Original Article

J Korean Soc Radiol 2021;82(5):1196-1206
https://doi.org/10.3348/jksr.2020.0177
eISSN 2288-2928

대한영상의학회지

Construction of a Standard Dataset for Liver Tumors for Testing the Performance and Safety of Artificial Intelligence-Based Clinical Decision Support Systems

인공지능 기반 임상의학 결정 지원 시스템 의료기기의 성능 및 안전성 검증을 위한 간 종양 표준 데이터셋 구축

Seung-seob Kim, MD[1], Dong Ho Lee, MD[2], Min Woo Lee, MD[3], So Yeon Kim, MD[4], Jaeseung Shin, MD[1], Jin-Young Choi, MD[1]*, Byoung Wook Choi, MD[1]

[1]Department of Radiology and Research Institute of Radiological Science, Severance Hospital, Yonsei University College of Medicine, Seoul, Korea
[2]Department of Radiology, Seoul National University Hospital, Seoul National University College of Medicine, Seoul, Korea
[3]Department of Radiology and Center for Imaging Science, Samsung Medical Center, Sungkyunkwan University School of Medicine, Seoul, Korea
[4]Department of Radiology, Asan Medical Center, University of Ulsan College of Medicine, Seoul, Korea

Purpose To construct a standard dataset of contrast-enhanced CT images of liver tumors to test the performance and safety of artificial intelligence (AI)-based algorithms for clinical decision support systems (CDSSs).

Materials and Methods A consensus group of medical experts in gastrointestinal radiology from four national tertiary institutions discussed the conditions to be included in a standard dataset. Seventy-five cases of hepatocellular carcinoma, 75 cases of metastasis, and 30–50 cases of benign lesions were retrieved from each institution, and the final dataset consisted of 300 cases of hepatocellular carcinoma, 300 cases of metastasis, and 183 cases of benign lesions. Only pathologically confirmed cases of hepatocellular carcinomas and metastases were enrolled. The medical experts retrieved the medical records of the patients and manually labeled the CT images. The CT images were saved as Digital Imaging and Communications in Medicine (DICOM) files.

Results The medical experts in gastrointestinal radiology constructed the standard dataset of contrast-enhanced CT images for 783 cases of liver tumors. The performance and safety of the AI algorithm can be evaluated by calculating the sensitivity and specificity for detecting and

Received October 13, 2020
Revised November 24, 2020
Accepted February 4, 2021

***Corresponding author**
Jin-Young Choi, MD
Department of Radiology and Research Institute of Radiological Science, Severance Hospital, Yonsei University College of Medicine, 50-1 Yonsei-ro, Seodaemun-gu, Seoul 03722, Korea.

Tel 82-2-2228-7400
Fax 82-2-2227-8337
E-mail gafield2@yuhs.ac

ORCID iDs
Seung-seob Kim
https://orcid.org/0000-0001-6071-306X
Dong Ho Lee
https://orcid.org/0000-0001-8983-851X
Min Woo Lee
https://orcid.org/0000-0001-9048-9011
So Yeon Kim
https://orcid.org/0000-0001-6853-8577
Jaeseung Shin
https://orcid.org/0000-0002-6755-4732
Jin-Young Choi
https://orcid.org/0000-0002-9025-6274
Byoung Wook Choi
https://orcid.org/0000-0002-8873-5444

characterizing the lesions.

Conclusion The constructed standard dataset can be utilized for evaluating the machine-learning-based AI algorithm for CDSS.

Index terms Liver Neoplasms; Aritificial Intelligence; Machine Learning; Deep Learning; Datasets as Topic

서론

인공지능 알고리즘은 지난 10년간 괄목할만한 발전을 이루고 있다. 일상생활의 수많은 분야에 혁명적인 변화를 일으켰으며, 의료 분야에 있어서도 업무처리 흐름을 최적화하고 환자의 진단 및 치료 정확도를 향상시키는 데에 기여하고 있다(1). 특히 영상의학은 보다 정확한 예후 예측 모델의 정립과 패턴 인식, 그리고 이를 위해 많은 양의 데이터를 필요로 하므로 인공지능의 발전으로부터 가장 많은 이득을 기대할 수 있는 의료 분야이다(1). 의료 영상 인공지능의 전 세계적인 시장 가치는 2018년도의 215억 달러에서 2026년에는 2649억 달러로 상승할 것으로 예측되고 있다(2).

그러나 실제로 인공지능 소프트웨어를 활용하여 일상적인 판독 업무를 수행하는 경우는 아직 드문 실정이다. 비교적 널리 사용되는 소프트웨어들은 대부분 병변이 있을 가능성이 있는 부위를 자동으로 발견하여 표시해 주는 정도에 그치는, 소위 컴퓨터 보조 검출(computer-aided detection) 수준에 머물러 있고, 발견된 병변의 감별진단 및 치료법 제안, 더 나아가 예후 예측까지 가능한 진정한 의미의 임상의학 결정 지원 시스템으로 승인된 제품은 현재까지 존재하지 않는다(1).

