Received: 13 April 2020 | Revised: 30 June 2020 | Accepted: 12 July 2020

DOI: 10.1111/vru.12912

ORIGINAL INVESTIGATION

Artificial intelligence evaluating primary thoracic lesions has an overall lower error rate compared to veterinarians or veterinarians in conjunction with the artificial intelligence

Emilie Boissady[1] | Alois de La Comble[1] | Xiaojuan Zhu[2] | Adrien-Maxence Hespel[3]

[1] Maisons-Alfort, France

[2] Office of Information Technology, The University of Tennessee, Knoxville, Tennessee, USA

[3] Department of Small Animal Clinical Science, University of Tennessee, Knoxville, Tennessee, USA

Correspondence
Adrien-Maxence Hespel, Department of Small Animal Clinical Science, University of Tennessee, Knoxville, Tennessee, USA.
Email: ahespel@utk.edu

Abstract

To date, deep learning technologies have provided powerful decision support systems to radiologists in human medicine. The aims of this retrospective, exploratory study were to develop and describe an artificial intelligence able to screen thoracic radiographs for primary thoracic lesions in feline and canine patients. Three deep learning networks using three different pretraining strategies to predict 15 types of primary thoracic lesions were created (including tracheal collapse, left atrial enlargement, alveolar pattern, pneumothorax, and pulmonary mass). Upon completion of pretraining, the algorithms were provided with over 22 000 thoracic veterinary radiographs for specific training. All radiographs had a report created by a board-certified veterinary radiologist used as the gold standard. The performances of all three networks were compared to one another. An additional 120 radiographs were then evaluated by three types of observers: the best performing network, veterinarians, and veterinarians aided by the network. The error rates for each of the observers was calculated as an overall and for the 15 labels and were compared using a McNemar's test. The overall error rate of the network was significantly better than the overall error rate of the veterinarians or the veterinarians aided by the network (10.7% vs 16.8% vs17.2%, $P = .001$). The network's error rate was significantly better to detect cardiac enlargement and for bronchial pattern. The current network only provides help in detecting various lesion types and does not provide a diagnosis. Based on its overall very good performance, this could be used as an aid to general practitioners while waiting for the radiologist's report.

KEYWORDS
computer vision-based decision support system, convolutional neural networks, deep learning, small animal thoracic radiology

1 | INTRODUCTION

Past decade breakthroughs in deep learning techniques brought computer vision-based decision support system to the forefront of medical imaging. The development of convolutional neural networks (CNN), able to autonomously identify complicated patterns by training on large datasets; can now provide radiologists in human medicine with computer vision algorithms that are accurate for all imaging

Abbreviations: AI, artificial intelligence; AUC, area under the curve; CNN, convolutional neural network; NLP, natural language processing.

EQUATOR network disclosure: No EQUATOR network checklist was used.

Previous presentation disclosure: None

modalities.[1–4] Convolutional neural networks also help smooth radiological workflow and absorb the increasing need for imaging exams analysis.[5] One of the greatest assets of this technology is the ability to automatically detect abnormalities contained in medical images. For instance, in human radiology, algorithms have been designed to screen thoracic radiographs for tuberculosis or to detect critical findings such as pneumothorax or fractures.[1,2,4] Similar algorithm has also shown potential for the detection of ischemic stroke on MRI.[3] For some specific tasks, machines are even already outperforming humans, such as the artificial intelligence (AI) system for breast cancer detection in digital mammography developed by Rodriguez-Ruiz et al that showed higher performance than the average of the radiologists who participated in the study.[6]

Those deep learning approaches have yet to be evaluated thoroughly in veterinary diagnostic imaging. To the best of the authors' knowledge, only one feasibility study has, to this date, been published on the subject in 2018, comparing two strategies to separate normal versus abnormal thoracic radiographs.[7] This technology has the potential to be a daily tool for the general practitioner. Such AI could provide instantaneous preliminary information while waiting for a radiology report to be generated by a specialist. In this context, an algorithm non-specialized in the detection of one specific lesion, but rather able to screen the whole radiograph for multiple common radiological patterns would be of great value. Furthermore, as many of the misdiagnosis made on radiographs are due to pattern oversight rather than misinterpretation,[8] such an AI could potentially limit the risk of false negatives by providing a systemic double reading. Thoracic radiographs are pertinent for an implementation of deep learning technics in small animal radiology, as their interpretation's paradigm can be challenging to veterinarians.[9,10] In the proof of concept study from Yoon et al,[7] the authors performed a careful image selection prior to training the AI, as only radiographs for which three radiologists' reports were in agreement were included in the study. Thus, potentially biasing the dataset toward more unequivocal cases. Furthermore, each of the classification tasks were broken down between several algorithms[7] rather than one large network.

This approach of data simplification might have been partly motivated by the challenges inherent to veterinary radiology datasets (ie, large patients diversity based on species and breed), which make it harder to perform learning tasks when working on small datasets. However, it is essential for the popularization that an AI be able to work with non-tailored images and that one single algorithm be able to recognize multiple type of lesions at once.

The aims of the current study were to develop a method for overcoming usual veterinary datasets limitations and create a unique deep neural network able to exhaustively screen thoracic radiographs of feline and canine patients.

2 | MATERIALS AND METHODS

The study was a retrospective, exploratory design. Based on a dataset of radiographs read by board-certified veterinary radiologists, we tested three pretraining strategies and built a CNN predicting 15 primary patterns trained. The best performing algorithm was then confronted with real-life condition usage by measuring its error rate and comparing it to the error rates of veterinarians and veterinarians aided by the AI on a new set of thoracic radiographs.

