개요: 질병진단 이미지(유방암) 데이터셋이란?

전 세계적으로 지속적으로 증가하고 있는 유방암 질병진단을 위해서 활용할 수 있는 학습데이터셋으로 국립암센터에서 구축했으며, 7,500명 대상으로 3만건의 유방촬영술(Mammography) 이미지와 7,500건의 임상데이터, 1,500건의 악성병변 어노테이션 정보로 구성되어 있다.

질병진단 기술은 진단 보조 서비스에 주로 사용되며, 유방촬영술 영상 이미지를 통해 유방암 정상, 악성을 판단하는 기술로, 방대한 유방촬영영상 이미지에 대한 유방암 판독 및 유방 악성 병변 segmentation 등에 활용하기 위해 활발한 연구 및 상용화를 진행하고 있다.

데이터셋의 구성

본 데이터셋은 정상, 양성, 악성 7,500명의 유방촬영영상 4-view (RCC, RMLO, LCC, LMLO) paired 유방촬영영상으로 한 환자당 4개의 영상이미지 30,000장과 각 환자별 임상데이터 7,500건, 그리고 전체데이터에서 악성 환자에서만 악성 병변의 어노테이션 정보 1,500건이 추가로 구성되어 있다.

기존의 유방암 관련 공개 데이터셋은 비교적 적은 샘플 수 또는 임상데이터가 없이 영상이미지만 공개하고 있어 인공지능 학습용 데이터셋으로 활용하기에는 한계점이 있으며, 7,500명의 유방촬영영상 이미지와 임상데이터, 악성 병변의 어노테이션 정보는 다양한 인공지능 학습용 모델 구축하는데 활용 가능하다.

데이터 종류	포함 내용	제공 방식
유방암 악성환자 데이터셋	유방촬영술 영상 이미지	DCM 포맷 파일
	악성병변 어노테이션	JSON 포맷 파일
	임상데이터	EXCEL 포맷 파일
유방암 양성환자 데이터셋	유방촬영술 영상 이미지	DCM 포맷 파일
	임상데이터	EXCEL 포맷 파일
유방암 정상 데이터셋	유방촬영술 영상 이미지	DCM 포맷 파일

	임상데이터	EXCEL 포맷 파일

데이터셋의 설계 기준과 분포

데이터셋을 설계할 때 가장 중요하게 고려했던 점은 데이터 밸런스이다. 유방암 질환 분류를 정상, 양성, 악성으로 구성할 때, 데이터의 편향성 최소화와 정확성을 위해 아래와 같이 설계하였다.