의료 영상 관련 인공지능 소프트웨어를 개발하는 데에 있어서 가장 큰 걸림돌은 질 좋은 데이터를 확보하는 일이다(3). “Garbage in, garbage out” 이라는 표현으로 묘사되는 것처럼, 낮은 질의 데이터를 기반으로 개발된 알고리즘은 본질적으로 낮은 성능을 보일 수밖에 없다. 여러 종류의 공개 데이터셋들이 알려져 있고 실제 알고리즘 개발에 활용되기도 했지만, 최근 이 데이터셋들의 라벨링(labeling)이 부족하고, 부정확하다는 우려가 제기되었다(4). 모델 개발과정에서 과최적화(overfitting)를 방지하기 위해서는 학습(training) 데이터셋과 함께 검증(validation) 데이터셋이 필요하고, 개발된 모델을 임상에 이용하기 위해서는 임상 현장에서 일어날 수 있는 모든 상황을 반영할 수 있는 최고 수준의 테스팅(testing) 데이터셋이 필수 불가결하다(3).

본 연구에서는 임상의학 결정 지원 시스템(clinical decision support system)을 위한 간 종양의 조영증강 컴퓨터단층촬영(이하 CT) 영상에 관한 인공지능 알고리즘의 성능과 안전성을 검증할 수 있는 표준 테스팅 데이터셋을 구축하고자 하였다.

대상과 방법

연구에 참여한 4개 의료기관의 연구심의위원회로부터 모두 동의서면제 승인을 받았다(IRB number: 4-2020-0136, H-2004-134-1117, 2020-05-022, 2020-0819).

컨센서스 그룹 구성

국내 4개 3차 의료기관의 복부 영상의학 전문가 4인이 모여 간 종양 진단 알고리즘의 성능과 안전성을 검증하기 위해 표준 데이터셋이 갖춰야 할 조건을 논의하였다. 첫째, 본 연구의 데이터셋에서는 자기공명영상(이하 MRI)이 아닌 CT만을 대상으로 하기로 하였다. 촬영 프로토콜, 장비, 그리고 조영제 등에 있어 표준화가 어려운 MRI보다는 CT를 기반으로 한 알고리즘이 먼저 개발될 가능성이 높다는 합의가 있었기 때문이다. 둘째, 발견된 간 병변의 양성과 악성 여부를 얼마나 정확히 감별진단 해내는지 평가하기 위해, 양성 종양과 악성 종양의 데이터를 적정 비율로 포함시키기로 하였다. 셋째, 알고리즘이 간 병변을 발견하는 데에 있어 병변의 크기에 영향을 받을 수 있으므로, 1 cm 이하의 병변은 배제하고 5 cm 이상의 병변은 악성인 경우가 더 많으므로 배제하기로 하였다. 품질 관리를 위해 복부 영상의학 전문가들이 직접 임상정보 추출 및 수기 라벨링(manual labeling)을 하고, 여러 기관에서 다양한 장비들로 촬영된 복부 CT 영상을 모두 포함함으로써, 대상 모델의 과최적화 정도 확인과 일반적인 적용 가능성을 검증하기로 하였다.

연구 대상

국내 3차 의료기관 4곳에서 촬영된 조영증강 복부 CT 영상을 대상으로 하였다. 데이터셋에 포함할 증례의 개수에 대해서는 명확한 근거를 찾을 수 없으므로, 컨센서스 그룹에서 현실적으로 수집 및 분석 가능하고 알고리즘의 성능을 판정할 수 있을 정도로 추산하였다. 간 종양 진단 알고리즘은 악성 병변을 특성화하는 것이 가장 중요하고 우선적인 역할이므로 악성과 양성의 비율은 3:1로 정하였다. 이는 실제 임상 상황을 그대로 반영하지는 못하지만, 간 종양의 악성과 양성을 구분하는 성능을 판단하기 위해 컨센서스 그룹에서 임의로 설정한 비율이다. 각 기관마다 간세포암 75예, 전이암 75예, 그리고 양성 병변 30–50예씩 수집하여, 총 간세포암 300예, 전이암 300예, 양성 병변 120–200예, 총 700–800명 환자의 CT 영상을 대상으로 하였다. 영상이 촬영된 연도 및 CT 장비의 종류는 가능한 다양하게 포함되도록 하되, 2012년 이전에 촬영된 CT 영상은 영상의 품질이 낮을 가능성이 높으므로 배제하였다. 전이암 및 양성 병변의 경우 최소한 문맥기 영상이 포함되어 있는 조영증강 CT 영상을 대상으로 하였고, 간세포암의 경우에는 동맥기, 문맥기, 지연기가 모두 포함되어 있는 영상만을 대상으로 하였다. 1차 의료기관에서 촬영된 CT 영상을 3차 의료기관에 등록했던 경우도 포함할 수 있도록 하였다. 항암 혹은 방사선 치료를 받았던 환자의 경우에는 치료 전 영상을 사용하도록 하였다.

