

RapidRead: Global Deployment of State-of-the-art Radiology AI for a Large Veterinary Teleradiology Practice

Michael Fitzke¹, Conrad Stack¹, Andre Dourson¹,
Rodrigo M. B. Santana¹, Diane Wilson², Lisa Ziemer²,
Arjun Soin³, Matthew P. Lungren³, Paul Fisher², Mark Parkinson¹

¹Mars Digital Technologies

²Antech Imaging Services

³Stanford University, Center for Artificial Intelligence in Medicine & Imaging

Abstract

This work describes the development and real-world deployment of a deep learning-based AI system for evaluating canine and feline radiographs across a broad range of findings and abnormalities. We describe a new semi-supervised learning approach that combines NLP-derived labels with self-supervised training leveraging more than 2.5 million x-ray images. Finally we describe the clinical deployment of the model including system architecture, real-time performance evaluation and data drift detection.

1 Introduction

Radiographic imaging is the most common clinical imaging modality in the world and important for clinical evaluation of patients both in human and veterinary medicine. The global veterinary imaging market size was valued at 2.01 billion in 2018 and increase in utilization for veterinary diagnostics is expected to be driven largely by rising demand for pet insurance and growing animal healthcare expenditure, increasing companion animal population, and growth in the number of veterinary practitioners globally [3]. Currently, while medical imaging use increases, the shortage of veterinary radiologists remains a critical problem globally as less than 5% of practicing veterinarians have formal training in imaging interpretation [3]. Further challenges in veterinary radiology arise chiefly due to the wide variety of species, breeds, size, as well as inconsistent positioning and poor standardization, all of which contribute to misinterpretation and degraded clinical care [7, 5, 1, 18, 39, 43].

Recent advancements in machine learning in medical imaging applications have demonstrated expert-level performance for automated diagnoses across a variety of modalities and diseases in human radiology. Similar applications have been described in veterinary radiology, yet suffer from limitations stemming from dataset size, lack of multi-site data informing generalizability, and chiefly for the field of deep learning in medical imaging broadly, lack experimental data with real-world constraints via deployment of automated systems in practice which vastly limit clinical translation of these promising technologies. Ultimately, to realize the potential for modern supervised deep

learning in automated medical image diagnosis, large real-world, multi-center, multi-class datasets with human-expert annotations and consensus benchmarks are needed; further a framework for model deployment in practice at scale is ideal to facilitate translation of these technologies in medical imaging practice broadly [14].

The purpose of this study is to explore state-of-the-art computer vision techniques leveraging a large multi-center dataset consisting of more than 2.5M veterinary radiographs, human expert supervision, multi-class end-to-end diagnostic tasks, and describe model best model performance in real-time deployment for a high-volume global teleradiology practice. Since the clinical workflows we aim to model change dynamically, this study also outlines data drift detection analysis that is tailored to a vision-based use case and presents a viable approach towards monitoring of large models deployed across multiple institutions, without requiring immediate access to expert-annotated labels at inference time.

This work could potentially serve as a vital framework for accurate clinical machine vision applications in radiology at scale; addressing critical limitations in the use of this technology to advance the health of pets while also providing insights for medical imaging human medicine [22, 6, 25, 20, 35, 34, 15, 33].

2 Data

2.1 Images

Over 3.9 million veterinary radiographs were available for modeling. The majority of these radiographs had previously been archived as lossy (quality 89) JPEG images, down-sampled to a fixed width of 1024 pixels (px) ('set 1' in Table 1). Of the remaining radiographs most were provided as (lossless) PNG images that were down-sampled so the smaller dimension (width or height) was 1024 px ('set 2' in Table 1). The final subset of images were collected as part of the current clinical workflow (Figure 9), where DICOM images were down-sampled to a fixed height of 512 px, then converted to PNG ('silver' in Table 1). In all cases, the downsampling process preserved original aspect ratios. All images were provided by Antech Imaging Services (AIS) along with a subset of the original DICOM tags as metadata. All images / studies cover real clinical cases from various client hospitals and clinics ($N > 3500$) received by AIS within the last 14 years (Figure 1).

A number of filtering steps were applied prior to modeling. First, duplicated and low-complexity images were removed using *imagemagick*, open-source software for displaying, creating, converting, modifying, and editing images. Second, imaging artefacts, and views / body parts irrelevant to our set of findings were filtered out using a CNN model trained for this purpose. Studies with more than 10 images were also excluded. Approximately 2.7 million images were left after filtering, representing over 725,000 distinct patients.

2.2 Annotation / Labeling

Images were annotated for the presence of 41 different radiological observations (Table 2). For the majority of images (pre-2020 studies) labels were extracted from corresponding (study-wise) radiological reports using an automated, natural language processing (NLP)-based algorithm. The radiology reports summarize all the images in a particular study, and were written by over 2000 different board-certified veterinary radiologists. Images from the most recent studies (years 2020-2021; 'silver' in Table 1) were individually labeled by a veterinary radiologist, immediately after

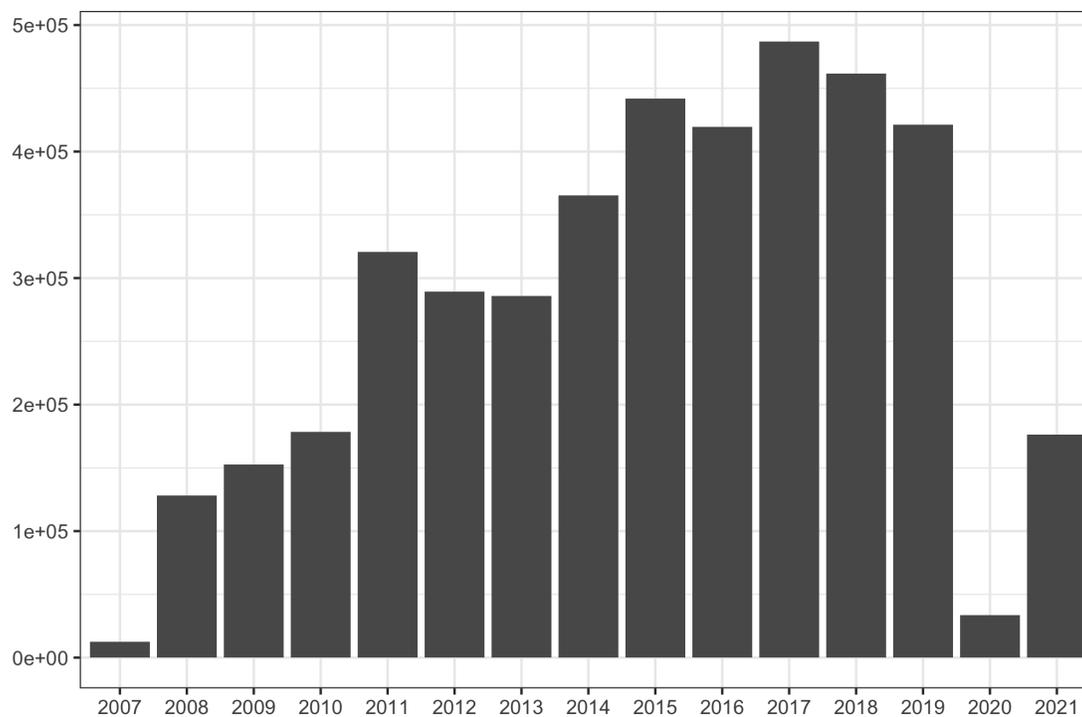


Figure 1: Number of images provided, by year

Table 1: Summary of x-ray image data

	Set 1	Set 2	Silver	(Total)
No. of images before filtering	3318579	455077	186563	3960219
No. of images after filtering	2202007	344050	186563	2732620
(% Filtered out)	0.336461	0.243974	0	0.309983
Image format(s)	JPEG	PNG	PNG	
Range of image widths	1024	[1024 - 40345]	[71 - 2516]	
Range of image heights	[258 - 3068]	[1024 - 35328]	512	

they finished evaluating the study.

