

유방영상에서 인공지능의 효과적인 활용 방안

함성원¹ · 배민선²

¹고려대학교 안산병원 의생명연구원, ²고려대학교 안산병원 영상의학과

서론

영상의학 분야에서 인공지능(artificial intelligence; 이하 AI) 기술은 새로운 진단 기법을 제공하며, 특히 영상 분석에서 판독의 정확도를 높이는 데 중요한 역할을 한다 (1). 최근 AI는 딥러닝(deep learning) 기반의 분할(segmentation)과 분류(classification) 기술을 통해 각종 질환의 양성 및 악성 병변을 구분하고, 놓치기 쉬운 미세한 병변까지 탐지(detection)하는 데 도움을 준다 (2, 3). AI를 활용하면 진단의 일관성이 높아지고 판독 시간은 단축되어, 의료진이 더욱 복잡한 사례에 집중할 수 있게 된다. 또한 AI와 임상 정보를 결합함으로써 보다 정밀하고 심층적인 진단 정보를 제공하며 진단의 고도화를 이끌어 내기도 한다.

이 종설에서는 유방 영상의학과 의사들이 AI 기술을 연구와 진단에 적용하는 방법을 보다 쉽게 이해하고 AI 연구를 시작할 수 있는 가이드 역할을 하고자 한다. 연구자들이 유방영상을 활용하여 다양한 연구에 AI를 쉽게 적용할 수 있도록 실질적인 정보를 제공하고자 한다.

1. 유방영상 연구에서 인공지능 활용의 다양성

유방 영상을 이용한 연구에서 AI가 적용되는 방식은 크게 분할(segmentation), 분류(classification), 탐지(detection) 등으로 나눌 수 있다. 유방 자기공명영상

(MRI)에서 종양의 경계와 크기를 자동으로 분할하거나 유방암 병변의 정확한 위치와 크기를 도출하고, 유방촬영술에서 석회화 병변의 정확한 위치와 경계를 추출하여 조직검사 및 후속 치료 계획에 중요한 정보를 제공하는 연구들이 있다 (5-7). 치밀 유방에서는 유방촬영술의 민감도가 감소하기 때문에 AI를 활용하여 유방촬영술에서 치밀한 정상 조직과 병변을 구분하고, 특히 유방암과 유사한 구조를 분할하는 연구가 수행되기도 한다. 그리고 유방촬영술 또는 초음파 영상을 이용하여 양성 및 악성 병변을 구분하는 모델을 학습하는 영상판독 연구, 위양성 판독을 줄이기 위해 병변의 악성도를 평가하는 연구, 유방 MRI에서 병변을 양성과 악성으로 분류하고 분석하는 연구 등이 있다 (8-11).

2. 인공지능 연구 시작을 위한 데이터 준비

AI 모델이 의료 영상처럼 복잡한 데이터를 학습할 때는 다양한 패턴을 충분히 이해해야 예측 정확도가 높아지기 때문에 많은 양의 레이블링(labeling)된 골드 스탠다드(gold standard) 데이터가 필요하다. 그러나 의료 영상의 레이블링 작업은 고도의 전문 지식과 경험이 요구되며, 다량의 데이터를 레이블링 하는 데 상당한 시간과 비용이 소요된다. 이를 해결하기 위해 최근에는 ‘휴먼인더루프(human-in-the-loop, HITL)’ 학습 방식을 도입하여 효율적으로 데이터를 구축하는 방법을 사용하고 있다 (12, 13). HITL 방식에서는 먼저 초기 학습을 위해 소량의 데이터만 수동으로 레이블링 한다. 예를 들어, 수백 개의 데이터를 수집하였다면 처음부터 수백 개 데이터 전부를 레이블링 할 필요 없이 약 20~30개의 데이터만으로 초기 모델 학습을 시작할 수 있다. 이 소량의 레이블링 된 데이터로 학습된 모델이 생성되면, 모델은 학습되지 않은 새로운 영상 데이터를 예측하고 전문가가 그 예측 결과를 검

이 종설은 정부(과학기술정보통신부)의 재원으로 한국연구재단의 지원을 받아 이루어진 것입니다(RS-2024-00361049, 2022R1A2C109302513).

