SIIM CAIMI26
Conference on Artificial Intelligence in Medical Imaging
OCT 26-27 | University of Pennsylvania, Philadelphia, PA

Uncertainty-Aware Triaging for Pediatric Bone Age Assessment

Bryan Luna, Cincinnati Children's Hospital Medical Center; Elanchezhian Somasundaram, PhD; Will Tepe; Rama Ayyala, MD; Jonathan Dillman, MD

Introduction/Background

This study aimed to improve pediatric bone age assessment by adding uncertainty-aware triaging to deep learning models, so routine cases could be auto-approved while uncertain or abnormal cases would be routed to radiologists. The framework was designed to align model confidence with the Brush Foundation clinical reference ranges.

Methods/Intervention

The authors evaluated 32 triage pipelines on more than 20,000 pediatric hand radiographs from an institutional cohort, alongside RSNA challenge data for model development. They compared EfficientNet and Xception backbones trained with either mean absolute error or heteroscedastic negative log-likelihood, then applied uncertainty methods including ensembles, sigma scaling, and conformal prediction. Two triage rules were tested: a strict containment strategy and a Gaussian overlap strategy against Brush Foundation reference intervals.

Results/Outcome

The study found a strong trade-off between automation and diagnostic precision. Strict containment achieved the highest precision, up to 98.0%, but automated only about 20% to 24% of cases. The best balance came from an EfficientNet late-fusion model trained with negative log-likelihood and calibrated with sigma scaling, which reached 46.5% automated approval while maintaining at least 95% precision. When the precision target was lowered to 90%, automation rose to 75.8%.

Conclusion

Calibrated uncertainty can make bone age AI safer and more clinically useful by enabling selective automation rather than fully manual review. Models that explicitly learn aleatoric uncertainty and are then recalibrated with sigma scaling provide the most practical balance of safety and efficiency. This approach could reduce radiologist workload while preserving high reliability for pediatric bone age reporting.

Statement of Impact

This work shows that bone age AI can move beyond simple prediction toward safe, selective automation when uncertainty is properly calibrated to clinical reference standards. By confidently routing routine cases and deferring ambiguous ones, the framework has the potential to improve radiology efficiency while preserving diagnostic safety.

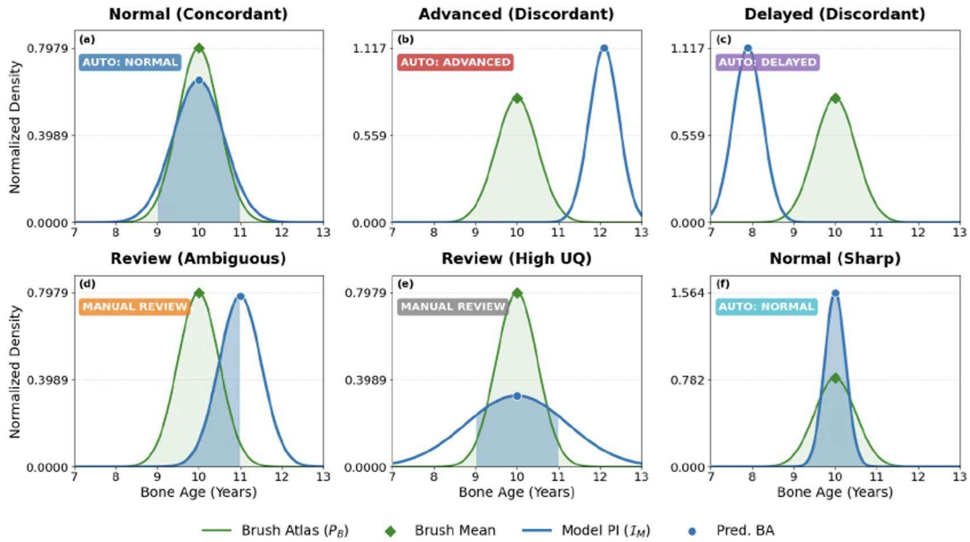

Fig. 1. Distributional Triage via Gaussian Overlap. Panels (a, b, c, f) are auto triage candidates and (d, e) require manual review. Green curves show the Brush atlas and blue curves the model

Distributional Triage via Gaussian Overlap. Panels (a, b, c, f) are auto triage candidates and (d, e) require manual review. Green curves show the Brush atlas and blue curves the model

Table 1: Optimal triage pipelines for specific clinical objectives. *Objectives:* **Max Auto.** (maximized automation), **Max Prec.** (maximized precision), **95%/90% Prec.** (maximized automation given precision ≥95%/90%). *Abbreviations:* **Thresh.** (overlap threshold). **Prec.** (Precision). **Auto.** (Auto approval).

