Comparing General Vision Transformer Embeddings versus Task-Specific Deep Learning Models for Knee Osteoarthritis Grading

Mohammadreza Chavoshi, MD, Postdoctoral Research Fellow, Department of Radiology, Emory University
Frank Li, PhD; Theo Dapamede, MD, PhD; Bardia Khosravi, MD, MPH; Aawez Mansuri, MS; Rohan Satya Isaac, MS; Janice Newsome, MD; Hari Trivedi, MD; Judy Gichoya, MS, MD, FSIIM

Introduction

Vision transformers are powerful tools for extracting image embeddings. However, the specificity and relevance of these embeddings for specialized diagnostic tasks remain unclear. The Kellgren-Lawrence Grade (KLG) scoring system for knee osteoarthritis (OA) assessment uses specific radiographic features -osteophytes, joint space narrowing, and bone deformity – to grade severity of knee OA, making it a suitable context to evaluate the performance of embeddings for specific diagnostic image analysis.

Hypothesis

To evaluate whether general-purpose medical image transformer embeddings (BioMedCLIP) can capture task-specific radiographic features as effectively as specialized deep learning models for KLG scoring.

Methods

We analyzed bilateral PA fixed-flexion knee radiographs from the NIH Osteoarthritis Initiative dataset. Bilateral images were cropped to unilateral. From 4,796 patients followed up at 12, 24, 36, 48, 72, and 96 months, we included 4,507 patients (38,199 images). We compared three approaches: (1) ConvNeXt and (2) ResNet18, both trained specifically for KLG classification, versus (3) a custom neural-network classifier trained on BioMedCLIP vision transformer embeddings. This design allowed us to contrast the performance of models learning task-specific features directly from images against a model using pre-extracted general medical image embeddings.

Results

Models trained directly on radiographs significantly outperformed the transformer embedding-based approach. ConvNeXt achieved the highest performance (quadratic weighted kappa: 0.8089, adjacent accuracy: 0.9413), followed by ResNet18 (kappa: 0.7474, adjacent accuracy: 0.9110). The BioMedCLIP embedding-based model showed notably lower performance (kappa: 0.5960, adjacent accuracy: 0.7890). All models performed better on extreme grades (0 and 4), with ConvNeXt achieving F1-scores of 0.709 and 0.835 respectively. Notably, the embedding-based approach completely failed to identify grade 1 cases (F1-score: 0.000) and showed substantial degradation in detecting intermediate grades (F1-scores: 0.407 and 0.389 for grades 2 and 3), suggesting limited capture of subtle radiographic features crucial for KLG scoring.

Conclusion

while general medical vision transformers like BioMedCLIP can recognize broad anatomical structures, their embeddings fail to capture the nuanced radiographic features essential for specialized tasks like KLG grading, particularly in distinguishing intermediate disease stages. The superior performance of task-specific deep learning models, especially in detecting subtle grade differences, emphasizes two key points: first, the current limitations of general-purpose medical image embeddings for fine-grained feature detection, and second, the need for developing specialized transformer architectures with focused pre-training on specific anatomical regions and imaging modalities. These results suggest that the successful application of vision transformers in medical imaging may require a shift from general to organ-specific pre-training strategies to ensure clinically relevant feature extraction.

Figure(s)

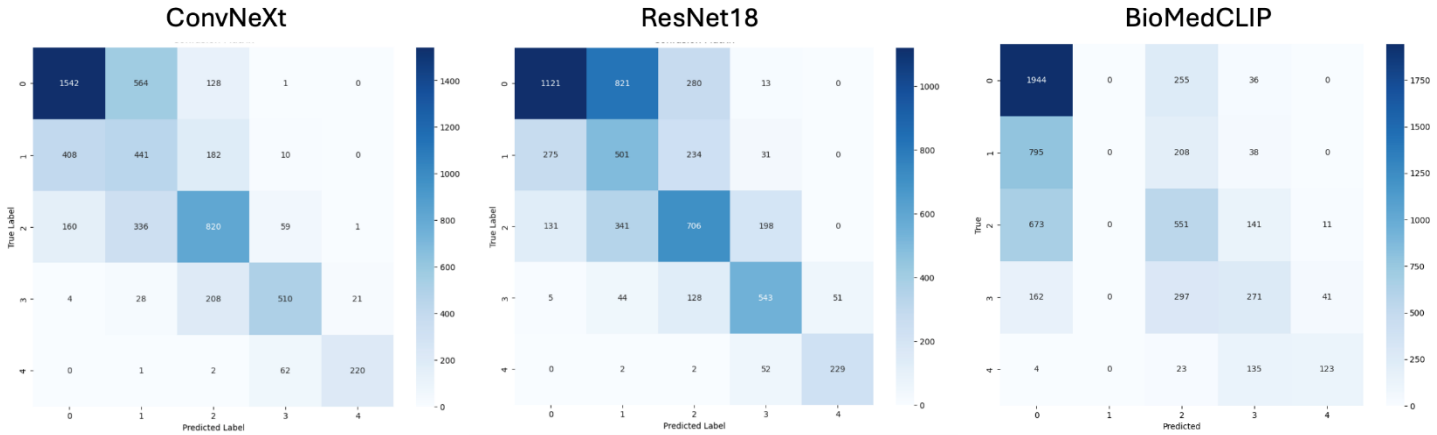

Figure 1. Confusion Matrix for 3 approaches

Metric	Grade	ConvNeXt	ResNet18	BioMedCLIP+NN
Precision	0	0.729	0.732	0.543
	1	0.322	0.293	0.000
	2	0.612	0.523	0.413
	3	0.794	0.649	0.436
	4	0.909	0.818	0.703
Recall	0	0.690	0.502	0.870
	1	0.424	0.481	0.000
	2	0.596	0.513	0.400
	3	0.662	0.704	0.351
	4	0.772	0.804	0.432
F1-score	0	0.709	0.595	0.669
	1	0.366	0.364	0.000
	2	0.604	0.518	0.407
	3	0.722	0.675	0.389
	4	0.835	0.811	0.535
Overall Metrics	Adj. Acc	0.941	0.911	0.789
	QW Kappa	0.809	0.747	0.596

Table 1. Performance metrics for each KLG score

Keywords

Applications; Artificial Intelligence/Machine Learning; Clinical Workflow & Productivity; Imaging Research