

웹기반 의료영상 표준 데이터셋 변환 및 관리 시스템 구축

김지언*, 임동욱*, 유영주*, 노시형*, 이충섭*, 김태훈*, 정창원*

*원광대학교 의료융합연구센터

e-mail:{kakasky112, dw316, yeriel.ryu, nosij123, cslee99, tae_hoonkim, mediblu}@wku.ac.kr

Construction of Web-Based Medical Image Standard Dataset Conversion and Management System

Ji-Eon Kim*, Dong Wook Lim*, Yeong Ju Yu*, Si-Hyeong Noh*, ChungSub Lee*, Tae-Hoon Kim*, Chang-Won Jeong**

*Medical Convergence Research Center, Wonkwang University

요 약

최근 4차 산업혁명으로 의료빅데이터 기반으로 한 AI 기술이 급속도로 발전하고 있다. 특히, 의료영상을 기반으로 병변을 탐색, 분할 및 정량화 그리고 자동진단 및 예측 관련된 기술이 AI 제품으로 출시되고 있다. AI 기술개발은 많은 학습데이터가 요구되며, 임상검증에 단일기관에서 2개 이상 기관의 검증이 요구되고 있다. 그러나 아직까지도 단일기관에서 학습용 데이터와 테스트, 검증용 데이터를 달리하여 기술개발에 활용하고 있다. 본 논문은 AI 기술개발에 필요한 영상데이터에 대한 표준화된 데이터셋 변환 및 관리를 위한 시스템에 대해 기술한다. 다기관 데이터를 수집하기 위해서는 각 기관의 의료영상 데이터 수집 및 저장하는 기준이 명확하지 않아 표준화 작업이 필요하다. 제안한 시스템은 기관 또는 다기관 연구 그룹의 의료영상데이터를 표준화하여 저장할 수 있을 뿐만 아니라 의료영상 뷰어 및 의료영상 리스트를 통해 연구자가 원하는 의료영상 데이터 셋을 검색하여 다양한 데이터셋으로 제공할 수 있기 때문에 수집 및 변환 그리고 관리까지 지원할 수 있는 시스템으로 영상기반의 머신러닝 연구에 활력을 불어넣을 수 있을 것으로 기대하고 있다.

1. 서론

국내의 인공지능 기술개발의 대부분이 의료영상 분석 분야에 중점을 두고 있다[1,2,3]. 하지만, 의료영상 데이터 기반의 임상연구와 더불어 인공지능 기술 개발은 많은 학습 데이터가 요구되며 부족한 데이터를 다기관을 통해 수집하더라도 표준화 작업에 많은 시간이 소요된다. 최근 학습 데이터 부족 문제를 해결하기 위해 GAN 기술 중심으로 합성의료 데이터를 생성하여 활용하고 있는 추세이다. 그러나 인공지능에 적용하기 위한 학습 데이터 구축을 위한 원데이터 수집에 관련된 수집과 관리하는 표준화된 방법이 부재한 실정이다[4]. 또한, AI 기술개발에 대한 검증 또한 엄격하게 진행되고 있어 단일기관 검증에서 2개이상의 기관에서 검증하여 신뢰성을 높이

기 위한 노력이 요구되고 있다. 그러나 매번 유사 영역의 기술개발을 위해 또 다른 데이터를 수집하고 선별하고 전처리하는 등의 과정을 거쳐야 한다. 더욱이 연구수행 종료후 관련 데이터는 더 이상 관리되지 않는다. 따라서, 본 논문은 이와 같은 문제점을 해결하기 위해서 의료영상 데이터 관리와 기존 OMOP-CDM과 연계하여 표준화된 데이터셋 제공뿐만 아니라 임상연구에 필요한 의료정보를 추출할 수 있는 웹기반 시스템을 제안한다. 이를 위해 DICOM 기반의 ETL 정의서를 개발하여 수집된 영상데이터에 대한 표준화된 관리와 다기관 공동 AI 연구에 필요한 다양한 포맷과 요청한 데이터셋을 제공하기 시스템을 제안한다.

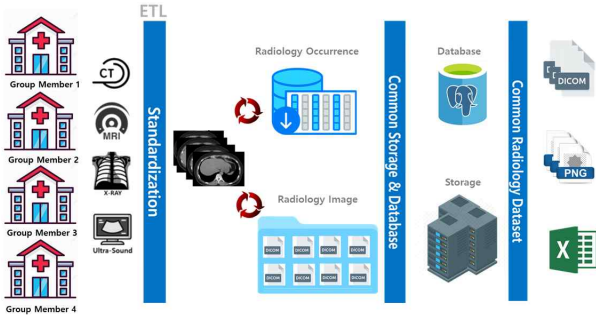
2. 제안한 시스템

2.1 다기관 데이터 수집 및 관리 방법

본 논문에서 제안하는 다기관 데이터를 수집 및 관리 방법은 그림 1과 같다. 의료영상 데이터는 X-ray, CT, MRI, US등의 다양한 의료영상진단기기

본 연구는 보건복지부의 지원으로 한국보건산업진흥원의 보건의료기술 연구개발사업(HI18C1216) 그리고 과학정보통신부의 지원으로 한국연구재단(NRF-2016M3A9A7918501)(NRF-2020R1H1A1A01074256) 지원에 의하여 이루어진 것임.

