

딥러닝 기반 상부위장관 내시경 이미지 자동 분류의 데이터 구성별 성능 분석 연구

서정민[†], 임상현^{††}, 김영재^{***}, 정준원^{****}, 김광기^{*****}

Performance analysis of deep learning-based automatic classification of upper endoscopic images according to data construction

Jeong Min Seo[†], Sang Heon Lim^{††}, Yung Jae Kim^{***},
Jun Won Chung^{****}, Kwang Gi Kim^{*****}

ABSTRACT

Recently, several deep learning studies have been reported to automatically identify the location of diagnostic devices using endoscopic data. In previous studies, there was no design to determine whether the configuration of the dataset resulted in differences in the accuracy in which artificial intelligence models perform image classification. Studies that are based on large amounts of data are likely to have different results depending on the composition of the dataset or its proportion. In this study, we intended to determine the existence and extent of accuracy according to the composition of the dataset by compiling it into three main types using larynx, esophagus, gastroscopy, and laryngeal endoscopy images.

Key words: Image Classification Model, Artificial Intelligence, Deep Learning, Dataset, Accuracy, Upper Endoscopy

1. 서 론

위암은 2018년 통계 자료의 주요 암 유병률에서 2위를 기록할 만큼 빈번히 발생하는 암으로, 조기

발견하여 치료를 받으면 생존율이 높다[1]. 하지만, 조기 위암은 뚜렷한 증상이 없어 육안을 이용한 진단에는 한계가 있다. 위궤양 또한 임상에서 자주 보이는 질병이며, 적절한 때에 치료받지 않으면 재발 또

※ Corresponding Author: Kwang Gi Kim, Address: Department of Biomedical Engineering Medical Center, Gachon University, College of Medicine, 21, Namdong-daero 774 beon-gil, Namdong-gu, Incheon, 21565, Republic of Korea, TEL : +82-32-820-4036, FAX : +82-32-460-2361, E-mail : kimkg@gachon.ac.kr

※ Corresponding Author: Jun Won Chung, Address: Department of Internal Medicine, Gachon University Gil Medical Center, 21 Namdong-daero 774 beon-gil, Namdong-gu, Incheon 21565, Korea, TEL : +82-32-460-3778, FAX : +82-32-460-3408, E-mail : junwonchung@daum.net
 Receipt date : Nov. 3, 2021, Revision date : Feb. 7, 2022
 Approval date : Feb. 10, 2022

[†] Dept. of Medicine, Gachon University College of Medicine
 (E-mail : jungmin00010@naver.com)

^{**} Dept. of Health Sciences and Technology, Gachon Advanced Institute for Health Sciences and Technology (GAIHST), Gachon University
 (E-mail : smion123@naver.com)

^{***} Dept. of Biomedical Engineering,, College of IT Convergence, Gachon University
 (E-mail : youngjae@gachon.ac.kr)

※ This work was supported by Institute of Information & communications Technology Planning & Evaluation (IITP) grant funded by the Korea government(MSIT) (No. 2020-0-00161-001, Active Machine Learning based on Open-set training for Surgical Video), and by National IT Industry Promotion Agency(NIPA) grant funded by the Korea government(MSIT) (No.S0252-21-1001, Development of AI Precision Medical Solution(Doctor Answer 2.0)).

는 합병증 발생으로 이어질 수 있다는 점에서 조기발견이 중요하다[2]. 한편, 식도에는 장막이 없고 림프관과 혈관이 많이 분포하여 식도암 발생 시에 위, 후두 등 주변 다른 기관으로 침범하거나 전이되기 쉽다[3,4]. 질병이 발견되었다면 주변 기관도 함께 검진받아야 질병의 확산을 막을 수 있다.

상부위장관 질병들을 초기에 치료하기 위해서는 후두, 식도, 위에서 발생하는 질병을 일찍 발견할 수 있도록 도와주는 검진의 역할이 중요하다. 상부위장관 검사 시에 사용하기 가장 좋은 방법 중 하나는 내시경이다. 하지만, 최근 임상에서 위내시경 검사 시 위암이나 식도암 등에 대한 간과율이 약 5~10% 비율로 보고됨으로써, 의사가 수행하는 내시경 검사 또한 한계가 존재한다는 것이 드러났다[5, 6]. 또한, 이는 검진의 정확도와 효율성을 향상시키기 위해 진단 기기 자동화가 필요하다는 것을 시사한다.

최근 내시경 데이터를 활용하여 질병을 발견할 수 있는 인공지능 프로그램이 활발하게 개발되고 있다[7,8,9]. 인공지능이 주어진 데이터에서 스스로 병변을 찾아내는 기술은 진단 기기의 자동화를 이루는 데에 중요한 기술이다. 하지만, 프로그램의 정확도를 높이고 진단 기기의 완전한 자동화를 이루기 위해서는 병변을 찾아내는 기술보다 현재 기기가 인체의 어느 위치에 있는지를 스스로 인지하는 기술이 선행되어야 한다. 진단 과정에서 병변 발견 이전에 이루어져야 하는 것이 위치 파악이며, 정확한 위치 파악이 되어야 위치에 맞는 최적의 알고리즘을 구현할 수 있기 때문이다.

