

An Assessment of Information Loss When Validating Radiology AI Models

Authors:

Patrick Doyle MS, Allison Spake DVM, DACVIM (Cardiology), Jessica Tagudar MS, Sean Turner MS, Camden Black-Ingersoll MS, Rebecca Baumruck DVM, DACVR, Mason Savage DVM, DACVR

IDEXX Departments: Diagnostic Imaging + Telemedicine, Data and AI Center of Excellence, Medical Data Insights Center of Excellence, Westbrook ME

Objective:

The purpose of this study was to quantify the effect of single, double, and triple read labeling strategies on the statistical assessment of radiography patient condition classification models.

Background:

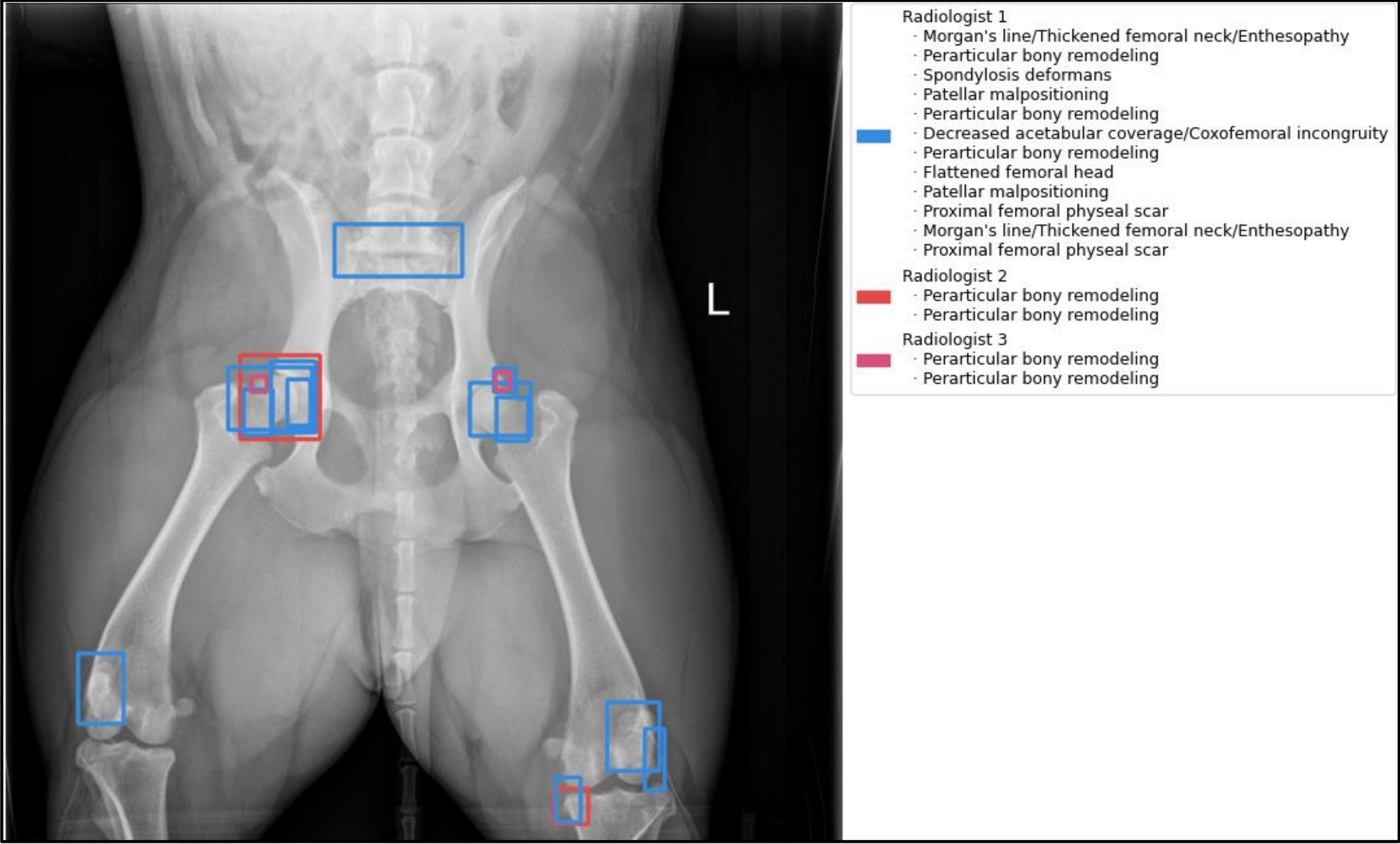
Validating machine learning models in veterinary radiology presents unique challenges. Low signal-to-noise ratios and inherent interpretive variability among radiologists mean that labels derived from a single reader may introduce systematic error, producing misleading performance estimates. Understanding how labeling strategy affects validation outcomes is essential for responsible AI development and deployment in veterinary imaging.

