



OPEN

Improving the classification of veterinary thoracic radiographs through inter-species and inter-pathology self-supervised pre-training of deep learning models

Weronika Celnia^{1,2}, Marek Wodziński^{1,2}, Artur Jurgas^{1,2}, Silvia Burti³, Alessandro Zotti³, Manfredo Atzori^{1,4,7}, Henning Müller^{1,5,6} & Tommaso Banzato³

The analysis of veterinary radiographic imaging data is an essential step in the diagnosis of many thoracic lesions. Given the limited time that physicians can devote to a single patient, it would be valuable to implement an automated system to help clinicians make faster but still accurate diagnoses. Currently, most of such systems are based on supervised deep learning approaches. However, the problem with these solutions is that they need a large database of labeled data. Access to such data is often limited, as it requires a great investment of both time and money. Therefore, in this work we present a solution that allows higher classification scores to be obtained using knowledge transfer from inter-species and inter-pathology self-supervised learning methods. Before training the network for classification, pretraining of the model was performed using self-supervised learning approaches on publicly available unlabeled radiographic data of human and dog images, which allowed substantially increasing the number of images for this phase. The self-supervised learning approaches included the Beta Variational Autoencoder, the Soft-Introspective Variational Autoencoder, and a Simple Framework for Contrastive Learning of Visual Representations. After the initial pretraining, fine-tuning was performed for the collected veterinary dataset using 20% of the available data. Next, a latent space exploration was performed for each model after which the encoding part of the model was fine-tuned again, this time in a supervised manner for classification. Simple Framework for Contrastive Learning of Visual Representations proved to be the most beneficial pretraining method. Therefore, it was for this method that experiments with various fine-tuning methods were carried out. We achieved a mean ROC AUC score of 0.77 and 0.66, respectively, for the laterolateral and dorsoventral projection datasets. The results show significant improvement compared to using the model without any pretraining approach.

Radiology is, by far, the most commonly used diagnostic imaging tool used for the investigation of thoracic pathology in cats and dogs¹. The increasing demand for high quality veterinary services has led to an increase in the demand for fast and accurate interpretation of diagnostic images². On the other hand, the reduced availability of trained veterinary radiologists struggles to meet the demand for accurate interpretation of diagnostic

¹University of Applied Sciences Western Switzerland (HES-SO), 3960 Sierre, Switzerland. ²Faculty of Electrical Engineering, Automatics, Computer Science and Biomedical Engineering, AGH University of Krakow, 30059 Kraków, Poland. ³Department of Animal Medicine, Productions, and Health, Legnaro (PD), University of Padua, 35020 Padua, Italy. ⁴Department of Neuroscience, University of Padua, 35121 Padua, IT, Italy. ⁵Medical Faculty, University of Geneva, 1206 Geneva, Switzerland. ⁶The Sense Research and Innovation Insitute, 1950 Sion, Switzerland. ⁷Padova Neuroscience Center, University of Padova, Via Orus 2/B, 35129 Padova, Italy. ✉email: wcelniak@agh.edu.pl

images³. In such a scenario, the role of AI-assisted tools is gaining popularity also in veterinary medicine, with an increasing number of publications on this topic^{4–11}.

The ever-increasing popularity of deep learning models due to their high performance in various domains has led to an increased interest in such methods in the field of computer-aided diagnosis. A number of studies exploring the use of these modern approaches in various fields of medicine, including medical imaging have been presented¹². Emerging solutions seek to create systems for lesion segmentation or classification in images from different imaging modalities. Some focus on detecting one specific disorder, while others attempt to annotate all of the lesions present in an image. A classification scenario in which one image can belong to several mutually non-exclusive classes is called multi-label classification. The presence of several conditions in a single examination is a common situation in medical imaging. Therefore, numerous methods dedicated to multi-label classification have already been proposed¹³. Majority of the methods use models initially pretrained on ImageNet database. The database consists of more than 14 million annotated images presenting everyday objects. The model pretraining usually improves the results but it is both computationally expensive and time-consuming since the architectures dedicated to computer vision are deeper and wider than the ones required in medical imaging. Additionally, the features presented in ImageNet database are not directly related to medical images which may reduce the impact of pretraining. Also a recently discussed topic is the carbon footprint left while training a model. Since ImageNet is a large dataset, the pretraining process takes a lot of time and memory, which results in high energy consumption. Therefore, taking into account all of the above-mentioned problems, in this work we present a different pretraining approach. We aim to improve the performance of models for classification of veterinary thoracic radiographs by self-supervised learning. We replace the standard ImageNet database by open, unlabelled inter-specie and inter-pathology XRay datasets. Using self-supervised learning not only allows to initialize the model using the same imaging modality but also does not require any expensive annotations for the pretraining.

The self-supervised setup includes variational encoders and contrastive learning frameworks. Variational encoders (VAEs) are neural networks consisting of an encoding and a decoding part, just like classical autoencoders. However, the main difference is that they aim to regularise the latent space. Thus, the encoder does not directly return a feature vector but instead the parameters of a desired distribution. As described in Ref.¹⁴ to balance reconstruction quality with latent space organization additional hyperparameter beta is needed. This approach allows for successful disentanglement of latent space and stable training. However, classical VAEs suffer from blurry image reconstruction. Several different solutions have been described in the literature to address this problem^{15–20}. One of the most successful methods described to date is Intro-VAE²¹, which is a combination of VAE and Generative Adversarial Network (GAN). It involves training the variational encoder in an introspective manner. During training the encoder part of the network minimizes the divergence between reconstruction and input images for real data samples and maximizes it for the generated samples, while the generator model minimizes the divergence of the generated samples. The results achievable with this method are promising, but the training process is difficult. The value above which the fake sample no longer affects the loss function is a manually tuneable hyperparameter. As a result, it is difficult to stabilize the network training. The problem was addressed by authors of SoftIntro-VAE²². The researchers proposed the soft exponential function over the evidence lower bound (ELBO) instead of the hard threshold, which is a lot easier to optimize and more stable during training.