최근 들어 기기의 위치 파악을 자동화하기 위해 내시경 영상 또는 이미지로 딥러닝을 진행하는 연구가 몇몇 진행되었다. 2018년 Hirotsoshi Takiyama 등은 GoogLeNet을 사용하여 위내시경 이미지를 기관별로 분류하고, 위의 경우는 일본 위암취급규약의 기준에 따라 세 부분으로 나누어 각각을 식별할 수 있도록 하는 딥러닝을 연구하였으며, 연구 결과 분류 및 식별을 함에 있어 0.99 이상의 정확도(accuracy)를 보여주었다[10]. 2020년 Qi He 등은 딥러닝 기반 분류 방법에 대한 효율성을 평가하고자 위내시경 이미지를 약 14가지 그룹으로 나누어 Alexnet을 이용해 학습시켰고, 학습에 대한 효과와 타당성을 입증하였다[11]. 이후 2021년에는 박중완 등이 이전 연구에서 사용된 딥러닝 모델과 최신 CNN 모델을 활용하

여 위내시경 시에 맹점 없이 관찰할 수 있도록 돕는 해부학적 분류모델을 제안하였고, 4% 미만의 에러율을 기록했다[12].

많은 양의 데이터를 기반으로 진행되는 딥러닝 연구는 편향된 데이터를 사용할 경우 잘못된 결과를 도출할 수 있기 때문에, 활용하는 데이터에 대해 고려하는 것이 중요하다[13]. 내시경 데이터를 학습시킬 때, 정상이라 판단되는 데이터만 사용한다면 질병이 관찰되는 데이터들을 구별하는 데 어려움이 있을 수 있다. 실제 임상에서는 건강한 환자뿐만 아니라 질병을 지닌 환자 또한 많이 접하기 때문에, 이러한 문제가 발생할 경우 딥러닝 모델의 보편성이 떨어진다. 한편, 학습 및 검증 데이터에 정상과 질병 상태의 데이터를 섞는 비율에 따라 인공지능의 정확도가 바뀔 가능성이 존재하기에 정확도가 높은 모델을 만들기 위해서는 이와 같은 데이터에 대한 구성을 고려해 보아야 한다[14]. 앞서 진행된 연구들은 데이터 세트의 구성에 따른 정확도 차이를 알아보고자 설계된 부분이 없고, 많은 양의 데이터를 확보하거나 기관을 세부적으로 분류하는 것, 최신의 분류 모델을 활용하는 것으로 판별 정확도를 향상시키는 것이 주요한 연구 목적이었다. 연구에 있어서 데이터 수집이 중요한 단계인 만큼, 학습 및 검증 데이터의 구성 차이에 따른 결과 변화를 수치로써 비교하는 과정이 필요하다[15,16].

따라서 본 연구에서는 후두, 식도, 위가 관찰되는 내시경 이미지를 이용하여 데이터 세트를 크게 세 가지로 구성하고 딥러닝 모델에 학습시켜 정확도를 비교함으로써, 데이터 세트 구성에 따른 정확도의 차이 유무 및 정도를 분석하였다. 데이터 세트 각각은 데이터의 구성과 비율이 다르며 데이터 세트 A, B, 그리고 C로 정의하였다. 데이터 세트 A는 정상 이미지(normal, 정상으로 판단되는 사진)로 훈련한 딥러닝 모델이 정상 이미지와 질병 이미지(abnormal, 질병이 관찰되는 사진)를 기관별로 분류하는 정확도를 알아볼 수 있도록 구성하였다. 데이터 세트 B는 의료 시설에서 별도의 조정 없이 데이터를 모아 딥러닝 모델을 학습시켰을 때의 정확도를 알아보고자, 내시경 검사를 받은 환자 중 정상으로 진단받은 환자와 궤양 또는 암으로 진단받은 환자의 평균 비율을 반영하여 훈련 데이터를 구성하였다. 데이터 세트 C는 데이터 세트 B의 훈련 데이터에서 정상 이미지와 질

Table 1. Number of images which are collected for data.

	Normal	Lesion	Total
Larynx	170	172	342
Esophagus	200	108	308
Stomach	200	135	335
Total	570	415	985

병 이미지 사이의 비율 차를 줄였을 때 딥러닝 모델이 가지는 정확도를 알아볼 수 있도록 구성하였다. 세 가지 데이터 세트에서 시험 데이터를 구성하는 이미지 종류와 종류별 이미지 수는 모두 동일하게 구성하였다.

2. 연구 방법

2.1 데이터 수집

가천대 길병원에서 2014년 7월 2일부터 2021년 7월 19일까지 내시경 검사를 받은 성인 환자를 대상으로 총 985장의 내시경 사진을 수집하였다. 수집한 데이터는 후두가 관찰되는 사진 342장, 식도가 관찰되는 사진 308장, 그리고 위가 관찰되는 사진 335장이며, 위 사진은 위몸통 또는 날문안뜰부위의 사진만 취합하였다. 후두, 식도, 위 각각 정상 이미지와 질병 이미지를 모았고 구체적인 수치는 Table 1과 같다. 질병은 궤양 또는 암에 한정하였다. 위의 방법으로

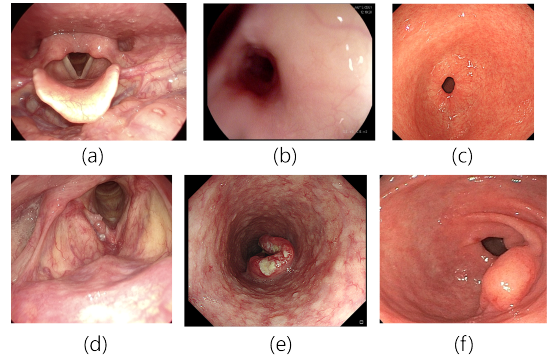


Fig. 1. Examples of collected image data from endoscope. (a) normal larynx image, (b) normal esophagus image, (c) normal stomach image, (d) larynx lesion image, (e) esophagus lesion image, and (f) stomach lesion image.

