Grad-CAM Interpretability Analysis

1. Objective

To ensure that the trained DenseNet121 model makes clinically meaningful decisions rather than relying on irrelevant image artifacts, Gradient-weighted Class Activation Mapping (Grad-CAM) was applied to visualize the regions contributing to model predictions.

Grad-CAM provides spatial heatmaps indicating which regions of the X-ray image most strongly influenced the classification decision.

This interpretability step is critical for medical AI systems to:

- Improve clinician trust
- Validate anatomical focus
- Analyse misclassification behaviour
- Support model transparency

---

2. Experimental Setup

- Model: DenseNet121 (ImageNet pre-trained)
- Input resolution: 320 × 320
- Best validation macro-F1: 0.6659
- Test macro-F1: 0.6623
- Target layer: Final convolutional feature block
- Dataset: Knee Osteoarthritis KL grading (5 classes)

Grad-CAM heatmaps were generated for:

- Correct predictions
- Incorrect predictions
- Predicted class CAM
- True class CAM (for error analysis)

A total of 12 representative examples were selected.

---

3. Overall Visualization Summary

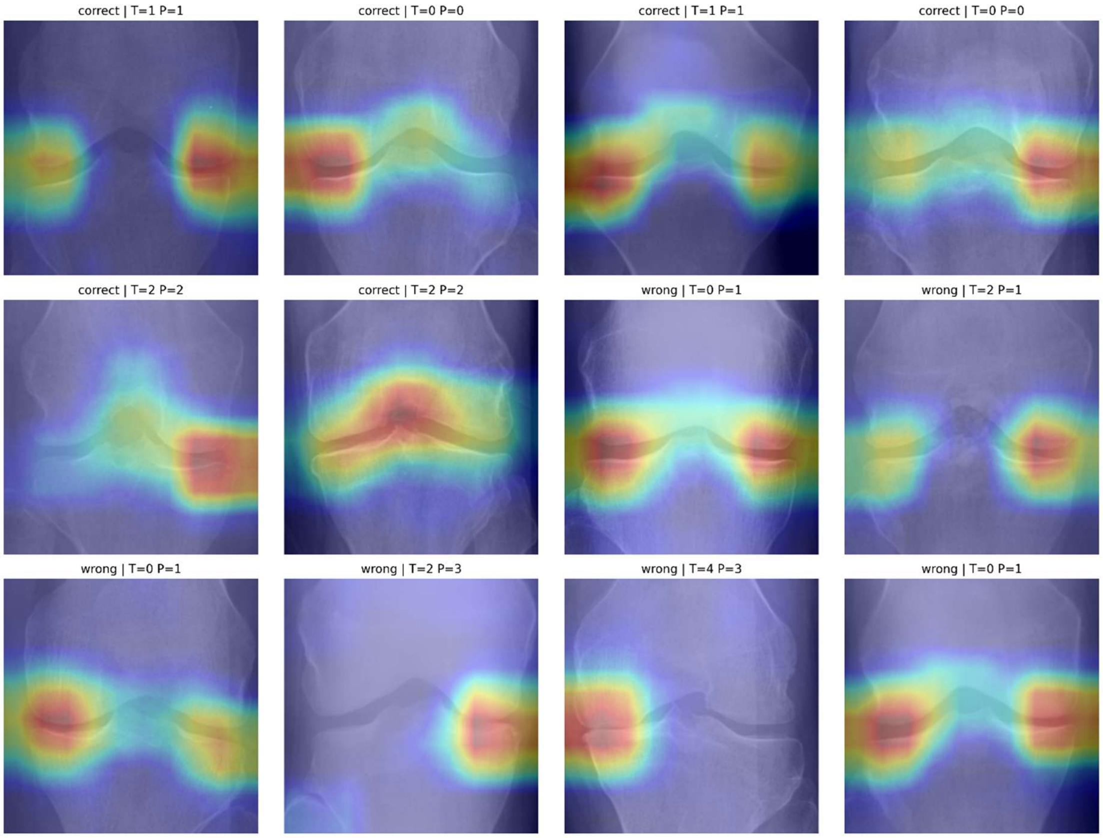


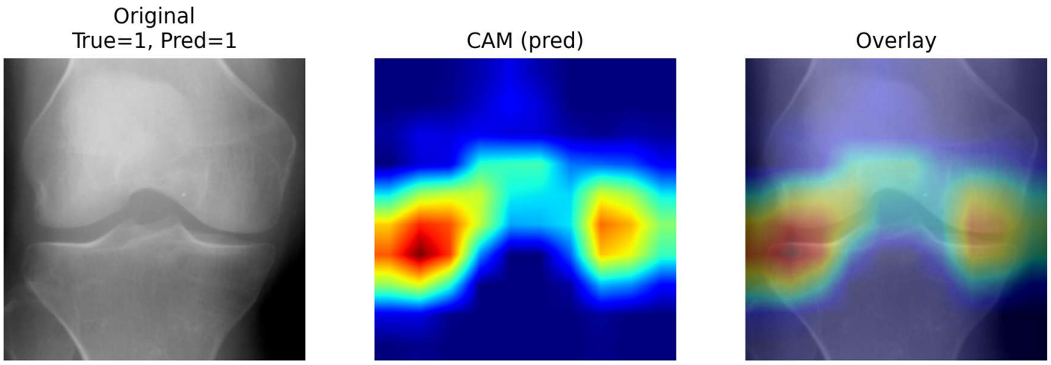
Figure 1 shows a grid of correct and incorrect predictions with their corresponding Grad-CAM overlays.

