Review Article

A literature review of artificial intelligence (AI) for medical image segmentation: from AI and explainable AI to trustworthy AI

Zixuan Teng[1#], Lan Li[2#], Ziqing Xin[1], Dehui Xiang[3#], Jiang Huang[4#], Hailing Zhou[5], Fei Shi[3], Weifang Zhu[3], Jing Cai[6], Tao Peng[1,6,7], Xinjian Chen[3,8]

[1]School of Future Science and Engineering, Soochow University, Suzhou, China; [2]Healthy Inspection and Testing Institute, The Center for Disease Control and Prevention of Huangshi, Huangshi, China; [3]MIPAV Lab, the School of Electronic and Information Engineering, Soochow University, Suzhou, China; [4]Department of Ophthalmology, the Second Affiliated Hospital of Soochow University, Suzhou, China; [5]Department of Mechanical Engineering and Product Design Engineering, Swinburne University of Technology, Melbourne, Australia; [6]Department of Health Technology and Informatics, The Hong Kong Polytechnic University, Hong Kong, China; [7]Department of Radiation Oncology, UT Southwestern Medical Center, Dallas, TX, USA; [8]The State Key Laboratory of Radiation Medicine and Protection, Soochow University, Suzhou, China

Contributions: (I) Conception and design: Z Teng; (II) Administrative support: T Peng, X Chen; (III) Provision of study materials or patients: Z Xin, D Xiang, J Cai; (IV) Collection and assembly of data: L Li, J Huang; (V) Data analysis and interpretation: L Li, H Zhou, F Shi, W Zhu; (VI) Manuscript writing: All authors; (VII) Final approval of manuscript: All authors.

[#]These authors contributed equally to this work.

Correspondence to: Tao Peng, PhD. School of Future Science and Engineering, Soochow University, No. 1, Jiuyongxi Road, Suzhou 215222, China; Department of Health Technology and Informatics, The Hong Kong Polytechnic University, Hong Kong, China; Department of Radiation Oncology, UT Southwestern Medical Center, Dallas, TX, USA. Email: sdpengtao401@gmail.com; Xinjian Chen, PhD. MIPAV Lab, the School of Electronic and Information Engineering, Soochow University, No. 1 Shizi Street, Suzhou, China; The State Key Laboratory of Radiation Medicine and Protection, Soochow University, Suzhou, China. Email: xjchen@suda.edu.cn.

Background and Objective: Medical image segmentation is a vital aspect of medical image processing, allowing healthcare professionals to conduct precise and comprehensive lesion analyses. Traditional segmentation methods are often labor intensive and influenced by the subjectivity of individual physicians. The advent of artificial intelligence (AI) has transformed this field by reducing the workload of physicians, and improving the accuracy and efficiency of disease diagnosis. However, conventional AI techniques are not without challenges. Issues such as inexplicability, uncontrollable decision-making processes, and unpredictability can lead to confusion and uncertainty in clinical decision-making. This review explores the evolution of AI in medical image segmentation, focusing on the development and impact of explainable AI (XAI) and trustworthy AI (TAI).

Methods: This review synthesizes existing literature on traditional segmentation methods, AI-based approaches, and the transition from conventional AI to XAI and TAI. The review highlights the key principles and advancements in XAI that aim to address the shortcomings of conventional AI by enhancing transparency and interpretability. It further examines how TAI builds on XAI to improve the reliability, safety, and accountability of AI systems in medical image segmentation.

Key Content and Findings: XAI has emerged as a solution to the limitations of conventional AI by providing greater transparency and interpretability, allowing healthcare professionals to better understand and trust AI-driven decisions. However, XAI itself faces challenges, including those related to safety, robustness, and value alignment. TAI has been developed to overcome these challenges, offering a more reliable framework for AI applications in medical image segmentation. By integrating the principles of XAI with enhanced safety and dependability, TAI addresses the critical need for TAI systems in clinical settings.

Conclusions: TAI presents a promising future for medical image segmentation, combining the benefits

of AI with improved reliability and safety. Thus, TAI is a more viable and dependable option for healthcare applications, and could ultimately lead to better clinical outcomes for patients, and advance the field of medical image processing.

Keywords: Medical image segmentation; artificial intelligence (AI); explainable AI (XAI); trustworthy AI (TAI)

Submitted Apr 08, 2024. Accepted for publication Sep 18, 2024. Published online Nov 29, 2024.
doi: 10.21037/qims-24-723
View this article at: https://dx.doi.org/10.21037/qims-24-723

Introduction

Background

Medical image segmentation, a foundational technique, involves subdividing an image into uniform subcomponents to extract information about regions and contours (1). It has garnered significant attention for its vital role in image analysis phases after medical image segmentation, including object representation and feature (e.g., shape contour) extraction (2). Image segmentation enhances physicians' decision-making accuracy, playing a crucial role in disease diagnosis and treatment (3). Various strategies for medical image segmentation have been explored, including threshold segmentation (4), region segmentation (5), and contour extraction (6). However, in these traditional methods, the process of feature extraction often requires manual characterization, which is affected by the subjective experience of clinicians, and requires a great deal of energy and significant effort.

Rationale and objectives

Artificial intelligence (AI) has led to significant advances in this field, reducing the workload of physicians, and improving the accuracy and efficiency of disease diagnosis (7,8). Common AI methods for medical image segmentation include convolutional neural networks (CNNs) (9-11), deep CNNs (12,13), graph convolutional networks (14,15), generative adversarial networks (GANs) (16-18), and transformers (19). Many research teams have tried to combine these methods to harness their respective strengths and solve more complex problems. However, there are still some issues with AI and explainable AI (XAI) (20). Traditional AI methods face issues such as inexplicable, unknowable, and uncontrollable decision-making processes. Their "black-box" nature leads to a lack of transparency and understandability in the medical field. XAI was developed to make the decision-making process more transparent and understandable, address the "black-box" flaws of AI, and improve ethical and safety considerations; however, several issues arise in relation to XAI that need to be addressed (21). For example, XAI has limitations, notably in its ability to deliver consistent, accurate judgments, and in its lack of security and robustness. Trustworthy AI (TAI) was introduced to enhance the trustworthiness of AI, and ensure it adheres to principles related to safety, robustness, interpretability, accountability, human rights, and value consistency (22). TAI is indispensable in medical image processing, as it enhances the accuracy and reliability of physicians' decision making, and AI-driven diagnoses and recommendations.

Thia article traces and analyzes the evolution of AI in medical image segmentation, focusing on the evolution from traditional AI to XAI and TAI. It addresses the critical need for transparency, interpretability, and trustworthiness in AI systems used for medical image segmentation. It aims to provide a comprehensive overview of the latest advances in XAI and TAI, highlighting the major challenges and potential future directions in medical image segmentation. Through a detailed discussion of AI, XAI, TAI, and a more general discussion, and conclusion, this article aims to clarify the importance of the transition from AI to XAI and TAI to improve the accuracy, transparency, and credibility of medical image segmentation. The application of this technology in medical image segmentation is widely accepted and used. We present this article in accordance with the Narrative Review reporting checklist (available at https://qims.amegroups.com/article/view/10.21037/qims-24-723/rc).

Methods

This review gathers and analyzes existing research on the progression from traditional segmentation methods to AI-based approaches, with particular attention to the development of XAI and TAI. It explores how XAI

Table 1 Search strategy summary

Items	Specification
Dates of searches	March 10, 2023 (first search), and March 10, 2024 (second search)
Databases and other sources searched	Google Scholar
Search terms used	AI-based segmentation, explainable AI (XAI), trustworthy AI (TAI), medical image segmentation, etc.
Timeframe	January 1, 2017 to January 1, 2024
Inclusion criteria	Only articles written in the English language were considered for inclusion
Selection process	Conducted independently by the author

AI, artificial intelligence.

was created to address the limitations of conventional AI by enhancing the interpretability and transparency of AI systems. Additionally, it examines the principles underlying TAI, which extend the goals of XAI by focusing on the dependability, safety, and ethical alignment of AI applications in medical image segmentation. A summary of the search strategy employed in this study is set out in *Table 1*.

Medical image segmentation based on AI

The term "AI" refers to the technologies and systems that simulate human intelligence through computer systems, including machine-learning algorithms and CNN methods. It was first proposed at the 1956 Dartmouth conference (1). Since then, traditional AI has evolved significantly and has been widely integrated into various domains, particularly medical imaging analysis, where its importance continues to increase (23,24). Due to its accuracy and universality, AI is of great significance in the field of medical image segmentation. In this section, we introduce a number of aspects of AI, such as its advantages and limitations (*Figure 1*).

Problems in conventional segmentation models

Early traditional medical image segmentation methods largely focused on edge detection, template matching technology, region growing, graph cutting, and other mathematical methods. These traditional medical image segmentation methods are widely used in practice, but they still face many challenges, including those related to the complexity of anatomical structures, and the diversity of tissues or organs, which vary in shape and size among different patients. Consequently, the segmentation accuracy of these traditional approaches is often inadequate, particularly when anatomical lesions are present (25). Additionally, the reliance of these methods on prior manual delineation complicates the segmentation process. For example, the threshold-based method is not effective in areas where gray-level changes are not obvious, and cannot deal with complex image structure and object overlap. The method based on edge detection has an unsatisfactory segmentation effect when the edge is discontinuous, and it also cannot deal with fuzzy boundaries. The method based on region growth requires a large amount of computation and has a slow processing speed, which is prone to over- or under-segmentation (26). Moreover, while many traditional methods perform well on specific datasets and tasks, they exhibit limited generalization capabilities when applied to new datasets or tasks. Therefore, there is a pressing need to develop more general algorithms that can be applied to various modalities of medical datasets and different organ types to address the challenges of practical application. To address these challenges, AI-based segmentation methods have emerged in recent years. These AI-based techniques are promising, and effectively address the limitations of traditional methods and enhance segmentation accuracy. The implementation of AI technology allows models to learn the features of anatomical tissues, thereby improving the segmentation of complex structures.

Methods of traditional AI algorithms

Some of the AI methods discussed in this section are listed in *Table 2*.

To reduce the workload of physicians in manually delineating the region of interest (ROI) and to enhance the accuracy and efficiency of diagnoses, various research groups have adopted conventional AI models, which

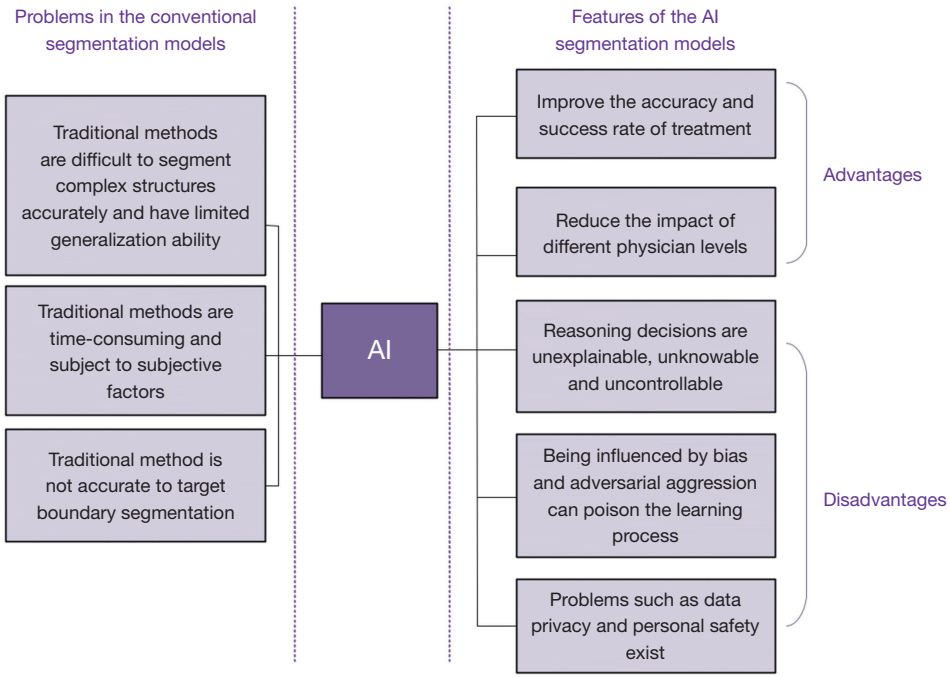

Figure 1 Medical image segmentation based on AI. AI, artificial intelligence.

Table 2 Research teams and their models

Team	Model	Running time	Year introduced
Diaz-Pinto *et al.* (27)	MONAI Label	Training time: ~14 hours	2024
Huang *et al.* (28)	SAM	Training time: ~60 hours	2024
Gao *et al.* (29)	DeSAM	Training time: 1.2–2.7 hours	2024
Ruan *et al.* (30)	VM-UNet	–	2024
Tragakis *et al.* (31)	FCT	Training time: ~14 hours	2023
Miao *et al.* (32)	CauSSL	Training time: ~5 hours	2023
Gaillochet *et al.* (33)	TAAL	Training time: ~9 hours	2023
Butoi *et al.* (34)	UniverSeg	–	2023
Wang *et al.* (35)	Mamba-UNet	Total running time: ~5 hours	2023
Zunair *et al.* (36)	MaskSup	–	2022
Verma *et al.* (37)	Dataset: MoNuSAC2020	–	2021
Zunair *et al.* (38)	Sharp U-Net	–	2021

are designed to automatically detect and segregate areas of interest. K-means (39) clustering is an unsupervised learning algorithm commonly used in image segmentation that simplifies the detection of the ROI by dividing image pixels into multiple clusters. The decision tree (40) is easy to understand and interpret. It constructs a tree model to classify data points. It is also used in some medical image analysis tasks. Support vector machines (41) work well with high-dimensional data and small samples by finding the best hyperplane to separate different classes of data.

With the advancement of technology, deep-learning methods have gradually become mainstream. These include CNNs, residual networks (ResNets), and GANs. The CNN (42) is a type of deep-learning model designed to process mesh-like data, especially images and videos. Features of input data are extracted through convolution and pooling operations. Its basic structure includes a convolutional layer, activation function, pooling layer, fully connected layer, and output layer. Unlike traditional feature engineering and machine-learning methods, which require manual design and may not capture complex patterns effectively, deep-learning methods can automatically learn complex image features and hierarchical representations without manually extracting features, thus improving the segmentation accuracy and generalization ability of models. In addition, deep-learning models perform well in terms of scalability on large-scale data sets and have achieved many leading results in medical image processing, especially in lesion detection, organ segmentation, and disease diagnosis. Conversely, CNNs excel in automatically learning hierarchical features directly from data, and thus are popular for their ability to handle diverse and intricate information in tasks, such as image classification and segmentation. ResNets were first proposed by He *et al.* (43), and use residual learning to train extremely deep networks. The core concept of residual learning is the introduction of skip connections, which allow the network to learn the residual mapping directly. The collection of these ResNets introduced by them achieved a 3.57% error on the ImageNet test set (44). This result earned the ResNet first place in the ImageNet Large-Scale Visual Recognition Challenge 2015 classification task (2). GANs were first proposed by Goodfellow *et al.* (45). GANs consists of two main parts: a generator, and a discriminator. The generator attempts to generate a realistic sample of the image while the discriminator evaluates the differences between the sample generated by the generator and the real sample. In the competition and game between the generator and discriminator during adversarial training, the generator gradually learns to generate more realistic samples, while the discriminator learns to distinguish between the generated samples and real samples. In the field of medical image segmentation, Ma *et al.* (18) proposed a novel and universal bidirectional GAN named the structure and illumination-constrained GAN (StillGAN) for medical image quality enhancement. The StillGAN treats low- and high-quality images as two different domains, and introduces the local structure and lighting constraints to learn the overall features and local details (*Figure 2A*). Recently, transformer-based methods have emerged as powerful tools in medical image analysis (46). Chen *et al.* (47) innovatively incorporated a transformer model into medical image segmentation and created the TransUNet. This model efficiently encodes robust global contexts by treating image features as sequences, and integrates low-layer CNN features through a U-shaped hybrid design, synergizing the strengths of both the transformer and U-Net (*Figure 2B*). The ability of transformers to model long-range relationships and their scalability has made them highly effective in tasks such as lesion detection, organ segmentation, and disease diagnosis, setting new benchmarks in the field.

In response to the growing need for accurate segmentation, many researchers have proposed variations on existing structures over the decades. Milletari *et al.* (10) proposed a novel three-dimensional (3D) image segmentation approach using fully CNNs coupled with a unique objective function based on Dice coefficient maximization for end-to-end training. This method enables the segmentation of entire volumes, and has proven to be particularly beneficial in magnetic resonance imaging (MRI) analysis. Kayalibay *et al.* (11) proposed a 3D filtering technique for hand and brain MRI using CNNs. They made the following two modifications to their original CNN design: (I) the combination of multiple segmentation maps generated at different scales; and (II) the use of element-wise summation to pass feature maps from one network stage to another. These modifications were implemented to enhance the ability of the network to capture multi-scale features, and to improve segmentation accuracy. To address the challenge of class imbalance in medical image segmentation, they adopted a loss function based on the Jaccard similarity index, which is effective in handling uneven class distributions typical in medical images. Additionally, instead of segmenting high-resolution 3D images directly, they downsampled the images as necessary to reduce memory requirements, ensuring the feasibility of their approach for resource-constrained platforms. This approach not only improves segmentation performance but also addresses practical challenges that arise in deploying CNN-based methods for medical image analysis.