Another approach to self-supervised learning is contrastive learning. The idea behind this group of methods is relatively simple: the most similar images are separated by the smallest distance while the distance between all other images is being maximised. In Ref.²³ authors proposed a Simple Framework for Contrastive Learning of Visual Representations (SimCLR). The algorithm works as follow: (i) on a batch of images data augmentation techniques are applied to create two versions of each image in a batch, (ii) the augmented images are given as an input to CNN network to obtain a 1D feature vector, (iii) a mapping with a multilayer perceptron that acts as a projection head of the encoder is performed. During training process the InfoNCE loss is used as a contrastive loss. In Ref.²⁴ authors presented a mechanism for building dynamic dictionaries for contrastive learning called Momentum Contrast or MoCo. The model consist of two encoders, an encoder and momentum encoder that produce vector representations of data: queries and keys respectively. A key encoder works as a momentum-based moving average of the query encoder for maintaining consistency. During a training step positive pairs of images are constructed from queries and keys from current mini-batch meanwhile negative pairs consist of query from current mini-batch and keys from previous mini batch that are stored in queue. Another recently introduced self-supervised learning method is called Bootstrap Your Own Latent (BYOL)²⁵. In this method we also create two augmented images and then we pass them as input one to the online network and the other to the target network. During the training process, the prediction outputs from the two networks are compared and the loss between them is calculated. At the same time, the target network parameters are updated so that they represent an exponential moving average of the online network parameters. The most notable difference between BYOL and other state-of-the-art methods is that it does not rely on the use of negative image pairs and, at the same time, allows comparably high results to be achieved. A solution that also does not require pairwise comparisons at all is the Swapping Assignments Between Views (SwaV) method²⁶. It combines clustering with contrastive learning to make it more computationally efficient. In this approach, the two augmented versions of the image are passed to the encoders and the resulting features, rather than being compared directly, are mapped to their nearest neighbour in the set of clusters, thus producing 'codes' that are then used to predict the features of the opposite image.

The advantage that all of the previously mentioned self-supervised learning methods bring is the use of data without the need for annotation, which is far more difficult to access than raw data. According to the National Health Service England, more than 20 million radiographs were taken in October 2020–October 2021 alone in England²⁷. Given how many of these exams are performed globally each year, using some of the collected data to pretrain self-supervised learning models seems like a beneficial solution, as it eliminates the obvious domain shift that occurs when pretraining on databases such as ImageNet where the features of the images are distinctly

different from those found in radiographic images. The impact of pretraining models on human X-ray images from one database and the application of these pretrained models to a different database have already been elucidated in Ref.²⁸. However, the possibility of inter-species knowledge transfer has yet to be fully described. Despite anatomical differences in human and veterinary images, the features present in both types of data are similar. The use of self-supervised learning models for pretraining provides an option to improve the results obtained by the deep learning model without the need for a larger database of annotated data. This is important because the annotation process, especially for medical data that requires specialised knowledge, is a labour-intensive and therefore also time- and cost-intensive task.

Contribution

In this work we propose and compare self-supervised pretraining strategies on open XRay datasets to improve the performance on the downstream task related to classification of veterinary thoracic radiographs. More concretely the study: (i) compares the effectiveness of different self-supervised learning strategies pretrained on open inter-specie and inter-pathology X-Ray datasets transferred to a small, annotated veterinary dataset, (ii) explores the latent spaces for various self-supervised learning method, prior and after fine-tuning, (iii) develops a system for multi-label classification of veterinary thoracic radiographs. The study confirms that the self-supervised pretraining improves the classification accuracy without any additional data labeling.

Results

Database

The database used for pretraining the self-supervised models consists of over 600k images from 23 open XRay datasets of varying size and quality containing different anatomical structures. The increased research interest in respiratory diseases in recent years due to the joint attempts to combat the Covid-19 pandemic and its aftermath has facilitated the availability of human chest XRay images, that is why most of the collected sets contain images of this type. The detailed overview of the open datasets is provided in Table 1 and example images are presented in Fig. 1.

Database name	Description	Number of images
PadChest ²⁹	Annotated high resolution chest x-ray images	160,000
MURA ³⁰	Large dataset of musculoskeletal radiographs	40,561
ChestX-ray14 ³¹	Frontal orientation chest X-Ray images	112,120
IRMA ³²	Dataset containing anonymised X-ray images that reflect different age groups, genders, exposure positions and pathologies.	14,410
TCGA-BLCA ³³	The Cancer Genome Atlas Urothelial Bladder Carcinoma X-Ray images data	74
COVID-19-NY-SBU ³⁴	Collection of chest X-Ray images from Covid-19 patients	902
CPTAC-PDA ³⁵	Images from the Clinical Proteomic Tumor Analysis Consortium containing imaging data of patients with pancreatic ductal adenocarcinoma	7
CPTAC-LSCC ³⁶	Collection of chest X-Ray images from Clinical Proteomic Tumor Analysis Consortium containing images with lung squamous cell carcinoma	2
MIDRC-RICORD-1C ³⁷	Imaging data from patients who suffered from Covid-19	296
COVID-19-AR ³⁸	Radiology imaging from Covid-19 positive population	192
CPTAC-UCEC ³⁹	Imaging data from National Cancer Institute's Clinical Proteomic Tumor Analysis Consortium taken from patient with uterine corpus endometrial carcinoma.	1
ACRIN-NSCLC-FDG-PET ⁴⁰	Data from multicenter clinical trial hold by American College of Radiology Imaging Network	20
LIDC-IDRI ⁴¹	Dataset collected to help with developing computer-aided methods of cancer detection	109
Chest X-Ray Images (Pneumonia) ⁴²	Dataset of chest X-rays from both healthy and pneumonia patients	5856
UNIFESP X-ray ⁴³	DICOM image database containing radiographs displaying 22 different parts of the body	2480
Aseptic Loose Hip Implant X-Ray ⁴⁴	A collection of radiographs including at least a stem and a cup of the hip implant collected from available sources such as medical journals and radiology websites	200
Shoulder Implant X-Ray ^{45,46}	X-Ray images of patients with shoulder implants collected at Biomedical Image and Data Analysis Lab in San Francisco State University	597
Osteoporosis Knee X-ray ⁴⁷	Collection of radiographs of the knee of healthy patients as well as those with osteoporosis	372
Shoulder X-ray	Shoulders X-Ray images from patients at Sun Yat-sen Memorial Hospital of Sun Yat-sen University, Guangzhou	1049
COVID-19 Radiography Database ^{48,49}	Chest X-ray images database consisted of covid-19 positive cases as well as pneumonia and healthy lung images	33,920
Bone Age ⁵⁰	Dataset from Stanford University and the University of Colorado consisting of hand radiographs	14,236
CheXpert ⁵¹	Large dataset with presence of 14 different radiology observations	224,316
Radiographic Dataset for VHS determination learning process ⁵²	Dataset of canine laterolateral thoracic radiographs	156
	Total number of images	511,876

Table 1. Pre-training database overview with the names of the individual data sets, a brief description and the number of images it contains.



Figure 1. Example images from open X-Ray datasets.

The internal, small database containing veterinary thoracic radiographs was used to fine-tune the pretrained models. It consists of 17869 annotated images they were acquired using scanners by 3 different manufacturers: Kodak, Isomedic and Fujifilm. Due to the poor quality 5453 images were excluded. The radiographs were acquired in laterolateral (LL), ventrodorsal (VD) and dorsoventral (DV) projections. The example images in different projections are presented in Fig. 2. A total of 20 different radiology findings were present in the dataset. Number of samples for each class varied significantly from 12 for edema to 2603 for cardiomegaly and 6047 for unremarkable (this class consists of images in which no abnormalities were detected), which were the most frequent tags. The exact distribution of the number of cases for each class is shown in Table 2. Good classification results and satisfactory generalisation are not achievable if the number of samples for each class is not sufficient, which is why we selected the 9 most numerous lesions as the target for classification. The division into training, validation and test set was made with a ratio: 0.75:0.05:0.25 for each class.

