Received: 1 March 2024 | Revised: 27 June 2024 | Accepted: 5 July 2024

DOI: 10.1111/vru.13413

ORIGINAL RESEARCH

WILEY

Automatic classification and grading of canine tracheal collapse on thoracic radiographs by using deep learning

Hathaiphat Suksangvoravong[1] | Nan Choisunirachon[1] | Teerawat Tongloy[2] | Santhad Chuwongin[2] | Siridech Boonsang[3] | Veerayuth Kittichai[4] | Chutimon Thanaboonnipat[1]

[1]Department of Veterinary Surgery, Faculty of Veterinary Science, Chulalongkorn University, Bangkok, Thailand

[2]College of Advanced Manufacturing Innovation, King Mongkut's Institute of Technology Ladkrabang, Bangkok, Thailand

[3]Department of Electrical Engineering, Faculty of Engineering, King Mongkut's Institute of Technology Ladkrabang, Bangkok, Thailand

[4]Faculty of Medicine, King Mongkut's Institute of Technology Ladkrabang, Bangkok, Thailand

Correspondence
Chutimon Thanaboonnipat, Department of Veterinary Surgery, Faculty of Veterinary Science, Chulalongkorn University, Henri-Dunant Road, Pathumwan, Pathumwan, Bangkok 10330, Thailand.
Email: chutimon.th@chula.ac.th

Abstract

Tracheal collapse is a chronic and progressively worsening disease; the severity of clinical symptoms experienced by affected individuals depends on the degree of airway collapse. Cutting-edge automated tools are necessary to modernize disease screening using radiographs across various veterinary settings, such as animal clinics and hospitals. This is primarily due to the inherent challenges associated with interpreting uncertainties among veterinarians. In this study, an artificial intelligence model was developed to screen canine tracheal collapse using archived lateral cervicothoracic radiographs. This model can differentiate between a normal and collapsed trachea, ranging from early to severe degrees. The you-only-look-once (YOLO) models, including YOLO v3, YOLO v4, and YOLO v4 tiny, were used to train and test data sets under the in-house XXX platform. The results showed that the YOLO v4 tiny-416 model had satisfactory performance in screening among the normal trachea, grade 1–2 tracheal collapse, and grade 3–4 tracheal collapse with 98.30% sensitivity, 99.20% specificity, and 98.90% accuracy. The area under the curve of the precision–recall curve was >0.8, which demonstrated high diagnostic accuracy. The intraobserver agreement between deep learning and radiologists was $\kappa = 0.975$ ($P < .001$), with all observers having excellent agreement ($\kappa = 1.00$, $P < .001$). The intraclass correlation coefficient between observers was >0.90, which represented excellent consistency. Therefore, the deep learning model can be a useful and reliable method for effective screening and classification of the degree of tracheal collapse based on routine lateral cervicothoracic radiographs.

KEYWORDS
deep learning, dogs, radiographs, screening, tracheal collapse

1 | INTRODUCTION

Canine tracheal collapse is a chronic, progressive, and irreversible disease of the trachea that commonly occurs in middle-aged, small-breed dogs, including Yorkshire terriers, Miniature poodles, Pug, Maltese, Chihuahuas, and Pomeranians.[1] The prevalence of tracheal collapse in normal and underweight dogs was approximately 0.1%[2] with no sex predisposing.[3] The prevalence was reported to increase to 0.2%–0.4% in overweight and obese dogs.[2] Several factors have been reported to be the causes of tracheal collapse; however, most of the cases have unknown etiology. The common clinical symptoms of canine tracheal collapse include persistent dry, harsh cough or goose-honking, and respiratory distress.[3,4] The severity of clinical symptoms depends on the degree of collapsed airway, which can lead to life-threatening severe respiratory distress.[5]

Radiography is a popular clinical diagnostic imaging test used for screening tracheal collapse in dogs by comparing the tracheal luminal height (TH) with the thoracic inlet height (TI) on the lateral view of a cervicothoracic radiograph.[6,7] The TH is measured from the lesser dorsoventral diameter of the trachea at the level of TI as it is the area with the smallest diameter.[8] The TI is measured from the cranioventral aspect of the first thoracic vertebrae (T1) to the highest point of the craniodorsal manubrium.[5] In normal dogs, the TH/TI ratio should be greater than 0.2:1.[9] However, it has been reported that brachycephalic breeds and English bulldogs normally have lower ratios than others, which are 0.16:19,[10] and 0.11:1,[11] respectively. In dogs with tracheal collapse, the TH is decreased, resulting in a decreased ratio.