Methods:

A total of 25,000 canine radiographic studies were independently interpreted and labeled by board-certified veterinary radiologists. Three binary classification models were selected to represent a range of interobserver agreement profiles:

- Low agreement — Redundant tracheal membrane (29.3% unanimous positive agreement)
- Medium agreement — Cardiomegaly (61.7% unanimous positive agreement)
- High agreement — Pulmonary nodule/mass/miliary pattern (71.5% unanimous positive agreement)

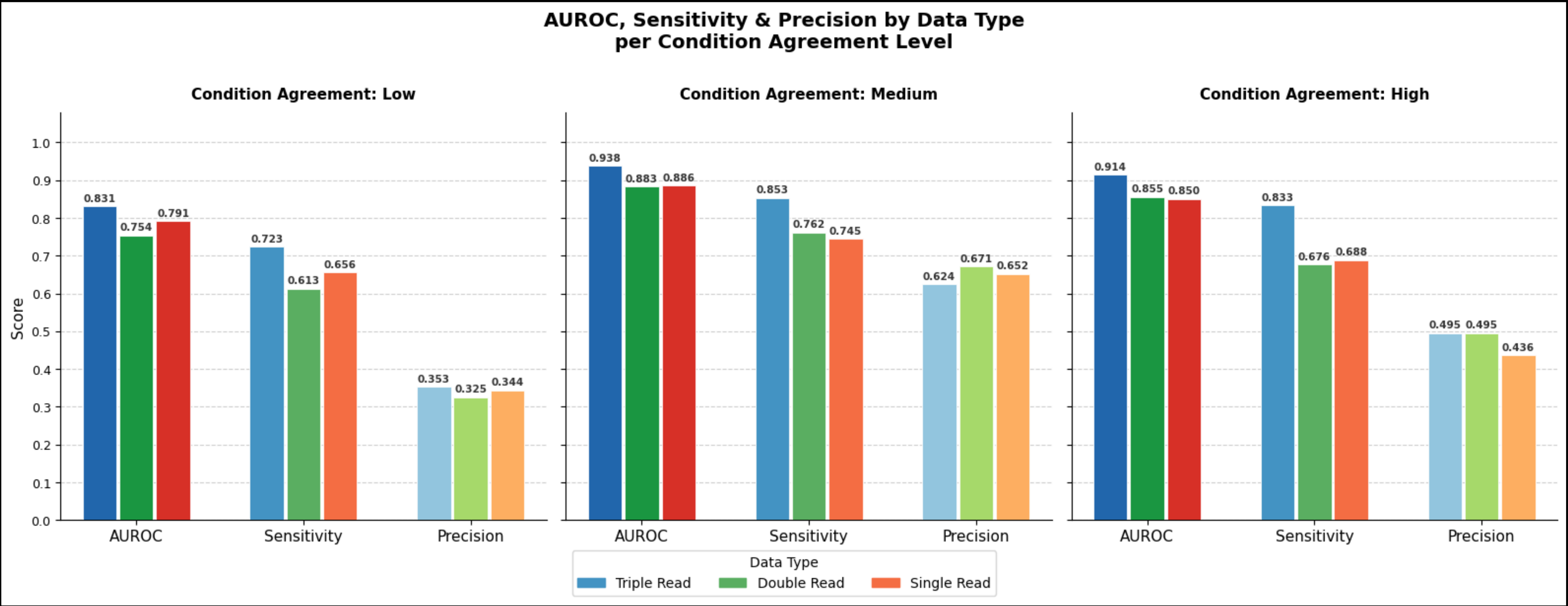
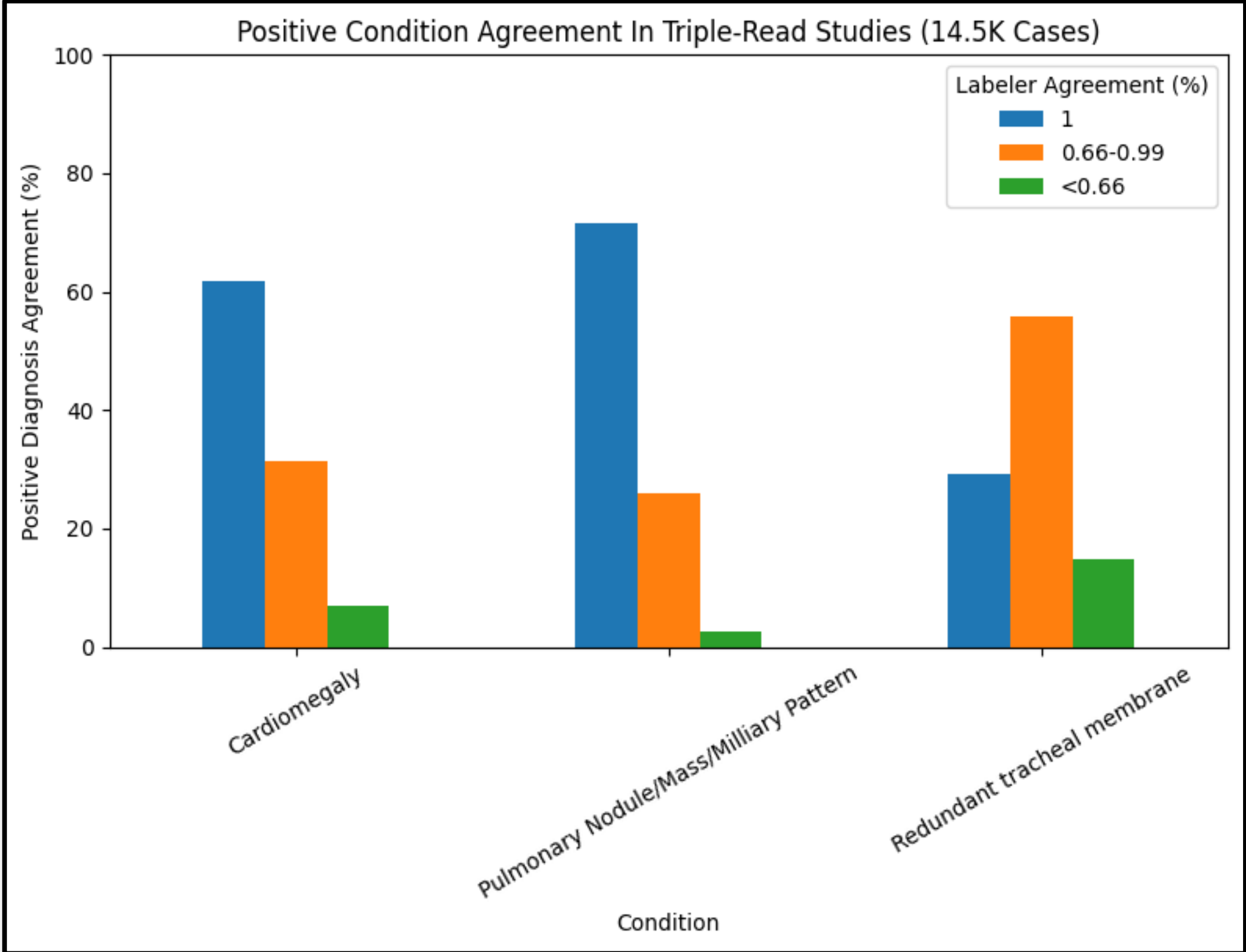
Each model was evaluated using single-reader, dual-reader, and triple-reader consensus reference standards. Performance was assessed across accuracy, AUROC, sensitivity, precision, specificity, NPV, and MCC.