정답(Ground Truth)의 정의

간세포암과 전이암의 경우 조직학적 확진이 된 병변만을 대상으로 하였다. 간세포암의 경우에는 수술적 절제(간 이식 제외)로 병리 결과가 확인된 경우만을 대상으로 하였고, 전이암의 경우에는 수술적 절제 혹은 생검으로 확인된 경우 모두를 대상으로 하였다. 양성 병변의 경우에는 전형적인 영상 소견과 함께 2년 이상의 기간 동안 변화 없이 관찰되는 경우로 정의하였다.

데이터 라벨링과 수집항목

각 기관의 복부 영상의학 전문가들이 직접 환자의 임상 기록 및 CT 촬영 관련 정보를 추출하였고, 추출할 항목에 대해서는 컨센서스 그룹에서 논의하였다(Table 1). 환자의 성별, 나이, CT 촬영 날짜 및 장비명, 그리고 조영증강 촬영 프로토콜에 대한 세부사항을 기록하기로 하였다. 촬영 프로토콜에 대해서는 동맥기, 문맥기, 지연기 중 어떠한 시기의 영상들이 포함되어 있는지와 관전압 수치를 기록하였다. CT 영상에 대한 라벨링도 복부 영상의학 전문가들이 직접 시행하였다. 환자별로 간내 병변의 개수가 1개인 경우와 2개 이상인 경우를 다양하게 포함하도록 하였고, 한 환자에서 5개 이상의 병변이 있을 경우에는 최대 5개까지만 기록하기로 결정하였다. 각각의 병변들이 위에서 정의하였던 정답 판단 기준 중 어떤 방법으로 간세포암, 전이암, 혹은 양성 병변으로 분류되었는지를 기록하였고, 병변의 간내 위치(퀴노 분절, couinaud segmentation), 크기, 그리고 영상 소견이 전형적인지 비전형적인지에 대해서도 함께 기록하기로 하였다. 각 병변이 가장 대표적으로 보이는 레벨의 CT 이미지를 캡처한 후 빨간색의 네모 박스로 표시함으로써 라벨링한 병변이 어느 것이었는지 명확하게 알 수 있도록 하였다(Fig. 1). CT 이미지는 의료용 디지털 영상 및 통신(Digital Imaging and Communications in Medicine, DICOM) 파일로 저장한 후, 기관에 따라

Table 1. Case Report Form for the CT Data for the Liver Tumors

Identification number	[Institution number] + [Anonymized patient number (1–200)]
Sex	0: Male 1: Female
Age	
CT date	Not earlier than year 2012
CT vendor	
CT model	
CT voltage (kVp)	
CT protocol	0: Portal venous phase 1: Late arterial phase + portal venous phase 2: Late arterial phase + portal venous phase + delayed phase (liver dynamic protocol)
Lesion number	Record up to the fifth lesion, even when the patient had more lesions
Ground truth	0: Hepatocellular carcinoma 1: Metastasis 2: Cyst 3: Hemangioma 4: Other benign lesions (focal nodular hyperplasia, adenoma, abscess, eosinophilic infiltration, arterioportal shunt, etc.)
Method of confirmation	0: Surgical resection 1: Needle biopsy 2: No interval change on follow-up images for more than two years
Size (mm)	Minimum: 10 Maximum: 50
Couinaud segment	
Imaging pattern	0: Typical 1: Atypical

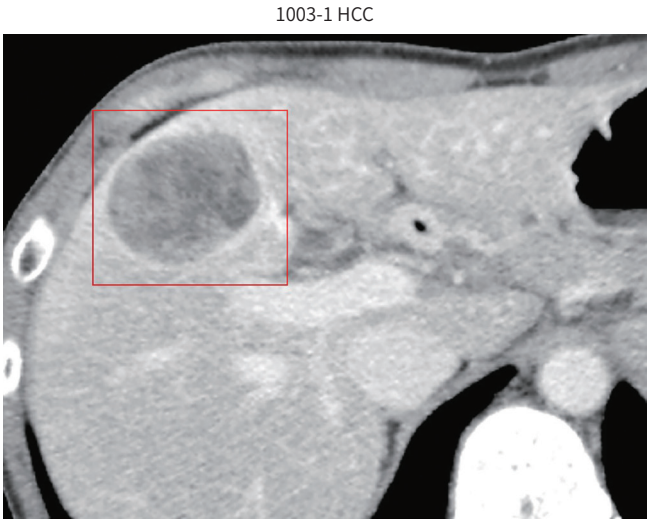

Fig. 1. Example of the captured representative image of the lesion.
An anonymized patient number and a lesion number are shown, along with the ground-truth for this data ('1003', '1', and 'HCC' in this example, respectively). The red box delineates the lesion for more accurate identification and localization.
HCC = hepatocellular carcinoma

상용 소프트웨어(Aview; Coreline Soft, Seoul, Korea), 혹은 공개 소프트웨어를 사용하여 헤더(header)에 기록되어 있던 실제 환자의 이름 및 의료기관명은 삭제하였고, 환자 번호는(사전에 정의된 의료기관번호[1–4] + 익명화된 환자 번호[1–200])의 규칙에 따라 생성된 네 자리의 숫자로 변경하는 방식으로 비식별화하였다.