2.1 | Characteristics of the dataset used to train the model

Canine and feline orthogonal thoracic radiographs and their associated radiology reports generated between 2015 and 2019 were extracted retrospectively from the medical records of a referral center institution. All reports were generated by at least one board-certified veterinary radiologist (either ACVR or ACVR and ECVDI). A total of 15 780 radiographs (lateral and dorsoventral/ventrodorsal projections) from 6584 patients were collected. All radiographs were exported individually as DICOM (Digital Imaging and Communications in Medicine) and paired reports were exported as a Word document (Microsoft Office, Redmond, WA, USA). There was no preselection regarding medical conditions, age, gender, or breed. The resulting dataset contained 87.5% of dogs radiographs (ie, 12.5% of cats) and 35.4% of ventrodorsal projections (ie, 64.6% of lateral projections). This dataset was split randomly following a Bernoulli distribution of parameter 0.9 into a training set (90% of the radiographs) and a validation set (10% of the radiographs). Thus, a study from the same patient appeared only once in either the training or validation set to offer a fair comparison baseline.

2.2 | Label extraction from radiographs' reports

To avoid manual relabelization of the radiographs, the authors developed a novel, specific Natural Language Processing (NLP) algorithm using a previous software algorithm dealing with human language to produce various tasks (Python software, Python 3.8.5, Python Software Foundation, 9450 SW Gemini Dr., ECM# 90772, Beaverton, OR 97008, USA). This algorithm works on the same principle as the one described by Irvin et al[11] and extracts labels from the existing unstructured reports written by the radiologists. The author's NLP algorithm was able to identify the presence or absence of 61 thoracic labels (lesions) and 26 attributes for each label. Attributes provided additional information on the veterinary radiologist's confidence level, as well as if the label could only be identified on specific views.

Of the 61 labels identified in the veterinary radiologists' reports, we chose to keep 15 to train the final CNN, which could be further divided into cardiovascular patterns (cardiomegaly, left, and right ventricular enlargement, left atrial enlargement), lung patterns (interstitial, bronchial, alveolar, and vascular opacities), pleural space abnormalities (pneumothorax, pleural effusion, interlobar fissures), airways abnormalities (tracheal deviation or collapse) and others (mass, esophageal dilation). Those 15 labels correspond to elementary thoracic lesions (such as "enlarged left atrium" or "alveolar pattern"), as the aim of the CNN was to detect radiographic features, but not to provide

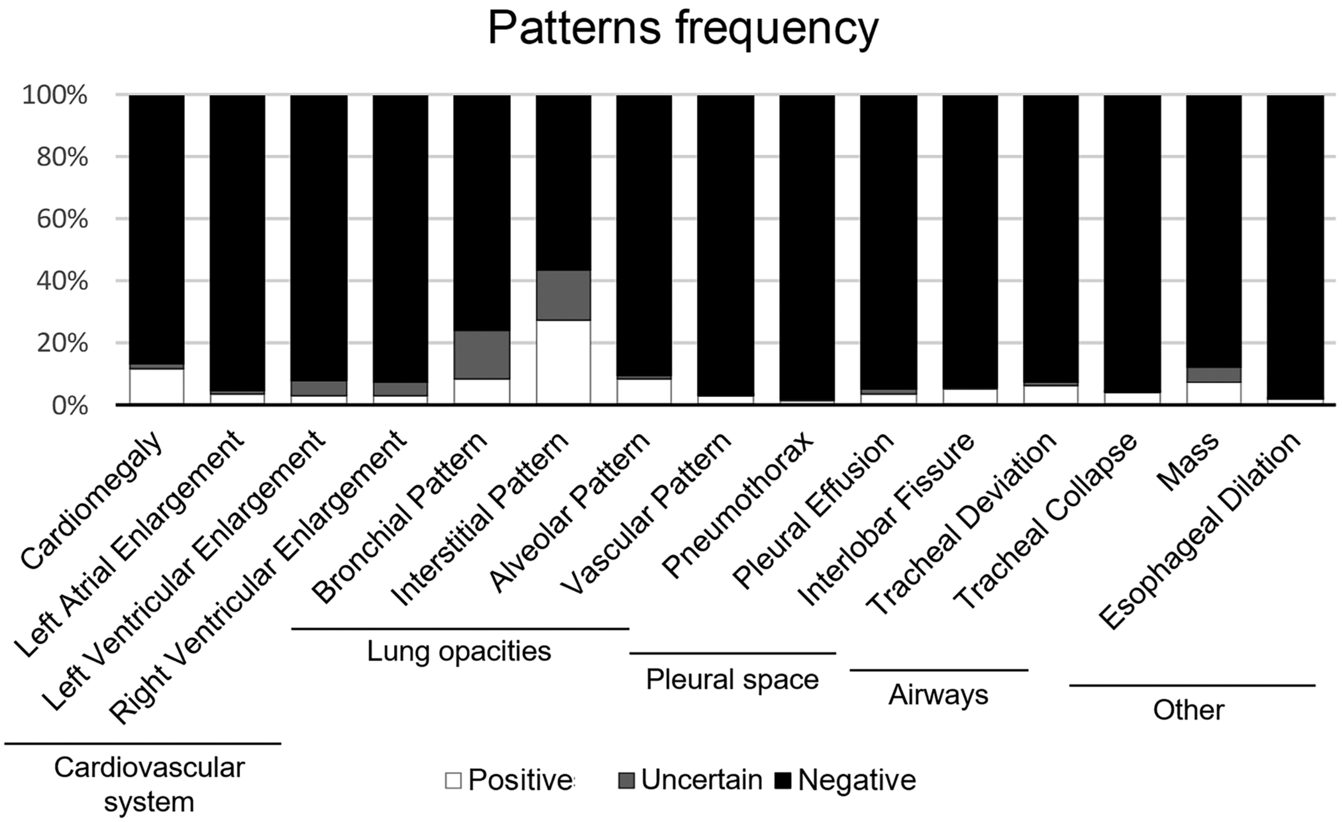

FIGURE 1 Frequency of occurrence of each of the labels throughout the overall veterinary dataset (training and validation sets). Labels were characterized as either positive, negative, or uncertain

interpretative analysis (ie, labels such as "left-sided congestive heart failure" or "aspiration pneumonia" were not included). Labels were selected based on human literature about CNNs[11,12] and according to their frequency in the dataset, but for low-frequency patterns, if they appeared easy to detect (strong radiographic characteristics), as for esophageal dilation or pneumothorax.

