

Prioritizing Inter-Labeler Agreement Improves Machine Learning Models for Canine Thoracic Diseases

Authors: Patrick Doyle, Allison Spake, Jessica Tagudar, Sean Turner, Ben Schestopol, Rebecca Baumruck, Mason Savage

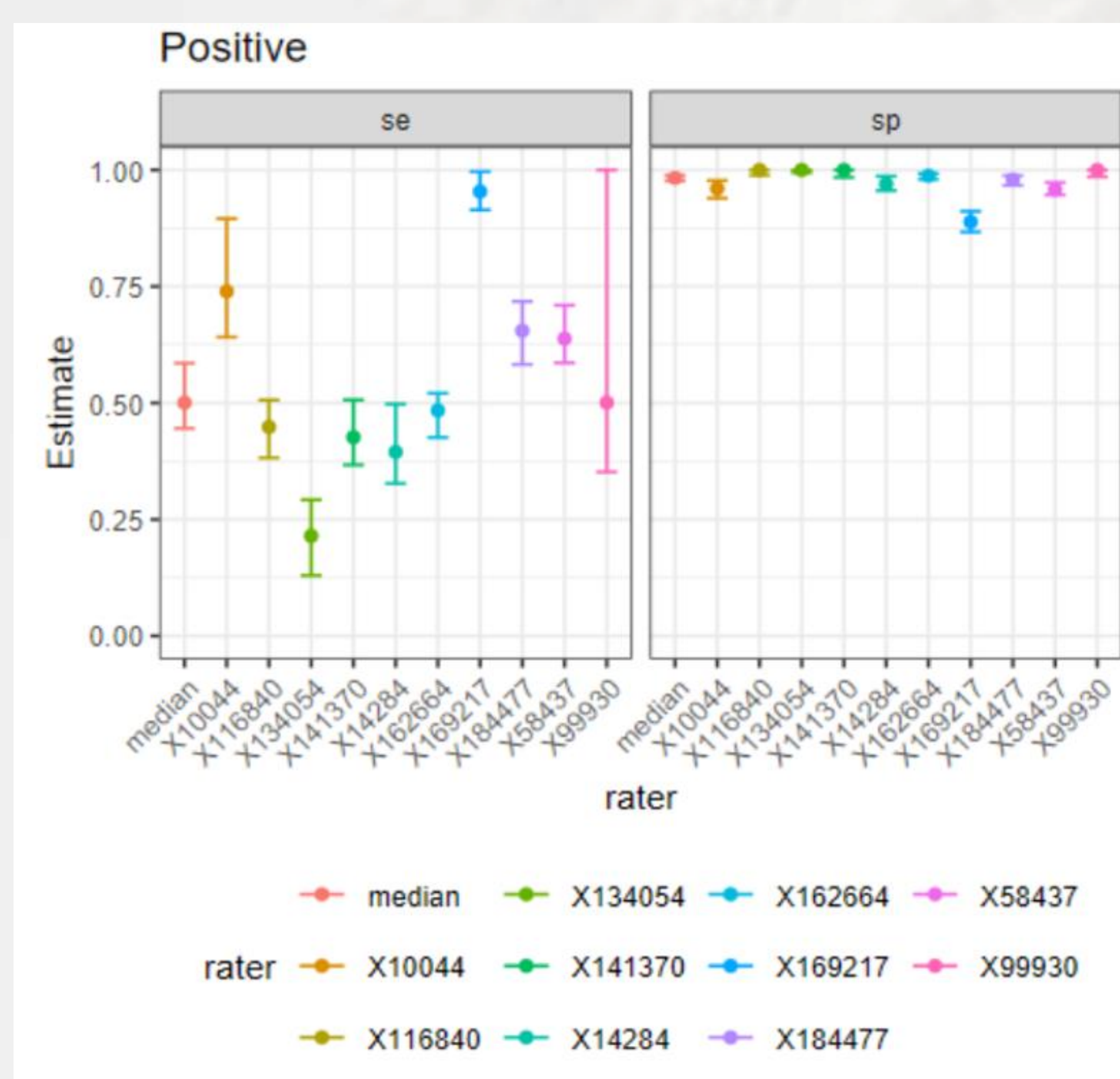
Objective

To investigate the importance of intra-labeler and inter-labeler studies in understanding the feasibility of machine learning (ML) projects in assessing canine thoracic diseases and to discuss possible solutions when positive and/or equivocal variability is high.