Zunair *et al.* (38) proposed a simple and efficient end-to-end deep encoder-decoder full convolutional network architecture called Sharp U-Net for binary and multiclass biomedical image segmentation. Sharp U-Net uses deep convolution with an encoder feature map with a sharpened

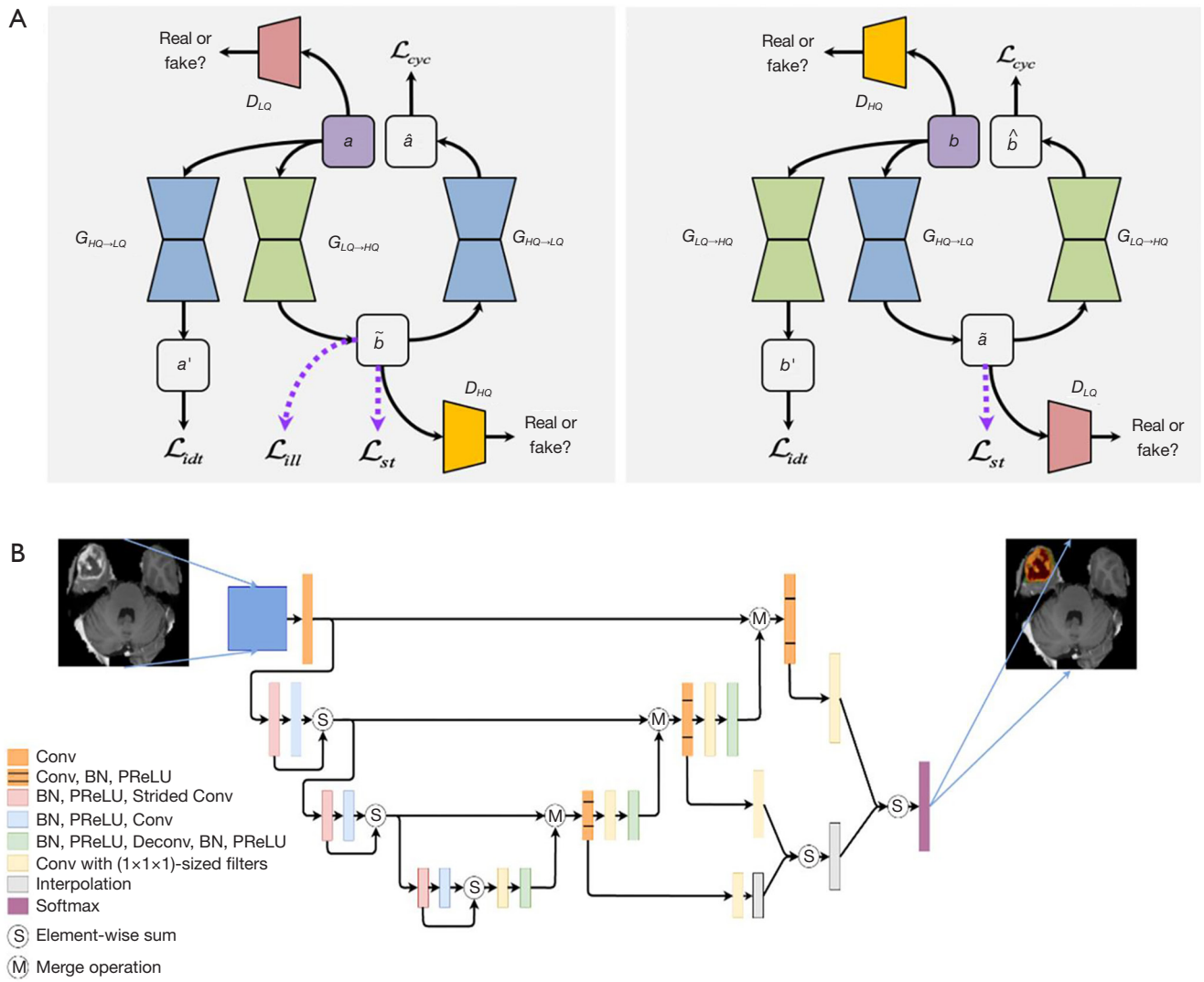

Figure 2 Some AI models: (A) an overall structure diagram of StillGAN; (B) an overview of TransUNet; the actual architecture is 3D. A 2D image is used here for the purpose of simplicity. AI, artificial intelligence; 3D, three-dimensional; 2D, two-dimensional; Conv, convolution; BN, batch normalization.

kernel filter before merging encoder and decoder features (instead of applying normal skip joins) to generate a sharpened intermediate feature map of the same size as the encoder map. Using this sharpening filter layer, it is not only possible to fuse semantically less similar features but also to smooth the artifacts of the entire network layer at an early stage of training (*Figure 3A*). Verma *et al.* (37) prepared a large and diverse dataset of nuclear boundary annotations and class labels, named MoNuSAC2020. The dataset contains more than 46,000 nuclei from 37 hospitals, 71 patients, four organs, and four cell types for the task of automating the detection, segmentation, and classification of nuclei. Zunair *et al.* (36) proposed a method called MaskSup to enhance semantic segmentation by modeling short- and long-context in images. By using random masking during training, MaskSup captures contextual relationships between pixels and improves segmentation performance, especially in fuzzy regions and minority classes. The method is computationally efficient, with a 10% increase in the average intersection-over-sum, and also uses three times fewer learnable parameters compared with the existing strong baselines model, and can be easily

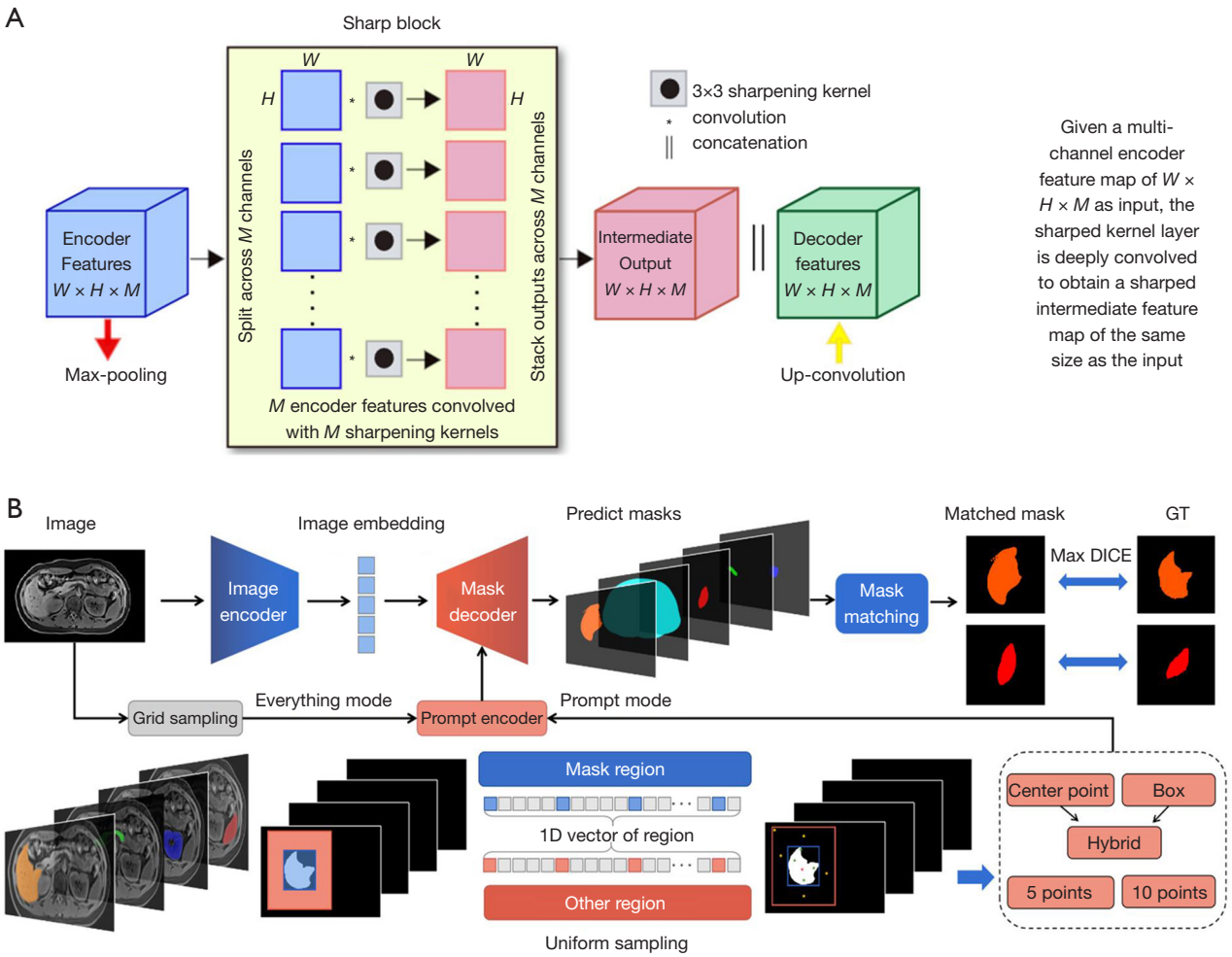

Figure 3 Some variations on existing structures: (A) Block of the Sharp U-Net; (B) testing pipeline of the SAM. W, width; M, mask; H, height; GT, ground truth; SAM, segmentation arbitrary model.

integrated into various semantic segmentation methods.

Miao *et al.* (32) proposed a new causal graph model to provide a theoretical basis for mainstream semi-supervised medical image segmentation methods. Based on this causal diagram, they introduce a causal heuristic semi-supervised learning method called Causal Self-Supervised Learning (CauSSL), which improves the existing co-training framework. CauSSL highlights the importance of algorithmic independence between networks or branches in self-supervised learning (SSL), enhanced by a new statistical quantification method and a minimum-maximum optimization process. Tragakis *et al.* (31) proposed a full convolution transformer (FCT) for medical image segmentation. It combines the strengths of CNNs and transformers to process inputs in two stages to capture remote semantic dependencies and hierarchical global properties. Compact and efficient, the FCT significantly outperforms existing transformer architectures on various datasets such as Automated Detection and Assessment of Cancer (ACDC), Synapse, Spleen, and International Skin Imaging Collaboration (ISIC) 2017 (4).

In a similar vein, Butoi *et al.* (34) proposed UniverSeg, which was designed to solve previously unseen medical image segmentation tasks without additional training. Traditionally, deep-learning models have performed well in medical image segmentation, but in the face of new anatomical structures, image modes, or labels, they often need to be retrained or fine-tuned, which is not only time

consuming but also challenging for clinical investigators. By introducing a new cross-block mechanism, UniverSeg is able to directly generate accurate segmentation graphs without additional training in the case of a given query image and a sample image-label pair that defines a new segmentation task. In addition, Gaillochet *et al.* (33) proposed a method that combines test time enhancement with active learning to improve the segmentation performance of a four-layer U-Net model. This approach improves the model's ability to generalize from finite labeled data by increasing test samples and incorporating indeterminate-based sample selection during training.

Continuing these innovations, Wang *et al.* (35) introduced Mamba-UNet, a new medical image segmentation architecture. It combines the advantages of U-Net (known for its encoder-decoder structure and skip connections) with the Mamba architecture, a state-space model (SSM) that is adept at handling long sequences and global context information. Mamba-UNet uses a pure visual Mamba (VMamba)-based encoder-decoder structure combined with an innovative integration mechanism in the VMamba block to ensure seamless connectivity and information flow. This design captures complex details and broader semantic context in medical images, thereby enhancing segmentation performance. The ACDC MRI cardiac and synaptic computed tomography (CT) abdominal segmentation datasets showed the effectiveness of Mamba-UNet over several U-Net variants.

Subsequently, Diaz-Pinto *et al.* (27) proposed the medical open network for AI (MONAI) Label, a free and open-source framework that facilitates application development based on AI models, and aims to reduce the time required to annotate radiology datasets. With the MONAI Label, researchers can develop AI annotation applications that focus on their area of expertise. It enables researchers to easily deploy their applications as services that can be made available to clinicians through their preferred user interface.

Huang *et al.* (28) studied the application of the segmentation arbitrary model (SAM) in medical image segmentation. The authors created the COSMOS 1050K dataset, consisting of 18 modes and 84 objects, to comprehensively evaluate the performance of the SAM (its test pipeline is shown in *Figure 3B*). They found that, especially when fine-tuned and used with manual prompts such as boxes, the SAM performed well on specific tasks but its performance in other tasks varied. This study highlighted the potential of SAM to improve segmentation accuracy and efficiency in medical imaging applications. Gao *et al.* (29) proposed a decoupled SAM (DeSAM) approach to address the performance degradation of deep-learning medical image segmentation models in domain migration. By modifying the SAM's mask decoder, the DeSAM introduced two new modules: the prompt-related Intersection over Union (IoU) module (PRIM); and the prompt-decouple mask module (PDMM). The PRIM predicts IoU scores and generates mask embeddings, while the PDMM extracts multi-scale features from the middle layer of the image encoder and fuses them with the PRIM's mask embeddings to generate the final segmentation mask. This decoupling design enables the DeSAM to use pre-training weights while reducing performance degradation due to bad prompts. Experimental results have shown that the DeSAM significantly improves performance in cross-site prostate and cross-modal abdominal image segmentation tasks, outperforming existing domain generalization methods.

Ruan and Xiang (30) proposed the Vision Mamba-UNet (VM-UNet), a U-shaped architecture based on the SSM, for medical image segmentation. The VM-UNet introduces visual state-space blocks as a base module to capture a wide range of contextual information and builds an asymmetric encoder-decoder structure. Through comprehensive experiments on ISIC17, ISIC18, and synapse datasets, the results showed that the VM-UNet performed well in medical image segmentation tasks. This is the first medical image segmentation model built on a pure SSM model, and the research findings provided benchmarks and valuable insights that could be used to develop more efficient SSM-based segmentation systems in the future.

Advantages of traditional AI algorithms

The incorporation of AI into medical image segmentation represents a transformative approach, yielding more accurate and detailed segmentation outcomes. This improved accuracy aids healthcare providers in making more informed decisions, thereby enhancing the precision and success rates of medical treatments. For example, in oncology, AI-driven segmentation can precisely delineate tumor boundaries, enabling targeted radiation therapy and reducing damage to surrounding healthy tissues. In urology, accurate segmentation of the prostate gland from imaging can assist in diagnosing and planning treatments for prostate cancer patients, ensuring that biopsies and surgeries are performed with high precision (29). AI-driven medical image segmentation is increasingly becoming an essential tool in both biomedical research and clinical

practice, significantly contributing to enhanced patient outcomes (48). For instance, in radiology, AI can standardize the interpretation of imaging results, ensuring consistency across different practitioners and institutions (11). In pathology, AI algorithms can provide uniform the analysis of biopsy samples, reducing the subjectivity and variability that may occur in manual evaluations (49). Further, the synergy between AI and healthcare introduces a systematic and standardized approach to medical treatments. This harmonization effectively reduces the variances related to the different expertise levels of medical practitioners, fostering a more uniform and reliable healthcare delivery system.

Disadvantages of traditional AI algorithms

Despite the advantages traditional AI methods have bought to segmentation efficiency, including reducing the segmentation time, diminishing subjective biases, and relieving doctors of the labor-intensive process of image segmentation (50), these methods are not without their challenges. First, traditional AI approaches are often criticized for their "black-box" nature, which includes problems such as inexplicable, unknowable, and uncontrollable reasoning decisions (51). Second, the lack of transparency in the decision-making processes hinders the understanding, debugging, and trust of AI applications in medical segmentation. Third, existing iterations of traditional AI exhibit vulnerabilities, including susceptibility to bias and adversarial attacks. These flaws not only undermine the integrity of the learning and reasoning processes of AI systems (52) but also raise significant concerns in relation to data privacy and the safety of individual personal data during extensive data processing tasks. For example, if the data storage or transmission channels are not secure, unauthorized access can lead to data breaches (53). Additionally, the emergence of AI algorithms may threaten the de-identification or anonymization of patient health data, thereby increasing the risk of patient data held under private data management (54). Therefore, transforming AI from an opaque "black box" into a transparent "white box" is a critical challenge in the field of medical segmentation. This transformation is crucial to ensure that end-users can understand, trust, and confidently adopt AI-driven decisions in medical image segmentation. Addressing these inherent issues of understanding and trust is essential to the widespread and effective application of AI in medical segmentation (55).

Medical image segmentation based on XAI

Owing to its interpretability and security, AI is of great significance in the field of medical image segmentation. In this section, we introduce a number of aspects of XAI (as detailed in *Figure 4*).

