테크니컬 리포트

2020년 1차

인공지능

학습용

데이터 구축

헬스케어 영역

질병진단(유방암 조직) 이미지

개요: 질병진단 (유방암 조직) 이미지 데이터셋이란?

전 세계적으로 지속적으로 증가하고 있는 유방암 질병진단을 위해 활용할 수 있는 학습 데이터셋으로 국립암센터에서 구축하였으며, 암환자 4,000명 대상으로 10만건의 유방암 조직(pathology)이미지와 4,000건의 메타데이터로 구성되어 있다.

질병진단 기술은 진단 보조 서비스에 주로 사용되며, 유방암 조직 이미지를 통해 유방암 양성, 악성을 판단하는 기술로, 방대한 유방암 조직 이미지에 대한 유방암 판독 및 유방암 조직의 병변 segmentation 등에 활용하기 위해 활발한 연구를 진행하고 있다.

데이터셋의 구성

본 데이터셋은 양성 1,000case, 악성 3,000case 총 4,000case의 유방암 조직 이미지로 한 case당 25장의 이미지를 추출하여 총 10만장의 PNG 이미지와 각 case에 해당하는 정보, 예를 들어 연령대, 검사일시, 검사장비 제조사, 검사장비 모델명, 양성 악성 여부, 종양의 병기 등의 데이터로 구성되어 있다.

기존의 유방암 관련 공개 데이터셋은 비교적 적은 샘플 수 또는 임상데이터 없이 영상이미지만 공개하고 있어 인공지능 학습용 데이터셋으로 활용하기에는 한계점이 있으며, 4,000case에 대한 10만장의 유방암 조직 이미지와 메타데이터, 어노테이션 정보는 다양한 인공지능학습용 모델을 구축하는데 활용가능하다.

데이터 종류	포함 내용	제공 방식
유방암 양성(Benign)환자 데이터셋	유방암 조직 이미지	PNG 포맷 파일
	메타데이터	EXCEL 포맷 파일
유방암 악성(Malignant)환자 데이터셋	유방암 조직 이미지	PNG 포맷 파일
	메타데이터	EXCEL 포맷 파일

데이터셋의 설계 기준과 분포

데이터셋을 설계할 때 가장 중요하게 고려했던 점은 데이터 밸런스이다. 유방암 조직 이미지의 분류를 양성(Benign), 악성(Malignant)으로 구성할 때 데이터의 편향성 최소화와 정확성을 위해 아래와 같이 설계하였다.