유효성과 안전성 테스트의 검증

구축한 데이터셋을 이용하여 개발된 알고리즘의 유효성과 안전성을 검증하고 결과 분석이 임상적 유효성을 가지는지와 통계적으로 유의한지 검증을 위한 지표를 선정하였다. 알고리즘의 성능은 병변의 발견 여부 및 특성화의 정확도를 병변 단위로 부여한 정답과의 비교를 통해 민감도(sensitivity)와 특이도(specificity)를 계산하여 평가할 수 있다.

$$\text{민감도} = \frac{\text{진양성(true positive)}}{\text{진양성(true positive)} + \text{위음성(false negative)}}$$

$$\text{특이도} = \frac{\text{진음성(true negative)}}{\text{진음성(true negative)} + \text{위양성(false positive)}}$$

알고리즘이 특정 병변일 확률값(probability)을 연속적인 값으로 함께 제시하는 경우에는, 다양한 역치값(threshold)에 대해 수신자 조작 특성 곡선(receiver operating characteristic [이하 ROC] curve) 분석을 수행하여 곡선하면적(area under the ROC curve; AUC) 값을 계산할 수 있고, 교정 그림(calibration plot)과 Hosmer-Lemeshow goodness-of-fit test를 이용하여 알고리즘의 확률값 계산 모델이 본 데이터셋에 얼마나 적합(fit)한지를 평가할 수 있다. 이러한 통계 지표들은 유병률(prevalence)의 영향을 받지 않으므로, 본 연구에서의 데이터셋에 적용 가능하다(5).

결과

4개 의료기관의 복부 영상의학 전문가들이 직접 데이터 라벨링을 시행한 총 783 증례로 이루어진 간 종양 조영증강 CT의 표준 데이터셋을 구축하였다(Table 2). 악성 병변의 경우 모두 병리 조직 결과를 바탕으로 정답을 부여하였고, 양성 병변의 경우에도 전형적인 영상 소견과 함께 2년 이상의 기간 동안 변화 없이 관찰되는 경우만을 포함하였으므로, 높은 신뢰도의 정답 수준이라고 할 수 있다. 병변의 정확한 위치를 표시한 대표적 이미지 캡처 데이터 또한 구축하였다.

알고리즘의 유효성 검증을 위해 선정된 민감도, 특이도, 곡선하면적, 그리고 Hosmer-Leme-

Table 2. Structure of the Dataset

Target institutions	Four national tertiary institutions
Total number of CT exams (patients)	783 (100%)
CT scanning protocol	
Portal venous phase	213 (27%)
Late arterial and portal venous phases	127 (16%)
Late arterial, portal venous, and delayed phases	443 (57%)
CT vendor	
Siemens	299 (38%)
GE	278 (36%)
Philips	137 (17%)
Others	69 (9%)
Sex (male:female)	497:286
Age (mean ± standard deviation)	61 ± 11
Exams (patients) with malignancy	600 (77%)
Exams (patients) without malignancy	183 (23%)
Total number of hepatic lesions	1160 (100%)
Hepatocellular carcinoma	333 (28%)
Metastasis	567 (49%)
Benign lesion (cyst)	80 (7%)
Benign lesion (hemangioma)	102 (9%)
Benign lesion (others)	78 (7%)
Size of hepatic lesion (mm)*	21 (15–30)
Couinaud segment	
S1	24 (2%)
S2	111 (9%)
S3	114 (10%)
S4	161 (14%)
S5	153 (13%)
S6	199 (17%)
S7	136 (12%)
S8	262 (23%)
Imaging pattern (typical:atypical)	1006:154

*Data are represented as median with the interquartile range in parentheses.