Post-processing was then performed on the extracted labels to verify consistency and apply correction rules. For instance, some lesions may not be visible on all projections. This is the case for tracheal collapse, dorsal, and ventral tracheal deviation on dorsoventral/ventrodorsal views. On lateral views, lateral tracheal deviations were assumed to be nonvisible. As a result, for training and comparison purposes, we considered that the associated labels in those cases were negative (0). As shown in Figure 1, there was an imbalance in the prevalence of most patterns. However, to set a fair baseline, and to be representative of what a real clinical caseload might be we elected not to use any class imbalance correction strategy, although there exist some in the literature.[13]

2.3 | Image processing and data augmentation

As per standard with the creation of CNNs, the DICOM images were converted into a smaller matrix (224 × 224) using a python script (Python Software Foundation, 9450 SW Gemini Dr., ECM# 90772, Beaverton, OR 97008, USA). To improve robustness of the CNN, the input was perturbed by randomly flipping, rotating (±10°), and zooming (0.8-1.2) each radiograph.[14]

2.4 | Model building

For the model to be easily reproducible, we used a state-of-the-art architecture of CNN with no specific improvement. Using the open-source deep-learning library Pytorch (https://pytorch.org/),[15] we chose to train a 121-layers densenet architecture (see Figure 2) using similar hyper-parameter settings as those described by Huang et al,[16] that is a dropout rate of 5%, a stochastic gradient descent optimizer with a momentum of 0.9 and a weight decay of 1×10^{-4}. The optimized loss function was binary cross-entropy. The learning rate was initially set to 0.1 and divided by 10 every time a new epoch did not improve the validation loss.

2.5 | Pretraining strategies

Three pretraining strategies were implemented to train three different CNNs before their exposure to the training set of thoracic veterinary radiographs. Pretrainings were performed using the same densenet architecture as described above. The first CNN did not receive any

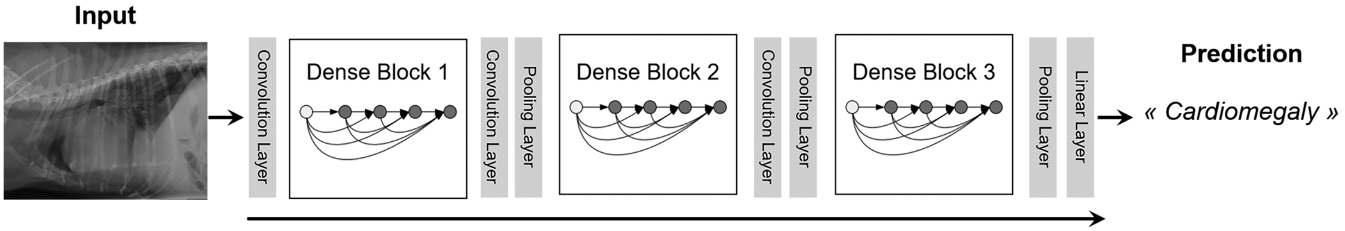

FIGURE 2 Architecture of the densenet model. Each convolution operation is learnable and extracts features from the previous convolution output. Inside a dense block, each dot represents a convolution operation followed by a batch normalization and an activation function. The arrows between two dots represent the reuse of features from previous layers which is the key improvement from the densenet architecture. Max pooling layers reduce feature map size to increase the ratio filter size (3 × 3)/input size in the following convolutions. Inspired from Huang et al[16]

pretraining. The second CNN underwent pretraining using ImageNet (Stanford Vision Lab, Stanford University, Princeton University, USA), which is the most commonly used pretraining dataset and contains over 14 million of categorized images.[17] The third CNN underwent pretraining using ImageNet followed by an open-source database of human thoracic radiographs (CheXpert, Stanfordmlgroup, Stanford University, USA).

2.6 | Evaluation of the algorithms

All statistical analyses were performed by an independent statistician (X.Z.) using commercially available statistics software (IBM SPSS, IBM, New York, USA) and the *P*-value was considered statistically significant if <.05. The three CNNs were then exposed to the training set of veterinary thoracic radiographs (14 202 radiographs). Subsequently, they labeled all 1458 radiographs from the validation dataset using the 15 labels. The area under the receiver operator characteristic curve (AUC) that provides an aggregate measure of performance across all possible classification thresholds when comparing CNNs to one another[18] was calculated for each CNNs using the radiologist's report as the gold standard. This was statistically compared using Chi-Square test and Fisher's exact test. Only the best performing of the three CNNs was retained and used as a comparison to veterinarians in the next stage of the experiment. For this purpose, the CNN was calibrated using a Platt Scaling,[19] which was fitted on the whole validation dataset. Out of the 1458 validation radiographs, 120 were drawn randomly following a uniform distribution.

