Chinese Medical Journal®

Artificial intelligence in medical imaging: From task-specific models to large-scale foundation models

Yueyan Bian[1,2,3], Jin Li[1,2,3], Chuyang Ye[4], Xiuqin Jia[1,2,3], Qi Yang[1,2,3]

[1]Department of Radiology, Beijing Chaoyang Hospital, Capital Medical University, Beijing 100020, China;
[2]Key Lab of Medical Engineering for Cardiovascular Disease, Ministry of Education, Beijing 100020, China;
[3]Laboratory for Clinical Medicine, Capital Medical University, Beijing 100020, China;
[4]School of Integrated Circuits and Electronics, Beijing Institute of Technology, Beijing 100081, China.

Abstract

Artificial intelligence (AI), particularly deep learning, has demonstrated remarkable performance in medical imaging across a variety of modalities, including X-ray, computed tomography (CT), magnetic resonance imaging (MRI), ultrasound, positron emission tomography (PET), and pathological imaging. However, most existing state-of-the-art AI techniques are task-specific and focus on a limited range of imaging modalities. Compared to these task-specific models, emerging foundation models represent a significant milestone in AI development. These models can learn generalized representations of medical images and apply them to downstream tasks through zero-shot or few-shot fine-tuning. Foundation models have the potential to address the comprehensive and multifactorial challenges encountered in clinical practice. This article reviews the clinical applications of both task-specific and foundation models, highlighting their differences, complementarities, and clinical relevance. We also examine their future research directions and potential challenges. Unlike the replacement relationship seen between deep learning and traditional machine learning, task-specific and foundation models are complementary, despite inherent differences. While foundation models primarily focus on segmentation and classification, task-specific models are integrated into nearly all medical image analyses. However, with further advancements, foundation models could be applied to other clinical scenarios. In conclusion, all indications suggest that task-specific and foundation models, especially the latter, have the potential to drive breakthroughs in medical imaging, from image processing to clinical workflows.

Keywords: Medical imaging; Artificial intelligence; Task-specific model; Foundation model; Generalized representations

Introduction

Artificial intelligence (AI), particularly with emerging deep learning techniques, has shown remarkable performance in the acquisition and interpretation of medical imaging. Its applications span all major imaging modalities, including X-ray, computed tomography (CT), magnetic resonance imaging (MRI), ultrasound, positron emission tomography (PET), and pathological imaging.[1–7] Most existing state-of-the-art AI algorithms or medical imaging systems primarily focus on specific tasks using single or limited imaging modalities.[8] In clinical practice, however, radiologists must identify all potential abnormal features across medical images, which is a comprehensive and multifactorial task.[9,10] In addition, for relatively rare diseases, task-specific AI algorithm training often encounters heavily imbalanced datasets, leading to significant performance degradation.[11,12] Therefore, developing medical foundation techniques that can manage various tasks and image modalities is crucial for advancing broader AI applications in medical imaging.

Recently, large-scale foundation models, such as ChatGPT and the Segment Anything Model (SAM), have achieved significant success in the fields of natural language processing (NLP) and computer vision.[13–15] Pretrained on extensive datasets of natural images and/or language, these models exhibit strong zero-shot or few-shot generalization capabilities.[13–16] Compared with natural images, however, medical images have their unique characteristics. First, medical images are acquired from diverse imaging devices with different physical properties and energy sources, including light, X-rays, ultrasound, nuclear imaging, and magnetic resonance.[11,17–19] Second, medical

Access this article online

Quick Response Code:

Website: www.cmj.org

DOI: 10.1097/CM9.0000000000003489

Correspondence to: Qi Yang, Department of Radiology, Beijing Chaoyang Hospital, Capital Medical University, No. 8 Gongti South Road, Chaoyang District, Beijing 100020, China
E-Mail: yangyangqiqi@gmail.com

Chinese Medical Journal 2025;138(6)

Received: 06-09-2024; **Online:** 26-02-2025 **Edited by:** Sihan Zhou and Xiuyuan Hao

images present information on a range of scales, from cells to organ systems to the whole body.[20,21] Therefore, medical foundation model development cannot be achieved by simply fine-tuning the foundation models designed for natural images. Instead, domain-specific foundation models adapted to the medical field must be designed to achieve optimal performance.

Similar to previous development trends, AI advancements for natural images would rapidly transition into the medical imaging field. The SAM in medical images (MedSAM) has been proposed to enhance performance in universal medical imaging segmentation.[22] Furthermore, modality-specific foundation models for X-ray, CT, MRI, and ultrasound have been developed.[23–26] These models can learn generalized representations of medical images and are applied to downstream medical tasks using zero-shot or few-shot fine-tuning.[8] As a result, foundation models trained on medical images may offer more robust solutions to critical clinical challenges, driving advancements in medical imaging and improving disease diagnosis and treatment efficacy and efficiency.

In this review, we aimed to provide a comprehensive overview of existing task-specific models and recently proposed foundation models in medical imaging, comparing their differences and mutual relationship and describing their future directions. We conducted a literature search using Google Scholar and PubMed for related works and published articles after the year 2000, employing the following keywords: task-specific model, transfer learning, segmentation, classification, registration, enhancement, image generation, foundation model, medical imaging, vision foundation model, vision-language foundation model, artificial intelligence, and deep learning. To create a thorough overview, we selected and organized the relevant and recent studies based on the following criteria: (1) covering typical image modalities in the medical imaging field; (2) primarily focusing on typical imaging analysis tasks; and (3) excluding the analog work in the same field. The articles only published in English were included.

Task-Specific Models

Task-specific medical image analysis is one of the most important fundamental thinking pathways in medical image processing due to the limited capabilities of computational learning to acquire generalized knowledge. Complex and comprehensive medical tasks are often divided into multiple, distinct specific tasks to achieve better performance. The most common tasks include segmentation, classification, enhancement, and registration.[27–30] Traditional machine learning approaches address these tasks by relying on human-based feature extraction.[31] With the rapid advancements in recent years, deep learning techniques demonstrate strong capabilities for automatic feature extraction, without the need for predefined rules or manual intervention, thereby enabling the handling of these tasks more efficiently and accurately. Deep learning techniques have shown promising performance in automating these tasks. Table 1 illustrates examples of task-specific models in medical imaging.

Segmentation

Medical image segmentation tasks involve delineating or labeling specific regions or structures within medical images. This process includes identifying and extracting regions of interest, such as organs, tumors, or lesions, from complex medical imaging data. Image segmentation plays a crucial role in medical image analysis by accurately delineating structures, enabling precise localization and quantitative assessment, and supporting diagnosis and personalized treatment. In traditional machine learning, image segmentation methods mainly rely on gray intensity, edge, morphological, and textural features of each distinct region, including statistical model-based,[32] boundary detection-based,[33] and similarity criteria-based methods.[34,35] However, these approaches often struggle with images containing irregular shapes and complex structures.

With the development of deep learning, fully convolutional neural networks and various modifications of these architectures are among the most effective deep learning models for medical image segmentation.[36] They offer the advantage of analyzing large portions of the image simultaneously, thereby reducing repeated convolution operations.[36] For example, in brain tumor segmentation, Havaei *et al*[37] proposed a 2D two-pathway cascading deep neural network—the local and global pathways—specifically for glioblastoma segmentation in multi-modality MRIs, including T1, T2, T1C, and fluid attenuation inversion recovery (FLAIR). This deep cascading network effectively exploits both local and global contextual features to overcome the lack of 3D information, achieving a state-of-the-art Dice score of 0.85. Another study developed a multi-label 3D U-Net to segment subcortical volume, including the choroid plexus, lateral posterior ventricle horn, cerebellum, and cavum septum pellucidum et vergae, from 3D fetal ultrasound. This network also achieved a Dice score exceeding 0.85, demonstrating its effectiveness in accurately segmenting these complex structures.[38] Intracranial hemorrhage segmentation is another typical clinical application of AI. Islam *et al*[39] introduced a novel dilated convolution neural network integrated with hypercolumn features, trained on 89 CT scans, for automated intracerebral hemorrhage segmentation in CT images. The model achieved a state-of-the-art Dice score of 89.04%. In addition to these examples, deep learning techniques have been widely applied to various segmentation tasks, including brain anatomy,[40–42] ischemic lesions,[43–45] cancer,[46–48] liver tumors,[49–51] retinal vessel[52], and among others.

Classification

Classification tasks in medical imaging primarily involve determining the presence or absence of abnormalities, distinguishing between benign and malignant conditions, and predicting prognostic information. These tasks enable healthcare professionals to make informed decisions about patient management by providing a clear assessment of imaging findings. Moreover, accurate classification can streamline the diagnostic workflow, allowing for faster identification of critical cases that require immediate

Table 1: Examples of task-specific models in medical imaging.

