

Evaluating Pre-Trained Human Radiology Foundation Models on Veterinary Radiology Data

Authors:

Sean Turner MS, Camden Black Ingersoll MS, Patrick Doyle MS, Rebecca Baumruck DVM, DACVR, Jessica Tagudar MS, Mason Savage DVM, DACVR

Objective:

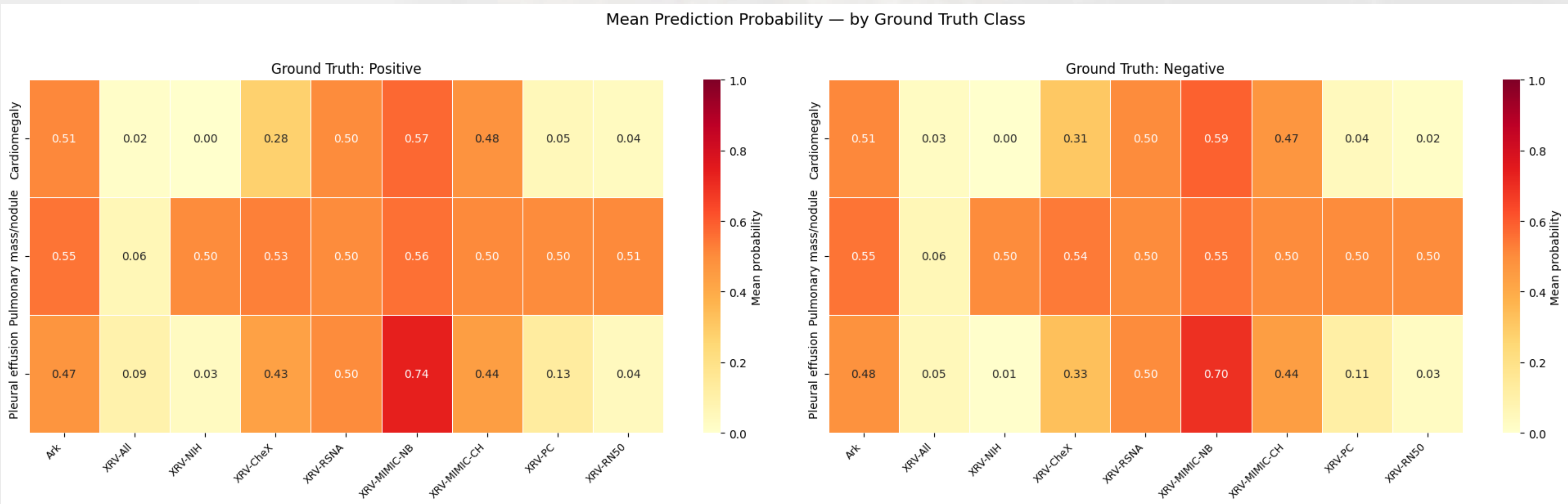
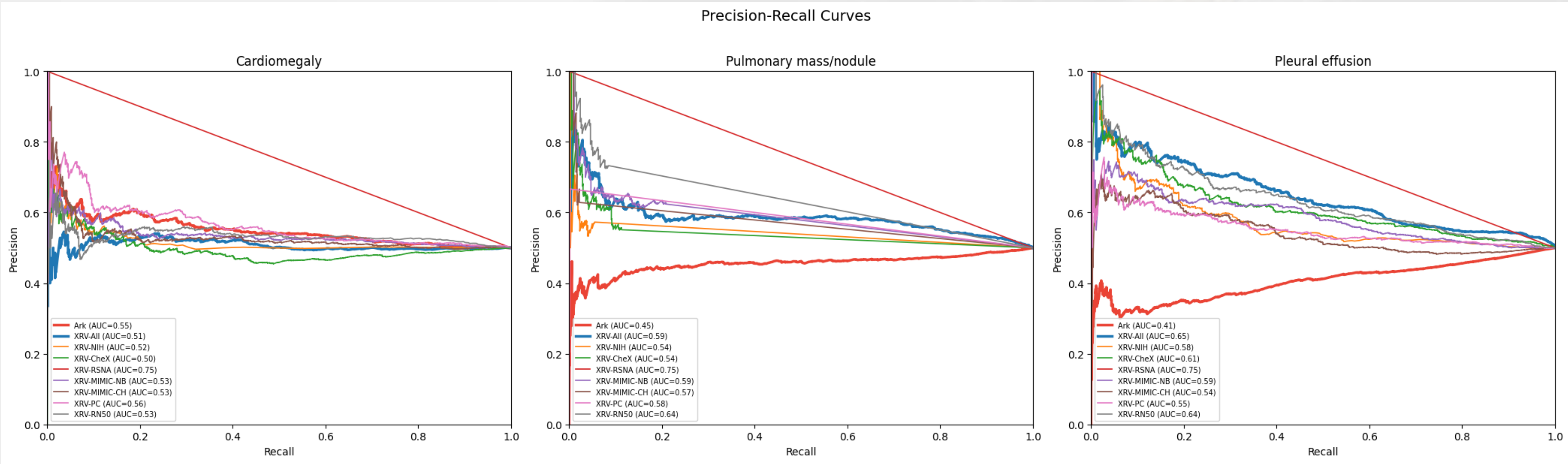
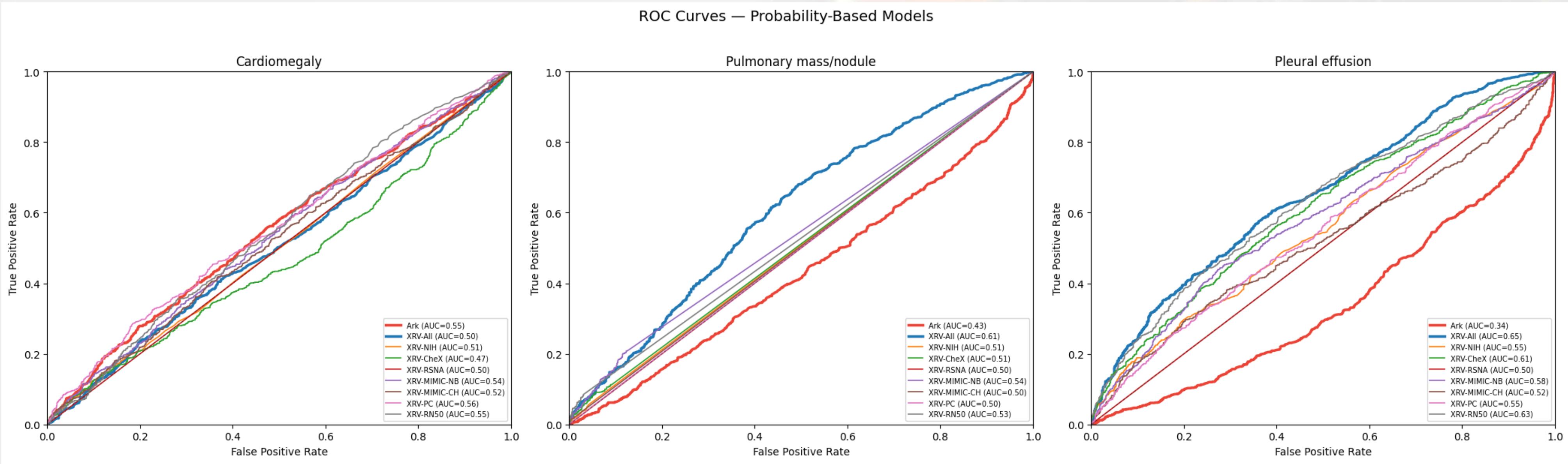
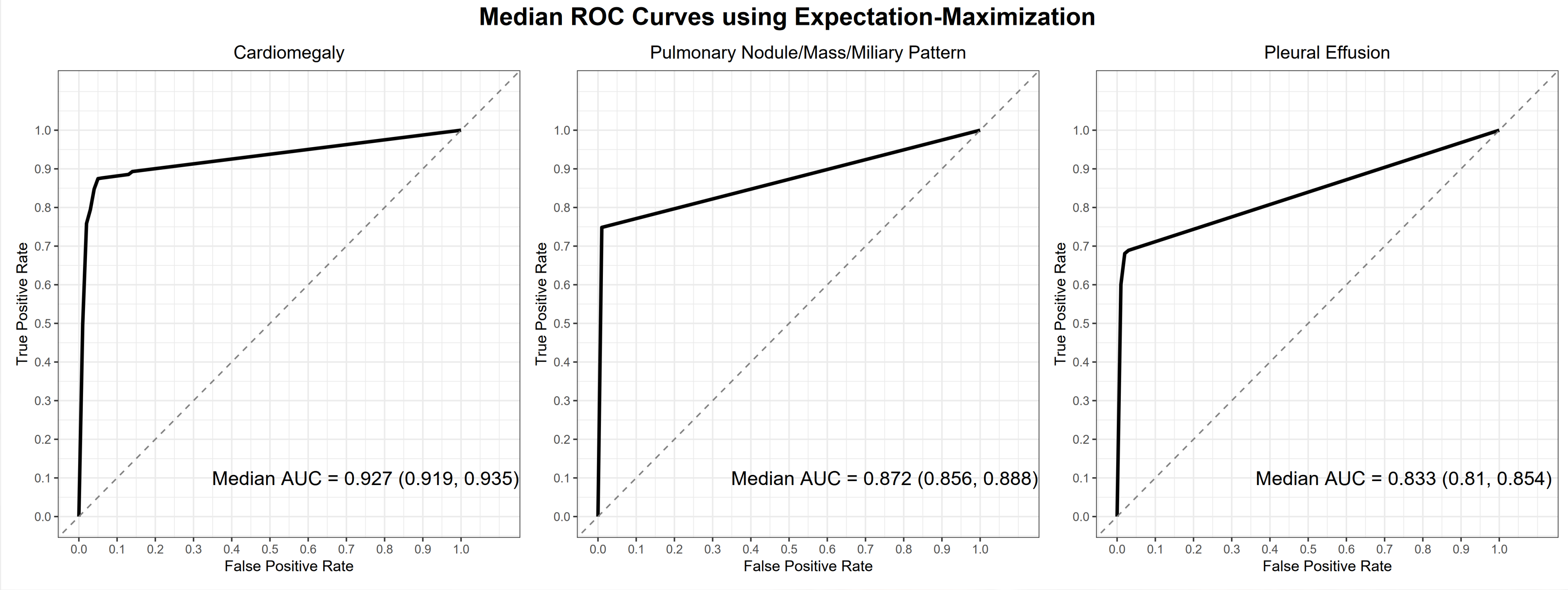
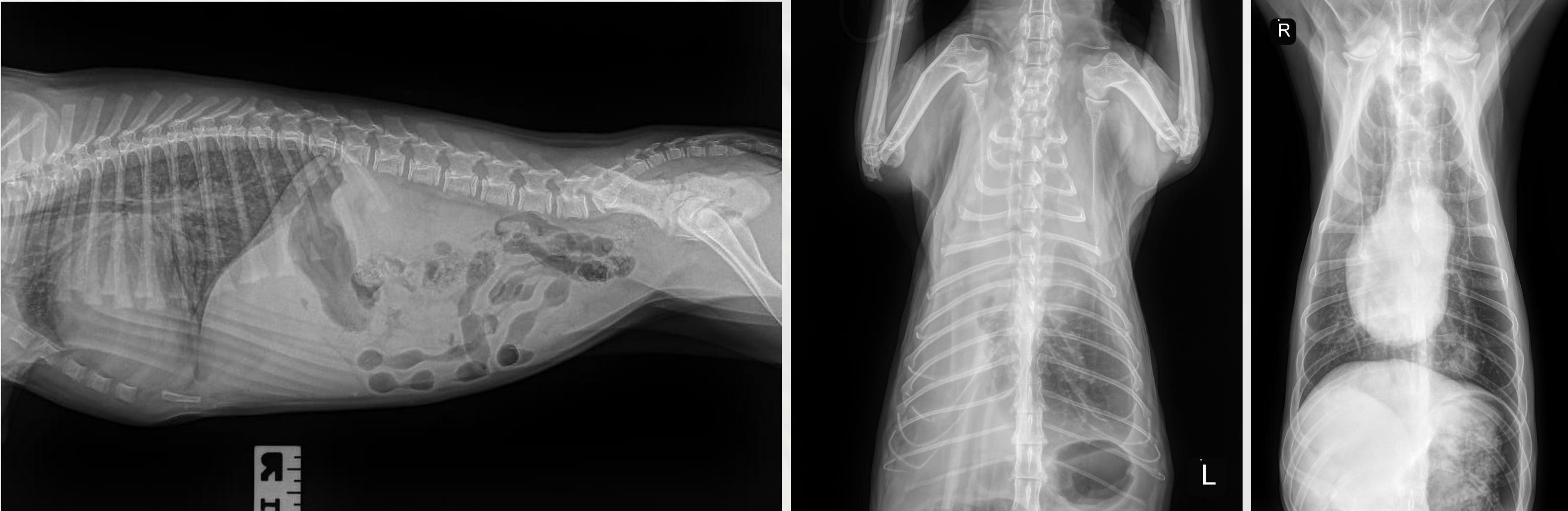
The purpose of this study was to evaluate the performance of open-source and open-weight human radiology foundation models on a curated, expert-annotated veterinary imaging dataset and to identify any performance limitations when these models are applied outside the human domain.