A group of 10 veterinarians, composed of five veterinarians with more than 2 years of experience and five veterinarians with less than 2 of experience, was asked to label the radiographs using the same labels as the CNN. Each veterinarian had to label 10 radiographs on their own and 10 radiographs side by side with concurrent access to the CNN analysis. A round-robin fashion process was used to distribute batches of 10 images so that the same batch is labeled by two different veterinarians in the two different configurations. No history or signalment was provided to the veterinarians. Using the veterinary radiologist's report as the gold standard, the overall error rate and specific error rate per label of the CNN, veterinarians, and veterinarians aided by the CNN were calculated. Those error rates were compared to one another using McNemar's test. To confirm that the NLP for label extraction was accurate, the corresponding reports of the 120 radiographs were relabeled manually by an additional veterinarian (not participating in the lecture of the radiographs). Comparison of those labels to the NLP's results was performed using the McNemar test.

3 | RESULTS

3.1 | Pretraining strategies

Among all three methods of pre-training, the AUC was statistically better for the CNN trained with ImageNet alone ($P < .05$) with an AUC of 0.86. This was therefore established as the best performing CNN and was retained for the rest of the experiment. The error rate of the NLP was measured at 0.3% when comparing NLP's labeling to the labels present in the radiologists' report when verified by hand ($P < .001$).

When comparing the overall error rates between the two-level of experience amongst veterinarians, there was no statistical difference ($P = .547$), with the error rate of veterinarians with less than 2 years of experience at 17.5% and the error rate of veterinarians with more than 2 years of experience at 16.9%. The error rates per category of the two groups of veterinarians were also not statistically different when compared to one another (Figure 3). Therefore, for the subsequent comparison of CNN versus veterinarians versus veterinarians aided by the CNN, the two groups of veterinarians were pooled together to further increase the statistical power of the analysis.

When looking at all 15 categories, the overall error rates were as follow: CNN 10.7%, veterinarians 17.2%, and veterinarians aided by the CNN 16.8%. Statistically, the CNN's error rate was significantly lower compared to the two others ($P = .001$). There was no statistically significant difference between the overall error rates of veterinarians and veterinarians aided by the CNN (Figure 3).

When observing each category separately (Figure 4), the CNN's error rate was significantly better for all evaluated cardiac categories (cardiomegaly, left atrial dilatation, left ventricular dilation, right ventricular dilation). The error rate of the CNN was also statistically the lowest for the detection of "bronchial pattern." For the label "alveolar pattern" *P*-value approached statistical significance at $P = .053$. For the remaining categories, the CNN's error rate was always the lowest

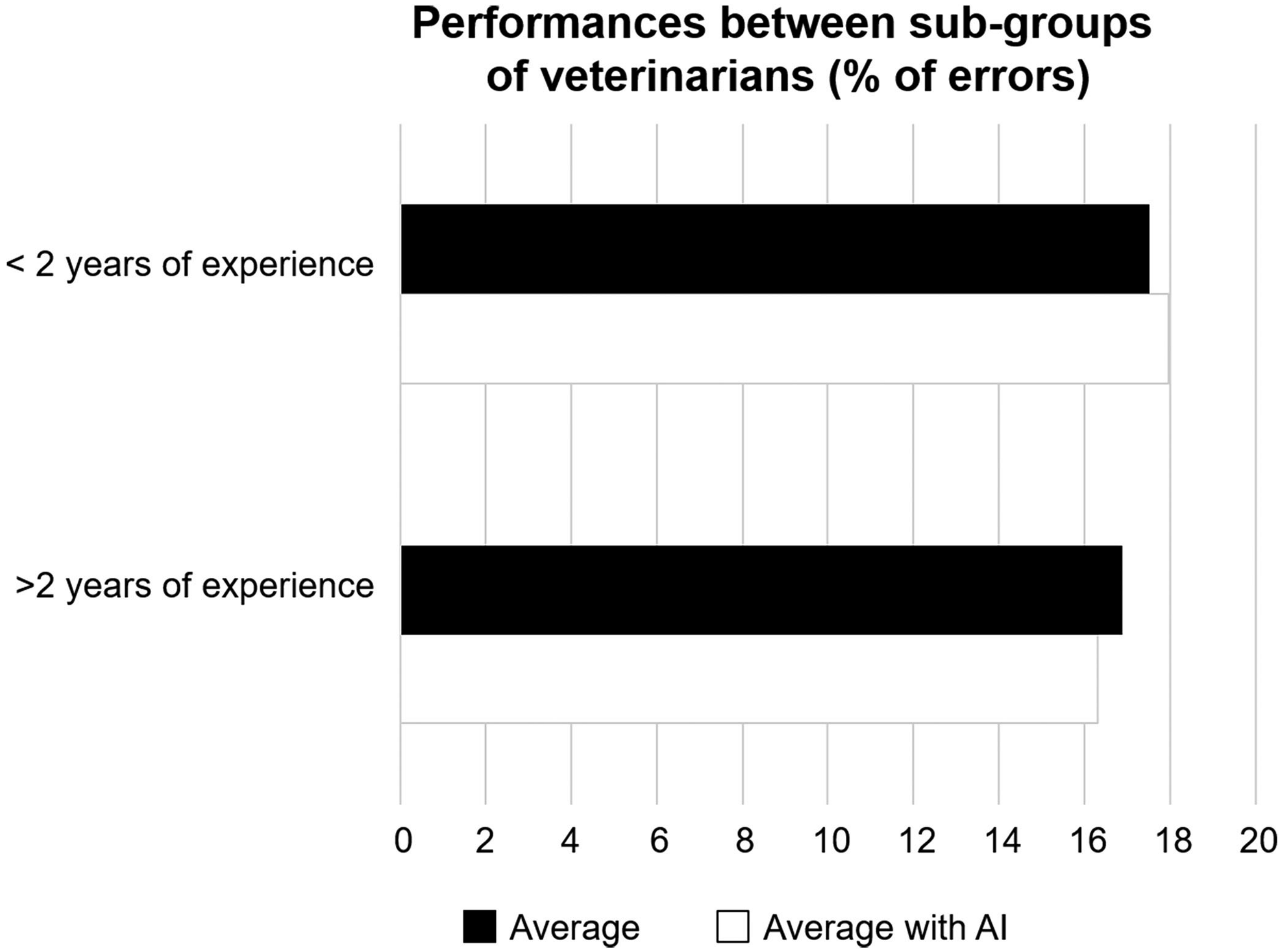

FIGURE 3 Comparison of the performance between veterinarians with less than 2 years of experience and more than 2 years of experience. Data are expressed as a percentage of errors. No statistically significant differences were noted between the two groups

compared to veterinarians or veterinarians aided by the CNN, but those results did not reach statistical significance.