Task type	Image modality	Training datasets	Performances	References
Segmentation				
Brain glioblastomas tumor	MR	BRATS2013	Dice score of 0.85; Specificity of 0.93; Sensitivity of 0.80.	[37]
Subcortical tissue	Ultrasound	537 fetal images from gestation between 18 and 26 weeks	Dice scores of 0.85, 0.85, 0.78, and 0.90 for choroid plexus, lateral posterior ventricle horn, cavum septum pellucidum et vergae, and cerebellum, respectively.	[38]
ICH	CT	89 CT images from ICH patients	Dice score of 89.04%; Specificity of 99.83%; Sensitivity of 91.51%.	[39]
Brain anatomy	MR, CT	70 T1w and T2w MRI; 840 T1w MRI; 62 CT scans	Dice score of above 0.91,0.97, 0.98, and 0.96 for the entire brain, GM, WM, and CSF respectively; Dice scores of above 0.97 and 0.95 for WM and GM, respectively; Overall Dice score of 0.84 for eleven intracranial structures.	[40–42]
Ischemic lesions	MR, CT	2348 DWI; 103 CT perfusion; ISLES 2015 and ISLES 2017	Best Dice score of 0.88 for acute and subacute ischemic lesions; Dice score of 0.68; Dice score of 0.85.	[43–45]
Cancer	MR, X-ray	140 T2w and DWI; 139 T1w; MIAS dataset	Best Dice score of 0.70 for rectal cancer; Dice score of 0.80 for endometrial cancer; Dice score and Jaccard score of 0.82 and 0.89 for breast cancer, respectively.	[46–48]
Liver tumor	CT, histology images	20 CT scans; IRCADB01; 80 hematoxylin and eosin stained histopathology images	Accuracy of 98.8%; Dice score of 99.2%; F1 score of 83.59%	[49–51]
Retinal blood vessel	Retinal images	120 retinal images	Accuracy of 0.978	[52]
Classification				
Alzheimer's disease	MR	489 MR scans	Best accuracy rates are 95.87% and 97.15% on 1.5T and 3T datasets, respectively.	[58]
Breast cancer	Histology images	7909 microscopic biopsy images	Accuracy of 95.4%	[59]
Lung adenocarcinoma	CT	480 CT scans	Best accuracy of 90% and AUC of 0.935	[60]
Brain hemorrhage	CT	8855 CT images	Accuracy of 93.48%	[71]
Brain tumor	MR	233 CE-MRI	Accuracy of 94.82%	[73]
Normal cardiovascular tissues and organs	Histology images	6000 blocks of images	F-score between 0.717 and 0.928 for elastic, heart, muscular, connective, vein, and light	[73]
Other tasks				
Image registration	CT and MR	45 short-axis cine MRIs and 2060 chest CT	Dice scores of 0.89 and 0.88 for cine MRI and chest CT, respectively.	[74]
Image registration	CT and MR	45 short-axis cine MRIs and 2060 lung CT	Normalized cross-correlation of 0.947 and 0.956 for chest CT and cine MRI, respectively.	[75]
Enhancement of signal-to-noise ratio	CT	Close to 6000 2D full-dose and corresponding low-dose (1/4 of the full-dose) CT images	PSNR of 47.90 and SSIM index of 0.9753	[76]
Enhancement of signal-to-noise ratio	MR	49 high-field-strength MRI and corresponding 64mT MRI with T1-weighted, T2-weighted and FLAIR	ΔNCC of 0.041, 0.044, and 0.027 for T1w, T2w, and FLAIR.	[77]
Image generation: generating synthetic FLAIR	MR	1416 DWI and FLAIR	Synthetic FLAIR with diagnostic performances similar to real FLAIR	[78]
Image generation: generating synthetic CTA	CT	1749 non-contrast CT and corresponding CTA	Synthesizes CTA with diagnostic performances similar to real CTA images	[79]

AUCs: Area of under curves; CE-MRI: Contrast-enhanced MRI; CSF: Cerebrospinal fluid; CT: Computed tomography; CTA: CT angiography; DCE-MR: Dynamic contrast-enhanced MR; DWI: Diffusion-weighted images; FFDM: Full-field digital mammography; FLAIR: Fluid attenuation inversion recovery; GM: Gray matter; ICH: Intracerebral hemorrhage; MIAS: Mammographic Image Analysis Society; MR: Magnetic resonance; MRI: Magnetic resonance imaging; NCC: Normalized cross-correlation; PSNR: Peak signal-to-noise ratio; SSIM: Structural similarity; WM: White matter.

attention. It also facilitates the development of risk stratification models, helping to prioritize patient treatment based on the likelihood of disease progression. Similar to segmentation, traditional classification machine learning techniques also rely on feature engineering, where the differences between categories are manually extracted, and clustering methods are then used to distinguish between these categories. Common classification methods include support vector machine (SVM),[53,54] decision trees,[55] random forest,[56] and k-nearest neighbors.[35,57]

Deep learning-based classifiers typically use convolutional neural networks to extract features, followed by fully connected layers for classification. Similar to segmentation tasks, many deep learning classifiers are trained from scratch or use pretrained features to optimize performance. For example, in classifying Alzheimer's disease (AD), a convolutional neural network based on GoogleNet architecture was trained to identify AD patients using diffusion-tensor images with a high accuracy of over 96% when distinguishing them from normal controls.[58] To distinguish benign from malignant breast cancer, Nawaz *et al*[59] proposed a DenseNet architecture for multi-class breast cancer classification using histopathological images. This model was trained on 7909 microscopic biopsy images and achieved high performance, with an accuracy of 95.4%. For prognostication in non-small cell adenocarcinoma lung cancer, Paul *et al*[60] developed a VGGNet-based network to extract deep features from CT images to predict short- and long-term survival outcomes, achieving state-of-the-art performance with an accuracy of 90% and an AUC of 0.935.

Another effective deep learning approach for classification tasks is transfer learning, a novel strategy for training models with limited datasets, based on the notion that knowledge can be transferred at the parametric level of deep learning models.[61] The deep learning models were trained on pretraining datasets, allowing the pretrained model parameters to be utilized for new tasks. Networks such as LeNet,[62] AlexNet,[63] VGGNet,[64] ResNet,[65] GoogleNet,[66] DenseNet,[67] XceptionNet,[68] and SqueezeNet[69] were pretrained on the ImageNet dataset[70] before being fine-tuned for other classification tasks. A study by Dawud *et al*[71] detected brain hemorrhage on CT images using a SVM based on a pretrained AlexNet. The findings showed that the pretrained AlexNet-based SVM outperformed a convolutional neural network created from scratch. In addition, Swati *et al*[72] reported similar results for brain tumor detection on multi-modality MRIs. In cardiac image analysis, Mazo *et al*[73] proposed using pretrained ResNet, VGGNet, VGG16, and Inception transfer learning models to identify normal cardiovascular tissues and organs, achieving *F*-score values ranging from 0.717 to 0.928.

Other imaging-related tasks

While most deep learning applications in medical imaging have focused on classification and segmentation tasks, other imaging-related tasks, such as registration, enhancement, and image generation, have also seen significant advancements using deep learning techniques. Image registration addresses the spatial alignment of different medical images, either from different modalities or from different time points. By ensuring that corresponding anatomical structures align correctly, image registration allows for improved treatment planning, monitoring disease progression, and multimodal data integration for comprehensive assessments. Vos *et al*[74] proposed an unsupervised deep learning multi-modality image registration framework based on traditional affine and deformable similarity metrics. This framework was trained on 45 cardiac cine MRI scans and 2060 chest CT images, providing outstanding registration performance for both cardiac cine MRI and chest CT images. In another approach, probability distribution knowledge was incorporated to enhance registration performance through Bayesian deep learning techniques. The architecture achieved improved performance in cardiac MRI and lung CT scans.[75]

In addition to registration, image enhancement techniques have significantly contributed to improving the spatial resolution and signal-to-noise ratio of medical images. Enhanced images enable healthcare professionals to make more accurate diagnoses and assessments because they provide better visibility of anatomical structures and abnormalities. Li[76] developed a cycle-consistent generative adversarial network (CycleGAN) to enhance the signal-to-noise ratio in low-dose CT images. This model demonstrated comparable performance in the peak signal-to-noise ratio and structural similarity index between full- and denoised low-dose CT images. A key advantage of this model is that it does not require aligned full- and low-dose CT image pairs for training. A similar deep learning architecture also demonstrated outstanding enhancement performance for low-field-strength MRI. Lucas *et al*[77] used the CycleGAN architecture to improve 64mT low-field-strength MRI quality. Trained on paired low- and high-field-strength MRIs from 49 multiple sclerosis patients, CycleGAN exhibited state-of-the-art performance in evaluating brain morphometry and white matter lesions.

