Review Article

J Korean Soc Radiol 2021;82(1):12-28
https://doi.org/10.3348/jksr.2020.0205
eISSN 2288-2928

대한영상의학회지

Applications of Artificial Intelligence in Mammography from a Development and Validation Perspective

유방촬영술에서 인공지능의 적용: 알고리즘 개발 및 평가 관점

Ki Hwan Kim, MD*, Sang Hyup Lee, MD

Lunit Inc., Seoul, Korea

Received December 16, 2020
Revised January 21, 2021
Accepted January 26, 2021

***Corresponding author**
Ki Hwan Kim, MD
Lunit Inc., 15F,
27 Teheran-ro 2-gil, Gangnam-gu,
Seoul 06241, Korea.

Tel 82-2-2138-0827
Fax 82-2-6919-2702
E-mail khkim@lunit.io

ORCID iDs
Ki Hwan Kim
https://orcid.org/0000-0001-7684-235X
Sang Hyup Lee
https://orcid.org/0000-0002-6773-2965

Mammography is the primary imaging modality for breast cancer detection; however, a high level of expertise is needed for its interpretation. To overcome this difficulty, artificial intelligence (AI) algorithms for breast cancer detection have recently been investigated. In this review, we describe the characteristics of AI algorithms compared to conventional computer-aided diagnosis software and share our thoughts on the best methods to develop and validate the algorithms. Additionally, several AI algorithms have introduced for triaging screening mammograms, breast density assessment, and prediction of breast cancer risk have been introduced. Finally, we emphasize the need for interest and guidance from radiologists regarding AI research in mammography, considering the possibility that AI will be introduced shortly into clinical practice.

Index terms Mammography; Artificial Intelligence; Breast Cancer

유방암 검진과 유방촬영술

유방암은 국내뿐만 아니라 소득수준이 높은 북미, 서유럽 국가들에서 가장 흔한 여성암이며, 2018년 WHO 보고에 따르면 한해 전 세계적으로 약 2088만 명의 유방암 환자가 발생하고 약 63만 명이 유방암으로 인해 사망한다고 한다(1). 국내의 경우 보건복지부의 국가암등록사업 보고에 따르면 2017년 기준으로 새롭게 진단받은 유방암 환자는 약 2만 2천 명이었으며, 이는 전체 여성암의 20.3%를 차지했다(2). 이는 2007년에 약 1만 2천 명의 유방암 환자가 새롭게 진단된 것과 비교 시, 환자의 수가 빠르게 증가하고 있음을 시사하며(3), 2017년 기준으로 10만 명당 유방암 연령 표준화 발생률이 55.6명으로 아시아 국가들 중에서 발생률

최상위 그룹에 속한다(1).

유방암 선별검사는 증상이 없는 여성에서 유방암을 조기에 발견함으로써 유방암에 의한 사망률을 낮출 수 있는 것으로 보고되어, 약 20~30년 전부터 유방촬영술을 이용한 유방암 선별검사가 선진국들을 위주로 도입되었다. 우리나라는 1999년부터 유방암조기검진사업을 시작하였고, 현재는 국가암검진 사업을 통해 만 40세 이상 여성을 대상으로 2년마다 유방촬영 검사를 시행하고 있다. 2018년 기준으로 유방암 검진 대상 인원 6371080명 중에서 4069694명(63.9%)이 유방촬영술을 시행하였다(4).

유방촬영술은 유방암의 조기 검진을 위한 효과적인 검사 방법이지만 몇 가지 문제점을 가지고 있다. 첫째로 유방촬영술은 단독으로 유방암을 진단하기에는 민감도(sensitivity)가 부족한 검사이다. 국내 유방암 선별검사에서의 유방촬영술 민감도는 2014년을 기준으로 76.3%로 보고되었으며(5), 2007년부터 2013년 사이 미국의 95개 병원에서 시행된 유방암 선별검사를 분석한 연구 결과에서는 민감도가 86.9%로 보고되었다(6). 유방촬영술의 민감도는 병변이 유방실질조직에 가려짐으로 인한 인지 어려움 또는 양성과 악성의 구분이 제한됨 등과 같은 여러 요인들에 영향을 받는다(7). 이와 같이 낮은 민감도를 극복하기 위한 노력이 때로는 높은 위양성률 문제를 야기하기도 한다. 미국의 검진 유방촬영술의 판독 결과를 분석한 연구에 따르면 25%의 의사가 권장되는 소환율(12%)보다 높은 소환율을 보였으며(6), 국내 연구에서도 2015년 유방암 선별검사에서 판정유보 판정을 받은 비율이 14.7%로 보고되었다(8). 치밀유방은 유방촬영술의 판독을 더욱 어렵게 하는 요인들 중에 한 가지이다. 아시아인에서는 치밀유방의 비율이 높은데, 국가암검진사업 수검자들 중에 약 절반이 치밀유방인 것으로 보고되었고, 특히 40대의 경우 치밀유방의 수검자가 전체의 70%를 넘었다(8). 이와 같은 이유들로 유방촬영술은 높은 수준의 민감도를 확보하기 어려우며, 판독에 있어서 높은 수준의 전문성을 필요로 하고, 따라서 판독자 사이에 판독 성능의 차이가 상당히 날 수 있는 검사법이다.

전통적인 CAD 시스템의 도입

이처럼 판독이 어려운 유방촬영술에서 유방암의 검출 정확도를 향상시키기 위한 목적으로 유방암이 의심되는 병변을 자동으로 검출하는 컴퓨터 보조 발견(computer-aided detection; 이하 CAD) 또는 컴퓨터 보조 진단(computer-aided diagnosis; CADx)이라는 알고리즘들이 1970년대부터 개발이 되기 시작했으며, 1998년에 처음 FDA 허가를 받고 미국 병원들에 도입이 되기 시작했다(9). 이러한 전통적인 CAD들을 이용한 초기 임상 연구들은 CAD 없이 판독한 후 CAD가 표지한 부분을 다시 한번 확인하는 방식으로 판독을 하면 유방암을 더 많이 찾을 수 있다는 결과들을 보고하였으며(10, 11), 미국에서는 보험수가로 인정받아 2016년을 기준으로 미국의 약 90% 이상 병원들에 CAD가 보급되었다.

대부분의 전통적인 CAD 시스템들은 크게 두 단계를 거쳐서 분석 결과를 만든다. 첫 번째는 영상에서 정상조직이 아닌 것으로 추정되는 영역을 찾아내는 것이며, 이는 픽셀의 값, 주변 픽셀과 해당 픽셀의 값 차이, 텍스처, 그리고 모양 등의 특징들을 분석하는 과정을 통해서 이루어진다

(12). 이 과정에서는 석회화 의심 소견과 연조직(soft tissue) 의심 소견이 각각 따로 검출이 된다. 두 번째는 위양성 판정을 줄이기 위한 과정으로 첫 번째 단계에서 검출된 여러 후보 병변들을 대상으로 분포나 모양, 텍스처 등을 고려하여 군집화를 이루어 분석한다(13). 두 번째 과정에서는 특이도를 높이는 것을 목적으로 하며, 최종적으로 암이 의심되는 위치를 표기하여 사용자에게 보여주게 된다. 전통적인 영상처리 기술 기반의 CAD 시스템들도 시간이 지나면서 지속적으로 성능이 개선되었으며, 한쪽 유방의 두 방향 영상정보를 모두 활용하는 기술로도 발전되었다(14).

그러나 여러 대규모 전향적 연구들에서 CAD의 효용성은 확실하게 입증되지 못하였다(15-17). 이는 CAD의 높은 빈도의 위양성 판정으로 인해서 판독자로 하여금 피로감을 유발하게 하여 실제적으로 참고를 하지 않게 되거나, CAD로 인해 위양성이 증가하는 등의 여러 요인들이 복합적으로 작용한 것으로 해석되고 있다(18). 실제 유방암 검진환경에서는 유방암의 유병률이 낮기 때문에, CAD 사용에 의해서 한 개의 유방암을 더 찾기 위해서 훨씬 더 많은 수의 위양성을 감수해야 하는 상황이 발생하기 쉬우며, 따라서 CAD 사용은 유방암을 더 찾을 수 있도록 보조하는 역할뿐만 아니라 위양성이 과도하게 증가되는 효과도 모두 고려되어야 한다.

비록 전통적인 영상처리기술 기반의 CAD 시스템은 실제 의료 현장에서 신뢰를 얻지 못했지만, 전 세계적으로 유방촬영술을 높은 수준으로 판독할 수 있는 전문가의 부족 문제는 여전히 해결되지 않았기 때문에 더 나은 진단 정확도를 갖춘 CAD 시스템에 대한 수요는 여전히 존재한다.