In order to assess the trained model’s accuracy and inter-annotator variability, a small number of images (N=615) were randomly selected from the ‘silver’ set to be labeled by twelve additional radiologists. These data were not used in training or validation. To produce ground truth labels for ROC and PR analyses the labels for each image were aggregated by a majority-rules vote (i.e., a majority of the 12 radiologists indicated the finding was present). Point estimates of the False Positive Rate (FPR) and Sensitivity for a particular radiologist were calculated by comparing their labels to the majority-rules vote of the 11 others.

Automated extraction of labels from the “Findings” section of radiology reports was done using a modified version of the rules-based labeling software, CheXpert [22, 30]. We extended the number of labels used by CheXpert to 41, and adapted its syntactic patterns and key-phrases to reflect the language found in our report data, and veterinary medicine more generally.

While these reports contain observations from all images in a study, they never explicitly link these observations with specific image files. Therefore, the per-study labels extracted from a report were initially applied to all images in the corresponding study, and then masked using a set of expert-provided rules. These rules ensure that labels are only applied to images that show the corresponding body parts (e.g., ‘Cardiomegaly’ labels are removed from images of the pelvis).

3 Methods

3.1 Noisy Radiology Student

Using our data sets of individually labeled data $(x_1, y_1), \dots, (x_n, y_n)$ and data labeled by the CheXpert labeler [22] adapted to veterinary radiology reports $(\tilde{x}_1, \tilde{y}_1), \dots, (\tilde{x}_m, \tilde{y}_m)$. For each finding \tilde{y}_i has as entry either 0 (negative), 1 (positive) or u (uncertain) and is determined on the level of the radiology report instead on the level of the individual image. Hence label noise is present on the latter data set.

To combine both data sets into a single mode we combine the distillation approaches of [26] and [44] and adapt them to our multi-label use case. We first train a teacher model θ_*^t with the individually labeled data set and add noise to the training process:

$$\frac{1}{n} \sum_{i=1}^n l(y_i, f^{noise}(x_i, \theta^t)).$$

After that we use our teacher model to infer soft pseudo labels for the images $\tilde{x}_1, \dots, \tilde{x}_m$. In the inference step we do not use any noise:

$$\hat{y}_i = f(x_i, \theta_*^t), \forall i \in (1, \dots, m).$$

We then combine the soft pseudo labels with the NLP derived labels using the rule:

$$\bar{y}_i = \begin{cases} \lambda \hat{y}_i + (1 - \lambda) 0.5 & \text{if } \tilde{y}_i = u \\ \lambda \hat{y}_i + (1 - \lambda) \tilde{y}_i & \text{otherwise} \end{cases}, \forall i \in (1, \dots, m).$$

Table 2: Radiological labels

Anatomical Grouping	Observation
Cardiovascular	Cardiomegaly
	Left Atrial Enlargement
	Left Ventricular Enlargement
	Right Ventricular Enlargement
	Right Atrial Enlargement
	Main Pulmonary Artery Enlargement
	Aortic Abnormality
	Heart Base Mass Effect
Microcardia	
Pulmonary Structures	Bronchial pattern
	Interstitial Unstructured
	Pulmonary Alveolar
	Pulmonary Interstitial - Nodule (Under 1 cm)
	Pulmonary Vascular
Pleural Space	Pulmonary Mass (Over 1 cm)
	Sign(s) of Pleural Effusion
	Pleural Mass Effect
	Pneumothorax
Mediastinal Structures	Thin Pleural Fissure Lines
	Esophageal Dilation
	Intrathoracic Tracheal Narrowing
	Tracheal Deviation
	Mediastinal Mass
Extra Thoracic	Mediastinal Widening
	Mediastinal Lymph Node Enlargement
	Spondylosis
	Liver Abnormality
	Abdominal mass
	Intervertebral Disc Disease
	Gastric Foreign Material
	Cervical Tracheal Narrowing or Opacity
	Degenerative Joint Disease
	Decreased serosal detail
	Gastric Distention
	Aggressive Bone Lesion
	Fracture and/or Luxation
	Splenomegaly
	Gastric Dilatation Volvulus
Subcutaneous Nodule	
Subcutaneous Mass	
Fat Opacity Mass (e.g. lipoma)	

With those derived labels we train an equal-or-larger student model θ_*^s with noise added to the student:

$$\frac{1}{n} \sum_{i=1}^n l(y_i, f^{noise}(x_i, \theta^s)) + \frac{1}{m} \sum_{i=1}^m l(\tilde{y}_i, f^{noise}(\tilde{x}_i, \theta^s)).$$

In our production setting we set the student model as new teacher and re-do the process without the first step whenever we have collected a significant amount of new single-labeled data.

Training details. We train different artificial neural network architectures, specifically Densenet-121 [19], Inception-v4 [40] and Models from the EfficientNet-Family [41]. For each model we maximize the number of images that fit into memory to determine batch-size which lead to batch-sizes between 32 and 256. We use different image-input-sizes (ranging from 224×224 to 456×456) and train both with reshaping the original images to the input dimension as well as zero-padding the image to a square and then resizing to keep the ratios of the original image intact. We perform a multitude of image augmentation techniques when training such as in [10]. All our models were pre-trained on Imagenet [11] and are trained for a maximum of 30 epochs with early stopping if the validation loss is not decreasing two epochs in a row.

Probability Calibration In order to calibrate the probability for each finding, we apply the piecewise linear transformation [9] :

$$t_{\phi, opt_{\phi}}(x) = \begin{cases} \frac{x}{2 \cdot opt_{\phi}}, & \text{if } x \leq opt_{\phi} \\ 1 - \frac{1-x}{2 \cdot (1-opt_{\phi})}, & \text{else} \end{cases}$$

for all findings ϕ . We set opt_{ϕ} to optimize Youden’s J Statistic [45] on an independent validation set.

3.2 Drift Analysis

In veterinary radiology at the scale described in this work we could well have multiple factors leading to a difference between data distributions seen pre and post model deployment. These factors include changes in breeds, different radiology equipment or differences in clinical practice in different regions. Such changes can have implications on performance metrics and necessitate checking whether assumptions made during the development of models still hold while models are in production [16, 17, 32].