수집한 데이터 중 이미지를 종류별로 한 장씩 임의로 골라 Fig. 1에 제시하였다. 본 연구에서 사용되는 데이터는 가천대 길병원의 임상시험심사위원회(IRB) (2021-286) 의 허가를 받았다.

2.2 데이터 전처리

내시경 종류, 데이터 수집 시의 내시경기기 설정 상태 등의 차이로 인해 데이터마다 테두리에 있는 여백의 유무 및 두께에 차이가 있었다. 딥러닝 학습 시에 데이터의 여백이 가지는 불필요한 정보를 제거

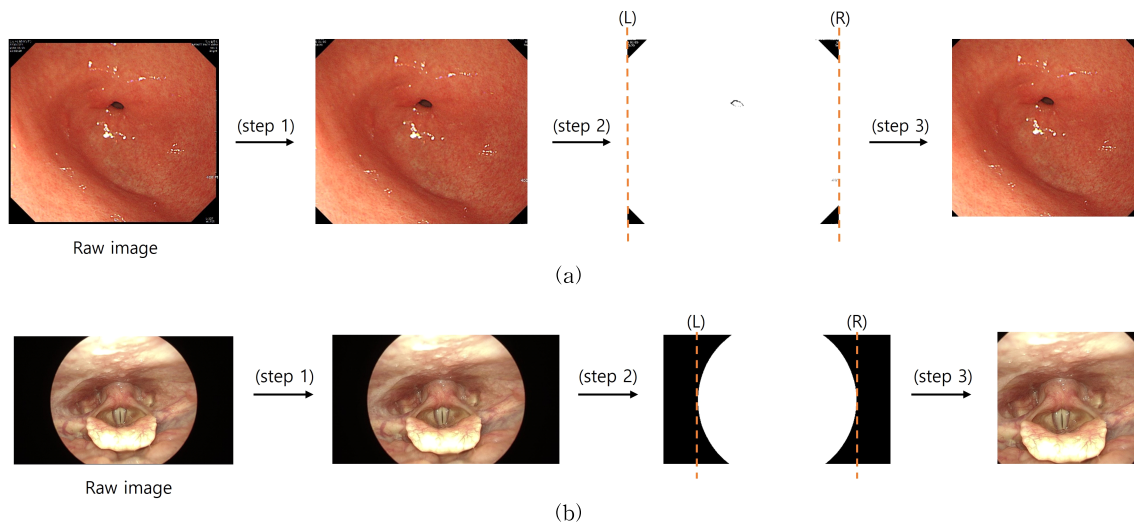


Fig. 2. Image preprocessing (No marks for (a) or (b) if a side of an image corresponds to them). (a) preprocessing normal stomach image and (b) preprocessing normal larynx image.

하기 위하여 Fig. 2와 같이 필요한 부분만 추출하였다. 대부분의 데이터 테두리에 검은 여백이 존재했으므로, 모든 데이터의 테두리를 같은 두께로 잘라내었다(step 1). 이후 데이터마다 가로 방향과 세로 방향에 남아있는 여백의 두께가 달랐기 때문에, 두 방향에서의 여백과 추출하려는 부분의 경계 지점을 파악하고자 하였다. 이를 위해 이미지 내에서 검은색이 아닌 부분은 모두 흰색으로 변경하여, 이미지가 흑백으로 표현되도록 하였다(step 2). 왼쪽 모서리에서 오른쪽으로 이동했을 때 최초로 검은색에서 흰색으로 바뀌는 지점(L)과 오른쪽 모서리에서 왼쪽으로 이동했을 때 최초로 검은색에서 흰색으로 바뀌는 지점(R)을 찾고, (L)에서 (R)까지의 거리에 0.05를 곱한 값을 구하였다. 흑백으로 변형하기 전의 이미지의 좌, 우에 위치한 모서리에서 그 값을 더하거나 뺀을 때 도달하는 지점에서 세로 방향으로 데이터를 잘라내었다. 위, 아래의 모서리에서도 동일한 방법으로 여백을 제거하였다. (L)에서 (R)까지의 거리에 0.05를 곱한 값을 사용한 이유는 두 가지이다. 먼저, 이미지마다 각 변의 길이와 이미지의 넓이가 달라, 수치의 차이와 무관하게 각 변에서 동일한 길이만큼 제거할 시 같은 조건 내에서 전처리 과정이 이루어지지 않는다. 각 변 길이의 일정한 비율에 해당하는 길이를 제거하여야 각 이미지의 너비와 높이 비는 유지하며 제거되는 사진 넓이의 비율이 통일된다. 또한, 이미지의 네 꼭지점에 존재하는 여백을 줄이되 이미지의 손실은 낮추기 위해서 작은 값인 0.05를 전처리에 활용할 비율 값으로 지정하였다. 위에서 정의한 (L) 또는 (R)이 존재하지 않는 데이터일 경우, 데이터의 모서리를 (L) 또는 (R)로 설정하여 진행하였다(step 3). 이후에는 딥러닝 학습을 위해 데이터의 너비와 높이가 같도록 조정하였다.