인공지능 기반의 새로운 CAD의 소개

전통적인 영상처리 기술 기반의 CAD 시스템들은 전문가가 정의한 유방암의 특징들을 유방촬영술에서 검출하도록 개발되었다. 이러한 유방암의 특징적인 영상 소견들은 오랜 시간에 걸쳐 여러 전문가들의 수많은 연구 결과들에 의해 만들어져 왔다. 유방촬영술에서 관찰되는 유방암의 모양 및 특징을 최대한 객관적으로 평가하기 위해서, 여러 가지 이상 소견들을 묘사하기 위한 설명자(descriptor)들은 유방영상 판독 및 데이터 시스템(Breast Imaging Reporting and Data System; 이하 BI-RADS)을 통해서 정의하여 사용해 왔다(19). 의사들은 많은 수의 유방암 증례와 유방암이 아닌 정상 또는 양성 질환들의 증례들의 영상의학적 소견에 대해 육안으로 평가하고, 그 소견들을 설명자들을 이용해 분석하는 방식으로 유방암과 연관성이 높은 특징적인 영상 소견들을 밝혀냈다(20, 21). 이렇게 밝혀진 유방암의 특징적인 영상 소견들을 검출하는 방식으로 전통적인 CAD 시스템들이 개발되어 왔다.

인공지능 기술의 한 종류로써 최근 영상 분석에 있어서 많은 관심을 받고 있는 딥러닝(deep learning)은 이와 같은 전통적인 영상 분석 방법과는 분석의 시작부터 방법을 달리하는 것이 큰 차이점이다. 여러 딥러닝 기법들 중에서도 합성곱 신경망(convolutional neural network)이라는 기법이 영상자료 해석에 많이 사용되고 있으며, 여러 층의 신경망이 연속적으로 쌓여서 만들어지는 합성곱 신경망은 영상의 다양한 공간적 정보로부터 특정한 작업을 수행하는 데 필요한 주요 정보들을 스스로 배울 수 있다고 알려져 있다(22-24). 즉, 흔히 딥러닝 알고리즘을 학습시킨다는 것은 영상자료로부터 풀고자 하는 문제를 풀기 위한 특징들을 알고리즘 스스로 정의하고 찾아가는

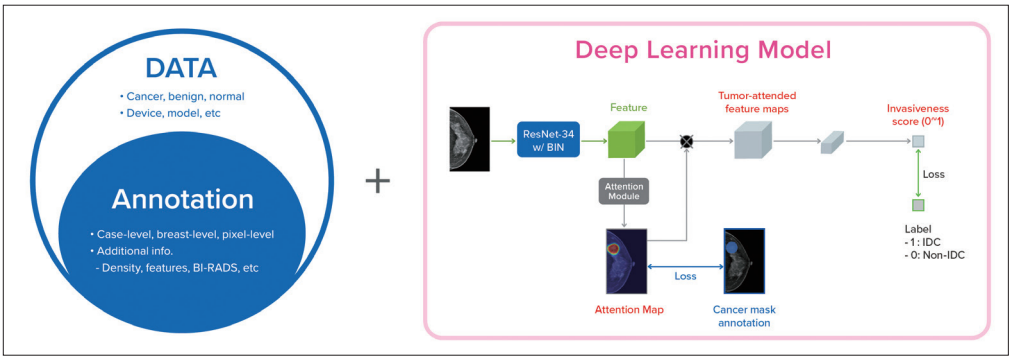
과정을 의미하며, 딥러닝을 이용하여 유방암 검출 알고리즘을 학습시키는 과정 또한 알고리즘이 데이터로부터 유방암과 관련된 다양한 특징들을 스스로 찾아내는 과정을 말한다.

유방촬영술을 위한 최적의 딥러닝 알고리즘 개발 방법

딥러닝 알고리즘이 유방암의 특징들을 효과적으로 배우기 위해서 필요한 요소로 어떤 것들이 있을지 있는지 살펴보자. 첫 번째로는 많은 수의 유방암 영상들을 학습 데이터로 확보해야 한다(Fig. 1). 딥러닝 알고리즘의 성능은 철저히 학습 과정에 의존하기 때문에, 학습에서 보지 않은 것에 대해서도 높은 성능을 기대하기는 어렵다. 현재 상용화에 이르는 여러 영상자료 분석 알고리즘들은 대개 지도학습기반(supervised learning)인 경우가 많으며, 유방촬영술에서 유방암을 찾은 딥러닝 기술들도 대부분 정답지가 있는 데이터를 이용하여 개발된다(25). 지도학습에서는 높은 수준의 정확도를 확보하기 위하여 유방촬영술에서 보일 수 있는 모든 범위를 최대한 포함할 수 있는 대규모의 자료가 필요하며, 이 자료에 다양한 크기, 다양한 위치, 다양한 영상 소견들을 포함할수록 이를 이용해 학습된 딥러닝 알고리즘은 더욱 정확한 성능을 확보할 가능성이 높다. 또한 유방암 증례의 다양성을 넘어서, 딥러닝의 일반화(generalization)를 잘 이루어낼 수 있도록 다양한 장비, 검사 기관, 연령대, 인종 등을 폭넓게 포함한 자료를 확보하는 것이 중요하다.

두 번째로는 수집된 데이터를 최대한 효율적으로 활용하기 위한 전략이 필요하다. 데이터의 숫자는 늘 제한적이기 마련이므로, 연구자들은 대개 수집한 자료의 활용을 극대화하기 위한 방법들을 고려해야 한다. 효율적인 자료 활용을 위해서 가장 먼저 하는 일은 여러 곳에서 수집된 자료를 일정한 기준으로 중앙화(centralization) 하고 이를 정제(cleansing) 하는 전략을 세우는 것이다. 해당 분야의 전문가와 딥러닝 연구자는 적절한 딥러닝 알고리즘 개발을 위한 비용 대비 효과가 극대화될 수 있는 자료 정리 전략을 세우는 과정이 필요하다. 그다음으로는 수집된 데이터들을 학습에 이용하기 위한 정답지를 정의하는 것이다. 영상 데이터의 활용을 극대화하기 위한 대표적인 방법은 병변의 위치를 영상에 표기함으로써 전체 영상에서 실제 병변이 위치하는 곳의 정보를 학습

Fig. 1. Development of deep learning models. The first step is to collect a large-scale dataset, followed by dataset annotation by radiologists, to improve the utility of the dataset. Deep learning models are trained using the collected data and annotations.
BI-RADS = Breast Imaging Reporting and Data System, IDC = invasive ductal carcinoma, ResNet-34 w/BIN = RestNet-34 with Batch Instance Normalization



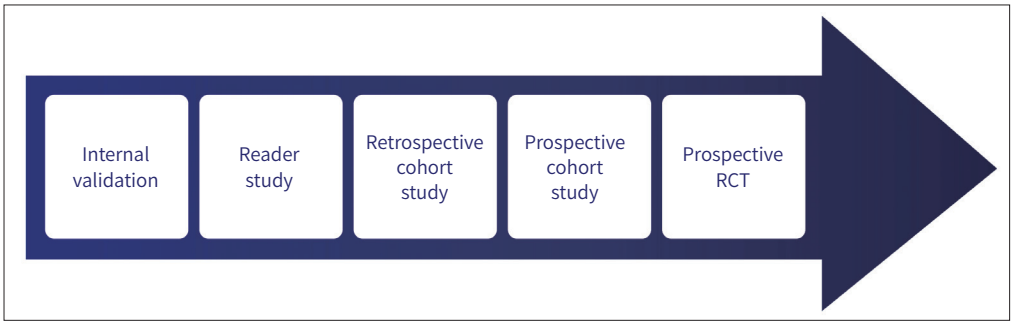
에 활용하는 것이다. 만일 딥러닝 알고리즘을 학습시키는 과정에서 각 케이스 별로 유방암의 존재 여부의 정보만을 이용해서 알고리즘을 학습시킨다면, 알고리즘은 학습과정에서 4개의 영상들을 모두 살펴보고, 4장의 영상 중에서 어디에 있을지 모르는 유방암의 특징을 찾기 위한 노력해야 한다. 4장의 영상에서 유방암이 차지하는 면적이 매우 작은 영역이라는 것을 감안하면, 위치 정보 없이 유방암을 찾는 것이 아주 쉽지 않은 일이라는 것을 짐작할 수 있을 것이다. 따라서 자료활용 효율을 높이기 위해서는 유방암 증례에서 병변이 어디에 있는지에 대한 위치 정보에 대한 정답지(label)를 만들어서 사용하는 하는 것이 효율적이다. 위치에 대한 정보를 제공하는 방법도 여러 단계를 생각해 볼 수 있는데, 가장 단순하게는 유방암이 좌 우중에서 어느 쪽에 유방암이 존재하는지를 활용하는 방법이 있으며, 더 나아가 영상에서 유방암에 해당하는 위치를 표기하여 활용하는 방법도 있다. 알고리즘 학습에 사용하는 정답지(label)가 더욱 세밀할수록 같은 양의 자료를 더욱 효과적으로 활용할 수 있다. 다만 이 과정에서 추가적으로 드는 비용(cost)을 고려하여, 적절한 수준에서 타협을 하고 전략을 세우게 된다. 지도학습(supervised learning)에서는 이처럼 정답지(label)가 얼마나 구체적이고 정확한지에 따라 학습 효율이 달라지며, 이는 사람이 무언가를 새롭게 배우는 방식과도 비슷하다고 볼 수 있다. 유방촬영술에서 유방암을 검출하는 알고리즘을 학습시킬 때는 다른 영상 검사 결과들 또는 조직검사 결과들을 참고한 유방암의 실제 위치 정보를 학습에 이용하거나, 유방 치밀도 또는 의사의 BI-RADS 판정 결과 등의 정보들을 함께 활용함으로써 제한된 자료의 활용효율을 높일 수 있으나, 이에 따른 비용(cost)과 적절한 타협점을 찾고는 것이 현실에서는 중요한 의사결정이 된다.

