BRITISH JOURNAL OF HOSPITAL MEDICINE

REVIEW

Evaluating artificial intelligence for medical imaging: a primer for clinicians

Shivank Keni[1,2]

Author details can be found at the end of this article

Correspondence to:
Shivank Keni (shivank.keni2@nhs.scot)

Abstract

Artificial intelligence has the potential to transform medical imaging. The effective integration of artificial intelligence into clinical practice requires a robust understanding of its capabilities and limitations. This paper begins with an overview of key clinical use cases such as detection, classification, segmentation and radiomics. It highlights foundational concepts in machine learning such as learning types and strategies, as well as the training and evaluation process. We provide a broad theoretical framework for assessing the clinical effectiveness of medical imaging artificial intelligence, including appraising internal validity and generalisability of studies, and discuss barriers to clinical translation. Finally, we highlight future directions of travel within the field including multimodal data integration, federated learning and explainability. By having an awareness of these issues, clinicians can make informed decisions about adopting artificial intelligence for medical imaging, improving patient care and clinical outcomes.

Key words: Artificial intelligence; Deep learning; Machine learning; Medical imaging; Radiomics

Submitted: 5 May 2024; Revised: 24 June 2024; Accepted: 1 July 2024

Introduction

Artificial intelligence (AI) is being increasingly investigated across myriad areas in medical imaging, including image analysis, diagnostic accuracy, and workflow efficiency. However, the effective integration of AI into medical practice requires clinicians to understand its capabilities and limitations. This paper provides a practical guide for clinicians on how to evaluate AI technologies in medical imaging, beginning with an overview of key clinical use cases. We cover foundational concepts in machine learning (ML), evaluation frameworks, barriers to clinical translation and future directions. Understanding these issues will enable clinicians to make informed decisions about adopting AI tools for medical imaging, improving patient care and clinical outcomes.

Applications in medical imaging

Virtually no area of medical imaging has remained unexplored when it comes to the development of AI tools. Clinical use cases are highly diverse. An exhaustive overview of AI applications in medical imaging is beyond the scope of this paper and a number of excellent reviews already exist, covering historical context, technical factors and clinical applications (Hosny et al, 2018; Barragán-Montero et al, 2021; Castiglioni et al, 2021; Varoquaux and Cheplygina, 2022; Najjar, 2023). Nonetheless, we briefly review key use cases relevant to the practising clinician. A visual summary of these is shown in Figure 1.

Detection and classification

Diagnostic errors are unfortunately relatively common occurrences, with error rates ranging from 3%–5% for all imaging studies (Itri et al, 2018). Diagnostic errors account for 40,000–80,000 deaths per year in the United States and 75% of malpractice lawsuits in radiology; they are estimated to cost $38 billion annually (Lee et al, 2013). Unsurprisingly, there has been considerable interest in leveraging AI tools to assist with detection and classification of abnormalities on medical images. The idea that machines can help diagnostic decision-making is not new. Computer-assisted diagnosis (CAD) has been explored for over two decades and traditional CAD software aims to identify the same features that radiologists

How to cite this article:
Keni S. Evaluating artificial intelligence for medical imaging: a primer for clinicians. Br J Hosp Med. 2024. https://doi.org/10.12968/hmed.2024.0312