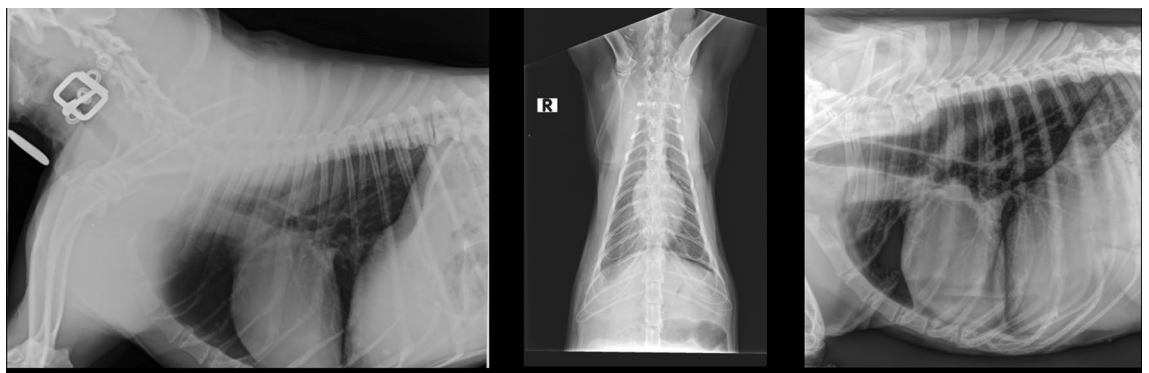


Figure 2. Example images from internal veterinary X-Ray dataset.

Radiographic finding	Number of samples
Cardiomegaly	2603
Alveolar pattern	1818
Bronchial pattern	1308
Interstitial pattern	997
Pleural efusion	893
Mass	735
Pneumothorax	326
Megaesophagus	230
Unremarkable	6047
Pneumoderma	121
Foreign body	114
Hernia	99
Suture	76
Fracture	64
Pneumomediastinum	47
Tracheal collapse	44
Edema	12

Table 2. Veterinary data overview.

Experimental setup

The described method was implemented using Python 3.9.7 with PyTorch framework and basic, publicly available packages, including numpy, pandas, sklearn, torchvision, pydicom or kornia. The workstation on which the analysis was performed includes four NVIDIA Tesla V100 GPUs and an Intel Xeon E5-2698 v4 2.2 GHz processor.

Self-supervised learning

For each of the pretrained models we used Uniform Manifold Approximation and Projection for Dimension Reduction (UMAP) to create visualizations of the latent spaces. We have noticed that images taken in 'DV'/'VD' and 'LL' projection are clearly separated as shown in Fig. 3. Figure 4 present distribution of images representing different classes of lesions in the latent space. The need for further finetuning for classification purposes is apparent as for neither of the models latent space is clearly organised.

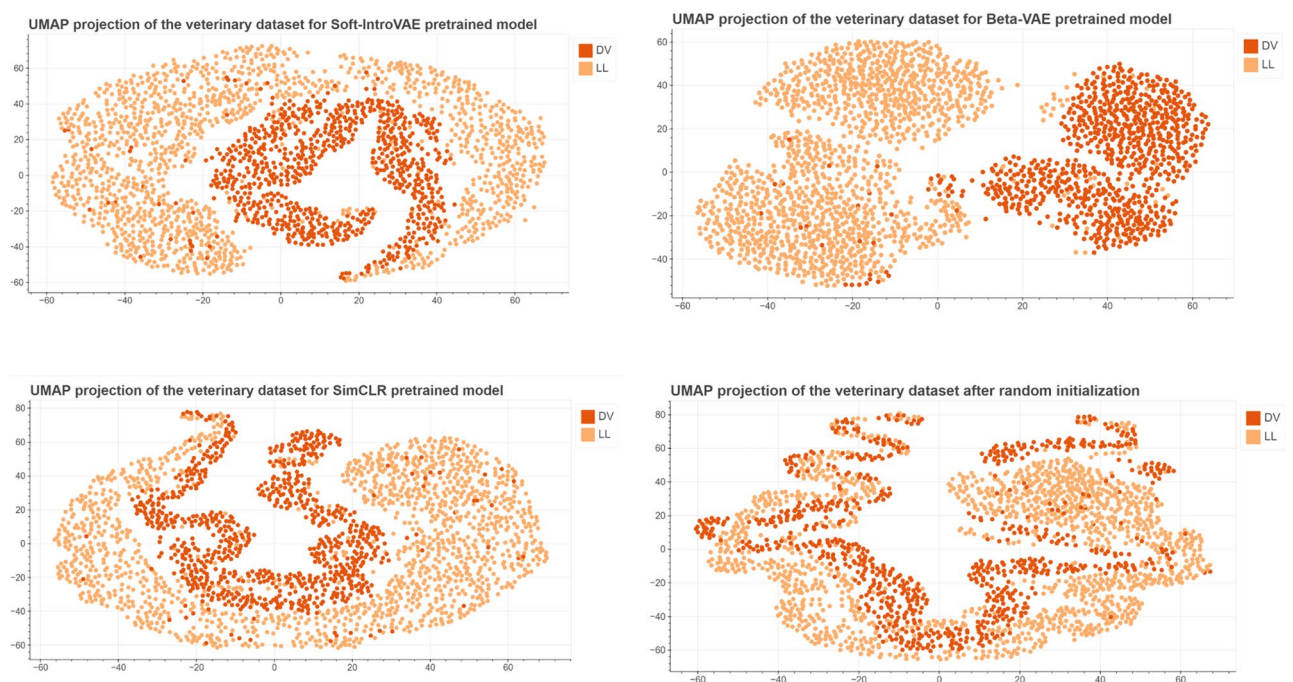


Figure 3. UMAP visualizations of distribution of images with different projections in the latent spaces for all used pretraining approaches: VAE, SoftIntro-VAE and SimCLR.

