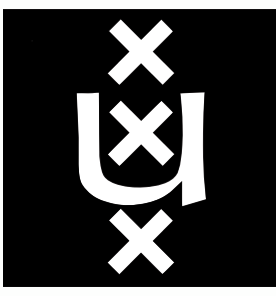


FRAC-MAS : A Safe and Explainable Multi-Agent System for Fracture Diagnosis

Hardik Iyer, Tirath Bhathawala, Mihir Panchal, Ying-Jung Chen, Kiran Bhowmick, Pankaj Sonawane, Meera Narvekar



The Problem : Diagnosis Without Oversight

Bone fracture diagnosis is high-stakes: 2 to 3% of orthopedic injuries are missed at first presentation, and plain radiographs fail to reveal 10 to 20% of some skull and skeletal injuries. Deep vision models can flag fractures accurately, but a single monolithic model gives one opaque answer, with no independent check, no grounding in clinical guidelines, and no way for a clinician to see why it decided what it decided.

Research Question

How do we build diagnostic AI that verifies itself before a human ever sees it? By decomposing the pipeline into specialized, coordinating agents: one retrieves clinical guidelines, one blindly critiques the diagnosis before seeing it and escalates uncertain cases, and one translates the verified result into a report a patient and clinician can actually trust.

Contributions

Multi Agent Triage

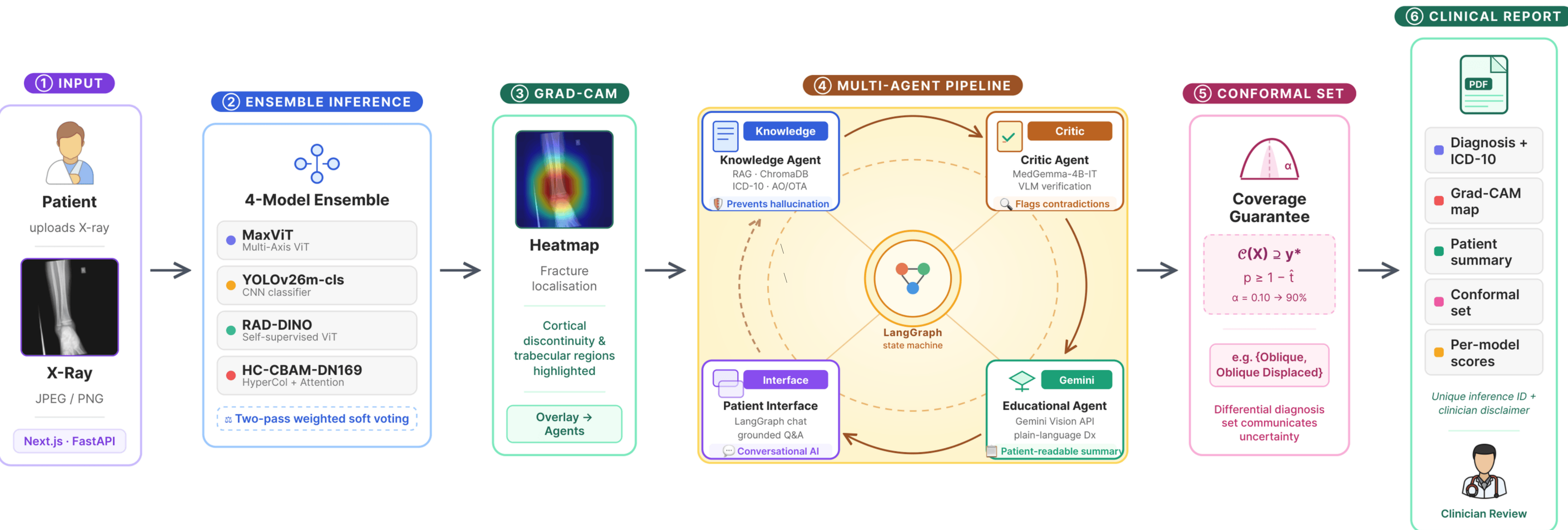
Blind Critic verification cuts escalations 7.5x while boosting confirmed accuracy to 93.8%.