show goodness-of-fit test의 *p*값 등의 지표를 전체 데이터셋을 대상으로 계산할 수 있고, 이 값들이 병변의 크기와 위치(퀴노 분절), 그리고 전형적/비전형적인 영상 소견에 따른 하위 그룹(subgroup)에서 어떤 양상으로 변하는지 파악할 수 있다. 알고리즘의 안전성 평가는 병변 단위가 아닌 환자 단위에서의 악성 병변의 유무를 예측하는 정확도에 대해 상기 통계 지표들을 계산하여 평가할 수 있다. 예를 들어, 국소 간 병변의 CT 영상 소견을 기반으로 발견된 국소 간 병변이 간세포암일 확률을 제시해주는 인공지능 알고리즘을 개발했다면, 본 데이터셋을 이용하여 우선 병변 단위로 각각의 국소 간 병변들이 간세포암일 확률에 대한 수신자 조작 특성 곡선을 그려 적절한 민감도와 특이도의 역치값을 구할 수 있고, 곡선하면적 값을 계산하여 모델의 전체적인 성능을 확인할 수 있다. 본 데이터셋은 병변 단위뿐 아니라 환자 단위로도 라벨링을 시행하였으므로, 알고리즘이 제시한 병변 단위의 확률을 바탕으로 최종적으로 악성 병변이 하나 이상 존재하는 검사와 양성 병변만 존재하는 검사로 얼마나 정확하게 나눌 수 있는지에 대해서도 민감도와 특이도를 계산할 수 있다.

고찰

기계학습(machine learning) 또는 심층학습(deep learning) 기반의 알고리즘 개발에 있어서 가장 중요한 부분은 질 좋은 데이터의 확보에 있다(1). 특히, 그 목표가 특정 질환을 일정 수준 이상의 특이도로 진단하려는 것일 경우, 해당 분야 전문가들에 의해 직접 라벨링된 데이터셋과 높은 수준의 정답을 확보하는 것이 권고된다(6). 본 연구에서는 높은 수준의 라벨링이 부여된, 4개 의료기관에서 촬영한 다양한 간 종양 조영증강 CT 이미지의 표준 데이터셋을 구축하였다.

인공지능 알고리즘은 학습 데이터셋을 선택하는 시작 단계에서부터 본질적으로 여러 편향(bias)을 내포하게 된다(7, 8). 지역 혹은 인종에 따라 큰 유병률의 차이를 보이는 질환들이 있고(9), 같은 지역 내에서도 의료기관의 층위에 따라 진료 및 판독 양식에 차이를 보일 수 있으며(10), CT 장비나 촬영 프로토콜 등에 의해서도 이미지가 일률적인 영향을 받을 수 있다(3). 이 모든 것이 잠재적인 편향의 요소가 되므로, 본 연구에서는 국내의 3차 의료기관에서 촬영되었거나 1차 의료기관에서 촬영된 후 전송된, 다양한 촬영 프로토콜의, 다양한 CT 장비에서의 이미지들을 모두 데이터셋에 포함함으로써, 알고리즘이 이러한 편향을 극복하고 국내 환자를 대상으로 일반적으로 적용될 수 있는지를 테스팅 할 수 있도록 하였다.

기계학습 알고리즘의 데이터 질을 논할 때, 편향과 더불어 가장 중요한 것은 데이터에 부여된 정답의 수준이다. 많은 공개 데이터셋의 경우에서 실제 환자의 진단병리검사 결과나 이후의 추적검사(follow-up) 영상 소견에 대한 고려 없이, 데이터셋에 포함된 단일 시점의 영상 소견만이 라벨링 되어 있다(4, 11). 해당 의료기관에서 실제 작성되었던 판독문 텍스트를 불러와 이를 자연어 분석(natural language processing)을 통해 필요한 정보를 추출한 후 라벨링하는 방법도 사용되고 있지만, 이 방법은 자연어 분석 자체에 내재된 고유의 한계로부터 자유롭지 못하고, 판독문 텍스트의 오류와 자연어 분석 과정상의 오류를 모두 고려해야 한다는 점에 주의해야 한다(12). 따라서 간 종양의 감별진단에 있어 이상적인 정답은 병리 결과라 할 수 있다. 본 연구에서 간세포암과 전이암의 진단은 모두 병리 결과에 기반하도록 하였다. 간세포암의 경우, 경피적 조직 생검으로는

표본 획득 오류(sampling error)로 인해 병합형 간세포암-담관암종(combined hepatocellular-cholangiocarcinoma)이 간세포암으로 오진되는 가능성을 배제할 수 없기에, 오직 수술적 절제로 진단된 경우만을 포함하도록 하였다(13).

간 내부를 해부학적으로 구획 짓는 방식으로 가장 널리 이용되고 있는 것은 퀴노 분절로, 실제 판독문 텍스트나 구조적 보고서(structured report)에서도 이 방식이 주로 사용되고 있다. 그러나 병변이 구획의 경계 부위에 위치하거나 혹은 비슷한 크기의 병변 여러 개가 같은 구획 내에 위치할 경우, 퀴노 분절만으로는 병변을 정확히 특정하는 데에 불충분할 수 있다(14). 본 연구에서는 복부 영상의학 전문가들이 직접 데이터셋의 병변이 대표적으로 나타나 있는 CT 이미지를 캡처한 후 해당 병변을 네모 박스로 표시함으로써, 명확하게 병변을 특정하였다. 이는 추후 병변을 모두 포함한 영상 단면별 라벨링을 하고자 할 때 지표로 활용될 수 있다.