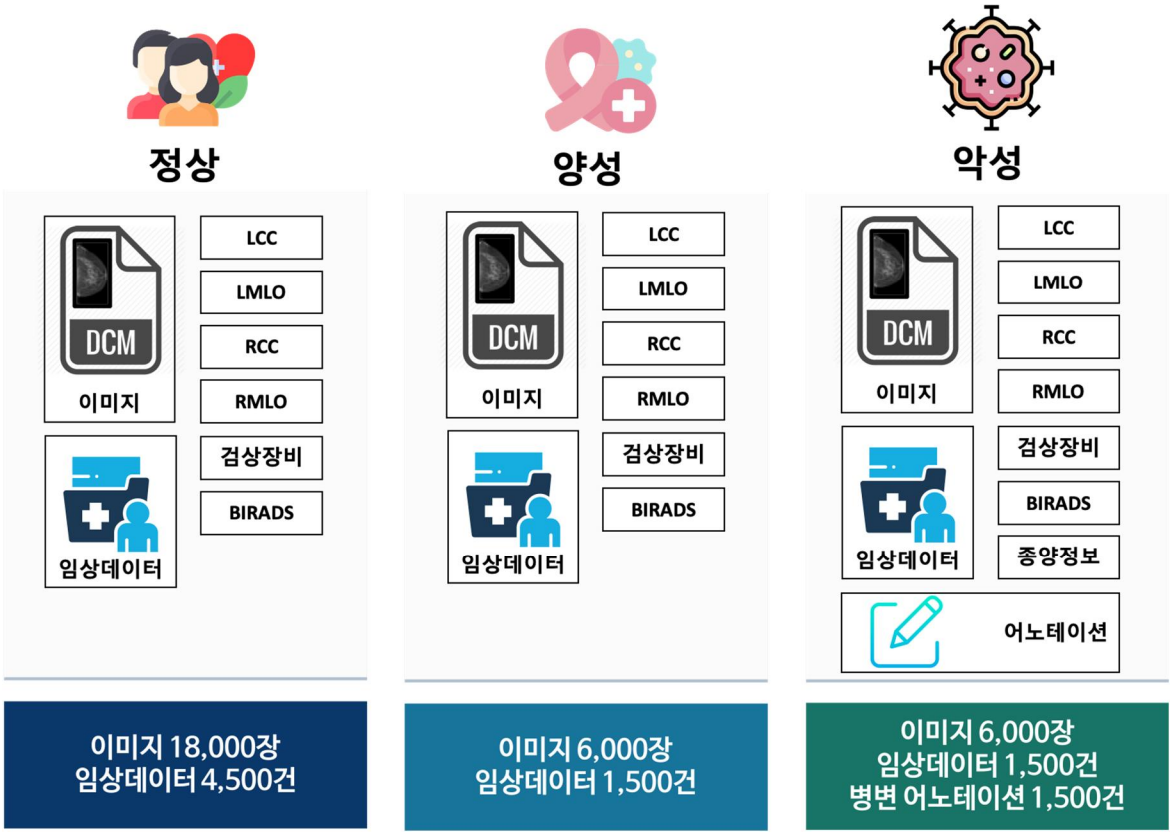

그림 1 데이터셋 구성 개요

정상, 양성, 악성 판정의 선정 기준은 다음과 같다.

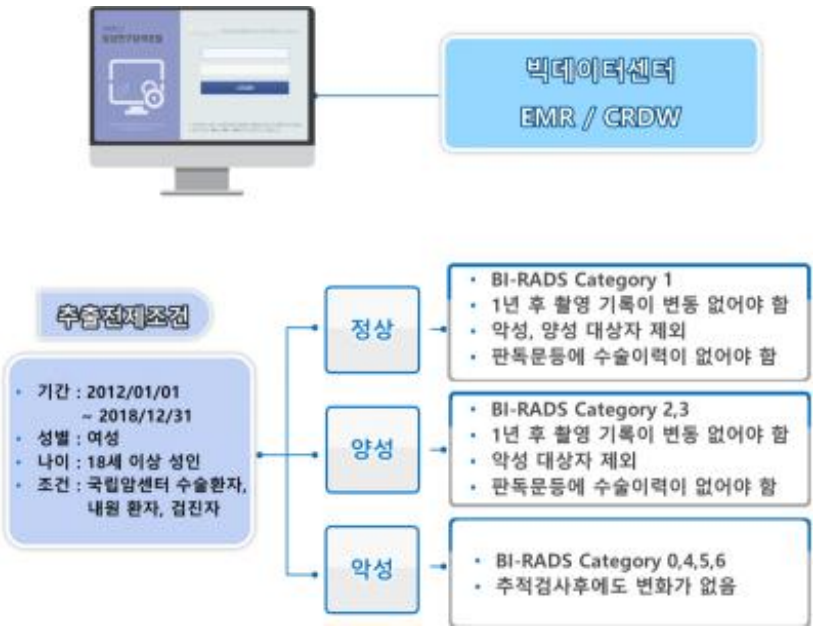

그림 2 데이터 설계 기준

데이터 구조

데이터셋에 따른 항목과 해당 값은 아래 테이블과 같다.

분류		정상	양성	악성
대상자수		4,500명	1,500명	1,500명
이미지 DICOM 파일		18,000장	6,000장	6,000장
어노테이션 JSON 파일		-	-	1,500건
임상데이터		4,500건	1,500건	1,500건
임상 데이터 항목		포함여부		
검사장비	varchar(50)	Y	Y	Y
검사장비모델	varchar(50)	Y	Y	Y
BIRADS category (양쪽)	varchar(15)	Y	Y	Y
BIRADS category (좌)	varchar(15)	Y	Y	Y

BIRADS category (우)	varchar(15)	Y	Y	Y
악성 종양의 위치	text			Y
T stage	int			Y
N stage	int			Y
악성 종양의 크기	int			Y
조직형 진단	varchar(50)			Y

어노테이션	
User_ID	어노테이션 판독의 식별용 ID
Case_ID	유방촬영영상 이미지 ID
Contour_list	각 view (rcc, rmlo, lcc, lmlo) 에 그려지는 lesion들 (lesion01, 02, ...)을 list of coordinates 형태로 저장
Discard_yn	해당 영상에 문제가 있어서 사용하지 못하게 될 케이스임을 표시하는 용도

데이터 예시

이 데이터는 악성 데이터 기준이며, 정상, 양성 데이터셋은 아래 예시에서 어노테이션 json가 없으며, 임상데이터에서 종양의 정보는 없는 구조이다.

