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

Dimana Stoyanova

Supervisors: Jesse Krijthe, Gijs van Tulder

[1]**EEMCS, Delft University of Technology, The Netherlands**

A Thesis Submitted to EEMCS Faculty Delft University of Technology,
In Partial Fulfilment of the Requirements
For the Bachelor of Computer Science and Engineering
June 22, 2025

Name of the student: Dimana Stoyanova
Final project course: CSE3000 Research Project
Thesis committee: Jesse Krijthe, Gijs van Tulder, Michael Weinmann

An electronic version of this thesis is available at http://repository.tudelft.nl/.

Abstract

Self-supervised learning (SSL) is a promising approach for medical imaging tasks by reducing the need for labeled data, but most existing SSL methods treat each scan as an isolated sample and overlook the fact that patients often have multiple radiographs taken over time. These longitudinal sequences—multiple scans of the same hip acquired at different visits—encode the natural progression of osteoarthritis (OA) and thus could enrich representation learning. In this study, we evaluate whether incorporating temporal information from these longitudinal radiographic sequences into SSL pretraining yields more transferable representations and leads to improved downstream classification of hip OA severity. We focus on a temporal contrastive task (Contrastive Predictive Coding, CPC), which learns to predict future scan representations from earlier ones, and compare it to a SimCLR-based pretraining that treats each radiograph independently. We also investigate a multitask framework that combines both objectives — either by sequentially pretraining with CPC then SimCLR, or by interleaving the two tasks. Experiments on the Osteoarthritis Initiative (OAI) dataset for binary classification of KL-grade severity show that CPC alone does not surpass SimCLR-based pretraining. However, both the sequential and interleaved multitask approaches significantly improve classification accuracy over either single-task method. These findings demonstrate that even though temporal prediction by itself isn't sufficient — combining temporal and within-scan contrastive learning can yield stronger models for hip OA severity assessment.

1 Introduction

Osteoarthritis (OA) is a progressive joint disorder that impacts millions globally, often leading to pain, limited mobility, and decreased quality of life. In clinical practice, its severity is commonly evaluated through radiographic images, such as hip or knee X-rays—using ordinal grading systems such as the Kellgren–Lawrence (KL) scale [9]. However, manual KL grading is known to be subjective, time-consuming, and particularly challenging for early-stage OA, when radiographic indicators such as joint-space narrowing and osteophyte development are only subtly different from healthy anatomy [17].

In an effort to overcome the subjectivity and labor-intensity of manual KL grading, deep-learning techniques have been developed to automate KL-grade classification. Although these methods have demonstrated high predictive accuracy for KL-grade classification on hip X-rays, they depend on the availability of large, accurately annotated datasets, which are costly and difficult to obtain in the medical imaging domain [17].

Self-supervised learning (SSL) offers a promising alternative by learning image representations from unlabeled data. It does so via auxiliary "pretext" tasks—such as contrasting different augmented views of the same X-ray, predicting masked regions, or reconstructing corrupted inputs—which require no manual labels but force the model to discover visual patterns (e.g., anatomical structures, texture variations, and subtle disease markers) that transfer well to downstream objectives and enable fine-tuning with far fewer annotations. [1]. Recent surveys [19, 8] demonstrate that SSL improves both label efficiency and downstream generalization across a range of medical-imaging tasks. However, most existing SSL (and supervised) studies still treat each radiograph in isolation—discarding the temporal relationships in longitudinal series that can reveal gradual cartilage loss, persistent anatomical landmarks, and scanner-specific biases. In the next section, we explain why leveraging these temporal dynamics could improve representation learning for OA severity prediction and state our hypothesis.

Most single-scan SSL methods learn only "snapshot" features—such as local bone texture or joint geometry—without any context for distinguishing pathological progression from benign anatomical or imaging differences. This limitation can cause the model to overlook subtle disease markers like early cartilage thinning or misinterpret scanner-specific artifacts as clinical indicators. We propose that longitudinal sequences could address this gap by providing two complementary signals:

1. **Progression cues**: gradual cartilage loss and joint-space narrowing trends that highlight the structural changes associated with worsening osteoarthritis, and
2. **Invariant signals**: patient-specific bone morphology and consistent imaging characteristics across visits that help the model discount non-disease-related variations.