통신저자: 배민선

(15355) 경기도 안산시 단원구 적금로 123,

고려대학교 안산병원 영상의학과

Tel. +82-31-412-5228 Fax. +82-31-412-5224

E-mail: minsunb@gmail.com

토한다. 검토 과정에서 예측이 잘못되었거나 보완이 필요한 경우 전문가가 수동으로 일부분 수정하여 정확한 레이블을 완성한다. 이렇게 수정된 레이블은 골드 스탠다드 레이블이 되며, 이후 레이블링 된 데이터를 추가하여 모델이 재학습을 하면서 모델의 성능이 점진적으로 개선된다. 이 과정은 반복적으로 진행되어, 모델이 예측한 결과를 전문가가 검토하고 수정하는 피드백 루프를 통해 점점 더 많은 고품질의 데이터를 빠른 시간 안에 효율적으로 구축하게 된다. 레이블링 소프트웨어로는 LabelMe, 3D Slicer, ITK-Snap 등이 있으며, 각 소프트웨어의 특성에 따라 적합한 도구를 선택하여 사용할 수 있다 (14-16). 특히 3D Slicer의 Level Tracing 기법은 분할 레이블 시 병변의 경계선을 반자동으로 빠르고 정확하게 추출할 수 있어 추천할 만하다 (그림 1).

3. 모델 선택

분할, 분류, 탐지 등의 AI 모델은 각각의 목적과 구조에

따라 서로 다른 레이어(layer) 구성을 갖는다. 분할 모델은 영상의 픽셀별 예측을 수행하여 각 픽셀에 대해 정상과 비정상을 구분해 특정 부위를 정확하게 분할한다. U-Net은 의료 영상에서의 세밀한 분할을 위해 개발된 대표적인 네트워크이다. 특히 nnU-Net은 U-Net 구조를 기반으로 다양한 데이터셋에 맞게 자동으로 최적화되어 데이터가 적더라도 좋은 성능을 보이며 병변의 정확한 경계 추출에 용이하다 (17, 18). 분류 모델은 이미지 전체를 하나의 클래스 값으로 요약해 특정 카테고리로 분류한다. 예를 들어, 영상이 양성인지 악성인지, 정상인지 비정상인지 예측하며 최종적으로 하나의 클래스 확률을 출력한다. 대표적으로 ResNet, DenseNet, InceptionV3 등의 분류 네트워크가 있다 (19-21). 탐지 모델은 이미지 내 여러 객체(예: 병변)의 위치를 찾고, 각 객체의 종류를 분류하는 작업을 수행한다. 이를 위해 탐지 모델은 바운딩 박스(bounding box) 좌표를 예측하는 레이어를 포함하여 객체의 위치를 지정하고, 바운딩 박스 내 객체가 어떤 클래스(예: 특정 병변)에 속하는지 판단하는 레이어도 포함하여, 탐지된 영

