Evaluating the Value of Longitudinal Hip Radiographs in Self-Supervised Pretraining for Osteoarthritis Classification

Author
Dimana Stoyanova - dstoyanova@tudelft.nl

Supervisors
Jesse Krijthe1, Gijs van Tulder

INTRODUCTION

Background

- Osteoarthritis (OA) progressively degrades joints, causing pain and disability.
- Hip OA is graded on X-rays with the Kellgren–Lawrence (KL) [1] scale, but manual scoring is slow and subjective.
- Deep-learning graders work well but demand large annotated datasets that are hard to obtain [2].
- Self-supervised learning (SSL) cuts the label burden, yet most SSL one use one scan per patient.
- Longitudinal data: some dataset offer serial hip X-rays (baseline, 5 yr, 10 yr) that capture progression.

Research Question

Determine whether integrating longitudinal information during representation learning leads to richer embeddings—and, in turn, to higher downstream KL-grade classification performance—while keeping labeled data requirements low.

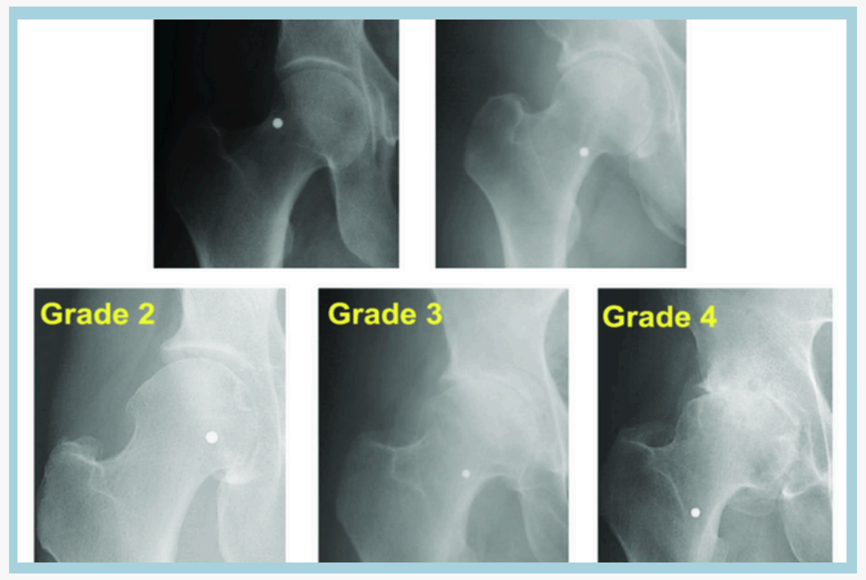

Figure 1:
Source: Jarraya et al., Radiographics (2021)
Reproduced for academic use

METHODOLOGY

- Data assumption: Patients have three visits (baseline, 5y, 10y), and disease progression occurs gradually and detectably across them.
- Contrastive Predictive Coding (CPC) [3] based longitudinal pretraining (see Fig. 2): Two past scans → ResNet-18 encoder → GRU context model predicts the future visit's embedding. InfoNCE loss aligns the prediction with the actual future embedding while repelling unrelated patients.
- CPC forces the model to learn progression-aware representations that reflect meaningful changes in joint structure over time.
- Multi-task pretraining: Combining CPC with SimCLR [4] may yield representations that are both temporally informative and robust to irrelevant variation.
- • Sequential CPC→SimCLR: First learn progression, then refine with per-scan contrast.
- • Interleaved CPC+SimCLR: Alternate CPC and SimCLR in each minibatch to co-train both objectives.