Model (Loss)	Objective	Strategy	UQ	Thresh.	Prec.	Auto.
EffNet-LF (MAE)	Max Auto.	Gaussian	Ens-MAE	60%	84.2%	96.1%
	Max Prec.	Strict	CP-Age	100%	97.4%	21.0%
	95% Prec.	Gaussian	CP-Age	90%	95.5%	43.4%
	90% Prec.	Gaussian	CP-Age	70%	90.2%	74.5%
EffNet-LF (NLL)	Max Auto.	Gaussian	Ens-NLL	60%	85.4%	92.9%
	Max Prec.	Strict	CP-Sigma	100%	98.0%	22.7%
	95% Prec.	Gaussian	Sigma-Age	90%	96.0%	46.5%
	90% Prec.	Gaussian	Ens-NLL	80%	90.0%	75.8%
Xcep-EF (MAE)	Max Auto.	Gaussian	Ens-MAE	60%	84.1%	95.8%
	Max Prec.	Strict	CP-Age	100%	97.1%	20.5%
	95% Prec.	Gaussian	CP-Age	90%	96.0%	40.9%
	90% Prec.	Gaussian	Ens-MAE	95%	90.0%	73.7%
Xcep-EF (NLL)	Max Auto.	Gaussian	Sigma-Age	60%	86.3%	88.2%
	Max Prec.	Strict	Sigma-Age	100%	98.0%	24.2%
	95% Prec.	Gaussian	Sigma-Age	90%	96.0%	44.7%
	90% Prec.	Gaussian	Ens-NLL	70%	90.1%	73.7%

Optimal Triage Pipelines across Clinical Objectives. Objectives: Max Auto.: Maximizes the automated approval rate; Max Prec.: Maximizes diagnostic precision; 95% / 90% Prec.: Maximizes auto-approval while strictly maintaining a precision lower bound of ≥95% or ≥90%. Abbreviations: Gauss.: Gaussian overlap strategy; Thresh.: Probability mass overlap threshold.

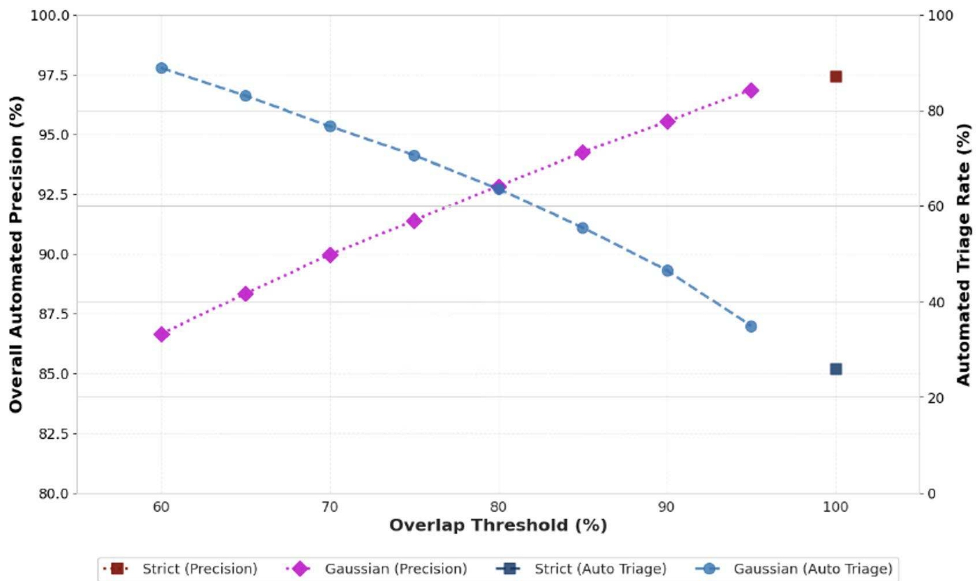

Fig. 2. Precision-automation trade-off for the EfficientNet-LateFusion (NLL) model. The *Gaussian* strategy (lines) enables tunable overlap thresholds: increasing the threshold improves overall precision (pink, left axis) but reduces the automated triage rate (blue, right axis). The *Strict* strategy (squares) operates rigidly at 100%, maximizing precision at the expense of automation.

Precision-automation trade-off for the EfficientNet-LateFusion (NLL) model. The Gaussian strategy (lines) enables tunable overlap thresholds: increasing the threshold improves overall precision (pink, left axis) but reduces the automated triage rate (blue, right axis). The Strict strategy (squares) operates rigidly at 100%, maximizing precision at the expense of automation.

Keywords
uncertainty quantification; automation; bone age; deep learning; triaging