의료기기로의 승인에 있어 성능 못지않게 중요한 것은 안전성이다. 미국 FDA는 스스로 학습하며 변화하는 성질을 지닌 기계학습 인공지능 의료기기의 경우 기존의 의료기기 승인 절차가 적절하지 않을 수 있다는 사실을 인식했다(15). 시판된 이후 추가적으로 축적된 데이터들로 알고리즘의 성능을 향상시키는 과정에서 오히려 안전성이 더 감소할 가능성을 배제할 수 없는 것이다. 본 연구의 표준 데이터셋에서는 간세포암 및 전이암과 같은 악성 병변을 갖고 있는 환자와 오직 양성 병변만을 갖고 있는 환자의 데이터를 함께 포함시킴으로써, 국소 병변의 발견 및 감별진단과 같은 모델의 세부적인 성능뿐 아니라 환자 단위에서의 악성 병변의 유무를 구별해내는 모델의 안전성 측면도 함께 평가할 수 있도록 하였다. 뿐만 아니라, 병변의 영상의학적 소견이 전형적인지 비전형적인지도 함께 라벨링 함으로써, 모델의 성능 및 안전성에 변화가 생겼을 때 어떤 방향으로 미세 튜닝을 해야 할지에 대한 실마리도 제공할 수 있도록 하였다.

CT 영상을 대상으로 간 종양을 감별진단하는 인공지능 소프트웨어에 관한 연구는 아직 많이 보고되어 있지 않고(16), 그중 많은 수가 알고리즘 자체의 공학적 모델의 개발 및 개선에 중점을 둔 연구들이기 때문에(17-24), 실제 데이터에 적용할 수 있도록 공개되어 있는 소프트웨어는 드물다. 간 전이암을 대상으로 원발암의 종류를 예측하는 기능의 심층학습 알고리즘이 보고되어 있으나, 원발성 간 종양이나 양성 간 병변이 배제되어 있다는 점에서 한계가 있다(25). 국소 간 병변의 CT 영상을 대상으로 하여 간세포암, 간세포암을 제외한 악성 간 종양, 조기 간세포암이나 이형성 결절, 혈관종, 그리고 물혹의 다섯 가지 카테고리 중 하나로 감별진단하는 기능의 심층학습 알고리즘도 보고되어 있으나, 대상 간 병변이 포함된 CT 단면 영상을 수기로 잘라내는 과정이 추가로 필요하므로 실제 임상 현장에서 사용하기는 어렵다(26). 따라서 본 연구의 데이터셋으로 성능을 검증해보기에 적합한 CT 기반의 간 종양 감별진단 모델은 아직 없는 실정이다. 향후 실제 판독 업무에서 보조적으로 활용할 수 있고, 본 연구의 데이터셋으로 그 성능을 테스팅 할 수 있는 모델은 수기로 병변을 표시해야 하거나 CT 단면을 잘라내는 등의 추가 작업 없이 병변의 자동 발견 및 분할(auto segmentation)이 되어야 하고, 간에 발생하는 다양한 종류의 병변들에 대해 모두 학습되어야 하며, 양성과 악성 여부를 구별할 수 있어야 한다(27).

본 연구의 데이터셋에는 몇 가지 제한점이 있다. 첫째, 본 데이터셋의 구성에는 실제 인구 집단의 유병률이 반영되어 있지 않고 악성 병변과 양성 병변의 비율이 3:1로 인위적으로 설정되어 있

다. 민감도/특이도와 곡선하면적 등의 통계 지표들은 유병률에 영향을 받지 않지만, 알고리즘이 제시하는 확률값 자체는 유병률의 영향을 받으므로, 진단의 컷오프(cut-off) 값은 본 데이터셋에서의 ROC curve 분석이 아닌 실제 인구집단의 유병률을 고려하여 설정되어야 할 것이다. 둘째, 환자-대조군 연구 방식의 후향적인 디자인으로 구성된 데이터셋이기 때문에, 선택 편향(selection bias)과 범주 편향(spectrum bias)이 있을 수 있고, 계산된 민감도/특이도, 곡선하면적 값 등은 실제 인구집단을 대상으로 하였을 때 기대되는 값보다 더 과장되었을 수 있다. 셋째, 본 연구의 데이터셋에 영상의학과 의사의 판독은 포함되어 있지 않으므로, 인공지능 소프트웨어를 사용함으로써 추가적으로 향상되는 진단 정확도 및 환자의 예후에 관해서는 검증할 수 없다. 궁극적으로 인공지능 알고리즘의 실제 임상적인 유효성을 가장 높은 수준으로 검증하기 위해서는, 전향적으로 모집되고, 시간/공간적으로 무작위로 배정된 환자군에 대해 인공지능 소프트웨어를 활용한 그룹과 그렇지 않은 그룹으로 나눠서 비교하는 임상시험이 필요하겠다.