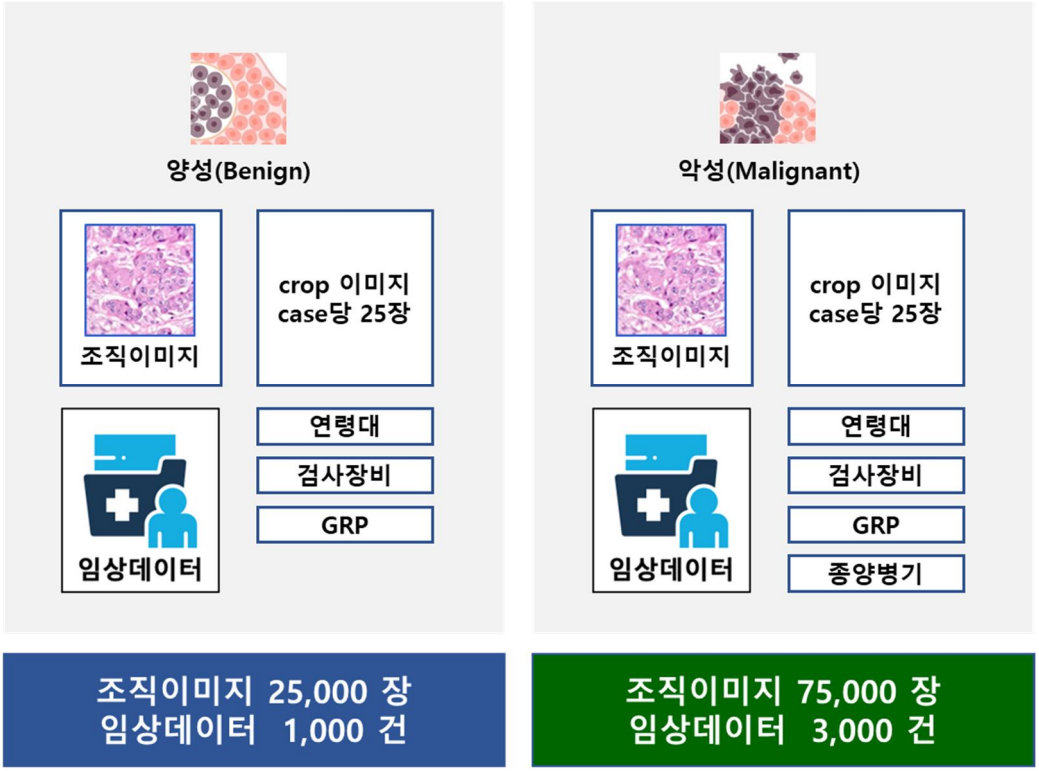

그림 1 데이터셋 구성 개요

양성, 악성의 특징은 다음과 같다.

- 양성 : 전체 100,000장의 조직이미지 중 25,000장이며 특정 병변 영역에 대한 설정 없이 이미지가 추출되며 양성인 관계로 종양병기 데이터는 없다.
- 악성 : 전체 100,000장의 조직이미지 중 75,000 장이며 악성병변에 해당하는 특정 영역에 대해 이미지가 추출된 것으로 해당 case에 대한 임상데이터에는 종양병기 정보를 포함한다.

데이터 구조

- 데이터셋에 따른 항목과 해당 값은 아래 테이블과 같다.

분류	양성	악성

대상자수		1,000 case	3,000 case
이미지 PNG 파일		25,000 장	75,000 장
항목		포함여부	
익명화 환자 ID	char	Y	Y
연령대	char	Y	Y
검사일시	date	Y	Y
검사장비 제조사	char	Y	Y
검사장비 모델명	char	Y	Y
GRP	char	Y	Y
종양의 병기	char		Y

데이터 예시

본 데이터는 유방암 악성 조직 이미지로, 비식별화 과정을 거친 후 512x512 사이즈로 crop 처리하여 여러 개의 tile이미지를 생성하였다. 이 후 인공지능 학습에 적합한 25개의 이미지를 선정하였다

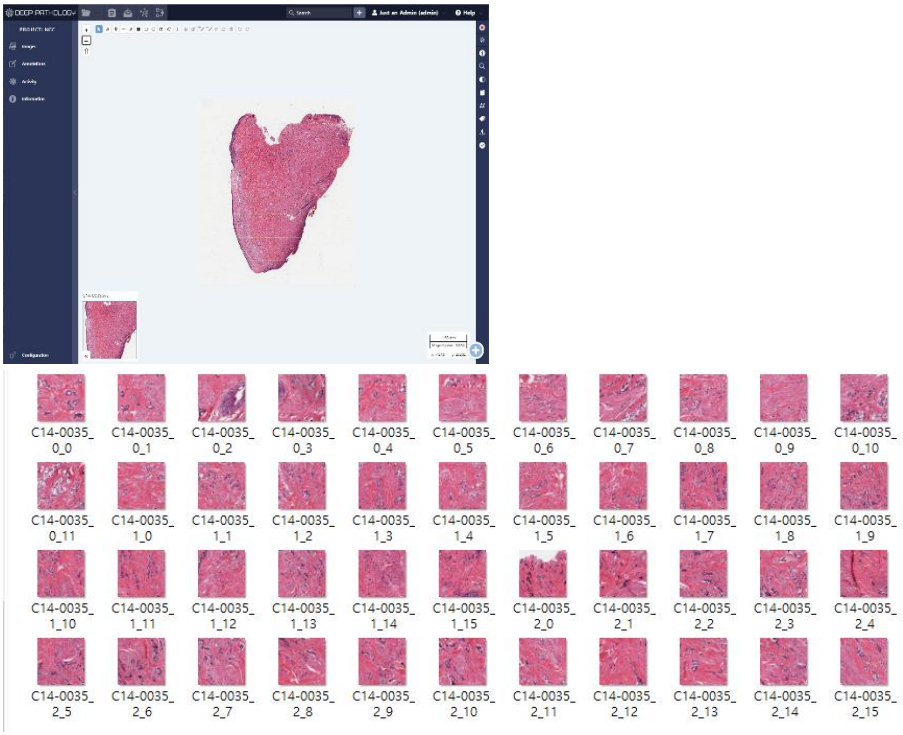

그림 2 유방암 조직 이미지 (좌) Whole Slide Image, (우) Crop 이미지

임상데이터 예시는 아래와 같다.