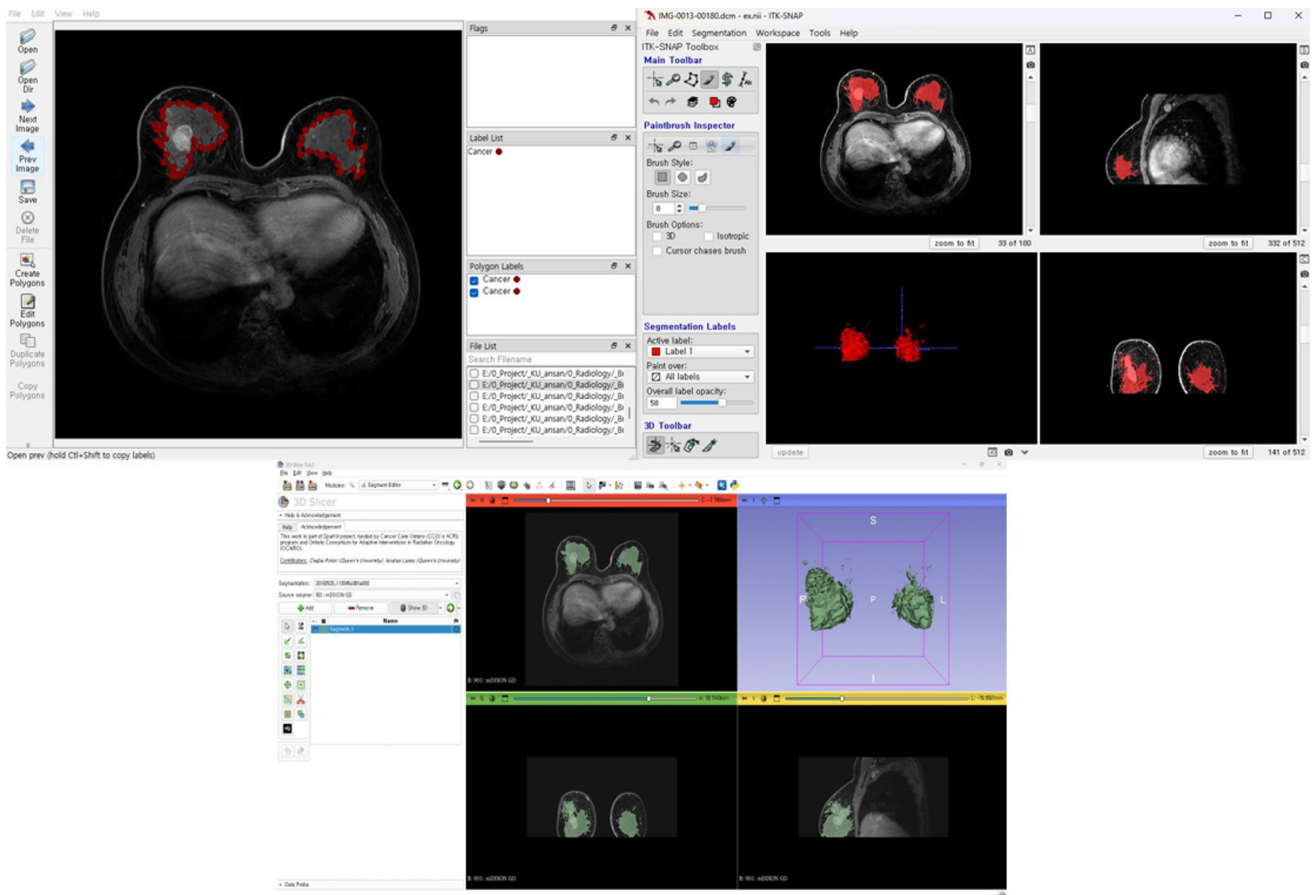


그림 1. 레이블링 소프트웨어. LabelMe(왼쪽 위), ITK-Snap(오른쪽 위), 3D Slicer(아래)

역이 어떤 종류의 병변인지도 분류할 수 있도록 설계한다. 관련 네트워크로는 YOLO, Faster R-CNN, RetinaNet 등이 있다 (22-24).

위에 열거된 네트워크들은 대표적이고 자주 사용되는 예시일 뿐, 여러 상황에 따라 다양한 네트워크가 존재하고 활용될 수 있다. 네트워크는 지속적으로 빠르게 업데이트되고 있으며, 레이어 구성과 변형을 통해 특정 목적에 맞게 얼마든지 수정도 가능하다. 그러므로 앞서 언급된 네트워크들에 국한되지 않고, 여러 옵션을 탐색하여 연구 목적에 맞는 최적의 모델을 선택할 것을 권장한다.

4. 모델 검증 및 평가

학습을 통해 인공지능 모델이 생성되었다면, 이 모델의 성능을 정확히 평가하는 것 또한 중요하다. 인공지능 모델의 성능을 평가할 때, 분할, 분류, 탐지 모델 각각의 특성에 맞는 지표를 사용하는 것이 중요하다. 다음은 각 모델에 적합한 주요 평가 지표들이다.