By pretraining on temporal sequences to predict how a hip changes over successive visits, the encoder discovers a latent progression manifold in feature space—one where movement along specific dimensions corresponds to increasing OA severity. As a result, even when fine-tuned on a single scan from a new patient, the model's representations are already organized by disease severity, making KL-grade classification more accurate. **We therefore hypothesize that using the longitudinal relationship between scans during pretraining helps the model learn more meaningful features. A classifier built on top of these features should achieve better KL-grade accuracy compared to one that uses features learned without any information about the longitudinal relationship between scans of the same patient across different years.**

To test our hypothesis, we adopt Contrastive Predictive Coding (CPC) [18] as our temporal SSL method, which trains an encoder to predict latent representations of future radiographs from earlier ones and thus captures progression cues in a feature space. CPC was chosen because its sequence-prediction objective directly leverages longitudinal data to model disease dynamics. We benchmark CPC against a static contrastive learning baseline inspired by SimCLR [4], which enforces consistency between multiple augmentations of individual scans but ignores temporal context. Beyond these single-task approaches, we introduce two hybrid pretrain-

ing schemes—sequentially applying CPC then static contrastive learning, and interleaving both objectives within each minibatch—to explore whether combining temporal progression modeling with robust snapshot features leads to richer representations. However, this setup leaves open whether any gains come specifically from temporal data or simply from multi-task training, so we add a third hybrid approach that pairs static contrastive learning with a spatial patch-prediction task using the same encoder as CPC but without longitudinal inputs. All SSL pretraining is performed on unlabeled longitudinal hip X-rays, and the resulting encoders are fine-tuned and evaluated on downstream KL-grade classification to assess their relative effectiveness.

The remainder of this paper is organized as follows. In Section 2, we review relevant work on SSL and longitudinal pretext tasks in medical imaging. Section 3 describes our methodology, including CPC, the SimCLR–inspired baseline, and our three hybrid pretraining schemes. In Section 4, we evaluate these approaches on the Osteoarthritis Initiative (OAI) hip X-ray dataset [12] through downstream KL-grade classification under varying label regimes. Section 5 presents our quantitative and qualitative results. In Section 6, we discuss ethical considerations and responsible use of SSL in clinical imaging. Finally, Section 7 summarizes our findings and outlines directions for future work.

Our main contributions are:

1. We compare temporal (CPC) and non-temporal (SimCLR-style) SSL pretraining on the OAI hip X-ray dataset to see which better supports OA grading.
2. We introduce a spatial hybrid control—combining contrastive learning with a patch-prediction task—to show whether improvements come from using multiple tasks or specifically from temporal data.
3. We demonstrate that interleaving temporal and non-temporal objectives during pretraining gives the best KL-grade classification accuracy when only limited labels are available.

2 Related Work

Supervised deep-learning models achieve high accuracy in KL-grade classification but require large annotated datasets [11]. For example, Tiulpin *et al.* [17] trained a multimodal CNN on hundreds of labeled knee X-rays, and Thomas *et al.* [16] evaluated Inceptionv3 and DenseNet backbones for hip OA grading under a similar supervised setup. Chen *et al.* [7] applied style-based manifold extrapolation to capture temporal progression and predict future OA severity, yet it too depends on manually assigned KL labels. This reliance on extensive annotation motivates our exploration of self-supervised approaches that can learn disease dynamics without per-scan labels.

To reduce annotation demands, self-supervised learning (SSL) has been adopted across medical imaging. Large-scale works (e.g., Azizi *et al.* [2]) and surveys (Shurrab *et al.* [14]; Zhang *et al.* [20]; Wang *et al.* [19]) report that contrastive methods (SimCLR [4], BYOL) and reconstruction tasks improve downstream performance with far fewer labels. However, these methods treat each radiograph as a standalone sample, discarding temporal links that as expressed in our aforementioned hypothesis could provide progression cues or stabilize representations against patient-specific anatomy and scanner effects.

Contrastive Predictive Coding (CPC) [18] brings sequence modeling into SSL by predicting future latent states from past ones. CPC has succeeded in modalities ranging from audio to histopathology [15], but its application to longitudinal radiographs—where images of the same joint are taken at multiple visits—has not been systematically evaluated. In osteoarthritis, supervised models have shown that using longitudinal scans can boost prediction [13], yet no work has directly compared CPC-based temporal pretraining to static contrastive SSL on hip X-rays.

Our work fills this gap: we benchmark CPC against a SimCLR-style baseline, and we introduce hybrid schemes to isolate the benefit of temporal supervision from that of multi-task training. This side-by-side evaluation on the OAI hip X-ray cohort clarifies when—and why—leveraging longitudinal sequences enhances self-supervised pretraining for KL-grade classification.

3 Methodology

3.1 Data Assumptions

Our SSL methods assume access to longitudinal hip radiographs, where each patient has multiple anteroposterior X-rays taken at distinct visits (e.g. baseline, mid-follow-up, final follow-up). These scans must share a common view and sufficient resolution for femoral landmark detection, allowing us to form ordered triplets (or longer sequences) for temporal pretext tasks.

3.2 Self-Supervised Pretraining Tasks

We design three self-supervised learning (SSL) strategies to learn representations from unlabeled hip radiographs: a static contrastive task (SimCLR), a temporal predictive task (Contrastive Predictive Coding, CPC), and a multitask configuration combining both. Each strategy is implemented with a dedicated architecture and contrastive loss.