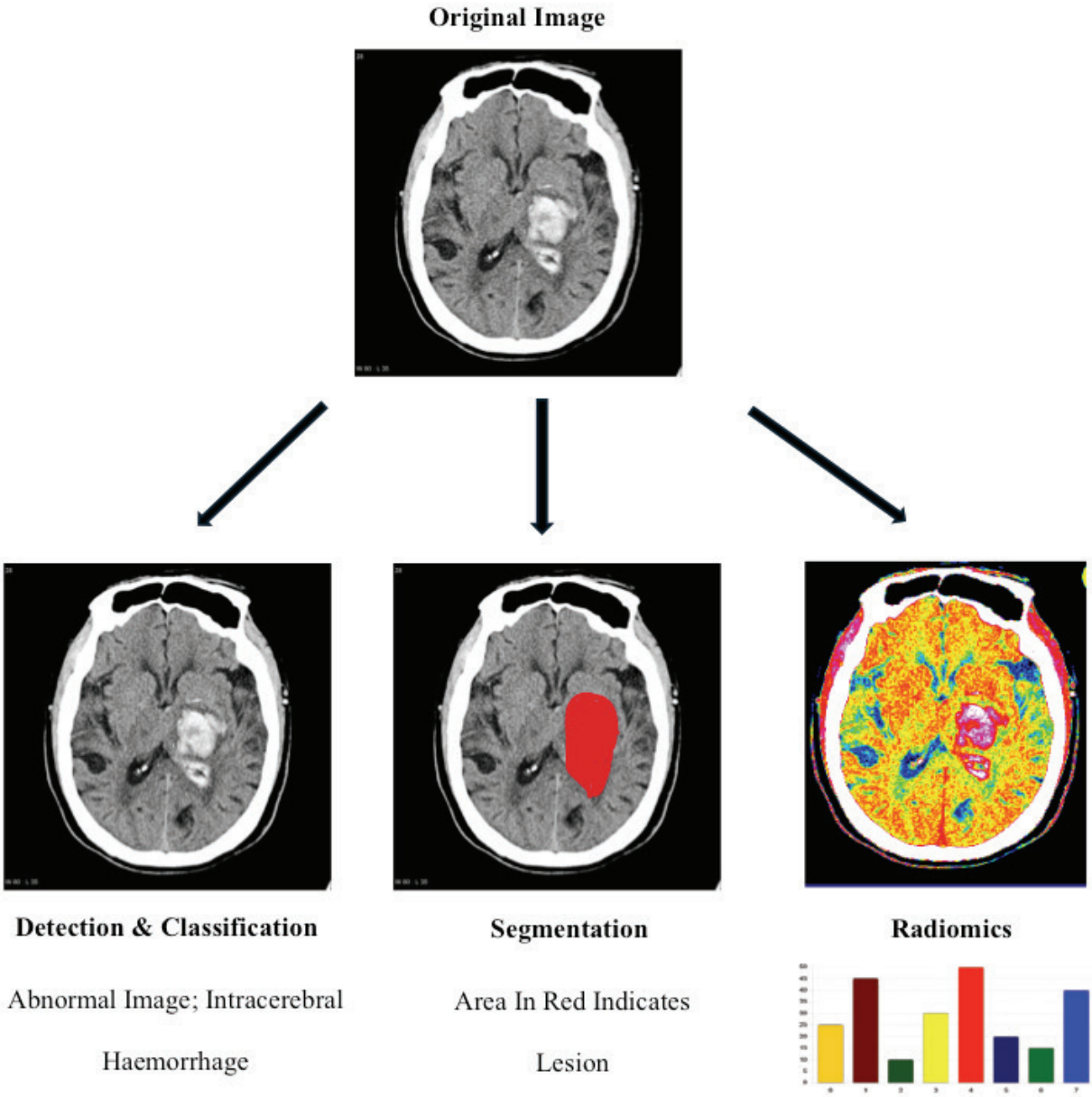

Figure 1. Overview of key clinical use cases for artificial intelligence in medical imaging. Radiomic features, such as texture, shape or intensity, are quantitatively derived (represented by the bar chart); each colour represents a different feature. Source image data: 'Hypertensive basal ganglial bleed' by Frank Gaillard is licenced under CC BY-NC-SA 3.0. Available at: https://radiopaedia.org/cases/hypertensive-basal-ganglial-bleed.

search for when building decision frameworks (Castellino, 2005). However, CAD software have had comparatively less clinical uptake than had been hoped, attributed to their subhuman performance, lack of generalisability and lack of demonstrable improvement to reporting accuracy, when integrated into diagnostic imaging workflows (Cole et al, 2014; Lehman et al, 2015; Hosny et al, 2018).

Artificial intelligence-enabled CAD software (AI-CAD) has shown considerably more promise. AI-CAD has been investigated across numerous pathologies and diagnostic modalities, particularly within ophthalmological, breast and respiratory imaging, with excellent results (Liu et al, 2019; Aggarwal et al, 2021). For the interested reader there are several relevant systematic reviews appraising AI-CAD performance (Aggarwal et al, 2021; Kelly et al, 2022; Kumar et al, 2023). In some instances, AI-CAD has been shown to have comparable diagnostic accuracy to human experts (Liu et al, 2019; Shen et al, 2019), even

surpassing radiologists for the detection of breast cancer in mammography (McKinney et al, 2020) and dermatologists for classification of skin lesions (Esteva et al, 2017).

Segmentation

The characterisation of radiological features and regions of interest (ROI) is a key step in the interrogation of medical images. It is essential for diagnosis, treatment planning and disease monitoring. However, manual annotation is time and labour intensive, as well as being prone to error (Ma et al, 2024). Artificial intelligence software can be used to automatically or semi-automatically delineate anatomical structures and abnormalities on images, segmenting organs and lesions (Akkus et al, 2017). Clinical use cases include automated evaluation of tumour treatment response on brain magnetic resonance imaging (MRI) (Kickingereder et al, 2019) and cardiac segmentation for disease analysis on cardiac MRI (Khened et al, 2019). For comprehensive reviews on the role of AI in medical image segmentation the reader is encouraged to refer to the following papers: (Shen et al, 2017; Hesamian et al, 2019; Cai et al, 2020).

There remain significant methodological challenges for the development of clinically useful AI segmentation software. For instance, the efficacy of such software may be limited to the task it is trained on, leading to calls for the development of task-agnostic 'generalisable learners' (Antonelli et al, 2022). Moreover, the effectiveness of AI segmentation software is contingent on the quality of the training labels, and is subject to systematic errors in annotation as well as inter-annotator variability; it is estimated that at least three expert annotators need to be used to generate accurate labels (Joskowicz et al, 2019).

Radiomics

Radiomics is a relatively nascent field, which involves the derivation of quantitative features at scale and correlation with biological or clinical endpoints; this provides a richer interpretation then that which is discernible to the human eye (Lambin et al, 2012; Avanzo et al, 2020). Supplementing a radiologist's interpretative skill with quantitative measures could enable predictions of treatment response, heralding a new era in personalised medicine (Parekh and Jacobs, 2019; Miles, 2020). Traditional radiomic approaches use 'hand-crafted' features such as intensity, shape or texture, which are applied to specific areas of images (Guiot et al, 2022). However, the utility of 'hand-crafted' features is constrained by difficulties in feature design and an inability to effectively capture image information (Avanzo et al, 2020). Artificial intelligence-based radiomic approaches have garnered significant attention as they can interrogate large number of feature combinations, facilitate the discovery of relationships not considered in 'hand-crafted' methods and remove human bias in feature development (Zhang et al, 2022).

Artificial intelligence-based radiomics has most extensively been studied in oncological imaging and has been explored in all stages of the clinical pathway (Li and Zhou, 2022; Zhang et al, 2023). In terms of diagnosis, AI-extracted radiomic features have been shown to discriminate lung cancer histological subtypes on computed tomography (CT) imaging (Guo et al, 2021; Marentakis et al, 2021). Moreover, AI-derived radiomic signatures have been demonstrated to predict treatment efficacy across modalities, including radiotherapy (Jiao et al, 2022), mutation-targeted therapies (Mu et al, 2020; Song et al, 2021) and immunotherapies (He et al, 2020; Park et al, 2020a).

Fundamentals of machine learning

Machine learning is the process that underlies the development of most AI tools. This involves the use of algorithms that enable computers to learn from and make predictions or decisions based on data, without explicit programming (Najjar, 2023). Machine learning algorithms are generally divided into supervised, unsupervised, and reinforcement learning approaches (Barragán-Montero et al, 2021). In supervised learning, models are generated from labelled data, where each input image is paired with the correct output (Barragán-Montero et al, 2021). In unsupervised learning, models are constructed by identifying patterns in data without labelled inputs (Raza and Singh, 2021). This might involve clustering similar

images or groups (Raza and Singh, 2021). Finally, in reinforcement learning, models are built iteratively by receiving feedback from correct and incorrect actions (Hu et al, 2023).

Hybrid frameworks also exist such as self-supervised learning, in which the machine produces supervisory signals itself from unlabelled data (Huang et al, 2023). Given the computational costs of training ML models, pre-trained models are commonly used in a process known as transfer learning (Kim et al, 2022). These have often been trained on very large datasets such as ImageNet, and fine-tuned on medical imaging tasks (Kim et al, 2022). Models may also be combined with each other to generate better performance, in a paradigm called ensemble learning (Castiglioni et al, 2021).

