Leveraging Pre-trained Medical Image Embeddings for Knee Pain Score Prediction: A Comparative Analysis of Vision Transformer and CNN Approaches

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

Introduction

Knee radiographs are the most commonly used imaging modality to assess osteoarthritis. Despite their widespread use, correlating radiographic findings with patient-reported pain remains challenging due to the complex and subjective nature of pain experience. Studies show variable associations between radiographic osteoarthritis severity and pain intensity.

Hypothesis

We hypothesized that general pre-trained medical image embedding extractors can capture subtle radiographic features associated with pain as effectively as task-specific convolutional networks, potentially offering insights into subtle imaging characteristics that correlate with pain reporting.

Methods

Using the eMoRy Knee Radiograph (MRKR) dataset of 83,011 patients (503,261 knee radiographs), we extracted a subset of 17,157 patients (33,138 unilateral knee images) after excluding cases with inflammatory arthritis, recent trauma, infection, or prior arthroplasty. Pain scores (0-10 scale) were collected during clinical care within 7 days of imaging. We compared four deep learning approaches: ConvNeXt (a CNN architecture), and two vision transformer-based embedding extractors - RAD-DINO (trained via DINOv2 self-supervised learning) and BiomedCLIP (evaluated in both image-only and image-text modes). For BiomedCLIP's multimodal analysis, we generated standardized descriptions based on automatically extracted Kellgren-Lawrence grades.

Results

All models demonstrated comparable performance in predicting pain scores (RMSE 2.47-2.60, MAE 2.02-2.16). Detailed subgroup analyses revealed consistent prediction patterns across age groups (&#60; 45, 45-70, >70), sex, and race. The pain score distribution was relatively uniform across demographic subgroups, with median scores ranging from 4 to 6. All models showed similar error patterns: slight underprediction for higher age groups and minor variations in prediction bias across sex and racial subgroups. Notably, BiomedCLIP with image and text inputs showed the most balanced error distribution across subgroups, suggesting that multimodal analysis may help mitigate demographic biases in pain prediction.

Conclusion

General-purpose medical image embedding models can match task-specific CNNs in predicting knee pain from radiographs, with consistent performance across demographic subgroups. The balanced error distribution in multimodal analysis suggests potential advantages in combining imaging features with structured clinical information. However, moderate prediction errors across all approaches underscore both the inherent complexity of pain assessment from imaging alone and the need for comprehensive clinical evaluation beyond radiographic findings.

Figure(s)

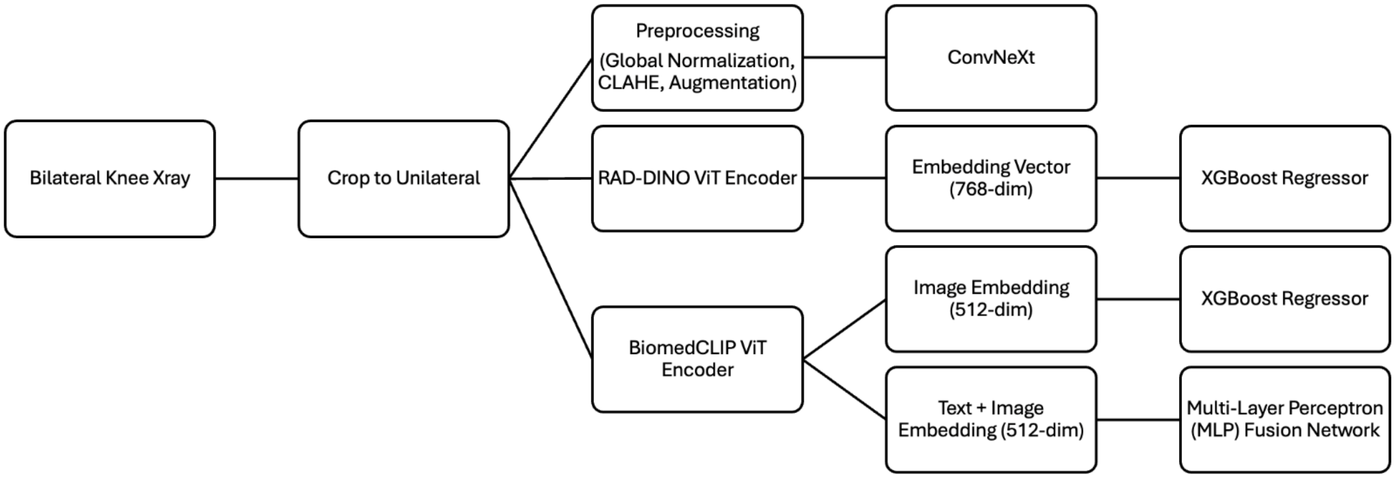
Bilateral Knee Xray
Crop to Unilateral
Preprocessing
(Global Normalization, CLAHE, Augmentation)
ConvNeXt
RAD-DINO ViT Encoder
Embedding Vector (768-dim)
XGBoost Regressor
BiomedCLIP ViT Encoder
Image Embedding (512-dim)
XGBoost Regressor
Text + Image Embedding (512-dim)
Multi-Layer Perceptron (MLP) Fusion Network

