Large-scale Long-tailed Disease Diagnosis on Radiology Images

Qiaoyu Zheng[*,1,2], Weike zhao[*,1,2], Chaoyi Wu[*1,2], Xiaoman Zhang[1,2],, Lisong Dai[1,3], Hengyu Guan[4,5], Yuehua Li[1,3], Ya Zhang[1,2], Yanfeng Wang[1,2,†], and Weidi Xie[1,2,†]

[1] Shanghai Jiao Tong University, Shanghai, China
[2] Shanghai AI Laboratory, Shanghai, China
[3] Shanghai Sixth People's Hospital, Affiliated to Shanghai Jiao Tong University
[4] Department of Reproductive Medicine, Ren Ji Hospital, Shanghai Jiao Tong University School of Medicine
[5] Shanghai Key Laboratory for Assisted Reproduction and Reproductive Genetics

Abstract. Developing a generalist radiology diagnosis system can significantly improve clinical diagnostics. We present **RadDiag**, a foundational model supporting 2D and 3D inputs across various modalities and anatomies, using a transformer-based fusion module for comprehensive disease diagnosis. Due to privacy concerns and data scarcity, we utilize high-quality, clinician-reviewed online radiological images with diagnosis labels. Our dataset, **RP3D-DiagDS**, contains 40,936 cases with 195,010 scans covering 5,568 disorders. **RadDiag** achieves 95.14% AUC on internal evaluation and demonstrates state-of-the-art results when applied to external datasets. This work highlights the potential of publicly shared medical data in developing generalist AI for healthcare.

Keywords: Diagnosis · Radiology images · Multi-modal.

1 Introduction

Radiology techniques have revolutionized medical diagnosis, but existing AI models often struggle with complex clinical scenarios. This paper aims to build a foundational radiology diagnostic model that fuses information from 2D and 3D inputs across different modalities and anatomic sites at the case level.

The main challenges are the lack of datasets, unified architecture, and benchmarks. We address these with 3 contributions: 1) We collect an internet dataset of 40,936 cases with 195,010 scans covering 5568 disorders and 930 ICD-10-CM codes across 7 anatomy regions and 9 modalities. 2) We propose a model architecture supporting multi-modal 2D and 3D inputs, with a transformer-based fusion module and knowledge-enhanced training strategy. 3) We build a comprehensive benchmark to evaluate diagnosis model performance.

Our model serves as a general-purpose image encoder for radiology, enabling efficient fine-tuning or zero-shot transfer learning, improving performance across diverse diagnosis tasks on numerous datasets.

Table 1. Zero-shot results on 6 external datasets. We report the supervised training results with partially down-stream task external data, denoting as "S.T.", for demonstrating what diagnosis performance our model can achieve zero-shotly.

Dataset	Anatomy	Modality	Dim	Method	External Data	F1 ↑	MCC ↑	ACC ↑
Brain-Tumor	Head&Neck	MRI	2D	RadIN	Zero-shot	22.29	3.23	63.63
				BiomedCLIP		56.25	32.81	57.21
				Ours		61.09	47.37	73.38
				S.T.	n=571(10%)	61.77	46.69	72.17
POCUS	Chest	US	Video	RadIN	Zero-shot	2.22	0.17	66.67
				BiomedCLIP		50.33	11.73	45.82
				Ours		65.02	47.97	79.84
				S.T.	n=223(1%)	64.20	48.31	80.58
CT-KIDNEY	Abdom&Pelvis	CT	2D	RadIN	Zero-shot	38.1	17.57	64.75
				BiomedCLIP		49.38	29.58	72.21
				Ours		56.50	43.18	69.96
				S.T.	n=995(10%)	55.73	42.28	66.95
MURA	Limb	X-ray	2D	RadIN	Zero-shot	31.26	4.37	52.94
				BiomedCLIP		68.36	14.96	54.70
				Ours		67.64	40.15	77.51
				S.T.	n=3680(10%)	66.71	41.98	78.32
VinDr-Spine	Spine	X-ray	2D	RadIN	Zero-shot	4.34	0.91	87.49
				BiomedCLIP		26.49	12.79	61.3
				Ours		25.93	19.40	83.07
				S.T.	n=84(1%)	24.73	20.61	84.46
TCGA	Chest&Breast	CT&MG	3D	RadIN	Zero-shot	-	-	-
				BiomedCLIP		-	-	-
				Ours		72.18	57.20	83.47
				S.T.	n=173(30%)	70.33	52.26	77.41

[*] All indicators are presented as percentages, with the % omitted in the table.

2 Results

First, on internal evaluations, our experiments show that adding the fusion module (FM) enhances performance across all diagnostic classes and levels. The knowledge enhancement (KE) , using a pre-trained natural language encoder, further boosts diagnostic accuracy, highlighting the importance of domain knowledge in improving diagnosis.

Second, in zero-shot evaluations across six anatomies and five imaging modalities, our model, RadDiag, outperforms RadIN and BiomedCLIP, demonstrating strong transfer learning capabilities and occasionally matching the performance of models trained on specific datasets as shown in **Table** 1.

Third, finetuning RadDiag on 22 external datasets results in significant performance improvements over models trained from scratch, often surpassing specialist SOTAs. This underscores the value of publicly available medical data as a resource for large-scale supervised training in the medical domain.