Temporal Contrastive Predictive Coding (CPC)

CPC [18] learns to predict the representation of a future scan from earlier scans of the same patient. Let (x_t, x_{t+1}, x_{t+2}) be three chronologically ordered radiographs. If a patient has four or more scans, we can extend the prediction horizon to $k > 1$, training the model to forecast embeddings multiple steps ahead (e.g. using (x_t, x_{t+1}) to predict (x_{t+2}, x_{t+3}), thereby exploiting richer temporal context when the data permit.

Images are encoded as $z_t = g_{\text{enc}}(x_t)$. A context network g_{ar} summarises past embeddings into $c_t = g_{\text{ar}}(z_t, z_{t+1})$, from which a prediction head ψ estimates the future embedding $\hat{z}_{t+2} = \psi(c_t)$. InfoNCE is applied between $\hat{z}_{t+2}$ and the true z_{t+2}; negatives are embeddings from other time-points and patients. After training, g_{enc} is retained, discarding g_{ar} and ψ.

Static Contrastive Learning (SimCLR Baseline)

As a non-temporal benchmark, we use a SimCLR-style contrastive learning approach that learns from individual radio-

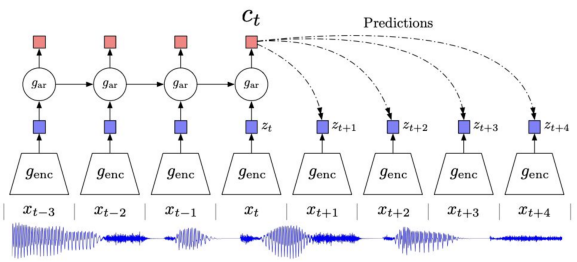

Figure 1: Original CPC architecture illustration reproduced from Van den Oord *et al.* [18]. Although the schematic depicts an audio sequence, the same encoder–context–predictor design is used in our study for longitudinal hip radiographs.

graphs in isolation. During pretraining, we select only the most recent X-ray from each patient—on the assumption that it exhibits the greatest range of OA-related changes—and randomly augment it twice to create a positive pair, while all other images in the batch act as negatives. The objective encourages representations of the same image's augmentations to be similar and those of different images to be distinct, teaching the model to capture invariant features of hip anatomy without any temporal or label information. After this stage, the learned encoder is fine-tuned on the downstream KL-grade classification task using available labels.

Combined / Multitask Setup

To capture both progression dynamics and per-scan invariances in a single model, we hypothesize that jointly training on temporal (CPC) and static (SimCLR) objectives will yield richer representations than either task alone. Our primary hybrid schemes therefore combine CPC and SimCLR in two ways:

- **Interleaved CPC+SimCLR:** alternate CPC and SimCLR losses on successive minibatches, so the encoder continuously learns from both tasks.
- **Sequential CPC→SimCLR:** first pretrain with CPC to instill temporal progression knowledge, then continue with SimCLR to reinforce snapshot consistency.

This design allows us to test whether combining temporal progression modeling with contrastive snapshot learning improves downstream KL-grade classification more than single-task pretraining.

If these hybrids outperform single-task pretraining, it could be due either to temporal information from CPC or simply to the benefit of multi-task learning. To disentangle these effects, we introduce a non-temporal control task that mirrors CPC's architecture and loss but uses only the latest scan per patient. We split each X-ray into three horizontal equal-width bands, encode the first two bands to predict the third via the same context and prediction heads used in CPC, and apply the InfoNCE loss between predicted and true embeddings—matching CPC's complexity without longitudinal data. We then combine this patch-prediction task with SimCLR in a third sequential hybrid:

- **Sequential Patch→SimCLR:** pretrain first on patch prediction, then fine-tune the encoder with the SimCLR contrastive loss.

With this third multitask setup, we are effectively copying our CPC→SimCLR hybrid but replacing the temporal CPC objective with the patch-prediction task. This allows us to match CPC's model complexity and multi-task training regime without any longitudinal data, isolating the unique contribution of temporal supervision. We limit our investigation to sequential multitask training—rather than both interleaved and alternating schemes—because our primary goal is to assess whether adding a second task (temporal or non-temporal) in a straightforward "train-then-train" pipeline influences performance, and sequential scheduling is both simpler to implement and faster to set up

All three multitask schemes share the same ResNet-18 encoder and maintain separate projection heads for each task during pretraining. After pretraining, we discard these heads and retain only the shared encoder for downstream KL-grade classification. By comparing performance across all hybrids and single-task baselines, we can determine whether improvements stem specifically from temporal progression modeling or from the act of combining multiple pretext tasks. This ensures we correctly attribute any gains to longitudinal supervision rather than mere multi-task learning.

3.3 Implementation Details

All models are implemented in PyTorch 2.7 with PyTorch Lightning 2.5, and our full codebase and configuration files are publicly available at https://gitlab.tudelft.nl/osteoarthritis-2025-bsc/contrastive-and-longitudinal-ssl-xray-comparison.