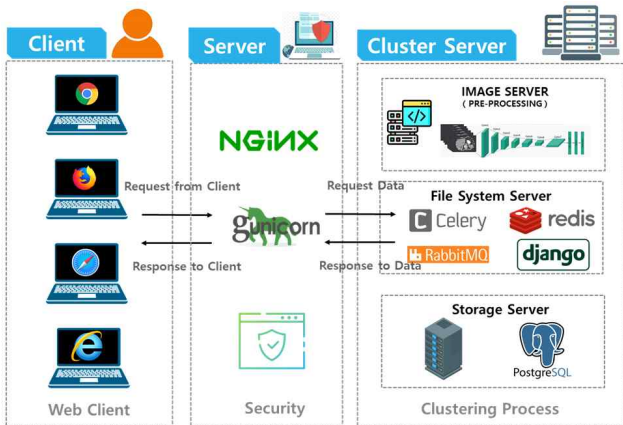
로부터 생성된다. 그러나, 다기관 의 경우 서로 상이한 조건에 따라 의료영상 데이터가 생성되고, 서로 다른 메타정보를 저장하여 관리하기 때문에 표준화 처리과정이 필요하다. 본 논문에서 제안하는 웹 기반 환경에 데이터를 업로드 하게 되면 ETL정의에 따라서 의료영상 데이터에 저장된 이미지와 헤더정보를 표준 데이터셋으로 변환하여 저장한다.



(그림 1) Multi-Center Data Management Method

2.2 시스템 소프트웨어 아키텍처

본 논문에서 제안하는 웹 기반의 표준 데이터셋 변환 시스템은 그림 2와 같다. 웹을 기반으로 구성되어 있어 다양한 웹브라우저 클라이언트로 의료영상 데이터를 업로드 할 수 있다. 또한, 소규모 및 대용량 의료영상 데이터에 대한 트래픽을 안정적으로 처리할 수 있도록 서버 구성이 하였을 뿐만 아니라 데이터 유출에 대한 민감성 문제를 해결하기 위해 HL7(Health Level 7)에 기준에 의한 의료데이터에 대한 보안 요소 까지 설계하였다.



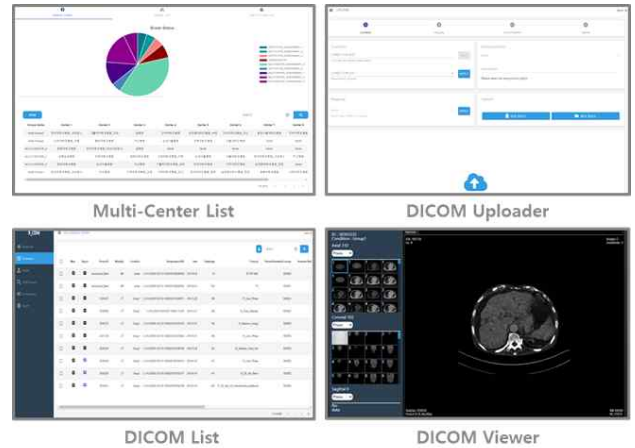
(그림 2) System Architecture

이와 관련하여, 의료영상 데이터는 표준 데이터셋으로 변환하기 위해서는 Cluster Server의 전처리 과정을 통해 표준 데이터셋으로 변환되어 저장 및 관리된다.

3. 시스템 구현 결과

3.1 Web 구현 결과

웹 기반의 시스템 구현 결과는 그림 3과 같다. 개발한 내용 가운데 주요 기능으로서 Multi-Center List에서는 기관 별로 다기관 연구 그룹의 생성리스트와 의료영상 업로드 결과를 확인할 수 있다. 또한 DICOM List를 통해 업로드 결과를 확인할 수 있을 뿐만 아니라 DICOM Viewer를 통해 DICOM Image 정보까지 확인할 수 있다.



(그림 3) System Results

3.2 ETL 변환 결과

본 논문은 ETL에 대한 신뢰성과 제안한 시스템 설계에 대한 평가를 위해 그림 4와 같이 다기관 데이터를 수집하여 표준 데이터셋으로 변환하는 테스트를 진행하였다. 다기관에 해당되는 의료정보 데이터는 보건복지부 지정의 상급대학병원 기관을 대상으로 다기관 연구 그룹을 생성 및 데이터를 수집하여 업로드하여 표준 데이터셋으로 변환하였다.