결론적으로 본 연구에서는 국내 4개 의료기관 783 증례로 구성된 간 종양 조영증강 CT 영상의 표준 데이터셋을 구축하였고, 가능한 최고 수준의 데이터 질을 확보하기 위하여 복부 영상의학 전문가들이 직접 데이터 라벨링을 시행하였다. 구축된 표준 데이터셋은 임상의학 결정 지원 시스템을 위한 기계학습 인공지능 기반 알고리즘의 성능 및 안전성을 평가하는 데에 활용될 예정이고, 알고리즘의 개발 단계에서 이 데이터셋의 정보가 활용되어서는 안 되므로 외부에 공개하지 않을 것이다.

Author Contributions

Conceptualization, all authors; data curation, K.S., L.D.H., L.M.W., K.S.Y., S.J.; formal analysis, K.S., C.J.; funding acquisition, C.J., C.B.W.; investigation, K.S., L.D.H., L.M.W., K.S.Y., S.J.; methodology, all authors; project administration, C.J., C.B.W.; resources, C.J., C.B.W.; software, K.S., C.J., C.B.W.; supervision, C.J., C.B.W.; visualization, K.S., S.J., C.J.; writing—original draft, K.S.; and writing—review & editing, C.J.

Conflicts of Interest

The author has no potential conflicts of interest to disclose.

Funding

This research was supported by a grant (18173의료평331-1[DY0002258200]) from Ministry of Food and Drug Safety in 2020.

REFERENCES

1. Wichmann JL, Willemink MJ, De Cecco CN. Artificial intelligence and machine learning in radiology: current state and considerations for routine clinical implementation. *Invest Radiol* 2020;55:619-627
2. Data Bridge Market Research. Global artificial intelligence in medical imaging market-industry trends-forecast to 2026. Available at. https://www.databridgemarketresearch.com/reports/global-artificial-intelligence-in-medical-imaging-market. Accessed Sep 22, 2020
3. Willemink MJ, Koszek WA, Hardell C, Wu J, Fleischmann D, Harvey H, et al. Preparing medical imaging data for machine learning. *Radiology* 2020;295:4-15
4. Oakden-Rayner L. Exploring large-scale public medical image datasets. *Acad Radiol* 2020;27:106-112
5. Park SH, Han K. Methodologic guide for evaluating clinical performance and effect of artificial intelligence technology for medical diagnosis and prediction. *Radiology* 2018;286:800-809
6. Weikert T, Cyriac J, Yang S, Nesic I, Parmar V, Stieltjes B. A practical guide to artificial intelligence-based im-