Data acquisition and splitting

The performance of an AI model heavily relies on the quality and quantity of training data. In medical imaging, this means acquiring high-resolution, well-annotated images that provide detailed information for the model to learn from (Varoquaux and Cheplygina, 2022). Once acquired, the data is typically split into three main sets. The training dataset is the largest portion of the data and is used to build a model (Prinzi et al, 2024). During the training process, the machine learns how features of the data interact together and generates predictions (Castiglioni et al, 2021). Parameters are internal variables that are used to create predictions, whereas hyperparameters are configurations which govern the learning process (Prinzi et al, 2024). The validation dataset is used to prevent the model from being overfitted, where the model performs well on training data but poorly on new, unseen data (Erickson et al, 2017). Validation involves finding the best hyperparameters that allow the fitting of the data, while ensuring generalisability (Prinzi et al, 2024). The testing dataset is used to evaluate the performance of the fully trained model (Castiglioni et al, 2021). It is crucial that this data is not used in any way during the training or validation phases to ensure that the evaluation is unbiased and reflects the model's ability to generalise to new data (Varoquaux and Cheplygina, 2022).

Deep learning

Deep learning is a subset of ML which uses artificial neural networks, complex architectures with multiple layers (Miotto et al, 2018). The incorporation of different processing layers enables the machine to develop various levels of learning representations with increasing abstraction (LeCun et al, 2015). Layers consist of individual nodes which receive information from other nodes and have summated weighted outputs (Currie et al, 2019). These are compared to the correct reference outputs and, during the training process, weights are adjusted to minimise error (Currie et al, 2019). Convolutional Neural Networks (CNNs) are a popular deep learning approach in medical image analysis and draw inspiration from the structure of the animal visual cortex (Yamashita et al, 2018). Convolutional Neural Networks can automatically identify and learn features from images, such as edges, textures, and shapes, with increasing levels of complexity (Kourounis et al, 2023). Hence, they are particularly good at developing an understanding of spatial hierarchies (Yamashita et al, 2018).

Convolutional Neural Networks have been extensively studied across myriad medical imaging tasks (Litjens et al, 2017). For instance, CNNs can detect a range of pulmonary pathologies on chest radiographs, including pneumonia, Coronavirus Disease 2019 (COVID-19) and tuberculosis, with excellent performance (Lakhani and Sundaram, 2017; Hashmi et al, 2020; Breve, 2022). Convolutional Neural Networks also perform well in cross-sectional imaging such as CT and MRI scans, identifying appendicitis, pulmonary embolisms and liver tumours (Huang et al, 2020; Park et al, 2020b; Wu et al, 2020). Moreover, 2-dimensional and 3-dimensional CNNs can segment organs and lesions such as tumours, with state-of-the-art accuracy across imaging modalities (Yamashita et al, 2018; Ilesanmi et al, 2024).

Model evaluation

Numerous metrics are used to quantify model performance, depending on whether the ML model is used for classification or segmentation. Unfortunately, there is no consensus on precisely which metrics to use and practice varies widely (Varoquaux and Cheplygina, 2022). Nonetheless, we give a brief summary of the most widely used performance metrics below.

Classification metrics

Similar to other diagnostic test accuracy studies, evaluation of ML classification is done through the creation of a confusion matrix (Erickson and Kitamura, 2021). This is a 2×2 table which outlines the number of true positives, false positives, true negatives and false negatives identified by the ML model (Erickson and Kitamura, 2021). The following commonly used metrics can then be derived:

Accuracy: The proportion of correctly identified cases out of all cases (Hicks et al, 2022).

Sensitivity (recall): The ability of the model to correctly identify positive cases (Hicks et al, 2022).

Specificity: The ability of the model to correctly identify negative cases (Hicks et al, 2022).

Precision: The proportion of true positive cases among all cases predicted as positive (Hicks et al, 2022).

Receiver operating characteristics curve: A graphical plot of true positive and false positive rates, illustrating the ability of the model to distinguish classes across all thresholds; the area under the curve represents the overall performance of the test (Erickson and Kitamura, 2021).

Segmentation metrics

Performance metrics for ML medical imaging segmentation are broadly divided into overlap and distance methods (Erickson and Kitamura, 2021). The most popular of these are described below.

Dice coefficient: The degree of overlap between the predicted segmentation and the actual segmentation, providing a score between 0 and 1, where 1 indicates perfect overlap (Zou et al, 2004). The Jaccard Index is a similar type of measure but is calculated differently (Eelbode et al, 2020).

Average hausdorff distance: The average distance between the points of the boundaries of two shapes, such as the predicted segmentation and the true object (Erickson and Kitamura, 2021).

Evaluating evidence for artificial intelligence technologies

Despite the transformative potential of AI in medical imaging, there is unfortunately no singular approach in evaluating ML models for clinical practice. This is especially problematic given widespread concerns about poor study design and lack of external validation for many ML tools (Kim et al, 2019; Liu et al, 2019; Nagendran et al, 2020; Yusuf et al, 2020). A variety of consensus guidelines have been developed, which address pre-clinical, observational and interventional studies; interested readers are encouraged to refer to a number of excellent recent reviews (Ibrahim et al, 2021; Shelmerdine et al, 2021; Kolbinger et al, 2024). Authors producing clinical research in AI for medical imaging should refer to the Checklist for Artificial Intelligence in Medical Imaging (CLAIM), which can also serve as a starting point for a critical appraisal (Mongan et al, 2020).

Several proposals have also been made to modify existing reporting guidelines to be specific to AI technologies. This includes diagnostic test accuracy studies (Sounderajah et al, 2021a), predictive studies (Collins et al, 2021) and systematic reviews (Cacciamani et al, 2023). Modifications to popular study assessment tools have been suggested, including the Prediction Model Risk of Bias Assessment Tool (PROBAST) and Quality Assessment of Diagnostic Accuracy Studies (QUADAS), but these have yet to be published (Collins et al, 2021; Sounderajah et al, 2021b). The multitude of reporting guidelines can be overwhelming; rather than providing an exhaustive overview, the following framework draws upon some key themes from these papers and can serve as a useful aide-memoire when evaluating research studies for AI tools in medical imaging. It sequentially addresses the internal validity, i.e., the robustness of study design, as well as the broader generalisability of study findings.