2.3 데이터 구성

데이터 세트를 총 세 가지로 구성하였고, 데이터 세트 A, B, 그리고 C로 정의하였다. 각 데이터 세트는 Table 2와 같이 데이터 구성과 비율에 차이가 있다. 시험 데이터(test data)는 세 가지 데이터 세트 모두에서 정상 이미지 34장과 질병 이미지 34장으로 통일했으며 같은 데이터를 사용하였다. 시험 데이터에 대한 결과 분석은 정상 이미지와 질병 이미지에 대해 각각 수행할 것이므로 훈련 데이터, 검증 데이터 그리고 전체 시험 데이터의 비율은 3:1:2, 시험 데이터에서 정상 이미지와 질병 이미지 각각으로 보았을 때의 비율은 3:1:1로 구성하였다. 비율을 데이터의 실제 장 수로 변환하는 과정에서 소수점이 나온 경우 반올림하여 처리하였고, 훈련 데이터와 검증 데이터의 합은 세 가지 데이터 세트에서 일정하도록 구성하였다. 단, 데이터 세트 B에서 질병 데이터의 검증 데이터는 합을 일정하게 맞출 경우 장 수가 매우 적어지기 때문에, 데이터의 장 수를 계산할 때 올림을 해주어 다른 데이터 세트보다 검증 데이터 수가 한 장 더 많게 구성하였다.

(1) 데이터 세트 A

정상 이미지로 훈련 데이터와 검증 데이터를 구성하고, 시험 데이터는 정상 이미지와 질병 이미지를 1:1로 혼합하여 구성했다. 정상 이미지와 질병 이미지가 모두 있는 시험 데이터를 통해 딥러닝 모델의 성능을 검증하고, 정상 이미지만으로 학습했을 때 질병 이미지를 얼마나 정확하게 분류할 수 있는지를 각각 분석하였다.

(2) 데이터 세트 B

임상에서의 내시경 데이터를 특별한 조정 없이 모아 학습 데이터와 검증 데이터로 활용했을 때, 인공지능이 정상 이미지와 질병 이미지에서 기관별로 분

Table 2. Data composition for training deep neural networks.

	Dataset A		Dataset B		Dataset C	
Type of image data	Normal	Lesion	Normal	Lesion	Normal	Lesion
Train data	102	-	92	10	68	34
Validation data	34	-	31	4	23	11
Test data	34	34	34	34	34	34
Total	204		205		204	

류하는 정확도를 알아보고자 하였다. 2018년 통계자료에 따르면, 내시경 검사를 받으러 온 환자 중에서 정상으로 판단되는 환자와 궤양 또는 암이 발견되는 환자의 비율은 9:1이다. 이를 반영하기 위해 훈련 데이터와 검증 데이터를 구성할 때 정상 이미지와 질병 이미지 비율을 9:1이 되도록 하였다. 시험 데이터 결과를 통해 학습 데이터와 검증 데이터의 비율에 따라 정상 이미지와 질병 이미지 각각에 대한 분류 정확도를 평가하도록 하였다.

(3) 데이터 세트 C

임상에서 자연스럽게 모이는 데이터가 아닌, 의도적으로 질병 이미지에 대한 비율을 늘려 인공지능 학습 데이터와 검증 데이터를 구성하였을 때 인공지능이 시험 데이터의 기관을 분류하는 정확도에 대한 변화 유무 및 정도를 알아보고자 하였다. 훈련 데이터와 검증 데이터를 구성할 때 정상 이미지와 질병 이미지 비율이 2:1이 되도록 구성하였다.

2.4 딥러닝 모델의 학습

실험에 사용한 딥러닝 아키텍처에서는 2차원 컨볼루션 (convolution 2D) 커널, 배치정규화 (batch normalization), 렐루 (ReLU, rectified linear unit) 활성화 함수를 이용하였다. 또한, 맥스풀링 (maxpooling)을 이용한 총 5개의 해상도처리 단계를 포함한다. 내시경 위치에 대한 출력 값은 GAP (global average pooling) 및 30%의 드롭아웃 (drop out)을 거쳐 출력되도록 설계하였다(Fig. 3). 딥러닝 네트워크 학습에 사용 된 내시경 영상의 입력사이즈, 배치사이즈, epoch은 각각 $512 \times 512 \times 3$, 64, 500으로 설정하였다. 본 연구에서 딥러닝 네트워크의 학습은 파이썬 ver. 3.6.12에서 TensorFlow-GPU ver. 1.15.4와

Keras ver. 2.2.5 프레임워크를 이용하여 수행하였다.

2.5 통계 분석

시험 데이터를 통한 분류 결과 수치는 두 가지 방법을 통해 계산하였다. 첫 번째로, 정확도를 계산하였다. 딥러닝 모델이 시험 데이터의 각 이미지에 대해서 후두, 식도, 위일 확률을 각각 추론하도록 설계하였다. 이후, 확률 수치들을 0.5를 기준으로 0.5보다 크면 1로, 작으면 0으로 변환하였다. 변환 결과를 시험 데이터 답과 비교하고 정답률을 바탕으로 정확도를 계산하였다. 데이터 세트마다 정확도를 구하여, 전체적인 정확도 변화와 양상을 분석하였다.