Image generation has recently emerged as a research interest in deep learning. This modality facilitates the creation of new image modalities or modifies existing ones based on learned patterns and features from training datasets. These tasks can generate entirely new images, simulate variations of existing ones, or enhance certain features, aiding data augmentation, visualization, and personalized medicine. For example, deep learning-based image synthesis methods could reduce the MRI scan times. Benzakoun *et al*[78] proposed a deep learning method to generate synthetic FLAIR images from diffusion-weighted images (DWI) for ischemic stroke patients, significantly reducing MRI duration. The generated synthetic FLAIR images demonstrated comparable diagnostic performance to real FLAIR images in identifying the DWI-FLAIR mismatch, a critical marker for assessing ischemic stroke progression. Another example involves synthetic iodinated contrast agent-free CT angiography image generation using generative adversarial network (GAN)-based methods, reducing the allergy risk associated with iodinated contrast agents.[79]

Foundation Models

Foundation models have gained prominence in AI in recent years, largely due to their impressive performance in NLP. Generative pretrained transformer (GPT) models are one of the best-known foundation models and have demonstrated impressive performance across a variety of NLP tasks, such as question response, translation, and sentence comprehension. Inspired by GPT models, foundation models that learn universal image representation have been developed to tackle computer vision tasks, including segmentation, classification, and visual concepts understanding. These models are typically trained on vast datasets, enabling them to learn general representations and capabilities that can be effectively transferred across various domains and applications [Figure 1]. Most existing foundation models in medical imaging analysis are built upon transformer architecture, which primarily consists of three components: a vision transformer-based image encoder for extracting image features; a prompt encoder for integrating user interactions from different prompt modes, including point, bounding boxes, and masks; and a mask decoder for predicting segmentation results using image and prompt embeddings [Figure 2]. Foundation models differ significantly from traditional pretrained transfer learning models. Pretrained transfer learning models are designed to tackle specific downstream tasks closely related to the pretraining dataset and often require large amounts of labeled data for supervised fine-tuning.[80] However, foundation models can handle a broad range of tasks using a singular set of model weights by zero-shot or few-shot learning, or prompt engineering.[81,82] Based on pretrained dataset modalities and clinical scenarios, foundation models are typically classified into three categories: general vision foundation models, modality-specific vision foundation models, and vision-language foundation models [Table 2].

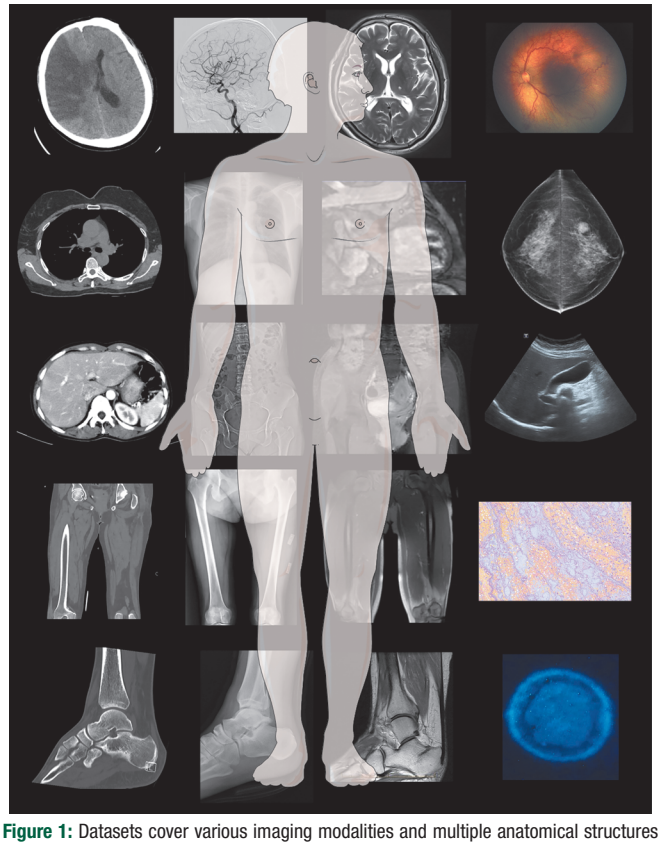

Figure 1: Datasets cover various imaging modalities and multiple anatomical structures in medical imaging.

General vision foundation models

SAM is a new general vision foundation model for image segmentation, pretrained on 11 million natural images with one billion masks. SAM can perform promptable segmentation tasks beyond the datasets used during training with zero-shot generalization, demonstrating impressive performance on natural images.[14] To evaluate its performance on medical images, Roy *et al*[83] used SAM to segment multiple abdominal organs on CT images using promptable points or boxes. Although its performance was not excellent, SAM showed significant potential as a promising candidate for downstream tasks. Deng *et al*[84] also employed SAM to segment tumor, non-tumor tissue, and cell nuclei on whole-slide images (WSI). Their findings revealed that SAM's performance was significantly inferior to that of task-specific models, particularly in cell nuclei segmentation where its performance was deemed unacceptable. Mazurowski *et al*[85] and Shi *et al*[86] observed similar results.

The SAM's significant performance differences on natural *vs*. medical images highlight the fundamental distinctions between these two image types. To address the knowledge gap between natural and medical images, MedSAM fine-tunes SAM using over one million 2D medical images for universal segmentation tasks.[22] MedSAM demonstrated better performance than UNet-based task-specific models, with median Dice scores of 94.0, 94.4, 81.5, and 98.4 for segmenting intracranial hemorrhage on CT, glioma on magnetic resonance (MR) T1, pneumothorax on X-ray, and polyps on endoscopy images, respectively. However, its out-of-the-box performance on vessel-like segmentation tasks remains unsatisfactory.[22] In addition, for 3D volumetric medical images, Wang *et al*[87] introduced SAM-Med3D, a modified version of SAM designed for 3D segmentation tasks. Trained from scratch on over 131,000 3D masks across 247 categories, SAM-Med3D outperforms both SAM and MedSAM in multiple organ and lesion segmentation tasks.

The aforementioned works demonstrate that a unified, generalized foundation model cannot achieve excellent performance across both natural and medical images due to the significant feature differences between them. However, these foundation models may potentially improve performance in specific medical imaging modalities. Currently, foundational models have been proposed primarily for segmentation tasks, but more comprehensive foundational models that span multiple tasks may be developed in the future.

Modality-specific foundation models

Due to the underperformance of general foundation models in medical imaging analysis tasks, modality-specific

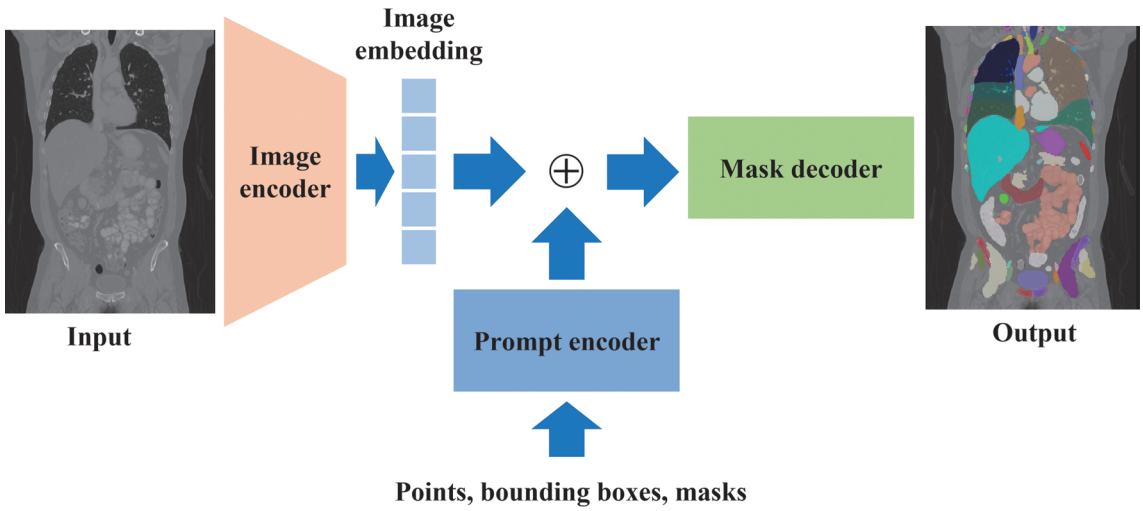

Figure 2: The typical architecture of an existing foundation model in medical imaging analysis. The image encoder extracts image features, and the prompt encoder integrates user interactions from different prompt modes, including points, bounding boxes, and masks. The mask decoder is used to predict segmentation outputs using image and prompt embeddings.

foundation models have been developed to address the unique challenges associated with particular imaging modalities. Modality-specific foundation models are typically built upon general foundation models and are trained on large-scale medical images from a specific modality to leverage the images' unique characteristics.