Problems in conventional AI segmentation models

As discussed further below, traditional AI methods face several challenges in the field of medical image segmentation. First, the AI models used in medical image segmentation often display inherent structural complexity. This complexity, combined with a significant lack of interpretability, hinders users' understanding of the models' decision-making mechanisms, thus limiting their clinical application. Second, AI-driven medical image segmentation methods typically require large volumes of labeled data for training. This demand presents a considerable challenge, as medical image data are subject to strict privacy protection protocols. Acquiring sufficient annotated data while adhering to privacy and ethical principles makes accessing large medical datasets difficult, which in turn poses a challenge to the development and refinement of AI models (56). Third, the presence of noise, artifacts, and various uncertainties in medical images further complicate matters. These imperfections can distort the predictive outcomes of AI models (57). Moreover, the diversity of complex lesions or disease states, where medical consensus is often lacking, presents inherent diagnostic challenges. In such cases, AI models may face challenges related to robustness in interpreting ambiguous and complex scenarios. Thus, AI-based approaches to medical image segmentation face several challenges, including model interpretation issues, age and racial bias, concerns related to robustness, and the ongoing problem of uncertainty in medical diagnoses. The advent of XAI offers potential solutions. XAI seeks to clarify AI decision-making processes, address these challenges, and improve reliability and trust in the use of AI for medical image segmentation.

XAI algorithm methods

Many traditional machine-learning algorithms, such as linear regression, decision trees, and logistic regression, generate inherently transparent model structures. However, dealing with complex real-world problems may require the use of more complex machine-learning models, such as deep neural networks (DNNs). These complex models may loose

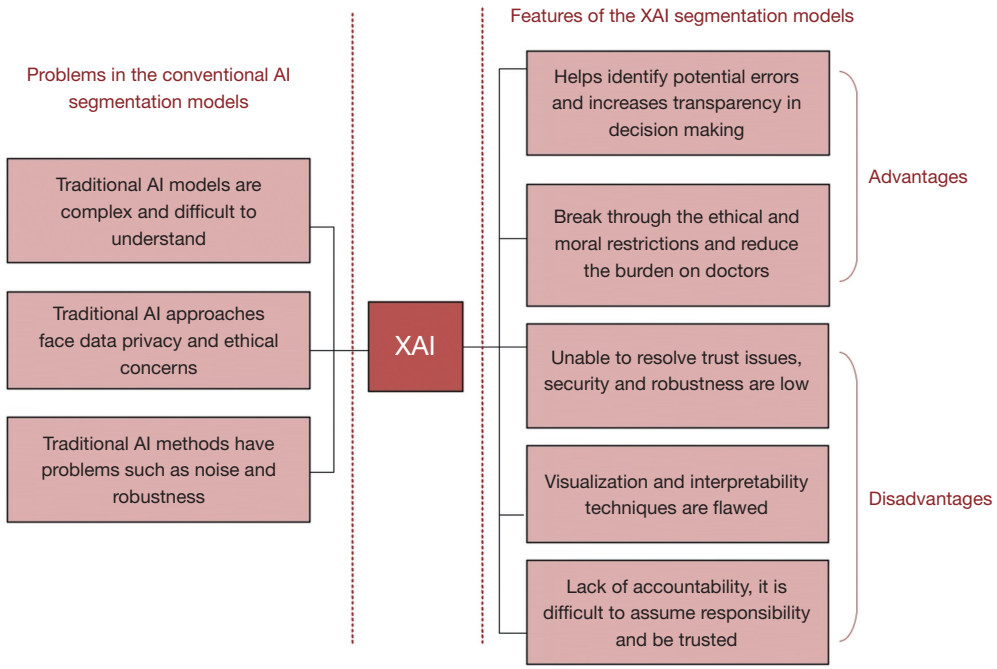

Figure 4 Medical image segmentation based on XAI. AI, artificial intelligence; XAI, explainable artificial intelligence.

a degree of interpretability because their internal structures and decision-making processes are not as intuitive and transparent as those of traditional models, preventing users from understanding their decision-making processes. This opacity is a significant barrier to the clinical application of such models. Additionally, the growing demand for ethical AI technologies that are transparent, manageable, and trustworthy intensifies these challenges (58-60). XAI emerged in response to these demands and to shed light on the "black box" of AI decision making. It is not only a model, it is also a set of methods and approaches designed to provide transparency and insight into the "black-box" nature of many AI models.

The interpretability approaches of XAI can be categorized as pre-, in-, and post-model, each corresponding to different stages of model development. Pre-model interpretability is independent of the model and applies exclusively to data, with techniques ranging from descriptive statistics to data visualization methods like principal component analysis, t-distributed stochastic neighbor embedding, and clustering. In-model interpretability (or intrinsic interpretability) clarifies the nature of the model, focusing on aspects such as sparsity, monotonicity, causality, external constraints, or the weights of the model. Interpretability can be categorized as either local or global based on its scope. Local interpretability focuses on specific prediction results, while global interpretability offers insights into the overall functioning of the model. In terms of methodology, interpretability is divided into alternative methods and visualization techniques. Alternative methods use different models for analyzing and clarifying "black-box" models, facilitating understanding through a comparative analysis of decision-making processes. Visualization techniques, while not providing alternative models, greatly assist in interpreting specific model components using visual elements, such as activation graphs (61). The above is summarized in *Figure 5*.

Several algorithms have been developed to enhance interpretability, including the local interpretable model-agnostic explanations (LIME) algorithm, saliency maps, layer-wise relevance propagation (LRP), integrated gradients (IGs), counterfactual explanation (CE), and gradient-weighted class activation mapping (Grad-CAM). The LRP method in particular has been widely applied in image-related fields and is increasingly used in natural language processing (62-64).

The LIME algorithm, introduced by Ribeiro *et al.* (65) in 2016, offers a unique approach to interpreting any classifier's prediction faithfully and understandably by approximating a local interpretable model around the

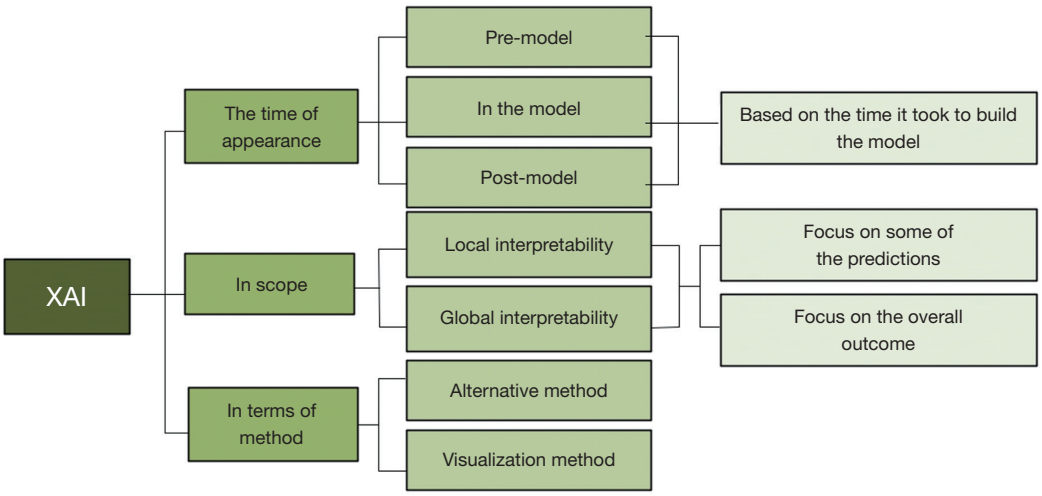

Figure 5 Classification of XAI. XAI, explainable artificial intelligence.

prediction. The core idea of LIME technology is to approximate the prediction results of the original model by generating a local linear model. It works by generating a set of "virtual samples" around a particular sample that resembles the original data, and using these virtual samples to train an interpretive model (such as a linear regression model). Then, by analyzing this explanatory model, an interpretation of the prediction for that sample can be obtained. However, this method shows instability, as a single prediction may lead to multiple interpretations (*Figure 6A*). The LIME assumption works for a small fraction around the instance, but for more complex data sets, the assumption may not hold as the local range increases. As a result, locally explainable models may not be able to explain the behavior of global linear models. In addition, LIME's interpretation lacks consistency due to sampling bias, similarity calculations, and neighborhood definitions.

Several researchers have further developed the LIME framework to address its instability and investigate its various features. Visani *et al.* (66) explored the trade-off between stability in interpretation and fidelity in the machine-learning model, and proposed a framework that maximizes stability while retaining a predefined level of fidelity. Zhang *et al.* (67) identified three sources of uncertainty in LIME related to the sampling process, sampling proximity changes, and variability in interpreted models across different data points. Many researchers have made suggestions to improve and enhance the LIME algorithm (68,69). LIME's image interpreter requires experts to manually adjust some parameters in advance, including the number of top-level features to be seen, and the number of superpixels in the segmented input image. However, parameter tuning is time consuming. Consequently, Nematzadeh *et al.* (70) developed an interpreter that automatically splits images for melanoma cancer detection; that is, the enhanced generative adversarial explainer (EGAE), which can automatically detect the informational part of the image and present it to the user. Compared to LIME, the EGAE also effectively improves interpretation accuracy.

Advances have also been made in using XAI technologies like LIME to explain predictions from complex image captioning models. As poor segmentation can compromise the consistency of interpretation and undermine the importance of segmentation, it affects overall interpretability. Duamwan and Bird (71) initially used CNN-based computer vision methods to detect Alzheimer's disease using Alzheimer's Disease Neuroimaging Initiative MRI datasets. They then implemented the LIME algorithm to reveal visual evidence supporting the predictions made by the model and used Felzenszwalb's segmentation algorithm to automatically visualize the fragments of the image that contributed to the predictions. Knab *et al.* (72) introduced data-driven segmentation LIME (DSEG-LIME), which is characterized by (I) data-driven segmentation for human recognition feature generation, and (II) hierarchical segmentation procedures by combination. Gaur *et al.* (73) proposed an interpretation-driven deep-learning model. CNNs, LIME, and SHapley Additive exPlanations (SHAP) were used to predict discrete subtypes of brain tumors

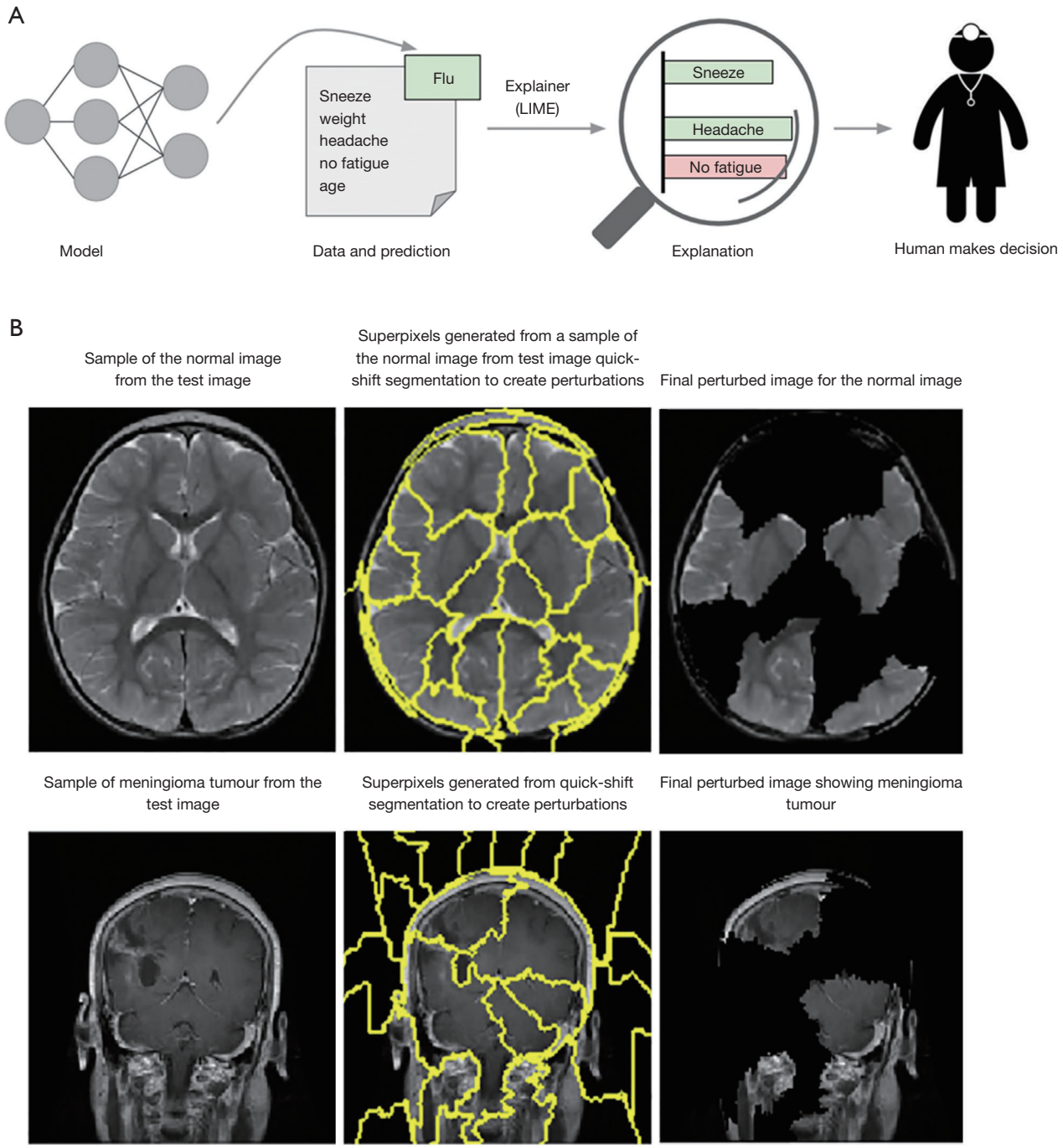

Figure 6 LIME. (A) LIME interpretation of individual prediction process; (B) interpretations generated by LIME. LIME, local interpretable model-agnostic explanations.

(meningiomas, gliomas, and the pituitary gland) using MRI image datasets. LIME builds sparse linear models around each prediction to illustrate how the model runs in the nearby region (*Figure 6B*). Ahsan *et al.* (74) used improved MobileNetV2 and LIME to detect coronavirus disease 2019 (COVID-19) patients from CT scans and chest X-ray (CXR) data, and the use of LIME helped to better understand which features in CT/X-ray images are characteristic of the

Table 3 LIME methods

Team	Technique	Advantage/s	Disadvantage/s
Ribeiro *et al.* (65)	LIME	Provides interpretability for any classifier's prediction	Instability—a single prediction can lead to multiple interpretations
Visani *et al.* (66)	Stability-Fidelity Framework (OptiLIME)	Balances stability and fidelity in interpretations	Needs a faster and more precise computation
Nematzadeh *et al.* (70)	EGAE	Automatically detects informational parts in images, which improves interpretation accuracy	Is a nonreproducible explainer, like most existing explainers
Duamwan *et al.* (71)	LIME with Felzenszwalb's segmentation algorithm	Reveals visual evidence supporting the predictions made by the model	On studying the confusion matrix, some errors are not equal to others when clinical implications are considered
Knab *et al.* (72)	DSEG-LIME	Predicts discrete subtypes of brain tumors	Segmentation is not possible when dealing with complex images
Gaur *et al.* (73)	Combination of CNN, LIME, and SHAP	Predicts discrete subtypes of brain tumors	Is a locally interpreted model with a model-agnostic explanation
Ahsan *et al.* (74)	Improved MobileNetV2 with LIME	Provides a better understanding of the features in CT/X-ray images characteristic of the onset of COVID-19	Findings must be validated in consultation with a healthcare professional

LIME, local interpretable model-agnostic explanations; EGAE, enhanced generative adversarial explainer; DSEG-LIME, data-driven segmentation LIME; CNN, convolutional neural network; SHAP, SHapley Additive exPlanations; CT, computed tomography; COVID-19, coronavirus disease 2019.

onset of COVID-19. These ongoing efforts and innovations aim to enhance the interpretability and trustworthiness of AI models across various applications. Some of the LIME methods are listed in *Table 3*.