At time of development of the system the images X_1, \dots, X_n and findings per image Y_1, \dots, Y_n can be viewed as draws from the joint distribution $P_{dev}(X, Y)$. In a real world production scenario we are presented with images and findings from a distribution $P_{prod}(X, Y)$. In this work we specifically investigate covariate shift, i.e a change in the marginal distribution:

$$\begin{aligned} P_{dev}(Y|X) &= P_{prod}(Y|X) \\ P_{dev}(X) &\neq P_{prod}(X) \end{aligned}$$

and its influence on model performance. Qualitatively, covariate shift refers to the change in the distribution of the input features present in the training and the test data. It occurs when some previously infrequent feature vectors become more frequent, and vice versa, while the relationship between the feature and the target classes remain the same [17, 32].

To detect covariate shift we train an autoencoder $f_A(\cdot, \Theta)$ at dev-time, i.e. try minimizing $\int (X - f_A(X|\Theta))^2 dP_{dev}$ with regards to Θ . Training was done using the Alibi-Detect software [42]. We then use the trained autoencoder to reconstruct images at production time and analyze the reconstruction errors to gauge drift.

While supervised drift detection is achieved by monitoring a performance metric (AUROC, F1, PR scores) of interest over time, it requires access to an abundance of labels at inference time to quantify a change in the system performance - a requirement that is cost-prohibitive and ideally requires expert radiologist annotators. Our approach does not make any assumptions about the model that has been deployed, and requires access to training data used to build the model as well as production data ingested by the model post-deployment. We build a machine-learned representation (autoencoder) of the training dataset and use this representation to reconstruct data that is presented to the model in the real world. By comparing reconstruction errors, we effectively detect changes in the input (X-Ray image data) seen in production.

3.3 Towards Scaling Laws for X-Ray Images

Academic datasets of X-Ray images that are categorized as large (e.g. [22], [23]) have under 500,000 individual images while our dataset has over 2 Million NLP-labeled images and 180,000 individually labeled images by board certified radiologists. This allows us to start investigating questions of data scaling and related questions on model capabilities.

3.4 Ensembling

Our final model (RapidReadNet) is an ensemble of individual, calibrated deep neural networks. We found that averaging the outputs leads to slightly better metrics than voting. We use the 8 best models based on our validation set and then used the best subset method to determine the optimal ensemble. Surprisingly, the best subset was the full set of 8 models, i.e. models that were under-performing against the rest of the ensemble still helped overall predictions.

3.5 Study level fusion

To determine the findings per study we merge individual findings per image by choosing the disease class mapping to the *argmax* value per finding over all images.

4 Results

4.1 Results on a highly curated data set

Figures 2, 3, 4, 5, 6 show the result ROC and PR analysis results for the set of 613 images labeled by multiple radiologists, comparing model predictions against the radiologists' labels. Each figure shows ROC (top) and PR (bottom) curves per-finding, and point estimates of each individual

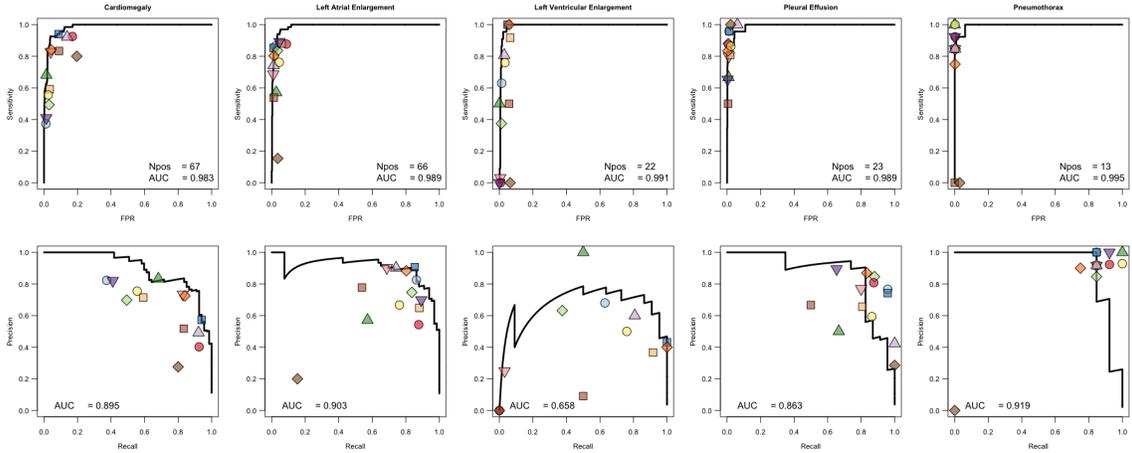


Figure 2: ROC and PR curves for Cardiovascular and Pleural Space findings

radiologist’s FPR, Precision, and Recall (Sensitivity). Findings with fewer than five positive labels were not analyzed.

For a majority of the findings, model accuracy (AUROC) is comparable to the accuracy (AU-ROC) of individual radiologists.

4.2 Longitudinal drift analysis

An autoencoder was trained using all archived images (sets 1 and 2 in Table 1), and then applied to the images from subsequent studies, checking the L2 reconstruction error of each. Figure 7 shows the distribution and quantiles of L2 errors, grouped by week.

We observe little-to-no difference between reconstruction errors; indicating that input data has remained distributionally consistent, and further suggesting that the the model is robust to data from new clients. We also show that a convolutional autoencoder is capable of quantifying differences in X-Ray image datasets on the basis of a reconstruction error. Formally, $P(X)$ and $P(Y|X)$ exhibit almost no change, and we expect the model’s performance on new data to remain comparable to testing baselines:

$$\begin{aligned} P_{dev}(Y|X) &\approx P_{prod}(Y|X) \\ P_{dev}(X) &\approx P_{prod}(X) \end{aligned}$$

This lack of perceived drift can be attributed, in part, the high diversity of organizations and animals represented in the training data. Figure 8 shows the distribution of reconstruction error as a function of the number of organizations represented. Data from a small number of organizations (1,6,...,16 orgs) does not provide a good representation of the total diversity of the image data.

4.3 Data Scaling

We observed a positive relationship between data size and model performance on an independent, hand labeled test set. We trained an Efficient-Net-b5 [41] on differently sized subsets of our data