Modality-specific foundation models for 2D medical imaging analysis were initially developed because the 2D image dimensions are similar to those of natural images. The primary 2D medical imaging modalities include computational pathology, retinal imaging, transparent medical images, and X-rays. Xu *et al*[6] developed a pathology foundation model for WSI images, supporting 9 cancer subtyping tasks and 17 pathomics tasks across 31 major tissue types. This model was trained on 1.3 billion 256 × 256 pathology images from 30,000 patients, achieving state-of-the-art performance in 25 out of 26 tasks. For example, for *EGFR* mutation prediction, this foundation model achieved improvements of 23.5% and 66.4% in area under the receiver operating characteristic curve (AUROC) and area under precision-recall curve (AUPRC), respectively, compared to the second-best model. Kim *et al*[88] introduced a dermatological image foundation model called MONET, trained on 105,550 images, and designed to address diagnostic challenges arising from the heterogeneity of diseases, skin tones, and imaging modalities. In 21 classification tasks, such as erythema, MONET achieved a mean AUROC of 0.766, outperforming the fully supervised task-specific ResNet-50 model, which had a mean AUROC of 0.692. Yu *et al*[89] proposed a retinal foundation model to learn universal representations from multimodal retinal images, including color fundus photography and optical coherence tomography. By training on over 180,000 retinal images, this model outperformed others in three diagnostic classification tasks: diabetic retinopathy grading, glaucoma detection, and multi-disease diagnosis, with the AUROC for each task exceeding 0.78. In the X-ray imaging field, X-ray foundation models have also been developed for various tasks, such as chest X-ray diagnosis[1,90,91] and X-ray image segmentation.[92,93]

Compared to 2D medical images, 3D medical images offer more accurate volume estimation, detailed anatomical information, and better spatial context. However, most SAM-based foundation models experience a decline in performance when applied to 3D medical image analysis.[9] To address these challenges, modality-specific foundation models for 3D medical images have recently been proposed. Cox *et al*[94] developed an MRI-specific foundation model for multimodal brain 3D segmentation tasks using a two-stage pretraining approach based on vision transformers. This model was trained on T1- and T2-weighted FLAIR from 41,400 participants. In the first stage, the model learns generalized features, such as the shapes and sizes of various brain structures. The second stage focuses on disease-specific features, such as the geometric shapes of tumors and lesions, as well as their spatial placements within the brain. The model demonstrated significantly superior performance compared to previously successful fully supervised deep learning models in both tumor and anatomical segmentation tasks, with mean Dice coefficients of 0.9115 and 0.721, respectively. Huang *et al*[26] designed a CT-specific foundation model for universal segmentation tasks, trained on 1204 images with 104 anatomical structures. This model also exhibited approximately 10% improvement in the mean Dice coefficients compared to existing supervised deep learning models in segmenting anatomical structures within the TotalSegmentator dataset. Zhang *et al*[95] proposed a brain lesion-specific foundation model for various brain lesion types on MRI. This model was trained on 6585 3D brain MRIs and achieved excellent performance for 14 brain lesion segmentation tasks. In addition to CT and MRIs, Liu *et al*[96] created a 3D optical coherence tomography (OCT) foundation model, named OCTCube, to generalize various diagnostic tasks for retinal disease, outperforming existing models.

Table 2: Examples of foundation models in medical imaging.

Foundation model type	Image modality	Training datasets	Performances	References
General vision foundation models				
SAM	Natural images	11 million natural images with 1 billion masks	Did not achieve state-of-the-art performance in medical imaging.	[14]
MedSAM	Medical images	Over 1 million 2D medical images	Median Dice scores of 94.0%, 94.4%, 81.5%, and 98.4% for segmenting intracranial hemorrhage on CT, glioma on T1w MR, pneumothorax on X-ray, and polyps on endoscopy images, respectively. However, unsatisfactory performance on vessel-like segmentation tasks.	[22]
SAM-Med3D	Medical images	Medical images with over 131,000 3D masks across 247 categories	Outperforms both SAM and MedSAM in multiple organ and lesion segmentation tasks	[87]
Modality-specific vision foundation models				
Pathology foundation model	Pathology images	1.3 billion 256 × 256 pathology images	Achieving state-of-the-art performance in 25 out of 26 tasks and significantly outperforming the second-best method on 18 tasks.	[6]
Dermatological foundation model	Dermatological images	105,550 dermatological images	Outperforming the fully supervised task-specific ResNet-50 model in 21 classification tasks.	[88]
Retinal foundation model	Retinal images	Over 180,000 retinal images	Outperforming others in three diagnostic classification tasks: diabetic retinopathy grading, glaucoma detection, and multi-disease diagnosis.	[89]
MRI-specific foundation model	MR	T1w and T2w FLAIR from 41,400 participants	Outperforming fully supervised learning deep learning models in both tumor and anatomical segmentation tasks.	[94]
CT-specific foundation model	CT	1204 images with 104 anatomical structures	Outperforming fully supervised learning deep learning models in segmenting anatomical structures.	[26]
Brain Lesion-specific foundation model	MR	6585 3D brain MR images	Achieving state-of-the-art performance for 14 brain lesion segmentation tasks.	[95]
3D OCT foundation model	OCT	26,605 3D OCT volumes	Outperforming fully supervised learning deep learning models on 27 out of 29 tasks	[96]
Vision-language foundation models				
Pathology image-text foundation model	X-ray	377,110 chest X-ray images and corresponding radiology reports	Outperformed a fully supervised model in the detection of three pathologies.	[1]
Histopathology vision-language foundation model	Histopathology images	Over 1.17 million histopathology images paired with corresponding captions	Achieving state-of-the-art performance across various downstream tasks, including histology image classification, segmentation, and captioning.	[99]
Retinal vision-language foundation model	Retinal images	1.6 million retinal images with explicit labels	Achieving state-of-the-art performance in the diagnosis and prediction of sight-threatening eye diseases, heart failure, and myocardial infarction.	[100]
CT vision-language foundation model	CT	6+ million images from 15,331 CTs, corresponding EHR diagnosis codes (1.8+ million codes), and corresponding radiology reports (6+ million tokens)	Achieving acceptable performance across 6 evaluation task types comprising 752 individual tasks.	[101]
Echocardiogram vision-language foundation model	Ultrasound	1,032,975 cardiac ultrasound videos and corresponding expert text	AUC of over 0.77 in identifying clinical transitions and assessing cardiac function.	[102]
Text-promptable segmentation foundation model	Medical images	22,000 3D medical images across 497 classes and 6502 anatomical terminologies	Outperforming MedSAM when driven by text prompts.	[103]
Text-promptable foundation model	Medical images	EndoVis2017 and EndoVis2018	Achieving state-of-the-art performance in surgical instrument segmentation.	[104]

AUC: Area of under curve; AUROC: Area under the receiver operating curve; CT: Computed tomography; FLAIR: Fluid attenuation inversion recovery; HER: Electronic health record; MedSAM: SAM in medical images; MRI: Magnetic resonance imaging; OCT: Optical coherence tomography; SAM: Segment Anything Model.

Although modality-specific foundation models sacrifice the ability to handle cross-modality tasks, their performance in handling tasks within a specific modality surpasses that of general foundation models and even exceeds fully supervised deep learning models. This is due to the ability of modality-specific models to leverage the unique characteristics of each imaging modality and optimize their architecture and training processes for those specific modalities. As a result, modality-specific foundation models demonstrate stronger potential in addressing downstream medical tasks more effectively.

Vision-language foundation models

Vision-language foundation models, which comprise a specialized subset of foundation models, combine visual and linguistic information to enhance medical image interpretation and analysis.[97] These models are trained to encode images and text into unified and compact representations, enabling them to perform a wide range of prediction tasks with zero-shot generalization. In clinical practice, medical images are often paired with corresponding radiology reports. These reports embed rich medical imaging knowledge, offering a strong foundation for the rapid development of vision-language foundation models in the field of medical imaging.[98] Existing research on vision-language foundation models primarily focuses on two key areas: automated medical image interpretation and text-prompted medical image analysis.

In the area of automated medical image interpretation, Tiu *et al*[1] presented a self-supervised image-text foundation model for pathology classification tasks. It was trained on chest X-ray images paired with radiology reports and demonstrated performance comparable to those of radiologists. Li *et al*[98] developed a vision-language foundation model to refine redundant descriptions in radiology reports. In histopathology, Lu *et al*[99] introduced a vision-language foundation model named CONCH, which was trained on over 1.17 million histopathology images paired with corresponding captions. CONCH achieved state-of-the-art performance across various downstream tasks, including histology image classification, segmentation, and captioning. Similarly, Zhou *et al*[100] designed a vision-language foundation model specifically for retinal images to overcome the limitations of task-specific models. By training on 1.6 million retinal images with explicit labels, this model exhibited superior performance in the diagnosis and prediction of sight-threatening eye diseases, heart failure, and myocardial infarction through zero-shot or few-shot fine-tuning. In addition to foundation models for 2D medical images, Blankemeier *et al*[101] and Christensen *et al*[102] developed vision-language foundation models for 3D CT images and echocardiogram videos, both of which achieved state-of-the-art performance.