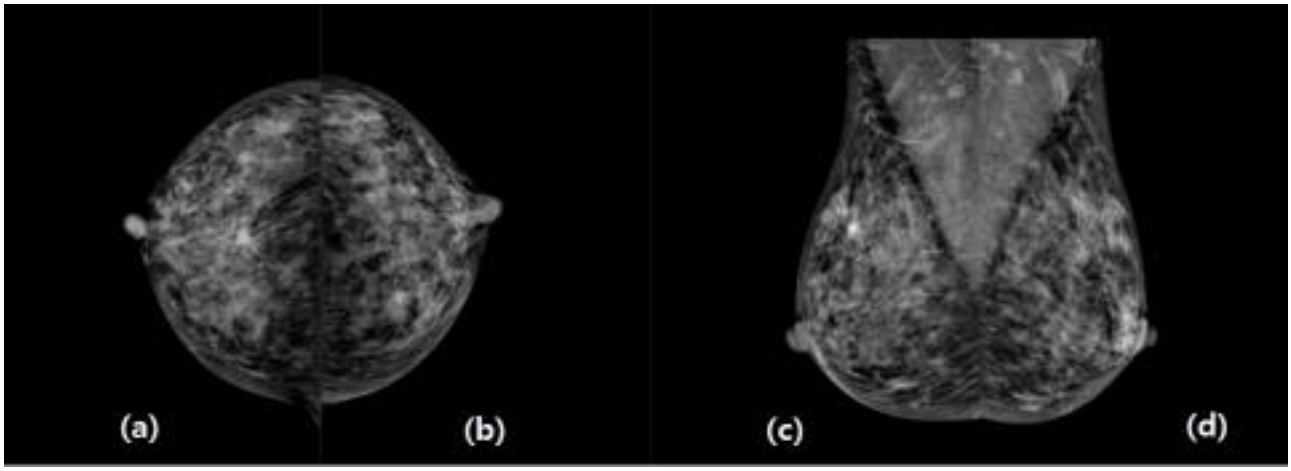

그림 3 악성 유방촬영영상 (a) LCC (b) RCC (c) LMLO (d) RMLO

ID	악성 종양의 위치	BIRADS category (양쪽)	BIRADS category (좌)	BIRADS category (우)	T stage	N stage	조직형 진단	검사장비	검사장비 모델	악성 종양의 크기 (cm)
1	Breast, left		5		1c	0	Invasive ductal carcinoma	GE MEDICAL SYSTEMS	Senograph DS VERSION ADS_53.10.10	1.8


```
{"user_id": "annotation_user1", "case_id": "Cancer_00001", "contour_list": {
        "cancer": {
                "lcc": {
                        "1565594002486": [
                                {"y": -162, "x": -506}, {"y": -149, "x": -519}, ... , {"y": -140, "x": -536}]
                },
                "lmlo": {
                        "1565594008811": [
                                {"y": 508, "x": -480}, {"y": 508, "x": -484}, ... , {"y": 504, "x": -489}]
                }
        }
}, discard_yn: 0}
```


※ LCC, RCC, RMLO, RCC 에 대해 병변이 여러 개일 수 있으며, 병변 번호(id)로 악성병변을 구분한다.

데이터 구축 과정

수집된 데이터는 개인정보에 대한 문제가 없도록 하기 위해서 기관생명윤리위원회 (Institutional Review Board, IRB)의 승인을 받고, 임상연구검색 포털시스템을 활용하여 2012.01.01 ~ 2018.12.31까지 만 18세 이상 여성에서 유방촬영영상을 받은 대상자에서 대상자군을 최종 확정하고 CRDW와 PACS에서 임상데이터와 영상이미지를 획득한다. 개인 식별화가 가능한 정보를 삭제하여 비식별화를 하고, 의료전문가가 악성 이미지에서 악성병변 어노테이션을 수행 후 2인의 의료전문가가 교차검증하여 검수를 진행하였다.