두 번째로, ROC를 활용하였다. 세 가지 데이터 세트에 대한 ROC 그래프를 한 곳에 그려 시각적으로 비교할 수 있도록 했다. 또한, 데이터 세트마다 정확도 변화의 통계적 유의성을 분석하기 위해 ROC 비교를 진행하였다.

3. 연구 결과

학습된 딥러닝 모델을 대상으로 시험 데이터에 대한 데이터 세트별 정확도, 민감도(sensitivity), 특이도(specificity) 그리고 AUC 값을 구하였다(Table 3). 이때, 시험 데이터에서 정상 이미지와 질병 이미지를 나누어 데이터의 종류가 정상일 때와 질병일 때 각각에 대한 값을 구하였다. 시험 데이터 중에서 정상 이미지에 대한 정확도는, 데이터 세트 A에서는 100.00%, 데이터 세트 B에서는 98.37%, 데이터 세트 C에서는 98.04%가 도출되었다. 한편, 질병 이미지는 정상 이미지에 대한 정확도보다 낮은 값으로, 데이터 세트 A에서는 81.37%, 데이터 세트 B에서는 83.99%, 데이터 세트 C에서는 85.95%가 도출되었다. 정상 이미지에 대한 AUC 값은, 데이터 세트 A에서 1.000,

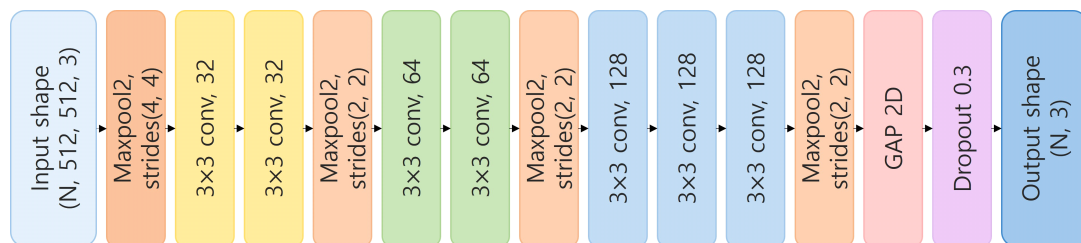


Fig. 3. The deep learning network architecture for anatomical site classification. Abbreviations : Convolution3, 3×3 convolution filter; Maxpool2, 2×2 maxpool filter; GAP 2D, Global Average Pool 2D.

Table 3. Accuracy, sensitivity, specificity and AUC(95% confidence interval) by each dataset.

Dataset	Accuracy	Sensitivity	Specificity	AUC
Dataset A, Normal	100.00% (98.80%~100.00%)	100.00% (96.45%~100.00%)	100.00% (98.21%~100.00%)	1.000 (0.988~1.000)
Dataset B, Normal	98.37% (96.23%~99.47%)	97.06% (91.64%~99.39%)	99.02% (96.50%~99.88%)	0.999 (0.987~1.000)
Dataset C, Normal	98.04% (95.78%~99.28%)	97.06% (91.64%~99.39%)	98.53% (95.76%~99.70%)	0.994 (0.977~0.999)
Dataset A, Lesion	81.37% (76.55%~85.58%)	69.61% (59.71%~78.33%)	87.25% (81.89%~91.50%)	0.873 (0.831~0.908)
Dataset B, Lesion	83.99% (79.39%~87.91%)	74.51% (64.92%~82.62%)	88.73% (83.56%~92.72%)	0.892 (0.852~0.925)
Dataset C, Lesion	85.95% (81.54%~89.64%)	78.43% (69.19%~85.96%)	89.71% (84.70%~93.51%)	0.914 (0.876~0.943)

데이터 세트 B에서 0.999, 그리고 데이터 세트 C에서 0.994를 기록하였다. 질병 이미지에서는, 데이터 세트 A에서 0.873, 데이터 세트 B에서 0.892, 그리고 데이터 세트 C에서 0.914의 AUC 값을 기록하였다.

정확도에서의 통계적 유의성을 분석하기 ROC 비교를 진행하였다. 위의 AUC 값을 바탕으로 Med

Calc를 이용하여 정상 이미지와 질병 이미지에 대한 ROC 그래프를 그렸고(Fig. 4), 아래의 Table 4과 같이 ROC 비교를 진행한 결과에 대한 p값을 얻었다.

ROC 비교 결과 질병 이미지에서는 데이터 세트 A와 데이터 세트 B 간에, 그리고 데이터 세트 A와 데이터 세트 C 간에 통계적으로 유의한 차이를 보였다($p=0.008$, $p=0.006$). 정상 이미지에서는 데이터 세트 B와 데이터 세트 C, 그리고 데이터 세트 A와 데이터 세트 C 간에 통계적으로 유의한 차이를 보였다($p=0.048$, $p=0.044$).

Table 4. p-value from ROC comparison of each dataset.