Although advanced imaging modalities, such as fluoroscopy and computed tomography, can precisely detect tracheal collapse, plain cervicothoracic radiography is preferable in clinical practice[12] because it has wide availability and is easy to use and inexpensive. Additionally, only lateral thoracic radiography is required for differentiating between a normal trachea and a collapsed trachea and for classifying the degree of tracheal collapse. The severity of tracheal collapse is classified into four grades[13]: grade 1 (mild condition; the TH is reduced by approximately 25%), grade 2 (mild to moderate condition; the TH is reduced by approximately 50%), grade 3 (moderate to severe condition; the TH is reduced by approximately 75%), and grade 4 (severe condition; the TH is reduced by >75%). In dogs, grade 1 tracheal collapse can be an accidental finding without any symptoms, whereas grade 3 and 4 tracheal collapse requires appropriate treatment.

An error in the interpretation of plain radiographs commonly occurs in both human medicine and veterinary medicine. The misinterpretation errors and misdiagnosis can significantly impact the treatment planning and overall quality of life of patients.[14] In human medicine, the error in interpretation has been reported to be 10%–15% for general radiographs[15] and increased up to 26% for thoracic radiographs.[16] In veterinary medicine, unlike human medicine, the reported data on this issue are limited. The accuracy of radiographic interpretation relies heavily on the expertise of radiologists,[17] and the variation in size and body conformation of dogs because of breed differences can lead to diverse radiographic appearances.[18] This variability can, in turn, affect the accuracy of radiographic interpretation. Given the scarcity of well-trained radiologists in veterinary medicine, general practitioners usually play a major responsibility in interpreting radiographs.[19] Moreover, the increasing workload can leave radiologists exhausted, fatigued, and stressed, potentially resulting in more mistakes and errors during radiographic interpretation.[20] Therefore, it becomes imperative to develop advanced alternatives that can minimize interpretation errors and enhance diagnostic accuracy in veterinary medicine.[21]

Deep learning algorithms, including convolutional neural networks (CNNs), have been effectively applied in several object detections and image recognitions in many medical applications.[22–24] Growing and rapidly evolving interest in deep learning algorithms has been seen in the field of veterinary radiology. These algorithms are gaining significant traction and attention because of their potentially revolutionizing approach to analyzing and interpreting radiographic images. CNNs have been used for automatic radiographic interpretation in screening lesions in veterinary medicine, such as the detection of cardiomegaly[19] and left atrial enlargement on canine thoracic radiographs,[25] classification of thoracic radiographs in cats,[26] comparison of the error rates between the CNNs and 13 board-certified veterinary radiologists on canine thoracic radiographs,[27] comparison of artificial intelligence (AI) with the veterinary radiologist's diagnosis on the detection of canine cardiogenic pulmonary edema,[28] and detection of confirmed pleural effusion in thoracic radiographs in dogs.[29] The you-only-look-once (YOLO) model, introduced by Redmon and Farhadi,[30] is recognized for its ability to perform object detection with remarkable speed and efficiency, utilizing just a single CNN. The YOLO v3 model has been reported to simultaneously classify and localize the images to identify the species and sex of mosquitoes with a mean average precision and sensitivity of 99% and 92.4%, respectively.[31] Additionally, several studies in human medicine have employed the YOLO network model. These studies have reported successful applications of this model in detecting and classifying abnormalities in radiographs, including evaluating C-spine injuries from lateral neck radiographs,[32] detecting hip fractures,[33] and classifying knee osteoarthritis and assessing its severity based on radiographic images.[34]