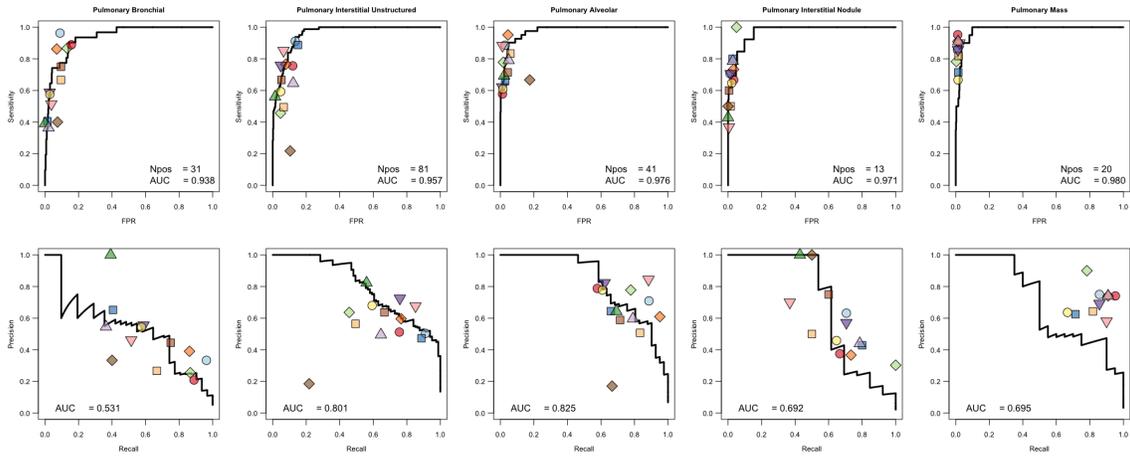


Figure 3: ROC and PR curves for Pulmonary findings

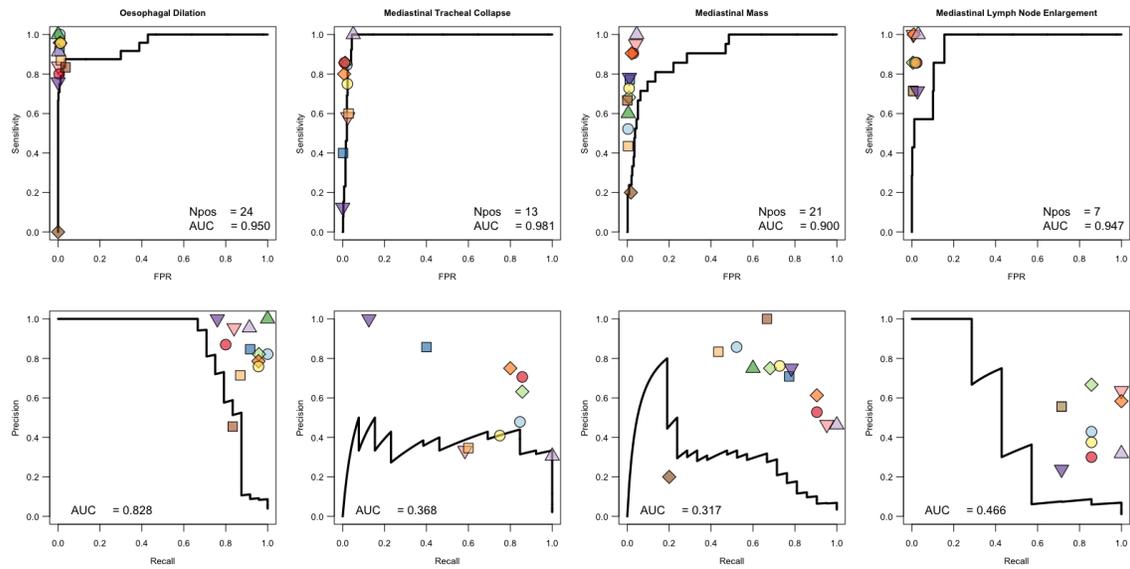


Figure 4: ROC and PR curves for Mediastinal findings

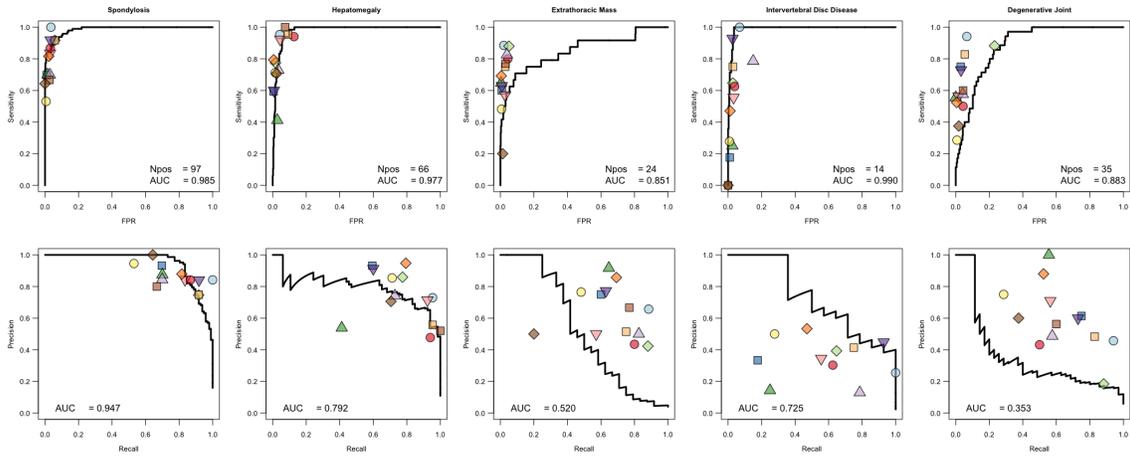


Figure 5: ROC and PR curves for Extrathoracic findings (part 1)

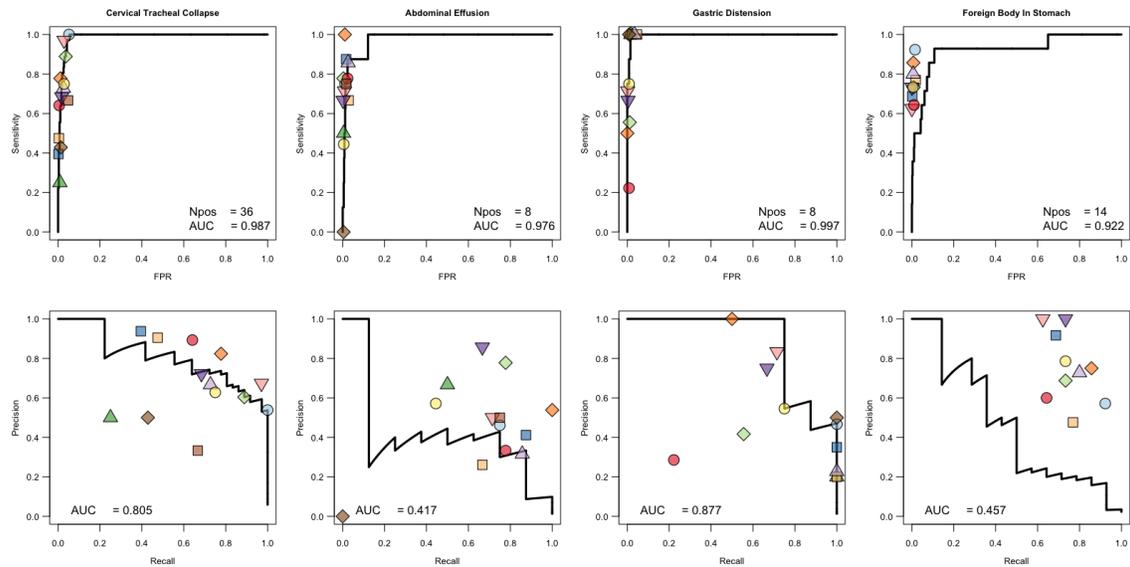


Figure 6: ROC and PR curves for Extrathoracic findings (part 2)