Type of image data	Two datasets compared		
	Dataset A Dataset B	Dataset B Dataset C	Dataset A Dataset C
Normal	0.174	0.048	0.044
Lesion	0.008	0.095	0.006

4. 고찰 및 결론

본 연구에서는 데이터 세트 구성에 따른 정확도의

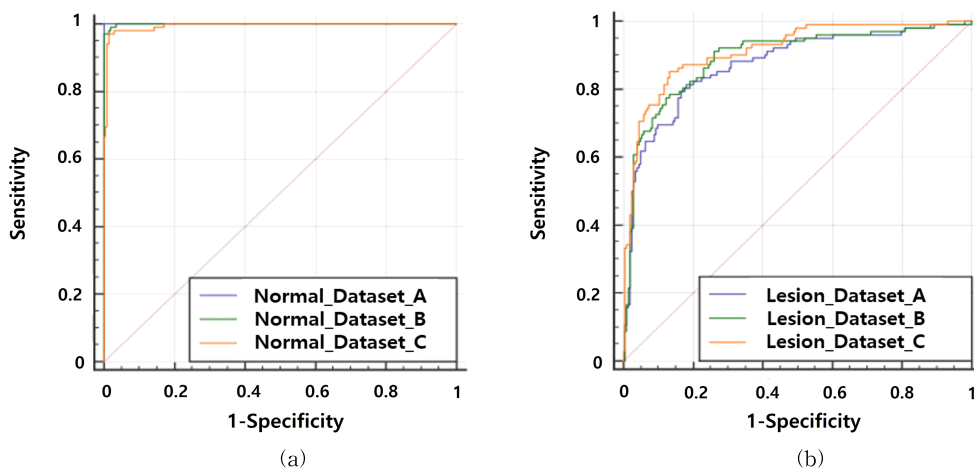


Fig. 4. ROC curve by dataset, (a) normal image and (b) lesion image.

차이 유무 및 정도를 알아보고자 하였다. 이를 위해 내시경 데이터에서 후두, 식도, 위가 관찰되는 이미지를 이용하여 데이터 세트를 세 가지로 구성하고 딥러닝 모델에 학습시켜 기관별로 데이터를 분류하는 정확도를 비교하였다. 세 데이터 세트는 학습 데이터와 검증 데이터의 구성에 차이를 두었다. 데이터 세트 A는 정상 이미지로 구성하였고, 데이터 세트 B는 정상 이미지와 질병 이미지를 9:1로 섞어 구성하였다. 데이터 세트 C는 정상 이미지와 질병 이미지를 2:1로 섞어 구성하였다. 세 데이터 세트의 시험 데이터는 정상 이미지와 질병 이미지를 1:1로 섞어 모두 함께 구성하였다. 결과를 도출할 때는 시험 데이터에서 정상 이미지와 질병 이미지를 나누어 진행했으며, 정확도를 비교하고 ROC 비교로 세부적인 분석을 하였다.

정확도 계산 결과, 정상 이미지에 대해서는 데이터 세트 A에서 1.00, 데이터 세트 B에서 0.984, 데이터 세트 C에서 0.980을 얻었다. 데이터 세트의 구성 차이가 정상 이미지를 분류하는 정확도에 영향을 주었지만, 세 데이터 세트 모두에서 0.980 이상의 수치를 보였다. 한편, 질병 이미지에 대해서는 데이터 세트 A에서 0.816, 데이터 세트 B에서 0.841, 데이터 세트 C에서 0.861이 도출되었다. 이를 통해, 학습 및 검증 데이터에서의 질병 이미지 유무가 새로운 질병 이미지를 분류하는 정확도에 영향을 끼치며, 질병 이미지가 높은 비율로 존재하면 새로운 질병 이미지를 더 정확하게 분류할 수 있음을 유추할 수 있다.

결과를 정확하게 비교하기 위해 ROC 비교를 진행하였다. 질병 이미지에서는 데이터 세트 A와 데이터 세트 B 사이, 데이터 세트 A와 데이터 세트 C 사이에서 유의미한 차이가 있었다($p=0.008$, $p=0.006$). 이를 통해, 학습 및 검증 데이터를 정상 이미지로만 구성하는 것보다 질병 이미지를 포함하는 것이 새로운 질병 이미지를 분류할 때 유의미한 차이로 더 정확하게 분류해낼 수 있음을 알 수 있다. 정상 이미지에 대해서는 데이터 세트 A와 데이터 세트 B 사이에 유의미한 차이가 존재하지 않았다. 학습 및 검증 데이터의 구성이 정상 이미지로 이루어질 때에서 정상 이미지와 질병 이미지가 9:1 비율로 존재할 때로 바뀌어도 정상 이미지를 기관별로 구분하는 데에는 정확도에 큰 영향을 주지 않았음을 추론할 수 있다.

한편, 정상 이미지에서는 데이터 세트 B과 데이터 세트 C 사이, 데이터 세트 A와 데이터 세트 C 사이에

서 유의미한 차이가 있었다($p=0.048$, $p=0.044$). 학습 및 검증 데이터 구성을 데이터 세트 A에서 데이터 세트 B으로 변경했을 때에는 의미 있는 변화가 없었지만, 데이터 세트 C에서는 정상 이미지와 질병 이미지의 비율이 상대적으로 크게 변화해 새로운 정상 이미지를 분류할 때 그 정확도에 영향을 준 것으로 추측된다.

비교실험결과, 학습 및 검증 데이터의 구성에 변화를 주었을 때 정상 이미지 또는 질병 이미지를 기관별로 정확하게 구별하는 정도에 유의미한 차이가 존재하였다. 특히 질병 이미지에 있어서는 학습 데이터와 검증 데이터에 질병 이미지가 없었을 때보다 포함되었을 때 유의미하게 정확도가 올라갔음을 볼 수 있었고, 새로운 정상 이미지에 있어서는 학습 및 검증 데이터에서 정상 이미지의 비율이 크게 감소했을 때 유의미하게 정확도가 감소하였다.