Itti *et al.* (75) introduced a visual attention system that effectively combines multi-scale feature maps into a saliency map of the terrain. The saliency map explains the decision-making process of the AI model by highlighting key areas in the input image. In medical images, saliency maps can reveal the features the model focuses on, showing which parts have the most impact on the diagnosis. By visualizing the model's area of attention, salience maps enable doctors and researchers to understand how the model works and identify possible biases or bad decisions in the model, thereby increasing the transparency and credibility of AI models. This system enables the rapid identification and prioritization of salient locations, facilitating efficient in-depth analysis (*Figure 7A*). Conversely, Simonyan *et al.* (76) presented a method for computing saliency maps for specific image classes, highlighting relevant image areas; however, this approach is prone to noise. Kim *et al.* (77) suggested that such noise might result from irrelevant features passing through the rectified linear unit activation function in the saliency graph. They introduced the rectified gradient, a technique designed to reduce this noise during backpropagation through layer-wise thresholding. This method outperformed other attribution techniques in experiments conducted on networks trained with CIFAR-10 and ImageNet. Using two extensive public radiology datasets, Bernal *et al.* (78) developed the Window Median Depth of Valleys (WM-DOVA) energy map for efficient polyp localization in colonoscopy images (*Figure 7B*). Li (79) showed that a pre-attention computation mechanism in the primary visual cortex generates a saliency map, explaining task difficulty based on the characteristics and arrangements of targets and distractors. Morch *et al.* (80) proposed saliency maps as a new way to understand and visualize the non-linearities embedded in feedforward neural networks, applying saliency maps to medical imaging (positron emission tomography scanning) to identify paradigm-related regions in the human brain. Gadgil *et al.* (81) proposed a method that combines high-quality pixel-level expert labeling with the saliency map scale generated by a coarse DNN to train a multi-label semantic segmentation model. Sun *et al.* (82) introduced a new architecture called the shape attentive U-Net (SAUNet), which aims to address the interpretability and robustness of traditional CNNs in medical image segmentation. The SAUNet captures rich

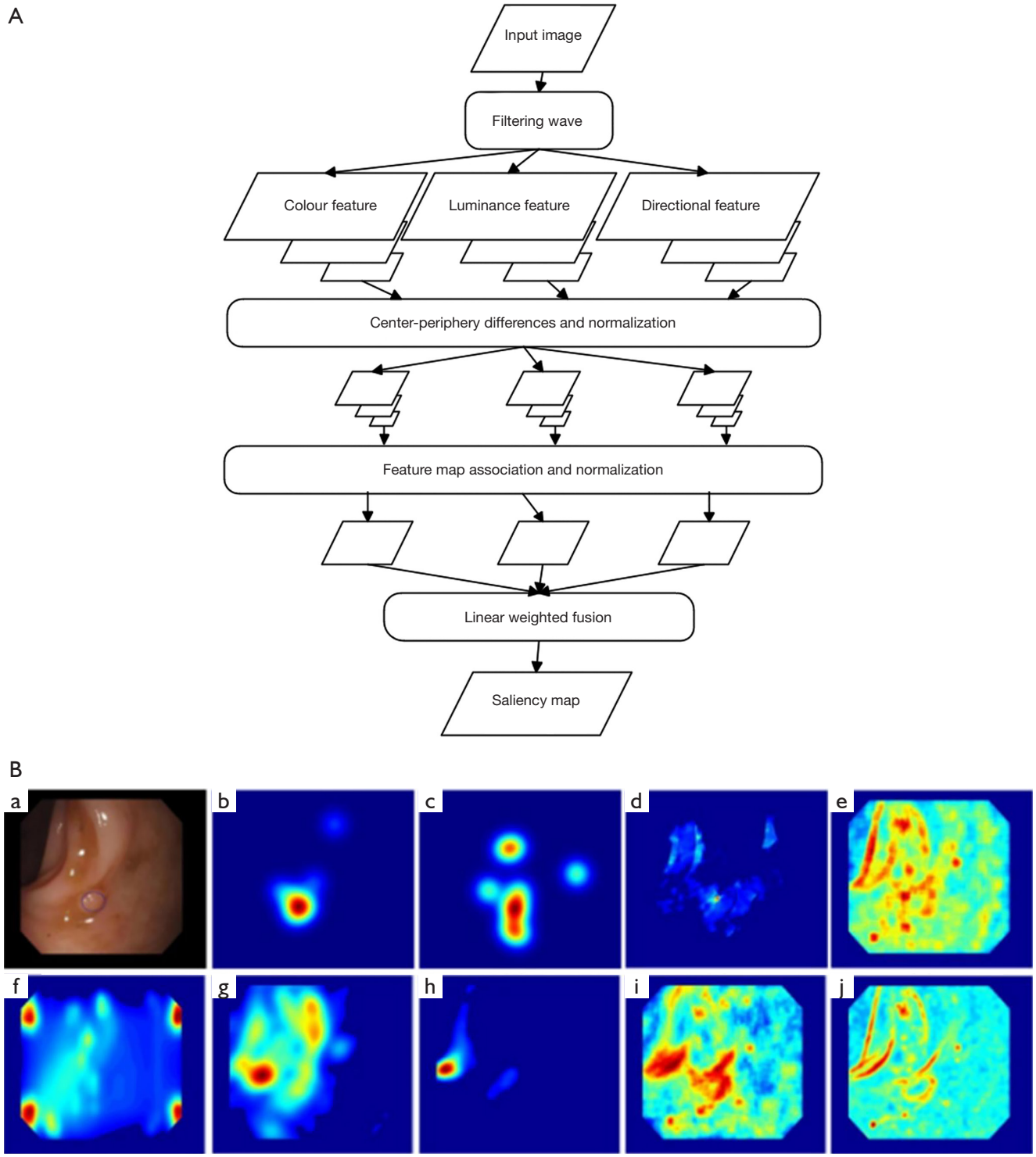

Figure 7 Saliency maps. (A) Overall architecture of saliency maps. (B) Comparison of saliency maps represented as heat maps: (a) original image with polyp mask superimposed; (b) average expert; (c) average novice; (d) WM-DOVA energy map; (e) Bruce and Tsotsos model; (f) GBVS; (g) Itti-Koch model; (h) Seo model; (i) SIM; (j) SUN. High saliency areas correspond to hot regions in the image. WM-DOVA, Window Median Depth of Valleys; GBVS, Graph-Based Visual Saliency; SIM, Saliency Image Model; SUN, Saliency Understanding Network.

Table 4 Saliency map methods

Team	Technique	Advantage/s	Disadvantage/s
Itti *et al.* (75)	Introduced saliency maps	Enables rapid identification and prioritization of salient locations, facilitating efficient analysis	Unimplemented feature types cannot be detected
Simonyan *et al.* (76)	Presented a method for computing saliency maps for specific image classes	Can be used to compute saliency maps for specific image classes	Prone to noise
Kim *et al.* (77)	Introduced the rectified gradient	Reduces noise during backpropagation through layer-wise thresholding, outperforming other attribution techniques	–
Bernal *et al.* (78)	Developed the WM-DOVA energy map for efficient polyp localization in colonoscopy images	Efficiently localizes polyps	Does not apply any kind of spatial or temporal coherence
Morch *et al.* (80)	Applied saliency maps to medical imaging (positron emission tomography scans) to identify paradigm-related areas	Visualizes non-linearities in neural networks	–
Gadgil *et al.* (81)	Combined high-quality pixel-level expert labeling with saliency map scale from coarse DNNs	Trains a multi-label semantic segmentation model with expert labeling and a coarse DNN saliency map	Reliance on expert pixel-level labeling and sensitivity to weakly supervised data may limit its wide applicability
Sun *et al.* (82)	Used a dual attention decoder module to learn multi-resolution saliency maps	Achieves multi-level interpretation capabilities and reduces additional calculations	–
Ning *et al.* (83)	Generated and used salience maps with low- and high-level image structures	Guides main and auxiliary networks to learn foreground and background salience feature representations	–

WM-DOVA, Window Median Depth of Valleys; DNN, deep neural network.

shape-related information by introducing parallel shape flows, which are combined with regular texture flows. In addition, the study proposed the use of a dual attention decoder module to learn multi-resolution saliency maps to achieve multi-level interpretation capabilities and reduce subsequent additional calculations. In experiments, SAUNet has achieved state-of-the-art results using public cardiac MRI image segmentation datasets such as SUN09 and AC17. Ning *et al.* (83) proposed a novel architecture called the saliency-guided morphology aware U-Net (SMU-Net) for lesion segmentation in breast ultrasound images. The SMU-Net first generates and uses salience maps, combined with low- and high-level image structures, to guide the main network and the auxiliary network in learning foreground and background salience feature representations, respectively. In addition, the SMU-Net also included an intermediate stream to effectively improve the ability of the network to learn morphological information in breast ultrasound images through background-assisted fusion, shape perception, edge perception, and position awareness units. Some of the saliency map methods are listed in *Table 4*.

Bach *et al.* (84) introduced the LRP method, a technique that visualizes each pixel's contribution to a kernel-based classifier's prediction on bag-of-words features and multi-layer neural networks. Current best practices for applying LRP rely on empirical human observation. To address this issue, Kohlbrenner *et al.* (85) empirically examined these practices in the context of feedforward neural networks engaged in visual object detection. Their findings confirm that recent research on LRP methods focused on specific layers and provided a more accurate representation of model reasoning, improving object localization and class discrimination. Ahmed and Ali (86) realized explainable medical image segmentation through GANs and LRP (*Figure 8A*). LRP is used to provide an interpretation of the prediction, specifying which input image pixels are

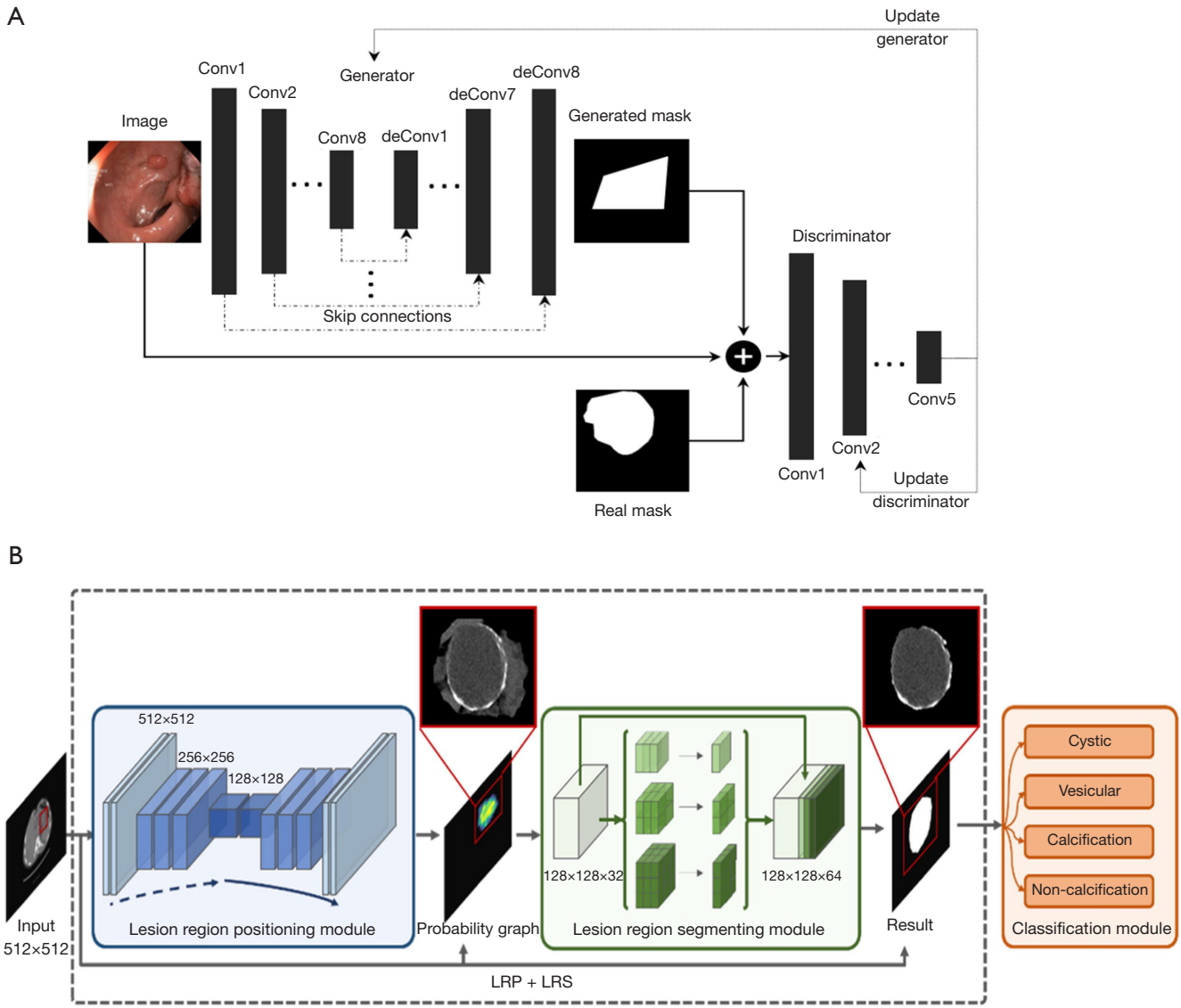

Figure 8 LRP. (A) Architecture of the model proposed by Aham *et al.* (B) The pipeline of the method, including the LRP module, the LRS module, and the classification module. LRP, layer-wise relevance propagation; LRS, lesion region segmentation.

relevant to the prediction and to what extent. In the polyp segmentation task, the accuracy of the model was 0.84, and the Jaccard index was 0.46. In the instrument segmentation task, the accuracy of the model was 0.96, and the Jaccard index was 0.70. In addition, Alam *et al.* (87) explored the application of the LRP algorithm in the interpretation of chest radiological images. LRP explains the decision-making process of DNNs by propagating importance layer by layer. By applying a multi-label classification model to the CheXpert dataset, they visualized the heat map results of LRP on chest radiology images. The results of the study showed that LRP produced a more fine-grained heat map than the Grad-CAM. Ni *et al.* (88) applied LRP rules to the 3D-CNN model and analyzed the model using a new decision decomposition strategy called (λ)n rules. They decomposed the classification decisions of the network layer by layer from the output layer, and finally obtained a contribution matrix with the same dimensions as the input data, highlighting the high contribution factors using visualization. Experiments were carried out using a 3D

Table 5 LRP methods

Team	Technique	Advantage/s	Disadvantage/s
Bach *et al.* (84)	LRP	Visualizes each pixel's contribution to the classifier's prediction; provides insight into model reasoning	Relies on empirical observation for best practices
Ahmed *et al.* (86)	LRP in chest radiological image interpretation	Provides interpretation of predictions in medical image segmentation; identifies relevant pixels and their contribution to predictions	Requires complex setup and tuning of GANs
Alam *et al.* (87)	LRP rules in a 3D-CNN model	Explains decision making in deep networks for chest radiology; produces fine-grained heat maps for better visualization	Interpretations can be subjective; heat maps might not always generalize well
Ni *et al.* (88)	LRP for liver echinococcosis lesion segmentation	Decomposes classification decisions layer by layer, providing detailed contribution matrices; improves model interpretability	Complexity increases with deeper networks' computational overhead for 3D data

LRP, layer-wise relevance propagation; 3D-CNN, 3D convolutional neural network; GAN, generative adversarial network.

mnist dataset and an additional 3D-MRI dataset, and their model achieved better interpretation results than previous work. Xin *et al.* (89) proposed a novel automated hepatic echinococcosis (HE) lesion segmentation and classification network, including LRP and lesion region segmentation (LRS) modules. The LRP module first determines the lesion location by generating a probability map of the lesion distribution and then uses this information to provide the basis for the LRS module to help the latter accurately segment HE lesions within the high-probability region (*Figure 8B*). Each of these varied approaches contributes unique insights and methodologies to the understanding and application of LRP in different contexts and use cases. Some of the LRP methods are listed in *Table 5*.

The IG method was proposed by Sundararajan *et al.* (90). The method aims to explain the behavior of deep-learning models by calculating the contribution of input features to model predictions. The IG method determines the importance of each feature to the model output by calculating the gradient integral of the input features. Specifically, it accumulates gradients along a straight path from the reference input (usually zero or the mean) to the actual input to generate feature contributions (*Figure 9A*). When IGs are used to interpret DNNs in medical image segmentation, noise is often generated in regions unrelated to the predicted class. To reduce these noises, Kapishnikov *et al.* (91) proposed adaptive path methods, such as guided IG. These methods reduce the accumulation of noise by adjusting the path so that it is dependent not only on the image but also on the model being interpreted. The experimental results show that guided IG can better align the model prediction with the input image, and the generated saliency map is more suitable for medical image segmentation (*Figure 9B*). Some of the IG methods are listed in *Table 6*.