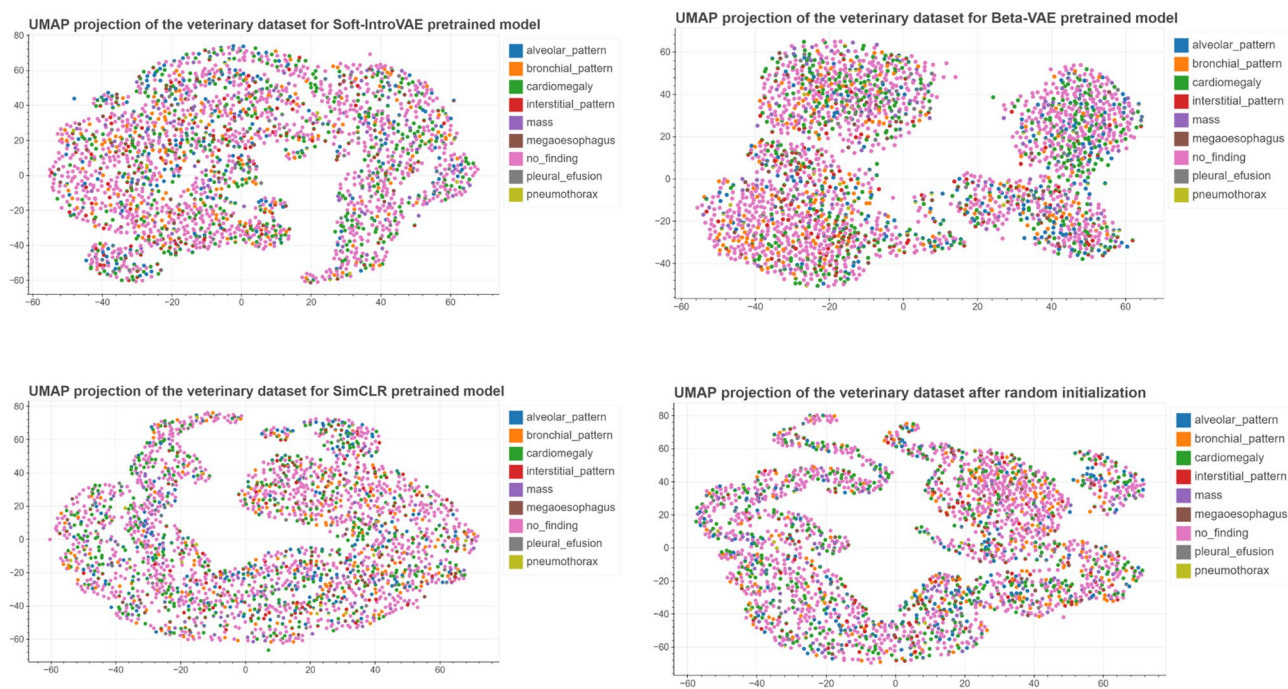


Figure 4. UMAP visualizations of distribution of images belonging to different classes of lesions in the latent spaces for all used pretraining approaches: VAE, SoftIntro-VAE and SimCLR.

Classification results

As mentioned above, the radiological images are clearly separated in the latent space in terms of projection. Given this as well as the significant difference in image number in the 'DV' and 'LL' projections, the model evaluation was carried out separately for image sets belonging to the two projection classes. In order to assess which of the pre-training approaches brings the most benefit in terms of using an encoder for classification, we performed a simple test: in each of the models we unfroze last layer and added additional linear layer serving as a classifier. Comparison of classification results obtained with different pretraining approaches is presented in Table 3. The best results were achieved with a model pretrained using contrastive learning framework, providing mean ROC AUC 0.74 for 'LL' projection dataset and 0.66 for 'DV' projection. In the next step we proceeded to try to get a better result using other finetuning methods only for the best of the models. Results of all conducted studies are presented in Table 4. The best results for both of the datasets were achieved while training model with 9 binary classifiers (one for each class) and last layer unfrozen. This method allowed us to obtain significant improvement in obtained results when compared to model without pretraining. Figure 5 present distribution of images representing different classes of lesions in the latent space after fine-tuning process, the distribution of images in the latent spaces for each class separately can be found as Supplementary Figs. S1–S9. Detailed results of classification consisting of ROC AUC score alongside sensitivity, specificity, PLR and NLR for each class in DV and LL projection datasets are presented in Tables 5 and 6. Additionally, to aid in the interpretation of the biological relevance of the obtained results, Fig. 6 showcases GradCAM image examples of different classes from both projections. It can be observed that the activation maps are spatially correlated with the classified lesions.

Discussion

An important consequence of this article is that it demonstrates how, thanks to approaches that do not require human labels such as self-supervised learning, it is possible to benefit from inter-specie and inter-pathology data to improve deep learning models for medical applications. This result paves the way to models that could

Model	AUC	
	Dorsoventral	Laterolateral
Beta-VAE	0.61	0.71
Soft-IntroVAE	0.64	0.73
SimCLR	0.66	0.74
Baseline	0.59	0.69

Table 3. Comparison of different pretraining approaches for classification of lesions belonging to 9 lesion classes. Highest achieved scores are in bold.

Model	AUC	
	DV	LL
1 classifier		
Last layer unfrozen	0.66	0.74
Last two layers unfrozen	0.65	0.76
All layer frozen with additional linear layer	0.63	0.72
9 binary classifiers		
Last layer unfrozen	0.66	0.77
Last two layers unfrozen	0.64	0.76
All layer frozen with additional linear layer	0.61	0.71
SVM classification	0.63	0.70

Table 4. Comparison of different finetuning approaches. Highest achieved scores are in bold.

Lesion	AUC	Sensitivity	Specificity	PLR	NLR
Cardiomegaly	0.77	0.78	0.58	1.86	0.37
Alveolar pattern	0.78	0.74	0.70	2.48	0.38
Bronchial pattern	0.63	0.68	0.50	1.36	0.64
Interstitial pattern	0.71	0.76	0.59	1.83	0.42
Pleural efusion	0.93	0.96	0.64	2.63	0.07
Mass	0.65	0.65	0.60	1.62	0.59
Pneumothorax	0.91	0.77	0.85	5.14	0.27
Megaesophagus	0.70	0.54	0.75	2.12	0.62
Unremarkable	0.73	0.83	0.50	1.66	0.34

Table 5. Results of classification with the best model (model with 9 binary classifiers and last layer unfrozen) for each class on dataset with 'LL' projection. Highest achieved scores are in bold.

Lesion	AUC	Sensitivity	Specificity	PLR	NLR
Cardiomegaly	0.58	0.81	0.28	1.02	0.93
Alveolar pattern	0.66	0.80	0.41	1.35	0.49
Bronchial pattern	0.54	0.82	0.27	1.50	0.40
Interstitial pattern	0.68	0.87	0.45	2.37	0.21
Pleural efusion	0.72	0.69	0.63	1.07	0.88
Mass	0.50	0.60	0.35	1.57	1.12
Pneumothorax	0.70	0.63	0.61	1.56	0.60
Megaesophagus	0.73	0.82	0.59	2.02	0.30
Unremarkable	0.61	0.92	0.21	1.15	0.41

Table 6. Results of classification with the best model (model with 9 binary classifiers and last layer unfrozen) for each class on dataset with dorsoventral ('DV') projection. Highest achieved scores are in bold.

benefit of increasingly available data on different species, allowing to create increasingly reliable models and potentially shifting the paradigm from classification to unsupervised inter-specie knowledge extraction problems.

