Evaluating Open-Source AI Models for Bone Fracture Detection Using the MURA Dataset: A Comparative Validation Study for Low-Resource Clinical Settings

Abdullah Jalal B.S[1], Fatimah Jalal B.S[1], Rami Dali, B.S[1], Umar Hashim, M.S[1], Varna Taranikanti, M.D., M.S., PH.D[1].

[1]Oakland University William Beaumont School of Medicine

Background

- Accurate fracture detection remains a major challenge, especially in low-resource and underserved healthcare settings
- Limited radiology expertise can delay diagnosis and treatment
- Recent advances in deep learning and open-source AI models offer potential solutions
- However, clinical reliability and validation remain uncertain
- This study evaluates two open-source models:
  - YOLOv7 Bone Fracture Detection
  - BoneFractureClassification (MohtashamMurshid)
- **Objective:** To assess diagnostic performance and clinical applicability

Methods

Study design: Retrospective validation study

Data Source: 50 randomly sampled MURA musculoskeletal radiographs, including fracture and non-fracture cases

Models tested:

- YOLOv7 Bone Fracture Detection
- BoneFractureClassification

Deployment: Models were integrated using Cursor IDE and tested through a Streamlit interface

Analysis: Binary predictions and confidence scores were evaluated using sensitivity, specificity, precision, F1 score, accuracy, AUC, and average true-positive confidence.

Results

Metric	YOLOv7 Bone Fracture Detection	MohtashamMurshid BoneFractureClassification
Sensitivity (Recall)	72	88
Specificity	92	64
Precision (PPV)	90	71
F1 Score	80	78.6
Accuracy	82	76
Average Confidence (True Positives)	78.4	72.7
AUC (ROC)	0.89	0.85

Figure 1. Comparative performance metrics of the YOLOv7 Bone Fracture Detection and MohtashamMurshid BoneFractureClassification models evaluated on the MURA radiograph dataset.

Key Findings

YOLOv7 Model

- Accuracy: **82%**
- Precision: **90%**
- Specificity: **92%**
- F1 Score: **80%**
- AUC: **0.89**
- Strong performance with **low false positives**

BoneFractureClassification Model

- Sensitivity: **88%**
- Accuracy: **76%**
- Precision: **71%**
- Specificity: **64%**
- F1 Score: **78.6%**
- AUC: **0.85**
- Higher detection rate but **more false positives**

Discussion

- Results show a trade-off between **false positives and missed fractures**
- **YOLOv7** demonstrated stronger precision and specificity, making it better suited for **confirmatory diagnostic support**
- **BoneFractureClassification** showed higher sensitivity, making it more useful for screening or triage
- Model choice may depend on clinical setting: prioritize **specificity** for confirmation or **sensitivity** when missed fractures are most concerning
- Both models show potential to support fracture detection in **low-resource settings**

Limitations

- Small sample size (n = 50)
- Single dataset (MURA), limited generalizability
- No comparison to radiologist gold standard
- No multi-site validation

References

- Rajpurkar P et al. MURA Dataset. Stanford ML Group, 2017
- Ciriello M. YOLOv7 Bone Fracture Detection (GitHub)
- Murshid M. BoneFractureClassification (GitHub)
- Rajpurkar P et al. Radiology, 2018