To the best of our knowledge, there has not been any information on the screening and staging of tracheal collapse in dogs using deep learning techniques. This highlights an opportunity for novel research and the application of deep learning algorithms to improve the diagnosis and management of tracheal collapse in veterinary medicine. Therefore, this study aimed to (1) develop a platform involving a deep learning approach to screen tracheal collapse in dogs and evaluate and stage the severity of tracheal collapse based on lateral cervicothoracic radiographs and (2) evaluate an agreement between deep learning and radiologists in detecting and staging the severity of tracheal collapse. The model was trained and tested using a processed sort of image quality corresponding to the radiographic images, which were verified by experienced radiologists. The model may serve as a promising deep-learning prototype for the screening and staging tracheal collapse in dogs to diagnose this condition with enhanced accuracy and efficiency. Nevertheless, further research is required to ensure the effectiveness of the model in real-world veterinary practice.

2 | MATERIALS AND METHODS

2.1 | Ethics statement

All data information including clinical demographic data and all radiographic images were allowed to be used by the committee of the Small Animal Hospital, Faculty of Veterinary Science, Chulalongkorn University (approval no. 252/2565). This study utilized cervicothoracic radiographs of dogs that were presented to the Diagnostic Imaging Unit, the Small Animal Hospital, Faculty of Veterinary Science, Chulalongkorn University from January 2011 to July 2021. The data were

TABLE 1 Clinical demographic data of all included dogs in this study.

Parameters	Overall ($n = 600$)	Normal ($n = 200$)	Grade 1–2 ($n = 200$)	Grade 3–4 ($n = 200$)
Age (years)	10.26 ± 3.98	11.05 ± 3.72[a]	9.52 ± 4.21	10.17 ± 3.87
BW (kg)	8.09 ± 8.47	9.47 ± 8.89[b]	9.95 ± 10.09[b]	3.93 ± 1.96
Female	288 (48.00%)	93 (46.50%)	102 (51.00%)	93 (46.50%)
Intact	136 (22.67%)	41 (20.50%)	52 (26.00%)	43 (21.50%)
Sterile	152 (25.33%)	52 (26.00%)	50 (25.00%)	50 (25.00%)
Male	312 (52.00%)	107 (53.50%)	98 (49.00%)	107 (53.50%)
Intact	216 (36.00%)	74 (37.00%)	59 (29.50%)	83 (41.50%)
Castrated	96 (16.00%)	33 (16.50%)	39 (19.50%)	24 (12.00%)

Note: Statistically significant difference between groups was made using the Kruskal–Wallis test followed by Dunn's multiple comparisons test. Data are presented as mean ± SD for age and BW and number of dogs (percent) for sex.
[a]Significant difference at $P < .001$ compared with grade 1–2 of tracheal collapse dog.
[b]Significant difference at $P < .0001$ compared with grade 3–4 of tracheal collapse dog.

retrieved from the Hospital Information System and Picture Archiving and Communication System.

2.2 | Data set collection and preparation

This study was designed as a retrospective study. The included radiographs were obtained from both the conventional radiograph using a high-detail film (T-MAT S/RA Film; Kodak) between 2011 and 2013 and digital radiograph (ETL; GE HealthCare) between 2013 and 2021. The inclusion criteria of radiographs were good positioning of the right lateral cervicothoracic radiographs in which the field of view covered the thoracic inlet to the last rib, no mass or tumor that could affect either the tracheal diameter or alignment and no abnormality of the first thoracic vertebra (T1) and manubrium. All included radiographs were collected as the Digital Imaging and Communications in Medicine (DICOM) file for the evaluation of the ratios of the TH and TI by DICOM viewer software (OSIRIX). The data set consisted of 600 cervicothoracic radiographs, which were classified into the (1) normal tracheal dog group (no previous history of respiratory problem; $n = 200$) and (2) tracheal collapse dog group (had previous history of respiratory problems; $n = 400$), which was subdivided according to the severity grades—grade 1 and 2 tracheal collapse group ($n = 200$) and grade 3 and 4 tracheal collapse group ($n = 200$). The ratios of the TH/TI of grade 1–2 tracheal collapse and grade 3–4 tracheal collapse were between 0.10 and 0.15 and ≤0.10, respectively.[35] All included radiographs in each group were divided into those for training-integrated validations (90%) and those for testing image sets (10%). All included cervicothoracic radiographs of each group in this study are summarized in Tables 1 and 2. All images were individually evaluated by three observers with different levels of experience, two of whom were well-trained veterinary radiologists experienced in radiographic interpretation for >20 (N.C.) and 8 (C.T.) years and a master's degree student (H.S.). All observers were blinded to the dog's history, signalment, and clinical signs to eliminate an analysis error from clinical bias.