When comparing veterinarians to veterinarians aided by the CNN, the error rates were not statistically different between all categories but two. For the categories "cardiomegaly" and "alveolar pattern," the veterinarians aided by the CNN performed better than the veterinarians alone (respectively 21.8% vs 32.8% for "cardiomegaly" $P = .041$, and 11.8% vs 20.2% for "alveolar pattern" $P = .041$).

4 | DISCUSSION

Results of this exploratory study showed that it is possible to produce a relevant automated lesion screening CNN on small animal thoracic radiographs with deep learning technics, using a relatively small dataset. The CNN we developed in this study is able to detect 15 primary thoracic lesions, allowing to get precise information on both the anatomical structures affected (ie, pulmonary parenchyma, cardiac silhouette, or mediastinal space) and the corresponding kind of radiographic anomaly involved (type of pulmonary pattern, specific cardiac chamber enlargement).

In deep learning, pretraining or transfer learning is a method consisting in using a large dataset to pretrain a neural network before training the same network on a smaller target dataset.[20] This enables the network to start learning on the target dataset with weight values, which are better than random.[20,21] The most used dataset for pretraining a CNN is ImageNet.[21,22] It contains more than 14 million images of all nature classified by type. There also exist multiple open-source datasets for human chest radiology like CheXpert, NIH,[23] or MIMIC-CXR databases. In order to make it possible for our CNN to learn from a rather small dataset (about 16 000 images in this study, in comparison to human datasets of more than a million images[11]) with a large patients size and shape diversity, we tested an original pretraining strategy, taking advantage of the abundance of human thoracic radiographs available on CheXpert.

All pretraining strategies had a positive outcome on the final performances of the algorithm as compared to no pretraining. However, pretraining with the large unspecialized dataset (ImageNet) resulted in a