분할 모델의 성능을 평가하는데 주로 Intersection over Union(IoU)이 사용된다. 분할된 영역과 실제 병변 간의 겹치는 정도를 측정하여 평가하는 지표로, 두 영

역이 완벽히 겹치면 1에 가까운 값을 갖는다. 유사한 개념으로 Dice similarity coefficient(DSC)와 Jaccard similarity coefficient(JSC)도 많이 사용한다. Hausdorff distance(HD)는 예측된 경계와 실제 경계 간의 최대 거리를 측정해 병변 경계의 정밀도를 평가한다 (25). 분류 모델의 성능 평가는 예측한 결과와 실제 값의 관계를 두 개의 행과 두 개의 열의 표로 보여주는 혼돈 행렬(confusion matrix)을 사용한다 (그림 2). 이 행렬은 진양성(true positive), 진음성(true negative), 위양성(false positive), 위음성(false negative)으로 이루어지며, 이 값을 통해 정확도(accuracy), 민감도(sensitivity), 특이도(specificity), 정밀도(precision)를 계산할 수 있다. AUC-ROC(area under the receiver operating characteristic curve)로 모델이 양성과 음성을 얼마나 잘 구분하는지를 평가하며, 값이 1에 가까울수록 우수한 성능을 나타낸다 (26). 탐지 모델의 성능은 탐지된 병변 중 실제 병변의 비율인 정밀도(precision)와 실제 병변 중에서 모델이 탐지한 비율인 재현율(recall)로 평가한다. 평균 정밀도(average precision, AP)는 precision-recall 곡선의 아래 영역을 계산하여 정밀도와 재현율의 조합을 측정하는 지표로, 다양한 임계값에서 모델이 병변을 얼마나 일관

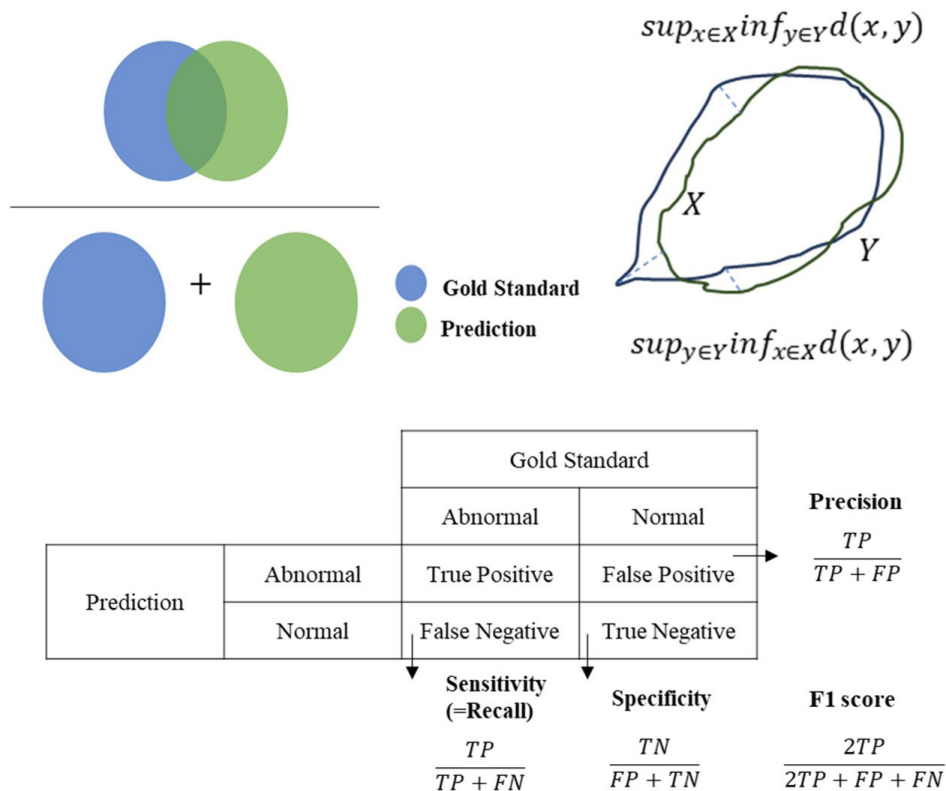


그림 2. AI 모델 성능 평가지표. IoU(왼쪽 위), Hausdorff distance(오른쪽 위), confusion matrix(아래)

되게 탐지하는지를 평가할 수 있다. F1 점수(F1 score)는 정밀도(precision)와 재현율(recall)의 조화 평균으로, 정밀도와 재현율의 균형을 고려하여 모델의 종합 성능을 평가할 수 있다 (27).

