results, they performed selection again. In the prediction of the primary class, the accuracy of deep learning model were 95.0%. Before knowing the deep learning result, the accuracy of professors, senior residents, and junior residents was 78.7%, 65.0%, and 54.4%. There was a significant difference between deep learning models and three groups. ($p<0.001$, respectively). After knowing the deep learning results, the accuracy of professors, senior residents, and junior residents were 89.7%, 93.8%, and 86.6%, respectively. The accuracy was statistically increased in all group compared to before knowing the deep learning results. ($p<0.001$, respectively). The diagnostic performance of all groups, especially residents groups,

improved after knowing the deep learning result. This suggests that deep learning can be helpful not only for doctors with little experience in diagnosing middle ear diseases, but also for resident in terms of education.