ID	익명화 환자 ID	연령대	검사일시	검사장비 제조사	검사장비 모델명	GRP	종양의 병기
1	SN110000049	30	1999-01-01	3DHistech	Panoramic1000	1	3

데이터 구축 과정

환자의 개인정보에 대한 문제가 없도록 하기 위해서 기관생명윤리위원회 (Institutional Review Board, IRB)의 승인을 취득하고, 임상연구검색 포털 시스템을 활용하여 2011.01.01 ~ 2018.12.31 까지 만19세 이상 성인 여성중 유방암 조직 검사를 수행한 환자로 대상군을 최종 확정하였다. 수집된 데이터는 조직 이미지와 임상데이터로 구분되는데, 조직 이미지의 경우 슬라이드를 스캐너 장비를 통해 스토리지 서버에 저장된 이미지로부터 획득하였다. 그리고 해당하는 환자의 임상데이터는 CRDW에서 수집하였다. 이때, 개인 식별화가 가능한 정보를 삭제하여 비식별화 하였고, 조직 이미지는 양성의 경우 그대로 crop하여 이미지를 추출하였고 악성의 경우 의료전문가가 전체 이미지에서 악성 병변에 해당하는 영역을 지정하였다. 지정한 악성 병변이 존재하는 범위를 일정한 크기로 잘라 인공지능 학습에 유의미한 데이터를 추출하고 검수하였으며 인공지능학습의 편의성을 위해 PNG파일로 저장하였다