4 Experiments

In this section, we describe our protocol for evaluating how different self-supervised pretraining strategies affect downstream osteoarthritis classification performance. We first detail the data sources, inclusion criteria, and preprocessing steps. Next, we outline each of the six SSL pretraining variants under comparison. We then present our standardized fine-tuning procedure—freezing the encoder and training only a lightweight classifier head—to isolate representation quality. Finally, we explain how we measure reproducibility and quantify run-to-run variability via multiple seeds, describe our ablation configurations, and define our evaluation metrics.

4.1 Data Acquisition and Preparation

We use 4,755 subjects from the Osteoarthritis Initiative (OAI) dataset, with the visit distribution shown in Table 1. Each subject was scheduled for up to three clinical visits: baseline (year 0), year 6, and year 10. Due to missing follow-ups, 2,745 subjects have complete 3-visit data. Only the left hip is used to avoid inter-hip correlation. This ensures each image represents an independent patient-level sample: since a person's two hips share anatomy and disease characteristics, including both could allow the model to learn patient-specific features rather than generalizable osteoarthritis patterns.

Table 1: Distribution of subjects by number of available clinic visits in the OAI dataset.

Number of visits	3 visits	2 visits	1 visit
Subjects	2,745	1,045	965

A 30% subset of patients is used as labeled data and split at the patient level into 70% training, 15% validation, and 15% test. The remaining 70% are used only for self-supervised pretraining. For the CPC setup, we restrict to subjects with three available visits, using the first two timepoints to predict the third. This design mirrors the temporal prediction goal and ensures temporal consistency across sequences.

All radiographs undergo a preprocessing pipeline to ensure consistency: images are normalized in intensity, femoral heads are automatically localized and cropped to a fixed region, and crops are resampled to a uniform pixel spacing. These prepared images are then used directly in both SSL pretraining and downstream evaluation. The preprocessing pipeline can be found at https://gitlab.tudelft.nl/osteoarthritis-2025-bsc/example-preprocessing-code

We follow clinical precedent by casting KL-grade prediction as a binary task—distinguishing mild/absent OA ($\mathrm{KL} < 2$) from moderate/severe OA ($\mathrm{KL} \geq 2$). This mirrors treatment decision thresholds and simplifies evaluation to a single, clinically meaningful decision boundary.

4.2 Self-Supervised Pretraining

We evaluate six SSL pretraining strategies:

- **Static SimCLR:** contrastive learning on each patient's most recent X-ray.
- **CPC-only:** sequence prediction by forecasting the third visit embedding from the first two.
- **Patch-only:** spatial predictive task on the latest scan, predicting the deepest band from the other two.
- **Interleaved CPC+SimCLR:** alternating CPC and SimCLR losses within each minibatch.
- **Sequential CPC→SimCLR:** first pretrain with CPC, then continue with SimCLR.
- **Sequential Patch→SimCLR:** first pretrain on patch prediction, then fine-tune with SimCLR.

All strategies share the same training setup: ResNet-18 encoder, 100 epochs, batch size 64, learning rate 1×10^{-3}, and InfoNCE temperature 0.07. We do not employ early stopping - in line with standard SSL protocols - to maintain equal training budgets and avoid validation-based tuning biases following standard SSL practice [4, 5]. After pretraining, we retain the final-epoch encoder weights for downstream evaluation. These encoders are then frozen and assessed on KL-grade classification using a uniform fine-tuning protocol.

4.3 Fine-Tuning Protocol

To assess the quality of each pretrained encoder, we adopt the standard linear evaluation protocol, in which a classifier is trained on top of a frozen base network and test accuracy serves as a proxy for representation quality [21, 18, 3]. This approach ensures that any differences in downstream performance reflect the effectiveness of the SSL pretraining rather than disparities in fine-tuning capacity.

Our classifier head consists of three linear layers interleaved with ReLU activations and an optional dropout layer. First, the encoder's output vector is projected down to 128 features. A ReLU activation introduces non-linearity, followed by a second linear layer that maps these 128 features to 64, and another ReLU. Finally, a dropout layer (rate tuned via validation) regularizes the 64-dimensional activations before a last linear layer reduces them to a single logit. We omit a softmax or sigmoid activation here because our loss function—PyTorch's `binary_cross_entropy_with_logits`—"combines a Sigmoid layer and the BCELoss in one single class. This version is more numerically stable than using a plain Sigmoid followed by a BCELoss"[1].

Hyperparameters for fine-tuning were determined by a one-time grid search on the SimCLR-pretrained encoder. The best combination—learning rate = 1e-3, weight decay = 0, batch size = 32, dropout = 0.0, cosine learning-rate scheduler, and seed = 0—was then applied uniformly across all pretrained encoders to ensure a fair comparison.

Because moderate-to-severe OA cases ($\mathrm{KL} \geq 2$) are less frequent, we apply minority oversampling during fine-tuning to mitigate class imbalance. High-grade examples are oversampled inversely to improve the learning signal for the positive class. We did not experiment with omitting this oversampling strategy, so its individual impact remains untested.