Performance Metrics by Read Type and Agreement Level

data_type	condition agreement	accuracy	AUROC	sensitivity	precision	specificity	npv	MCC	positive cases	total cases
triple_read	low	0.751	0.831	0.723	0.353	0.756	0.937	0.373	224	1440
double_read	low	0.719	0.754	0.613	0.325	0.741	0.904	0.285	243	1440
single_read	low	0.733	0.791	0.656	0.344	0.749	0.915	0.324	241	1440
triple_read	medium	0.85	0.938	0.853	0.624	0.849	0.952	0.636	307	1352
double_read	medium	0.833	0.883	0.762	0.671	0.859	0.906	0.599	370	1352
single_read	medium	0.822	0.886	0.745	0.652	0.852	0.899	0.573	368	1352
triple_read	high	0.889	0.914	0.833	0.495	0.896	0.978	0.587	60	551
double_read	high	0.864	0.855	0.676	0.495	0.893	0.947	0.501	74	551
single_read	high	0.86	0.85	0.688	0.436	0.883	0.956	0.472	64	551

triple_read: accuracy, AUROC, sensitivity, precision, specificity, npv, MCC



Results:

Triple-read consensus labeling consistently produced the strongest performance estimates across all agreement levels and metrics (Table 1). Compared to triple-read validation:

- Single- and dual-read strategies resulted in 10–30% reductions in key metrics including sensitivity and AUROC.
- Label error rates increased by 10% or greater under single- and dual-read conditions.
- Performance gaps were most pronounced for conditions with low interobserver agreement, where label noise has the greatest impact on metric reliability

Conclusion:

Triple-read consensus labeling substantially reduces validation error in veterinary radiology image classification models, particularly for conditions with high interobserver variability. These findings suggest that single- and dual-read reference standards may systematically underestimate true model performance. Triple-read labeling should be considered a minimum standard for both model validation and training data curation in veterinary diagnostic imaging.