Internal validity

Study design and population

Study design: Where is the study on the hierarchy of evidence (i.e., observational study, trial or meta-analysis?) (Murad et al, 2016)?

Population characteristics: What are the patient demographics and clinical characteristics? Were the inclusion and exclusion criteria appropriate to maintain representativeness and mitigate selection bias (Yu and Eng, 2020)?

Clinical Problem: Does the study design adequately address the clinical problem being investigated (Scott et al, 2021)?

Data

Data sources and structure: Where does the data come from? What is the quality and quantity of data, including the labels and annotations (Balki et al, 2019)?

Data preprocessing: What steps were taken to clean and prepare data, including handling missing values and outliers (Willemink et al, 2020)?

Performance

Performance metrics: Were appropriate metrics chosen to evaluate the model's effectiveness (Bluemke et al, 2020; Erickson and Kitamura, 2021; Hicks et al, 2022)?

Generalisability

Population

Diverse populations: Is the model applicable to diverse patient populations and clinical environments? There are numerous instances of models failing to generalise to new clinical settings due to imbalances arising from disease prevalence or severity (Zech et al, 2018; Scott et al, 2021).

Testing

External validation: Was the model tested against an external dataset? Has it been tested in the 'real world' (Kim et al, 2019; Nagendran et al, 2020)?

Benchmarking against human performance: How does the model compare to experienced clinicians (Liu et al, 2019; Shen et al, 2019; Nagendran et al, 2020)?

Head-to-Head comparisons: How does the model compare against other AI technologies (Litjens et al, 2017)?

Practical implications and integration

Clinical workflow: How well does the model integrate into existing workflows? Does it impair clinician efficiency and increase workload (Scott et al, 2021)?

Interpretation: Are the model's predictions easy to interpret? Are they explainable? Are there safeguards to handle anomalies (Singh et al, 2020)?

Impact on decision making: How does the model affect clinical decision-making processes, and does it lead to better patient outcomes (Scott et al, 2021)?

Robustness: Does the model adequately handle variations in imaging techniques or equipment, including scanner types and imaging protocols? Noise and variations in input can significantly affect model performance (Kulkarni et al, 2021; Anand et al, 2024).

Ethics

Fairness: Do the AI models ensure equitable treatment of minority patients and all subgroups within the population (Ricci Lara et al, 2022)?

Informed consent: Did patients consent for their data to be used in model development and, if clinically deployed, in their own care (Geis et al, 2019)?

Barriers to clinical translation

Despite the immense promise of AI for medical imaging, there has been comparatively little impact on clinical practice (Liu et al, 2019; Nagendran et al, 2020). There are several methodological constraints limiting effective clinical translation. Firstly, the development of robust ML models is impeded by the availability of high-quality data (Varoquaux and Cheplygina, 2022). This not simply a case of needing to use larger sample sizes; many existing datasets have an inherent bias due to a lack of diverse patient populations, measurement error or the presence of confounding features (Zech et al, 2018; Joskowicz et al, 2019; Winkler et al, 2019; Larrazabal et al, 2020). Discriminatory bias can arise during ML development for numerous reasons, including a selection of models which prioritise majority populations and a lack of training data for minorities (Kelly et al, 2019; Obermeyer et al, 2019). Models can also perform poorly across racial and demographic groups (Gichoya et al, 2022).

Secondly, there is a lack of robust evidence derived from prospective studies (Liu et al, 2019; Nagendran et al, 2020; Aggarwal et al, 2021). One recent systematic review found only two documented randomised studies in the most highly investigated specialities for medical imaging AI (Aggarwal et al, 2021). Unfortunately, the vast majority of research in healthcare AI is retrospective in nature and the interpretation of results is affected by the presence of confounding setting-specific variables (Kelly et al, 2019). Many studies also do not test ML models outside of the datasets they were trained on; one systematic review has found only 6% of 516 studies employed any external validation whatsoever (Kim et al, 2019).

Future directions

Multimodal artificial intelligence

Given the rapid pace of growth in the development of AI for medical imaging, predicting the future remains challenging. One area that is attracting attention is multimodal data integration (Topol, 2023). By combining imaging data with clinical, genomic, and pathological information, AI tools can offer a more holistic understanding of diagnostic and prognostic features, leading to individualised treatment plans (Acosta et al, 2022). For instance, training models on both imaging data and clinical parameters such as respiratory rate, systolic blood pressure and blood glucose measurements in intensive care patients, have been shown to improve diagnostic performance when classifying chest radiograph pathologies (Khader et al, 2023). This is hardly surprising as it mimics reasoning in daily practice, where clinicians draw upon multiple dimensions of inputs when making complex decisions (Kitamura and Topol, 2023).

Federated learning

Issues of privacy, fairness and bias may also be mitigated by employing federated learning approaches. In contrast to centrally developed models, which typically use multi-institutional datasets, federated learning leverages local institutional data (Kwak and Bai, 2023). This has the advantage of reducing the impact of confounding variables across settings and facilitates the tailoring of models to local patient populations (Sheller et al, 2020). This may be particularly helpful for institutions that experience substantial differences in disease prevalence or severity, as it addresses resulting class imbalances (Darzidehkalani et al, 2022a). However, federated learning faces several challenges including difficulties in processing heterogeneous data, selecting appropriate training paradigms and developing local systems architecture (Darzidehkalani et al, 2022b).

Explainability

Machine learning models face the ‘black box’ problem – the decision-making process is opaque and the reasons for mapping inputs to outputs cannot readily be explained (Liu et al, 2019). This makes human-machine interaction challenging and is especially concerning in healthcare, where decision-making is high stakes and has life-altering consequences

(Chen et al, 2022). There has been significant recent interest in developing explainable AI (XAI) systems, by incorporating techniques such as heat or saliency mapping (Singh et al, 2020). These are attribution-based methods, which aim to determine how specific image features contribute to model outcomes (Singh et al, 2020). While still in its infancy, XAI systems will be crucial for fostering clinician trust and subsequent integration into clinical workflows (Borys et al, 2023).

Conclusion

A systematic and thorough evaluation of AI tools is essential for safe and efficacious implementation in clinical practice. Clinical use cases in medical imaging are highly diverse and include detection, classification, segmentation and radiomics. Understanding foundational ML concepts, including evaluation metrics, as well as appraising the internal validity and generalisability of evidence, can help clinicians rigorously evaluate the potential of AI tools. Future developments in AI for medical imaging could involve the creation of multi-modal AI, the use of federated learning, and incorporation of explainability into ML models.