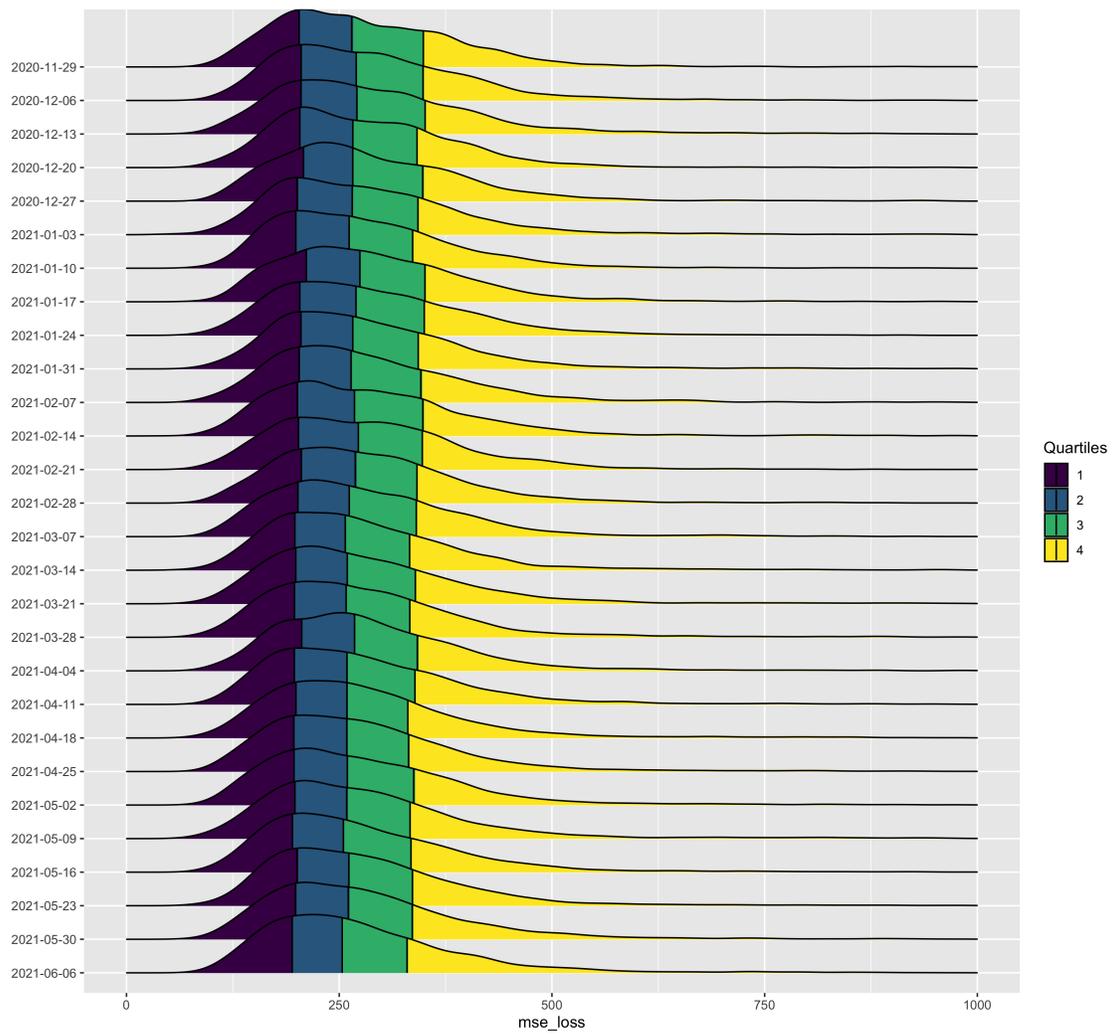


Figure 7: Weekly reconstruction error

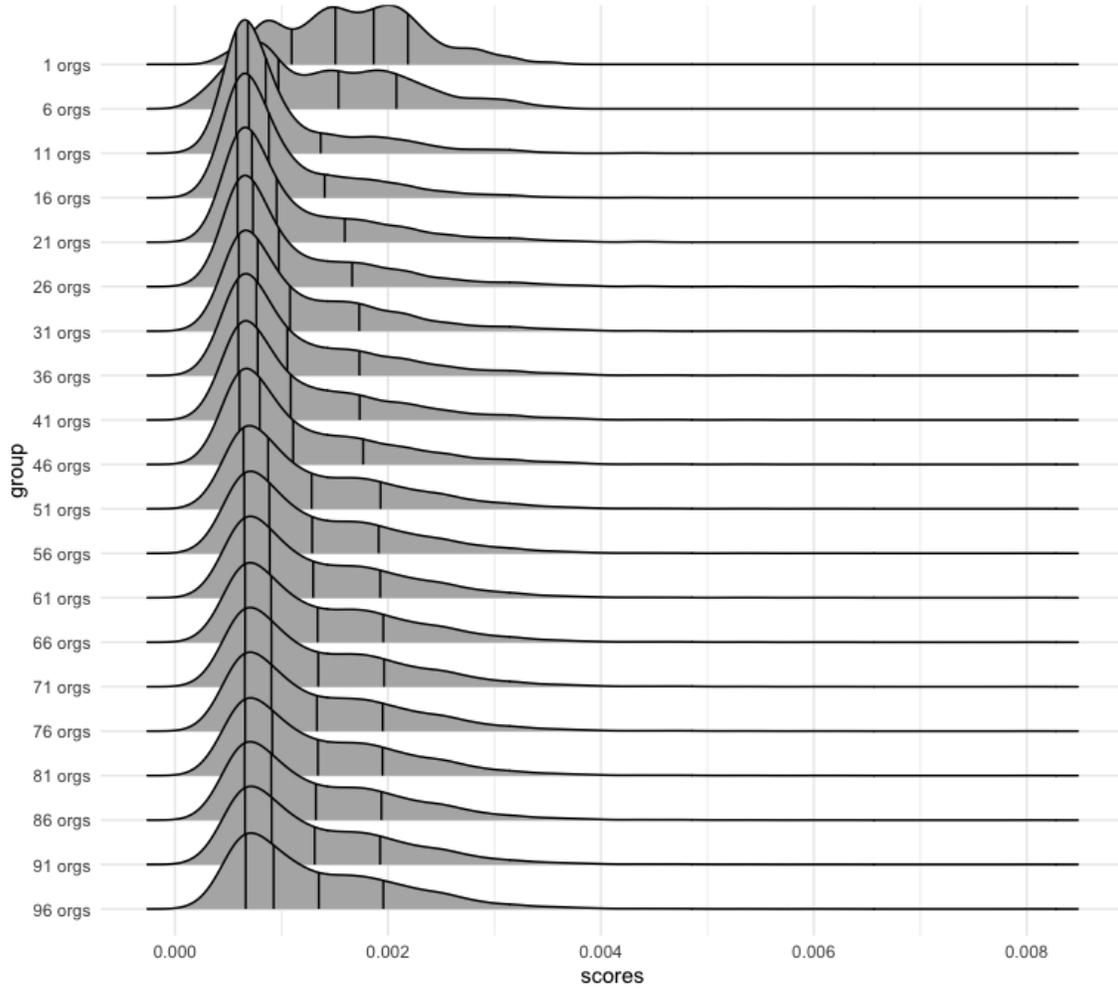


Figure 8: Loss distribution vs. number of organizations

and tested the resulting models on the same test set. Results can be observed in Table 3 and suggest the possibility of further performance gains with data scaling.