마지막으로는 특정 문제를 풀기 위한 최적의 딥러닝 알고리즘을 구현하는 것이다. 딥러닝 관련 기술들은 누구나 접근 가능한 형태로 공개되어 있어, 성능이 좋은 기본 알고리즘에 대한 접근성이 용이한 것은 사실이다. 하지만 일반 이미지를 분류하는 것과 의료영상의 특성은 다르며, 또한 문제에 따라 필요한 특징들이 다르다. 따라서 공개된 좋은 알고리즘으로 특정 목적의 알고리즘 개발을 시작할 수는 있지만, 목적에 맞는 최적의 알고리즘을 만들기 위해서는 또 다른 많은 노력들이 필요하다. 일반 이미지 분석과 대비해, 유방촬영술은 총 4개의 영상으로 구성되어 있으며, 한쪽 유방에 대해서 2개의 다른 자세로 촬영된 영상이 있고, 비록 양쪽 유방은 서로 다르지만 서로 약간의 연관성은 있으며, 이미지의 해상도는 매우 높은 특징을 가진다. 이러한 영상자료를 가장 잘 해석하기 위한 공개된 딥러닝 기술은 해당 도메인의 여러 연구들을 참고하여 연구자들이 스스로 개발해야 하며, 최적의 알고리즘을 개발하는 과정에는 학습에 사용할 데이터의 숫자 및 정답 자료의 형태 등도 함께 고려가 되어야 한다(26-29). 따라서 유방촬영술에서 유방암을 검출하는 최적의 딥러닝 알고리즘을 개발하는 과정은 수많은 반복 학습과정을 거쳐서 완성되게 되며, 이 과정에서 수집한 자료들의 사용을 극대화하기 위한 유방영상의학 전문의와 협업도 반드시 필요하다.

딥러닝 알고리즘의 단계별 임상 검증

유방암 검출을 위한 딥러닝 알고리즘을 실제 의료현장에 사용하기 위해서는 적절한 임상자료를 이용한 성능 평가 과정을 거쳐야만 하며, 이 과정에서 의료인들의 역할이 강조되고 있다(30-32).

Fig. 2. Process for clinical validation of deep learning algorithms.
RCT = randomized control study



첫 번째 단계로는 알고리즘 개발에 사용된 데이터 중에서 일부의 데이터를 이용한 내부 성능 평가(internal validation)이다(Fig. 2). 이 내부 검증에 사용된 데이터와 같은 특성을 가진 데이터가 이미 딥러닝 알고리즘의 학습에 사용이 되었기 때문에, 일반적으로 내부 검증에서는 높은 성능을 보이기 쉽다. 내부 검증에서의 성능은 크게 두 가지 관점에서 해석될 수 있으나, 이 둘을 구분할 수 없다는 점에서 근본적으로 한계가 있다. 첫째로 알고리즘이 학습과정에서 연구자의 의도대로 특정 문제를 잘 풀 수 있도록 학습이 된 경우이다. 두 번째로는 딥러닝 알고리즘이 학습 데이터에서는 해당 문제는 잘 풀어낼 수 있지만, 조금만 다른 특성의 데이터에 대해서는 같은 문제를 잘 풀지 못할 가능성이다. 이는 학습 데이터에 지나치게 의존적인 딥러닝 알고리즘이 개발될 수 있으며, 이는 과적합(overfitting)으로 불린다(33). 특히 딥러닝의 경우는 학습과정에서 학습되어야 하는 변수의 숫자가 매우 많기 때문에, 학습에 사용할 데이터의 규모에 비해서 깊은 층의 딥러닝 알고리즘을 사용할수록 학습 데이터에 지나치게 의존적인 알고리즘으로 학습될 가능성이 높다. 또한 풀고자 하는 문제가 얼마나 복잡한 문제인가 단순한 문제인가에 따라서도 과적합 발생 가능성이 영향을 받는다. 그 밖에도 과적합에 영향을 주는 요인들은 다양하며, 과적합 문제는 딥러닝 알고리즘을 실험실 밖에서 실제로 사용하는 데 있어서 매우 중요한 이슈이다(32-34). 따라서 딥러닝 알고리즘 관련 연구 결과를 평가할 때, 해당 알고리즘이 과적합 여부를 평가할 수 있는 연구 설계인지, 그리고 과적합으로부터 얼마나 영향을 받는지를 확인해야 한다.

내부 검증을 통해서 충분한 성능에 도달하였다면, 연구자는 학습에 사용되지 않은 자료를 이용하여 성능을 확인하게 되며, 이를 외부 검증(external validation)이라고 부른다. 이 외부 검증에서는 학습에 사용되지 않은 새로운 자료를 이용하는 것이며, 이 과정에서 과적합의 정도를 평가할 수 있다(35). 유방촬영술 분석을 위한 딥러닝 알고리즘의 외부 검증을 목적으로 자료를 수집할 때, 다음의 몇 가지 항목들을 고려해야 한다. 우선은 알고리즘의 사용목적에 맞는 다양한 영상촬영장비 또는 촬영방식을 포함하는 자료를 수집하는 것이다. 예를 들어 필름 유방촬영술은 사용목적에 포함되지 않았다면, 이 자료들로 검증을 하는 것은 의미가 없을 것이다. 반면 디지털 유방촬영술로 딥러닝 알고리즘의 사용목적을 지정하였다면, 다양한 제조사, 제품 모델, 촬영 조건, 영상 후처리(post-processing) 방식 등 촬영된 다양한 디지털 영상으로 평가하는 것이 외부 검증으로써 연구 결과의 가치를 높일 수 있다. 촬영 방식의 차이 외에도 환자의 구성의 차이도 고려될 수 있다. 가령 건강검진 자료처럼 유방암 유병률이 낮은 환경에서 수집된 자료인지, 대학병원에서 수집된

자료인지에 따라서 환자군의 특성이 차이가 다르며, 따라서 평가 결과도 다를 수 있다(36). 그 밖에도 동양인과 서양인의 유방실질의 치밀도 차이 또는 나이에 따른 유방실질의 차이 등 다양한 요소들이 데이터 세트의 차이를 만들 수 있다(37). 이상적인 딥러닝 알고리즘이 어떠한 조건에서 촬영된 영상자료에도 일정하게 좋은 성능을 보여주는 것이겠지만, 이는 영상의학과 전문의도 영상의 화질이나, 유방의 치밀도, 또는 환자군에 따른 판독 정확도의 성능이 발생할 수 있음을 감안하면, 어떠한 상황에서도 일정한 성능을 내는 알고리즘의 개발은 현실적으로 불가능하다(38, 39). 따라서 딥러닝 알고리즘의 성능 평가에 있어서 적절한 대조군의 선정이 필요하며, 알고리즘 사용 범위 안의 다양한 조건으로 촬영된 영상자료에 대해서 대조군과의 성능 비교를 통한 평가로 기술에 대한 신뢰도를 쌓을 수 있다.