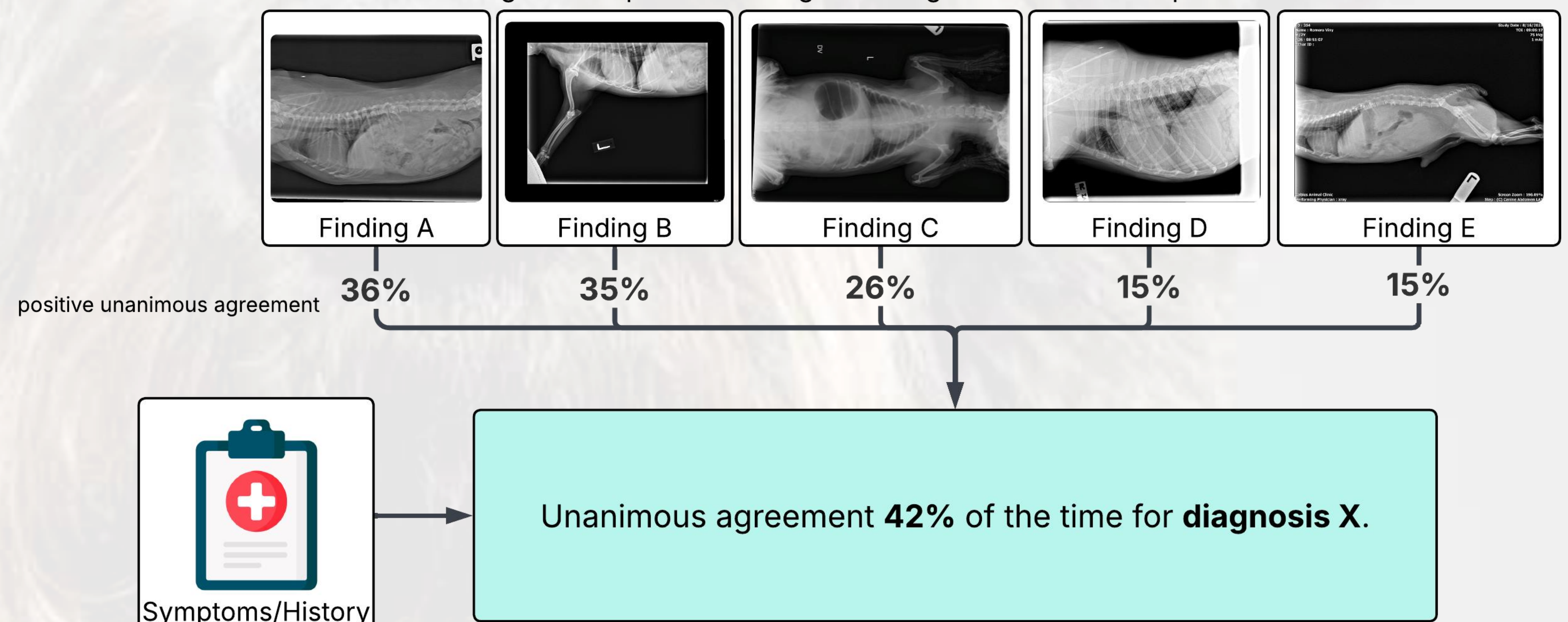
Background

As machine learning (ML) and artificial intelligence (AI) become more prevalent in assisting veterinarians in their workflow, including diagnoses, the need for consensus ground truth becomes critical. Traditional supervised ML requires ground truth labels, but in certain pathologies, trained labelers may systematically disagree due to differences in education/specialty training, professional experience, task specific training, biased ground-truth methods, or inherent equivocality. Our work investigates the importance of the intra-labeler and inter-labeler studies in understanding the feasibility of ML projects and discusses possible solutions when intra-labeler or inter-labeler variability is high.