Key points

- Artificial intelligence tools are increasingly being investigated for detection, classification, segmentation and radiomic approaches in medical imaging.
- A variety of performance metrics are used to evaluate classification and segmentation.
- When appraising evidence for AI tools, clinicians need to consider the internal validity and generalisability of studies.
- Future developments could involve multi-modal AI, federated learning and an increased emphasis on model explainability.

Author details

[1]Centre for Clinical Brain Sciences, University of Edinburgh, Edinburgh, UK

[2]Acute Medical Unit, Royal Infirmary of Edinburgh, Edinburgh, UK

Availability of data and materials

All the data of this study are included in this article.

Author contributions

SK was responsible for the conception, design and writing of this paper. SK contributed to important editorial changes in the manuscript. SK read and approved the final manuscript. SK has participated sufficiently in the work and agreed to be accountable for all aspects of the work.

Ethics approval and consent to participant

No consent or ethical approval was required as this was a narrative review that used publically available information and sources.

Acknowledgement

Not applicable.

Funding

No funding was used to support this work.

Conflict of interest

The author declares no conflict of interest.

References

Acosta JN, Falcone GJ, Rajpurkar P, Topol EJ. Multimodal biomedical AI. Nat Med. 2022;28(9):1773–1784. https://doi.org/10.1038/s41591-022-01981-2

Aggarwal R, Sounderajah V, Martin G et al. Diagnostic accuracy of deep learning in medical imaging: a systematic review and meta-analysis. NPJ Digit Med. 2021;4(1):65. https://doi.org/10.1038/s41746-021-00438-z

Akkus Z, Galimzianova A, Hoogi A, Rubin DL, Erickson BJ. Deep learning for brain MRI segmentation: state of the art and future directions. J Digit Imaging. 2017;30(4):449–459. https://doi.org/10.1007/s10278-017-9983-4

Anand A, Krithivasan S, Roy K. Romia: a framework for creating robust medical imaging AI models for chest radiographs. Front Radiol. 2024;3:1274273. https://doi.org/10.3389/fradi.2023.1274273

Antonelli M, Reinke A, Bakas S et al. The medical segmentation decathlon. Nat Commun. 2022;13(1):4128. https://doi.org/10.1038/s41467-022-30695-9

Avanzo M, Wei L, Stancanello J et al. Machine and deep learning methods for radiomics. Med Phys. 2020;47(5):e185–e202. https://doi.org/10.1002/mp.13678

Balki I, Amirabadi A, Levman J et al. Sample-size determination methodologies for machine learning in medical imaging research: a systematic review. Can Assoc Radiol J. 2019;70(4):344–353. https://doi.org/10.1016/j.carj.2019.06.002

Barragán-Montero A, Javaid U, Valdés G et al. Artificial intelligence and machine learning for medical imaging: a technology review. Phys Med. 2021;83:242–256. https://doi.org/10.1016/j.ejmp.2021.04.016

Bluemke DA, Moy L, Bredella MA et al. Assessing radiology research on artificial intelligence: a brief guide for authors, reviewers, and readers-from the radiology editorial board. Radiology. 2020;294(3):487–489. https://doi.org/10.1148/radiol.2019192515

Borys K, Schmitt YA, Nauta M et al. Explainable AI in medical imaging: an overview for clinical practitioners - saliency-based XAI approaches. Eur J Radiol. 2023;162:110787. https://doi.org/10.1016/j.ejrad.2023.110787

Breve FA. Covid-19 detection on chest x-ray images: a comparison of CNN architectures and ensembles. Expert Syst Appl. 2022;204:117549. https://doi.org/10.1016/j.eswa.2022.117549

Cacciamani GE, Chu TN, Sanford DI et al. Prisma AI reporting guidelines for systematic reviews and meta-analyses on AI in healthcare. Nat Med. 2023;29(1):14–15. https://doi.org/10.1038/s41591-022-02139-w

Cai L, Gao J, Zhao D. A review of the application of deep learning in medical image classification and segmentation. Ann Transl Med. 2020;8(11):713. https://doi.org/10.21037/atm.2020.02.44

Castellino RA. Computer aided detection (CAD): an overview. Cancer Imaging. 2005;5(1):17–19. https://doi.org/10.1102/1470-7330.2005.0018

Castiglioni I, Rundo L, Codari M et al. AI applications to medical images: from machine learning to deep learning. Phys Med. 2021;83:9–24. https://doi.org/10.1016/j.ejmp.2021.02.006

Chen H, Gomez C, Huang CM, Unberath M. Explainable medical imaging AI needs human-centered design: guidelines and evidence from a systematic review. NPJ Digit Med. 2022;5(1):156. https://doi.org/10.1038/s41746-022-00699-2

Cole EB, Zhang Z, Marques HS et al. Impact of computer-aided detection systems on radiologist accuracy with digital mammography. AJR Am J Roentgenol. 2014;203(4):909–916. https://doi.org/10.2214/AJR.12.10187

Collins GS, Dhiman P, Andaur Navarro CL et al. Protocol for development of a reporting guideline (TRIPOD-AI) and risk of bias tool (PROBAST-AI) for diagnostic and prognostic prediction model studies based on artificial intelligence. BMJ Open. 2021;11(7):e048008. https://doi.org/10.1136/bmjopen-2020-048008

Currie G, Hawk KE, Rohren E, Vial A, Klein R. Machine learning and deep learning in medical imaging: intelligent imaging. J Med Imaging Radiat Sci. 2019;50(4):477–487. https://doi.org/10.1016/j.jmir.2019.09.005

Darzidehkalani E, Ghasemi-Rad M, van Ooijen PMA. Federated learning in medical imaging: part I: toward multicentral health care ecosystems. J Am Coll Radiol. 2022a;19(8):969–974. https://doi.org/10.1016/j.jacr.2022.03.015

Darzidehkalani E, Ghasemi-Rad M, van Ooijen PMA. Federated learning in medical imaging: part II: methods, challenges, and considerations. J Am Coll Radiol. 2022b;19(8):975–982. https://doi.org/10.1016/j.jacr.2022.03.016