Conformal Diagnosis

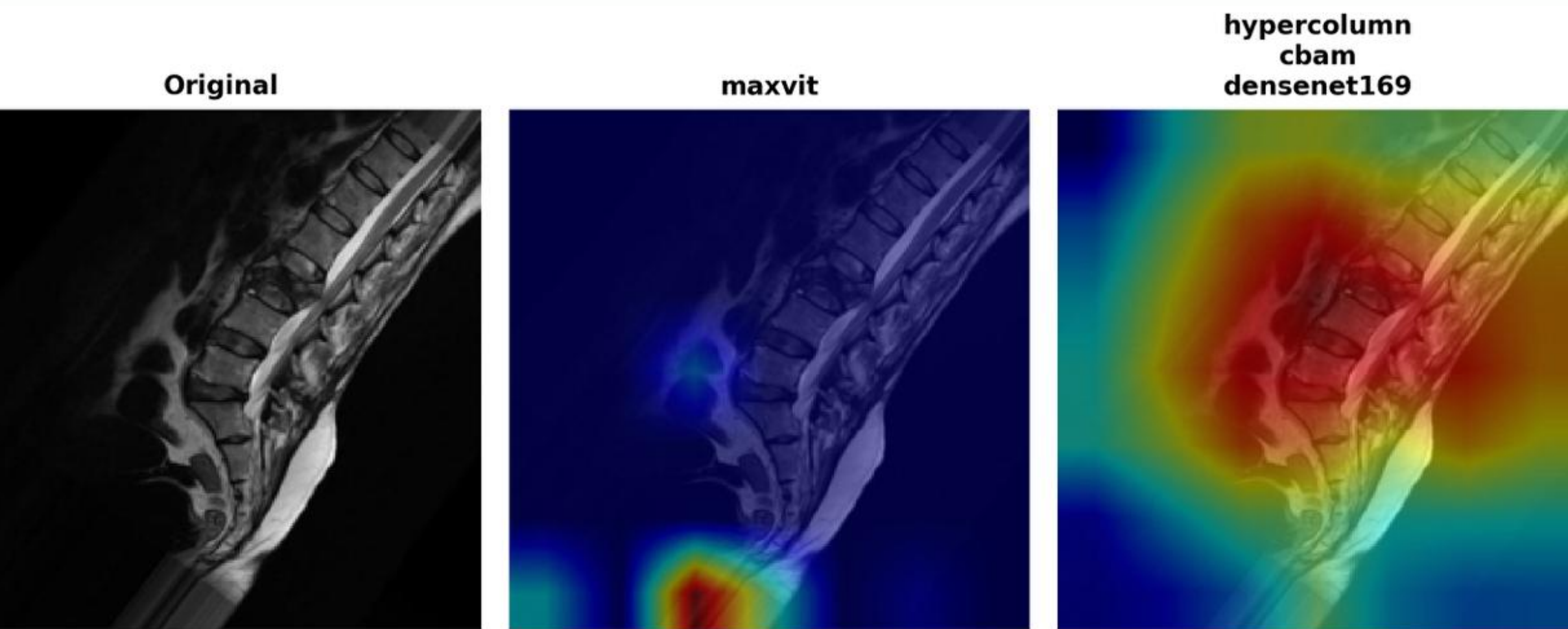
Ensemble output wrapped into statistically grounded differentials with 92.0% coverage.

Human Validated

Clinicians and patients both rate FRAC-MAS reports above Llama, MedGemma, and Gemini.



