Iraqi Journal of Intelligent Computing and Informatics (IJICI)
Vol. 4, 2. December 2025, pp. 194~211
ISSN:2791-2868,DOI: 10.52940/ijici.v4i2.147

Vision and Multimodal Foundation Models in Medical Imaging: A Comprehensive Review of Architectures, Clinical Trends, and Future Directions

Hikmat Z .Neima[1], Rana M. Ghadban[2], Ghaihab H. Adday[1]
[1] Department of Computer Science, College of CSIT, Basra University, Basra, Iraq
[2] Department of Intelligent Medical Systems, College of CSIT, Basra University, Basra, Iraq

Article Info

Article history:

Received Nov., 24, 2025
Revised Dec., 18, 2025
Accepted Dec., 29, 2025

Keywords:

Foundation Models , Medical Imaging , Multimodal Learning, Medicine, Promotable Segmentation , Precision Vision Transformers.

ABSTRACT

Foundation models (FMs) are revolutionizing medical imaging by transitioning from task-specific algorithms to large-scale , generalizable systems that can learn from a broad range of multimodal data. Recent advances in these fields—transformer-based visual encoders , promptable segmentation architectures , vision–language models , and parameter-efficient fine-tuning—have resulted in improved performance among segmentation , detection , classification and report generation techniques in a variety of modalities such as MRI , CT , ultrasound , X-ray , endoscopy , and digital pathology. Domain specific FMs (including prostate MRI, brain MRI , retinal , ultrasound and pathology models) have proved to be effective in providing high label efficiency and competitive or better performance with the mainstream deep learning models , in particular under low-annotation conditions. Trends in the research emphasize such techniques as large-scale pretraining, multimodal integration , cross-task generalization , data-efficient learning , and the development of universal feature encoders. Simultaneously , extensive benchmarking and external validation indicate performance variability , motivating the continued development of standardized evaluation protocols. Adoption by clinical practice has been restricted because of interpretability , bias, workflow integration, computational requirements , and regulatory uncertainty. New options such as personalizable AI , continual learning , federated model adaptation , and imaging–genomics integration , stand out to make FMs key for the future of precision medicine. This article consolidates architectural , pioneering foundation models , clinical evaluation , and translational advancements , drawing upon the current context and future direction of foundation-model medical imaging.

Corresponding Author:
Hikmat Z. Neima
Department of Computer Science, College of CSIT, Basra University, Basra, Iraq
Email: hikmat.taher@uobasrah.edu.iq

1. INTRODUCTION

Foundation models (FMs) have been a disruptive paradigm of AI , where focus moves away from task-driven pipelines to large pre-trained and generalizable to a wide variety of clinical tasks , modalities, and data types. This shift is particularly powerful in medical imaging. Historically , medical image analysis has been performed using models trained on a constrained set of problems – organ segmentation , lesion diagnosis and disease identification , for example – as these require large datasets of annotated images with heavy dependence on the expert's knowledge. The advance of the FMs , however , has shifted the focus of medical image analysis towards unified , multimodal and multitask models that learn from large volumes of highly heterogeneous medical and non-medical data and rapidly learn new tasks with low supervised workload [1] , [2] , [3].

Recent architectural advances (e.g. , transformer-based visual encoders , vision–language models , promptable segmentation , SAM) have increased the representational power of medical AI by allowing for cross-modality reasoning, robust feature extraction, and improved generalization over institutions and populations. There is a growing emphasis on domain-specific FMs , with systems for prostate MRI , brain MRI , retinal images , chest radiographs, ultrasound, endoscopy , pathology and ECG interpretation outperforming traditional approaches especially in regions lacking labels [4] , [5] , [6].

Simultaneously , multimodal and multitask learning is becoming extensively used in the research community , as FMs combine imaging with clinical notes , laboratory data, genomics , and physiological information to help more complete diagnosis and prognosis modelling [7] , [8]. Self-supervised learning , few-shot learning , and parameter-efficient fine-tuning are enabling unprecedented label efficiency , while federated learning and privacy-preserving frameworks offer pathways for large-scale , multi-institutional model development without compromising patient confidentiality [9] , [10] , [11].

Notwithstanding such progress , clinical translation remains limited. Obstacles include interpretability , bias , regulatory uncertainty , computational constraints , and challenges in integrating FMs into real-world