유방암 판독 인공지능 모델 개발을 위한 유방촬영술 의료지식베이스 구축방안

박현우 국립암센터 인공지능사업팀 산학협력교수

1. 머리말

유방암의 증가는 전 세계적인 현상이다. 세계암통계[1]에 의하면 연령표준 발생률에서 남녀를 모두 합쳐도 유방암이 가장 많이 발생하는 암종이다. 이를 토대로 국내 유방암 발생률은 앞으로도 계속 증가할 것으로 판단된다. 이런 상황에서 기존 건강관리 패러다임이 치료 중심에서 조기진단을 통한 예방 중심으로 변화함에 따라, 유방암을 조기에 검진하는 데 필요한 유방촬영술(Mammography)이 다기관 병원에서 수행되고 있다.

최근 모든 산업분야에 인공지능이 활용되고 있는데, 인공지능 기술로 유방촬영술 영상을 분석할 경우 암을 더 잘 찾아내고, 불필요한 검사를 줄일 수 있을 것으로 기대된다. 그에 따라 유방촬영술 영상에 인공지능을 잘 응용하려면 구조화되고 신뢰성 높은 의료영상 이미지와 메타데이터가 중요하다.

유방촬영술 이미지는 의료용 디지털 영상 및 통신(DICOM, Digital Imaging and Communications in Medicine) 표준을 기반으로 촬영 이미지와 이미지정보 태그를 포함해 수집되고 저장된다. 그런데 환자 개인정보의 비식별화 문제로 각 의료기관마다 태그 정보가 상이하게 저장되는 한편, 인공지능 기술을 활용해 유방암 판독모델을 개발하는 데 필요한 메타데이터 수준(level) 또한 병원마다 상이해서 여러 병원에서 수집되는 유방촬영술 이미지와 메타데이터 구조를 표준화해야 한다는 요구가 증대되고 있다.

본고에서는 다양한 의료기관에서 수집한 유방촬영술 이미지 데이터와 메타데이터 수준의 구조화 방법을 제시해 의료 인공지능 모델을 개발하는 데 필요한 인공지능 학습용 데이터 구축 방안을 정의하려 한다.

2. 유방촬영술 이미지 의료지식베이스 구축

2.1 데이터 수집 절차

병원에서 운영하는 의료데이터베이스(EMR, Electronic Medical Record), PACS(Picture

<표 1> 데이터 수집 절차, 상세 내역, 산출물

수집 절차	상세 내역	산출물
1. 병원 의료 데이터베이스 (EMR, PACS)에 질의	만 19세 이상 성인 여성환자의 유방암 진단을 위해 촬영한 유방촬영영상과 임상 정보 수집	
2. 4-view paired 유방촬영영상 자료 수집	RCC, RMLO, LCC, LMLO 유방촬영영상 수집	
3. 유방촬영영상 판독	영상 판독문 기준으로 유방촬영영상의 악성, 양성, 정상으로 구분	
3.1 유방촬영영상 악성 판정	1. 유방촬영영상에서 악성 병변이 발견되어 악성이 의심되는 의심 환자의 추적 조직검사를 수행 2. 추적 조직검사를 통해 유방암으로 판명된 유방암 환자의 유방촬영영상 데이터 수집	• 유방암 유방촬영영상 • 병리 검사 보고서 • BI-RADS 4, 5, 6
3.2 유방촬영영상 양성 판정	1. 유방촬영영상에서 양성 병변 확인 2. 1년 이상의 지난 시점에서 양성으로 판명된 환자의 유방촬영영상 데이터 수집	• 양성 유방촬영영상 이미지 • BI-RADS 2, 3
3.3 유방촬영영상 정상 판정	1. 유방촬영영상에서 음성으로 판단 2. 1년 이상의 지난 시점에서 여전히 음성으로 판명된 정상군의 유방촬영영상 데이터 수집	• 정상 유방촬영영상 데이터 • BI-RADS 1
4. 영상 비식별화 전처리	영상 데이터 내에 포함된 개인정보(환자번호, 성별, 나이 등)의 비식별화 처리	비식별화된 유방촬영영상 이미지 데이터
5. 비식별화된 영상 데이터와 메타 데이터의 결합	레이블(악성, 양성, 정상) 된 유방촬영영상 이미지 데이터와 관련한 임상 데이터의 결합	두 데이터가 결합된 최종 데이터