Throughout fine-tuning, the encoder weights remain frozen and only the classifier head is updated. This design preserves the learned SSL representations and prevents their degradation by overfitting to the limited labeled data. All downstream results are averaged over the three random seeds to account for variability in data splits, initialization, and training order.

To ensure that each classifier is evaluated at its optimal point, we select the checkpoint (epoch) that minimizes validation loss on the held-out validation split. All downstream metrics (AUROC, accuracy, precision, recall) are then computed on the test set using this best-epoch model. This strategy prevents evaluation at arbitrary or late epochs—when overfitting can inflate loss—thereby providing a fair and consistent comparison of representation quality across all pretrained encoders.

4.4 Experimental Variability and Random Seeds

To ensure exact reproducibility and assess variability, we repeat every SSL pretraining and downstream fine-tuning run with three fixed random seeds: 0, 1, and 2. Each seed controls:

- Patient-level splits into the unlabeled pretraining pool and the labeled train/validation/test pools.
- All data-loader operations (shuffling, augmentation order).
- Model weight initialization.

[1] https://pytorch.org/docs/stable/generated/torch.nn.functional.binary_cross_entropy_with_logits.html

All reported metrics are the mean and standard deviation over these three runs.

4.5 Ablation Configurations

To pinpoint the source of any downstream gains, we compare our SSL-pretrained encoders against a fully supervised baseline (no pretraining) and then evaluate six distinct pretraining settings: CPC-only, Patch-only, Static SimCLR, Interleaved CPC+SimCLR, Sequential CPC→SimCLR, Sequential Patch→SimCLR.

Evaluating this suite of models under identical data splits, preprocessing, and fine-tuning protocols allows us to attribute performance differences specifically to temporal dynamics, spatial forecasting, multi-task training, or simply the act of pretraining itself.

4.6 Evaluation Metrics

We report Area Under the ROC Curve (AUROC) and accuracy on the test set. AUROC is threshold-independent and robust to class imbalance [6]. Accuracy is reported for completeness.

5 Results

We report performance on the test set for each pretraining configuration in terms of AUROC and accuracy.

Figure 2 summarises the discriminative power of each model in terms of area under the ROC curve (AUROC). The *Interleaved CPC + SimCLR* and *Sequential CPC → SimCLR* hybrids achieve the highest AUROC scores (**0.87** each), followed by the *CPC-only* model (**0.74**). The best non-temporal alternatives trail behind: *Sequential Patch → SimCLR* attains **0.70**, and *SimCLR-only* reaches **0.69**. The fully supervised baseline manages **0.60**, while the *Patch-only* ablation records the lowest AUROC at **0.59**.

Because moderate–to–severe OA examples (KL 2) form the minority of the cohort, a naïve majority-class predictor would already attain an accuracy of 0.72. Overall accuracy can therefore overstate performance; throughout this section we regard AUROC as the primary metric and use accuracy only as a secondary, easily interpretable figure.

In terms of raw classification accuracy in Figure 3, the two hybrid models again lead with a mean accuracy of **0.92**, clearly outperforming the best single-task temporal model (*CPC-only*, **0.87**) and the supervised baseline (**0.89**). Purely spatial objectives fare worse: *SimCLR-only* reaches **0.85**, and adding patch prediction does not improve it (*Sequential Patch → SimCLR*: **0.85**). The *Patch-only* configuration lags behind at **0.82**.

Key observations.

- **Temporal information helps.** Even the *CPC-only* model exceeds the best non-temporal variant by 0.04 AUROC points (0.74 vs. 0.70), and combining CPC with SimCLR raises that gap to 0.17 (0.87 vs. 0.70).
- **Hybrid training is beneficial.** Interleaving or sequentially combining CPC with SimCLR lifts AUROC from 0.74 (CPC-only) to 0.87 and raises accuracy from 0.87 to 0.92.