In a notable development, Wachter *et al.* (92) introduced the concept of unconditional CE as a novel approach to clarify automated decision-making processes. This method addresses various challenges in algorithmic explainability and accountability. By using counterfactual interpretations, this approach not only adheres to the stringent requirements of the General Data Protection Regulation regarding automated decision making but also provides valuable insights, based on the local model's varying responsiveness to the choice of scale. Thiagarajan *et al.* (93) proposed the Trajectory-based Causal Explanation (TraCE) technique to reliably generate counterfactual interpretations using an indetermination-based calibration strategy, thereby demonstrating its superiority in radiology, especially in-depth models of CXR anomaly detection. In this way, TraCE helps detect shortcuts in model decisions and understand the relationship between patient attributes and disease severity (*Figure 10A*). Singh *et al.* (94) proposed a DeepLabv3+ optimized deep-learning method with ResNet-101 as the backbone for the counterfactually interpretable segmentation of gastrointestinal and colonoscopy images. This method showed stable performance in the segmentation of large and small medical objects. Zhou *et al.* (95) proposed Sparse CounteRGAN (SCGAN) to generate counterfactual instances to reveal causal relationships between image phenotypes, clinical information, molecular signatures (Integrated Clinical-Molecular signatures), and treatment responses. Through this method, it is possible to understand how different

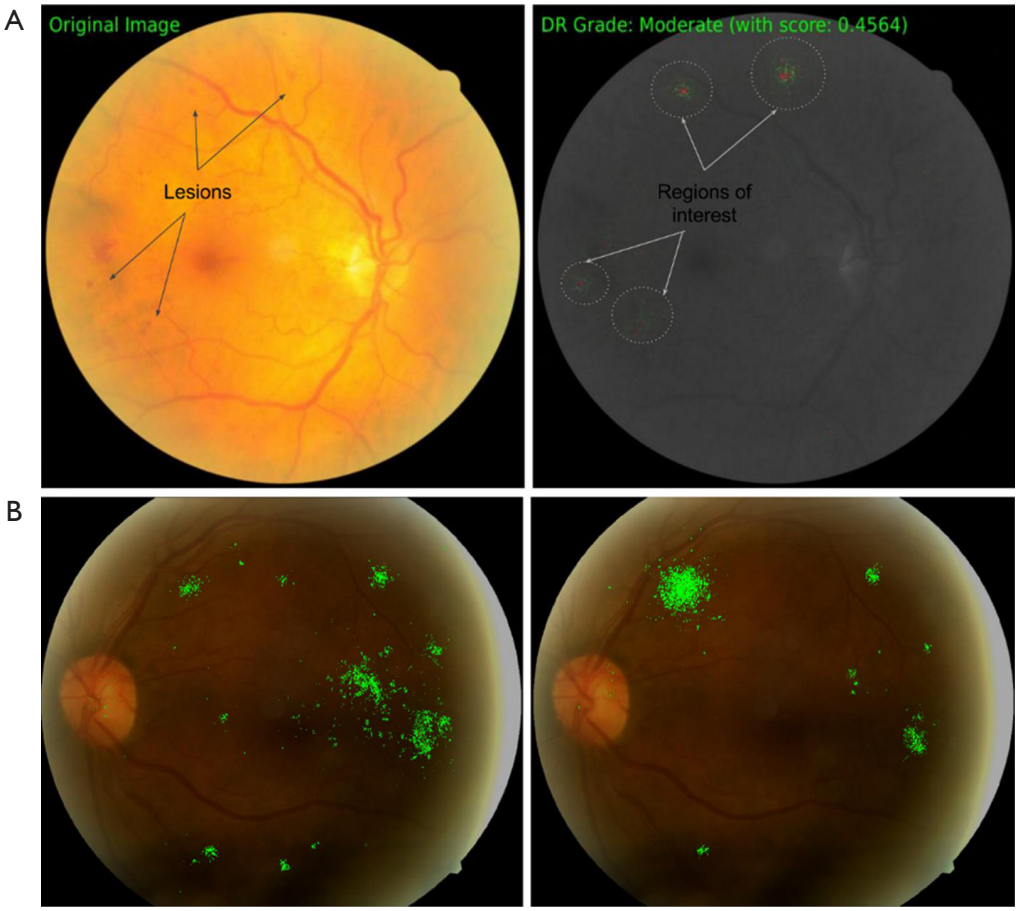

Figure 9 IG. (A) Prediction of grade attribution of diabetic retinopathy from retinal fundus images. The original image is displayed on the left, and the properties (overlaid in grayscale on the original image) are displayed on the right. In the original images, the study annotated the lesions visible to humans and confirmed that the attributes pointed to them. (B) Comparison of IG and guided IG in the diagnosis of diabetic retinopathy. IG, integrated gradient; DR, diabetic retinopathy.

Table 6 IG methods

Team	Technique	Advantage/s	Disadvantage/s
Sundararajan *et al.* (90)	IGs	Quantifies feature importance by gradient integration, widely used in the field of image segmentation	Generates noise in unrelated regions
Kapishnikov *et al.* (91)	Guided IG (adaptive path methods)	Reduces noise accumulation and better aligns predictions with input images' learning effectiveness; Enhances model interpretability	Potentially higher computational complexity

IG, integrated gradient.

features affect the treatment effect on the basis of medical image segmentation. SCGAN learns the distribution of the original image data, thus ensuring that the generated counterfactual instances are authentic and interpretable in medical image segmentation tasks. Lenis *et al.* (96) introduced a counterfactual impact analysis-based approach for interpreting medical image classification models. By measuring the influence of local image perturbation on the

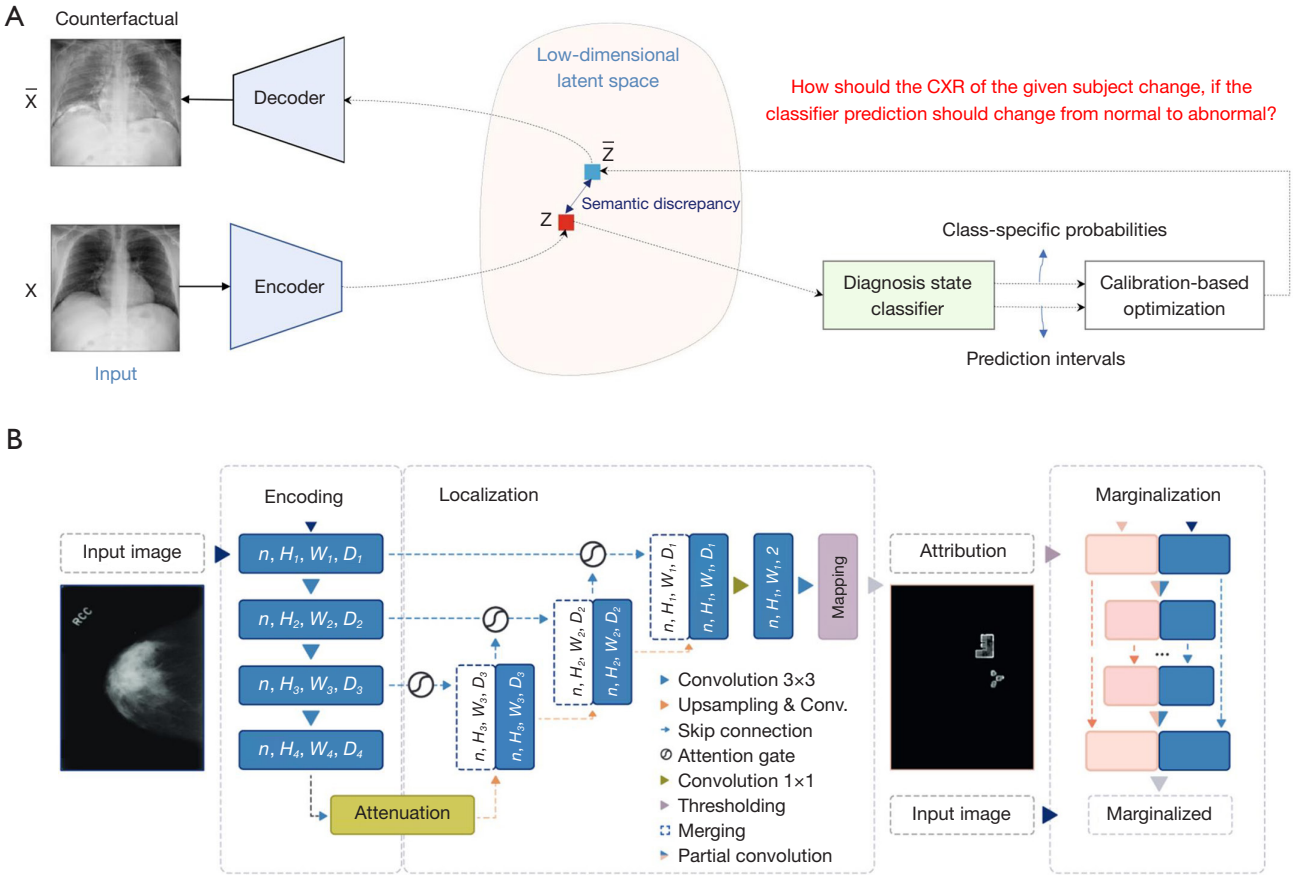

Figure 10 Counterfactual interpretation. (A) An overview of TraCE applied to the introspective analysis of CXR-based predictive models; (B) attribution framework. CXR, chest X-ray; H, height; W, width; D, depth; Conv., convolution.

model prediction results, the method significantly improves the accuracy and clarity of the interpretation results and avoids the time limitation, ambiguity, and misleading problems caused by traditional heuristic techniques, such as Gaussian noise and fuzzy processing (*Figure 10B*). Some of the CE methods are listed in *Table 7*.

Selvaraju *et al.* (97) proposed Grad-CAM, which uses the gradients of any target concept that flow into the final convolutional layer to generate a rough location map that highlights important areas in the image to predict concepts (*Figure 11A*). Grad-CAM is suitable for a variety of CNN model families, including: (I) CNNs with full connectivity layers; (II) CNNs used to generate structured outputs; and (III) CNNs used for tasks with multimodal inputs or reinforcement learning without any architectural changes or retraining. Xiao *et al.* (98) proposed an improved Grad-CAM visualization method. First, the output mask of the segmentation model is converted into a column vector, and the segmentation is carried out by setting a threshold strategy. Second, to generate the visualization result of the medical image segmentation model, the sum of pixels exceeding the threshold is backpropagated to obtain the contribution of the input pixels to the segmentation result. Third, the proposed method is applied to three medical image segmentation models (i.e., Double U-Net, R2U-Net, and MCGU-Net), and the effectiveness of the proposed method is verified in three medical image datasets, and accurate interpretable heat maps are generated. Chen *et al.* (99) proposed a new causal CAM (C-CAM) method to solve the problem of unclear foreground-background boundaries and serious co-occurrence in the weakly supervised semantic segmentation of medical images (*Figure 11B*). Using two causal chain models (i.e., the categorical causal chain and anatomical causal chain models), the C-CAM approach achieved leading pseudo-mask generation and organ segmentation performance

Table 7 Counterfactual explanation methods

Team	Technique	Advantage/s	Disadvantage/s
Wachter *et al.* (92)	Introduced the unconditional CE	Clarifies automated decision-making processes; Addresses General Data Protection Regulation requirements; provides valuable insights	Potential complexity in generating and interpreting counterfactuals accurately
Singh *et al.* (94)	Proposed a DeepLabv3+ optimized deep-learning method with ResNet-101	Shows stable performance in segmenting gastrointestinal and colonoscopy images	Computational complexity may limit real-time applications
Zhou *et al.* (95)	Developed a technique specifically for generating counterfactual visual interpretations	Produces interpretable and discriminative counterfactual interpretations for image classification	Dependent on the quality and diversity of generated counterfactuals for robust interpretation
Lenis *et al.* (96)	Introduced a counterfactual impact analysis-based approach	Improves accuracy and clarity of the interpretation results in medical image classification	May require significant computational resources for comprehensive impact analysis

CE, counterfactual explanation.

on multiple public medical image datasets and provides open-source code. Vinogradova *et al.* (100) proposed Segmentation-based Grad-CAM (SEG-GRAD-CAM), a gradient-based approach to interpreting semantic segmentation. The method is an extension of the widely used Grad-CAM method and has been applied locally to generate heat maps, showing the relevance of individual pixels to semantic segmentation. As SEG-Grad-CAM does not use spatial information when generating interpretations for regions in a segmented graph, drawing inspiration from HiResCAM, Hasany *et al.* (101) proposed Segmentation-based Explainable Residual Class Activation Mapping (Seg-XRes-CAM), which improves the generation of segmented graph interpretations by incorporating spatial context. The effectiveness of Seg-XRes-CAM was verified by a visual comparison with the SEG-GRAD-CAM and model-independent Randomized Input Sampling for Explanation (RISE) methods, which performed well in highlighting relevant areas in segmentation plots. Some Grad-CAM methods are listed in *Table 8*.

The Grad-CAM algorithm is noted for its robustness and effectiveness in visualizing and interpreting the decisions of CNNs, making it particularly useful in contexts in which understanding the spatial localization of features is crucial. Conversely, the LIME algorithm is less effective in comprehensively analyzing all models. While the Grad-CAM is highly effective for CNNs, algorithms such as IG, LRP, saliency maps, and CE also show high efficacy in specific architectural contexts (102).

In the field of medical applications, XAI has garnered substantial attention (103-105). Noteworthy contributions include Attention-based Lugnet Segmentation (A-LugSEG), a framework developed by Peng *et al.* (8,106,107) for lung segmentation in CXR images, and Hierarchical Progressive Segmentation (H-ProSeg), designed for prostate segmentation in transrectal ultrasound images. These frameworks address specific challenges in their respective domains. Using XAI instead of traditional CNNs in this dataset allows for greater transparency and interpretability, which is essential for understanding and validating the decision-making process in critical medical applications. In skin cancer classification, Young *et al.* (108) focused on enhancing interpretability. Their research involved the use of Grad-CAM for feature analysis in CNN models, specifically for melanoma detection. Further, various researchers have introduced innovative architectures and methods to improve the precision, interpretability, and robustness of medical image segmentation and classification. Significant advancements include the creation of Vid-NET for COVID-19 detection from CXR images (109), and the development of the comprehensive attention-based CNN (CA-Net) (110). Kaur *et al.* (111) proposed an XAI-based medical image segmentation model, GradXcepUNet, which combines the segmentation capability of the U-Net with the interpretability feature of Grad-CAM's Xception classification network. The Grad-CAM-trained image highlights key areas of the Xception classification network. Then, as a guide, the visualized results of the key regions are combined with the existing segmentation model (the U-Net) to produce the final segmentation results. Abeyagunasekera

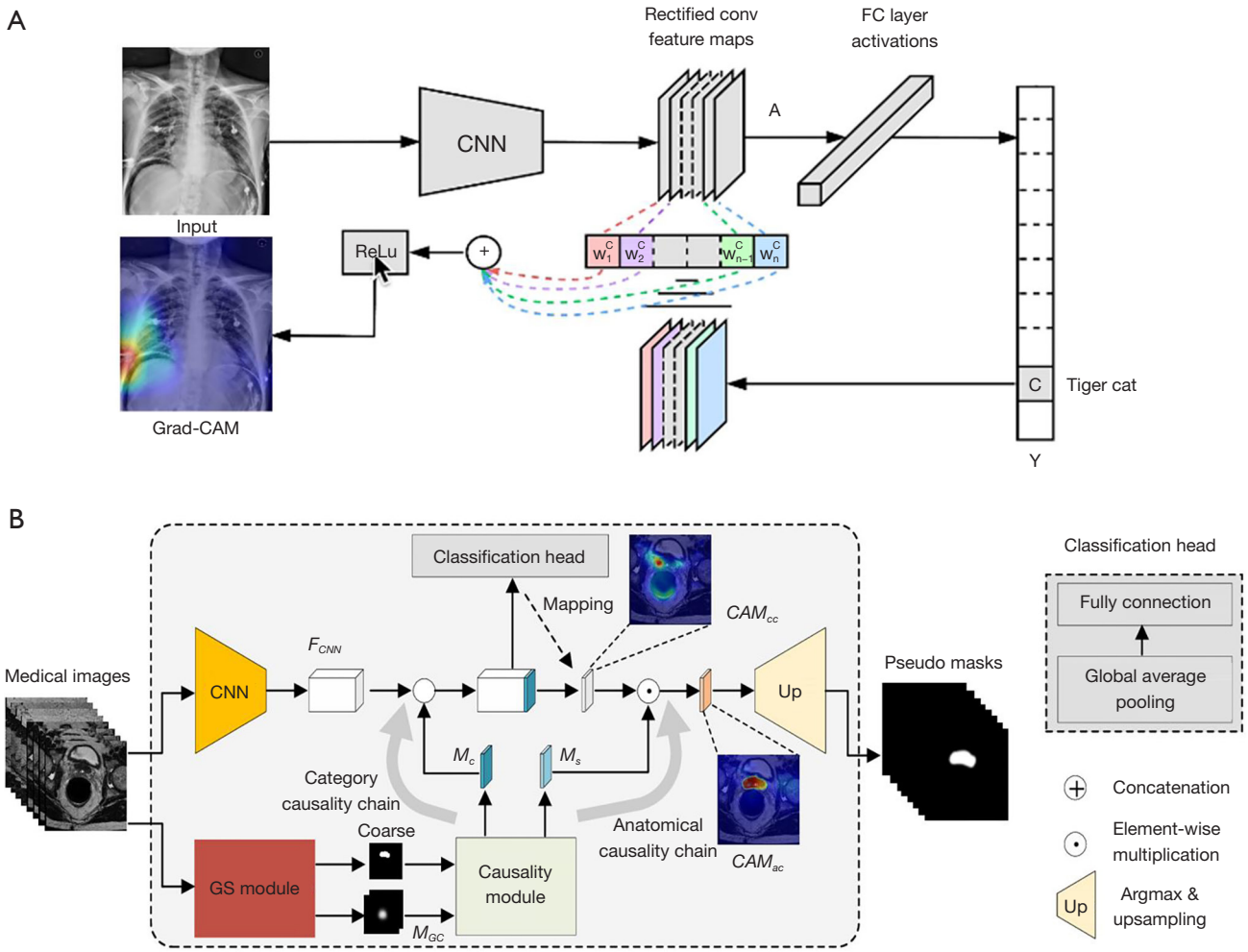

Figure 11 Grad-CAM. (A) Architecture of Grad-CAM. (B) Architecture of C-CAM. Grad-CAM, gradient-weighted class activation mapping, C-CAM, causal class activation mapping; CNN, convolutional neural network; RNN, recurrent neural network; LSTM, long short-term memory; FC, fully connected; GS, global structure.