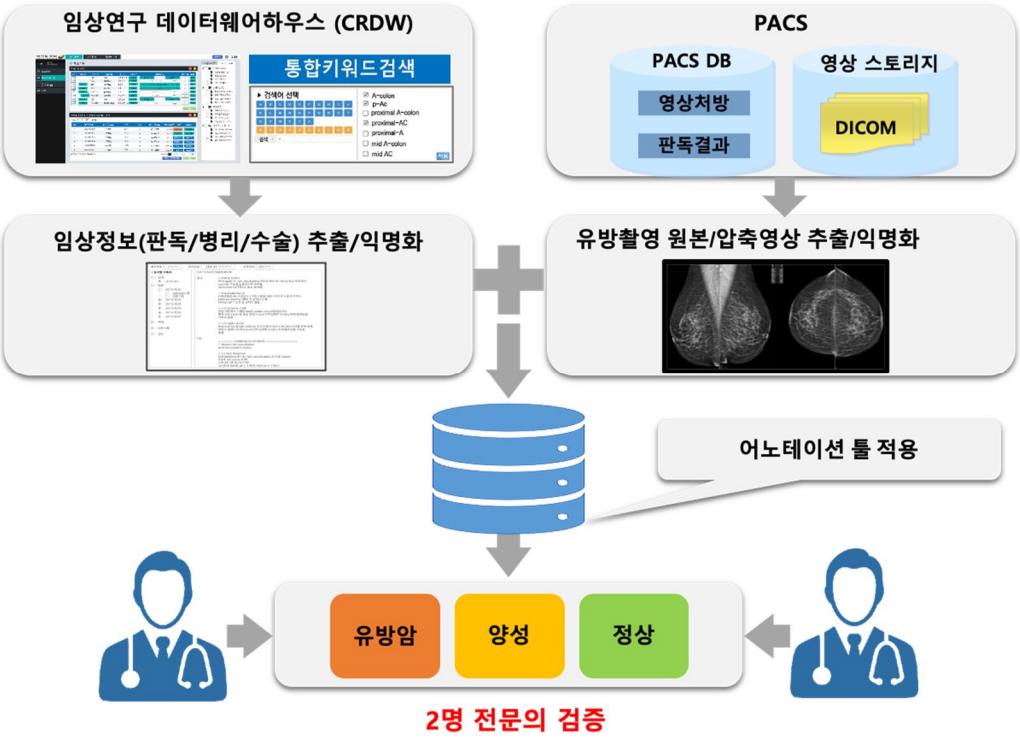

그림 1 데이터 구축 과정

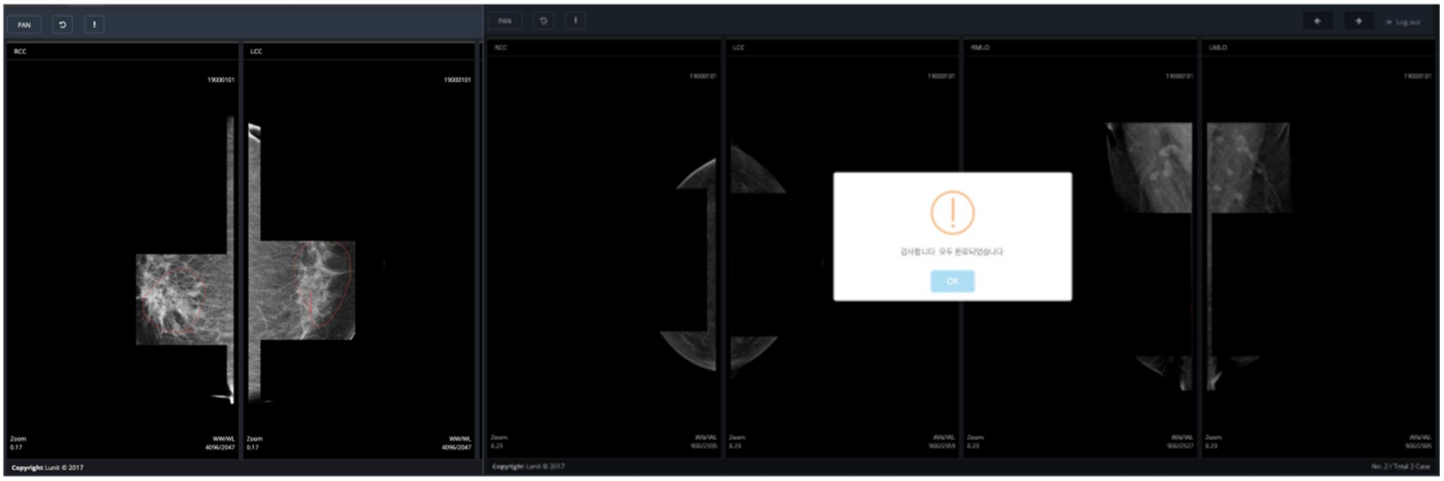

그림 5 악성병변 어노테이션 수행 프로그램

검수와 품질 확보

유방 촬영영상 이미지, 악성병변 어노테이션 정보, 임상정보 등을 질병진단 학습용 데이터로 활용하기 위해서 검수 프로세스를 정립하고 영상판독문 또는 병리판독문이 존재하는 영상에 대해서 두 명의 의료전문가가 교차판독하고 두 명의 소견이 100% 일치를 통해 데이터의 검수를 진행하여 신뢰도가 높은 데이터를 확보하였다.

· Option 1: 저작도구에 어노테이션 review 기능 구현:

- 지금까지 진행된 어노테이션에 대해, 데이터베이스에 저장된 결과파일 (json 혹은 binary mask image)을 영상자료상에 로드하여 해당 병변이 제대로 어노테이션 및 저장됐는지 확인