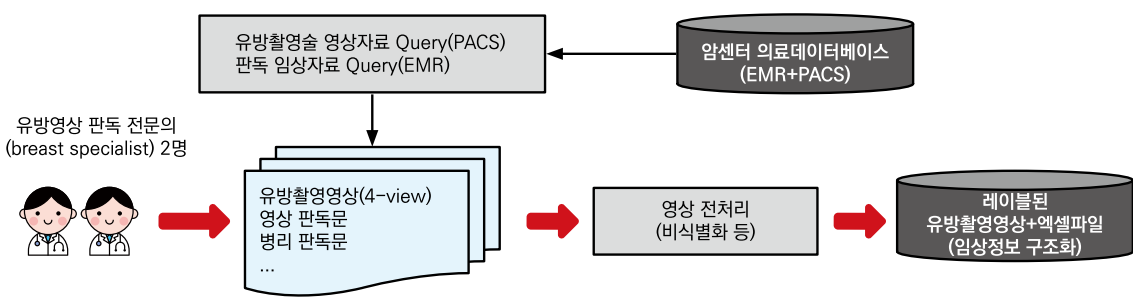
Archiving Communication System)에서는 데이터베이스 질의를 통해 유방암을 진단하고자 촬영한 유방촬영술 이미지(Mammography image) 및 임상정보 등을 수집한다. 유방암을 진단하기 위해 유방촬영술 이미지와 같이 이미지 판독문 및 병리판독문을 활용해 정확한 유방암 진단 기준을 정의한다.

유방촬영술 이미지의 진단 결과는 악성(Malignant), 양성(Benign), 정상(Normal) 유방 촬영술 이미지로 구분한다. 양성, 정상의 경우 판독 당시 오진이 있을 수 있으므로 추후 검사를 통해 변화가 없는지를 확인하는 것이 필요하며 일반적으로 1년 이상 추적검사를 통해 이상 소견이 없는지를 확인하여 최종 판단한다.

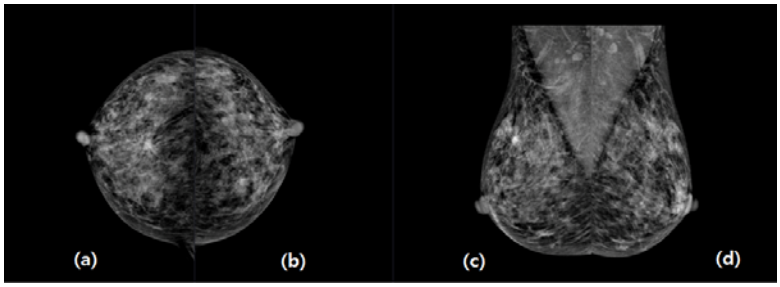
2.2 이미지 데이터 정의

2.2.1 이미지 데이터

유방촬영술 이미지 데이터는 국제 표준인 의료용 디지털 영상 및 통신(DICOM) 형식으로 수집한다. DICOM 파일은 태그가 있는 이미지 파일로 이미지와 이미지에 대한 데이터 정보가 모두 들어 있으며, 이미지 파일의 데이터는 개별 요소의 순서로 저장된다. 각 요소에는 이미지 또



[그림 1] 데이터 수집 절차



[그림 2] 악성 유방촬영영상 (a)LCC (b)RCC (c)LMLO (d)RMLO

는 이미지 자체에 대한 정보 항목이 태그 형식으로 존재한다.

- **태그** : 일반적으로 16진수 형식(xxxx, xxxx)으로 속성을 식별
- **DICOM** Value Representation(VR): 속성값의 데이터 유형 및 형식

4-view(RCC, Right CranioCaudal, RMLO; Right MedioLateral Oblique, LCC, Left CranioCaudal, LMLO, Left MedioLateral Oblique) paired 유방촬영술 이미지로 한 환자에서 4개의 이 미지 데이터 수집하며 [그림 2]와 같은 유방촬영술 이미지 데이터로 정의된다.

2.2.2 유방촬영술 이미지 데이터 비식별화

환자 ID는 비식별화를 위해 환자의 병변 상태를 기준으로 새로 아이디를 부여(예: 음성:00001, 양성:10001, 정상: 20001)한다.