5. 추가적인 평가 및 해석 가능성

AI 모델의 성능 평가 외에도 해석 가능성(interpretability)은 임상 적용에서 중요한 요소로 부각되고 있다. 특히, CAM(class activation map)은 모델이 특정 예측을 수행할 때 어떤 영역에 주목했는지를 시각적으로 보여주는 유용한 도구이다 (그림 3). CAM은 분류 모델이나 탐지 모델에서 자주 사용되며, 이미지 상의 특정 영역이 예측 결과에 얼마나 기여했는지를 직관적으로 확인할 수 있다. 이를 통해 모델이 실제 병변에 기반하여 합리적인 결정을 내렸는지, 혹은 병변 외의 불필요한 영역에

주목했는지를 확인할 수 있다. 예를 들어, CAM을 통해 모델이 정상 조직을 병변으로 잘못 분류한 이유를 파악하거나, 병변 경계에 대한 추가적인 정보를 얻는 데 도움을 줄 수 있다. CAM은 분할 모델에서도 적용 가능하며, 이 경우 예측된 병변 영역과 실제 병변 영역 간의 일치도를 시각적으로 검증할 수 있다. 이러한 시각화는 모델 성능 평가에서 정량적 지표와 더불어 임상적 타당성을 확보하는 데 중요한 역할을 한다. 따라서, CAM을 활용한 시각적 검증은 정량적 지표와 함께 AI 모델의 성능을 종합적으로 이해하고, 결과를 임상적으로 신뢰할 수 있는 형태로 전달하는 데 기여한다.