Text-promptable medical image analysis aims to streamline the semi-automated annotation process. High-quality and publicly available annotated medical data are essential for developing foundation models. However, the data collection is often expensive and time-consuming due to the high level of domain expertise required. Zhao *et al*[103] developed a text-promptable segmentation foundation model trained on 22,000 3D medical images across 497 classes and 6502 anatomical terminologies. Compared to MedSAM, this model demonstrates superior performance, scalability, and robustness when driven by text prompts. Similarly, Zhou *et al*[104] proposed a text-promptable foundation model for surgical instrument segmentation.

Vision-language foundation models, which integrate image and text knowledge, hold potential for healthcare applications. These applications range from generating descriptive captions for medical images to supporting decision-making systems. However, these models are still in their early stages, as they currently incorporate large-scale language models as a prompting module without achieving deep vision and linguistic knowledge integration. In the future, as information silos in hospitals are dismantled, vision-language foundation models are expected to play an increasingly critical role in healthcare applications.

Differences between Task-Specific and Foundation Models

As described above, task-specific models have been applied to a wide range of specialized tasks in medical image analysis, while foundation models have demonstrated potential for deeper integration into the medical imaging workflow, offering a more generalized and scalable approach. Although both task-specific and foundation models achieve state-of-the-art performance in medical imaging, they differ in their inherent characteristics.

First, task-specific models primarily focus on addressing particular tasks, often using a single or limited type of imaging modality to achieve optimal performance in a specific domain.[8] In contrast, foundation models aim to develop generalized representations that enable them to perform multitasking and multimodal analyses across diverse imaging modalities and tasks.[8] For example, although modality-specific foundation models are trained on a specific imaging modality, their datasets include a diverse range of subtypes within that modality and cover multiple human organs. However, some task-specific models are also trained using a specific image modality, but their training datasets include only a few modality-specific image subtypes and focus on a particular organ, such as ischemic lesion segmentation using brain DWI, FLAIR, and apparent diffusion coefficient, or breast cancer identification using mammography. In addition, there are significant differences between general foundation models and task-specific models. General foundation models emphasize managing multiple downstream tasks, such as classification, detection, segmentation, and registration, using a single set of model weights.[9,105] While transfer learning in task-specific models shares similarities with foundation models, there are natural distinctions between them. Pretrained transfer learning models are designed to address specific downstream tasks closely related to the pretraining dataset and often achieve outstanding performance in those areas.[106] These models focus on optimizing performance for particular tasks rather than expanding their applicability to

a wider range of tasks. Foundation models, on the other hand, focus more on handling a broader range of tasks through zero-shot or few-shot fine-tuning.[20]

Second, data form the cornerstone for both task-specific and foundation models, but the models' data requirements differ significantly. Task-specific models typically require a single or limited image modality type to achieve good performance. In contrast, foundation models, particularly general foundation ones, benefit from exposure to various image modalities to enhance their generalization capabilities. Regarding image count, while many task-specific models can perform well with relatively smaller datasets, foundation models often rely on extensive datasets to perform optimally. There has always been a shortage of publicly available, high-quality annotated datasets in medical imaging for training deep learning models. Vision-language foundation models provide a highly efficient text-promptable annotation approach, which is essential for overcoming the medical data bottleneck.

Next, task-specific and foundation models differ in their medical image analysis task performance. General foundation models often underperform compared to task-specific models when handling specific medical tasks due to the significant feature differences between natural and medical images, as well as variability within different medical image modalities. Natural images are primarily acquired from visible light imaging, while medical images are obtained from a variety of imaging devices with different physical properties and energy sources, including light, X-rays, ultrasound, nuclear imaging, and magnetic resonance. These differences in physical imaging principles lead to image feature variations. However, many modality-specific foundation models can outperform task-specific models, which may improve with large-scale datasets and reduce feature complexity.

Finally, task-specific models currently have a broader range of clinical applications compared to foundation models. While foundation models can perform multiple tasks using a single set of model weights, most are primarily developed for segmentation and classification analyses. In contrast, task-specific models encompass a wider array of medical image analysis tasks, including classification, segmentation, registration, localization, detection, enhancement, and prediction.

Complementarities between Task-Specific and Foundation Models

Although task-specific and foundation models have some natural differences, they also have a strong complementary relationship. Looking back at AI history, especially with the rise of deep learning techniques, many traditional machine learning methods have been replaced by deep learning in medical image analysis. During AI development from task-specific models to foundation models, the relationship between novel techniques and traditional methods is complementary.

Foundation models effectively compensate for the shortcomings of task-specific models with long-tailed data. In these scenarios, the datasets are often heavily imbalanced, with common disease cases coexisting with relatively few rare disease cases. This imbalance leads to significant performance degradation. In task-specific models, however, data augmentation techniques increase the few annotated cases to fully leverage the available supervised data. The techniques include rotation, cropping, and noise addition. However, the performance improvement remains limited because no additional information is introduced for training task-specific models. Compared to task-specific models, the few-shot capabilities of foundation models perfectly address the performance degradation seen with rare diseases with long-tailed data. Foundation models are pretrained on large-scale datasets to learn general representations, reducing the amount of labeled data required for training. As a result, foundation models can perform well on rare disease cases using zero-shot or few-shot fine-tuning.

Generally, both foundation and task-specific models handle medical image analysis tasks by gaining knowledge. Foundation and task-specific models can represent the initial and advanced stages of knowledge acquisition. Foundation models focus on learning generalized representations in the initial learning stage, while task-specific models concentrate on acquiring in-depth and domain-specific knowledge at the advanced learning stage. Integrating both generalized and specialized knowledge leads to a comprehensive understanding. Therefore, combining the strengths of foundation and task-specific models may further improve their performance in medical image analysis.

Task-specific and foundation models may also mutually reinforce each other in technological development. Due to the widespread use of task-specific models in medical imaging, large-scale and high-quality publicly available annotated datasets have accumulated. These datasets provide a solid foundation for foundation model development. In addition, the development of vision-language foundation models, which integrate visual and linguistic interpretations, significantly reduces the difficulties associated with annotating datasets and further enhances task-specific model performance. As a result, creating complementary deep learning models that combine multimodal data, such as images, text, and voice, could be a promising field for future research.

Future Directions and Challenges

Both task-specific and foundation models show strong potential to drive advancements in medical imaging, based on their respective performances.

One future research direction could be to develop foundation models with a stronger capability for learning generalized representations that can incorporate a range of scales (cell, tissue, organ, and the whole body) and various image types (light, electrons, lasers, X-rays, ultrasound, nuclear physics, and MR), as well as non-image data (text, voice, and video). These foundation models could provide a more comprehensive understanding of diseases and aid medical professionals in developing more accurate and personalized treatment plans. A potential

clinical application would make the radiology workflow more efficient and accurate. Image, text, and audio are the primary data modalities used throughout the radiology workflow. By leveraging the multi-modality foundation models that integrate these modalities, they could freely transform knowledge representations, benefiting the automatic drafting of structured radiology reports, describing possible abnormalities, and diagnosing diseases.

Another future research direction could be to design deep learning architectures with cross-task connections from foundation models to downstream task-specific models, which is essential for expanding clinical applications. For instance, segmentation foundation models can be used to refine a downstream registration-specific model. These cross-task architectures provide the capability to fine-tune a broader range of downstream task-specific models, making them adaptable to more clinical scenarios. One potential cross-task architecture is the integration of foundation and task-specific models. Foundation models are used to store generalized knowledge, while task-specific models are designed as output modes, similar to multi-plug outlets, that adapt to different scenarios. Existing foundation models, such as SAM, are designed to learn knowledge about segmenting objects from large-scale images. However, in cross-task architectures, foundation models are built to gain knowledge not specifically for segmentation, but for generalized knowledge without specific goals. The task-specific models in cross-task architectures are designed like multi-plug outlets, allowing the use of generalized knowledge stored in foundation models across different clinical scenarios.

However, advancing the use of task-specific and foundation models in medical imaging is challenging. Large-scale and high-quality open-source datasets not only drive the rapid development of these models but also raise more concerns about privacy preservation. Therefore, one of the future challenges is how to share large-scale knowledge flexibly and effectively. If foundation models can be developed to acquire generalized knowledge without specific goals, this knowledge could be stored within the models themselves, rather than relying on original datasets. This approach could help overcome current limitations. Furthermore, foundation models such as SAM and MedSAM have been integrated into medical image software for annotating datasets. It is still unclear whether AI-aided annotated datasets can continuously improve the performance of deep learning models. One study demonstrated that training on generated datasets may cause irreversible defects in large language models.[107] Consequently, another challenge is how to manage the use of AI-aided annotated datasets.