Expectation maximization of expert variability

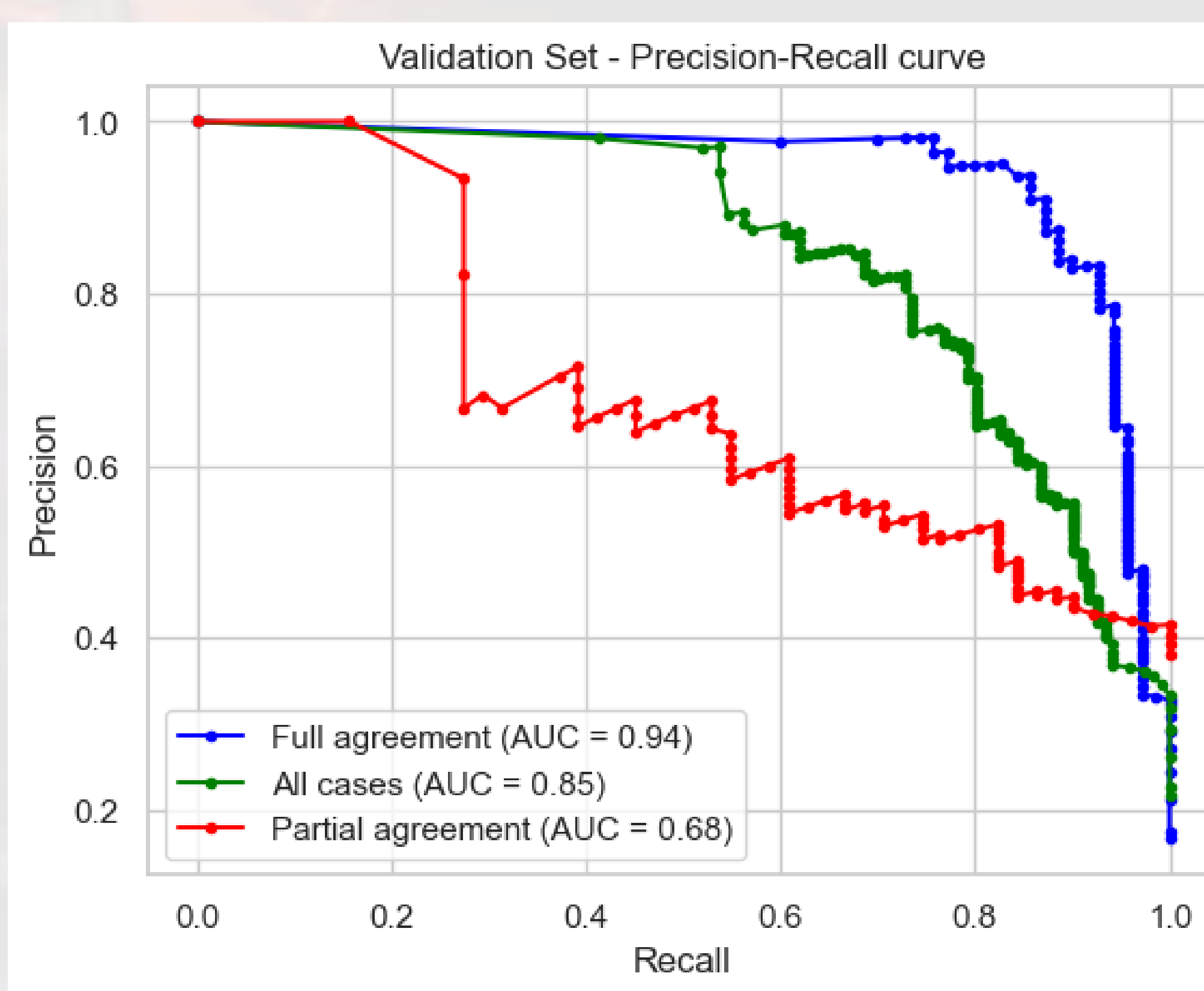
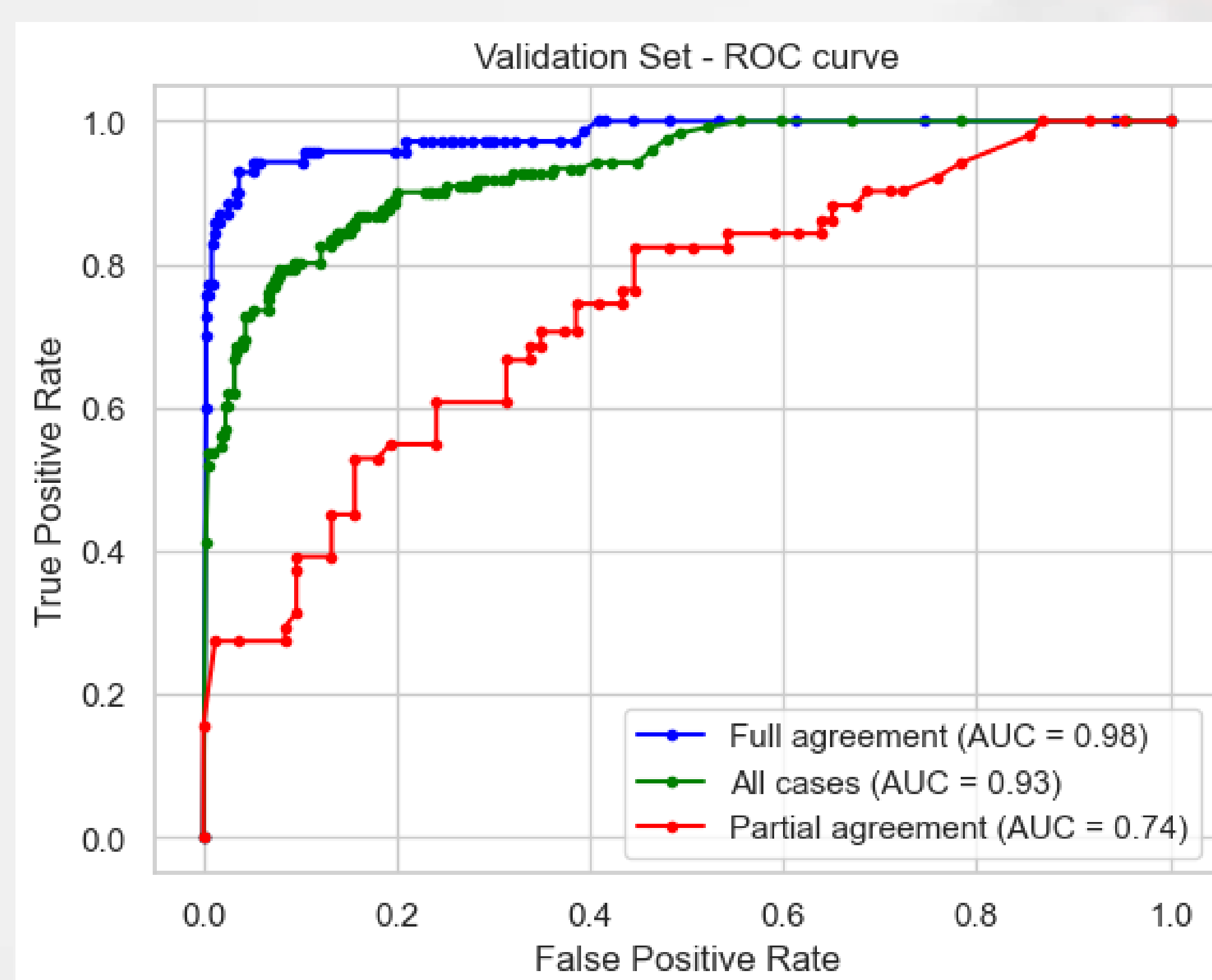