그림 3 유방암 조직 이미지 획득 과정

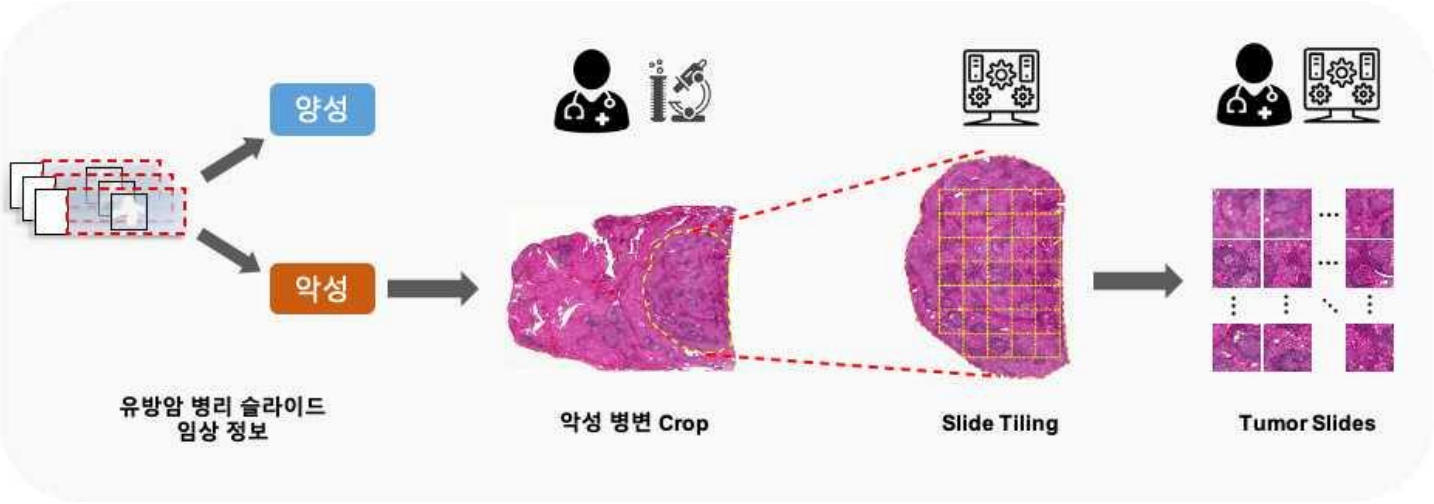

그림 1 인공지능 학습용 데이터 구축 과정

- 어노테이션 저작도구 개발

  - 어노테이션툴은 바이오 이미지 분석 소프트웨어로 디지털로 변환된 전체 슬라이드 이미지를 보여준다. 호환되는 디지털 이미지 종류의 범위를 넓히고 빠르고 자연스럽게 사용자들이 볼 수 있는 뷰어를 제공하며 직관적인 레이블링 도구와 수정 및 추출 방법으로 편리하게 사용할 수 있도록 개발하였다. 또한 병변의 다양한 형태를 고려하여 polygon 형태로 어노테이션을 붙일 수 있는 레이블링 방법을 구현하였다.

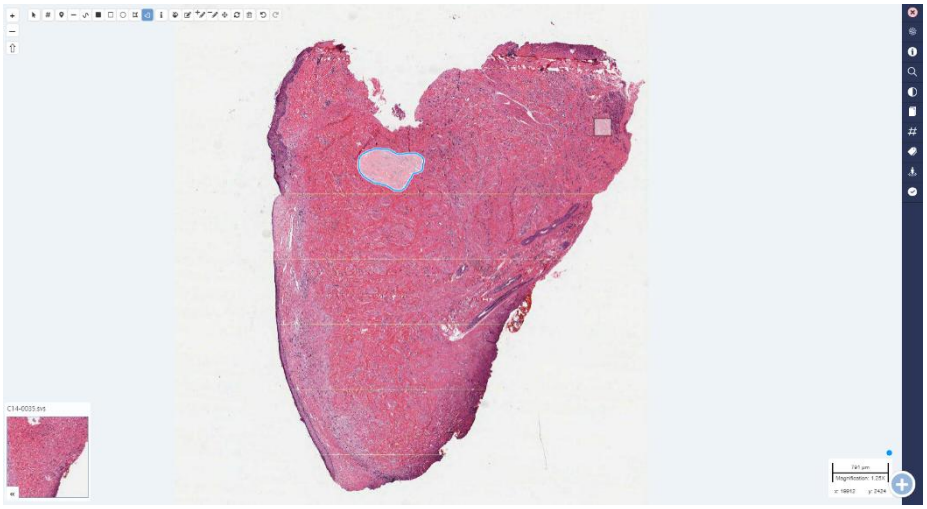

그림 5 어노테이션 수행 프로그램(polygon, rectangle 형태로 영역설정)