Conclusion

In conclusion, we introduced the clinical applications of task-specific and foundation models in medical imaging. Task-specific models have been widely applied to almost all medical image analysis tasks. Although foundation models currently focus on segmentation and classification tasks, they could expand into other clinical scenarios. We also elucidated the differences and complementarities between task-specific and foundation models. Unlike the relationship in which deep learning techniques have replaced traditional machine learning methods, task-specific models and foundation models have a more complementary relationship, despite some inherent distinctions. Given their outstanding performance in addressing medical image analysis tasks, we explored the models' future research directions and potential challenges. Foundation models with a stronger capability for learning generalized representations and creating novel deep learning architecture with cross-task connections between foundation models and downstream task-specific models may be developed. In summary, task-specific and foundation models have the potential to drive breakthroughs in medical imaging, from image processing to clinical workflows.

Acknowledgment

The authors acknowledge the valuable contributions of Chen Zhang at MR research collaboration, Siemens Healthineers, for her insightful comments and feedback.

Funding

This study was supported by grants from Beijing Hospitals Authority's Ascent Plan (No. DFL20220303) and Beijing Municipal Science & Technology Commission (No. Z221100003522008).

Conflicts of interest

None.

References

1. Tiu E, Talius E, Patel P, Langlotz CP, Ng AY, Rajpurkar P. Expert-level detection of pathologies from unannotated chest X-ray images via self-supervised learning. Nat Biomed Eng 2022;6:1399–1406. doi: 10.1038/s41551-022-00936-9.
2. Gillebert CR, Humphreys GW, Mantini D. Automated delineation of stroke lesions using brain CT images. Neuroimage Clin 2014;4:540–548. doi: 10.1016/j.nicl.2014.03.009.
3. Deng M, Yu R, Wang L, Shi F, Yap PT, Shen D. Learning-based 3T brain MRI segmentation with guidance from 7T MRI labeling. Med Phys 2016;43:6588–6597. doi: 10.1118/1.4967487.
4. Jun TJ, Kang SJ, Lee JG, Kweon J, Na W, Kang D, *et al*. Automated detection of vulnerable plaque in intravascular ultrasound images. Med Biol Eng Comput 2019;57:863–876. doi: 10.1007/s11517-018-1925-x.
5. Pieszko K, Shanbhag A, Killekar A, Miller RJH, Lemley M, Otaki Y, *et al*. Deep learning of coronary calcium scores from PET/CT attenuation maps accurately predicts adverse cardiovascular events. JACC Cardiovasc Imaging 2023;16:675–687. doi: 10.1016/j.jcmg.2022.06.006.
6. Xu H, Usuyama N, Bagga J, Zhang S, Rao R, Naumann T, *et al*. A whole-slide foundation model for digital pathology from real-world data. Nature 2024;630:181–188. doi: 10.1038/s41586-024-07441-w.
7. Liang XW, Cai YY, Yu JS, Liao JY, Chen ZY, Wang NN. Update on thyroid ultrasound: A narrative review from diagnostic criteria to artificial intelligence techniques. Chin Med J 2019;132:1974–1982. doi: 10.1097/CM9.0000000000000346.
8. Zhang S, Metaxas D. On the challenges and perspectives of foundation models for medical image analysis. Med Image Anal 2024;91:102996. doi: 10.1016/j.media.2023.102996.
9. Ren Z, Zhang Y, Wang S. Large foundation model for cancer segmentation. Technol Cancer Res Treat 2024;23:15330338241266205. doi: 10.1177/15330338241266205.

10. Kim S, Lee CK, Kim SS. Large language models: A guide for radiologists. Korean J Radiol 2024;25:126–133. doi: 10.3348/kjr.2023.0997.
11. Schneider J, Meske C, Kuss P. Foundation models: A new paradigm for artificial intelligence. Bus Inform Syst Eng 2024;66:221–231. doi: 10.1007/s12599-024-00851-0.
12. Yan Y, Chen M, Shyu ML, Chen SC. Deep learning for imbalanced multimedia data classification. 2015 IEEE International Symposium on Multimedia (ISM). IEEE; 2015: 483–488.
13. Brown TB. Language models are few-shot learners. arXiv preprint ArXiv:200514165 2020.
14. Kirillov A, Mintun E, Ravi N, Mao H, Rolland C, Gustafson L, *et al*. Segment anything. Proceedings of the IEEE/CVF International Conference on Computer Vision. 2023: 4015–4026. doi: 10.1109/ICCV51070.2023.00371
15. Zou X, Yang J, Zhang H, Li F, Li L, Wang J, *et al*. Segment everything everywhere all at once. Adv Neural Inf Process Syst 2023;36:1–14.
16. Wu C, Yin S, Qi W, Wang X, Tang Z, Duan N. Visual ChatGPT: Talking, drawing and editing with visual foundation models. 2023.
17. Wang G, Li W, Zuluaga MA, Pratt R, Patel PA, Aertsen M, *et al*. Interactive medical image segmentation using deep learning with image-specific fine tuning. IEEE Trans Med Imaging 2018;37:1562–1573. doi: 10.1109/TMI.2018.2791721.
18. Avanzo M, Wei L, Stancanello J, Vallières M, Rao A, Morin O, *et al*. Machine and deep learning methods for radiomics. Med Phys 2020;47:e185–e202. doi: 10.1002/mp.13678.
19. Ortiz Salvador M, Montero Hernández J, Castro Navarro V. Multimodal imaging in laser pointer maculopathy. Arch Soc Esp Oftalmol 2020;95:e44. doi: 10.1016/j.oftal.2019.12.015.
20. Zhang Y, Shen Z, Jiao R. Segment Anything Model for medical image segmentation: Current applications and future directions. Comput Biol Med 2024;171:108238. doi: 10.1016/j.compbiomed.2024.108238.
21. Greenwald NF, Miller G, Moen E, Kong A, Kagel A, Dougherty T, *et al*. Whole-cell segmentation of tissue images with human-level performance using large-scale data annotation and deep learning. Nat Biotechnol 2022;40:555–565. doi: 10.1038/s41587-021-01094-0.
22. Ma J, He Y, Li F, Han L, You C, Wang B. Segment anything in medical images. Nat Commun 2024;15:654. doi: 10.1038/s41467-024-44824-z.
23. Ghesu FC, Georgescu B, Mansoor A, Yoo Y, Neumann D, Patel P, *et al*. Self-supervised learning from 100 million medical images. arXiv 2022. arXiv preprint arXiv:220101283.
24. Vorontsov E, Bozkurt A, Casson A, Shaikovski G, Zelechowski M, Liu S, *et al*. Virchow: A million-slide digital pathology foundation model. arXiv 2023. arXiv preprint arXiv:230907778.
25. Chen S, Ma K, Zheng Y. Med3d: Transfer learning for 3d medical image analysis. arXiv 2019. arXiv preprint arXiv:190400625.
26. Huang Z, Wang H, Deng Z, Ye J, Su Y, Sun H, *et al*. Stu-net: Scalable and transferable medical image segmentation models empowered by large-scale supervised pre-training. arXiv 2023. arXiv preprint arXiv:230406716.
27. Sistaninejhad B, Rasi H, Nayeri P. A review paper about deep learning for medical image analysis. Comput Math Methods Med 2023;2023:7091301. doi: 10.1155/2023/7091301.
28. Zhou E, Lee D. Generative artificial intelligence, human creativity, and art. PNAS Nexus 2024;3:1–8. doi: 10.1093/pnasnexus/pgae052.
29. Zhang Q, Burrage MK, Lukaschuk E, Shanmuganathan M, Popescu IA, Nikolaidou C, *et al*. Toward replacing late gadolinium enhancement with artificial intelligence virtual native enhancement for gadolinium-free cardiovascular magnetic resonance tissue characterization in hypertrophic cardiomyopathy. Circulation 2021;144:589–599. doi: 10.1161/CIRCULATIONAHA.121.054432.
30. Deng B, Yao Y, Dyke RM, Zhang J. A survey of non-rigid 3D registration. Comput Graph Forum 2022;41:559–589. doi: 10.1111/cgf.14502.
31. Zhang YQ, Liu AF, Man FY, Zhang YY, Li C, Liu YE, *et al*. MRI radiomic features-based machine learning approach to classify ischemic stroke onset time. J Neurol 2022;269:350–360. doi: 10.1007/s00415-021-10638-y.
32. Hashimoto A, Kudo H. Ordered-subsets EM algorithm for image segmentation with application to brain MRI. 2000 IEEE Nuclear Science Symposium. Conference Record (Cat. No.00CH37149). (Vol. 3), 2000:118–118.
33. Wei-Ying MA, Manjunath BS. EdgeFlow: A technique for boundary detection and image segmentation. IEEE Trans Image Process 2000;9:1375–1388. doi: 10.1109/83.855433.
34. Jorge Cardoso M, Leung K, Modat M, Keihaninejad S, Cash D, Barnes J, *et al*. STEPS: Similarity and truth estimation for propagated segmentations and its application to hippocampal segmentation and brain parcelation. Med Image Anal 2013;17:671–684. doi: 10.1016/j.media.2013.02.006.
35. Lian Y, Song Z. Automated brain tumor segmentation in magnetic resonance imaging based on sliding-window technique and symmetry analysis. Chin Med J 2014;127:462–468. doi: 10.3760/cma.j.issn.0366-6999.20132554.
36. Shelhamer E, Long J, Darrell T. Fully convolutional networks for semantic segmentation. IEEE Trans Pattern Anal Mach Intell 2017;39:640–651. doi: 10.1109/TPAMI.2016.2572683.
37. Havaei M, Davy A, Warde-Farley D, Biard A, Courville A, Bengio Y, *et al*. Brain tumor segmentation with Deep Neural Networks. Med Image Anal 2017;35:18–31. doi: 10.1016/j.media.2016.05.004.
38. Hesse LS, Aliasi M, Moser F, INTERGROWTH-21(st) Consortium, Haak MC, Xie W, *et al*. Subcortical segmentation of the fetal brain in 3D ultrasound using deep learning. Neuroimage 2022;254:119117. doi: 10.1016/j.neuroimage.2022.119117.
39. Islam M, Sanghani P, See AAQ, James ML, King NKK, Ren H. ICHNet: Intracerebral hemorrhage (ICH) segmentation using deep learning. Springer International Publishing; 2019.
40. Zhang F, Breger A, Cho KIK, Ning L, Westin CF, O'Donnell LJ, *et al*. Deep learning based segmentation of brain tissue from diffusion MRI. Neuroimage 2021;233:117934. doi: 10.1016/j.neuroimage.2021.117934.
41. Rebsamen M, Rummel C, Reyes M, Wiest R, McKinley R. Direct cortical thickness estimation using deep learning-based anatomy segmentation and cortex parcellation. Hum Brain Mapp 2020;41:4804–4814. doi: 10.1002/hbm.25159.
42. Cai JC, Akkus Z, Philbrick KA, Boonrod A, Hoodeshenas S, Weston AD, *et al*. Fully automated segmentation of head CT neuroanatomy using deep learning. Radiol Artif Intell 2020;2:e190183. doi: 10.1148/ryai.2020190183.
43. Liu CF, Hsu J, Xu X, Ramachandran S, Wang V, Miller MI, *et al*. Deep learning-based detection and segmentation of diffusion abnormalities in acute ischemic stroke. Commun Med (Lond) 2021;1:61. doi: 10.1038/s43856-021-00062-8.
44. Soltanpour M, Greiner R, Boulanger P, Buck B. Improvement of automatic ischemic stroke lesion segmentation in CT perfusion maps using a learned deep neural network. Comput Biol Med 2021;137:104849. doi: 10.1016/j.compbiomed.2021.104849.
45. Kumar A, Upadhyay N, Ghosal P, Chowdhury T, Das D, Mukherjee A, *et al*. CSNet: A new DeepNet framework for ischemic stroke lesion segmentation. Comput Methods Programs Biomed 2020;193:105524. doi: 10.1016/j.cmpb.2020.105524.
46. Trebeschi S, van Griethuysen JJM, Lambregts DMJ, Lahaye MJ, Parmar C, Bakers FCH, *et al*. Deep learning for fully-automated localization and segmentation of rectal cancer on multiparametric MR. Sci Rep 2017;7:5301. doi: 10.1038/s41598-017-05728-9.
47. Hodneland E, Dybvik JA, Wagner-Larsen KS, Šoltészová V, Munthe-Kaas AZ, Fasmer KE, *et al*. Automated segmentation of endometrial cancer on MR images using deep learning. Sci Rep 2021;11:179. doi: 10.1038/s41598-020-80068-9.
48. Ramesh S, Sasikala S, Gomathi S, Geetha V, Anbumani V. Segmentation and classification of breast cancer using novel deep learning architecture. Neural Comput Appl 2022;34:16533–16545. doi: 10.1007/s00521-022-07230-4.
49. Almotairi S, Kareem G, Aouf M, Almutairi B, Salem MAM. Liver tumor segmentation in CT scans using modified SegNet. Sensors (Basel) 2020;20:1516. doi: 10.3390/s20051516.
50. Rahman H, Bukht TFN, Imran A, Tariq J, Tu S, Alzahrani A. A deep learning approach for liver and tumor segmentation in CT images using ResUNet. Bioengineering (Basel) 2022;9:368. doi: 10.3390/bioengineering9080368.
51. Chanchal AK, Kumar A, Lal S, Kini J. Efficient and robust deep learning architecture for segmentation of kidney and breast histopathology images. Comput Electri Eng 2021;92:104075. doi: 10.1016/j.compeleceng.2021.107177.
52. Boudegga H, Elloumi Y, Akil M, Hedi Bedoui M, Kachouri R, Abdallah AB. Fast and efficient retinal blood vessel segmentation method