2.3 | Deep learning model configuration, model training, and model testing

In this study, the YOLO v3, YOLO v4, and YOLO v4 tiny models were used to perform the object detection algorithm. A total of 540 labeled images were used as training and integrated validation sets consisting of the normal tracheal set ($n = 180$), grade 1–2 tracheal collapse set ($n = 180$), and grade 3–4 tracheal collapse set ($n = 180$). All images in the training sets were labeled by the same professional radiologist using the in-house deep learning software (CiRA CORE, https://www.facebook.com/groups/cira.core.comm/) to develop models for the screening and classification of the degree of collapsed trachea. The bounding boxes represented a potential region of interest that contained the tracheal location at the thoracic inlet area for object detection. The labels corresponding to the normal trachea and degree of tracheal collapse of each image were used to train the model (Figure 1).

Overall, 190,116 augmented images based on an eight-step rotation and nine brightness/darkness steps were used to train the model to improve its performance. The general setting of the data augmentation was performed to increase the number of data sets before training the model. Multiple data augmentation techniques were applied in this study to increase the variations of the data set, including image rotation and adjustment. First, the rotation of the images in different angles was set with the value ranging from −180° to 180°, varying at every step of 90°. Second, the adjustment of the brightness and contrast of the images was set by one step ranging from 0.2 to 1.2.

Moreover, this study included the cervicothoracic radiographs obtained from two radiographic machines that provided different image qualities. The data augmentation technique used in this study is one of the important parts that help to imitate the real-life scenario, increase more training sets, and alleviate overfitting problems.[36] This technique could increase the variety of radiographic images, including overexposure, underexposure, and oblique images. Adding different levels of noise and blur in this augmentation was not performed

TABLE 2 Breeds of included dog in this study.

Breeds	Data set ($n = 600$)	Normal ($n = 200$)	Grade 1–2 ($n = 200$)	Grade 3–4 ($n = 200$)
Pomeranian	194 (32.33%)	20 (10.00%)	74 (37.00%)	100 (50.00%)
Mixed breed	138 (23.00%)	50 (25.00%)	54 (27.00%)	34 (17.00%)
Chihuahua	72 (12.00%)	39 (19.50%)	18 (9.00%)	15 (7.50%)
Poodle	70 (11.67%)	35 (17.50%)	18 (9.00%)	17 (8.50%)
Yorkshire Terrier	40 (6.67%)	2 (10.00%)	12 (6.00%)	26 (13.00%)
Beagle	13 (2.17%)	9 (4.50%)	4 (2.00%)	0 (0.00%)
Miniature Pinscher	9 (1.5%)	5 (2.50%)	2 (1.00%)	2 (1.00%)
Siberian Husky	9 (1.50%)	6 (3.00%)	3 (1.50%)	0 (0.00%)
Dachshund	8 (1.33%)	8 (4.00%)	0 (0.00%)	0 (0.00%)
Golden Retriever	8 (1.33%)	3 (1.50%)	5 (2.50%)	0 (0.00%)
Labrador Retriever	7 (1.17%)	5 (2.50%)	2 (1.00%)	0 (0.00%)
Jack Russel Terrier	5 (0.83%)	3 (1.50%)	2 (1.00%)	0 (0.00%)
German Shepherd	4 (0.67%)	4 (2.00%)	0 (0.00%)	0 (0.00%)
Maltese	4 (0.67%)	1 (0.05%)	0 (0.00%)	3 (1.50%)
Cocker Spaniel	3 (0.50%)	2 (1.00%)	0 (0.00%)	1 (0.05%)
Bang Kaew	2 (0.33%)	1 (0.05%)	1 (0.05%)	0 (0.00%)
Schnauzer	2 (0.33%)	0 (0.00%)	2 (1.00%)	0 (0.00%)
Scottish Terrier	2 (0.33%)	2 (1.00%)	0 (0.00%)	0 (0.00%)
American Pitbull	1 (0.17%)	1 (0.05%)	0 (0.00%)	0 (0.00%)
Bernese Mountain dog	1 (0.17%)	1 (0.05%)	0 (0.00%)	0 (0.00%)
Chinese crested dog	1 (0.17%)	1 (0.05%)	0 (0.00%)	0 (0.00%)
Corgi	1 (0.17%)	1 (0.05%)	0 (0.00%)	0 (0.00%)
Dalmatian	1 (0.17%)	0 (0.00%)	1 (0.05%)	0 (0.00%)
Italian Greyhound	1 (0.17%)	0 (0.00%)	1 (0.05%)	0 (0.00%)
Japanese Spitz	1 (0.17%)	1 (0.05%)	0 (0.00%)	0 (0.00%)
Samoyed	1 (0.17%)	0 (0.00%)	1 (0.05%)	0 (0.00%)
Silky Terrier	1 (0.17%)	0 (0.00%)	0 (0.00%)	1 (0.05%)
Shetland Sheepdog	1 (0.17%)	0 (0.00%)	0 (0.00%)	1 (0.05%)