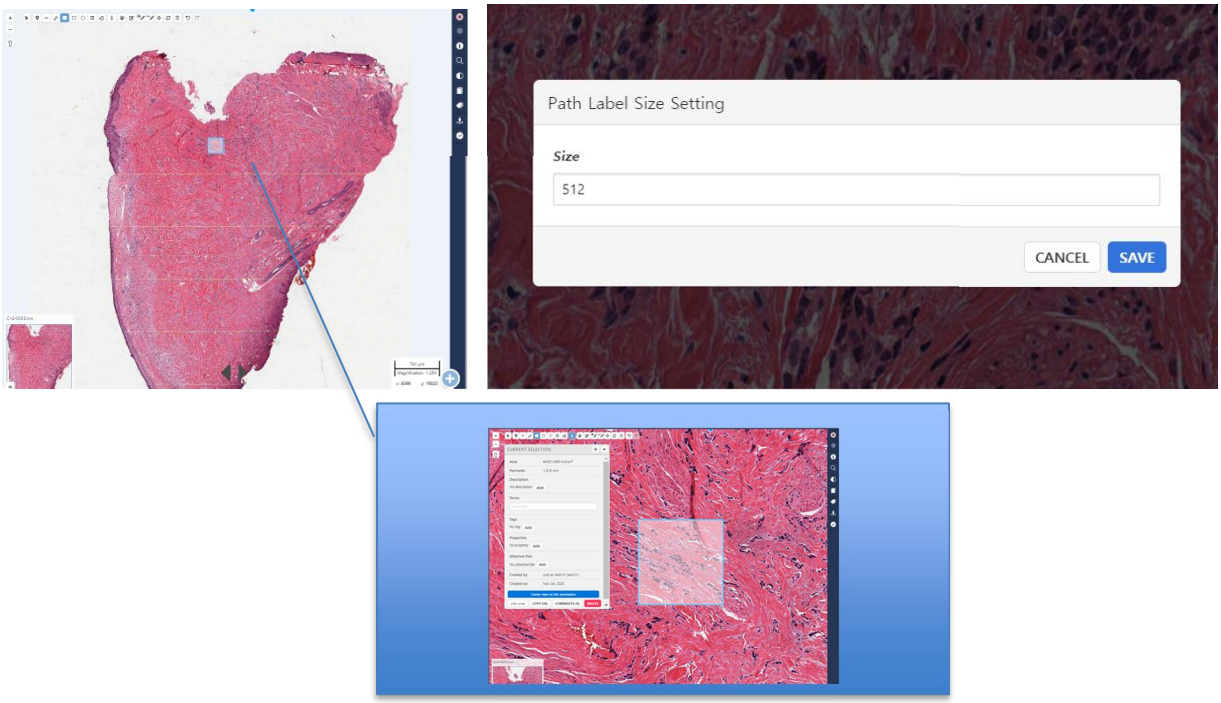

그림 6 어노테이션 크기(512x512) 사전 설정 후 위치선정 기능

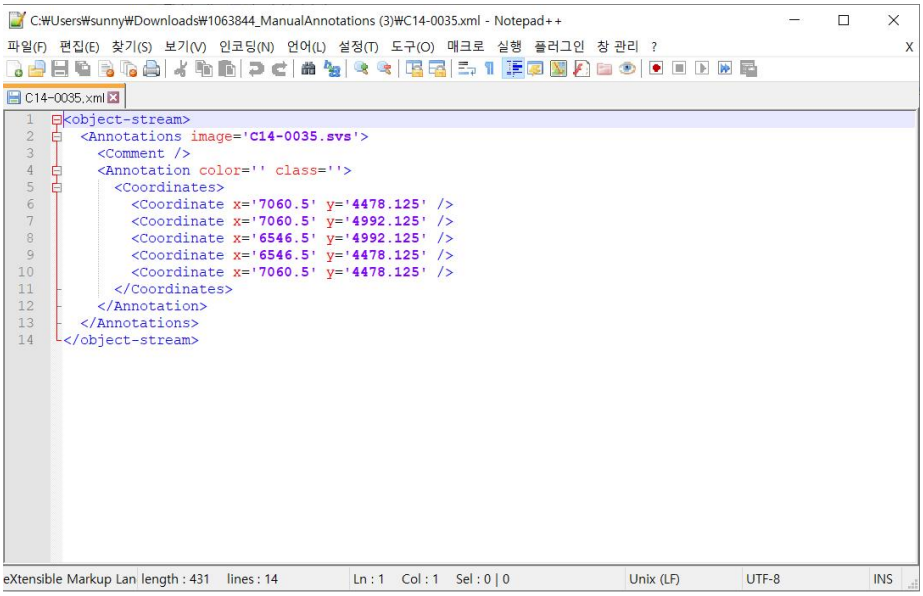

그림7 어노테이션 위치 정보(xml 포맷 파일)

검수와 품질 확보

구축한 유방암 조직 이미지, 악성병변 어노테이션 정보, 임상정보 등을 질병진단 학습용 데이터로 활용하기 위해서 검수 프로세스를 정립하였다. 이때 인공지능학습에 부적합한 이미지가 있을 경우 해당 이미지를 삭제하였으며, 이 후 해당 case당 이미지의 수가 25장에 미달하는 경우 그 case는 제외 처리하고, 새로운 case에 대한 데이터 확보를 재진행하여 신뢰도가 높은 데이터를 확보하였다.

그림 8 구축한 유방암 조직 이미지에서 인공지능학습에 부적합한 경우 삭제

데이터 구축 담당자

수행기관(주관) : ㈜국립암센터(전화: 031-920-0572), 이메일: healthcare_ai@ncc.re.kr