Figure 1. The pipeline for Four approaches

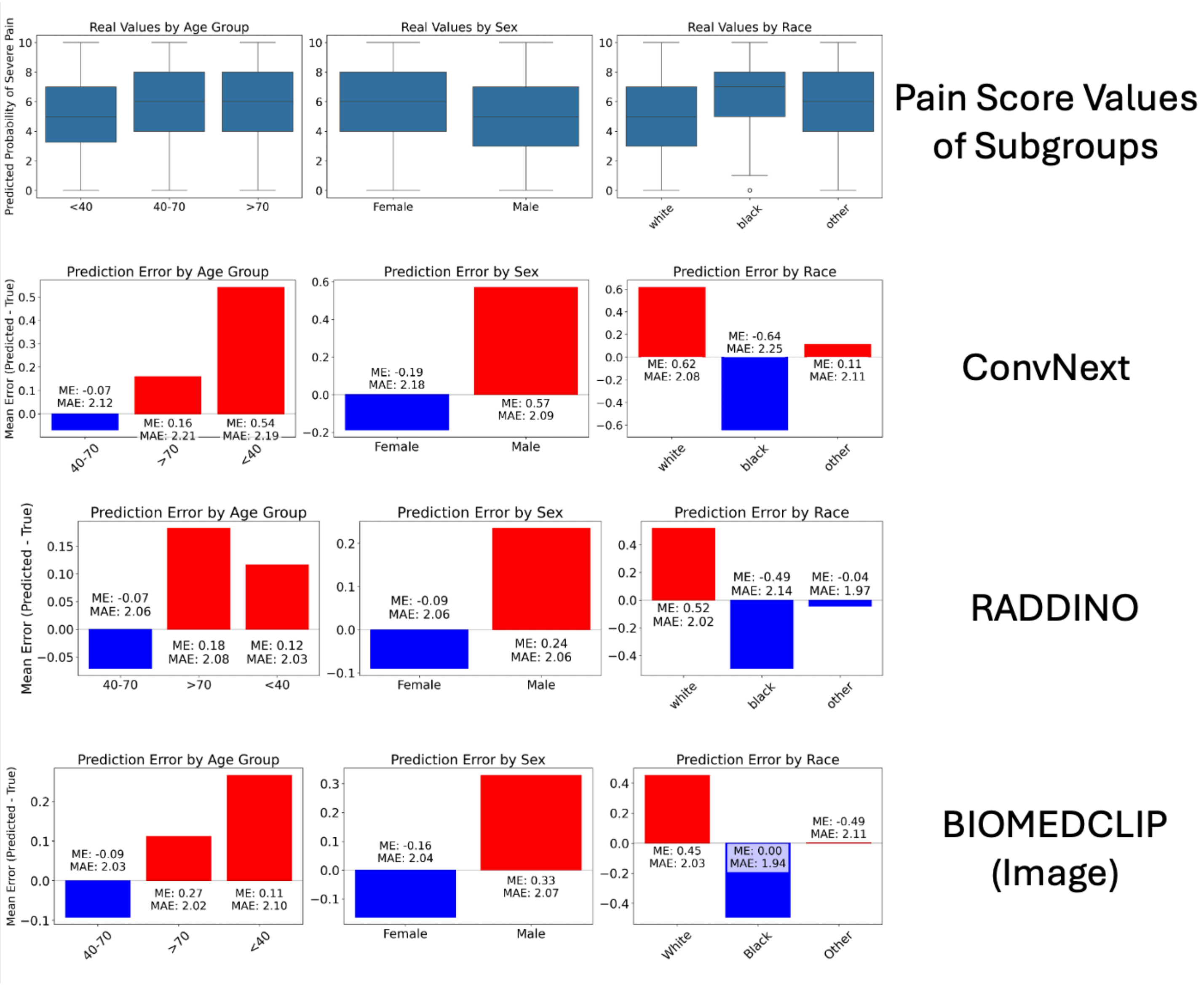

Figure 2. Prediction Error in subgroups

Keywords

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