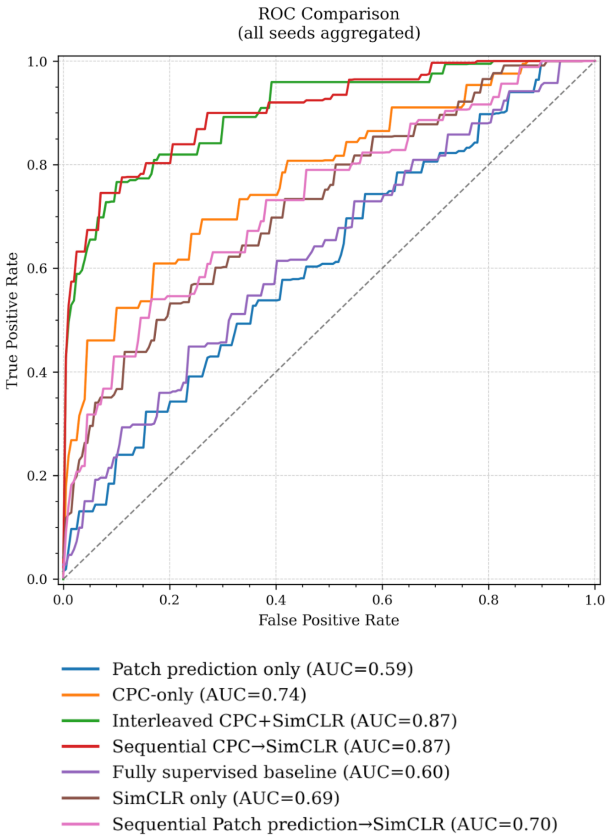

Figure 2: Area-under-the-ROC-curve (AUROC) obtained by each pre-training strategy, averaged over the three experimental seeds.

- **Patch prediction underperforms.** Despite matching CPC in model capacity, the *Patch-only* objective delivers the weakest performance (AUROC 0.59, ACC 0.82), suggesting that this pretext task fails to encourage the encoder to learn clinically useful features. A stronger non-temporal alternative—such as masked-image modelling—would likely provide a fairer comparison to CPC.

- **Adding patch prediction before SimCLR yields no benefit.** The sequential *Patch prediction → SimCLR* hybrid (AUROC 0.70, ACC 0.85) is virtually indistinguishable from *SimCLR-only* (AUROC 0.69, ACC 0.85).

As an additional experiment, to further validate the discriminative power of our multitask-pretrained encoder, we also examined the learned feature space by applying UMAP (Uniform Manifold Approximation and Projection [10]) to the test-set embeddings produced by our sequentially pretrained (CPC→SimCLR) encoder. As shown in Figure 4, the positive (red) and negative (blue) OA cases form two well-separated clusters, confirming that the multitask pretraining yields representations in which disease status is readily distinguishable in low dimensions.

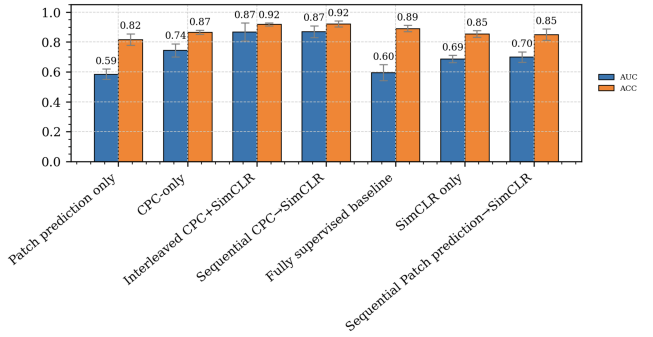

Figure 3: Overall classification accuracy (ACC) for the same models as in Fig. 2.

6 Responsible Research

This section discusses how ethical and reproducible research practices were applied throughout this classification study based on medical imaging data.

6.1 Ethical Considerations

The dataset used in this study comes from the publicly available Osteoarthritis Initiative (OAI) [12], which includes longitudinal X-ray images and associated clinical data. Although the data are de-identified and released for research purposes under specific terms, they still represent sensitive medical information. Therefore, we adhered strictly to the usage policies provided by the National Institute on Aging and the National Institutes of Health, ensuring that no attempt was made to re-identify individuals or share the data outside of approved research settings. Only aggregated, non-identifiable results and model performance metrics are reported in this paper.

When using deep learning in healthcare-related contexts, one must also consider the risks associated with biased learning and opaque model behavior. Our models were implemented using standard PyTorch and PyTorch Lightning components, including `pl.LightningModule`, `nn.Conv2d`, `nn.Module`, `nn.GRU`, and `nn.Linear`. While fairness or subgroup performance analysis was not conducted at this stage, we recognize this as a critical direction for future work. Furthermore, given the complexity of the models, interpretability remains a concern. As of now, the model is not explainable enough for clinical deployment.

Importantly, the models developed in this project are experimental and should not be used for clinical diagnosis or decision-making. Their outputs can only support expert judgment and are not intended to replace medical professionals.

6.2 Reproducibility of Methods

Reproducibility is a key aspect of trustworthy machine learning research. To support replicability, the report provides a detailed description of the experimental setup, including the dataset structure, preprocessing steps, training objectives, and evaluation procedures.

The codebase relies on widely used open-source frameworks and will be made publicly available in a version-controlled repository which can be found at the following link: https://gitlab.tudelft.nl/osteoarthritis-2025-bsc/contrastive-and-longitudinal-ssl-xray-comparison. This will allow other researchers to reproduce, verify, and build upon our work.