Note: Data are presented as number of dogs (percent).

because these variations already existed in normal clinical radiographic images.

All images included in the testing sets were transferred to a dedicated deep-learning server to establish ground truth labels and model training. This process is essential to ensure that the model can learn from accurately annotated data and subsequently make accurate predictions when presented with unseen images. The labeling tool available in the CiRA CORE platform was employed for this purpose. All images in the testing set were subjected to testing and validation using the model. The model automatically identified the location of the trachea lesions and classified them into distinct groups, which included normal trachea, grade 1 and 2 tracheal collapse, and grade 3 and 4 tracheal collapse, based on their respective probability percentages. This automated classification process can greatly expedite the evaluation of tracheal collapse in dogs. The models were evaluated by assessing their performance in detecting tracheal collapse and accurately classifying its degree. Such evaluations are crucial to determine how well the deep learning model can reliably identify and categorize the severity of tracheal collapse in dogs. The workflow and model architecture of the methodology are shown in Figure 2.

2.4 | Statistical analysis and performance evaluation

The performance assessment of the deep learning network models was carried out using a confusion matrix table. Several statistical metrics from these models were compared, including precision, recall (sensitivity), accuracy, F1 score, misclassification rate, and specificity. These metrics are commonly used to evaluate the performance of deep learning models and provide a comprehensive understanding of the performance of the model in terms of identifying and classifying tra-