MULTICENTER USER				
User Name	email	Institution	Multi-center group name	Grade
MULTICENTER_1 USER LIST				
msh001	msh001@multi.com	한양대학교병원	MI	
mka001	mka001@multi.com	강남대학교병원	MI	
mks001	mks001@multi.com	원광대학교병원	MI	
mko001	mko001@multi.com	전북대학교병원	MI	
mky001	mky001@multi.com	경희대학교병원	MI	
mgj001	mgj001@multi.com	한림대학교병원	MI	
mss001	mss001@multi.com	삼성서울병원	MI	
MULTICENTER_2 USER LIST				
mhy002	mhy002@multi.com			
mgf002	mgf002@multi.com			
myy002	myy002@multi.com			
MULTICENTER_3 USER LIST				
mgj002	mgj002@multi.com			
mcp001	mcp001@multi.com			
mca001	mca001@multi.com			
mka001	mka001@multi.com			
Multi-Group1 USER LIST				
mkt001	mkt001@multi.com			
myy001	myy001@multi.com			
mva001	mva001@multi.com			
mas002	mas002@multi.com			
msd002	msd002@multi.com			
Multi-Group2 USER LIST				
msj001	msj001@multi.com			
msu001	msu001@multi.com			
mss002	mss002@multi.com			
Multi-Group3 USER LIST				
mgf001	mgf001@multi.com	강병원	Multi-Group3	MULTICENTER_SUBSCRIBER_L3
mbd001	mbd001@multi.com	봉성서울대학교병원	Multi-Group3	MULTICENTER_SUBSCRIBER_L2
mco001	mco001@multi.com	공인대학교병원	Multi-Group3	MULTICENTER_SUBSCRIBER_L2
mho001	mho001@multi.com	간희대학교병원	Multi-Group3	MULTICENTER_SUBSCRIBER_L2
mic001	mic001@multi.com	가톨릭대학교병원, 인천	Multi-Group3	MULTICENTER_SUBSCRIBER_L2
user count				5
total user count				27

(그림 4) Multi-Center Data Upload Test

표준 데이터셋 변환에 대한 결과를 분석하기 위해

의료영상데이터에 저장되어 있는 데이터를 ETL정의서에 따라 변환 정보를 미리 정의하여 제안한 시스템을 통해 변환된 결과 값을 비교 및 분석하였다. 검색기준은 StudyInstanceUID를 기준으로 업로드한 각 기관에 해당하는 의료영상 데이터를 검색하였다. 검색한 결과를 검토하였을 때 의료영상 이미지와 헤더정보는 정상적으로 검색이 가능하였을 뿐만 아니라 미리 정의해둔 변환 결과에 따라 표준 데이터셋으로 변환되어진 결과를 확인할 수 있었다.

4. 결론

최근 4차 산업 혁명으로 의료분야에서는 빅데이터 및 인공지능과 같은 기술을 임상 연구에 적용하는 연구가 활발히 진행되고 있다. 하지만, 빅데이터와 인공지능에 적용하기에는 많은 연구데이터가 필요하며, 다기관 데이터를 함께 수집하더라도 서로 다른 기준으로 저장되어 있기 때문에 표준화 방법이 모호한 실정이었다. 따라서, 본 논문은 의료영상 데이터를 표준화하기 위한 ETL정의서를 개발하여 다양한 기관에서 의료영상 데이터를 수집하더라도 동일한 조건으로 의료영상 데이터를 표준화 할 수 있도록 개발하였을 뿐만 아니라 다양한 의료영상 기반의 임상연구 지원을 위해 웹 기술에 적용하여 의료영상 데이터를 수집 및 관리 그리고 다양한 데이터 타입으로 변환된 데이터 셋을 제공할 수 있도록 개발하였다. 향후, 표준화된 데이터셋을 기반으로 인공지능 기술과 같은 4차 산업 기술 연구에 적용하여 실제 임상연구에 지원할 수 있도록 평가하여 신뢰성 있는 표준화 기준이 될 수 있도록 연구를 진행할 예정이다.

참고문헌

- [1] 이고은, 서종현, 우동철, 오우용, 이종구, 정승채, & 김경원, “다기관 임상영상 관리 시스템: 다학제적 너연구의 효율적 수행을 위한 중요 성공 요소”, 대한의학영상정보학회지, 2018, 24(1), 11-19 .
- [2] 신수용. “비정형 헬스케어 데이터 표준화.” 한국통신학회지 (정보와통신) 35.2 (2018): 58-64.
- [3] 이승룡; 신수용. 의료 빅데이터 분석, 의료 빅데이터 표준화, 의료 온톨로지, 의료 보안, 상호호환성, CDSS, 정밀의료, 지능형 의료 플랫폼. 한국통신학회지 (정보와통신), 2018, 35.2: 2-2.
- [4] 이충섭, 김승진, 김지언, 노시형, 김태훈, 윤권하, & 정창원. (2020). 다기관 임상연구를 위한 의료 데이터 셋 관리 시스템. 한국정보처리학회 학술대회논문집, 27(1), 16-19.