based on deep learning network. Comput Med Imaging Graph 2021;90:101902. doi: 10.1016/j.compmedimag.2021.101902.
53. Othman MFB, Abdullah NB, Kamal NFB. MRI brain classification using support vector machine. 2011 Fourth International Conference on Modeling, Simulation and Applied Optimization. IEEE; 2011:1–4.
54. Xie YB, Liu Q, He F, Guo CG, Wang CF, Zhao P. Diagnosis of colon cancer with Fourier transform infrared spectroscopy on the malignant colon tissue samples. Chin Med J 2011;124:2517–2521.
55. Carneiro G, Georgescu B, Good S, Comaniciu D. Detection and measurement of fetal anatomies from ultrasound images using a constrained probabilistic boosting tree. IEEE Trans Med Imaging 2008;27:1342–1355. doi: 10.1109/TMI.2008.928917.
56. Nedjar I, Daho MEH, Settouti N, Mahmoudi S, Chikh MA. Random forest based classification of medical X-ray images using a genetic algorithm for feature selection. J Mech Med Biol 2015;15:1540025. doi: 10.1142/S0219519415400254.
57. Arbach L, Reinhardt JM, Bennett DL, Fallouh G. Mammographic masses classification: Comparison between backpropagation neural network (BNN), K nearest neighbors (KNN), and human readers. CCECE 2003 – Canadian Conference on Electrical and Computer Engineering. Toward a Caring and Humane Technology (Cat. No.03CH37436). (Vol. 3), IEEE;1441–1444.
58. Chen S, Zhang J, Wei X, Zhang Q. Alzheimer's disease classification using structural MRI based on convolutional neural networks. ACM International Conference Proceeding Series. 2020:7–13.
59. Nawaz M, Sewissy AA, Soliman THA. Multi-class breast cancer classification using deep learning convolutional neural network. Int J Adv Comp Sci Appl 2018;9:316–322. doi: 10.14569/IJACSA.2018.090645.
60. Paul R, Hawkins SH, Balagurunathan Y, Schabath MB, Gillies RJ, Hall LO, *et al*. Deep feature transfer learning in combination with traditional features predicts survival among patients with lung adenocarcinoma. Tomography 2016;2:388–395. doi: 10.18383/j.tom.2016.00211.
61. Pan SJ, Yang Q. A survey on transfer learning. IEEE Trans Knowl Data Eng 2009;22:1345–1359. doi: 10.1109/TKDE.2009.191.
62. LeCun Y, Bottou L, Bengio Y, Haffner P. Gradient-based learning applied to document recognition. Proc IEEE 1998;86:2278–2324. doi: 10.1109/5.726791.
63. Krizhevsky A, Sutskever I, Hinton GE. Imagenet classification with deep convolutional neural networks. Adv Neural Inf Process Syst 2012:25.
64. Qassim H, Verma A, Feinzimer D. Compressed residual-VGG16 CNN model for big data places image recognition. 2018 IEEE 8th Annual Computing and Communication Workshop and Conference (CCWC). IEEE; 2018:169–175.
65. He K, Zhang X, Ren S, Sun J. Deep residual learning for image recognition. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2016:770–778.
66. Szegedy C, Vanhoucke V, Ioffe S, Shlens J, Wojna Z. Rethinking the inception architecture for computer vision. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2016:2818–2826.
67. Huang G, Liu Z, Van Der Maaten L, Weinberger KQ. Densely connected convolutional networks. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2017:4700–4708.
68. Chollet F. Xception: Deep learning with depthwise separable convolutions. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2017:1251–1258.
69. Iandola FN, Han S, Moskewicz MW, Ashraf K, Dally WJ, Keutzer K. SqueezeNet: AlexNet-level accuracy with 50x fewer parameters and <0.5 MB model size. arXiv 2016. arXiv preprint arXiv:160207360.
70. Szegedy C, Liu W, Jia Y, Sermanet P, Reed S, Anguelov D, *et al*. Going deeper with convolutions. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2015:1–9.
71. Dawud AM, Yurtkan K, Oztoprak H. Application of deep learning in neuroradiology: Brain haemorrhage classification using transfer learning. Comput Intell Neurosci 2019;2019:4629859. doi: 10.1155/2019/4629859.
72. Swati ZNK, Zhao Q, Kabir M, Ali F, Ali Z, Ahmed S, *et al*. Brain tumor classification for MR images using transfer learning and fine-tuning. Comput Med Imaging Graph 2019;75:34–46. doi: 10.1016/j.compmedimag.2019.05.001.
73. Mazo C, Bernal J, Trujillo M, Alegre E. Transfer learning for classification of cardiovascular tissues in histological images. Comput Methods Programs Biomed 2018;165:69–76. doi: 10.1016/j.cmpb.2018.08.006.
74. de Vos BD, Berendsen FF, Viergever MA, Sokooti H, Staring M, Išgum I. A deep learning framework for unsupervised affine and deformable image registration. Med Image Anal 2019;52:128–143. doi: 10.1016/j.media.2018.11.010.
75. Deshpande VS, Bhatt JS. Bayesian deep learning for deformable medical image registration. International Conference on Pattern Recognition and Machine Intelligence. Springer; 2019:41–49.
76. Li Z. Investigation of low-dose CT image denoising using unpaired deep learning methods. Physiol Behav 2018;176:139–148. doi: 10.4049/jimmunol.1801473.The.
77. Lucas A, Campbell Arnold T, Okar SV, Vadali C, Kawatra KD, Ren Z, *et al*. Multi-contrast high-field quality image synthesis for portable low-field MRI using generative adversarial networks and paired data. medRxiv 2023:2023.12.28.23300409. doi: 10.1101/2023.12.28.23300409.
78. Benzakoun J, Deslys MA, Legrand L, Hmeydia G, Turc G, Hassen WB, *et al*. Synthetic FLAIR as a substitute for FLAIR sequence in acute ischemic stroke. Radiology 2022;303:153–159. doi: 10.1148/RADIOL.211394.
79. Lyu J, Fu Y, Yang M, Xiong Y, Duan Q, Duan C, *et al*. Generative adversarial network–based noncontrast CT angiography for aorta and carotid arteries. Radiology 2023;309:e230681. doi: 10.1148/radiol.230681.
80. Kora P, Ooi CP, Faust O, Raghavendra U, Gudigar A, Chan WY, *et al*. Transfer learning techniques for medical image analysis: A review. Biocybern Biomed Eng 2022;42:79–107. doi: 10.1016/j.bbe.2021.11.004.
81. Chen RJ, Ding T, Lu MY, Williamson DFK, Jaume G, Song AH, *et al*. Towards a general-purpose foundation model for computational pathology. US: Springer; 2024.
82. Moor M, Banerjee O, Abad ZSH, Krumholz HM, Leskovec J, Topol EJ, *et al*. Foundation models for generalist medical artificial intelligence. Nature 2023;616:259–265. doi: 10.1038/s41586-023-05881-4.
83. Roy S, Wald T, Koehler G, Rokuss MR, Disch N, Holzschuh J, *et al*. SAM. MD: Zero-shot medical image segmentation capabilities of the Segment Anything Model. 2023:1–4.
84. Deng R, Cui C, Liu Q, Yao T, Remedios LW, Bao S, *et al*. Segment Anything Model (SAM) for digital pathology: Assess zero-shot segmentation on whole slide imaging. 2023:1–6.
85. Mazurowski MA, Dong H, Gu H, Yang J, Konz N, Zhang Y. Segment Anything Model for medical image analysis: An experimental study. Med Image Anal 2023;89:102918. doi: 10.1016/j.media.2023.102918.
86. Shi P, Qiu J, Abaxi SMD, Wei H, Lo FP-W, Yuan W. Generalist vision foundation models for medical imaging: A case study of Segment Anything Model on zero-shot medical segmentation. Diagnostics (Basel) 2023;13:1947. doi: 10.3390/diagnostics13111947.
87. Wang H, Guo S, Ye J, Deng Z, Cheng J, Li T, *et al*. SAM-Med3D. 2023:1–17.
88. Kim C, Gadgil SU, DeGrave AJ, Omiye JA, Cai ZR, Daneshjou R, *et al*. Transparent medical image AI via an image–text foundation model grounded in medical literature. Nat Med 2024;30:1154–1165. doi: 10.1038/s41591-024-02887-x.
89. Yu K, Zhou Y, Bai Y, Da Soh Z, Xu X, Goh RSM, *et al*. UrFound: Towards universal retinal foundation models via knowledge-guided masked modeling. 2024;1:1–17.
90. Rajpurkar P, Irvin J, Zhu K, Yang B, Mehta H, Duan T, *et al*. CheXNet: Radiologist-level pneumonia detection on chest X-rays with deep learning. ArXiv 2017:3–9.
91. Chen Z, Varma M, Delbrouck JB, Paschali M, Blankemeier L, Van Veen D, *et al*. CheXagent: Towards a foundation model for chest X-ray interpretation. 2024:1–24.
92. Killeen BD, Wang LJ, Zhang H, Armand M, Mar CV. FluoroSAM: A Language-aligned Foundation Model for X-ray image segmentation.
93. Du W, Shen H, Fu J. Automatic defect segmentation in X-ray images based on deep learning. IEEE Trans Ind Electr 2020;68:12912–12920. doi: 10.1109/TIE.2020.3047060.
94. Cox J, Liu P, Stolte SE, Yang Y, Liu K, See KB, *et al*. BrainSegFounder: Towards 3D foundation models for neuroimage segmentation. Med Image Anal 2024;97:103301. doi: 10.1016/j.media.2024.103301.
95. Zhang X, Ou N, Basaran BD, Visentin M, Qiao M, Gu R, *et al*. A Foundation model for brain lesion segmentation with mix-