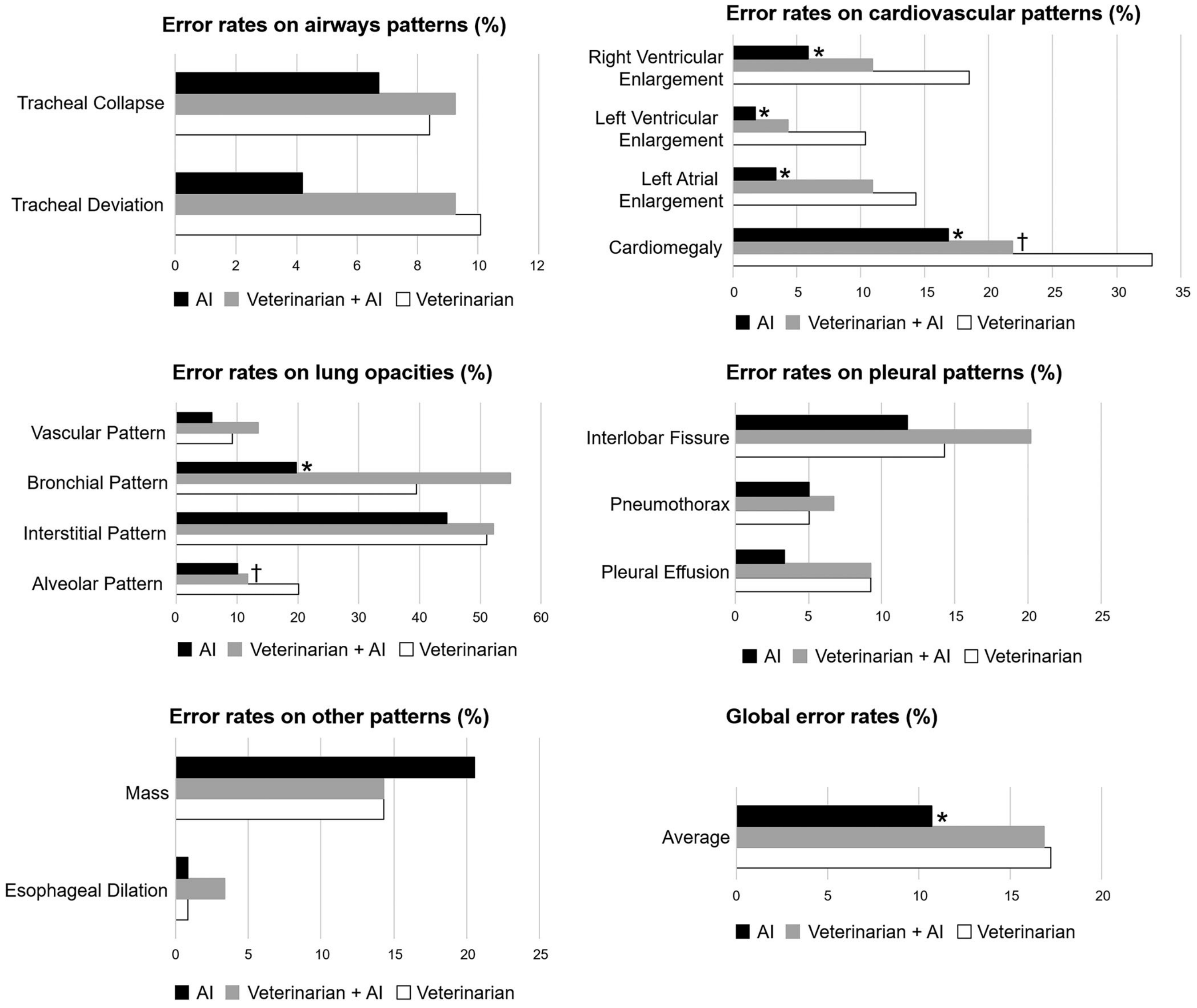

FIGURE 4 Error rates per label and per observers. Statistically significant better CNN's error rate compared to veterinarians and veterinarians aided by the CNN are denoted by a "*." Statistically significant better veterinarians aided by the CNN's error rate compared to veterinarian alone are denoted by a †

statistically better learning score (global superiority according to the average AUC scores) than pretraining with human chest radiographs. This might be explained by the difference of sample size: ImageNet contains 14 197 122 images whereas CheXpert contains only 187 641 images. Although the model trained on CheXpert had itself been pretrained on ImageNet, its training on a smaller dataset may have caused it to delete some discriminative features irrelevant for human chest X-ray classification, but that could have been the base for learning other features for veterinary radiographs.

In order to validate this hypothesis, we anecdotally performed another pretraining on CheXpert dataset, so-called "hard pretraining" with the same settings except for the learning rate which was increased at 0.1 instead (×1000). The point of testing a higher learning rate was to increase learning on the CheXpert dataset to the detriment of previously learned ImageNet features. The experiment indeed resulted in a catastrophic forgetting phenomenon such as described by McClosey et al[24] that harmed the transfer to our dataset as shown by results which exhibit poorer performances on animal patterns detection task.

It is important to highlight that the results presented in this study were obtained with only one CNN and are not an ensemble of several algorithms as usually presented to maximize scores,[25,26] as the aim was to produce an operational decision support system performing real-time inference. The confrontation of the resultant algorithm to general practitioners' abilities resulted in statistically significant improvement of the error rate for only two categories "alveolar pattern" and "cardiomegaly." However, by comparing CNN versus veterinarians alone versus veterinarians aided by the CNN (ie, veterinarians who had access to the results of the AI), we surprisingly highlighted that, on the majority of the studied patterns, the access to the algorithm did not result in statistically significant improvement of the veterinarian's error rate. This might suggest that veterinarians chose not to follow/trust the CNN's recommendations on these patterns. This can potentially be explained by the lack of experience of the veterinarian using AI as an aid.

One can also argue that the design of the study was not exactly representative of radiographic interpretation in practice as the veterinarians had to annotate the images blindly, with no access to the patient's history. Thus, without a sense of clinical relevance of some patterns, they might have had more difficulties than in real conditions. However, this allowed for a fair comparison between veterinarians and AI, as the current CNN does not process the patient's history or clinical signs.

Lower performances of the CNN on some categories identification could have been due to limitations of the NLP approach used to extract labels from the radiologist' reports. Indeed, if such strategy makes it possible to exploit a large dataset for multi-label tasks in a minimal amount of time, it also introduces noise in the dataset due to NLP errors. However, in our current study, the NLP's error rate was negligible at 0.3%. It would also be highly impractical and would potentially lead to more errors to have a veterinarian manually extract the labels for all the included radiographs. High in-class variance of certain labels associated with a low number of positive samples could also explained lower detection performances on some pattern, as for the label "Mass". We chose to gather all appearances from large mediastinal masses to millimetric nodules as done in some human lung nodules detection algorithm.[27]

Our results are encouraging as they were obtained with a simple learning strategy. Consistent improvement can be expected with a customized approach, focusing on veterinary radiographs dataset particularities. Our algorithm was trained on a standard matrix size for deep learning (224 × 224 images).[16,28–31] Such small images are currently the standard dimensions because the CNN architectures used are designed to work with images of this size.[16] Importantly, most current algorithms developed for human radiology were fed with similar-size images.[11,31] Using a significantly larger image size would require a specific architecture and is currently an open research topic in computer vision.[32–34] This, however, could potentially increase performances on spatially restricted patterns such as interlobar fissures or to make a further distinction between different pulmonary patterns. More simply, using a dataset imbalance strategy could also potentially improve general performances,[35] but would have misrepresented a real clinical case load.

Nevertheless, this study has limitations. The algorithm is not exhaustive as the analysis is restricted to thoracic radiographs and some lesion categories were ignored for our proof of concept. For instance, the addition of extra-thoracic structures screening such as lesions affecting the musculoskeletal system or cranial abdomen would be pertinent for clinical practice. The error rates of the three observers based on species (feline vs canine) or position of the patient (lateral vs ventrodorsal/dorsoventral radiograph) were not specifically evaluated.