Eelbode T, Bertels J, Berman M et al. Optimization for medical image segmentation: theory and practice when evaluating with dice score or Jaccard index. IEEE Trans Med Imaging. 2020;39(11):3679–3690. https://doi.org/10.1109/TMI.2020.3002417

Erickson BJ, Kitamura F. Magician's corner: 9. Performance metrics for machine learning models. Radiol Artif Intell. 2021;3(3):e200126. https://doi.org/10.1148/ryai.2021200126

Erickson BJ, Korfiatis P, Akkus Z, Kline TL. Machine learning for medical imaging. Radiographics. 2017;37(2):505–515. https://doi.org/10.1148/rg.2017160130

Esteva A, Kuprel B, Novoa RA et al. Dermatologist-level classification of skin cancer with deep neural networks. Nature. 2017;542(7639):115–118. https://doi.org/10.1038/nature21056

Geis JR, Brady AP, Wu CC et al. Ethics of artificial intelligence in radiology: summary of the joint European and North American multisociety statement. Radiology. 2019;293(2):436–440. https://doi.org/10.1148/radiol.2019191586

Gichoya JW, Banerjee I, Bhimireddy AR et al. AI recognition of patient race in medical imaging: a modelling study. Lancet Digit Health. 2022;4(6):e406–e414. https://doi.org/10.1016/S2589-7500(22)00063-2

Guiot J, Vaidyanathan A, Deprez L et al. A review in radiomics: making personalized medicine a reality via routine imaging. Med Res Rev. 2022;42(1):426–440. https://doi.org/10.1002/med.21846

Guo Y, Song Q, Jiang M et al. Histological subtypes classification of lung cancers on CT images using 3d deep learning and radiomics. Acad Radiol. 2021;28(9):e258–e266. https://doi.org/10.1016/j.acra.2020.06.010

Hashmi MF, Katiyar S, Keskar AG, Bokde ND, Geem ZW. Efficient pneumonia detection in chest xray images using deep transfer learning. Diagnostics (Basel). 2020;10(6):417. https://doi.org/10.3390/diagnostics10060417

He B, Dong D, She Y et al. Predicting response to immunotherapy in advanced non-small-cell lung cancer using tumor mutational burden radiomic biomarker. J Immunother Cancer. 2020;8(2):e000550. https://doi.org/10.1136/jitc-2020-000550

Hesamian MH, Jia W, He X, Kennedy P. Deep learning techniques for medical image segmentation: achievements and challenges. J Digit Imaging. 2019;32(4):582–596. https://doi.org/10.1007/s10278-019-00227-x

Hicks SA, Strumke I, Thambawita V et al. On evaluation metrics for medical applications of artificial intelligence. Sci Rep. 2022;12(1):5979. https://doi.org/10.1038/s41598-022-09954-8

Hosny A, Parmar C, Quackenbush J, Schwartz LH, Aerts H. Artificial intelligence in radiology. Nat Rev Cancer. 2018;18(8):500–510. https://doi.org/10.1038/s41568-018-0016-5

Hu M, Zhang J, Matkovic L, Liu T, Yang X. Reinforcement learning in medical image analysis: concepts, applications, challenges, and future directions. J Appl Clin Med Phys. 2023;24(2):e13898. https://doi.org/10.1002/acm2.13898

Huang SC, Kothari T, Banerjee I et al. PENet-a scalable deep-learning model for automated diagnosis of pulmonary embolism using volumetric CT imaging. NPJ Digit Med. 2020;3(1):61. https://doi.org/10.1038/s41746-020-0266-y

Huang SC, Pareek A, Jensen M et al. Self-supervised learning for medical image classification: a systematic review and implementation guidelines. NPJ Digit Med. 2023;6(1):74. https://doi.org/10.1038/s41746-023-00811-0

Ibrahim H, Liu X, Denniston AK. Reporting guidelines for artificial intelligence in healthcare research. Clin Exp Ophthalmol. 2021;49(5):470–476. https://doi.org/10.1111/ceo.13943

Ilesanmi AE, Ilesanmi TO, Ajayi BO. Reviewing 3D convolutional neural network approaches for medical image segmentation. Heliyon. 2024;10(6):e27398. https://doi.org/10.1016/j.heliyon.2024.e27398

Itri JN, Tappouni RR, McEachern RO, Pesch AJ, Patel SH. Fundamentals of diagnostic error in imaging. Radiographics. 2018;38(6):1845–1865. https://doi.org/10.1148/rg.2018180021

Jiao Z, Li H, Xiao Y et al. Integration of deep learning radiomics and counts of circulating tumor cells improves prediction of outcomes of early stage NSCLC patients treated with stereotactic body

radiation therapy. Int J Radiat Oncol Biol Phys. 2022;112(4):1045–1054. https://doi.org/10.1016/j.ijrobp.2021.11.006

Joskowicz L, Cohen D, Caplan N, Sosna J. Inter-observer variability of manual contour delineation of structures in CT. Eur Radiol. 2019;29(3):1391–1399. https://doi.org/10.1007/s00330-018-5695-5

Kelly BS, Judge C, Bollard SM et al. Radiology artificial intelligence: a systematic review and evaluation of methods (raise). Eur Radiol. 2022;32(11):7998–8007. https://doi.org/10.1007/s00330-022-08784-6

Kelly CJ, Karthikesalingam A, Suleyman M, Corrado G, King D. Key challenges for delivering clinical impact with artificial intelligence. BMC Med. 2019;17(1):195. https://doi.org/10.1186/s12916-019-1426-2

Khader F, Müller-Franzes G, Wang T et al. Multimodal deep learning for integrating chest radiographs and clinical parameters: a case for transformers. Radiology. 2023;309(1):e230806. https://doi.org/10.1148/radiol.230806

Khened M, Kollerathu VA, Krishnamurthi G. Fully convolutional multi-scale residual DenseNets for cardiac segmentation and automated cardiac diagnosis using ensemble of classifiers. Med Image Anal. 2019;51:21–45. https://doi.org/10.1016/j.media.2018.10.004

Kickingereder P, Isensee F, Tursunova I et al. Automated quantitative tumour response assessment of MRI in neuro-oncology with artificial neural networks: a multicentre, retrospective study. Lancet Oncol. 2019;20(5):728–740. https://doi.org/10.1016/S1470-2045(19)30098-1