유방촬영술 CAD의 경우, 알고리즘의 단독 성능을 평가하는 것 외에도 CAD 사용에 의한 영상의학과 의사의 판독 정확도 향상을 입증하는 평가가 필요하다. 이는 일반적으로 리더스터디(reader study)라고 불리며, 여러 명의 의사들이 수백 장의 유방촬영술을 판독함에 있어서 CAD 또는 딥러닝 알고리즘 분석 결과를 참고하지 않았을 때 대비 분석 결과를 참고하였을 때의 판독 결과가 얼마나 향상되는지 비교 평가하는 방식으로 소프트웨어 사용에 의한 추가적인 가치를 입증하는 방식이다. 이러한 경우 유방암 환자군과 유방암이 아닌 환자군의 자료를 각각 일정 수 수집한 환자-대조군 연구의 방식으로 진행되며, 실험적인 판독 조건하에 딥러닝 알고리즘 사용에 의한 판독 성능 향상을 입증할 수 있다. 이는 국내나 미국 등 여러 국가에서 의료기기 허가를 위한 임상시험을 시행할 때 흔히 사용되는 연구 방식이기도 하며, 의료기기로써 성능을 검증하는 목적으로 흔히 시행되고 있다(40, 41). 다만 이러한 연구 결과는 딥러닝 알고리즘에 의한 판독 정확도 향상을 주장하기 위한 첫 번째 단계의 검증이며, 실제 의료환경에 사용되었을 때의 기대되는 효과와는 다름을 인지할 필요가 있다.

유방촬영술에서 진단 보조 소프트웨어의 주요 사용목적은 주로 유방암 선별검사 환경에서의 유방암 검출률 향상, 소환율 감소를 목표로 하기 때문에, 실제 유방암 검진 코호트에서의 성능 평가가 이루어져야만 한다. 이전에 수집된 유방암 검진 코호트 자료의 활용이 가능하다면, 유방암 검진 환경에서 딥러닝 알고리즘의 예민도와 특이도를 후향적으로 평가할 수 있으며(42, 43), 이를 바탕으로 유방암 검진에서의 유방암 검출률 및 소환율을 예측할 수 있다. 또한 해당 코호트 자료에 판독의사의 판독 결과가 존재할 경우, 의사들의 판독 결과와 딥러닝 알고리즘의 판독 결과를 비교함으로써 잠재적으로 기대되는 유방암의 추가 검출 효과를 예상할 수 있다(43). 이러한 후향적 코호트 연구에서는 소규모 데이터에서의 성능 평가 결과가 실제 유방암 검진환경에서 재현될 수 있는지에 대한 가능성을 확인할 수 있으며, 선택 편향에 의한 효과를 최소화한 상태에서의 임상적 진단성능 평가가 가능하다(32). 후향적 코호트 연구를 통해서 딥러닝 알고리즘에 의한 추가적인 임상적 가치에 대한 근거가 확보된다면, 마지막 단계인 전향적 연구를 시행하게 될 것이다.

전향적 연구의 주목적은 리더스터디 및 후향적 코호트 연구에서 확인한 유방암 검출률 및 소환율의 변화를 실제 유방암 검진 코호트에서 입증하는 것이다. 이는 단일 또는 다기관으로 진행될 수 있으며, 전향적 코호트 연구 또는 무작위 배정연구 방식으로 진행될 수 있다. 전향적 연구의 경우 연구 대상자 모집에만 1~2년이 걸리고, 간격암(interval cancer)에 대한 평가까지 이루어지기

위해서는 1~2년의 경과 관찰기간까지 필요하기 때문에, 실제 연구 결과를 획득하기까지 3~4년이 걸린다. 뿐만 아니라 전향적 연구는 각국의 유방암 검진 방식에 따라서 진행되어야 하기 때문에, 여러 국가들에서 인정을 받기 위해서는 그만큼 많은 연구 결과들이 필요할 가능성이 높다. 전향적 연구들을 통한 연구 결과까지 확보하게 된다면 실제 딥러닝 알고리즘이 유방암 선별검사에 어떠한 긍정적인 결과를 미칠 수 있는지에 대한 근거를 더욱 탄탄하게 마련할 수 있을 것이다.

딥러닝 알고리즘의 유방촬영술 적용 사례

유방암 진단 보조 소프트웨어

유방촬영술 자료에 대한 인공지능 기술 적용의 가장 대표적인 적용 사례는 딥러닝 기반 CAD의 개발이다(44-48). 딥러닝 알고리즘은 많은 수의 데이터를 이용하여 유방촬영술에서 유방암이 의심되는 위치를 검출하도록 학습된다. 이렇게 학습된 딥러닝 알고리즘은 영상 내에서 유방암이 의심스러운 정도에 대한 확신도 값을 출력할 수 있으며, 학습 방식에 따라서 검사별로, 이미지(view)별로, 또는 픽셀 수준에서 얻을 수 있다. 대부분의 딥러닝 알고리즘의 분석 결과는 병변 또는 검사별 점수와 함께 이상 소견에 대한 위치 정보를 포함하는 형태로 제공된다. 병변의 위치에 대한 표기는 고전적인 CAD처럼 선(line)을 이용해서 병변의 주변부를 그려주거나, 다양한 색으로 병변의 위치를 표기하기도 한다(Fig. 3).

유방촬영술을 판독하는 의사는 원본 영상을 모두 확인하고 딥러닝 기반 CAD의 분석 결과를 확인하는 방식으로 사용하거나, 원본 영상을 확인하면서 딥러닝 기반 CAD의 분석 결과를 확인하는

Fig. 3. Visualization methods of deep learning models. The suspected region is highlighted using a color-map (left) or contour lines (right).

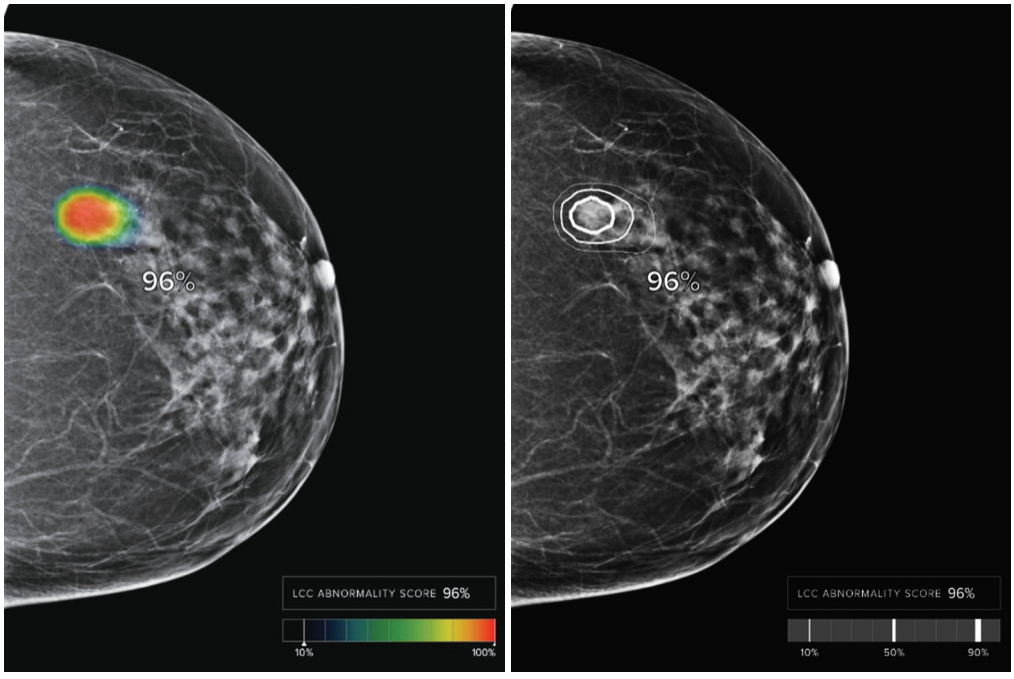


Table 1. Summary of Results for AI Applications in Digital Mammography

References	Comparison	Cases	N*	Sensitivity (%)		Specificity (%)		ROC AUC	
				Test	Control	Test	Control	Test	Control
Rodríguez-Ruiz et al. (2019) (45)	Reader study: Reader + AI (test) vs. Reader (control)	Cancer 100 Non-cancer 140	14	86	83	79	77	0.89	0.87
Wu et al. (2020) (48)	Reader study: Reader + AI (test) vs. Reader (control)	Cancer 62 Non-cancer 658	14	-	-	-	-	0.891	0.876
McKinney et al. (2020) (47)	Historical comparison AI (test) vs. Original report (control)	Total = 25856, cancer = 414 (UK) Total = 3097, cancer = 686 (US)		65 (UK) 56 (US)	63 (UK) 48 (US)	94 (UK) 87 (US)	93 (UK) 80 (US)	0.889 (UK) 0.811 (US)	-
	Reader study: AI (test) vs. Reader (control)	Cancer 113 Non-cancer 352	6	-	-	-	-	0.740	0.625
Kim et al. (2020) (46)	Reader study: Reader + AI (test) vs. Reader (control)	Cancer 160 Non-cancer 160	14	75	85	72	75	0.881	0.81
Salim et al. (2020) (43)	Historical comparison AI (test) vs. Original report (control)	Cancer 739 Non-cancer 8066		82 (AI-1)† 67 (AI-2) 67 (AI-3)	77	96.6 (AI-1) 96.6 (AI-2) 96.6 (AI-3)	97	-	-