Achieved results suggest that features learned from images presenting particular specie can be used to improve classification result for another specie. These clinical insights could potentially have significant repercussions. One of the potential applications of the presented self-supervised learning model could be in the diagnosis of those diseases that are rare in humans but have a relatively higher incidence in animals⁵³. By pre-training the model on a multi-species dataset and then fine-tuning it on a much smaller human dataset, it could be used to accurately diagnose diseases that some dog breeds are more prone to, such as histiocytic sarcoma in Bernese mountain dogs, gliomas in boxers and juvenile dilated cardiomyopathy in Portuguese water dogs. The potential to employ a cross-species strategy could revolutionize the relationship between human and animal medical care, thereby progressing towards a One-Health medicine.

Using more balanced dataset for fine-tuning self-supervised model would probably lead to achieving better performance, as we can see significantly difference in results for 'DV' and 'LL' projection datasets. Also, to improve the performance of the presented approach, other encoder architectures can be implemented, for

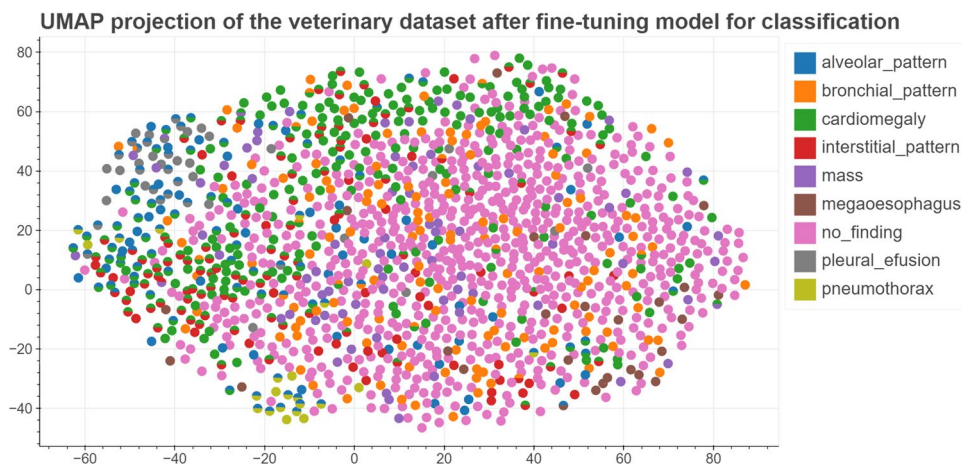


Figure 5. Latent space after fine-tuning the model for classification.

example based on Vision Transformer (ViT) instead of CNN, as literature shows its promising results in image recognition tasks⁵⁴.

Methods

Data acquisition

Open, universally available databases containing human radiographs were used for the pretraining part of presented solution. Fine-tuning, on the other hand, was carried out using an internal veterinary dataset. The methods used to create the database for this study are described elsewhere⁵. Briefly, all the radiographs included in the database were evaluated by three experienced veterinary radiologist (AZ, TB, and SB), for several different radiographic abnormalities (alveolar pattern, interstitial pattern, bronchial pattern, mass, cardiomegaly, pleural effusion, pneumothorax, hernia, megaesophagus, fracture, pneumomediastinum, tracheal collapse). If no radiographic findings were evident, the image was classified as unremarkable. All the tags on the images were applied following a consensus discussion. Image quality was also assessed and only technically correct images were included in the database.

Overview

The first step in image preprocessing was to resize images to equal size, in our case: 224x224 pixels. The images were then normalised so that the intensity was in a range of 0-1. In the next step three different self-supervised architectures were pretrained using unlabelled data from open database. For this purpose, the following were selected: BetaVAE, Soft-IntroVAE and SimCLR. Each of the models was later fine-tuned using 20% of available annotated veterinary data. After fine-tuning latent space exploration using UMAP was carried out. In the latent space, a clear division was evident in terms of features such as image projection, but additional fine tuning was needed to obtain better division results for lesions visible on radiographs. Due to the possibility of several lesions being present on a single image, a multilabel classification approach had to be used. In the classification step the encoding part of the model was used which was extended with linear layer to classify features extracted by the encoder. Three approaches were compared: the introduction of one additional linear layer with 9 outputs, the addition of 9 binary classifiers, one for each lesion type, and the use of an SVM as a classifier. A brief summary of the pipeline is presented in Fig. 7.

Self-supervised learning

For all pretraining approaches we used the same network architecture. The encoder was built with 5 residual blocks with 64, 128, 256, 512 and 512 channels. Additionally for VAE-based approaches the same setting was used as a decoder part. For Beta Variational Autoencoder training beta parameter was set to 0.01 and L1 loss was chosen as a cost function. Soft-Intro VAE was trained using MSE loss function with beta neg equal to 256. During SimCLR Info Nce Loss was minimized. For all of the approaches batch size was set to 32. A comparison of how each approach was trained is presented in Fig. 8.

Classification fine-tuning

For fine-tuning purposes, we conducted a number of tests with different modifications of the classification model. However, firstly, in order to establish a baseline for comparing the different approaches, we used the encoder network without any pre-training to see what results it was able to achieve in classifying 9 selected lesions. As the dataset was highly imbalanced to increase the effectiveness of the model we implemented a balancing sampler so that the representation of each class was closer to equal. In addition, to improve performance on the training set, we also used image augmentation techniques such as vertical and horizontal rotation as well as brightness, contrast and saturation augmentations. We used these steps for each finetuning method tested. We tested 4