Table 8 Grad-CAM methods

Team	Technique	Advantage/s	Disadvantage/s
Selvaraju *et al.* (97)	Grad-CAM	Highlights important image areas for concept prediction	May not capture fine details
Xiao *et al.* (98)	Improved Grad-CAM visualization method	Enhances the visualization of critical image regions	Specific advantages and disadvantages were not detailed in the text
Chen *et al.* (99)	C-CAM method	Improves foreground-background boundaries in segmentation	Detailed pros and cons were not specified in the provided information
Vinogradova *et al.* (100)	SEG-GRAD-CAM	Provides local interpretations for segmentation using gradients	May not fully use spatial information for interpretation
Hasany *et al.* (101)	Seg-XRes-CAM	Enhances segmented graph interpretations with spatial context	Requires careful adjustment for spatial context and clarity

Grad-CAM, gradient-weighted class activation mapping; C-CAM, causal class activation mapping.

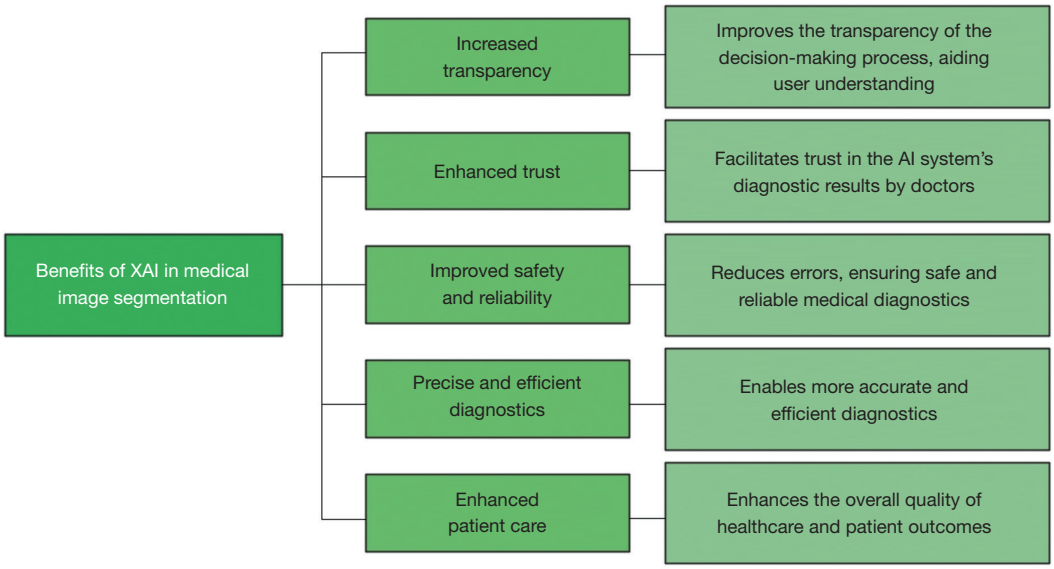

Figure 12 Benefits of XAI on medical image segmentation. XAI, explainable artificial intelligence; AI, artificial intelligence.

et al. (112) used currently available XAI techniques to enhance the interpretability of CNN predictions on medical images. A CXR classification model for identifying COVID-19 patients was trained using transfer learning to show the applicability of XAI techniques and a Unified Approach (Locally Interpretable Model-agnostic Explanations: LISA) to interpret model predictions. Ghnemat *et al.* (113) introduced an XAI model for medical image classification to enhance the interpretability of decision-making processes. The image segmentation-based method provides a better understanding of how AI models arrive at their results.

Advantages of XAI algorithms

In AI, interpretability is crucial for identifying and correcting potential errors, thus improving the model through systematic debugging (114). XAI facilitates a set of tools, techniques, and algorithms that can produce high-quality, interpretable, intuitive, human-understandable explanations of AI decisions (24). This shift in understanding allows users to grasp the rationale behind AI's choices, methodologies, and the content involved in decision making, effectively transforming the "black-box" nature of AI systems into more transparent "white-box" models. In the field of medical image segmentation, XAI technology can improve the transparency of the model and gain the trust of doctors, thus increasing the safety and reliability of medical imaging (115). Consequently, the improved comprehension and application of AI decisions in medicine lead to more precise and efficient diagnostics and treatments, enhancing patient care and healthcare outcomes (61). *Figure 12* shows the effect of XAI on medical image segmentation as described by Abeyagunasekera *et al.* in studies (61,112).

Disadvantages of XAI algorithms

Despite the valuable insights provided by XAI in interpreting AI decisions and addressing biases, it does not consistently engender trust among users (116). First, numerous challenges inherent to XAI systems hinder their ability to deliver accurate judgments uniformly across varied tasks. The ability to explain is often the most important benefit offered by expert systems, but these systems are built entirely on subject matter expertise, and while powerful, are somewhat inflexible and difficult to use (117). Second, the visualization and interpretability techniques used in XAI have inherent limitations. Specifically, while useful in decision making, the interpretative graphs produced by XAI are constrained in terms of their completeness and accuracy, challenges that defy quantitative assessment (24). These intrinsic limitations prevent human users from fully using these visual tools for informed decision making. Third, the implementation of XAI in medicine brings unique considerations and challenges, especially in relation

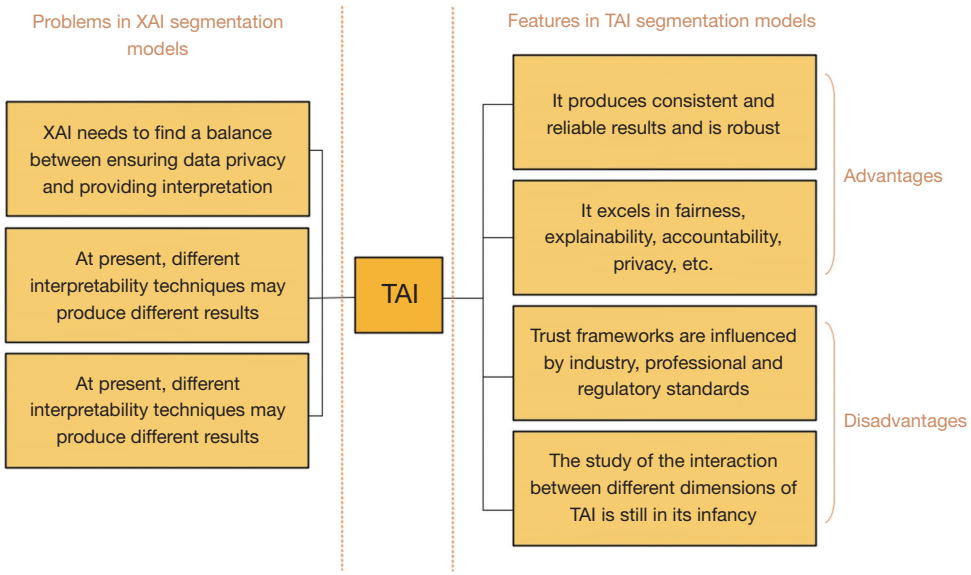

Figure 13 Medical image segmentation based on TAI. XAI, explainable artificial intelligence; TAI, trustworthy artificial intelligence.

to risk and liability (118). The concept of accountability is particularly crucial in the medical field, where decisions can have significant and often irreversible impacts. However, current XAI systems lack sufficient mechanisms for accountability, impeding their ability to ensure users and operators accept responsibility, and thereby undermining the trust placed in them by end-users (119).

Delegating critical decisions to systems lacking transparent accountability frameworks equates to a complete abdication of responsibility. Therefore, cultivating user trust in XAI systems presents a significant challenge that requires urgent and thoughtful consideration. This is essential if the full potential of XAI is to be harnessed in various application domains, especially in healthcare.

Medical image segmentation based on TAI

Due to advantages related to its robustness, accountability, and fairness, TAI is of great significance in the field of medical image segmentation. In this section, we will introduce a number of aspects of TAI, as detailed in *Figure 13*.

Problems in XAI conventional segmentation models

In the medical sector, AI algorithms analyze extensive personal data, raising significant privacy concerns. It is vital for AI to maintain a careful balance in interpreting sensitive data and providing insightful explanations without compromising confidentiality and privacy. A system's owners and operators must establish user trust by convincingly demonstrating its commitment to safety, security, fairness, and privacy (120). Moreover, there is a recognized bias in the training data for medical image segmentation models, largely sourced from specific demographic groups or centers. This selectivity inadvertently introduces age, gender, and racial biases, which could lead to inconsistent model performance across diverse patient groups. Such inconsistencies could negatively impact the learning and reasoning processes (52). Due to their inherent diversity and complexity, medical images, which cover various pathological types, organs, and data formats, require algorithms with high robustness and generalization capabilities for effective segmentation. However, the intricate nature of the data implies that different interpretability techniques might yield varying interpretation results, leading to inconsistencies. Such inconsistencies can undermine physicians' trust and acceptance of AI models, presenting significant barriers to their practical application. Additionally, the field of XAI has not adequately addressed critical aspects such as risk, responsibility, and accountability, resulting in continued user reluctance to fully trust these systems (119).

In response to these challenges, the development of TAI, which is characterized by reliable features and supported by trusted technology, has become a key focus in AI research

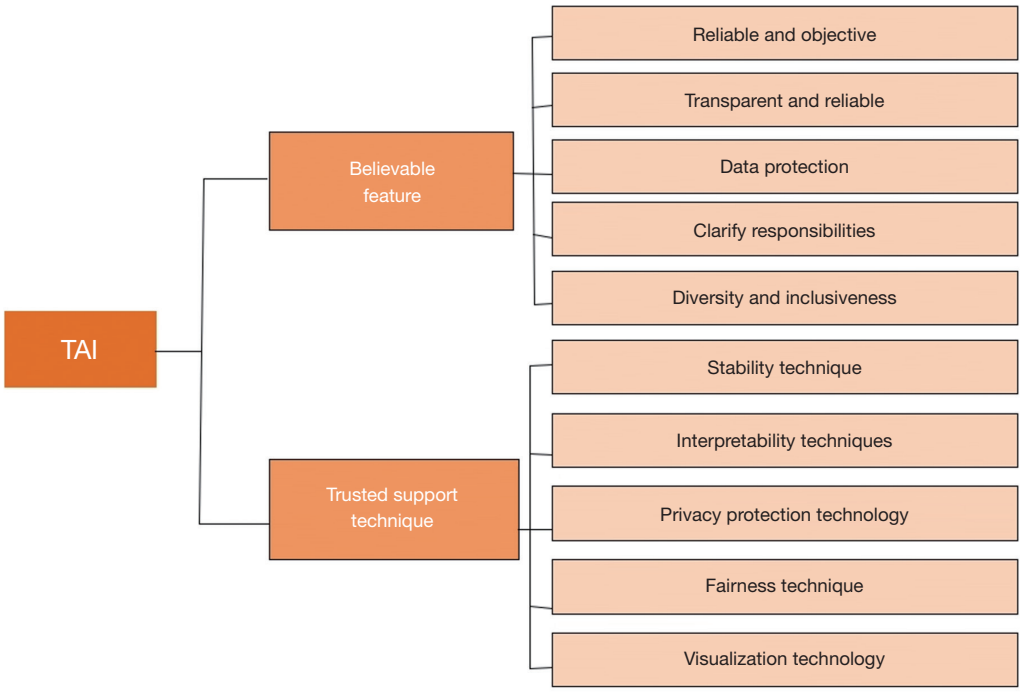

Figure 14 Characteristics of TAI. TAI, trustworthy artificial intelligence.

and development. Prioritizing trustworthiness is essential to ensure that AI systems are not only robust and effective, but also responsible, thereby building user confidence, particularly in sensitive areas like healthcare.

TAI algorithm methods

As AI progresses, the creation of appropriate datasets and pipelines for developing and evaluating AI models has emerged as a significant challenge. Recently, publicly accessible automated AI model builders have been developed to optimize performance across various applications. However, the design and engraving of the data used to develop AI often rely on custom manual work, which greatly affects the credibility of the models. This section discusses key considerations at each stage of the AI data pipeline—from data design to data sculpting (e.g., cleaning, valuation, and annotation) and data evaluation—to make AI more reliable (121). In its July 2021 white paper (4), the China academy of Information and Communications Technology introduced a framework for TAI. TAI is characterized by reliability, controllability, transparency, and inclusiveness, and supported by technologies that ensure stability, interpretability, privacy protection, fairness, and visualization (*Figure 14*) (122). Stability ensures the stable operation of the system in different environments; interpretability enables the decision-making process to be traced and understood; privacy protection ensures data security and compliance; fairness works to reduce algorithmic bias; visualization provides intuitive data presentation and decision support. The European Union's recent ethical guidelines for AI stipulate that TAI systems must adhere to the following four ethical principles: respect for human autonomy; prevention of harm; fairness; and explicability (123). Various researchers have proposed specific dimensions for TAI that align with these ethical principles (120). At present, many researchers have proposed evaluation criteria for the reliability of TAI (124,125). Further, numerous initiatives are addressing bias and fairness in AI (126,127).

Several researchers have incorporated TAI principles into medical image segmentation. Ricci Lara *et al.* (128) identified three sources of system bias in medical imaging AI and suggested mitigation strategies. Chen *et al.* (129) developed a semi-supervised semantic segmentation method, increasing the confidence in the predicted class probability graph. Zou *et al.* (130) introduced EvidenceCap, a reliable deep-learning segmentation model that enhances reliability,

robustness, and computational efficiency by assessing the uncertainty of image contour features.

In the field of medical AI, researchers have developed models and services to accelerate the creation of reliable medical-assisted diagnostic models for healthcare institutions. These initiatives enhance transparency, trust, and the adoption of AI in multicenter research (131,132). Together, these efforts represent significant progress in establishing trustworthy and ethically responsible AI systems, which are crucial for the successful integration and acceptance of AI, especially in sensitive areas like healthcare.

Advantages of TAI algorithms

Trust, a vital aspect of human interactions, is also important in our relationship with technology, including everyday tools and advanced systems like AI (133). First, TAI algorithms, which clarify their decision-making and reasoning processes, provide transparency for users in understanding how specific results and decisions are reached (134). This clarity increases the trust placed in these algorithms. Second, TAI is also adaptable to variations in input data, noise, and other disruptions. By consistently providing reliable results and demonstrating robustness in various environments, TAI ensures users can rely on its performance, further enhancing trust (22). Third, TAI proactively addresses fairness, privacy, and bias mitigation (135). Through careful balancing and the elimination of data bias, and by employing transparent, reviewable algorithm designs, TAI takes deliberate steps to ensure fairness, safeguard privacy, and prevent discrimination. A TAI model can be created using previous XAI methods to explain the decision process, and a framework that supports TAI principles, including model fairness and privacy. Finally, the effect is measured by specific methods for evaluating TAI. TAI encompasses security, robustness, non-discrimination, fairness, interpretability, accountability, auditability, privacy protection, and environmental sustainability, significantly increasing the trust in AI decisions among developers and users (136). When applied to medical imaging, TAI not only maintains the security and robustness of algorithms but also processes medical image data in a reliable and trustworthy manner. This increase in reliability is vital for clinical diagnoses, enabling intelligent systems to communicate their operational processes to healthcare professionals. Ultimately, this fosters a collaborative and trust-based relationship between humans and computers, laying the groundwork for the successful integration and application of AI in various fields (124).

Challenges of TAI algorithms

TAI has significant potential, and integrates both trusted characteristics and technologies; however, it still faces challenges. First, its research and development are still in the early stages, and associated regulations are not yet fully developed. Second, the trust architecture in AI is shaped by various factors, including industry practices, professional and regulatory standards, and considerations of accuracy, security, bias, risk, and auditability (137). Third, there are potential conflicts among the different dimensions of TAI. The study of the interactions between these dimensions has only just begun. For example, while a XAI can promote fairness by providing transparency in its decision-making processes, the techniques used to enhance a model's explainability might unintentionally create disparities in explainability across different groups, leading to fairness concerns (132). Understanding the complex interactions between these various dimensions presents a significant challenge. Therefore, the development of standardized ethical guidelines for TAI, the creation of robust frameworks supporting TAI in AI, and the effective integration of TAI in the medical sector are considerable challenges in the field of TAI development. Addressing these challenges is essential to fully realize the potential and benefits of TAI in diverse applications and industries (138).

General discussion and future research directions

General discussion

This article provides a review of the literature on traditional AI, XAI, and TAI algorithms, particularly focusing on their application in medical image segmentation. Traditional machine-learning methods, such as decision trees and linear regression, offer inherent explainability through clear visualizations and feature importance scores, which are crucial in medical imaging. However, these methods struggle with high-dimensional data and complex relationships. It would be beneficial to discuss the explainability of traditional machine-learning methods, as they provide a foundation for understanding more complex models. Conversely, XAI builds on these foundations, using techniques like saliency maps to make complex models

more interpretable, while TAI emphasizes trustworthiness alongside explainability. Our study highlights the strengths of XAI and TAI but acknowledges their limitations, particularly the insufficient analysis of TAI applications in medical imaging. As TAI develops, its integration with medical image segmentation will be crucial, but it will also face challenges, such as those related to the standardization of ethical guidelines and the building of a robust TAI framework.