Observation:

- The model consistently focuses on the tibiofemoral joint space.
- Activation maps are centered around:
    - Medial and lateral joint compartments
    - Regions showing joint space narrowing
    - Osteophyte-prone margins

This indicates that the model is learning clinically relevant radiographic features rather than background noise.

4. Correct Prediction Example



Example: True = 1, Predicted = 1

Observation:

- Heatmap is concentrated along the joint line.
- Strong activation near narrowing regions.
- No activation on irrelevant areas (background, soft tissue).

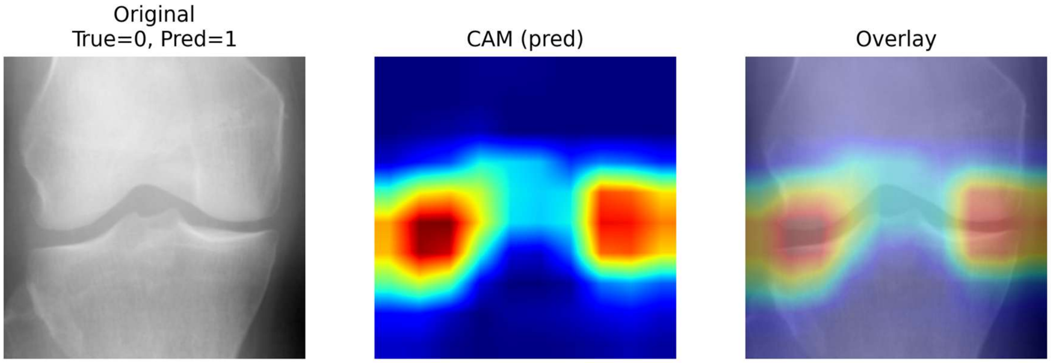
Interpretation:

The model correctly identifies early degenerative features and bases its decision on anatomically meaningful regions.

This supports the clinical validity of the learned representation.

---

5. Incorrect Prediction Example (Predicted Class CAM)



Example: True = 0, Predicted = 1

Observation:

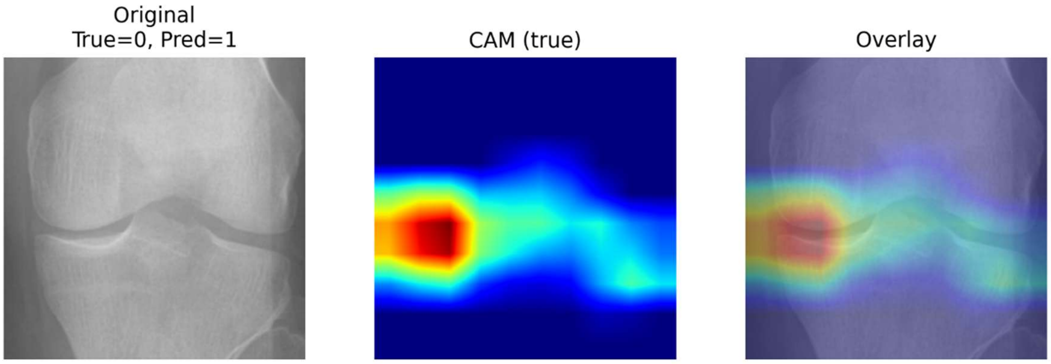
- Activation is still concentrated at the joint space.
- Model highlights subtle joint irregularities.
- Suggests confusion between KL0 and KL1.

Interpretation:

The error is not random.
The model appears to interpret minimal structural changes as early osteoarthritis.

This reflects the inherent difficulty in distinguishing adjacent KL grades.

6. Incorrect Prediction Example (True Class CAM)



Example: True = 0, Predicted = 1

True-class Grad-CAM was computed for comparison.

Observation:

- True-class CAM shows weaker activation.
- Predicted-class CAM shows stronger joint emphasis.

Interpretation:

The model’s internal feature representation is more aligned with mild OA characteristics than completely healthy anatomy.

This indicates class boundary ambiguity rather than pathological reasoning failure.

7. Key Findings

1. The model consistently attends to clinically meaningful anatomical regions.
2. Focus is primarily on tibiofemoral joint space.
3. No evidence of reliance on image corners, borders, or artifacts.
4. Misclassifications mainly occur between adjacent KL grades.
5. Errors are due to subtle structural overlap rather than incorrect region focus.

8. Clinical Relevance

The Grad-CAM visualizations demonstrate that:

- The model behaviour aligns with radiological reasoning.
- Joint space narrowing and marginal bone regions drive predictions.
- Interpretability confirms reliability of the learned features.

This strengthens the credibility of the model for automated KL grading tasks.

9. Conclusion

The Grad-CAM analysis confirms that the DenseNet121 model:

- Learns anatomically meaningful features
- Focuses on clinically relevant joint regions
- Demonstrates interpretable decision-making patterns

Although classification errors exist, they stem from the intrinsic challenge of grading adjacent KL stages rather than spurious feature reliance.

This interpretability validation supports the deployment potential of the model in computer-assisted knee osteoarthritis assessment systems.