age analysis in radiology. *Invest Radiol* 2020;55:1-7
7. Lloyd, K. Bias amplification in artificial intelligence systems. *ArXiv preprint* 2018;arXiv:1809.07842
8. Yu AC, Eng J. One algorithm may not fit all: how selection bias affects machine learning performance. *Radiographics* 2020;40:1932-1937
9. Song TJ, Fong Y, Cho SJ, Gönen M, Hezel M, Tuorto S, et al. Comparison of hepatocellular carcinoma in American and Asian patients by tissue array analysis. *J Surg Oncol* 2012;106:84-88
10. Pasquinelli MM, Kovitz KL, Koshy M, Menchaca MG, Liu L, Winn R, et al. Outcomes from a minority-based lung cancer screening program vs the National Lung Screening Trial. *JAMA Oncol* 2018;4:1291-1293
11. Jaeger S, Candemir S, Antani S, Wáng YX, Lu PX, Thoma G. Two public chest X-ray datasets for computer-aided screening of pulmonary diseases. *Quant Imaging Med Surg* 2014;4:475-477
12. Pons E, Braun LM, Hunink MG, Kors JA. Natural language processing in radiology: a systematic review. *Radiology* 2016;279:329-343
13. Gera S, Ettel M, Acosta-Gonzalez G, Xu R. Clinical features, histology, and histogenesis of combined hepatocellular-cholangiocarcinoma. *World J Hepatol* 2017;9:300-309
14. Fasel JH, Selle D, Evertsz CJ, Terrier F, Peitgen HO, Gailloud P. Segmental anatomy of the liver: poor correlation with CT. *Radiology* 1998;206:151-156
15. FDA. Artificial intelligence and machine learning in software as a medical device. Available at. https://www.fda.gov/medical-devices/software-medical-device-samd/artificial-intelligence-and-machine-learning-software-medical-device. Accessed Sep 24, 2020
16. Azer SA. Deep learning with convolutional neural networks for identification of liver masses and hepatocellular carcinoma: a systematic review. *World J Gastrointest Oncol* 2019;11:1218-1230
17. Wang W, Iwamoto Y, Han X, Chen YW, Chen Q, Liang D, et al. Classification of focal liver lesions using deep learning with fine-tuning. Proceedings of the 2018 International Conference on Digital Medicine and Image Processing; 2018 Nov; Okinawa, Japan: Association for Computing Machinery; 2018:56-60
18. Liang D, Lin L, Hu H, Zhang Q, Chen Q, Han X, et al. *Combining convolutional and recurrent neural networks for classification of focal liver lesions in multi-phase CT images. International Conference on Medical Image Computing and Computer-Assisted Intervention.* Cham: Springer 2018:666-675
19. Das A, Acharya UR, Panda SS, Sabut S. Deep learning based liver cancer detection using watershed transform and Gaussian mixture model techniques. *Cogn Syst Res* 2019;54:165-175
20. Gletsos M, Mougiakakou SG, Matsopoulos GK, Nikita KS, Nikita AS, Kelekis D. A computer-aided diagnostic system to characterize CT focal liver lesions: design and optimization of a neural network classifier. *IEEE Trans Inf Technol Biomed* 2003;7:153-162
21. Cao SE, Zhang LQ, Kuang SC, Shi WQ, Hu B, Xie SD, et al. Multiphase convolutional dense network for the classification of focal liver lesions on dynamic contrast-enhanced computed tomography. *World J Gastroenterol* 2020;26:3660-3672
22. Xu Y, Lin L, Hu H, Wang D, Zhu W, Wang J, et al. Texture-specific bag of visual words model and spatial cone matching-based method for the retrieval of focal liver lesions using multiphase contrast-enhanced CT images. *Int J Comput Assist Radiol Surg* 2018;13:151-164
23. Mougiakakou SG, Valavanis IK, Nikita A, Nikita KS. Differential diagnosis of CT focal liver lesions using texture features, feature selection and ensemble driven classifiers. *Artif Intell Med* 2007;41:25-37
24. Nayantara PV, Kamath S, Manjunath KN, Rajagopal KV. Computer-aided diagnosis of liver lesions using CT images: a systematic review. *Comput Biol Med* 2020;127:104035
25. Ben-Cohen A, Klang E, Diamant I, Rozendorn N, Raskin SP, Konen E, et al. CT image-based decision support system for categorization of liver metastases into primary cancer sites: initial results. *Acad Radiol* 2017;24:1501-1509
26. Yasaka K, Akai H, Abe O, Kiryu S. Deep learning with convolutional neural network for differentiation of liver masses at dynamic contrast-enhanced CT: a preliminary study. *Radiology* 2018;286:887-896
27. Park HJ, Park B, Lee SS. Radiomics and deep learning: hepatic applications. *Korean J Radiol* 2020;21:387-401

대한영상의학회지

인공지능 기반 임상의학 결정 지원 시스템 의료기기의 성능 및 안전성 검증을 위한 간 종양 표준 데이터셋 구축

김승섭[1] · 이동호[2] · 이민우[3] · 김소연[4] · 신재승[1] · 최진영[1]* · 최병욱[1]

목적 간 종양의 조영증강 컴퓨터단층촬영(이하 CT) 영상에 관한 인공지능 알고리즘의 성능과 안전성을 검증할 수 있는 표준 테스팅 데이터셋을 구축하고자 하였다.

대상과 방법 국내 4개 3차 의료기관의 복부 영상의학 전문가 4인이 모여 간 종양 진단 알고리즘의 성능과 안전성을 검증하기 위해 표준 데이터셋이 갖춰야 할 조건을 논의하였다. 각 기관마다 간세포암 75예, 전이암 75예, 그리고 양성 병변 30–50예씩 수집하여, 총 783명 환자의 CT 영상을 대상으로 하였다. 간세포암과 전이암의 경우 병리학적으로 확진된 경우만을 대상으로 하였다. 각 기관의 복부 영상의학 전문가들이 직접 환자의 임상정보를 추출하고 CT 영상에 관한 데이터 라벨링(labeling)을 수기로 시행하였다. CT 영상은 의료용 디지털 영상 및 통신(Digital Imaging and Communications in Medicine, DICOM) 파일로 저장하였다.

결과 복부 영상의학 전문가들이 수기 데이터 라벨링을 시행한 총 783 증례의 간 종양 조영증강 CT의 표준 데이터셋을 구축하였다. 알고리즘의 성능 및 안전성은 병변의 발견 여부 및 특성화의 정확도에 대해 민감도와 특이도를 계산하여 평가할 수 있다.

결론 본 연구에서 구축한 간 종양 조영증강 CT 영상의 표준 데이터셋은 임상의학 결정 지원 시스템을 위한 기계학습 기반 인공지능 알고리즘을 평가하는 데에 활용될 수 있다.

[1]연세대학교 의과대학 세브란스병원 영상의학과, 방사선의과학연구소,
[2]서울대학교 의과대학 서울대학교병원 영상의학과,
[3]성균관대학교 의과대학 삼성서울병원 영상의학과,
[4]울산대학교 의과대학 서울아산병원 영상의학과