Future research directions

The convergence of AI technology with medical image segmentation represents an ongoing journey of innovation and progress. AI approaches tailored for medical image segmentation have made significant progress over time, meeting the escalating needs and expectations for AI in the field. This evolution included the shift from traditional AI approaches to XAI and TAI. Ensuring the reliability of AI is critical, especially given its expanding applications in the field of medical image segmentation. Key challenges include addressing data bias in the training and decision-making processes of AI systems, where uneven or biased training data can distort results, lead to unfair or inaccurate results, negatively impact patient diagnosis, and pose a threat to patient safety. Mitigating these biases is critical to improving the credibility of AI. In addition, processing complex medical image data is an important challenge when applying TAI to medical image segmentation. Medical images can involve different anatomical structures and pathological features, and the quality and consistency of these data are critical to the accuracy of algorithms. Ensuring data quality and dealing with data biases, such as dealing with unbalanced data distribution or labeling errors, are key steps to ensuring the accuracy of algorithm training and segmentation results. Additionally, effectively interpreting and communicating segmentation results to clinicians is a challenging task. Despite the emphasis of TAI on the interpretability of algorithms, it remains a technical and communication challenge to present complex segmentation results in an intuitive and understandable way to help physicians make accurate diagnostic and treatment decisions. In the context of AI in medicine, information fusion technology is expected to improve the combined effect of diagnosis and treatment by integrating information from different medical data modes. However, the challenges of achieving multimodal interpretation and causal analysis require researchers to focus on deeper data understanding and model interpretation capabilities to ensure the trustworthiness and application value of healthcare AI systems. Innovative human-machine interfaces and supportive visualization technologies are also important factors driving the development of medical AI. These technologies can help physicians intuitively understand and interact with segmentation results, further enhancing the accuracy and efficiency of medical decisions.

Conclusions

Medical image segmentation involves complex algorithms designed to accurately identify and delineate anatomical structures or pathological areas in medical images. In this review, we introduced some traditional AI methods (i.e., CNN, ResNets, and GANs). In the introduction to XAI, we focused on LIME, LRP, DeepLIFT, and Grad-CAM. We started with the authors of these methods and summarized the development and improvement of these methods. Among these methods, LRP and Grad-CAM are more effective in the field of medical image segmentation. Finally, we introduced the concept and some applications of TAI. It is expected that medical image segmentation will increasingly adopt TAI technology, marking a transformative phase in the development of the trusted AI medical field. This shift not only represents a major shift in the field of medical image segmentation but also underscores the need for ongoing research and development in the emerging field of trusted AI.

Future research directions include exploring new AI architectures, integrating multimodal data fusion techniques, and advancing TAI methods to improve the reliability and trustworthiness of AI in medical image segmentation. Emphasis should be placed on developing standardized evaluation indicators and benchmarks to facilitate fair comparisons between different studies.

By delving into these technical details, discussing specific case studies, and highlighting insights and research directions, we aimed to provide a comprehensive understanding of where AI, XAI, and TAI stand in medical image segmentation. Our approach not only addressed current challenges but also revealed a promising future for AI in transforming medical diagnostics through reliable and precise image analysis techniques.

Acknowledgments

Our figures were drawn by Visio, which was purchased by

the school and provided to us. It is licensed software.
Funding: This study was supported by the China Postdoctoral Science Foundation (certificate No. 2023M742568).

Footnote

Reporting Checklist: The authors have completed the Narrative Review reporting checklist. Available at https://qims.amegroups.com/article/view/10.21037/qims-24-723/rc

Conflicts of Interest: All authors have completed the ICMJE uniform disclosure form (available at https://qims.amegroups.com/article/view/10.21037/qims-24-723/coif). The authors have no conflicts of interest to declare.

Ethical Statement: The authors are accountable for all aspects of the work in ensuring that questions related to the accuracy or integrity of any part of the work are appropriately investigated and resolved.

References

1. Lee LK, Liew SC, Thong WJ. A review of image segmentation methodologies in medical image. Adv Comput Commun Eng Technol 2015;1069-80.
2. Hesamian MH, Jia W, He X, Kennedy P. Deep Learning Techniques for Medical Image Segmentation: Achievements and Challenges. J Digit Imaging 2019;32:582-96.
3. Chen A, Zhu L, Zang H, Ding Z, Zhan S. Computer-aided diagnosis and decision-making system for medical data analysis: a case study on prostate MR images. J Manag Sci Eng 2019;4:266-78.
4. Abdel-Basset M, Chang V, Mohamed R. A novel equilibrium optimization algorithm for multi-thresholding image segmentation problems. Neural Comput Appl 2021;33:10685-718.
5. Yang D, Xu Z, Li W, Myronenko A, Roth HR, Harmon S, Xu S, Turkbey B, Turkbey E, Wang X, Zhu W, Carrafiello G, Patella F, Cariati M, Obinata H, Mori H, Tamura K, An P, Wood BJ, Xu D. Federated semi-supervised learning for COVID region segmentation in chest CT using multi-national data from China, Italy, Japan. Med Image Anal 2021;70:101992.
6. Savant S. A review on edge detection techniques for image segmentation. Int J Comput Appl 2014;5:101.
7. Peng T, Wu Y, Zhao J, Yang Y, Li Q, Tang H, et al. Coarse-to-fine tuning knowledgeable system for boundary delineation in medical images. Appl Intell 2023;53:30642-60.
8. Peng T, Gu Y, Ruan SJ, Wu QJ, Cai J. Novel Solution for Using Neural Networks for Kidney Boundary Extraction in 2D Ultrasound Data. Biomolecules 2023;13:1548.
9. Chen L, Bentley P, Mori K, Misawa K, Fujiwara M, Rueckert D. DRINet for Medical Image Segmentation. IEEE Trans Med Imaging 2018;37:2453-62.
10. Milletari F, Navab N, Ahmadi SA. V-Net: Fully convolutional neural networks for volumetric medical image segmentation. In: 2016 Fourth International Conference on 3D Vision (3DV). IEEE; 2016:565-71.
11. Kayalibay B, Jensen G, van der Smagt P. CNN-based segmentation of medical imaging data. Med Image Anal 2017;42:15-25.
12. Jafari M, Auer D, Karimi D. DRU-Net: An efficient deep convolutional neural network for medical image segmentation. In: 2020 IEEE 17th International Symposium on Biomedical Imaging (ISBI). IEEE; 2020:1144-8.
13. Hassanzadeh T, Essam D, Sarker R. 2D to 3D Evolutionary Deep Convolutional Neural Networks for Medical Image Segmentation. IEEE Trans Med Imaging 2021;40:712-21.
14. Meng Y, Zhang H, Wang Z, Chen M, Hu P, Du Y, et al. BI-GCN: Boundary-aware input-dependent graph convolution network for biomedical image segmentation. Med Image Anal 2021;72:102135.
15. Kazi A, Farghadani S, Aganj I, Navab N. IA-GCN: Interpretable attention based graph convolutional network for disease prediction. Mach Learn Med Imaging 2024:382-92.
16. Sun Y, Yuan P, Sun Y. MM-GAN: 3D MRI data augmentation for medical image segmentation via generative adversarial networks. In: 2020 IEEE

International Conference on Knowledge Graph (ICKG). IEEE; 2020:227-34.
17. Kazeminia S, Baur C, Kuijper A, van Ginneken B, Navab N, Albarqouni S, Mukhopadhyay A. GANs for medical image analysis. Artif Intell Med 2020;109:101938.
18. Ma Y, Liu J, Liu Y, Fu H, Hu Y, Cheng J, Qi H, Wu Y, Zhang J, Zhao Y. Structure and Illumination Constrained GAN for Medical Image Enhancement. IEEE Trans Med Imaging 2021;40:3955-67.
19. Li S, Sui X, Luo X, Xu X, Liu Y, Goh R. Medical image segmentation using squeeze-and-expansion transformers. arXiv preprint 2021; arXiv:2105.13677.
20. Barredo Arrieta A, Díaz-Rodríguez N, Ser JD, Bennetot A, Tabik S, Barbado A, et al. Explainable artificial intelligence (XAI): Concepts, taxonomies, opportunities, and challenges toward responsible AI. Inf Fusion 2020;58:82-115.
21. Das A, Rad P. Opportunities and challenges in explainable artificial intelligence (XAI): A survey. arXiv preprint 2020; arXiv:2006.11371.
22. Thiebes S, Lins S, Sunyaev A. Trustworthy artificial intelligence. Electron Mark 2021;31:447-64.
23. McCarthy J, Minsky ML, Rochester N, Shannon CE. A proposal for the Dartmouth summer research project on artificial intelligence, August 31, 1955. AI Mag 2006;27:12.
24. Buchanan BG. A (very) brief history of artificial intelligence. AI Mag 2005;26:53.
25. Wang X, Feng C, Huang M, Liu S, Ma H, Yu K. Cervical cancer segmentation based on medical images: a literature review. Quant Imaging Med Surg 2024;14:5176-204.
26. Liu X, Song L, Liu S, Zhang Y. A review of deep-learning-based medical image segmentation methods. Sustainability 2021;13:1224.
27. Diaz-Pinto A, Alle S, Nath V, Tang Y, Ihsani A, Asad M, Pérez-García F, Mehta P, Li W, Flores M, Roth HR, Vercauteren T, Xu D, Dogra P, Ourselin S, Feng A, Cardoso MJ. MONAI Label: A framework for AI-assisted interactive labeling of 3D medical images. Med Image Anal 2024;95:103207.
28. Huang Y, Yang X, Liu L, Zhou H, Chang A, Zhou X, Chen R, Yu J, Chen J, Chen C, Liu S, Chi H, Hu X, Yue K, Li L, Grau V, Fan DP, Dong F, Ni D. Segment anything model for medical images? Med Image Anal 2024;92:103061.
29. Gao Y, Xia W, Hu D, Wang W, Gao X. DeSAM: Decoupled Segment Anything Model for Generalizable Medical Image Segmentation. arXiv preprint 2024; arXiv:2306.00499.
30. Ruan J, Xiang S. VM-UNet: Vision Mamba UNet for Medical Image Segmentation. arXiv preprint 2024; arXiv:2306.00499.
31. Tragakis A, Kaul C, Murray-Smith R, Husmeier D. The fully convolutional transformer for medical image segmentation. In: 2023 IEEE/CVF Winter Conference on Applications of Computer Vision (WACV). IEEE; 2023:3649-58.
32. Miao J, Chen C, Liu F, Wei H, Heng PA. CauSSL: Causality-inspired semi-supervised learning for medical image segmentation. In: 2023 IEEE/CVF International Conference on Computer Vision (ICCV). IEEE; 2023:21369-80.
33. Gaillochet M, Desrosiers C, Lombaert H. TAAL: Test-time augmentation for active learning in medical image segmentation. arXiv preprint 2023; arXiv:2304.05534.
34. Butoi VI, Ortiz JJG, Ma T, Sabuncu MR, Guttag J, Dalca AV. UniverSeg: Universal medical image segmentation. arXiv preprint 2023; arXiv:2302.00700.
35. Wang Z, Zheng JQ, Zhang Y, Cui G, Li L. Mamba-UNet: UNet-Like Pure Visual Mamba for Medical Image Segmentation. arXiv preprint 2024; arXiv:2402.05079.
36. Zunair H, Hamza AB. Masked supervised learning for semantic segmentation. In: Proceedings of the British Machine Vision Conference (BMVC); 2022.
37. Verma R, Kumar N, Patil A, Kurian NC, Rane S, Graham S, et al. MoNuSAC2020: A Multi-Organ Nuclei Segmentation and Classification Challenge. IEEE Trans Med Imaging 2021;40:3413-23.
38. Zunair H, Ben Hamza A. Sharp U-Net: Depthwise convolutional network for biomedical image segmentation. Comput Biol Med 2021;136:104699.
39. Lloyd S. Least squares quantization in PCM. IEEE Trans Inf Theory 1982;28:129-37.
40. Quinlan JR. Induction of decision trees. Mach Learn 1986;1:81-106.
41. Cortes C, Vapnik V. Support-vector networks. Mach Learn 1995;20:273-97.
42. Lecun Y, Bottou L, Bengio Y, Haffner P. Gradient-based learning applied to document recognition. Proc IEEE 1998;86:2278-324.
43. He K, Zhang X, Ren S, Sun J. Deep residual learning for image recognition. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition; 2016:770-8.
44. Russakovsky O, Deng J, Su H, Krause J, Satheesh S, Ma

S, Huang Z, Karpathy A, Khosla A, Bernstein M, Berg AC, Fei-Fei L. ImageNet Large Scale Visual Recognition Challenge. International Journal of Computer Vision 2015;115:211-52.
45. Goodfellow IJ, Pouget-Abadie J, Mirza M, Xu B, Warde-Farley D, Ozair S, et al. Generative adversarial networks. Commun ACM 2020;63:139-44.
46. Vaswani A, Shazeer N, Parmar N, Uszkoreit J, Jones L, Gomez AN, et al. Attention is all you need. Adv Neural Inf Process Syst 2017;30:5998-6008.
47. Chen J, Lu Y, Yu Q, Luo X, Adeli E, Wang Y, et al. TransUNet: Transformers make strong encoders for medical image segmentation. arXiv preprint 2021; arXiv:2102.04306.
48. Lu Z, Zou Q, Wang M, Han X, Shi X, Wu S, Xie Z, Ye Q, Song L, He Y, Feng Q, Zhao Y. Artificial intelligence improves the diagnosis of human leukocyte antigen (HLA)-B27-negative axial spondyloarthritis based on multi-sequence magnetic resonance imaging and clinical features. Quant Imaging Med Surg 2024;14:5845-60.
49. Xiong Z, Qiu J, Liang Q, Jiang J, Zhao K, Chang H, Lv C, Zhang W, Li B, Ye J, Li S, Peng S, Sun C, Chen S, Long D, Shu X. Deep learning models for rapid discrimination of high-grade gliomas from solitary brain metastases using multi-plane T1-weighted contrast-enhanced (T1CE) images. Quant Imaging Med Surg 2024;14:5762-73.
50. Wu Y, Xiaoqin Z, Artificial intelligence in medical image processing: progress and prospect, Journal volume & issue 2021;43:1707-12.
51. Rai A. Explainable AI: from black box to glass box. J Acad Mark Sci 2020;48:137-41.
52. Nassar M, Salah K, Ur Rehman MH, Svetinovic D. Blockchain for explainable and trustworthy artificial intelligence. WIREs Data Min Knowl Discov 2020;10:e1340.
53. Wang K, Dong J, Wang Y, Yin H. Securing Data With Blockchain and AI. IEEE Access 2019;7:77981-9.
54. Zuo Z, Watson M, Budgen D, Hall R, Kennelly C, Al Moubayed N. Data Anonymization for Pervasive Health Care: Systematic Literature Mapping Study. JMIR Med Inform 2021;9:e29871.
55. Lekadir K, Osuala R, Gallin C, Lazrak N, Kushibar K, Tsakou G, et al. FUTURE-AI: Guiding principles and consensus recommendations for trustworthy artificial intelligence in medical imaging. arXiv preprint 2023. arXiv:2109.09658.
56. Dilmaghani S, Brust MR, Danoy G, Cassagnes N, Pecero J, Bouvry P. Privacy and security of big data in AI systems: A research and standards perspective. In: Proceedings of the 2019 IEEE International Conference on Big Data (Big Data). IEEE; 2019:5737-43.
57. Karimi D, Dou H, Warfield SK, Gholipour A. Deep learning with noisy labels: Exploring techniques and remedies in medical image analysis. Med Image Anal 2020;65:101759.
58. Goodman B, Flaxman S. European Union regulations on algorithmic decision-making and a "right to explanation". AI Mag 2017;38:50-7.
59. Zhu J, Liapis A, Risi S, Bidarra R, Youngblood GM. Explainable AI for designers: A human-centered perspective on mixed-initiative co-creation. In: Proceedings of the 2018 IEEE Conference on Computational Intelligence and Games (CIG). IEEE; 2018:1-8.
60. Preece A, Harborne D, Braines D, Tomsett R, Chakraborty S. Stakeholders in explainable AI. arXiv preprint 2018; arXiv:1810.00184.
61. van der Velden BHM, Kuijf HJ, Gilhuijs KGA, Viergever MA. Explainable artificial intelligence (XAI) in deep learning-based medical image analysis. Med Image Anal 2022;79:102470.
62. Strubell E, Ganesh A, McCallum A. Energy and policy considerations for deep learning in NLP. arXiv preprint 2019; arXiv:1906.02243.
63. Lopez MM, Kalita J. Deep learning applied to NLP. arXiv preprint 2017; arXiv:1703.03091.
64. Li J, Chen X, Hovy E, Jurafsky D. Visualizing and understanding neural models in NLP. arXiv preprint 2016; arXiv:1506.01066.
65. Ribeiro MT, Singh S, Guestrin C. "Why Should I Trust You?": Explaining the predictions of any classifier. In: Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. ACM; 2016:1135-44.
66. Visani G, Bagli E, Chesani F. OptiLIME: Optimized LIME explanations for diagnostic computer algorithms. arXiv preprint 2022; arXiv:2006.05714.
67. Zhang Y, Song K, Sun Y, Tan S, Udell M. "Why should you trust my explanation?" Understanding uncertainty in LIME explanations. arXiv preprint 2019; arXiv:1904.12991.
68. Zhao X, Huang W, Huang X, Robu V, Flynn D. BayLIME: Bayesian local interpretable model-agnostic explanations. In: Proceedings of the Thirty-Seventh

