1. Introduction

The objective of this phase was to establish a reproducible baseline for automated Knee Osteoarthritis (KOA) grading using deep learning. The baseline model serves as a reference point for subsequent replication and improvement experiments.

This phase focuses on:

- Dataset preparation and preprocessing
- Class distribution analysis
- Implementation of a transfer learning–based CNN (ResNet18)
- Evaluation using clinically relevant metrics

---

2. Dataset Description

2.1 Dataset Source

Dataset: **Knee Osteoarthritis Dataset with Severity Grading**
Source: Public Kaggle repository

The dataset contains knee radiographs labelled according to the **Kellgren–Lawrence (KL) grading system**, with severity levels:

- KL 0 — No osteoarthritis
- KL 1 — Doubtful OA
- KL 2 — Mild OA
- KL 3 — Moderate OA
- KL 4 — Severe OA

2.2 Dataset Structure

The dataset is organized into:

train/

val/

test/

Total samples:

- Training: 5778 images
- Validation: 826 images

- Test: 1656 images

Total classes: 5

2.3 Class Distribution (Imbalance)

Training class counts:

- KL 0: 2286
- KL 1: 1046
- KL 2: 1516
- KL 3: 757
- KL 4: 173

Observation:

- Severe OA (KL 4) is underrepresented.
- Dataset exhibits moderate class imbalance.
- Adjacent KL grades have overlapping visual patterns.

To address imbalance, a **WeightedRandomSampler** was employed during training.

---

3. Preprocessing Pipeline

All images were resized to:

224 × 224 pixels

Preprocessing steps:

- Resize to fixed resolution
- Random horizontal flipping
- Mild rotation augmentation
- Normalization using ImageNet mean and standard deviation

Normalization parameters:

mean = [0.485, 0.456, 0.406]

std = [0.229, 0.224, 0.225]

These match the pretraining statistics of ImageNet.

---

4. Baseline Model Architecture

4.1 Network

Model: **ResNet18**

- Pretrained on ImageNet
- Final fully connected layer replaced with 5-class classifier
- Transfer learning applied

Two-stage training strategy:

1. Train classifier head only
2. Fine-tune deeper layers

4.2 Training Configuration

- Optimizer: Adam
- Learning rate (initial): 1e-3
- Weight decay: 1e-4
- Loss function: CrossEntropyLoss
- Scheduler: ReduceLROnPlateau
- Total epochs: 18

Hardware:

- GPU: Tesla T4

---

5. Evaluation Metrics

Evaluation was performed on the held-out test set using:

- Accuracy
- Precision
- Recall
- Macro-F1 score
- Confusion matrix

Macro-F1 was emphasized due to class imbalance.

---

6. Baseline Results (ResNet18 @224)

Test Performance:

- Test Accuracy ≈ **0.59**
- Macro-F1 ≈ **0.62**

Observations:

- Strong performance on KL 0 and KL 4
- Frequent confusion between adjacent grades (KL 1 vs 2, KL 2 vs 3)
- Class imbalance impacts minority class performance

The confusion matrix revealed that most misclassifications occur between neighboring KL grades, consistent with known difficulty in fine-grained radiographic grading.

---

7. Baseline Analysis

The ResNet18 baseline confirms:

- Transfer learning is effective for KOA grading.
- The dataset is suitable for supervised deep learning.
- Adjacent KL grades remain challenging.
- Class imbalance affects minority class recall.

This baseline establishes a quantitative reference for subsequent replication experiments using more advanced architectures.

---

8. Conclusion of Baseline Phase

The ResNet18 baseline provides:

- A reproducible training pipeline
- A measurable performance benchmark
- Insight into dataset imbalance
- Foundation for literature-based replication

This model is not intended as the final system but serves as a structured starting point for further experimentation.