	Roc-AUC	PR-AUC
20,000 images	0.8765	0.3140
50,000 images	0.8974	0.3706
100,000 images	0.907	0.393
500,000 images	0.912	0.404
1,000,000 images	0.916	0.433
1,500,000 images	0.920	0.443
2,000,000 images	0.922	0.477
2,500,000 images	0.926	0.488

Table 3: Metrics of our models on 30477 unseen test data points against ground truth of one board certified radiologist. All data was labeled in clinical production setting.

4.4 Study level results

Results are shown in Table 4. Additional information can be found in the supplemental file, `ROC-report_silver_2021-09-13_STUDYWISE.pdf`, including ROC plots broken down by species (ie ‘canine’, ‘feline’).

5 Deployment

Deployment pipeline and workflow: We now describe the system architecture of RapidRead-Net. Our infrastructure pipeline is exclusively relying on micro-services (rest-APIs) deployed using docker containers [21]. Each container is using the FastApi [36] framework in order to deploy a rest API module, and each of them are serving as per best practices for such unique and specialized tasks.

Given the amount of images processed each day (around 15,000+) we adopted an asynchronous processing approach for our production pipeline using a message broker. This was achieved using both Redis [37], which is a NoSQL database component where we store each individual incoming requests, and use Redis Queue [13] as a background processing mechanism to consume each of the stored requests in parallel.

More specifically, each request consists of an image and its corresponding metadata. Thus, incoming requests are processed by a component named Redis Queuer which takes care of caching the data locally and registering the request into the Redis DB’s queue. Next, RQ workers are processing in parallel each of the requests stored in the queue and sending them to the model serving module called AI Orchestrator.

All the predictions from our models are gathered in JSON format [31] and are directly stored into a MongoDB [27] database (for long-term archiving) and a Redis JSON [38] database (for short term archiving). The short term storage (Redis JSON) allows us to aggregate results at the study-level along with inclusion of some contextualisation aspects.

Table 4: Study-wise ROC results for each finding. The Npositive column lists the number of studies with at least one positive label (out of 9311 total studies). Area under the Receiver-Operating Curve (AUROC), False positive rate (FPR), and Sensitivity were not calculated for findings with fewer than 10 positive instances (Npositive).

	Npositive	AUROC	FPR @ 0.6	Sensi @ 0.6
Cardiomegaly	915	0.986	0.025	0.923
Left Atrial Enlargement	738	0.988	0.016	0.912
Left Ventricular Enlargement	479	0.994	0.005	0.939
Right Ventricular Enlargement	157	0.971	0.005	0.694
Right Atrial Enlargement	90	0.984	0.003	0.767
Main Pulmory Artery Enlargement	13	0.826	0	0.308
Aortic Abnormality	20	0.687	0	0
Heart Base Mass Effect	1	n/a	n/a	n/a
Spondylosis	2133	0.993	0.010	0.962
Liver Abnormality	1294	0.969	0.041	0.934
Ex. Thoracic or abdomil mass	502	0.960	0.017	0.843
Sign(s) of IVDD	810	0.951	0.018	0.806
Gastric Foreign Material	357	0.903	0.011	0.650
Cervical Tracheal narrowing or Opacity	632	0.983	0.015	0.897
Degenerative Joint Disease	900	0.945	0.037	0.810
Decreased serosal detail	410	0.964	0.010	0.793
Gastric Distention	500	0.975	0.012	0.912
Aggressive Bone Lesion	41	0.848	0.000	0.049
Fracture and/or Luxation	101	0.753	0	0
Esophageal Dilation	213	0.986	0.010	0.897
Intrathoracic Tracheal rrowing	323	0.974	0.010	0.796
Tracheal Deviation	252	0.985	0.003	0.825
Mediastil Mass	99	0.965	0.007	0.778
Mediastil Lymph Node Enlargement (any)	56	0.911	0.001	0.536
Sign(s) of Pleural Effusion	238	0.985	0.010	0.924
Pleural Mass Effect	4	n/a	n/a	n/a
Pneumothorax	44	0.990	0.001	0.750
Bronchial pattern	1745	0.959	0.057	0.890
Interstitial Unstructured	1300	0.974	0.026	0.906
Pulmory Alveolar	798	0.986	0.018	0.910
Pulmory Interstitial - Nodule (Under 1 cm)	218	0.934	0.010	0.638
Pulmory Vascular	137	0.938	0.004	0.599
Pulmory Mass (Over 1 cm)	206	0.956	0.008	0.738
Splenomegaly	194	0.941	0.009	0.732
Gastric Dilatation Volvulus	11	0.970	0.000	0.818
Microcardia	75	0.968	0.005	0.773
Mediastil Widening	131	0.962	0.004	0.748
Pleural Fissure Lines	579	0.981	0.038	0.957
Subcutaneous Nodule	9	n/a	n/a	n/a
Subcutaneous Mass	77	0.950	0.001	0.506
Fat Opacity Mass (e.g. lipoma)	253	0.984	0.006	0.893

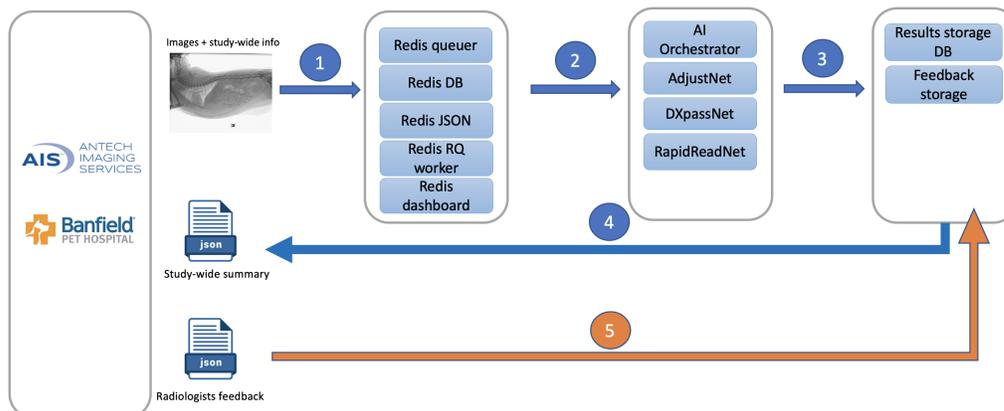


Figure 9: Infrastructure of X-Ray system

Last but not least, for monitoring purposes, we deployed a Redis dashboard instance [28], where each request can be tracked and re-processed if needed.

The micro-services are managed through docker compose [12] and organized using the following categories:

- Message broker: Pre-processing of incoming requests, monitoring and dispatch
- Models serving: AI Orchestrator module, and individual models serving. All models are using Pytorch framework[29]
- Results and feedback loop storage: Aggregation and contextualization of models results at study-level, sending back of the results

An infrastructure diagram is depicted in Figure 9

Models serving

For serving models, the central element is the AI Orchestrator container, which is responsible to coordinate the execution of the inferences from our different AI modules.

The first AI model from which predictions are collected is the *AdjustNet*. This model checks the orientation of the radiograph image, and proceeds with adjustment if required (rotation, horizontal flipping, vertical flipping etc.).