*Number of readers.
†Three AI algorithms (AI-1, AI-2, AI-3) were tested.
AI = artificial intelligence, AUC = area under the curve, ROC = receiver operation characteristic curve

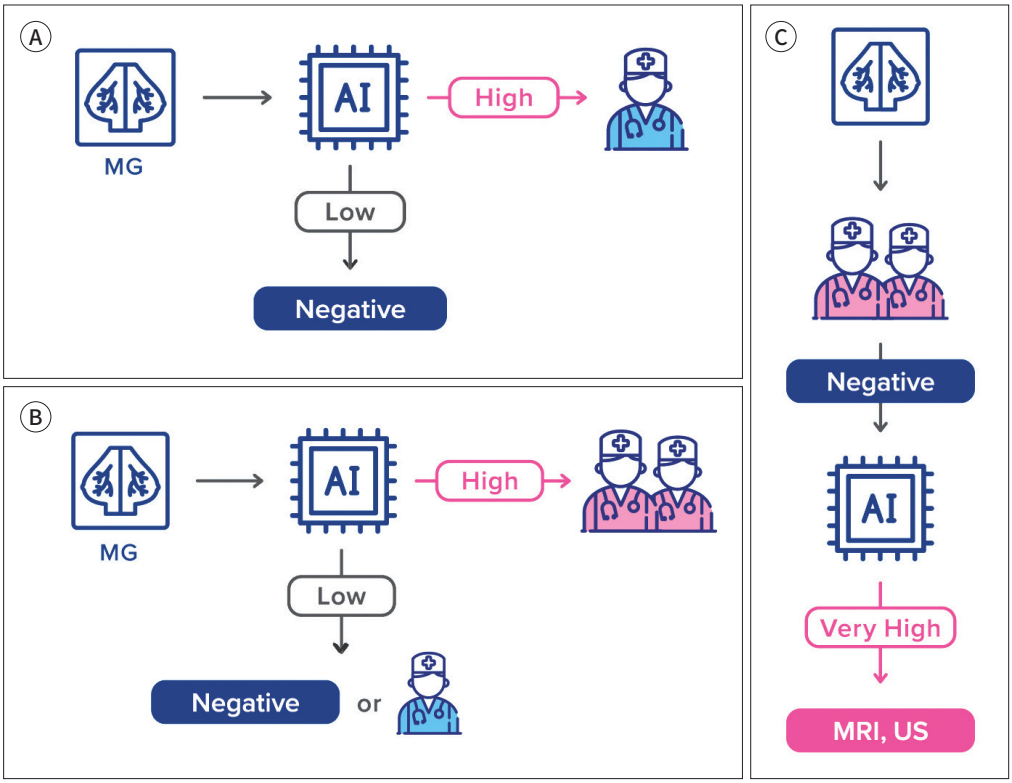
방식으로 사용할 수 있다. 최근 몇 년 사이에 출간된 리더스터디 연구 결과들에서 딥러닝 기반 CAD의 도움을 받았을 때 영상의학과 전문의들의 판독 성능이 향상되는 것을 확인할 수 있다(Table 1) (45-47). 다만 대부분의 연구에서 성능 평가의 척도로 예민도 또는 area under the receiver operating characteristic (이하 AUROC)를 이용하기 때문에, AUROC 수치의 향상이 실제 유방암 검진 환경에서 유방암 검출률 및 소환율에 어떠한 영향을 줄 것인지에 대해서는 추가 연구들이 더욱 시행되어야 한다. 정확도의 향상뿐만 아니라, 딥러닝 기반 CAD 결과를 함께 보면서 판독하면 판독 정확도는 향상되면서 판독시간은 오히려 감소한다는 연구 결과들이 있으며, 특히 이러한 효과는 breast tomosynthesis를 이용한 연구들에서 여러 차례 보고되었다(49-52). 이상적으로 딥러닝 알고리즘의 성능이 더욱 고도화될수록 딥러닝 기반 CAD 사용에 의한 판독 정확도 및 효율성 향상은 둘 다 더욱 개선이 될 수 있을 것으로 기대된다.

검진 유방촬영술의 분류(Triage)

딥러닝 알고리즘 분석 결과를 기반으로 유방암 검진 목적으로 촬영된 유방촬영술을 효과적으로 분류(triage) 하려는 시도들이 연구 수준에서 논의되고 있다. 딥러닝 알고리즘의 성능이 이전의 CAD에 비해서 향상되면서, 진단보조 외에도 유방암 검진을 효율적으로 운영하는 데 이를 이용하려는 시도들이 있다. 비록 국가마다 차이는 있지만, 검진 대상자 중에서 유방암의 비율이 1%에 미치지 못한다는 것을 감안하면(53), 유방암 검진 코호트 자료를 이용해서 가능성을 확인해 볼 만한 시도이다. 특히 이러한 움직임은 검진 유방촬영술을 두 명의 의사가 판독을 하는 double reading 시스템을 운영하는 유럽의 국가들에서 논의되고 있으며, 특히 영국처럼 유방영상의학과 전문의가 부족한 국가에서 많은 관심을 가진다(54).

딥러닝 알고리즘 분석 결과를 이용해서 검진 유방촬영술을 분류하는 다양한 제안들이 있다. 첫 번째로는 역치(threshold) 값을 기준으로 점수가 낮은 그룹에 대해서는 자동으로 음성(negative) 판정을 내리고, 점수가 높은 그룹의 검사들에 대해서는 의사가 판독을 수행하는 방식을 고려할 수 있다(Fig. 4A) (42). 이 같은 분류를 위해서 연구자들은 매우 높은 수준의 예민도를 확보하는 역치 값을 선정한다. 스웨덴의 유방암 검진 자료를 이용한 연구에서는 검진에서 의사에 의해 발견된 유방암들을 한 개도 놓치지 않을 수 있는 가장 높은 값을 역치 값으로 지정하고, 해당 점수보다 낮은 검사들을 모두 자동으로 음성 판정 처리를 할 경우 전체 검사 중에서 약 60% 검사를 자동으로 음성 판정을 내릴 수 있다고 보고했다(42). 다른 연구에서는 전체 검진 검사의 양을 절반으로 줄일 수 있는 역치 값을 기준으로 해당 역치 값보다 낮은 검사들을 모두 음성 처리를 할 경우, 전체 유방암 케이스들 중에서 8개(11.3%)의 유방암을 놓치게 된다는 시뮬레이션 결과를 보고했다(55). 미국의 검진 코호트 자료에서도 다른 알고리즘을 이용하여 자동으로 음성 판정을 내리는 시뮬레이션 연구가 시행되었는데, 전체 검사의 20%를 줄이더라도 민감도 측면에서는 열등하지 않다(non-inferior)는 것을 보여주었다(56). 이와 같은 연구 결과들은 딥러닝 분석 결과를 기준으로 의사의 판독 여부를 결정할 수 있다는 가능성을 시사하지만, 딥러닝 알고리즘이 단독으로 판독을 수행하는 것에 대해서 FDA 등의 규제기관의 허가를 받은 사례는 존재하지 않기 때문에, 실제 이러한 형태로 사용되기가 어려우며, 알고리즘 단독으로 판단한 것에 대한 책임을 누가 어떻게 질 것인가에

Fig. 4. Diagram illustrates potential scenarios for triaging mammograms in breast cancer screening. In the standard scenario, radiologists read all mammograms.
A. In a rule-out scenario, radiologists only read mammograms above a rule-out threshold.
B. In a double reading scenario, mammograms below a certain threshold are read by one radiologist and mammograms above the threshold are interpreted by two radiologists.
C. In the rule-in scenario, mammograms are triaged into an enhanced assessment when the score is above a rule-in threshold (after negative double reading by radiologists).
AI = artificial intelligence, MG = mammography



대한 윤리적인 문제들이 충분히 논의가 되어야 이러한 가설들이 실제 사용될 수 있을지에 대한 논의가 가능할 것이다.