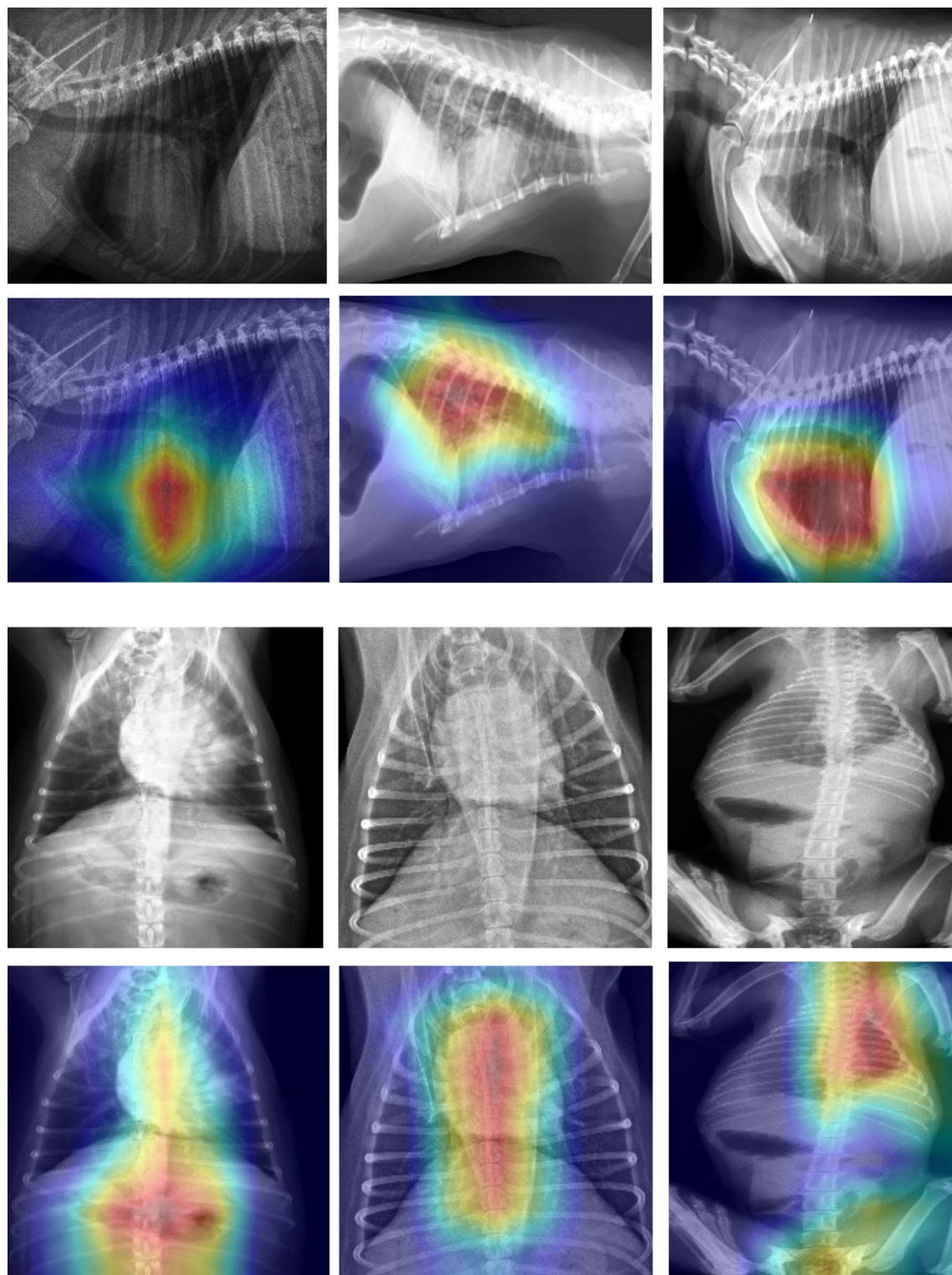


Figure 6. GradCAM image examples alongside original images. The first two rows show (from left) cardiomegaly, alveolar pattern, and bronchial pattern, while the last two rows (from left) show mass, megaesophagus, and pleural effusion.

different fine-tuning approaches including: classification with unfreezing last layer of the model, unfreezing last two layers of the model, adding additional linear layer to model with all weights frozen. Tests were performed for one classification layer with 9 outputs corresponding to each lesion as well as with 9 binary classifiers one for each lesion. For all the approaches considered, the initial value of the learning rate was set to 0.01. A lambda

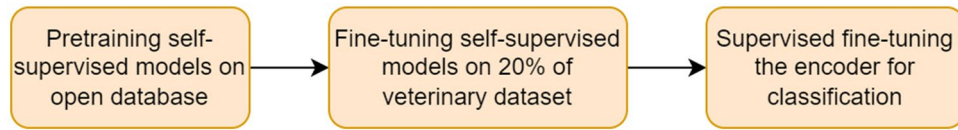


Figure 7. Brief overview of proposed approach.

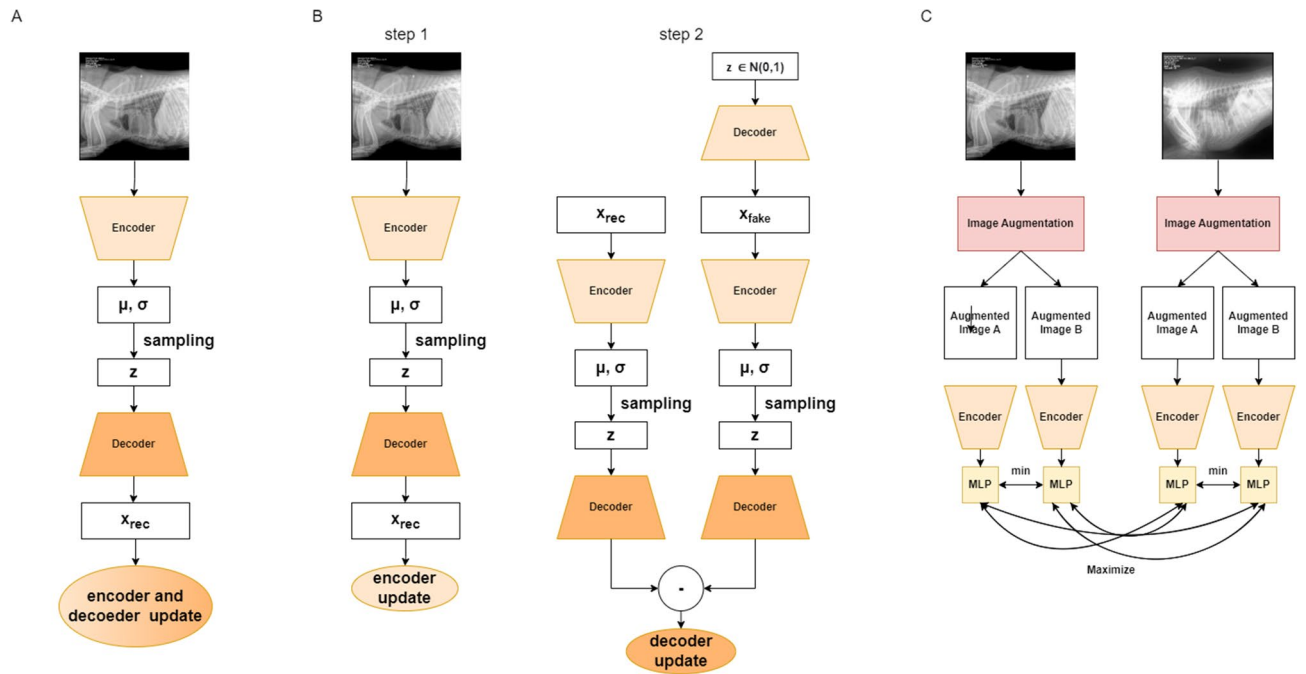


Figure 8. Block diagrams of self-supervised pretraining techniques.

learning rate scheduler was employed, causing the learning rate to decrease by a decay rate (set to 0.95) raised to the power of the current epoch's number. Additionally, we also performed classification using features extracted with pretrained model with all weights frozen and Support Vector Machine (SVM).