Conference on Uncertainty in Artificial Intelligence. PMLR; 2021:887-96.
69. Lee E, Braines D, Stiffler M, Hudler A, Harborne D. Developing the sensitivity of LIME for better machine learning explanation. In: Pham T, editor. Artificial Intelligence and Machine Learning for Multi-Domain Operations and Applications. SPIE; 2019:55.
70. Nematzadeh H, García-Nieto J, Navas-Delgado I, Aldana-Montes JF. Ensemble-based genetic algorithm explainer with automized image segmentation: A case study on melanoma detection dataset. Comput Biol Med 2023;155:106613.
71. Duamwan LM, Bird JJ. Explainable AI for medical image processing: A study on MRI in Alzheimer's disease. In: Proceedings of the 16th International Conference on Pervasive Technologies Related to Assistive Environments. ACM; 2023:480-4.
72. Knab P, Marton S, Bartelt C. DSEG-LIME: Improving image explanation by hierarchical data-driven segmentation. arXiv preprint 2024; arXiv:2403.07733.
73. Gaur L, Bhandari M, Razdan T, Mallik S, Zhao Z. Explanation-Driven Deep Learning Model for Prediction of Brain Tumour Status Using MRI Image Data. Front Genet 2022;13:822666.
74. Ahsan MM, Nazim R, Siddique Z, Huebner P. Detection of COVID-19 Patients from CT Scan and Chest X-ray Data Using Modified MobileNetV2 and LIME. Healthcare (Basel) 2021;9:1099.
75. Itti L, Koch C, Niebur E. A model of saliency-based visual attention for rapid scene analysis. IEEE Trans Pattern Anal Mach Intell 1998;20:1254-9.
76. Simonyan K, Vedaldi A, Zisserman A. Deep inside convolutional networks: Visualising image classification models and saliency maps. arXiv preprint 2014;arXiv:1312.6034.
77. Kim B, Seo J, Jeon S, Koo J, Choe J, Jeon T. Why are saliency maps noisy? Cause of and solution to noisy saliency maps. In: 2019 IEEE/CVF International Conference on Computer Vision Workshop (ICCVW). IEEE; 2019:4149-57.
78. Bernal J, Sánchez FJ, Fernández-Esparrach G, Gil D, Rodríguez C, Vilariño F. WM-DOVA maps for accurate polyp highlighting in colonoscopy: Validation vs. saliency maps from physicians. Comput Med Imaging Graph 2015;43:99-111.
79. Li Z. A saliency map in primary visual cortex. Trends Cogn Sci 2002;6:9-16.
80. Morch NJS, Kjems U, Hansen LK, Svarer C, Law I, Lautrup B, et al. Visualization of neural networks using saliency maps. In: Proceedings of ICNN95 - International Conference on Neural Networks. IEEE; 1995:2085-90.
81. Gadgil SU, Endo M, Wen E, Ng AY, Rajpurkar P. Proceedings of the Fourth Conference on Medical Imaging with Deep Learning. PMLR; 2021:190-204.
82. Sun J, Darbehani F, Zaidi M, Wang B. SAUNet: Shape attentive U-Net for interpretable medical image segmentation. In: Martel AL, Abolmaesumi P, Stoyanov D, Mateus D, Zuluaga MA, Zhou SK, Racoceanu D, Joskowicz L, editors. Medical Image Computing and Computer-Assisted Intervention - MICCAI 2020. Springer International Publishing; 2020:797-806.
83. Ning Z, Zhong S, Feng Q, Chen W, Zhang Y. SMU-Net: Saliency-Guided Morphology-Aware U-Net for Breast Lesion Segmentation in Ultrasound Image. IEEE Trans Med Imaging 2022;41:476-90.
84. Bach S, Binder A, Montavon G, Klauschen F, Müller KR, Samek W. On Pixel-Wise Explanations for Non-Linear Classifier Decisions by Layer-Wise Relevance Propagation. PLoS One 2015;10:e0130140.
85. Kohlbrenner M, Bauer A, Nakajima S, Binder A, Samek W, Lapuschkin S. Towards best practice in explaining neural network decisions with LRP. In: 2020 International Joint Conference on Neural Networks (IJCNN). IEEE; 2020:1-7.
86. Ahmed AMA, Ali LAM. Explainable medical image segmentation via generative adversarial networks and layer-wise relevance propagation. arXiv preprint 2021; arXiv:2111.01665.
87. Alam MU, Baldvinsson JR, Wang Y. Exploring LRP and Grad-CAM visualization to interpret multi-label-multi-class pathology prediction using chest radiography. In: 2022 IEEE 35th International Symposium on Computer-Based Medical Systems (CBMS). IEEE; 2022:258-63.
88. Ni C, Wan J, Leng T. Interpret 3D-CNN through a new decision decomposition strategy based on LRP. In: 2021 6th International Conference on Computational Intelligence and Applications (ICCIA). IEEE; 2021:218-23.
89. Xin S, Shi H, Jide A, Zhu M, Ma C, Liao H. Automatic lesion segmentation and classification of hepatic echinococcosis using a multiscale-feature convolutional neural network. Med Biol Eng Comput 2020;58:659-68.
90. Sundararajan M, Taly A, Yan Q. Axiomatic attribution for deep networks. Proceedings of the 34th International Conference on Machine Learning. PMLR 2017;70:3319-28.

91. Kapishnikov A, Venugopalan S, Avci B, Wedin B, Terry M, Bolukbasi T. Guided integrated gradients: An adaptive path method for removing noise. Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2021:5050-8.
92. Wachter S, Mittelstadt B, Russell C. Counterfactual explanations without opening the black box: Automated decisions and the GDPR. Harv J Law Technol 2017;31:841.
93. Thiagarajan JJ, Thopalli K, Rajan D, Turaga P. Training calibration-based counterfactual explainers for deep learning models in medical image analysis. Sci Rep 2022;12:597.
94. Singh D, Somani A, Horsch A, Prasad DK. Counterfactual explainable gastrointestinal and colonoscopy image segmentation. In: 2022 IEEE 19th International Symposium on Biomedical Imaging (ISBI). IEEE; 2022:1-5.
95. Zhou S, Islam UJ, Pfeiffer N, Banerjee I, Patel BK, Iquebal AS. SCGAN: Sparse CounterGAN for counterfactual explanations in breast cancer prediction. IEEE Transactions on Automation Science and Engineering; 2023.
96. Lenis D, Major D, Wimmer M, Berg A, Sluiter G, Bühler K. Domain aware medical image classifier interpretation by counterfactual impact analysis. In: Martel AL, Abolmaesumi P, Stoyanov D, Mateus D, Zuluaga MA, Zhou SK, Racoceanu D, Joskowicz L, editors. Medical Image Computing and Computer Assisted Intervention – MICCAI 2020. Lecture Notes in Computer Science, vol 12261. Springer; 2020:315-25.
97. Selvaraju RR, Cogswell M, Das A, Vedantam R, Parikh D, Batra D. Grad-CAM: Visual explanations from deep networks via gradient-based localization. Int J Comput Vis 2020;128:336-59.
98. Xiao M, Zhang L, Shi W, Liu J, He W, Jiang Z. A visualization method based on Grad-CAM for medical image segmentation model. In: 2021 International Conference on Electronics, Information Engineering and Computer Science (EIECS). IEEE; 2021:242-7.
99. Chen Z, Tian Z, Zhu J, Li C, Du S. C-CAM: Causal CAM for weakly supervised semantic segmentation on medical images. In: 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). IEEE; 2022:11666-75.
100. Vinogradova K, Dibrov A, Myers G. Towards interpretable semantic segmentation via gradient-weighted class activation mapping. In: Proceedings of the AAAI Conference on Artificial Intelligence; 2020;34:13943-4.
101. Hasany SN, Petitjean C, Mériaudeau F. Seg-XRes-CAM: Explaining spatially local regions in image segmentation. In: 2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshop (CVPRW). IEEE; 2023:3733-8.
102. Schlegel U, Arnout H, El-Assady M, Oelke D, Keim DA. Towards a rigorous evaluation of XAI methods on time series. In: 2019 IEEE/CVF International Conference on Computer Vision Workshop (ICCVW). IEEE; 2019:4197-201.
103. Peng T, Gu Y, Ye Z, Cheng X, Wang J. A-LugSeg: Automatic and explainability-guided multi-site lung detection in chest X-ray images. Expert Syst Appl 2022;198:116873.
104. Peng T, Xu D, Wu Y, Zhao J, Yang C, Zhang L, Cai J. A mathematical and neural network-based hybrid technique for detecting the prostate contour from medical image data. Biomed Signal Process Control 2023;86:105337.
105. Peng T, Wu Y, Qin J, Wu QJ, Cai J. H-ProSeg: Hybrid ultrasound prostate segmentation based on explainability-guided mathematical model. Comput Methods Programs Biomed 2022;219:106752.
106. Peng T, Wu Y, Zhao J, Zhang B, Wang J, Cai J. Explainability-guided mathematical model-based segmentation of transrectal ultrasound images for prostate brachytherapy. In: 2022 IEEE International Conference on Bioinformatics and Biomedicine (BIBM). IEEE; 2022:1126-31.
107. Peng T, Wu Y, Zhao J, Wang C, Wang W, Shen Y, Cai J. Coarse-to-fine tuning knowledgeable system for boundary delineation in medical images. Appl Intell 2023;53:30642-60.
108. Young K, Booth G, Simpson B, Dutton R, Shrapnel S. Deep neural network or dermatologist? In: Suzuki K, Reyes M, Syeda-Mahmood T, Konukoglu E, Glocker B, Wiest R, Gur Y, Greenspan H, Madabhushi A, editors. Interpretability of Machine Intelligence in Medical Image Computing and Multimodal Learning for Clinical Decision Support. Springer; 2019:48-55.
109. Wang L, Lin ZQ, Wong A. COVID-Net: a tailored deep convolutional neural network design for detection of COVID-19 cases from chest X-ray images. Sci Rep 2020;10:19549.
110. Zhang Y, Tiňo P, Leonardis A, Tang K. A survey on neural network interpretability. IEEE Trans Emerg Top Comput Intell 2021;5:726-42.
111. Kaur A, Dong G, Basu A. GradXcepUNet: Explainable AI

based medical image segmentation. In: Berretti S, Su GM, editors. Smart Multimedia. Springer; 2022:174-88.
112. Abeyagunasekera SHP, Perera Y, Chamara K, Kaushalya U, Sumathipala P, Senaweera O. LISA: Enhance the explainability of medical images unifying current XAI techniques. In: 2022 IEEE 7th International Conference on Convergence Technology (I2CT). IEEE; 2022:1-9.
113. Ghnemat R, Alodibat S, Abu Al-Haija Q. Explainable Artificial Intelligence (XAI) for Deep Learning Based Medical Imaging Classification. J Imaging 2023;9:177.
114. Medina J, Ojeda-Aciego M, Verdegay JL, Pelta DA, Cabrera IP, Bouchon-Meunier B, Yager RR. Information Processing and Management of Uncertainty in Knowledge-Based Systems. Springer International Publishing; 2020:99-113.
115. Zhang R, Hong M, Cai H, Liang Y, Chen X, Liu Z, Wu M, Zhou C, Bao C, Wang H, Yang S, Hu Q. Predicting the pathological invasiveness in patients with a solitary pulmonary nodule via Shapley additive explanations interpretation of a tree-based machine learning radiomics model: a multicenter study. Quant Imaging Med Surg 2023;13:7828-41.
116. Stoyanov D, Taylor Z, Kia SM, Oguz I, Reyes M, Martel A, Maier-Hein L, Marquand AF, Duchesnay E, Löfstedt T, Landman B, Cardoso MJ, Silva CA, Pereira S, Meier R. Understanding and interpreting machine learning in medical image computing and applications. In: Machine Learning in Medical Imaging. Springer International Publishing; 2018:106-14.
117. Cassel CK, Jameton AL. Dementia in the elderly: an analysis of medical responsibility. Ann Intern Med 1981;94:802-7.
118. Xie Y, Gao G, Chen X. Outlining the design space of explainable intelligent systems for medical diagnosis. arXiv preprint 2019; arXiv:1902.06019.
119. Tjoa E, Guan C. A Survey on Explainable Artificial Intelligence (XAI): Toward Medical XAI. IEEE Trans Neural Netw Learn Syst 2021;32:4793-813.
120. Brundage M, Avin S, Wang J, Belfield H, Krueger G, Hadfield G, et al. Toward trustworthy AI development: Mechanisms for supporting verifiable claims. arXiv preprint 2020; arXiv:2004.07213.
121. Liang W, Tadesse GA, Ho D, Fei-Fei L, Zaharia M, Zhang C, Zou J. Advances, challenges and opportunities in creating data for trustworthy AI. Nat Mach Intell 2022;4:669-77.
122. China Academy of Information and Communications Technology. White paper on trusted artificial intelligence. Available online: http://www.caict.ac.cn/english/research/whitepapers/202110/P020211014399666967457.pdf
123. Smuha NA. The EU approach to ethics guidelines for trustworthy artificial intelligence. Comput Law Rev Int 2019;20:97-106.
124. Zicari RV, Brodersen J, Brusseau J, Düdder B, Eichhorn T, Ivanov T. Z-Inspection®: A process to assess trustworthy AI. IEEE Trans Technol Soc 2021;2:83-97.
125. Lv Z, Han Y, Singh AK, Manogaran G, Lv H. Trustworthiness in industrial IoT systems based on artificial intelligence. IEEE Trans Ind Inform 2021;17:1496-504.
126. Mehrabi N, Morstatter F, Saxena N, Lerman K, Galstyan A. A survey on bias and fairness in machine learning. arXiv preprint 2022; arXiv:1906.03750.
127. Salahuddin Z, Woodruff HC, Chatterjee A, Lambin P. Transparency of deep neural networks for medical image analysis: A review of interpretability methods. Comput Biol Med 2022;140:105111.
128. Ricci Lara MA, Echeveste R, Ferrante E. Addressing fairness in artificial intelligence for medical imaging. Nat Commun 2022;13:4581.
129. Chen H, Jin Y, Jin G, Zhu C, Chen E. Semisupervised Semantic Segmentation by Improving Prediction Confidence. IEEE Trans Neural Netw Learn Syst 2022;33:4991-5003.
130. Zou K, Yuan X, Shen X, Chen Y, Wang M, Goh RSM, Liu Y, Fu H. EvidenceCap: Towards trustworthy medical image segmentation via evidential identity cap. 2023. Available online: https://doi.org/10.21203/rs.3.rs-2558155/v1
131. Ma Y, Wang S, Derr T, Wu L, Tang J. Attacking graph convolutional networks via rewiring. arXiv preprint 2019; arXiv:1906.03750
132. Guo K, Ren S, Bhuiyan MZA, Li T, Liu D, Liang Z, Chen X. MDMaaS: Medical-Assisted Diagnosis Model as a Service with artificial intelligence and trust. IEEE Trans Ind Inform 2020;16:2102-14.
133. Danks D. The value of trustworthy AI. In: Proceedings of the 2019 AAAI/ACM Conference on AI Ethics and Society. ACM; 2019:521-2.
134. Chiao V. Fairness, accountability, and transparency: Notes on algorithmic decision-making in criminal justice. Int J Law Context 2019;15:126-39.
135. Varona D, Suárez JL. Discrimination, bias, fairness, and trustworthy AI. Appl Sci 2022;12:5826.
136. Liu H, Wang Y, Fan W, Liu X, Li Y, Jain S, Liu Y, Jain AK, Tang J. Trustworthy AI: A computational perspective. arXiv preprint 2021; arXiv:2107.06641.

137. Procter R, Tolmie P, Rouncefield M. Holding AI to account: Challenges for the delivery of trustworthy AI in healthcare. ACM Trans Comput-Hum Interact 2023;30:1-34.
138. Holzinger A, Dehmer M, Emmert-Streib F, Cucchiara R, Augenstein I, Del Ser J, Samek W, Jurisica I, Díaz-Rodríguez N. Information fusion as an integrative cross-cutting enabler to achieve robust, explainable, and trustworthy medical artificial intelligence. Inf Fusion 2022;79:263-78.

Cite this article as: Teng Z, Li L, Xin Z, Xiang D, Huang J, Zhou H, Shi F, Zhu W, Cai J, Peng T, Chen X. A literature review of artificial intelligence (AI) for medical image segmentation: from AI and explainable AI to trustworthy AI. Quant Imaging Med Surg 2024;14(12):9620-9652. doi: 10.21037/qims-24-723