Kim DW, Jang HY, Kim KW, Shin Y, Park SH. Design characteristics of studies reporting the performance of artificial intelligence algorithms for diagnostic analysis of medical images: results from recently published papers. Korean J Radiol. 2019;20(3):405–410. https://doi.org/10.3348/kjr.2019.0025

Kim HE, Cosa-Linan A, Santhanam N et al. Transfer learning for medical image classification: a literature review. BMC Med Imaging. 2022;22(1):69. https://doi.org/10.1186/s12880-022-00793-7

Kitamura FC, Topol EJ. The initial steps of multimodal AI in radiology. Radiology. 2023;309(1):e232372. https://doi.org/10.1148/radiol.232372

Kolbinger FR, Veldhuizen GP, Zhu J, Truhn D, Kather JN. Reporting guidelines in medical artificial intelligence: a systematic review and meta-analysis. Commun Med. 2024;4(1):71. https://doi.org/10.1038/s43856-024-00492-0

Kourounis G, Elmahmudi AA, Thomson B et al. Computer image analysis with artificial intelligence: a practical introduction to convolutional neural networks for medical professionals. Postgrad Med J. 2023;99(1178):1287–1294. https://doi.org/10.1093/postmj/qgad095

Kulkarni V, Gawali M, Kharat A. Key technology considerations in developing and deploying machine learning models in clinical radiology practice. JMIR Med Inform. 2021;9(9):e28776. https://doi.org/10.2196/28776

Kumar Y, Koul A, Singla R, Ijaz MF. Artificial intelligence in disease diagnosis: a systematic literature review, synthesizing framework and future research agenda. J Ambient Intell Humanz Comput. 2023;14(7):8459–8486. https://doi.org/10.1007/s12652-021-03612-z

Kwak L, Bai H. The role of federated learning models in medical imaging. Radiol Artif Intell. 2023;5(3):e230136. https://doi.org/10.1148/ryai.230136

Lakhani P, Sundaram B. Deep learning at chest radiography: automated classification of pulmonary tuberculosis by using convolutional neural networks. Radiology. 2017;284(2):574–582. https://doi.org/10.1148/radiol.2017162326

Lambin P, Rios-Velazquez E, Leijenaar R et al. Radiomics: extracting more information from medical images using advanced feature analysis. Eur J Cancer. 2012;48(4):441–446. https://doi.org/10.1016/j.ejca.2011.11.036

Larrazabal AJ, Nieto N, Peterson V, Milone DH, Ferrante E. Gender imbalance in medical imaging datasets produces biased classifiers for computer-aided diagnosis. Proc Natl Acad Sci USA. 2020;117(23):12592–12594. https://doi.org/10.1073/pnas.1919012117

LeCun Y, Bengio Y, Hinton G. Deep learning. Nature. 2015;521(7553):436–444. https://doi.org/10.1038/nature14539

Lee CS, Nagy PG, Weaver SJ, Newman-Toker DE. Cognitive and system factors contributing to diagnostic errors in radiology. AJR Am J Roentgenol. 2013;201(3):611–617. https://doi.org/10.2214/AJR.12.10375

Lehman CD, Wellman RD, Buist DS et al. Diagnostic accuracy of digital screening mammography with and without computer-aided detection. JAMA Intern Med. 2015;175(11):1828–1837. https://doi.org/10.1001/jamainternmed.2015.5231

Li S, Zhou B. A review of radiomics and genomics applications in cancers: the way towards precision medicine. Radiat Oncol. 2022;17(1):217. https://doi.org/10.1186/s13014-022-02192-2

Litjens G, Kooi T, Bejnordi BE et al. A survey on deep learning in medical image analysis. Med Image Anal. 2017;42:60–88. https://doi.org/10.1016/j.media.2017.07.005

Liu X, Faes L, Kale AU et al. A comparison of deep learning performance against health-care professionals in detecting diseases from medical imaging: a systematic review and meta-analysis. Lancet Digit Health. 2019;1(6):e271–e297. https://doi.org/10.1016/S2589-7500(19)30123-2

Ma J, He Y, Li F et al. Segment anything in medical images. Nat Commun. 2024;15(1):654. https://doi.org/10.1038/s41467-024-44824-z

Marentakis P, Karaiskos P, Kouloulias V et al. Lung cancer histology classification from CT images based on radiomics and deep learning models. Med Biol Eng Comput. 2021;59(1):215–226. https://doi.org/10.1007/s11517-020-02302-w

McKinney SM, Sieniek M, Godbole V et al. International evaluation of an AI system for breast cancer screening. Nature. 2020;577(7788):89–94. https://doi.org/10.1038/s41586-019-1799-6

Miles K. Radiomics for personalised medicine: the long road ahead. Br J Cancer. 2020;122(7):929–930. https://doi.org/10.1038/s41416-019-0699-8

Miotto R, Wang F, Wang S, Jiang X, Dudley JT. Deep learning for healthcare: review, opportunities and challenges. Brief Bioinform. 2018;19(6):1236–1246. https://doi.org/10.1093/bib/bbx044

Mongan J, Moy L, Kahn CE. Checklist for artificial intelligence in medical imaging (CLAIM): a guide for authors and reviewers. Radiol Artif Intell. 2020;2(2):e200029. https://doi.org/10.1148/ryai.2020200029

Mu W, Jiang L, Zhang J et al. Non-invasive decision support for NSCLC treatment using PET/CT radiomics. Nat Commun. 2020;11(1):5228. https://doi.org/10.1038/s41467-020-19116-x

Murad MH, Asi N, Alsawas M, Alahdab F. New evidence pyramid. Evid Based Med. 2016;21(4):125–127. https://doi.org/10.1136/ebmed-2016-110401

Nagendran M, Chen Y, Lovejoy CA et al. Artificial intelligence versus clinicians: systematic review of design, reporting standards, and claims of deep learning studies. BMJ. 2020;368:m689. https://doi.org/10.1136/bmj.m689

Najjar R. Redefining radiology: a review of artificial intelligence integration in medical imaging. Diagnostics (Basel). 2023;13(17):2760. https://doi.org/10.3390/diagnostics13172760

Obermeyer Z, Powers B, Vogeli C, Mullainathan S. Dissecting racial bias in an algorithm used to manage the health of populations. Science. 2019;366(6464):447–453. https://doi.org/10.1126/science.aax2342