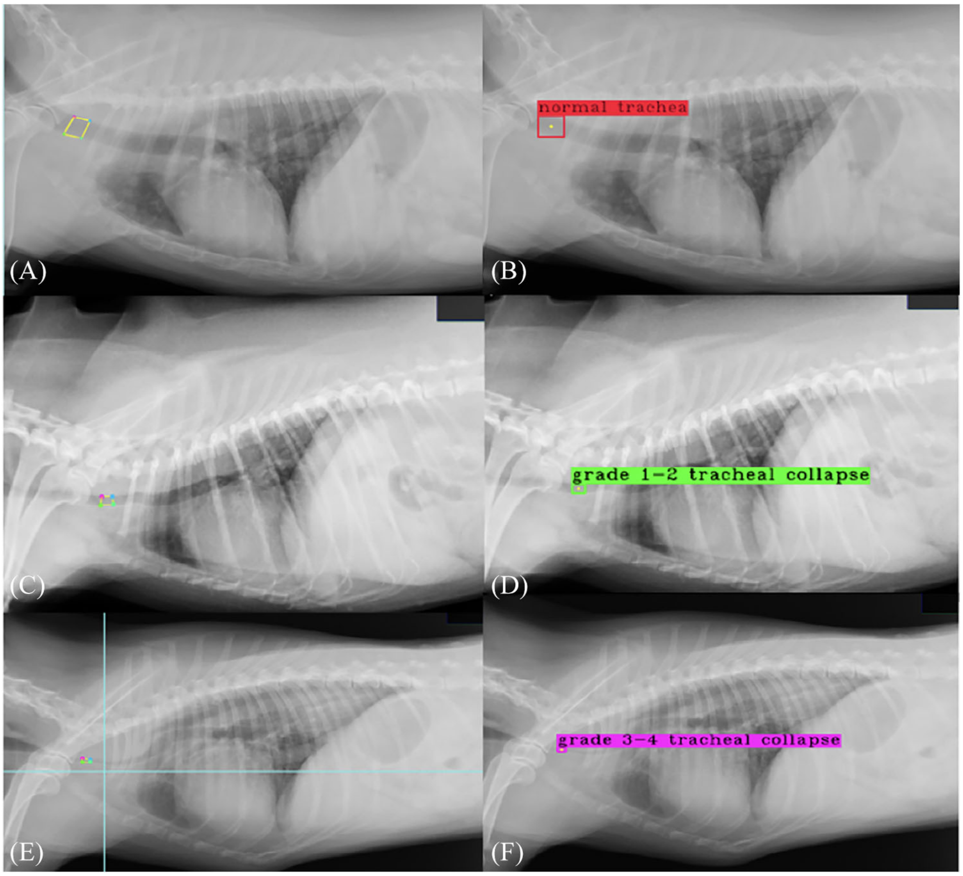

FIGURE 1 The labeling of training methodology. Normal tracheal (A and B), grade 1–2 tracheal collapse (C and D), and grade 3–4 tracheal collapse (E and F).

cheal collapse in this context. The formulas for these parameters are as follows:

$$\text{Sensitivity/ Recall} = \frac{\text{TP}}{\text{TP} + \text{FN}} \tag{1}$$

$$\text{False negative rate} = \frac{\text{FN}}{\text{FN} + \text{TP}} \tag{2}$$

$$\text{Specificity} = \frac{\text{TN}}{\text{TN} + \text{FP}} \tag{3}$$

$$\text{False positive rate} = \frac{\text{FP}}{\text{TP} + \text{FP}} \tag{4}$$

$$\text{Precision} = \frac{\text{TP}}{\text{TP} + \text{FP}} \tag{5}$$

$$\text{Accuracy} = \frac{\text{TP} + \text{TN}}{\text{TP} + \text{TN} + \text{FP} + \text{FN}} \tag{6}$$

$$\text{Misclassification rate} = \frac{\text{FP} + \text{FN}}{\text{TP} + \text{TN} + \text{FP} + \text{FN}} \tag{7}$$

$$\text{Negative predictive value} = \frac{\text{TN}}{\text{TN} + \text{FN}} \tag{8}$$

$$\text{F1 score} = 2 \times \frac{(\text{precision} \times \text{recall})}{(\text{precision} + \text{recall})} \tag{9}$$

where, TP, TN, FP, and FN refer to the number of true positive, true negative, false positive, and false negative, respectively. The testing sets were used to evaluate the performance of the models using a confusion matrix to analyze the sensitivity (true-positive rate and recall), false-negative rate, specificity (true-negative rate), false-positive rate, precision (positive predictive value), accuracy, misclassification rate, and negative predictive value (NPV). In addition, the precision–recall curve (PRC) was a graphical representation used to evaluate the object detection model. It displayed the trade-off between precision and recall at different thresholds used for generalized accuracy. The area under the curve (AUC) of 1 indicates perfect prediction, whereas that of 0.5 suggests random prediction. The performance of the model was summarized in terms of the area under the PRC