ture of modality experts. arXiv 2024. arXiv preprint arXiv: 240510246.
96. Liu Z, Xu H, Woicik A, Shapiro LG, Blazes M, Wu Y, *et al*. OCTCube: A 3D foundation model for optical coherence tomography that improves cross-dataset, cross-disease, cross-device and cross-modality analysis.
97. Gu J, Han Z, Chen S, Beirami A, He B, Zhang G, *et al*. A systematic survey of prompt engineering on vision-language foundation models. 2023:1–21.
98. Li C, Huang W, Yang H, Liu J, Liang Y, Zheng H, *et al*. Enhancing the vision-language foundation model with key semantic knowledge-emphasized report refinement. Med Image Anal 2024;97:103299. doi: 10.1016/j.media.2024.103299.
99. Lu MY, Chen B, Williamson DFK, Chen RJ, Liang I, Ding T, *et al*. A visual-language foundation model for computational pathology. Nat Med 2024;30:863–874. doi: 10.1038/s41591-024-02856-4.
100. Zhou Y, Chia MA, Wagner SK, Ayhan MS, Williamson DJ, Struyven RR, *et al*. A foundation model for generalizable disease detection from retinal images. Nature 2023;622:156–163. doi: 10.1038/s41586-023-06555-x.
101. Blankemeier L, Cohen JP, Kumar A, Van Veen D, Gardezi SJS, Paschali M, *et al*. Merlin: A vision language foundation model for 3D computed tomography. Res Sq 2024:3.rs–4546309.rs. doi: 10.21203/rs.3.rs-4546309/v1.
102. Christensen M, Vukadinovic M, Yuan N, Ouyang D. Vision–language foundation model for echocardiogram interpretation. Nat Med 2024;30:1481–1488. doi: 10.1038/s41591-024-02959-y.
103. Zhao Z, Zhang Y, Wu C, Zhang X, Zhang Y, Wang Y, *et al*. One model to rule them all: Towards universal segmentation for medical images with text prompts. 2023:1–60.
104. Zhou Z, Alabi O, Wei M, Vercauteren T, Shi M. Text promptable surgical instrument segmentation with vision-language models. Adv Neural Inf Process Syst 2023;36:1–13.
105. Liu C, Jin Y, Guan Z, Li T, Qin Y, Qian B, *et al*. Visual–language foundation models in medicine. Vis Comput 2024. doi: 10.1007/s00371-024-03579-w.
106. Kim HE, Cosa-Linan A, Santhanam N, Jannesari M, Maros ME, Ganslandt T. Transfer learning for medical image classification: A literature review. BMC Med Imaging 2022;22:69. doi: 10.1186/s12880-022-00793-7.
107. Shumailov I, Shumaylov Z, Zhao Y, Gal Y, Papernot N, Anderson R. The curse of recursion: Training on generated data makes models forget. arXiv 2023. arXiv preprint arXiv:230517493.

How to cite this article: Bian YY, Li J, Ye CY, Jia XQ, Yang Q. Artificial intelligence in medical imaging: From task-specific models to large-scale foundation models. Chin Med J 2025;138:651–663. doi: 10.1097/CM9.0000000000003489