두 번째로 시도되는 분류 방식은 주로 두 명이 판독을 하는 유럽에서 연구되는 것으로서, 딥러닝 분석 결과 유방암 관련 점수가 낮은 유방촬영술에 대해서는 두 명이 아닌 한 명의 의사가 단독으로 판독을 하거나 자동으로 음성으로 처리하는 방법이다(Fig. 4B). 이는 딥러닝 알고리즘이 낮은 점수로 평가한 검사들에서의 유방암 유병률은 유방암 검진 코호트의 그것보다 낮기 때문에, 두 명의 의사가 판독을 시행할 필요 없이 한 명의 의사가 판독을 하여도 유방암 검진의 성능이 저하되지 않는다는 가설이다. 이러한 방법에 대해서는 현재 연구들이 진행 중이며, 관련 연구 결과들이 근 시일 내에 출간될 것으로 생각된다. 이제까지는 딥러닝 알고리즘 분석 결과에서 유방암의 위험도가 낮은 그룹에 대한 효율성 향상 관점에서의 적용 사례에 대해서 소개했다.

반대로 두 명의 의사의 판독 결과가 음성인 케이스들 중에서, 딥러닝 분석 점수가 매우 높은 일부에 대해서 초음파나 MRI와 같은 추가검사를 시행하는 방식도 고려해볼 수 있다(Fig. 4C). 비록 두 명의 의사가 판독을 하더라도 유방암을 놓칠 수 있으며, 이처럼 잠재적으로 놓칠 수 있는 유방

암들을 딥러닝 알고리즘이 검출을 할 수도 있다. 최근에 발표된 후향적 시뮬레이션 연구에 따르면 두 명의 의사 판독은 음성이었지만 딥러닝 점수가 매우 높은 상위 1%의 수검자들의 추적검사 결과를 확인해보니, 이들 중에서 약 10%에서 간격암 또는 다음 검진에서 유방암으로 확인이 되었다고 한다(42). 이는 두 의사의 판독은 음성이지만 딥러닝 알고리즘 분석 결과에서 유방암이 의심되는 수검자들을 고위험군으로 적절히 선정하면, 해당 집단에 대해서는 초음파나 MRI와 같은 추가 검사를 권유함으로써 유방암을 조기에 추가로 검출할 수 있으며, 이러한 사용법에 대해서 가격 대비 효과 측면에서 경쟁력도 주장할 수 있을 것으로 예상된다.

유방 치밀도의 평가

유방촬영술에서의 유방 치밀도는 유방암에 대한 독립적인 위험인자로서(57), 유방촬영술 판독 시 반드시 평가되어야 하는 항목이다. 유방 치밀도는 BI-RADS에 따라 평가되지만, 실제로는 판독 의사에 따라 평가 차이가 크다고 잘 알려져 있다. 이러한 문제를 해결하고자 유방 치밀도를 정량적으로 평가하기 위한 노력이 오랫동안 시도되었으나 의사와 유방 치밀도 평가 프로그램의 분석 결과에 대한 일치도가 높지 않은 문제가 있었다. 그래서 딥러닝 기술을 이용하여 기존의 유방 치밀도 분석 프로그램보다 더욱 의사와의 일치도가 높은 알고리즘을 개발하고자 하는 노력들이 시도되고 있다(58, 59). 딥러닝 기반의 치밀도 평가 알고리즘을 두 곳의 병원에서 실제 설치해서 사용해본 결과, 전체 검사 중 약 90~94%에서 딥러닝 알고리즘의 분석 결과를 수정 없이 그대로 판독문에 사용한 것으로 보고되었다(36). 비록 앞으로 더 많은 연구 결과가 필요하겠지만, 딥러닝 기반의 치밀도 평가 알고리즘이 실제 의료현장에서 널리 사용될 수 있는 가능성을 보여주었다.

유방암 위험도 예측 모델

최근에는 유방촬영술에서 유방 조직의 패턴 분석을 통해서 유방암 위험도를 예측하고자 하는 시도들이 소개되고 있다. 이러한 연구들은 유방 치밀도 외에도 유방실질 조직의 텍스처(texture) 패턴과 유방암의 발생 사이에 연관관계가 있을 것이라는 가정이 있으며, 치밀도는 이 패턴의 가장 단순화된 표현형으로써 치밀도 평가보다 더욱 높은 수준의 위험도 요인이 유방 패턴에 내재되어 있을 것으로 생각하는 것이다. 딥러닝을 이용하면 유방실질의 복잡한 패턴을 더욱 고차원적인 수준에서 분석이 가능하기 때문에, 유방암 발생전의 유방촬영술로 향후 5년 내에 유방암이 발생할 가능성에 대해서 딥러닝 기반의 예측 모형을 만들어보려는 시도들이 보고되고 있다(60, 61). 약 8만 건의 유방촬영술 자료를 이용하여 학습 및 내부 평가를 시행한 연구 결과에 따르면 딥러닝 모델의 5년 위험도 예측 AUROC가 0.68 (0.64~0.73)로 임상자료 기반의 위험도 모델인 Tyrer-Cuzick model (version 8)의 AUROC 0.62 (0.57~0.66) 보다 높았으며, 임상지표들과 영상기반의 딥러닝 모델의 결과를 모두 활용하면 AUROC가 0.7 (0.66~0.75)로 향상되었음을 입증했다. 비록 기존 유방암 예측 모델들에 비해서 월등한 차이를 보이지는 않았지만, 유방실질의 패턴이 유방 치밀도와는 독립적인 위험인자로 간주될 수 있는 가능성을 보여준 연구들로서, 앞으로 관련된 더욱 많은

연구들이 이루어질 것으로 기대된다. 유방실질의 패턴 정보를 이용한 단기(short-term) 위험도 예측 모형들의 개발에 대한 궁극적인 목적은 개인별 위험도에 따라 적절한 검사를 권유하거나 검진 주기를 권유하기 위함이다(62).

의료현장에서의 도입

유방촬영술에서 유방암의 검출을 위한 딥러닝 기반 CAD는 국내외에서 의료현장에서의 사용을 위한 허가를 받기 시작했으며, 점차 그 사용 범위를 넓혀가고 있다. 다만 현재는 보험수가를 적용받지 못하여, 병원에서 비용을 지불하는 방식으로 판매되고 있어 그 파급 속도가 빠르다고 볼 수 없다. 그러나 여러 상용화된 소프트웨어들이 전 세계의 여러 기관들에서 후향적 코호트 자료를 이용한 연구 결과들을 출간하기 시작했으며(43), 앞으로도 코호트 연구 결과들이 향후 몇 년간 지속적으로 보고될 것으로 보인다. 이 과정에서 다양한 상용 알고리즘들에 대해서 더욱 엄밀한 검증이 이루어지고, 여러 딥러닝 기반의 CAD들에 대한 '옥석 가리기'가 이루어질 것이다.

나아가 각 국가의 유방암 검진 방식에 맞는 전향적 연구들이 내년부터는 여러 곳에서 시행될 예정이며, 3~5년 후에는 이러한 연구들의 결과들이 결실을 맺기 시작할 것으로 예상된다. 특히 유럽의 유방암 선별검사의 경우는 한 검사를 두 명의 영상의학과 의사가 판독해야 하는 상황인데, 의사의 수마저 부족한 국가들이 많아, 두 번째 의사 역할을 딥러닝 알고리즘이 대신하는 방식을 검증하기 위한 전향적 연구가 논의되고 있다. 우리나라처럼 의사 한 명이 판독을 하는 국가들에서는 판독의사가 단독으로 판독을 하는 경우와 딥러닝 기반 CAD의 결과를 참고하여서 판독하는 경우를 비교하는 방식으로 전향적 연구들이 진행될 것이다. 그 외에도 최근 여러 연구들에서 소개된 다양한 분류(triage) 방식들 중 일부는 전향적 연구들을 통해서 검증이 될 예정이다. 이러한 임상연구 결과들을 통해 실제 의료현장에서의 임상적 가치를 입증하고 비용 대비 효과성까지 인정을 받는다면, 해당 국가의 보험수가를 받을 가능성이 높아질 것으로 기대된다.