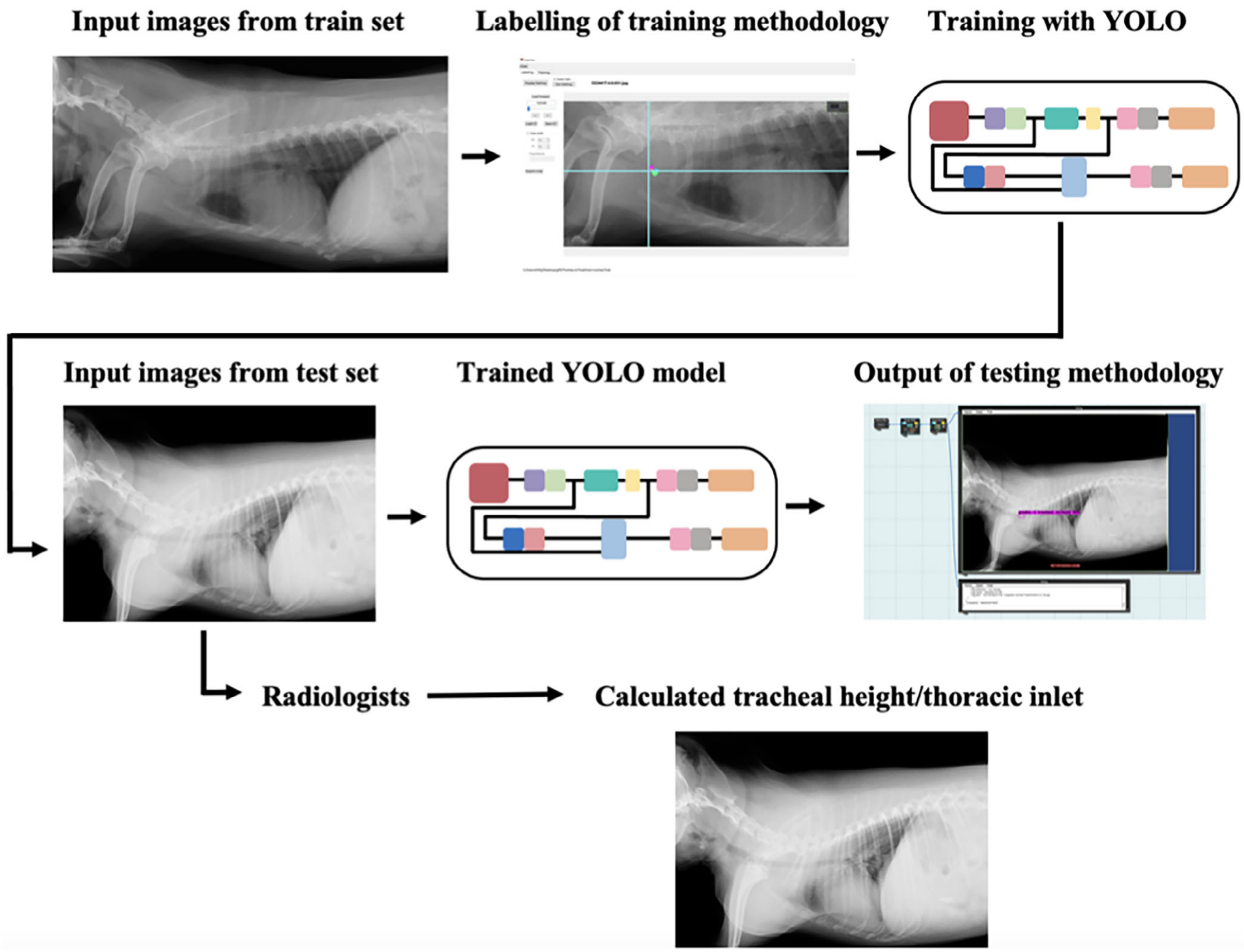

FIGURE 2 The dataset consisted of right lateral cervicothoracic radiographs that were labeled with a bounding box. These images were used to train the YOLO algorithm. After training, the YOLO model was tested and validated using images from the test set. The output of the model was presented as a predicted bounding box that outlined the regions of interest. Additionally, the model assigned a class "grade 1–2 tracheal collapse" through CiRA CORE software.

(AUC). The mean average precision (mAP), a widely used metric, was applied to evaluate the performance of the object detection models, particularly in computer vision as image recognition. A higher mAP indicated better object detection performance across all classes. The performance of the proposed models in screening tracheal collapse and grading its degree was compared with the radiologists' report.