GRAD-CAM Comparison Across Ensemble Models



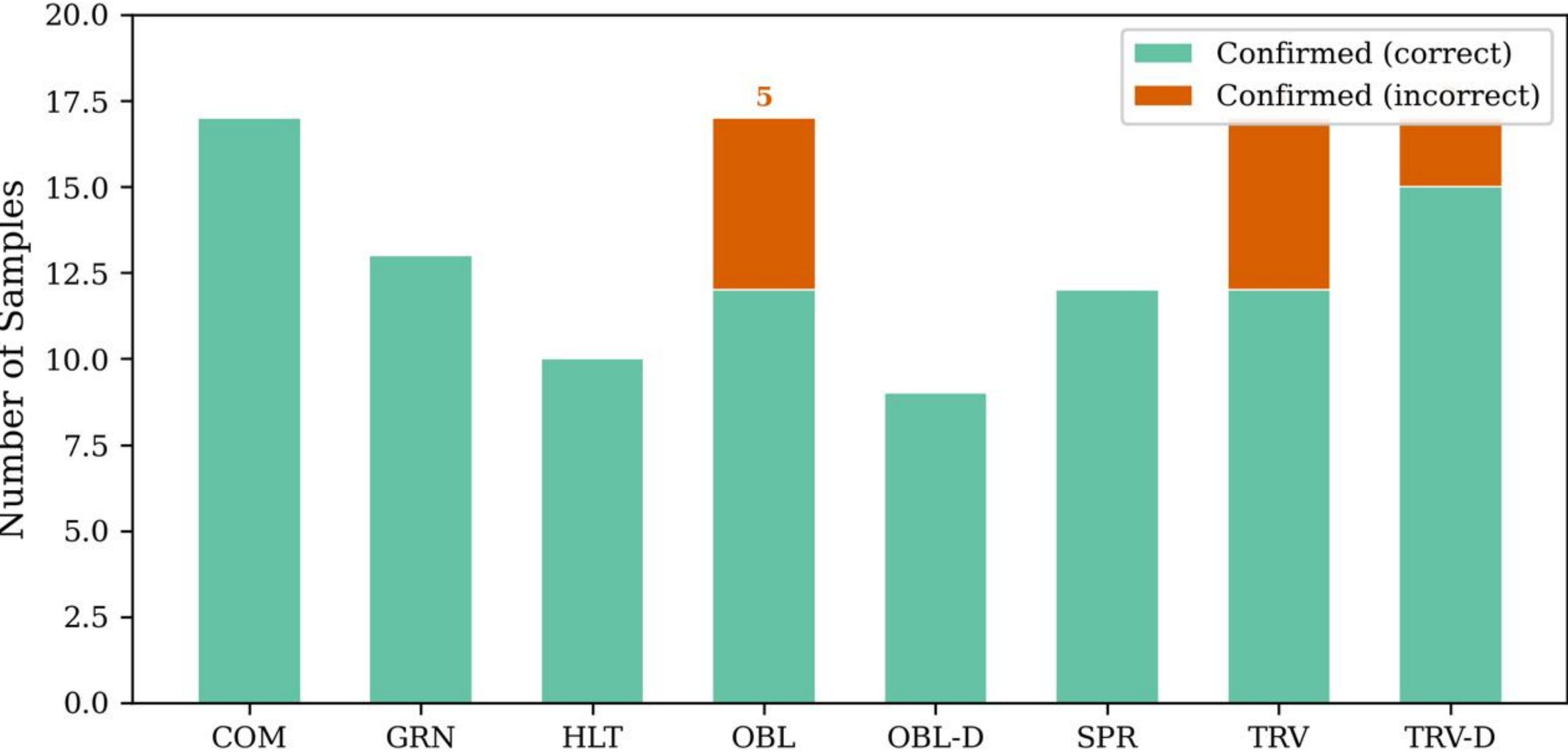
Single Model Ablation Results

Model	Accuracy	F1 Macro
MaxViT	96.2%	96.6%
HC-CBAM DenseNet-169	93.4%	93.6%
YOLOv26m-cls	93.1%	93.8%
RAD-DINO	92.5%	93.1%

External Dataset Validation Results

Dataset	Configuration	Result
Roboflow (external)	Ensemble (Stacking)	96.4% detection F1 = 0.982
FracAtlas (OOD)	Best threshold	AUC = 0.652

Confirmed Case Accuracy by Fracture Class



Key Insights

- High-Confidence Triage**: Critic auto-confirms 86.6% of cases while escalating the rest for review.
- Fewer Unnecessary Escalations**: 7.5x reduction in escalations vs. a single-agent baseline.
- Complementary Attention**: MaxViT isolates fracture lines; HC-CBAM captures broader structural context.
- Confusable Classes Struggle Most**: Oblique and Transverse show the most incorrect confirmations (see chart).
- Strong Generalization**: 96.4% detection on Roboflow; AUC drops to 0.652 on OOD FracAtlas.
- Preferred by Patients and Clinicians**: Top win rate vs. Llama, MedGemma, and Gemini; 4.14/5 clinician accuracy rating.

Conclusion & Limitations

FRAC-MAS combines a vision ensemble, conformal prediction, and a multi-agent critic to deliver auditable, clinician-verifiable fracture diagnoses, confidently triaging most cases while escalating ambiguous ones. Limitations include a small calibration pool and reduced accuracy under domain shift, motivating future work on multi-anatomy training and domain adaptation.

Demo



Github



Paper