Unanimous agreement positive findings and diagnoses across multiple readers can be scarce.

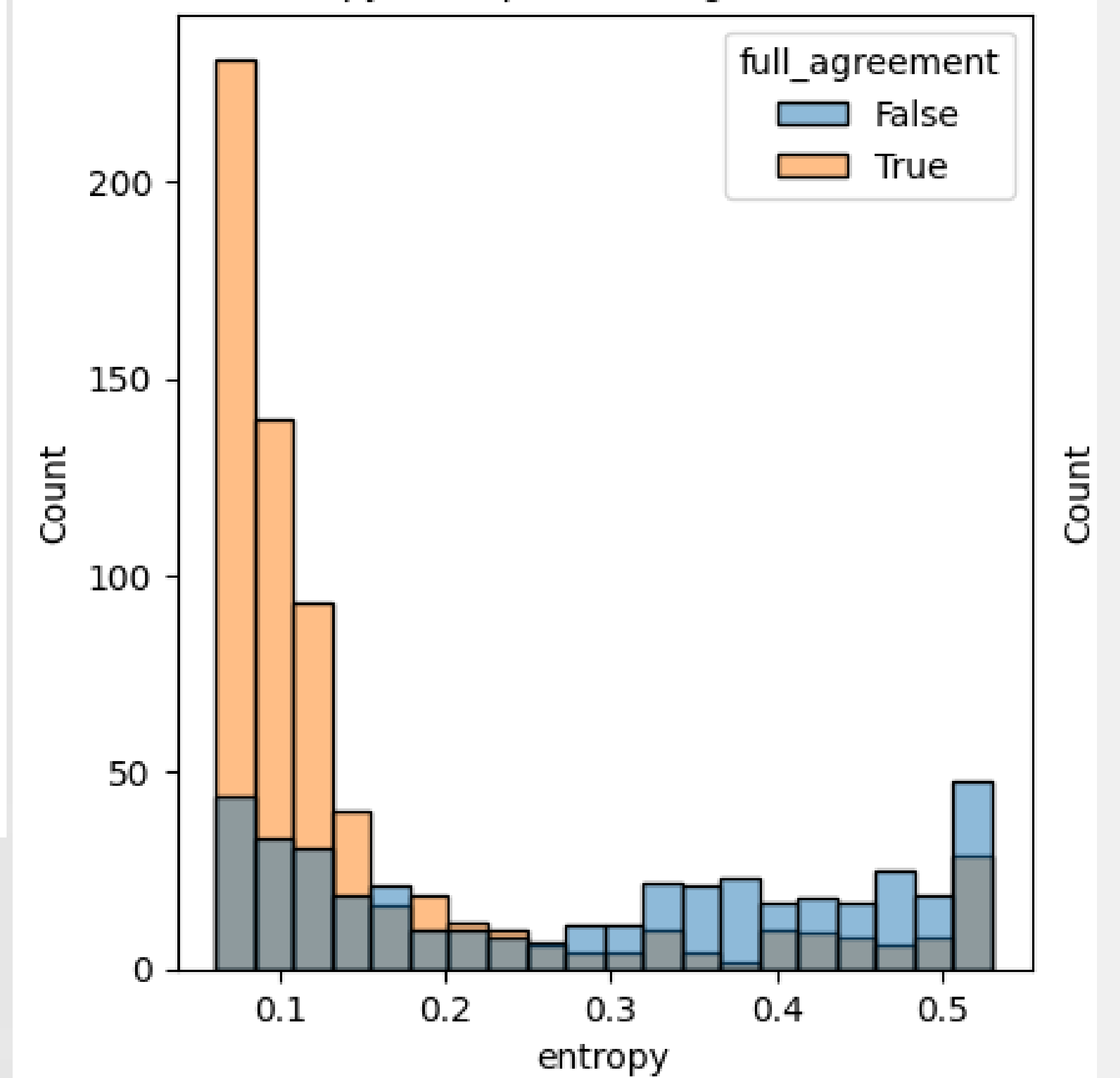


Methods and Results

When modeling data with conflicting or ambiguous signal, statistics can be misleading. Low prevalence rates of a diagnosis in a dataset, coupled with expert disagreement, causes numerous standard measures to appear highly performant while producing a sub-optimal customer experience. Analysis below shows that unanimous agreement positive cases have a significant signal advantage over discordant cases.



Model entropy on all positive diagnosis labeled cases.



Model Metrics					
Dataset	Sensitivity	Specificity	Precision	Accuracy	AUROC
Entire Set	0.95	0.96	0.51	0.96	0.93
Unanimous	1	0.99	0.9	0.99	0.98
Discordant	0.88	0.79	0.38	0.72	0.74

Conclusions

Understanding and addressing intra-labeler and inter-labeler variability is crucial for the successful implementation of ML in veterinary diagnostics. This understanding could lead to the adoption of diagnostic methods that result in greater agreement among experts. While long term solutions to inter-labeler variability include standardized expert training protocols, the development of more robust ground-truth methods, and the use of consensus-building techniques among labelers, training data including a high proportion of unanimous agreement cases could improve the reliability of ML models particularly in the short term.