Result analysis

We evaluated the performance of the individual classification methods on both the data set containing the 'LL' projection and the data set containing the 'DV' projection by calculating values such as area under the receiver operating characteristic curve (AUC), sensitivity, specificity, positive likelihood ratio (PLR) and negative likelihood ratio (NLR) according to the formulas in Eqs. (1), (2), (3) and (4).

$$Sensitivity = \frac{true\ positive}{true\ positive + false\ negative} \tag{1}$$

$$Specificity = \frac{true\ negative}{true\ negative + false\ positive} \tag{2}$$

$$PLR = \frac{sensitivity}{1 - specificity} \tag{3}$$

$$NLR = \frac{1 - sensitivity}{specificity} \tag{4}$$

Conclusions

The impact of pretraining self-supervised models on the accuracy of automatic multi-label classification of lesions in veterinary radiographs was studied. The detection accuracy for each class was higher for the model that had been pretrained, compared to the model without pretraining. Further studies, with a larger dataset and

possibly a different encoder architecture, could lead to the development of a system that can aid veterinarians in their diagnosis.

Data availability

The datasets generated and/or analysed during the current study are not publicly available due to privacy restrictions but are available from the corresponding author on reasonable request.

Received: 27 January 2023; Accepted: 31 October 2023

Published online: 09 November 2023

References

- Thrall, D. E. Principles of radiographic interpretation of the thorax. In *Textbook of Veterinary Diagnostic Radiology* 7th edn (ed. Thrall, D. E.) 568–582 (Elsevier, 2018). <https://doi.org/10.1016/B978-0-323-48247-9.00040-1>.
- Wilson, D. U., Bailey, M. Q. & Craig, J. The role of artificial intelligence in clinical imaging and workflows. *Vet. Radiol. Ultrasound* **63**, 897–902. <https://doi.org/10.1111/vru.13157> (2022).
- Lungren, M. P. & Wilson, D. U. Artificial intelligence in veterinary care will be a major driving force behind AI advancements in healthcare. *Vet. Radiol. Ultrasound* **63**, 913–915. <https://doi.org/10.1111/vru.13161> (2022).
- Banzato, T., Cherubini, G. B., Atzori, M. & Zotti, A. Development of a deep convolutional neural network to predict grading of canine meningiomas from magnetic resonance images. *Vet. J.* **235**, 90–92. <https://doi.org/10.1016/j.tvjl.2018.04.001> (2018).
- Banzato, T. *et al.* Automatic classification of canine thoracic radiographs using deep learning. *Sci. Rep.* **11**, 1–8. <https://doi.org/10.1038/s41598-021-83515-3> (2021).
- Banzato, T. *et al.* An AI-based algorithm for the automatic classification of thoracic radiographs in cats. *Front. Vet. Sci.* **8**, 1–7. <https://doi.org/10.3389/fvets.2021.731936> (2021).
- Burti, S., Osti, V. L., Zotti, A. & Banzato, T. Use of deep learning to detect cardiomegaly on thoracic radiographs in dogs. *Vet. J.* **262**, 105505. <https://doi.org/10.1016/j.tvjl.2020.105505> (2020).
- Wodzinski, M. *et al.* Training deep neural networks for small and highly heterogeneous MRI datasets for cancer grading. *Proc. Annu. Int. Conf. IEEE Eng. Med. Biol. Soc. EMBS* **2020**, 1758–1761. <https://doi.org/10.1109/EMBC44109.2020.9175634> (2020).
- Boissady, E., de La Comble, A., Zhu, X. & Hespel, A. M. Artificial intelligence evaluating primary thoracic lesions has an overall lower error rate compared to veterinarians or veterinarians in conjunction with the artificial intelligence. *Vet. Radiol. Ultrasound* **61**, 619–627. <https://doi.org/10.1111/vru.12912> (2020).
- Adrien-maxence, H. *et al.* Comparison of error rates between four pretrained DenseNet convolutional neural network models and 13 board-certified veterinary radiologists when evaluating 15 labels of canine thoracic radiographs. *Vet. Radiol. Ultrasound* <https://doi.org/10.1111/vru.13069> (2022).
- Biercher, A. *et al.* Using deep learning to detect spinal cord diseases on thoracolumbar magnetic resonance images of dogs. *Front. Vet. Sci.* **8**, 1–9. <https://doi.org/10.3389/fvets.2021.721167> (2021).
- Yanase, J. & Triantaphyllou, E. A systematic survey of computer-aided diagnosis in medicine: Past and present developments. *Expert Syst. Appl.* **138**, 112821. <https://doi.org/10.1016/j.eswa.2019.112821> (2019).
- Maksoud, E. A. A., Barakat, S. & Elmogy, M. Medical images analysis based on multilabel classification. In *Machine Learning in Bio-Signal Analysis and Diagnostic Imaging* (eds Maksoud, E. A. A. *et al.*) 209–245 (Elsevier, 2019). <https://doi.org/10.1016/b978-0-12-816086-2.00009-6>.
- Higgins, I. *et al.* beta-VAE: Learning basic visual concepts with a constrained variational framework. In *International Conference on Learning Representations* (2017).
- Cao, S., Li, J., Nelson, K. P. & Kon, M. A. Coupled VAE: Improved Accuracy and Robustness of a Variational Autoencoder. arXiv e-prints [arXiv:1906.00536](https://arxiv.org/abs/1906.00536) (2019). eprint1906.00536.
- Zhao, S., Song, J. & Ermon, S. Infvae: Balancing learning and inference in variational autoencoders. *Proc. AAAI Conf. Artif. Intell.* **33**, 5885–5892 (2019).
- Cai, L., Gao, H. & Ji, S. Multi-stage variational auto-encoders for coarse-to-fine image generation. In *Proceedings of the 2019 SIAM International Conference on Data Mining* (eds Cai, L. *et al.*) 630–638 (SIAM, 2019).
- Larsen, A. B. L., Sønderby, S. K., Larochelle, H. & Winther, O. Autoencoding beyond pixels using a learned similarity metric. In *International Conference on Machine Learning* (eds Larsen, A. B. L. *et al.*) 1558–1566 (PMLR, 2016).
- Makhzani, A., Shlens, J., Jaitly, N., Goodfellow, I. & Frey, B. Adversarial autoencoders. arXiv preprint [arXiv:1511.05644](https://arxiv.org/abs/1511.05644) (2015).
- Dieng, A. B., Kim, Y., Rush, A. M. & Blei, D. M. Avoiding latent variable collapse with generative skip models. In *The 22nd International Conference on Artificial Intelligence and Statistics* (eds Dieng, A. B. *et al.*) 2397–2405 (PMLR, 2019).
- Huang, H., He, R., Sun, Z., Tan, T. *et al.* Introvae: Introspective variational autoencoders for photographic image synthesis. *Adv. Neural Inform. Process. Syst.* **31** (2018).
- Daniel, T. & Tamar, A. Soft-introvae: Analyzing and improving the introspective variational autoencoder. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 4391–4400 (2021).
- Chen, T., Kornblith, S., Norouzi, M. & Hinton, G. A simple framework for contrastive learning of visual representations. In *International Conference on Machine Learning* (eds Chen, T. *et al.*) 1597–1607 (PMLR, 2020).
- He, K., Fan, H., Wu, Y., Xie, S. & Girshick, R. Momentum contrast for unsupervised visual representation learning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 9729–9738 (2020).
- Grill, J.-B. *et al.* Bootstrap your own latent—a new approach to self-supervised learning. *Adv. Neural. Inf. Process. Syst.* **33**, 21271–21284 (2020).
- Caron, M. *et al.* Unsupervised learning of visual features by contrasting cluster assignments. *Adv. Neural. Inf. Process. Syst.* **33**, 9912–9924 (2020).
- England, N. & Improvement, N. *Diagnostic imaging dataset statistical release* (Department of Health, 2022).
- Cho, K. *et al.* CheSS: Chest x-ray pre-trained model via self-supervised contrastive learning. *J. Digit. Imaging* **36**, 902–910. <https://doi.org/10.1007/s10278-023-00782-4> (2023).
- Bustos, A., Pertusa, A., Salinas, J.-M. & de la Iglesia-Vayá, M. Padchest: A large chest x-ray image dataset with multi-label annotated reports. *Med. Image Anal.* **66**, 101797 (2020).
- Rajpurkar, P. *et al.* Mura: Large dataset for abnormality detection in musculoskeletal radiographs. arXiv preprint [arXiv:1712.06957](https://arxiv.org/abs/1712.06957) (2017).
- Wang, X. *et al.* Chestx-ray8: Hospital-scale chest x-ray database and benchmarks on weakly-supervised classification and localization of common thorax diseases. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2097–2106 (2017).
- Deserno, T. & Ott, B. 15, 363 irma images of 193 categories for imageclefmed 2009, <https://doi.org/10.18154/RWTH-2016-06143> (2009).