Then comes next the *DxpassNet* which validates that the image presented to the model is appropriate for the diagnostic tasks provided by RapidReadNet, a multi-label classifier associated to 41 findings corresponding to various pathologies.

Lastly, a contextualization of the results is done using the study-wide metadata which were provided along with the image during the initial upload to the service. To achieve this aggregation at study level, a record of all study-wide image inferences is temporarily stored in the Redis JSON module, and we are using a rule-based expert system tool called CLIPS [24] to manage them from this contextualization. CLIPS (C Language Integrated Production System) was developed at NASA's Johnson Space Center in order to generate an AI rule-based system for space shuttle flight controllers. With that tool we were able to accommodate the rules applied against the outputs of

our models and animal metadata in order to get context for the radiologist reports. In order to productionize CLIPS we use a python library called clipspy [8] which interacts with C tools, while we save all the rules in a mongoDb database in order to have a rule dynamism as new rules are created periodically.

Results and feedback loop storage

All records are stored in a mongoDb database in JSON format.

2. Embedded pre- and post- deployment infrastructure
3. Number of clinics, radiologist using system
4. Workflow for feedback on labels for radiologists (screenshot for structured labels)
5. Methods for adding in new labels as data is acquired - semi-supervised approach
6. Canary/shadow performance description of process

6 Discussion

The purpose of this study was to explore the development of a deep learning-based system to automate the detection of pre-defined clinical findings in canine and feline radiographs. We introduced a new method for data distillation that combines automated labeling with distillation and showed performance enhancements against the automated-labeling-only approach. We explored data scaling and its interaction with different models. We then look into the model’s performance over time and compare it with input data drift over time. The input drift method uses a machine-learned representation of the image data to analyze whether subsequent X-Ray instances seen in production have drifted over time, and does not assume access to expert labels. Lastly we present the deployment process and the embedding of the model in a larger deep learning-based platform for X-ray image processing. This work has salient implications for the development, deployment and lifecycle management of deep learning-based systems in radiology practices.

Human medical technological advancements have been driven by animal models for centuries; leading to the discovery of vitamins, hormones, antibiotics, transfusion, organ transplantation, vaccines, insulin, hemodialysis, chemotherapy, advanced diagnostic imaging techniques, and countless others. Animal medical research has saved millions of lives, human and non-human, and has transformed the world many times over. The utilization of high-performance deep learning diagnosis systems at scale in veterinary care provide critical insights that can serve to address the gap in translation for these promising technologies in human and veterinary medical imaging diagnostics into clinical practice.

This work has a larger scope than any previous work in veterinary radiology. For example, a retrospective evaluation of a multi-class CNN trained with a dataset of only 3800 single-view lateral canine thoracic radiographs demonstrated modest performance on hold out data [4]. The lack of external validation, limited data size, narrow use case, and lack of deployment insights diminish real-world applicability. Another retrospective study was recently reported leveraging both canine and feline thoracic radiographs with two views (frontal and lateral) and three clinical labels in a dataset of 2800 images reported less than state-of-the-art performance compared to human radiology machine learning clinical models [2] of comparable scale. The fact that the dataset presented in this paper is thousand orders of magnitude larger than those in aforementioned studies leads to significant gains in generalizability and robustness effects.

In conclusion, this work represents a first-of-its-kind, extensive analysis of the development and real-world application of deep learning methods in veterinary health. It shows promising potential towards using deeper architectures when scaling up the size of the data, as seen particularly in how the Noisy Radiologist Student boosts robustness in X-Ray image prediction. Furthermore, we analyze a large longitudinal dataset of medical imaging data for covariate shift and model performance and demonstrate the usefulness of clinical variability in helping avoid data drift. Lastly, we hope to have contributed to the best practices of deploying deep learning-based radiology systems at scale with findings that provide definitive insights towards advancing the use of such tools in radiology practices broadly.

References

- [1] Kate Alexander. Reducing error in radiographic interpretation. *The Canadian Veterinary Journal*, 51(5):533, 2010.
- [2] Pattaramanee Arsomngern, Nichakorn Numcharoenpinij, Jitpinun Piriataravet, Wutiwong Teerapan, Woranich Hinthong, and Phond Phunchongharn. Computer-aided diagnosis for lung lesion in companion animals from x-ray images using deep learning techniques. In *2019 IEEE 10th International Conference on Awareness Science and Technology (iCAST)*, pages 1–6. IEEE, 2019.
- [3] American Veterinary Medical Association et al. Veterinary Specialists, 2018. Accessed: 2021-9-28.
- [4] Tommaso Banzato, Marek Wodzinski, Silvia Burti, Valentina Longhin Osti, Valentina Rossoni, Manfredo Atzori, and Alessandro Zotti. Automatic classification of canine thoracic radiographs using deep learning. *Scientific Reports*, 11(1):1–8, 2021.
- [5] Leonard Berlin. Accuracy of diagnostic procedures: has it improved over the past five decades? *American Journal of Roentgenology*, 188(5):1173–1178, 2007.
- [6] Nicholas Bien, Pranav Rajpurkar, Robyn L Ball, Jeremy Irvin, Allison Park, Erik Jones, Michael Bereket, Bhavik N Patel, Kristen W Yeom, Katie Shpanskaya, et al. Deep-learning-assisted diagnosis for knee magnetic resonance imaging: development and retrospective validation of mrnet. *PLoS medicine*, 15(11):e1002699, 2018.
- [7] Michael A Bruno, Eric A Walker, and Hani H Abujudeh. Understanding and confronting our mistakes: the epidemiology of error in radiology and strategies for error reduction. *Radiographics*, 35(6):1668–1676, 2015.
- [8] Matteo Cafasso. Clips python bindings, 2021. Accessed: 2021-06-01.
- [9] Joseph Paul Cohen, Paul Bertin, and Vincent Frappier. Chester: A web delivered locally computed chest x-ray disease prediction system. *arXiv preprint arXiv:1901.11210*, 2019.
- [10] Ekin D Cubuk, Barret Zoph, Jonathon Shlens, and Quoc V Le. Randaugment: Practical automated data augmentation with a reduced search space. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, pages 702–703, 2020.

