Reeve Gonsalves

Progress Report – Case Study (Phase 2)

Project Title:

AI-Driven Detection and Grading of Knee Osteoarthritis in Radiographic Data

Date: 8th February 2026

AI-Driven Detection and Grading of Knee Osteoarthritis from X-ray Images

---

1. Introduction

Following the completion of the expose and the first progress report, Phase 2 of this case study focuses on the transition from planning and dataset investigation to practical implementation and experimentation. The primary objective of this phase was to establish a working deep learning pipeline for knee osteoarthritis (KOA) grading using radiographic images and to validate the feasibility of the proposed approach through exploratory analysis and a baseline model.

This phase does not aim to achieve optimal performance. Instead, it is intended to (i) finalize a suitable dataset, (ii) understand its characteristics through exploratory data analysis, and (iii) implement and evaluate an initial baseline convolutional neural network using transfer learning.

---

2. Dataset Selection and Finalization

2.1 Dataset Access Considerations

In the initial project plan, controlled-access datasets such as the Osteoarthritis Initiative (OAI) and MOST were identified as primary and secondary candidates. However, during Phase 2, technical and administrative constraints associated with NDA-based and approval-dependent datasets made it impractical to rely on them within the project timeline.

To ensure steady progress and full reproducibility, a fully public knee X-ray dataset with Kellgren–Lawrence (KL) severity grading was selected for implementation and experimentation in this phase.

2.2 Final Dataset Used

The dataset used in Phase 2 is a **public knee osteoarthritis X-ray dataset with KL severity grading (0–4)**, hosted on Kaggle. The dataset provides radiographic knee images organized into predefined training, validation, and test splits, with KL grades encoded directly via folder structure.

The dataset follows this structure:

- train/0–4

- val/0–4
- test/0–4
- (optional auto_test/ folder)

This organization enables direct compatibility with standard deep learning pipelines (e.g., PyTorch ImageFolder) and avoids manual label processing.

2.3 Dataset Characteristics

An initial inspection revealed a **strong class imbalance**, with early-stage or non-osteoarthritic cases (KL 0–2) being significantly more frequent than severe osteoarthritis cases (KL 4). This imbalance reflects real-world clinical distributions but poses challenges for supervised learning and evaluation, particularly for minority classes.

---

3. Exploratory Data Analysis (EDA)

3.1 Class Distribution Analysis

Exploratory analysis was performed to quantify the distribution of samples across KL grades in the training, validation, and test sets. The results confirmed a consistent imbalance across all splits, with KL 4 being the least represented class.

This observation is important for model design and evaluation, as overall accuracy alone may not adequately reflect performance on underrepresented severity grades. Consequently, macro-averaged metrics were considered in later evaluation.

3.2 Visual Inspection of Radiographs

Representative X-ray samples from multiple KL grades were visualized to qualitatively assess image quality and label consistency. Clear radiographic differences were observed across severity levels, including joint space narrowing, osteophyte formation, and structural deformities in higher KL grades.

Visual overlap between adjacent grades (e.g., KL 1 vs. KL 2, KL 3 vs. KL 4) was also evident, highlighting the inherent difficulty of fine-grained osteoarthritis grading from 2D radiographs. This qualitative inspection confirmed that the dataset is clinically meaningful and suitable for supervised deep learning analysis.

---

4. Baseline Model Implementation

4.1 Model Architecture

A baseline convolutional neural network was implemented using **ResNet-18**, a widely adopted architecture in medical image analysis. Transfer learning was applied by initializing the model

with ImageNet-pretrained weights, allowing the network to leverage previously learned generic visual features.

The final fully connected layer was replaced to output predictions for **five classes**, corresponding to KL grades 0–4.

4.2 Training Configuration

The baseline experiment was configured as follows:

- Framework: PyTorch
- Input image size: 224×224
- Loss function: Cross-entropy loss with **class weighting** to mitigate class imbalance
- Optimizer: Adam
- Learning rate: 1e-4
- Training epochs: 5
- Evaluation metrics: Accuracy and macro-averaged F1-score

Training was performed on local hardware with limited computational resources. Therefore, the number of epochs was intentionally kept small, as the purpose of this phase was to establish a reference baseline rather than to optimize performance.

---

5. Training and Validation Results

During training, the model demonstrated stable learning behaviour, with decreasing loss values and improving performance metrics across epochs. Validation accuracy and macro F1-score improved gradually and remained stable, indicating that the model was learning meaningful patterns without severe overfitting.

After five training epochs, validation performance reached approximately:

- Validation accuracy: ~57%
- Validation macro F1-score: ~0.60

Given the multi-class nature of the task, strong class imbalance, and limited training duration, these results are considered reasonable for a first baseline model.

---

6. Test Set Evaluation

To obtain an unbiased estimate of performance, the trained baseline model was evaluated on the held-out test set. The following results were obtained:

- **Test accuracy:** 59.8%

- **Test macro F1-score:** 0.629

A confusion matrix analysis revealed that most misclassifications occurred between adjacent KL grades, particularly between KL 1 and KL 2, as well as KL 3 and KL 4. This behaviour is expected due to subtle visual differences between neighbouring severity stages and is consistent with observations reported in the literature.

---

7. Discussion and Observations

The results of Phase 2 confirm that the proposed deep learning pipeline is technically viable and capable of learning clinically relevant features from knee X-ray images. The baseline model establishes a meaningful reference point for further refinement.

Key challenges identified during this phase include:

- Strong class imbalance, especially for severe osteoarthritis (KL 4)
- Visual overlap between adjacent KL grades
- Limited training duration due to computational constraints

These challenges highlight the importance of improved imbalance handling, more advanced architectures, and interpretability analysis in subsequent phases.

---

8. Planned Next Steps (Phase 3)

Based on the outcomes of Phase 2, the next phase of the project will focus on:

- Implementing Grad-CAM to visualize model attention and support interpretability
- Exploring stronger backbone architectures (e.g., DenseNet)
- Investigating improved imbalance mitigation strategies
- Conducting more detailed per-class performance analysis

---

9. Conclusion

Phase 2 successfully achieved its objectives by finalizing a suitable dataset, conducting exploratory data analysis, and implementing a baseline transfer learning model for knee osteoarthritis grading. The obtained results demonstrate meaningful learning and provide a solid foundation for further methodological improvements and interpretability analysis in the next phase of the project.