Background:

Significant advances in open-source and open-weight human radiology foundation models (e.g., RAD-DINO, ARK+, MedGemma) have not been matched by equivalent progress in veterinary radiology, likely due to a lack of high-quality, open-source veterinary imaging datasets available to academic and industry researchers. If high-performing human foundation models generalize to veterinary patients, they could serve as powerful transfer learning starting points - reducing labeled data requirements and accelerating model development in a domain where curated open-source datasets remain scarce.

Methods and Results:

A reference dataset of triple-reader-labeled canine radiographs labeled by board-certified veterinary radiologists was used to assess sensitivity, specificity, precision, and accuracy of selected foundation models across a subset of veterinary imaging findings chosen based on compatibility with all the selected offerings. Model performance was evaluated against a veterinary reference standard and compared across models. Cardiomegaly, pleural effusion, and pulmonary masses/nodules are common thoracic findings in canines and human chest radiographs, thus were selected for this evaluation. MedGemma, ARK+, and the XRV suite were selected as they produce state of the art human radiograph predictions. The evaluation dataset was comprised of 6000 images: 1000 negative and 1000 positives for each of the three findings.



	Finding	Model	Accuracy	Precision	Recall	F1	N
0	Cardiomegaly	MedGemma	0.507	0.510	0.393	0.444	2000
1	Cardiomegaly	Ark	0.505	0.503	0.908	0.647	2000
2	Cardiomegaly	XRV-AII	0.500	1.000	0.001	0.002	2000
3	Cardiomegaly	XRV-NIH	0.500	0.000	0.000	0.000	2000
4	Cardiomegaly	XRV-CheX	0.500	0.500	0.218	0.304	2000
5	Cardiomegaly	XRV-RSNA	0.500	0.500	1.000	0.667	2000
6	Cardiomegaly	XRV-MIMIC-NB	0.529	0.524	0.644	0.578	2000
7	Cardiomegaly	XRV-MIMIC-CH	0.518	0.518	0.530	0.524	2000
8	Cardiomegaly	XRV-PC	0.501	1.000	0.002	0.004	2000
9	Cardiomegaly	XRV-RN50	0.500	0.500	0.006	0.012	2000
10	Pulmonary mass/nodule	MedGemma	0.506	0.773	0.017	0.033	2000
11	Pulmonary mass/nodule	Ark	0.493	0.356	0.016	0.031	2000
12	Pulmonary mass/nodule	XRV-AII	0.500	0.600	0.003	0.006	2000
13	Pulmonary mass/nodule	XRV-NIH	0.500	0.500	1.000	0.667	2000
14	Pulmonary mass/nodule	XRV-CheX	0.500	0.500	1.000	0.667	2000
15	Pulmonary mass/nodule	XRV-RSNA	0.500	0.500	1.000	0.667	2000
16	Pulmonary mass/nodule	XRV-MIMIC-NB	0.500	0.500	1.000	0.667	2000
17	Pulmonary mass/nodule	XRV-MIMIC-CH	0.500	0.500	1.000	0.667	2000
18	Pulmonary mass/nodule	XRV-PC	0.500	0.500	1.000	0.667	2000
19	Pulmonary mass/nodule	XRV-RN50	0.500	0.500	1.000	0.667	2000
20	Pleural effusion	MedGemma	0.551	0.763	0.148	0.248	2000
21	Pleural effusion	Ark	0.383	0.369	0.332	0.350	2000
22	Pleural effusion	XRV-AII	0.501	1.000	0.003	0.006	2000
23	Pleural effusion	XRV-NIH	0.500	1.000	0.001	0.002	2000
24	Pleural effusion	XRV-CheX	0.573	0.614	0.392	0.479	2000
25	Pleural effusion	XRV-RSNA	0.500	0.500	1.000	0.667	2000
26	Pleural effusion	XRV-MIMIC-NB	0.515	0.509	0.839	0.634	2000
27	Pleural effusion	XRV-MIMIC-CH	0.524	0.528	0.460	0.491	2000
28	Pleural effusion	XRV-PC	0.509	0.656	0.040	0.075	2000
29	Pleural effusion	XRV-RN50	0.501	1.000	0.003	0.006	2000

Conclusion:

No evaluated open-source human radiology foundation model transferred meaningfully to canine thoracic radiographs. Across all three findings, accuracy clustered near 0.50, consistent with coin-flip level classification. The pattern of extreme recall values (many models achieving recall of 0.000 or 1.000) indicates that most classifiers defaulted to blanket positive or negative predictions rather than learning discriminative features. Performance among the best-available models for each finding further illustrates the ceiling. For cardiomegaly, the top-performing model (Ark, F1=0.647, Recall=0.908) showed high recall paired with near-chance accuracy and precision, suggesting a positive-call bias rather than true classification ability. For pleural effusion, the leading model (MedGemma, F1=0.248, Recall=0.148) correctly identified fewer than 1 in 7 positive cases. Pulmonary mass/nodule detection was effectively non-functional across all models, with the top performer (MedGemma, F1=0.033, Recall=0.017) offering no clinically meaningful utility.

These results highlight a critical domain gap between human and veterinary imaging and underscore the need for veterinary-specific training data and model development.