Using the radiologist's report as the gold standard is the best choice but also has pitfalls. It is possible that some patterns that were studied might not have been systematically mentioned in the radiologist's reports. For instance, it might have been the case for tracheal deviations, when concomitant to a major etiological anomaly (such as a cardiomegaly), as the information might not have added value for comprehension of the case. Ideally, the 6,584 reports would have been reevaluated to ensure absolute completeness with the goal of CNN training in mind, but this was not logistically feasible. The veterinary radiographs used in this study came from a referring institution. This creates a strong bias toward the quality of the radiographs. Issues such as over/underexposure, patient's motion, poor collimation, patient's rotation, or radiation safety violation (ie, human's body parts in the primary beam) were absent from this dataset. This could potentially lead the CNN to errors when using radiographs of lower quality. The algorithm presented in this study only allows for the detection of primary lesions. It does not correlate the presence of lesions on orthogonal projections, nor does it link the different findings together to provide an interpretation of the underlying disease. For instance, if the CNN identifies simultaneously cardiomegaly with venous congestion and alveolar pattern, it does not suggest the presence of left-sided heart failure. Indeed, the aim here was to aid in lesion detection, letting the final interpretation, which also depends on external elements such as commemoratives or clinical exams, to the clinician and the radiologist. The 15 labels evaluated here were chosen out of the 61 that the NLP was able to extract from the reports. Those 15 labels were chosen as they represented primary patterns and are also patterns that have been evaluated in creating CNNs for human radiology. Categories such as pleural effusion, pneumothorax, alveolar pattern, bronchial pattern, interstitial pattern, cardiomegaly, specific cardiac chamber enlargement, tracheal deviation, tracheal collapse, and esophageal dilation were evaluated by Chexpert[11] and Qure[12]; pneumothorax was also studied by ChexPert (See Table S1).[11] Regarding validation of the software, it has not been yet confronted to board certified radiologists, but the gold standard was constituted of images with prior annotation by referral center veterinary radiologists.

In conclusion, findings from this preliminary study indicated that a CNN to identify 15 thoracic canine and feline primary radiographic patterns was feasible using a novel approach developed by the authors. To push further the evaluation of the algorithm, it should be confronted to other datasets acquired using different equipment or different operators and collimation habits for instance. It would also be interesting to reconduct the comparison between veterinarians and veterinarians aided by the CNN, also, informing the veterinarians ahead of time, that the overall error rate of the CNN is extremely low. The lack of improvement we noticed in the current study between veterinarians and veterinarians aided by the CNN might be dramatically changed, providing that the veterinarians trusted the AI's interpretation. To further enhance the statistical analysis, a larger number of cases could have been used. This potentially could have highlighted error rate's differences for which we did not achieve statistical significance in the current study. From this study, it appears that using such a deep learning tool for assisting veterinarians in everyday life radiograph reading could improve the quality of care. Such technology, if trusted and used systematically, could help the veterinarian detect thoracic lesions more efficiently while waiting for the report of the specialist who can provide a global and diagnostic opinion of the radiographs.

ACKNOWLEDGMENTS

The authors would like to acknowledge all the veterinary radiologists whose reports throughout the years might have contributed to the pool of cases.