DICOM 태그에서 개인정보 비식별 조치는 ISO/IEC 20889에서 다른 값으로 대체하는 가명처리, 데이터 aggregation, 일부 식별 요소를 삭제, 데이터 범주화, 데이터 마스킹 등을 활용한다. 환자 생년월일과 수행된 프로세스 시작 날짜와 같은 형식의 태그 정보는 비식별화를 위해 '19000101'로 고정한다.

DICOM 데이터의 비식별화를 위해 개인 식별이 가능한 모든 태그는 삭제했으며, 구축하 는 의료지식베이스의 DICOM 데이터에서 필요한 태그는 <표 2>와 같다. 또한, DICOM 태그들이 가지는 속성 값의 정의는 <표 3>과 같다.

2.3 메타 데이터 정의

영상 이미지 데이터 환자의 메타정보(<표 4>)를 수집하며, 악성 환자에 한해 병변에 대한 추가적인 정보를 얻기 위해 수술 후 병리보고서를 수집한다. 또한 추후에 데이터의 품질 검수에서 활용함으로써 구축한 의료지식베이스의 신뢰성을 확보하도록 한다.

2.4 어노테이션 구조

유방촬영영상에서 악성 병변의 위치를 어노테이션을 수행해 유방암 악성판독에 활용 가능하다. (<표 5>)

2.5 데이터 정제

한 환자당 최적의 4-view paired 영상자료를 보장하기 위하여 최종 4개(RCC, RMLO, LCC, LMLO) view를 선별한다.

- 4-view paired (RCC, RMLO, LCC, LMLO) 중 하나라도 없는 유방촬영 영상은 제거
- 각 4개의 view 중에서 영상이 2장 이상 존재하는 경우, 유방암 판독에 적합한 영상을 제외하고 중복되는 영상은 제거 (예: spot compression 영상, magnification view 영상 등)
- DR (digital radiography) 영상자료 수집만 수집하며 CR (computed radiography) 영상은 제거한다.

<표 2> 영상이미지 데이터의 비식별화 후 DICOM 태그 구조

Tag (Group, Element)	VR	TAG Description	Value (예시)
(0002,0000)	UL	FileMetaInformationGroupLength (파일의 메타 정보 그룹 길이)	188
(0002,0002)	UI	MediaStorageSOPClassUID (SOP 클래스의 고유식별자)	1.2.840.10008.5.1.4.1.1.1.2
(0002,0003)	UI	MediaStorageSOPInstanceUID (SOP 인스턴스의 고유식별자)	1.2.840.113681.2887049286.1491896736.4980.135780
(0002,0010)	UI	TransferSyntaxUID (데이터셋 인코딩에 사용되는 전송 구문의 고유식별자)	1.2.840.10008.1.2.4.91
(0002,0012)	UI	ImplementationClassUID (구현된 클래스의 고유식별자)	1.2.410.200003.2020819.5.2.1
(0008,0008)	CS	ImageType (이미지 타입)	DERIVED₩PRIMARY172.20.172172
(0008,0016)	UI	SOPClassUID (SOP 클래스의 고유식별자)	1.2.840.10008.5.1.4.1.1.1.2
(0008,0018)	UI	SOPInstanceUID (SOP 인스턴스의 고유식별자)	1.2.840.113681.2887049286.1491896736.4980.135780
(0008,0060)	CS	Modality (양식)	MG
(0008,0070)	LO	Manufacturer (제조사)	XXXXXX, Inc.
(0008,1090)	LO	ManufacturerModelName (제조사 모델명)	XXXXX Dimensions
(0028,135A)	CS	SpatialLocationsPreserved (공간 위치 보존 여부)	YES
(0010,0020)	PN	PatientID (환자 ID)	10001
(0010,0030)	DA	PatientBirthDate (환자 생년월일)	19000101
(0018,1164)	DS	ImagerPixelSpacing (픽셀 간격)	0.070000₩0.070000
(0018,5101)	CS	ViewPosition (이미지 대상의 방향)	CC
(0020,0062)	CS	ImageLaterality (이미지의 부위 위치: 왼쪽, 오른쪽)	L
(0028,0002)	US	SamplesPerPixel (픽셀당 샘플 크기)	1
(0028,0004)	CS	PhotometricInterpretation (광도 해석)	MONOCHROME2
(0028,0010)	US	Rows (행)	3328
(0028,0011)	US	Columns (열)	2560
(0028,0030)	DS	PixelSpacing (픽셀 간격)	0.065238₩0.065238
(0028,0100)	US	BitsAllocated (할당된 bit 단위)	16
(0028,0101)	US	BitsStored (저장된 bit 단위)	12
(0028,0102)	US	HighBit (픽셀 최상위 bit 값)	11
(0028,0103)	US	PixelRepresentation (픽셀 표현방식)	0
(0028,1050)	DS	WindowCenter (창 중심값)	2047
(0028,1051)	DS	WindowWidth (창 너비값)	4096
(0028,1052)	DS	RescaleIntercept (저장된 값과 출력 사이에 있는 값)	0
(0028,1053)	DS	RescaleSlope (재조정 스케일)	1
(0028,1054)	LO	RescaleType (재조정 유형)	US
(0040,0244)	DA	PerformedProcedureStepStartDate (수행된 프로세스 시작 날짜)	19000101
(0040,0245)	TM	PerformedProcedureStepStartTime (수행된 프로세스 시작 시간)	0
(0008,0100)	SH	CodeValue (코드 값)	R-10242
(0008,0102)	SH	CodingSchemeDesignator (코딩 스키마 지정자)	SNM3
(0008,0104)	LO	CodeMeaning (코드 의미)	cranio-caudal