- 오류가 있을 경우 저작도구를 활용하여 수정하고, 해당 내용에 대해 기록을 남겨 2명의 전문의의

합의하에 최종 판단

· Option 2: 저장된 어노테이션 결과 데이터를 포함한 영상을 별도로 생성하여 검수

- 지금까지 진행된 어노테이션에 대해, 데이터베이스에 저장된 결과파일을 영상자료상에 contour 형태로 표현하여, annotated 영상을 새로 생성하여 해당 병변이 제대로 어노테이션 및 저장됐는지 확인
- 오류가 있을 경우 해당내용에 대해 기록을 남겨 2명의 전문의의 합의하에 최종 판단하고, 이 데이터를 별도로 모아서 최종 어노테이션 수정

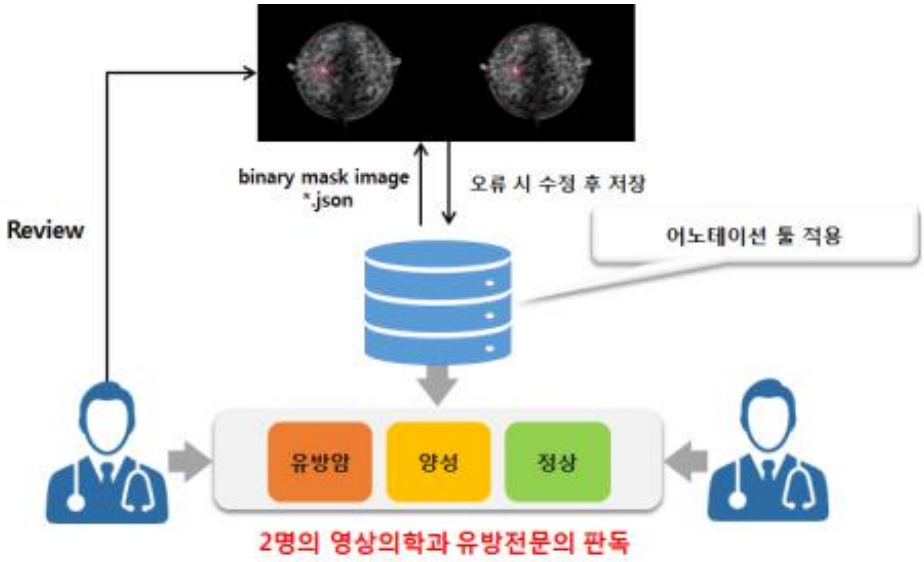

그림 6 품질 확보를 위한 품질 검수 체계

데이터 구축 담당자

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