결 론

이번 종설에서는 AI 기술이 유방영상의학 연구와 진단에 어떻게 적용될 수 있는지, 그리고 AI 연구를 수행하기

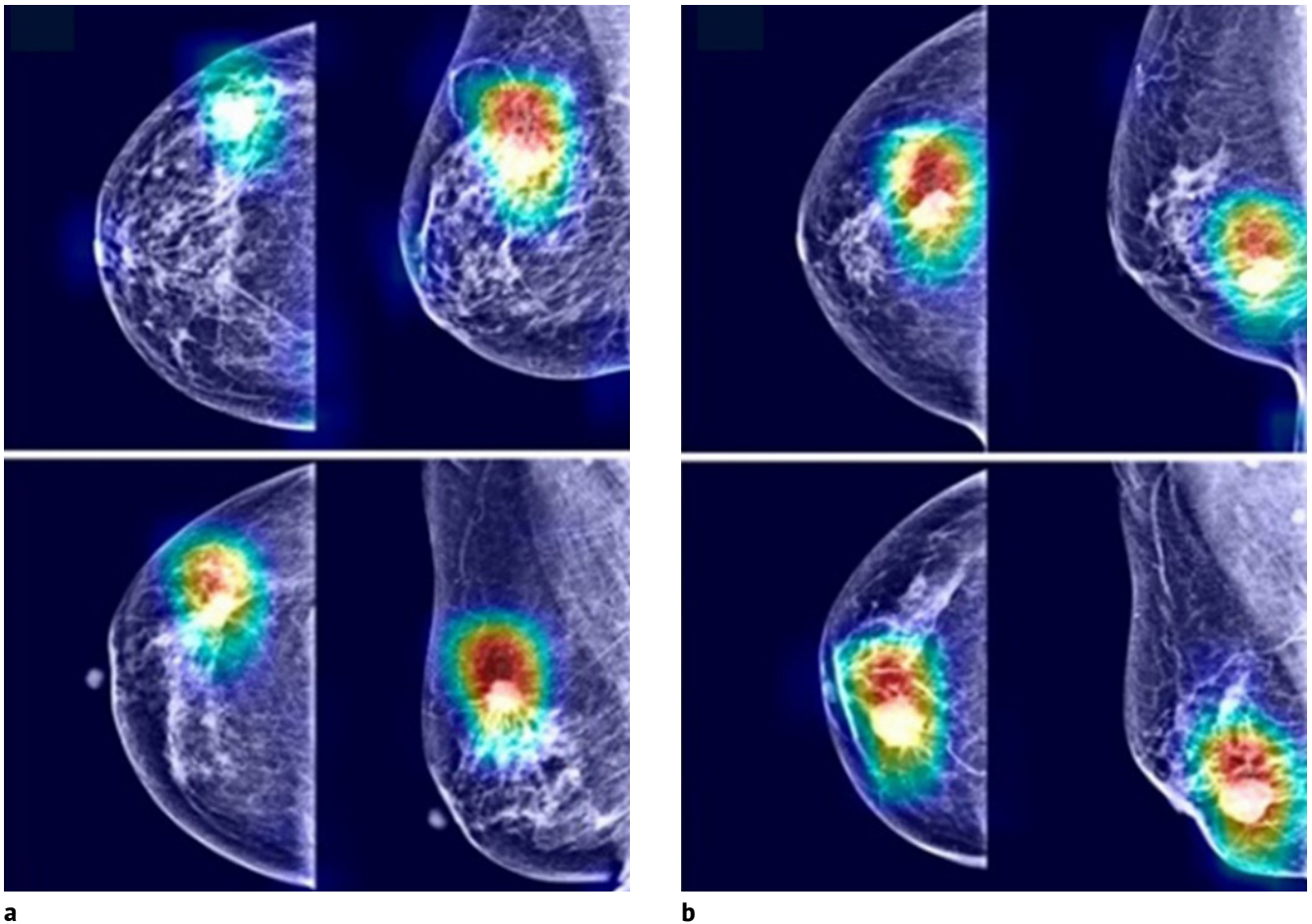


그림 3. 유방촬영술에서 CAM을 활용한 병변 시각화의 예(Suh et al. Automated breast cancer detection in digital mammograms of various densities via deep learning. J Pers Med 2020)

위해 필요한 구체적인 단계와 지표에 대해 다루었다. 앞으로 다양한 환경과 데이터셋에서 AI 모델의 성능을 검증하고 신뢰성을 높이기 위한 추가 연구는 지속적으로 이뤄질 필요가 있다. AI 기술이 유방영상의학 분야에서 안정적이고 신뢰할 수 있는 도구로 자리 잡을 수 있도록 지속적인 개선과 실증 연구가 이루어져야 한다. 이를 위해 AI 기술을 적극적으로 활용하여 유방영상 연구의 새로운 가능성을 탐색하고, 환자 맞춤형 의료를 위한 기반을 마련해 나가야 할 것이다. AI는 영상의학 분야의 발전을 촉진할 잠재력을 가지고 있으며, 유방영상 연구에도 큰 변화를 가져올 것으로 기대한다.