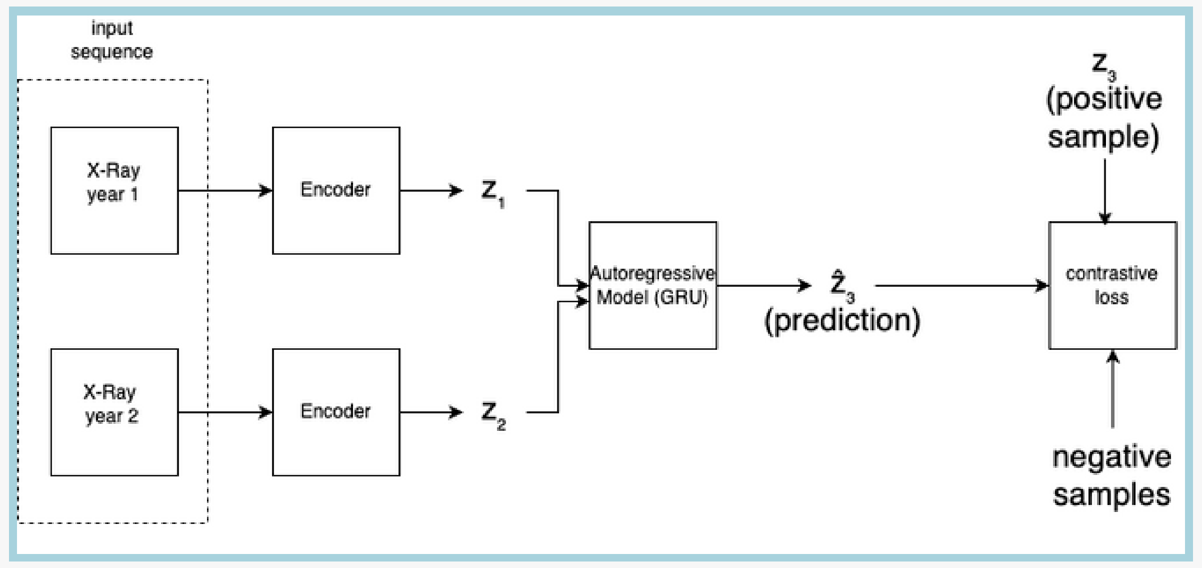

Figure 2: CPC-based self-supervised pretraining.
The model encodes a sequence of hip X-rays (e.g., Year 1 & 2) and uses an autoregressive model (GRU) to predict future representations (e.g., Year 3). A contrastive loss encourages alignment with the true future (z_3) while distinguishing from negatives drawn from other patients.

EXPERIMENTAL SETUP

Pretraining Modes:

1) CPC-only

2) Sequential CPC→SimCLR

3) Static SimCLR – contrastive learning on most recent X-ray

4) Interleaved CPC+SimCLR

5) Patch prediction*-only – predict the rightmost band from image split horizontally in 3 bands

6) Sequential Patch prediction*→SimCLR – pretrain with patch task, then SimCLR

Patch prediction has the same architecture as the one for CPC pretraining, but predicts within a single scan. This controls for multi-task effects without temporal input. Comparing Patch→SimCLR vs CPC→SimCLR helps isolate the effect of temporal modeling.*

Evaluation:

- Attach a classifier head to the frozen encoder
- Binary KL classification (KL < 2 vs. ≥ 2) using theWe use the Osteoarthritis Initiative dataset [5]
- Metric: AUROC over 3 seeds (to account for randomness in split, init, and training order)
- Additional comparison to a fully supervised baseline