영상의학과 전문의의 역할

유방영상 분야는 의료 인공지능 연구가 가장 활발하게 연구되는 분야이며, 앞으로도 많은 연구개발들을 통한 새로운 기회들이 창출될 가능성이 높다. 다른 장기 또는 영상 검사들에 비해서 감별진단해야 할 주요 질환들의 범주가 좁고, 유방암으로 한정하여 기술 개발 및 상용화를 하기 상대적으로 용이하기 때문이다. 따라서 이처럼 인공지능 관련 연구가 활발히 이루어지는 만큼 영상의학과 의사들이 더욱 관심을 가져야 할 분야가 유방영상 분야에서의 인공지능 활용이라고 할 수 있다. 현재 의료 현장에 도입되기 시작한 제품들은 대부분 딥러닝 기반 CAD들로써, 과거의 CAD에 비해서는 그 성능이 일반적으로 크게 향상되었다고 알려져 있다. 인공지능 기술의 특성상 향후 데이터가 더욱 축적될수록 그 성능은 더욱 향상될 것으로 예상되며, 따라서 유방촬영술을 판독하는 의사라면 이 분야에 지속적인 관심을 가질 필요가 있다.

또한 앞으로 유방암 위험도 예측 모형들과 같이 기존에 의사의 눈으로 하기 어려운 작업들에 대

한 기술 개발도 활발히 이루어질 것이며, 상용화가 가능한 수준의 예측력을 확보하게 되면 개인별 유방암 선별검사 워크플로우가 실현 가능해질 수도 있다. 따라서 영상의학과 의사 입장에서 이러한 기술을 어떻게 활용하는 것이 우리나라의 실정에 맞는지, 그리고 수검자 또는 환자들에게 도움을 줄 수 있을지에 대한 전문가 집단으로서의 지속적인 관심과 목소리가 필요하다.

결론

최근 다양한 분야에서 각광받고 있는 인공지능은 특히 유방촬영술의 분석에 있어서 활발히 연구되고 있으며, 상용 제품들도 병원에 도입되기 시작하였다. 딥러닝을 도구로 잘 활용하기 위해서는 딥러닝을 이용한 의료영상 기술의 개발에 대한 대략적인 이해도가 필요하며, 개발 과정 및 임상적 평가 과정에 대한 지식도 필요할 것이다. 뿐만 아니라 딥러닝을 이용하여 어떠한 문제들을 풀기 위한 노력들이 시도되고 있는지를 이해하고, 마지막으로 이들 기술을 어떻게 평가하고 사용하는 것이 국민건강증진에 도움이 될 것인지에 대한 전문가 집단으로서의 관심이 필요한 때이다.

Author Contributions

Conceptualization, all authors; investigation, all authors; methodology, K.K.H.; project administration, K.K.H.; resources, all authors; supervision, K.K.H.; visualization, all authors; writing—original draft, all authors; and writing—review & editing, all authors.

Conflicts of Interest

The authors have no potential conflicts of interest to disclose.

REFERENCES

1. International Agency for Research on Cancer. GCO Global Cancer Observatory. Available at. https://gco.iarc.fr/today/data/factsheets/cancers/20-Breast-fact-sheet.pdf. Published 2018. Accessed Jan 20, 2020
2. Hong S, Won YJ, Park YR, Jung KW, Kong HJ, Lee ES; Community of Population-Based Regional Cancer Registries. Cancer statistics in Korea: incidence, mortality, survival, and prevalence in 2017. *Cancer Res Treat* 2020;52:335-350
3. Ministry of Health and Welfare. 2017 cancer registration statistics. Available at. https://www.index.go.kr/potal/main/EachDtlPageDetail.do?idx_cd=2770. Accessed Nov 29, 2020
4. Korean Statistical Information Service. Health examination statistics, 2018 status of the number of cancer screening examinations by age and gender. Available at. https://kosis.kr/statHtml/statHtml.do?orgId=350&tblId=DT_35007_N010&conn_path=I2. Published 2020. Accessed Nov 29, 2020
5. Lee K, Kim H, Lee JH, Jeong H, Shin SA, Han T, et al. Retrospective observation on contribution and limitations of screening for breast cancer with mammography in Korea: detection rate of breast cancer and incidence rate of interval cancer of the breast. *BMC Womens Health* 2016;16:72
6. Lehman CD, Arao RF, Sprague BL, Lee JM, Buist DS, Kerlikowske K, et al. National performance benchmarks for modern screening digital mammography: update from the Breast Cancer Surveillance Consortium. *Radiology* 2017;283:49-58
7. Majid AS, de Paredes ES, Doherty RD, Sharma NR, Salvador X. Missed breast carcinoma: pitfalls and pearls. *Radiographics* 2003;23:881-895
8. Ministry of Health and Welfare National Cancer Center. *Second revision of the quality guidelines of breast cancer screening.* Goyang: Ministry of Health and Welfare National Cancer Center 2018
9. Giger ML, Chan HP, Boone J. Anniversary paper: history and status of CAD and quantitative image analysis: the role of Medical Physics and AAPM. *Med Phys* 2008;35:5799-5820
10. Freer TW, Ulissey MJ. Screening mammography with computer-aided detection: prospective study of 12,860

patients in a community breast center. *Radiology* 2001;220:781-786
11. Warren Burhenne LJ, Wood SA, D'Orsi CJ, Feig SA, Kopans DB, O'Shaughnessy KF, et al. Potential contribution of computer-aided detection to the sensitivity of screening mammography. *Radiology* 2000;215:554-562
12. Oliver A, Freixenet J, Martí J, Pérez E, Pont J, Denton ER, et al. A review of automatic mass detection and segmentation in mammographic images. *Med Image Anal* 2010;14:87-110
13. Masotti M, Lanconelli N, Campanini R. Computer-aided mass detection in mammography: false positive reduction via gray-scale invariant ranklet texture features. *Med Phys* 2009;36:311-316
14. Hupse R, Samulski M, Lobbes M, den Heeten A, Imhof-Tas MW, Beijerinck D, et al. Standalone computer-aided detection compared to radiologists' performance for the detection of mammographic masses. *Eur Radiol* 2013;23:93-100
15. Lehman CD, Wellman RD, Buist DS, Kerlikowske K, Tosteson AN, Miglioretti DL. Diagnostic accuracy of digital screening mammography with and without computer-aided detection. *JAMA Intern Med* 2015;175:1828-1837
16. Fenton JJ, Abraham L, Taplin SH, Geller BM, Carney PA, D'Orsi C, et al. Effectiveness of computer-aided detection in community mammography practice. *J Natl Cancer Inst* 2011;103:1152-1161
17. Fenton JJ, Taplin SH, Carney PA, Abraham L, Sickles EA, D'Orsi C, et al. Influence of computer-aided detection on performance of screening mammography. *N Engl J Med* 2007;356:1399-1409
18. Ikeda DM, Birdwell RL, O'Shaughnessy KF, Sickles EA, Brenner RJ. Computer-aided detection output on 172 subtle findings on normal mammograms previously obtained in women with breast cancer detected at follow-up screening mammography. *Radiology* 2004;230:811-819
19. American College of Radiology. *ACR BI-RADS atlas: breast imaging reporting and data system.* 5th ed. Reston: American College of Radiology 2013
20. Mercado CL. BI-RADS update. *Radiol Clin North Am* 2014;52:481-487
21. Rao AA, Feneis J, Lalonde C, Ojeda-Fournier H. A pictorial review of changes in the BI-RADS fifth edition. *Radiographics* 2016;36:623-639
22. Krizhevsky A, Sutskever I, Hinton GE. ImageNet classification with deep convolutional neural networks. *Communications of the ACM* 2017;60:84-90
23. He K, Zhang X, Ren S, Sun J. Deep residual learning for image recognition. Proceedings of the 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR); 2016 Jun 27-30; Las Vegas, NV, USA: IEEE; 2016:770-778
24. Simonyan K, Zisserman A. Very deep convolutional networks for large-scale image recognition. *ArXiv preprint* 2014;arXiv:1409.1556
25. Hamidinekoo A, Denton E, Rampun A, Honnor K, Zwiggelaar R. Deep learning in mammography and breast histology, an overview and future trends. *Med Image Anal* 2018;47:45-67
26. Rabidas R, Midya A, Chakraborty J. Neighborhood structural similarity mapping for the classification of masses in mammograms. *IEEE J Biomed Health Inform* 2018;22:826-834
27. Oyelade ON, Ezugwu AES. A state-of-the-art survey on deep learning methods for detection of architectural distortion from digital mammography. *IEEE Access* 2020;8:148644-148676
28. Samala RK, Heang-Ping Chan, Hadjiiski L, Helvie MA, Richter CD, Cha KH. Breast cancer diagnosis in digital breast tomosynthesis: effects of training sample size on multi-stage transfer learning using deep neural nets. *IEEE Trans Med Imaging* 2019;38:686-696
29. Sun L, Wang J, Hu Z, Xu Y, Cui Z. Multi-view convolutional neural networks for mammographic image classification. *IEEE Access* 2019;7:126273-126282
30. Park SH. Artificial intelligence in medicine: beginner's guide. *J Korean Soc Radiol* 2018;78:301-308
31. 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
32. Park SH, Choi J, Byeon JS. Key principles of clinical validation, device approval, and insurance coverage decisions of artificial intelligence. *J Korean Med Assoc* 2020;63:696-708
33. Mutasa S, Sun S, Ha R. Understanding artificial intelligence based radiology studies: what is overfitting? *Clin Imaging* 2020;65:96-99
34. Ghassemi M, Naumann T, Schulam P, Beam AL, Chen IY, Ranganath R. Practical guidance on artificial intelligence for health-care data. *Lancet Digit Health* 2019;1:e157-e159