LIST OF AUTHOR CONTRIBUTIONS

Category 1

(a) Conception and Design: Boissady, De La Comble, Hespel

(b) Acquisition of Data: Boissady, De La Comble

(c) Analysis and Interpretation of Data: Hespel, Zhu

Category 2

(a) Drafting the Article: Boissady, De La Comble, Hespel

(b) Revising Article for Intellectual Content: Boissady, De La Comble, Hespel

(c) Final Approval of the Completed Article: Boissady, De La Comble, Hespel, Zhu

Category 3

(a) Final Approval of the Completed Article: Boissady, De La Comble, Hespel, Zhu

CONFLICT OF INTEREST

E. Boissady and A. de La Comble are the developers of the Artificial Intelligence described in this study.

ORCID

Adrien-Maxence Hespel https://orcid.org/0000-0002-9060-2309

REFERENCES

1. Burns JE, Yao J, Muñoz H, Summers RM. Automated detection, localization, and classification of traumatic vertebral body fractures in the thoracic and lumbar spine at CT. *Radiology*. 2016;278(1):64-73.
2. Lakhani P, Sundaram B. Deep learning at chest radiography: automated classification of pulmonary tuberculosis by using convolutional neural networks. *Radiology*. 2017;284(2):574-582.
3. Maier O, Schröder C, Forkert ND, Martinetz T, Handels H. Classifiers for ischemic stroke lesion segmentation: a comparison study. *PLoS One*. 2015;10(12):e0145118.
4. Taylor AG, Mielke C, Mongan J. Automated detection of moderate and large pneumothorax on frontal chest X-rays using deep convolutional neural networks: a retrospective study. *PLoS Med*. 2018;15(11):e1002697.
5. Juliusson G, Thorvaldsdottir B, Kristjansson JM, Hannesson P. Diagnostic imaging trends in the emergency department: an extensive single-center experience. *Acta Radiol Open*. 2019;8(7):2058460119860404.
6. Rodriguez-Ruiz A, Lång K, Gubern-Merida A, et al. Stand-alone artificial intelligence for breast cancer detection in mammography: comparison with 101 radiologists. *J Natl Cancer Inst*. 2019;111(9):916-922. https://doi.org/10.1093/jnci/djy222
7. Yoon Y, Hwang T, Lee H. Prediction of radiographic abnormalities by the use of bag-of-features and convolutional neural networks. *Vet J*. 2018;237:43-48. https://doi.org/10.1016/j.tvjl.2018.05.009
8. Donald JJ, Barnard SA. Common patterns in 558 diagnostic radiology errors. *J Med Imaging Radiat Oncol*. 2012;56(2):173-178. https://doi.org/10.1111/j.1754-9485.2012.02348.x
9. Gaschen L. Pitfalls of thoracic radiography: don't let them trap you. *WSAVA Proc*. 2010. https://www.vin.com/apputil/content/defaultadv1.aspx?pId=11310&id=4516254&print=1
10. Thrall DE. *Textbook of Veterinary Diagnostic Radiology*. Amsterdam, the Netherlands: Elsevier Health Sciences; 2017.
11. Irvin J, Rajpurkar P, Ko M, et al. CheXpert: a large chest radiograph dataset with uncertainty labels and expert comparison. *ArXiv190107031 Cs Eess*. http://arxiv.org/abs/1901.07031. Accessed April 6, 2020.
12. Putha P, Tadepalli M, Reddy B, et al. Can artificial intelligence reliably report chest X-rays? Radiologist validation of an algorithm trained on 2.3 million X-rays. *ArXiv180707455 Cs*. http://arxiv.org/abs/1807.07455. Accessed April 7, 2020.
13. Giraldo Forero AF, Jaramillo-Garzón J, Ruiz-Muñoz J, Castellanos-Dominguez G. Managing imbalanced data sets in multi-label problems: a case study with the SMOTE algorithm. *Lect Notes Comput Sci*. 2013;8258:334-342.
14. Perez L, Wang J, The effectiveness of data augmentation in image classification using deep learning. *ArXiv171204621 Cs*. http://arxiv.org/abs/1712.04621. Accessed April 7, 2020.
15. Paszke A, Gross S, Chintala S, et al. Automatic differentiation in PyTorch.4. https://openreview.net/pdf/25b8eee6c373d48b84e5e9c6e10e7cbbbce4ac73.pdf
16. Huang G, Liu Z, van der Maaten L, Weinberger KQ. Densely connected convolutional networks. ArXiv160806993 Cs. http://arxiv.org/abs/1608.06993. Accessed April 6, 2020.
17. He K, Girshick R, Dollár P, Rethinking ImageNet pre-training. *ArXiv181108883 Cs*. http://arxiv.org/abs/1811.08883. Accessed April 7, 2020.
18. Ling CX, Huang J, Zhang H, AUC: a Statistically Consistent and more Discriminating Measure than Accuracy.6.
19. Platt J. Probabilistic outputs for support vector machines and comparisons to regularized likelihood methods. *Adv Large Margin Classif*. 2000:10.
20. Tan C, Sun F, Kong T, Zhang W, Yang C, Liu C, A survey on deep transfer learning. *ArXiv180801974 Cs Stat*. http://arxiv.org/abs/1808.01974. Accessed April 7, 2020.
21. Raghu M, Kleinberg J, Zhang C, Bengio S, Transfusion: Understanding Transfer Learning for Medical Imaging. 11. https://papers.nips.cc/paper/8596-transfusion-understanding-transfer-learning-for-medical-imaging.pdf
22. Deng J, Dong W, Socher R, Li L-J, Li Kai, Fei-Fei Li, ImageNet: a large-scale hierarchical image database. In: *2009 IEEE Conference on Computer Vision and Pattern Recognition*.; 2009:248-255. https://ieeexplore.ieee.org/document/5206848
23. Wang X, Peng Y, Lu L, Lu Z, Bagheri M, Summers RM, ChestX-ray8: hospital-scale chest X-ray database and benchmarks on weakly-supervised classification and localization of common thorax diseases. *2017 IEEE Conf Comput Vis Pattern Recognit CVPR*. 2017:3462-3471. https://arxiv.org/abs/1705.02315
24. McCloskey M, Cohen NJ. Catastrophic interference in connectionist networks: the sequential learning problem. In: Bower GH, ed. *Psychology of Learning and Motivation*. Cambridge, MA: Academic Press; 1989:109-165.
25. Dietterich TG. Ensemble methods in machine learning. *Multiple Classifier Systems*. Vol 1857. Lecture Notes in Computer Science. Berlin, Heidelberg: Springer; 2000:1-15.
26. Opitz D, Maclin R. Popular ensemble methods: an empirical study. *J Artif Intell Res*. 1999;11:169-198.
27. Shiraishi J, Li Q, Suzuki K, Engelmann R, Doi K. Computer-aided diagnostic scheme for the detection of lung nodules on chest radiographs: localized search method based on anatomical classification. *Med Phys*. 2006;33(7):2642-2653.

28. Howard AG, Zhu M, Chen B, et al. MobileNets: efficient convolutional neural networks for mobile vision applications. *ArXiv170404861 Cs*. 2017. http://arxiv.org/abs/1704.04861. Accessed April 7, 2020.
29. Zoph B, Vasudevan V, Shlens J, Le QV, Learning transferable architectures for scalable image recognition. *ArXiv170707012 Cs Stat*. http://arxiv.org/abs/1707.07012. Accessed April 7, 2020.
30. He K, Zhang X, Ren S, Sun J, Deep residual learning for image recognition. *ArXiv151203385 Cs*. http://arxiv.org/abs/1512.03385. Accessed April 7, 2020.
31. Krizhevsky A, Sutskever I, Hinton GE. ImageNet classification with deep convolutional neural networks. *Commun ACM*. 2017;60(6):84-90.
32. Maggiori E, Tarabalka Y, Charpiat G, Alliez P, High-resolution image classification with convolutional networks. In: *2017 IEEE International Geoscience and Remote Sensing Symposium (IGARSS)*; 2017:5157-5160. https://ieeexplore.ieee.org/document/8128163
33. Huang Y, Chung AC. Improving high resolution histology image classification with deep spatial fusion network. *ArXiv180710552 Cs*. 2018;11039:19-26.
34. Iftene M, Qingjie L, Yunhong W. Very high resolution images classification by fusing deep convolutional neural networks. In: 2017. https://www.clausiuspress.com/conferences/ACSS/ACSAT%202017/GACS57.pdf
35. Kotsiantis S, Kanellopoulos D, Pintelas P. Handling imbalanced datasets: a review. *GESTS Int Trans Comput Sci Eng*. 2005;30:25-36.

SUPPORTING INFORMATION

Additional supporting information may be found online in the Supporting Information section at the end of the article.

How to cite this article: 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*. 2020;61:619–627. https://doi.org/10.1111/vru.12912