RESULTS & DISCUSSION

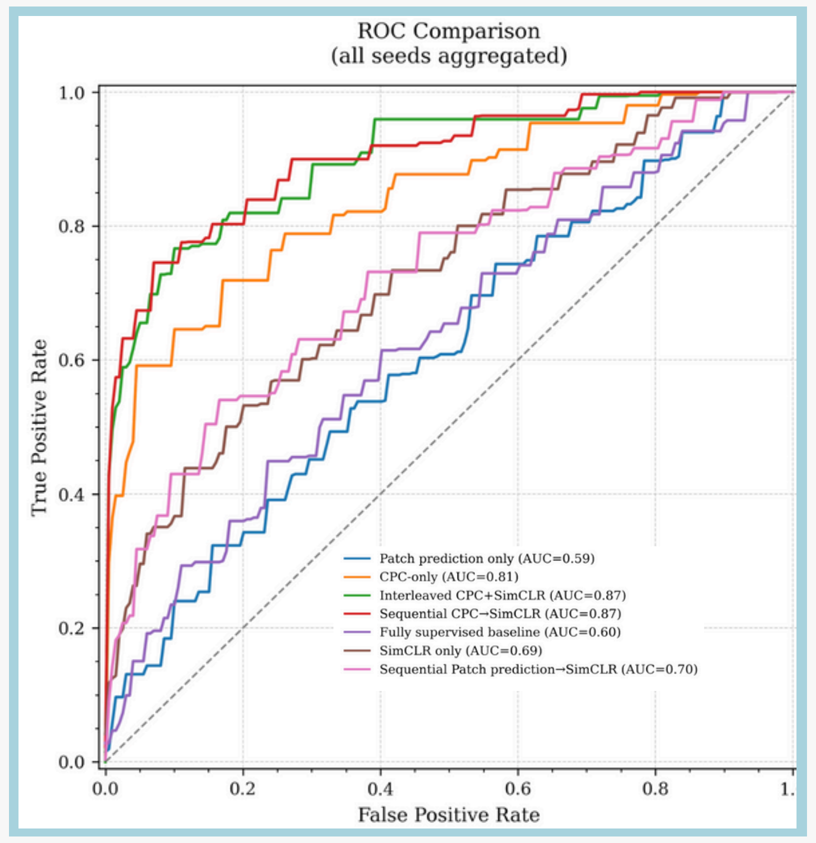

Figure 3: Area-under-the-ROC-curve (AUROC) obtained by each
pre-training strategy, averaged over the three experimental seeds

- Hybrid models combining temporal and static objectives achieved the best performance (AUROC: 0.87)
- CPC slightly outperformed SimCLR, suggesting a modest benefit from longitudinal information alone.
- Patch-based controls underperformed
- In the multitask setting, patch-based pretraining did not improve performance—AUROC remained similar between SimCLR-only and Sequential Patch→SimCLR (0.69 vs 0.70).
- It remains unclear whether gains stem from multitask learning itself or specifically from combining temporally-aware pretraining with SimCLR, as the patch task may be too weak for a fair comparison.

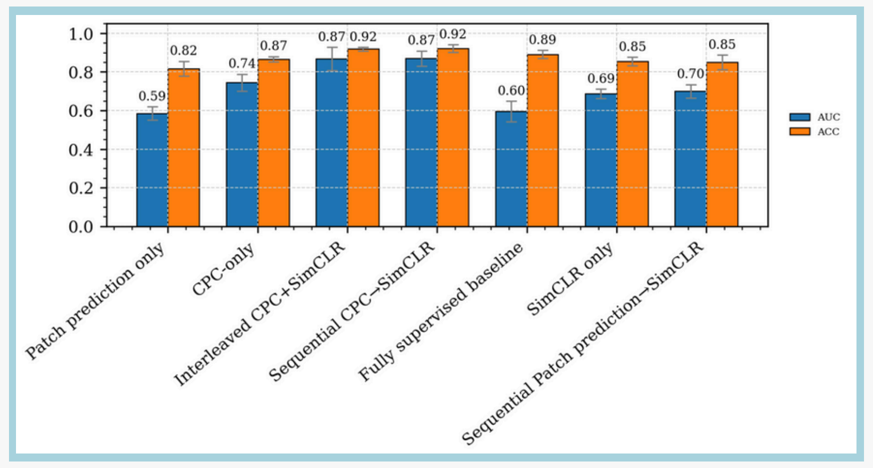

Figure 4: Classification performance (AUROC and Accuracy) for all pretraining strategies.

This study has the following limitations:

- Evaluation was limited to the OAI dataset; generalization to other cohorts remains untested.
- Each patient had only three visits—longer sequences might enhance the value of CPC.
- Patch prediction is a weak control task; stronger pretext tasks are needed for fair multitask comparisons.

CONCLUSION

- Using longitudinal information during SSL improves learned representations for KL-grade classification.
- Multitask pretraining approaches (CPC + SimCLR) exceed the performance of fully supervised performance without requiring labeled data for pretraining.
- Temporal pretraining alone offers modest gains, but performs best when combined with scan-level contrastive learning.
- Results support the value of incorporating progression cues into SSL pipelines for medical imaging.

[1] J. H. Kellgren and J. S. Lawrence, Radiological Assessment of Osteo-Arthrosis, Annals of the Rheumatic Diseases, 1957.

[2] A. Tiulpin, J. Thévenot, E. Rahtu, P. Lehenkari, and S. Saarakkala, Automatic knee osteoarthritis diagnosis from plain radiographs: A deep learning-based approach, IEEE Transactions on Medical Imaging, 2018.

[3] A. van den Oord, Y. Li, and O. Vinyals, Representation Learning with Contrastive Predictive Coding, arXiv preprint arXiv:1807.03748, 2018.

[4] T. Chen, S. Kornblith, M. Norouzi, and G. Hinton, A Simple Framework for Contrastive Learning of Visual Representations, Proceedings of the International Conference on Machine Learning (ICML), pp. 1597–1607, 2020.

[5] Osteoarthritis Initiative, Osteoarthritis Initiative (OAI) Database, 2006. Available at: https://nda.nih.gov/oai/