35. Wang X, Liang G, Zhang Y, Blanton H, Bessinger Z, Jacobs N. Inconsistent performance of deep learning models on mammogram classification. *J Am Coll Radiol* 2020;17:796-803
36. Dontchos BN, Yala A, Barzilay R, Xiang J, Lehman CD. External validation of a deep learning model for predicting mammographic breast density in routine clinical practice. *Acad Radiol* 2020 [in press] doi: https://doi.org/10.1016/j.acra.2019.12.012
37. Yamaguchi T, Inoue K, Tsunoda H, Uematsu T, Shinohara N, Mukai H. A deep learning-based automated diagnostic system for classifying mammographic lesions. *Medicine (Baltimore)* 2020;99:e20977
38. Milea D, Singhal S, Najjar RP. Artificial intelligence for detection of optic disc abnormalities. *Curr Opin Neurol* 2020;33:106-110
39. Leconte I, Feger C, Galant C, Berlière M, Berg BV, D'Hoore W, et al. Mammography and subsequent whole-breast sonography of nonpalpable breast cancers: the importance of radiologic breast density. *AJR Am J Roentgenol* 2003;180:1675-1679
40. U.S. Food and Drug Administration. *Computer-assisted detection devices applied to radiology images and radiology device data - premarket notification [510(k)] submissions.* Silver Spring: Food and Drug Administration 2012
41. Retson TA, Eghtedari M. Computer-aided detection/diagnosis in breast imaging: a focus on the evolving FDA regulations for using software as a medical device. *Curr Radiol Rep* 2020;8:1-7
42. Dembrower K, Wåhlin E, Liu Y, Salim M, Smith K, Lindholm P, et al. Effect of artificial intelligence-based triaging of breast cancer screening mammograms on cancer detection and radiologist workload: a retrospective simulation study. *Lancet Digit Health* 2020;2:e468-e474
43. Salim M, Wåhlin E, Dembrower K, Azavedo E, Foukakis T, Liu Y, et al. External evaluation of 3 commercial artificial intelligence algorithms for independent assessment of screening mammograms. *JAMA Oncol* 2020;6:1581-1588
44. Kooi T, Litjens G, van Ginneken B, Gubern-Mérida A, Sánchez CI, Mann R, et al. Large scale deep learning for computer aided detection of mammographic lesions. *Med Image Anal* 2017;35:303-312
45. Rodríguez-Ruiz A, Krupinski E, Mordang JJ, Schilling K, Heywang-Köbrunner SH, Sechopoulos I, et al. Detection of breast cancer with mammography: effect of an artificial intelligence support system. *Radiology* 2019;290:305-314
46. Kim HE, Kim HH, Han BK, Kim KH, Han K, Nam H, et al. Changes in cancer detection and false-positive recall in mammography using artificial intelligence: a retrospective, multireader study. *Lancet Digit Health* 2020;2:e138-e148
47. McKinney SM, Sieniek M, Godbole V, Godwin J, Antropova N, Ashrafian H, et al. International evaluation of an AI system for breast cancer screening. *Nature* 2020;577:89-94
48. Wu N, Phang J, Park J, Shen Y, Huang Z, Zorin M, et al. Deep neural networks improve radiologists' performance in breast cancer screening. *IEEE Trans Med Imaging* 2020;39:1184-1194
49. Conant EF, Toledano AY, Periaswamy S, Fotin SV, Go J, Boatsman JE, et al. Improving accuracy and efficiency with concurrent use of artificial intelligence for digital breast tomosynthesis. *Radiol Artif Intell* 2019;1:e180096
50. Pacilè S, Lopez J, Chone P, Bertinotti T, Grouin JM, Fillard P. Improving breast cancer detection accuracy of mammography with the concurrent use of an artificial intelligence tool. *Radiology: Artificial Intelligence* 2020;2:e190208
51. Balleyguier C, Arfi-Rouche J, Levy L, Toubiana PR, Cohen-Scali F, Toledano AY, et al. Improving digital breast tomosynthesis reading time: a pilot multi-reader, multi-case study using concurrent Computer-Aided Detection (CAD). *Eur J Radiol* 2017;97:83-89
52. Chae EY, Kim HH, Jeong JW, Chae SH, Lee S, Choi YW. Decrease in interpretation time for both novice and experienced readers using a concurrent computer-aided detection system for digital breast tomosynthesis. *Eur Radiol* 2019;29:2518-2525
53. Lee EH, Kim KW, Kim YJ, Shin DR, Park YM, Lim HS, et al. Performance of screening mammography: a report of the alliance for breast cancer screening in Korea. *Korean J Radiol* 2016;17:489-496
54. Gulland A. Staff shortages are putting UK breast cancer screening "at risk," survey finds. *BMJ* 2016;353:i2350
55. Rodriguez-Ruiz A, Lång K, Gubern-Merida A, Teuwen J, Broeders M, Gennaro G, et al. Can we reduce the workload of mammographic screening by automatic identification of normal exams with artificial intelligence? A feasibility study. *Eur Radiol* 2019;29:4825-4832

56. Yala A, Schuster T, Miles R, Barzilay R, Lehman C. A deep learning model to triage screening mammograms: a simulation study. *Radiology* 2019;293:38-46
57. Carney PA, Miglioretti DL, Yankaskas BC, Kerlikowske K, Rosenberg R, Rutter CM, et al. Individual and combined effects of age, breast density, and hormone replacement therapy use on the accuracy of screening mammography. *Ann Intern Med* 2003;138:168-175
58. Lehman CD, Yala A, Schuster T, Dontchos B, Bahl M, Swanson K, et al. Mammographic breast density assessment using deep learning: clinical implementation. *Radiology* 2019;290:52-58
59. Mohamed AA, Berg WA, Peng H, Luo Y, Jankowitz RC, Wu S. A deep learning method for classifying mammographic breast density categories. *Med Phys* 2018;45:314-321
60. Yala A, Lehman C, Schuster T, Portnoi T, Barzilay R. A deep learning mammography-based model for improved breast cancer risk prediction. *Radiology* 2019;292:60-66
61. Dembrower K, Liu Y, Azizpour H, Eklund M, Smith K, Lindholm P, et al. Comparison of a deep learning risk score and standard mammographic density score for breast cancer risk prediction. *Radiology* 2020;294:265-272
62. Eriksson M, Czene K, Pawitan Y, Leifland K, Darabi H, Hall P. A clinical model for identifying the short-term risk of breast cancer. *Breast Cancer Res* 2017;19:29

유방촬영술에서 인공지능의 적용: 알고리즘 개발 및 평가 관점

김기환* · 이상협

유방촬영술은 유방암 검진 및 진단을 위한 기본적인 영상 검사이지만, 판독이 어려우며 높은 숙련도를 필요로 한다고 잘 알려져 있다. 이러한 어려움을 극복하기 위해 최근 몇 년 사이에 인공지능을 이용한 유방암 검출 알고리즘들이 활발히 연구되고 있다. 본 종설에서 저자는 고전적인 computer-aided detection 소프트웨어 대비 최근 많이 사용되는 딥러닝의 특징을 알아보고, 딥러닝 알고리즘의 개발 방법과 임상적 검증 방법에 대해서 기술하였다. 또한 딥러닝 기반의 검진 유방촬영술의 판독 방법 분류, 유방 치밀도 평가, 그리고 유방암 위험도 예측 모델 등을 위한 딥러닝 연구들도 소개하였다. 마지막으로 유방촬영술 관련 인공지능 기술들에 대한 영상의학과 전문의의 관심과 의견의 필요성을 기술하였다.

루닛, 서울, 한국