참 고 문 헌

1. Winston PH. Artificial intelligence. Addison-Wesley Longman Publishing Co., Inc., 1992.
2. LeCun Y, Bengio Y, Hinton G. Deep learning. *Nature* 2015;521(7553):436-444.
3. Jiang H, Diao Z, Shi T, et al. A review of deep learning-based multiple-lesion recognition from medical images: classification, detection and segmentation. *Comput Biol Med* 2023;157:106726.
4. Le EPV, Wang Y, Huang Y, et al. Artificial intelligence in breast imaging. *Clin Radiol* 2019;74(5):357-366.
5. Baccouche A, Garcia-Zapirain B, Olea CC, et al. Connected-UNets: a deep learning architecture for breast mass segmentation. *NPJ Breast Cancer* 2021;7(1):151.
6. Dembrower K, Liu Y, Azizpour H, et al. Comparison of a deep learning risk score and standard mammographic density score for breast cancer risk prediction. *Radiology* 2020;294(2):265-272.
7. Gerbasi A, Clementi G, Corsi F, et al. DeepMiCa: automatic segmentation and classification of breast MIncroCALcifications from mammograms. *Comput Methods Programs Biomed* 2023;235:107483.
8. Yue W, Zhang H, Zhou J, et al. Deep learning-based automatic segmentation for size and volumetric measurement of breast cancer on magnetic resonance imaging. *Front Oncol* 2022;12:984626.
9. Shen L, Margolies LR, Rothstein JH, et al. Deep learning to improve breast cancer detection on screening mammography. *Sci Rep* 2019;9(1):12495.
10. Kang J, Gwak J. Deep learning-based brain tumor classification in MRI images using ensemble of deep features. *Journal of the Korea Society of Computer and Information* 2021;26(7):37-44.
11. Li JW, Sheng DL, Chen JG, et al. Artificial intelligence in breast imaging: potentials and challenges. *Phys Med Biol* 2023;68(23).
12. Kim T, Lee KH, Ham S, et al. Active learning for accuracy enhancement of semantic segmentation with CNN-corrected label curations: evaluation on kidney segmentation in abdominal CT. *Sci Rep* 2020;10(1):366.
13. Zanzotto FM. Viewpoint: human-in-the-loop artificial intelligence. *Journal of Artificial Intelligence Research* 2019;64:243-252.
14. Russell BC, Torralba A, Murphy KP, et al. LabelMe: a database and web-based tool for image annotation. *Int J Comput Vis* 2008;77:157-173.
15. Kikinis R, Pieper SD, Vosburgh KG. 3D Slicer: a platform for subject-specific image analysis, visualization, and clinical support. *Intraoperative Imaging and Image-Guided Therapy*. Springer New York, 2013; 277-289.
16. Yushkevich PA, Gao Y, Gerig G. ITK-SNAP: an interactive tool for semi-automatic segmentation of multi-modality biomedical images. 2016 38th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBS). IEEE, 2016.
17. Ronneberger O, Fischer P, Brox T. U-net: convolutional networks for biomedical image segmentation. *Medical Image Computing and Computer-Assisted Intervention-MICCAI 2015: 18th International Conference, Munich, Germany, October 5-9, 2015, Proceedings, Part III* 18. Springer International Publishing, 2015.
18. Isensee F, Jaeger PF, Kohl SAA, et al. nnU-Net: a self-configuring method for deep learning-based biomedical image segmentation. *Nat Methods* 2021;18:203-211.
19. Sasha T, Almeida D, Lyman K. Resnet in Resnet: generalizing residual architectures. *ICLR 2016 Workshop Track*. arXiv:1603.08029v1
20. Zhang K, Guo Y, Wang X, et al. Multiple feature reweight DenseNet for image classification. *IEEE Access* 2019;7:9872-9880.
21. Xia X, Xu C, Nan B. Inception-v3 for flower classification. 2017 2nd International Conference on Image, Vision and Computing (ICIVC). IEEE, 2017.
22. Jiang P, Ergu D, Liu F, et al. A review of Yolo algorithm developments. *Procedia Computer Science*

- 2022;199:1066-1073.
23. Ren S, He K, Girshick R, et al. Faster R-CNN: towards real-time object detection with region proposal networks. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 2017;39:1137-1149.
24. Lin T.-Y, Goyal P, Girshick R, et al. Focal loss for dense object detection. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 2020;42(2):318-327.
25. Müller D, Iñaki S.-R, Kramer F. Towards a guideline for evaluation metrics in medical image segmentation. *BMC Res Notes* 2022;15:210.
26. Hossin M, Sulaiman MN. A review on evaluation metrics for data classification evaluations. *International Journal of Data Mining & Knowledge Management Process* 2015;5(2):1-11.
27. Fang J, Tao X, Tang Z, et al. Dataset, ground-truth and performance metrics for table detection evaluation. 2012 10th IAPR International Workshop on Document Analysis Systems. IEEE, 2012.