최적의 비율에 가장 근접한 데이터 세트는 데이터 세트 B이다. 데이터 세트 B는 데이터 세트 A보다 질병 이미지를 기관별로 더 정확하게 분류할 수 있고, 데이터 세트 C보다 정상 이미지를 기관별로 더 정확하게 분류할 수 있음을 ROC 비교 결과를 통해 확인할 수 있었다. 또한, 데이터 세트 A에서의 정상 이미지 비율이 더 높았음에도 불구하고 정상 이미지 분류 능력은 통계적으로 유의미한 차이가 없었다. 질병 이미지 분류 능력에 대해서는 데이터 세트 B와 질병 이미지 비율이 더 높은 데이터 세트 C를 비교했을 때, 통계적으로 유의미한 차이를 확인할 수 없었다.

데이터 구성에 대한 비율 설정이 연속적이지 않았기 때문에 이번 연구로 최대한 정확도를 높이는 최적의 비율을 확정할 수 없다. 하지만, 결과를 통해 최적의 비율에 근접한 수치를 확인하였고, 향후 연구를 정상 이미지와 질병 이미지 간 비율에 연속적으로 변화를 주면서 진행한다면 정상 이미지와 질병 이미지 모두에서 높은 정확도를 보이는 모델을 학습시키는 비율을 확정할 수 있을 것이라 판단된다.

이전에 시행된 연구들에는 내시경 이미지를 분류하는 프로그램을 개발할 때 데이터 세트의 구성에 따라 이미지 분류의 정확도가 달라지는지에 대해 집중적으로 분석하는 시각이 없었다. 본 연구를 통해 딥러닝을 통한 이미지 분류 인공지능을 만들 시에는 데이터 세트를 구성하는 데이터의 종류와 그 비율을 충분히 고려해야 하며, 이에 따라 모델의 정확도에

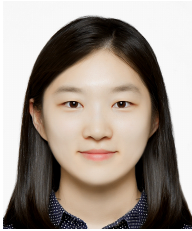
유의미한 차이가 있을 수 있다는 가능성을 수치를 통해 보였다. 임상에서 활용될 인공지능 모델을 학습시킬 때 최대한 많은 데이터를 수집하는 것이 좋으나, 아직은 이상적인 딥러닝 기술 실현을 위한 대규모 데이터 인프라 구축에 있어 한계가 많이 존재한다. 국가나 큰 기관에서 데이터 수집을 위해 다양한 프로젝트가 이루어지고 있지만, 많은 양의 데이터에 대한 딥러닝 연구자들의 접근 및 활용은 쉽지 않다. 이러한 한계점은 딥러닝 모델에 편향을 발생시킬 가능성이 높기 때문에, 확보한 데이터 내에서 높은 수준의 딥러닝 학습을 이루어내는 방법이 필요하다. 본 연구에서 제시한 결과를 바탕으로 제한된 데이터로 최고의 정확도를 가질 수 있는 최적의 데이터 구성 비율을 밝혀낸다면, 이미지 분류 인공지능을 연구 시 필요한 데이터 수집 규모의 벽을 낮추어 데이터 부족에서 발생하는 한계를 극복할 수 있을 것이다.

진단 기기와 치료기기의 자동화를 이루기 위해서는, 현재 연구가 활발하게 진행되고 있는 정상과 질병을 구분하는 기술 이전에 의료기기의 위치를 정확하게 파악하는 기술이 선행되어야 한다. 의료기기의 위치를 자동으로 파악하여 그 위치에 알맞은 최적의 진단 또는 치료 알고리즘을 스스로 실행하는 것이 의료기기의 완전한 자동화이기 때문이다. 이미지 분류의 성능을 최대로 하는 데이터 구성 비율을 알아내어 인공지능 프로그램의 학습에 반영한다면, 활용 가능한 데이터를 통해 최적의 학습을 진행함으로써 의료기기의 위치 파악이 가능해질 것이고, 산업적 측면에서 자동화가 이루어진 진단기기 또는 치료기기 개발이 본격화될 것이다.