Parekh VS, Jacobs MA. Deep learning and radiomics in precision medicine. Expert Rev Precis Med Drug Dev. 2019;4(2):59–72. https://doi.org/10.1080/23808993.2019.1585805

Park C, Na KJ, Choi H et al. Tumor immune profiles noninvasively estimated by FDG pet with deep learning correlate with immunotherapy response in lung adenocarcinoma. Theranostics. 2020a;10(23):10838–10848. https://doi.org/10.7150/thno.50283

Park JJ, Kim KA, Nam Y et al. Convolutional-neural-network-based diagnosis of appendicitis via CT scans in patients with acute abdominal pain presenting in the emergency department. Sci Rep. 2020b;10(1):9556. https://doi.org/10.1038/s41598-020-66674-7

Prinzi F, Currieri T, Gaglio S, Vitabile S. Shallow and deep learning classifiers in medical image analysis. Eur Radiol Exp. 2024;8(1):26. https://doi.org/10.1186/s41747-024-00428-2

Raza K, Singh NK. A tour of unsupervised deep learning for medical image analysis. Curr Med Imaging. 2021;17(9):1059–1077. https://doi.org/10.2174/1573405617666210127154257

Ricci Lara MA, Echeveste R, Ferrante E. Addressing fairness in artificial intelligence for medical imaging. Nat Commun. 2022;13(1):4581. https://doi.org/10.1038/s41467-022-32186-3

Scott I, Carter S, Coiera E. Clinician checklist for assessing suitability of machine learning applications in healthcare. BMJ Health Care Inform. 2021;28(1):e100251. https://doi.org/10.1136/bmjhci-2020-100251

Sheller MJ, Edwards B, Reina GA et al. Federated learning in medicine: facilitating multi-institutional collaborations without sharing patient data. Sci Rep. 2020;10(1):12598. https://doi.org/10.1038/s41598-020-69250-1

Shelmerdine SC, Arthurs OJ, Denniston A, Sebire NJ. Review of study reporting guidelines for clinical studies using artificial intelligence in healthcare. BMJ Health Care Inform. 2021;28(1):e100385. https://doi.org/10.1136/bmjhci-2021-100385

Shen D, Wu G, Suk HI. Deep learning in medical image analysis. Annu Rev Biomed Eng. 2017;19(1):221–248. https://doi.org/10.1146/annurev-bioeng-071516-044442

Shen J, Zhang CJP, Jiang B et al. Artificial intelligence versus clinicians in disease diagnosis: systematic review. JMIR Med Inform. 2019;7(3):e10010. https://doi.org/10.2196/10010

Singh A, Sengupta S, Lakshminarayanan V. Explainable deep learning models in medical image analysis. J Imaging. 2020;6(6):52. https://doi.org/10.3390/jimaging6060052

Song Z, Liu T, Shi L et al. The deep learning model combining CT image and clinicopathological information for predicting ALK fusion status and response to ALK-TKI therapy in non-small cell lung cancer patients. Eur J Nucl Med Mol Imaging. 2021;48(2):361–371. https://doi.org/10.1007/s00259-020-04986-6

Sounderajah V, Ashrafian H, Golub RM et al. Developing a reporting guideline for artificial intelligence-centred diagnostic test accuracy studies: the STARD-AI protocol. BMJ Open. 2021a;11(6):e047709. https://doi.org/10.1136/bmjopen-2020-047709

Sounderajah V, Ashrafian H, Rose S et al. A quality assessment tool for artificial intelligence-centered diagnostic test accuracy studies: QUADAS-AI. Nat Med. 2021b;27(10):1663–1665. https://doi.org/10.1038/s41591-021-01517-0

Topol EJ. As artificial intelligence goes multimodal, medical applications multiply. Science. 2023;381(6663):adk6139. https://doi.org/10.1126/science.adk6139

Varoquaux G, Cheplygina V. Machine learning for medical imaging: methodological failures and recommendations for the future. NPJ Digit Med. 2022;5(1):48. https://doi.org/10.1038/s41746-022-00592-y

Willemink MJ, Koszek WA, Hardell C et al. Preparing medical imaging data for machine learning. Radiology. 2020;295(1):4–15. https://doi.org/10.1148/radiol.2020192224

Winkler JK, Fink C, Toberer F et al. Association between surgical skin markings in dermoscopic images and diagnostic performance of a deep learning convolutional neural network for melanoma recognition. JAMA Dermatol. 2019;155(10):1135–1141. https://doi.org/10.1001/jamadermatol.2019.1735

Wu Y, White GM, Cornelius T et al. Deep learning li-rads grading system based on contrast enhanced multiphase MRI for differentiation between LR-3 and LR-4/LR-5 liver tumors. Ann Transl Med. 2020;8(11):701. https://doi.org/10.21037/atm.2019.12.151

Yamashita R, Nishio M, Do RKG, Togashi K. Convolutional neural networks: an overview and application in radiology. Insights Imaging. 2018;9(4):611–629. https://doi.org/10.1007/s13244-018-0639-9

Yu AC, Eng J. One algorithm may not fit all: how selection bias affects machine learning performance. Radiographics. 2020;40(7):1932–1937. https://doi.org/10.1148/rg.2020200040

Yusuf M, Atal I, Li J et al. Reporting quality of studies using machine learning models for medical diagnosis: a systematic review. BMJ Open. 2020;10(3):e034568. https://doi.org/10.1136/bmjopen-2019-034568

Zech JR, Badgeley MA, Liu M et al. Variable generalization performance of a deep learning model to detect pneumonia in chest radiographs: a cross-sectional study. PLoS Med. 2018;15(11):e1002683. https://doi.org/10.1371/journal.pmed.1002683

Zhang X, Zhang Y, Zhang G et al. Deep learning with radiomics for disease diagnosis and treatment: challenges and potential. Front Oncol. 2022;12:773840. https://doi.org/10.3389/fonc.2022.773840

Zhang YP, Zhang XY, Cheng YT et al. Artificial intelligence-driven radiomics study in cancer: the role of feature engineering and modeling. Mil Med Res. 2023;10(1):22. https://doi.org/10.1186/s40779-023-00458-8

Zou KH, Warfield SK, Bharatha A et al. Statistical validation of image segmentation quality based on a spatial overlap index. Acad Radiol. 2004;11(2):178–189. https://doi.org/10.1016/S1076-6332(03)00671-8