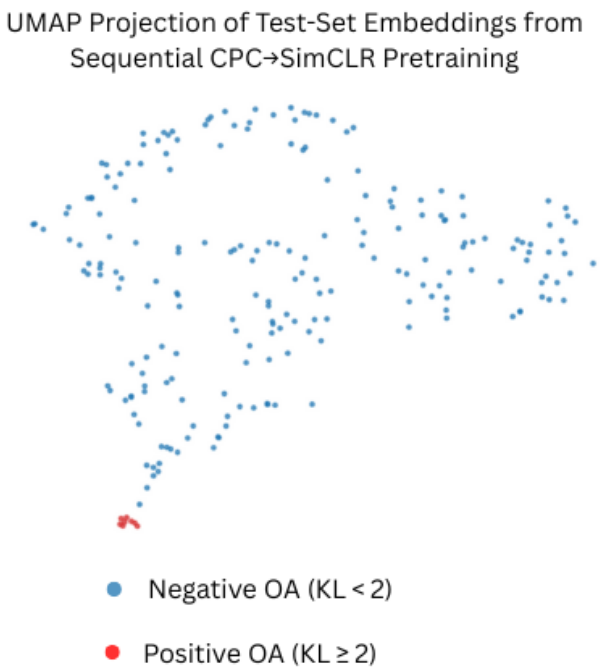

Figure 4: UMAP (Uniform Manifold Approximation and Projection) of test-set embeddings from the encoder pretrained sequentially with CPC then SimCLR. Red and blue points denote positive and negative OA cases, showing clear class separation.

7 Discussion and Conclusion

We began with the idea that giving the model several scans of the same hip taken in different years—instead of considering only a single image per patient—would help it learn richer features for classifying osteoarthritis (OA). The results support that view, but only partly. A model pretrained purely with a temporal prediction (CPC) task is a little better than the best single-image baseline (SimCLR: AUROC 0.69 $\rightarrow$ 0.74). Clear progress shows up only when the two ideas are combined: the multitask approaches, interleaved or sequential, reach AUROC 0.87 and ACC 0.92, well ahead of any single-task model and the fully supervised baseline.

Why does this mix help? Each pretext task supplies the encoder what the other cannot. CPC captures progression cues—gradual cartilage loss and joint-space narrowing that mark worsening OA over time—while SimCLR focuses on what stays the same in one scan and learns to ignore lighting, view angle, and scanner noise. CPC alone is still sensitive to those image quirks, and SimCLR alone knows nothing about change over time. Using both objectives together guides the model toward patterns that truly reflect OA severity while dis-

carding distracting variation, leading to the large gain in performance.

Our control experiment with patch prediction shows that simply adding a second task is not enough. Patch-only scores poorly (AUROC 0.59) and, even when paired with SimCLR, adds almost no benefit (0.70). Because the patch task is so weak, it does not let us fully separate the value of "any" multitask training from the value of temporal information. A stronger non-temporal control will be needed to tease apart those contributions more convincingly.

Many hospitals store years of hip X-rays but have few expert labels. A single self-supervised pretraining run that blends CPC and SimCLR can turn that unlabelled archive into a strong encoder—better than a supervised model trained from scratch.

This study has limits. First, all experiments used only the Osteoarthritis Initiative cohort; performance on other datasets is still unknown. Second, each patient has only three visits; longer timelines might let CPC do even better without SimCLR. Third, patch prediction is a very simple stand-in for "another task"; stronger pretext tasks could give a fairer comparison.

Next steps are to repeat the study on independent datasets, try other pretext tasks alongside SimCLR for the multitask pretraining, and include longer follow-up series.

In short, using the longitudinal relationship between scans during pretraining helps the model, but it really pays off only when it is combined with a another task - in our case the SimCLR one. This multitask approach clearly beats single-task SSL and a supervised baseline, supporting our original idea that temporal information helps and shows that combining temporal and within-scan contrastive learning can yield stronger models for hip OA severity assessment.

A Appendix One: Use of Large Language Models (LLMs)

Large Language Models (LLMs) were used to support the writing process of this report in a limited and responsible manner. The tool was used mainly for language refinement — to help improve the clarity and scientific tone of passages originally drafted by the authors. The content and structure of the report, including all sections were developed entirely by the authors.

Specifically, we used LLMs to rephrase and improve certain sentences or paragraphs that were already written.

For example, when drafting the abstract, we initially wanted to phrase the idea that *the method uses multiple years of scans of the same patient to capture disease progression*, but weren't satisfied with how it read.

We prompted the LLM with:

> *Paraphrase this in a writing style that is more proper for a scientific paper and so it is easier to understand: "representation learning utilising multiple years of a scan of the same patient that express disease progression."*

The LLM (ChatGPT) responded with:

> *"This study investigates how to leverage temporal information from radiographic scan sequences of the same patient—capturing disease progression over time—for improved SSL."*

This output was reviewed and manually edited before inclusion in the final version.

No content was generated autonomously without author review and/or edit. All use of LLMs was limited to surface-level editing and phrasing fixes. The scientific content, structure, analysis, and conclusions of the report reflect the independent work of the authors.

References

[1] Saleh Albelwi. Survey on self-supervised learning: Auxiliary pretext tasks and contrastive learning methods in imaging. *Entropy*, 24(4):551, 2022.