REFERENCE

- [1] Annual report of cancer statistics in Korea in 2018(2021), <https://ncc.re.kr/cancerStatsView.ncc?bbsnum=558&searchKey=total&searchValue=&pageNum=1> (accessed August 25, 2021).
- [2] J.G. Yoon, J.K. Kim, D.H. Lee, J.I. Kim, and S.W. Kim, "Clinical Study for Peptic Ulcer," *The Korean Journal of Gastroenterology*, Vol. 36, No. 2, pp. 336-347, 2000.
- [3] Y. Wang, L. Zhu, W. Xia, and F. Wang, "Anatomy of Lymphatic Drainage of the Esophagus and Lymph Node Metastasis of Thoracic Esophageal Cancer," *Cancer Management and Research*, Vol. 10, pp. 6295-6303, 2018.
- [4] J.W. Cho, S.C. Choi, and J.Y. Jang, "Lymph Node Metastases in Esophageal Carcinoma: An Endoscopist's View," *Clinical Endoscopy*, Vol. 47, No. 6, pp. 523-529, 2014.
- [5] A.R. Pimenta-Melo, M. Monteiro-Soares, D. Libânio, and M. Dinis-Ribeiro, "Missing Rate for Gastric Cancer During Upper Gastrointestinal Endoscopy: A Systematic Review and Meta-Analysis," *European Journal of Gastroenterology & Hepatology*, Vol. 28, No. 9, pp. 1041-1049, 2016.
- [6] E.R. Santiago, N. Hernanz, H.M. Marcos-Prieto, M.Á. De-Jorge-Turrión, E. Barreiro-Alonso, et al., "Rate of Missed Oesophageal Cancer at Routine Endoscopy and Survival Outcomes: A Multicentric Cohort Study," *United European Gastroenterology Journal*, Vol. 7, No. 2, pp. 189-198, 2019.
- [7] Y. Zhu, Q. Wang, M. Xu, Z. Zhang, J. Cheng, Y. Zhong, et al., "Application of Convolutional Neural Network in the Diagnosis of the Invasion Depth of Gastric Cancer Based on Conventional Endoscopy," *Gastrointestinal Endoscopy*, Vol. 89, No. 4, pp. 806-815, 2019.
- [8] H. Alaskar, A. Hussain, N. Al-Aseem, P. Liatsis, and D. Al-Jumeily, "Application of Convolutional Neural Networks for Automated Ulcer Detection in Wireless Capsule Endoscopy Images," *Sensors*, Vol. 19, No. 6, 1265, 2019.
- [9] Y.J. Seol, Y.J. Kim, K.H. Nam, and K.G. Kim, "Comparison on the Deep Learning Performance of a Field of View Variable Color Images of Uterine Cervix," *Journal of Korea Multimedia Society*, Vol. 23, No. 7, pp. 812-818, 2020.
- [10] H. Takiyama, T. Ozawa, S. Ishihara, M. Fujishiro, S. Shichijo, S. Nomura, et al., "Automatic Anatomical Classification of Esophagogastrroduodenoscopy Images Using Deep Convolutional Neural Networks," *Scientific Reports*,

- Vol. 8, No. 1, pp. 7497, 2018.
- [11] Q. He, S. Bano, O.F. Ahmad, B. Yang, X. Chen, P. Valdastrì, et al., "Deep Learning-based Anatomical Site Classification for Upper Gastrointestinal Endoscopy," *International Journal of Computer Assisted Radiology and Surgery*, Vol. 15, No. 7, pp. 1085-1094, 2020.
- [12] J.W. Park, Y. Kim, W.J. Kim, and S.J. Nam, "Automatic Anatomical Classification Model of Esophagogastroduodenoscopy Images Using Deep Convolutional Neural Networks for Guiding Endoscopic Photodocumentation," *Korean Society of Computer Information*, Vol. 26, No. 3, pp. 19-28, 2021.
- [13] L. Chen, A. Cruz, S. Ramsey, C.J. Dickson, J.S. Duca, V. Hornak, et al., "Hidden Bias in the DUD-E Dataset Leads to Misleading Performance of Deep Learning in Structure-based Virtual Screening," *PLoS One*, Vol. 14, No. 8, pp. e0220113, 2019.
- [14] P. Dufour, S. Bhartiya, P.S. Dhurjati, and F.J. Doyle III, "Neural Network-based Software Sensor: Training Set Design and Application to a Continuous Pulp Digester," *Control Engineering Practice*, Vol. 13, No. 2, pp. 135-143, 2005.
- [15] T. Tommasi, N. Patricia, B. Caputo, and T. Tuytelaars, *A Deeper Look at Dataset Bias*, In: Csurka G. (eds) Domain Adaptation in Computer Vision Applications. Advances in Computer Vision and Pattern Recognition. Springer, Cham, 2017.
- [16] S.W. Kwon, M.H. Kim, J.H. Kim, and S.W. Hong, "Changes in the Performance for Predicting Inappropriate Thermal Images according to the Composition of Datasets," *Transactions of the Korean Society of Mechanical Engineers A*, Vol. 44, No. 12, pp. 933-940, 2020.



서 정 민

2020년 3월~현재 가천대학교 의
과대학 재학 중
관심분야: 인공지능, 의료기기,
의료정보



임 상 현

2020년 2월 계명대학교 의용공학
과(학사)
2022년 2월 가천대학교 의용생체
공학과(석사)
관심분야: 인공지능, 기계학습, 인
공신경망, 의료영상



김 영 재

2013년 2월 광운대학교 플라즈마
바이오디스플레이학과
(석사)
2018년 2월 광운대학교 플라즈마
바이오디스플레이학과
(박사)

2008년 1월~2017년 4월 국립암센터 의공학연구과
2019년 9월~현재 가천대학교 의용생체공학과 조교수
관심분야: 의료영상, 인공지능, 빅데이터, 컴퓨터보조진단



정 준 원

1997년 2월 경희대 의학사
2008년 2월 울산대 석사
2012년 2월 울산대 학사
1997년 2월 서울아산병원 인턴
2002년 2월 서울아산병원 전공의
2007년 2월 서울아산병원 전임의

2009년 9월 서울아산병원 임상교수
2017년 2월 가천대 길병원 조교수
2020년 2월 가천대 길병원 부교수
2020년 3월~현재 가천대 길병원 교수
관심분야: 인공지능, 소화기 내시경, 헬리코박터 감염



김 광 기

2006년 2월 서울대학교 공학대학
박사
2007년 7월 국립암센터 선임/책
임 연구원
2017년 2월~현재 가천대학교 의
공학과 교수

2017년 2월~현재 가천대길병원 의료기기R&D센터 센터
장, 의용생체공학회/대한의료정보학회/대한의료로
봇학회/대한혁신의료기술학회 이사로 활동중임
관심분야: 의료기기, 의료로봇, 의료정보, 인공지능