<표3> DICOM 태그 속성값 정의

VR	속성명	길이
CS	Code String	최대 16 바이트
DA	Date	8 바이트 고정
DS	Decimal String	최대 16 바이트
LO	Long string	최대 64 문자
PN	Person Name	최대 64 문자
SH	Short String	최대 64 문자
TM	Time	16 바이트 고정
UI	Unique Identifier (UID)	최대 64 문자
UL	Unsigned Long	4 바이트 고정

<표 4> 메타 데이터 수집 목록

데이터명	상세내역
인구학적 정보 및 종양병기	• 촬영시점의 나이 • 종양의 병기(t-stage)
유방촬영영상의 판독보고서 (radiology report)	• 영상 촬영 일시 • 유방촬영 영상의 형태 (4-view paired 여부) • 유방촬영 영상의 기기 정보 (모델명, 제조사명) • BI-RADS 판독 결과
악성 유방촬영 영상의 조직검사 및 병리 검사 보고서 (pathology report)	• 조직 검사 일시 • 유방암 조직 검사 결과 • 조직 검사가 수행된 병소의 유방촬영영상에서의 위치
양성 및 정상 유방촬영 영상의 추적 유방촬영 영상 판독 보고서	• 추적 영상 촬영 일시 • BI-RADS 판독결과 (final assessment)

<표 5> 어노테이션 json 포함 정보

컬럼명	상세내역
user_id	수행 영상판독 전문의 ID
case_id	해당 유방촬영 이미지 ID
view_location	유방촬영영상 위치 (RCC, RMLO, LCC, LMLO)
lesion_id	악성병변 ID
contour coordinate	악성병변 coordinates list
discard_yn	해당 영상에 문제가 있어서 사용하지 못하게 될 케이스임을 표시하는 용도

```
{
  "user_id": 1,     #Annotation 전문의에게 부여되는 tool 로그인 ID
  "case_id": 100001,#해당 Mammography 케이스 ID
  "contour_list": { #각 view에 그려지는 lesion에 대한 정보가 list of coordinates 형태로 저장
    "cancer": {     #cancer case에 대한 lesion wjdqhrk view (roc, rmlo, lcc, lmlo)별로 저장
      "rmlo": {     #rmlo view에 대한 lesion 좌표 정보들
        "lesio01": [{"y": 78, "x": 195}, {"y": 78, "x": 191}, ... , {"y": 73, "x": 195}, {"y": 78, "x": 195}]
        "lesio02": [{"y": 48, "x": 341}, {"y": 48, "x": 355}, ... , {"y": 52, "x": 355}, {"y": 48, "x": 355}]
      },
      "rcc": {       #rcc view에 대한 lesion 좌표 정보들
        "lesio03": [{"y": 154, "x": 42}, {"y": 154, "x": 49}, ... , {"y": 142, "x": 42}, {"y": 154, "x": 42}]
      }
    },
    "discard_yn": 0  #해당 영상에 문제가 있어 사용이 불가능할 경우 제외 여부 yes or no 표기
  }
```