33. Kirk, S. et al. Radiology data from the cancer genome atlas urothelial bladder carcinoma [tcga-blca] collection. *Cancer Imaging Arch* 96–108 (2016).
34. Saltz, J. et al. Stony brook university covid-19 positive cases. The Cancer Imaging Archive (2021).
35. National Cancer Institute Clinical Proteomic Tumor Analysis Consortium (CPTAC). The clinical proteomic tumor analysis consortium pancreatic ductal adenocarcinoma collection (cptac-pda), <https://doi.org/10.7937/K9/TCIA.2018.SC20FO18> (2018).
36. National Cancer Institute Clinical Proteomic Tumor Analysis Consortium (CPTAC). The clinical proteomic tumor analysis consortium lung squamous cell carcinoma collection (cptac-lscc), <https://doi.org/10.7937/K9/TCIA.2018.6EMUB5L2> (2018).
37. Tsai, E. et al. Medical imaging data resource center (midrc) - rsna international covid radiology database (ricord) release 1c - chest x-ray, covid+ (midrc-ricord-1c), <https://doi.org/10.7937/91AH-V663> (2021).
38. Desai, S. et al. Chest imaging with clinical and genomic correlates representing a rural covid-19 positive population, <https://doi.org/10.7937/TCIA.2020.PY71-5978> (2020).
39. National Cancer Institute Clinical Proteomic Tumor Analysis Consortium (CPTAC). The clinical proteomic tumor analysis consortium uterine corpus endometrial carcinoma collection (cptac-ucec), <https://doi.org/10.7937/K9/TCIA.2018.3R3JUISW> (2019).
40. Kinahan, P., Muzi, M., Bialecki, B., Herman, B. & Coombs, L. Data from the acrin 6668 trial nslc-fdg-pet, <https://doi.org/10.7937/TCIA.2019.30ILQFCL> (2019).
41. Armato, I. I. I. et al. Data from lidc-idri <https://doi.org/10.7937/K9/TCIA.2015.LO9QL9SX> (2015).
42. Kermay, D. S. et al. Identifying medical diagnoses and treatable diseases by image-based deep learning. *Cell* 172, 1122–1131.e9. <https://doi.org/10.1016/j.cell.2018.02.010> (2018).
43. de Mattos Farina, E. M. J., Abdala, N. & Kitamura, F. C. The unifesp x-ray body part classification dataset, <https://doi.org/10.34740/KAGGLE/DSV/3399135> (2022).
44. Rahman, T. Aseptic loose hip implant x-ray database (2022).
45. Stark, M. Automatic detection and segmentation of shoulder implants in X-ray images. Ph.D. thesis, San Francisco State University San Francisco, CA, USA (2018).
46. Urban, G. et al. Classifying shoulder implants in x-ray images using deep learning. *Comput. Struct. Biotechnol. J.* 18, 967–972. <https://doi.org/10.1016/j.csbj.2020.04.005> (2020).
47. Osteoporosis knee x-ray dataset (2021).
48. Rahman, T. et al. Exploring the effect of image enhancement techniques on COVID-19 detection using chest x-ray images. *Comput. Biol. Med.* 132, 104319. <https://doi.org/10.1016/j.compbiomed.2021.104319> (2021).
49. Chowdhury, M. E. H. et al. Can AI help in screening viral and COVID-19 pneumonia?. *IEEE Access* 8, 132665–132676. <https://doi.org/10.1109/access.2020.3010287> (2020).
50. Halabi, S. S. et al. The RSNA pediatric bone age machine learning challenge. *Radiology* 290, 498–503. <https://doi.org/10.1148/radiol.2018180736> (2019).
51. Irvin, J. et al. Chexpert: A large chest radiograph dataset with uncertainty labels and expert comparison, <https://doi.org/10.48550/ARXIV.1901.07031> (2019).
52. Flores Duenas, C. A. Radiographic dataset for vhs determination learning process, <https://doi.org/10.17632/KTX4CJ55PN.1> (2020).
53. Arnold, C. Sick as a dog: How understanding canine diseases will save human lives. *Nat. Med.* 28, 1970–1973. <https://doi.org/10.1038/s41591-022-02025-5> (2022).
54. Dosovitskiy, A. et al. An image is worth 16x16 words: Transformers for image recognition at scale. arXiv preprint [arXiv:2010.11929](https://arxiv.org/abs/2010.11929) (2020).

Author contributions

W.C. conducted the experiment(s) and drafted the manuscript(s), additionally W.C. alongside M.W., A.J. developed the deep-learning model; S.B., A.Z., T.B. interpreted the radiographs, analyzed the experiment results and drafted the manuscript(s); M.A. and H.M. analyzed the experiment results and drafted the manuscript(s); all the authors reviewed the manuscript.

Competing interests

The authors declare no competing interests.

Additional information

Supplementary Information The online version contains supplementary material available at <https://doi.org/10.1038/s41598-023-46345-z>.

Correspondence and requests for materials should be addressed to W.C.

Reprints and permissions information is available at www.nature.com/reprints.

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.



Open Access This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit <http://creativecommons.org/licenses/by/4.0/>.

© The Author(s) 2023