[2] Shekoofeh Azizi, Basil Mustafa, Fiona Ryan, Zachary Beaver, Jan Freyberg, Jonathan Deaton, Aaron Loh, Alan Karthikesalingam, Simon Kornblith, Ting Chen, Vivek Natarajan, and Mohammad Norouzi. Big self-supervised models advance medical image classification. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 3478–3488, 2021.

[3] Philip Bachman, R. Devon Hjelm, and William Buchwalter. Learning representations by maximizing mutual information across views. *arXiv preprint arXiv:1906.00910*, 2019.

[4] Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. A simple framework for contrastive learning of visual representations. In *Proc. International Conference on Machine Learning (ICML)*, pages 1597–1607, 2020.

[5] Xinlei Chen, Haoqi Fan, Ross Girshick, and Kaiming He. Improved baselines with momentum contrastive learning. 2020.

[6] Jesse Davis and Mark Goadrich. The relationship between precision–recall and roc curves. pages 233–240, 2006.

[7] Tianyu Han, Jakob Nikolas Kather, Federico Pedersoli, Markus Zimmermann, Sebastian Keil, Maximilian Schulze-Hagen, Marc Terwoelbeck, Peter Isfort, Christoph Haarburger, Fabian Kiessling, Volkmar Schulz, Christiane Kuhl, Sven Nebelung, and Daniel Truhn. Image prediction of disease progression for osteoarthritis by style-based manifold extrapolation. *Nature Machine Intelligence*, 4(11):1–11, 2022.

[8] Shih-Cheng Huang, Anuj Pareek, Malte Jensen, Matthew P. Lungren, Serena Yeung, and Akshay S. Chaudhari. Self-supervised learning for medical image classification: a systematic review and implementation guidelines. *NPJ Digital Medicine*, 87:102846, 2023.

[9] J. H. Kellgren and J. S. Lawrence. Radiological assessment of osteo-arthrosis. *Annals of the Rheumatic Diseases*, 16(4):494–502, 1957.

[10] Leland McInnes, John Healy, and James Melville. Umap: Uniform manifold approximation and projection for dimension reduction. *arXiv preprint arXiv:1802.03426*, 2018.

[11] Huy Hoang Nguyen, Simo Saarakkala, Matthew B. Blaschko, and Aleksei Tiulpin. Semixup: In- and out-of-manifold regularization for deep semi-supervised knee osteoarthritis severity grading from plain radiographs. *IEEE Transactions on Medical Imaging*, 39(12):4346–4356, December 2020.

[12] Osteoarthritis Initiative. Osteoarthritis initiative (oai) database. https://nda.nih.gov/oai/, 2006. Accessed: June 10, 2025.

[13] Jean-Baptiste Schiratti, Rémy Dubois, Paul Herent, David Cahané, Jocelyn Dachary, Thomas Clozel, Gilles Wainrib, Florence Keime-Guibert, Agnes Lalande, Maria Pueyo, Romain Guillier, Christine Gabarroca, and Philippe Moingeon. A deep learning method for predicting knee osteoarthritis radiographic progression from mri. *Arthritis Research Therapy*, 23(1):262, 2021.

[14] Saeed Shurrab and Rehab Duwairi. Self-supervised learning methods and applications in medical imaging analysis: A survey. *PeerJ Computer Science*, 8:e1045, 2022.

[15] Karin Stacke, Claes Lundström, Jonas Unger, and Gabriel Eilertsen. Evaluation of contrastive predictive coding for histopathology applications. 136:328–340, 2020.

[16] Kevin A. Thomas, Łukasz Kidziński, Eni Halilaj, Scott L. Fleming, Guhan R. Venkataraman, Edwin H. G. Oei, Garry E. Gold, and Scott L. Delp. Automated classification of radiographic knee osteoarthritis severity using deep neural networks. *Radiology: Artificial Intelligence*, 2(2):e190065, 2020.

[17] Aleksei Tiulpin, Jérôme Thevenot, Esa Rahtu, Petri Lehenkari, and Simo Saarakkala. Automatic knee osteoarthritis diagnosis from plain radiographs: A deep learning-based approach. *Scientific Reports*, 8(1):1727, 2018.

[18] Aaron van den Oord, Yazhe Li, and Oriol Vinyals. Representation learning with contrastive predictive coding. *arXiv preprint arXiv:1807.03748*, 2018.

[19] Wei-Chien Wang, Euijoon Ahn, Dagan Feng, and Jinman Kim. A review of predictive and contrastive self-supervised learning for medical images. *Machine Intelligence Research*, 20(4):483–513, 2023.

[20] Chuyan Zhang and Yun Gu. Dive into self-supervised learning for medical image analysis: Data, models and tasks. *arXiv preprint arXiv:2209.12157*, 2022. v2, last revised 17 Apr 2023.

[21] Richard Zhang, Phillip Isola, and Alexei A. Efros. Colorful image colorization. 2016.