[그림 3] 악성병변 어노테이션 json 파일 level별 구성 예시

2.6 의료지식베이스 구조모델

<표 6> 의료지식베이스 데이터 구조

테이블명	컬럼명	데이터타입	상세정보
이미지자료	영상이미지 id	char(10)	영상이미지 id
	환자 개인식별 대체번호	char(10)	환자 대체키로 id로 비식별화
	촬영순서	int	촬영순서 표시
	4-view 유방촬영영상 위치	char(1)	RCC, RMLO, LCC, LMLO
	레이블	char(1)	악성, 양성, 정상
	유방촬영 영상의 기기모델	varchar(20)	모델 제품명
	유방촬영 영상의 기기모델번호	varchar(20)	모델번호

테이블명	컬럼명	데이터타입	상세정보
이미지자료	BIRADS category (both)	varchar(5)	양쪽 유방촬영영상이미지의 BIRARDS category
	BIRADS category (left)	char(2)	왼쪽 유방촬영영상이미지의 BIRARDS category
	BIRADS category (right)	char(2)	오른쪽 유방촬영영상이미지의 BIRARDS category
	환자 개인식별 대체번호	char(10)	환자 대체키로 id로 비식별화
	촬영순서	int	촬영순서 표시
	나이	int	환자의 나이를 구간화하여 연령대로 제공
	검사일	date	
	처방일	date	
	악성 종양의 크기	float	악성 종양의 크기 (cm)
	악성 종양의 위치	char(5)	악성 종양의 위치 (left, right)
	종양의 병기	char(1)	악성 종양의 병기 (t-stage)
어노테이션	임상의 id	char(10)	annotation 수행 영상판독 전문의 비식별화된 ID
	이미지 id	char(10)	해당 유방촬영 이미지 id
	유방촬영영상 위치	char(1)	1: RCC 2: RMLO 3: LCC 4: LMLO
	악성병변 id	char(5)	악성병변 ID
	악성병변 coordinate	string	악성병변 coordinates list
	discard_yn	binary	해당 영상에 문제가 있어서 사용하지 못하게 될 케이스임을 표시하는 용도

*BIRADS: Breast Imaging-Reporting and Data System

3. 맺음말

본고에서는 유방촬영술 영상 이미지를 인공지능 학습용 데이터로 활용하기 위한 표준화 방안을 소개했다. 본고에서 강조하는 표준화 방안을 통해 각 의료기관마다 서로 상이하게 수집·저장하는 유방촬영술 이미지 데이터를 유방암 판독 인공지능 모델 구축을 위한 인공지능 학습용 데이터로 활용할 수 있다. 이를 통해 유방암 조기 판독 인공지능모델을 구축하고 검증해 전 세계적으로 증가하는 유방암 환자를 조기에 진단함으로써 의료비를 절감하고 국민 건강을 증진할 수 있을 것이다.

※ 본 연구는 2019년도 정부(과학기술정보통신부)의 재원으로 한국지능정보사회진흥원(인공지능 학습용 데이터 구축사업)의 지원을 받아 수행된 연구임.

주요용어풀이

- **주석**(Annotation): 이미지 내 각종 사물을 알아보고 경계선을 구분 짓는 것
- **의료 영상 저장 전송 시스템**: 의료 영상을 기존 필름 대신에 디지털 형태로 저장하고 통신망을 통해 의료진들에게 전송하는 장치
- **유방영상보고 및 자료체계**(BI-RADS, Breast Imaging Reporting and Data System): 유방암 선별과 진단에 사용되는 유방촬영, 유방초음파 및 자기공명영상(MRI, Magnetic Resonace Imaging) 결과를 표준화된 방식으로 판정하고 보고하는 방법
- **디지털 의료 영상 전송 장치**(DICOM, Digital Imaging abd Communications in Medicine): 미국 방사선 학회와 전기 공업회가 합동으로 설립한 ACR/NEMA 위원회(1996년에 DICOM 위원회로 개칭)가 모체가 되어 의료 화상 전송을 중심으로 정한 규격. 현재는 데이터 보존 규격도 포함된 표준 규격

참고문헌

[1] http://gco.iarc.fr/