-
- [11] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*, pages 248–255. Ieee, 2009.
- [12] docker. Compose is a tool for defining and running multi-container docker applications, 2021. Accessed: 2021-06-01.
- [13] Vincent Driessen. Rq (redis queue) is a simple python library for queuing jobs and processing them in the background with workers, 2021. Accessed: 2021-06-01.
- [14] C Noelle Driver, Bradley S Bowles, Brian J Bartholmai, and Alexandra J Greenberg-Worisek. Artificial intelligence in radiology: a call for thoughtful application. *Clinical and translational science*, 13(2):216, 2020.
- [15] Jared A Dunnmon, Darvin Yi, Curtis P Langlotz, Christopher Ré, Daniel L Rubin, and Matthew P Lungren. Assessment of convolutional neural networks for automated classification of chest radiographs. *Radiology*, 290(2):537–544, 2019.
- [16] Samuel G Finlayson, Adarsh Subbaswamy, Karandeep Singh, John Bowers, Annabel Kupke, Jonathan Zittrain, Isaac S Kohane, and Suchi Saria. The clinician and dataset shift in artificial intelligence. *The New England Journal of Medicine*, pages 283–286, 2020.
- [17] João Gama, Indrė Žliobaitė, Albert Bifet, Mykola Pechenizkiy, and Abdelhamid Bouchachia. A survey on concept drift adaptation. *ACM Comput. Surv.*, 46(4), March 2014.
- [18] ME Gatt, G Spectre, O Paltiel, N Hiller, and R Stalnikowicz. Chest radiographs in the emergency department: is the radiologist really necessary? *Postgraduate medical journal*, 79(930):214–217, 2003.
- [19] Gao Huang, Zhuang Liu, Laurens Van Der Maaten, and Kilian Q Weinberger. Densely connected convolutional networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4700–4708, 2017.
- [20] Shih-Cheng Huang, Tanay Kothari, Imon Banerjee, Chris Chute, Robyn L Ball, Norah Borus, Andrew Huang, Bhavik N Patel, Pranav Rajpurkar, Jeremy Irvin, et al. Penet—a scalable deep-learning model for automated diagnosis of pulmonary embolism using volumetric ct imaging. *NPJ digital medicine*, 3(1):1–9, 2020.
- [21] Solomon Hykes. Use docker containers to build, share and run your applications, 2021. Accessed: 2021-06-01.
- [22] Jeremy Irvin, Pranav Rajpurkar, Michael Ko, Yifan Yu, Silvana Ciurea-Ilcus, Chris Chute, Henrik Marklund, Behzad Haghgoo, Robyn Ball, Katie Shpanskaya, et al. Chexpert: A large chest radiograph dataset with uncertainty labels and expert comparison. In *Proceedings of the AAAI conference on artificial intelligence*, volume 33, pages 590–597, 2019.
- [23] A Johnson, M Lungren, Y Peng, Z Lu, R Mark, S Berkowitz, and S Horng. Mimic-cxr-jpg-chest radiographs with structured labels (version 2.0. 0). physionet, 2019.
- [24] Gary D. Riley Joseph C. Giarratano. *Expert Systems: Principles and Programming*. Course Technology, 2004.

- [25] David B Larson, Matthew C Chen, Matthew P Lungren, Safwan S Halabi, Nicholas V Stence, and Curtis P Langlotz. Performance of a deep-learning neural network model in assessing skeletal maturity on pediatric hand radiographs. *Radiology*, 287(1):313–322, 2018.
- [26] Yuncheng Li, Jianchao Yang, Yale Song, Liangliang Cao, Jiebo Luo, and Li-Jia Li. Learning from noisy labels with distillation. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 1910–1918, 2017.
- [27] MongoDB. *Mongodb : The application data platform*, 2021. Accessed: 2021-06-01.
- [28] Inc. Parallels. *rq-dashboard is a general purpose, lightweight, flask-based web front-end to monitor your rq queues, jobs, and workers in realtime*, 2021. Accessed: 2021-06-01.
- [29] Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, et al. Pytorch: An imperative style, high-performance deep learning library. *Advances in neural information processing systems*, 32:8026–8037, 2019.
- [30] Yifan Peng, Xiaosong Wang, Le Lu, Mohammadhadi Bagheri, Ronald Summers, and Zhiyong Lu. Negbio: a high-performance tool for negation and uncertainty detection in radiology reports. *AMIA Summits on Translational Science Proceedings*, 2018:188, 2018.
- [31] Felipe Pezoa, Juan L Reutter, Fernando Suarez, Martín Ugarte, and Domagoj Vrgoč. Foundations of json schema. In *Proceedings of the 25th International Conference on World Wide Web*, pages 263–273. International World Wide Web Conferences Steering Committee, 2016.
- [32] Stephan Rabanser, Stephan Günnemann, and Zachary C. Lipton. Failing loudly: An empirical study of methods for detecting dataset shift, 2019.
- [33] Pranav Rajpurkar, Jeremy Irvin, Robyn L Ball, Kaylie Zhu, Brandon Yang, Hershel Mehta, Tony Duan, Daisy Ding, Aarti Bagul, Curtis P Langlotz, et al. Deep learning for chest radiograph diagnosis: A retrospective comparison of the chexnext algorithm to practicing radiologists. *PLoS medicine*, 15(11):e1002686, 2018.
- [34] Pranav Rajpurkar, Jeremy Irvin, Kaylie Zhu, Brandon Yang, Hershel Mehta, Tony Duan, Daisy Ding, Aarti Bagul, Curtis Langlotz, Katie Shpanskaya, et al. Chexnet: Radiologist-level pneumonia detection on chest x-rays with deep learning. *arXiv preprint arXiv:1711.05225*, 2017.
- [35] Pranav Rajpurkar, Allison Park, Jeremy Irvin, Chris Chute, Michael Bereket, Domenico Mastroiaca, Curtis P Langlotz, Matthew P Lungren, Andrew Y Ng, and Bhavik N Patel. Appendixnet: deep learning for diagnosis of appendicitis from a small dataset of ct exams using video pretraining. *Scientific reports*, 10(1):1–7, 2020.
- [36] Sebastián Ramírez. *Fastapi framework, high performance, easy to learn, fast to code, ready for production*, 2021. Accessed: 2021-06-01.
- [37] Redis. *Redis is an open source (bsd licensed), in-memory data structure store, used as a database, cache, and message broker*, 2021. Accessed: 2021-06-01.

-
- [38] Redis. The redisjson module provides in-memory manipulation of json documents at high velocity and volume, 2021. Accessed: 2021-06-01.
- [39] David J Reese, Eric M Green, Lisa J Zekas, Jane E Flores, Lawrence N Hill, Matthew D Winter, Clifford R Berry, and Norman Ackerman. Intra-and interobserver variability of board-certified veterinary radiologists and veterinary general practitioners for pulmonary nodule detection in standard and inverted display mode images of digital thoracic radiographs of dogs. *Journal of the American Veterinary Medical Association*, 238(8):998–1003, 2011.
- [40] Christian Szegedy, Sergey Ioffe, Vincent Vanhoucke, and Alex Alemi. Inception-v4, inception-resnet and the impact of residual connections on learning. *arXiv preprint arXiv:1602.07261*, 2016.
- [41] Mingxing Tan and Quoc V Le. Efficientnet: Rethinking model scaling for convolutional neural networks. *arXiv preprint arXiv:1905.11946*, 2019.
- [42] Arnaud Van Looveren, Giovanni Vacanti, Janis Klaise, Alexandru Coca, and Oliver Cobb. *Alibi Detect: Algorithms for outlier, adversarial and drift detection*, 2019.
- [43] Stephen Waite, Jinel Scott, Brian Gale, Travis Fuchs, Srinivas Kolla, and Deborah Reede. Interpretive error in radiology. *American Journal of Roentgenology*, 208(4):739–749, 2017.
- [44] Qizhe Xie, Minh-Thang Luong, Eduard Hovy, and Quoc V Le. Self-training with noisy student improves imagenet classification. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10687–10698, 2020.
- [45] William J Youden. Index for rating diagnostic tests. *Cancer*, 3(1):32–35, 1950.