3 | RESULTS

The performance and robustness of the proposed models were tested using the in-house deep learning software (CiRA CORE). In the initial and modified learning processes of the YOLO models, radiographic images verified by experienced radiologists were used in the training and testing sets, and the different network architectures of the models were assessed considering the same assignments. The statistical results indicated that the accuracy of the YOLO v4 tiny-416#3 model was 98.3%, which was higher than that of the other models, as shown in Table 3. In addition, all of the performance was satisfactory, with 98.3% sensitivity, 99.2% specificity, a 98.3% F1 score, and a 1.10% misclassification rate. The results indicated the high-level precision when the recall increases represents the satisfactoriness of performance. The PRC showed that the YOLO v4 tiny-416#3 model had the best result, which could classify normal trachea, grade 1 and 2 tracheal collapse, and grade 3 and 4 tracheal collapse with AUC values of 0.950, 0.841, and 0.934, respectively. All three curves were close to the top right of the graph, indicating very good performance. Moreover, the average precision (AP) of the YOLO v4 tiny-416#3 model exhibited the best performance in screening among normal trachea, grade 1 and 2 tracheal collapse, and grade 3 and 4 tracheal collapse with AP values of 0.99, 0.94, and 0.84, respectively, and the results showed satisfactory mAP at 0.92.

For the YOLO v4 tiny-416#3 *model*, the results showed that the *model* could automatically locate the tracheal location on all

TABLE 3 Model comparison.

Model	AP value			
	mAP	Normal trachea	Grade 1–2 tracheal collapse	Grade 3–4 tracheal collapse
YOLO v3	0.917 ± 0.046	0.942 ± 0.003	0.951 ± 0.090	0.857 ± 0.000
YOLO v4-64	0.386 ± 0.081	0.647 ± 0.032	0.209 ± 0.097	0.302 ± 0.114
YOLO v4-128	0.655 ± 0.091	0.825 ± 0.094	0.680 ± 0.089	0.459 ± 0.089
YOLO v4-256	0.863 ± 0.027	0.940 ± 0.021	0.934 ± 0.022	0.714 ± 0.037
YOLO v4-608	0.751 ± 0.041	0.821 ± 0.001	0.831 ± 0.062	0.600 ± 0.059
YOLO v4 tiny-416#1	0.876 ± 0.019	0.871 ± 0.014	0.980 ± 0.033	0.777 ± 0.011
YOLO v4 tiny-416#2	0.611 ± 0.037	0.871 ± 0.014	0.735 ± 0.062	0.280 ± 0.026
YOLO v4 tiny-416#3	0.923 ± 0.038	0.990 ± 0.057	0.943 ± 0.033	0.835 ± 0.023
YOLO v4 tiny-416#4	0.652 ± 0.018	0.937 ± 0.008	0.740 ± 0.021	0.279 ± 0.024

Note: AP value was represented as value and its 95% confidence interval.

TABLE 4 The results of YOLO v4 tiny-416#3 model.

Parameters	Normal trachea	Grade 1–2 Tracheal collapse	Grade 3–4 Tracheal collapse	Average
Sensitivity	1.000	0.950	1.000	0.983 ± 0.008
False Negative Rate	0.000	0.050	0.000	0.017 ± 0.008
Specificity	1.000	1.000	0.975	0.992 ± 0.004
False Positive Rate	0.000	0.000	0.025	0.008 ± 0.004
Precision	1.000	1.000	0.952	0.984 ± 0.007
Accuracy	1.000	0.983	0.983	0.989 ± 0.003
Misclassification rate	0.000	0.017	0.017	0.011 ± 0.003
F1 score	1.000	0.974	0.976	0.983 ± 0.004
Negative predictive value	1.000	0.976	1.000	0.992 ± 0.004

Performance metrics analyzed were represented as value and 95% confidence interval.

radiographs *in the testing sets* and could well differentiate among a normal trachea (60/60, 100%), grade 1 and 2 tracheal collapse (59/60, 95%), and grade 3 and 4 tracheal collapse (60/60, 100%) *with* the highest sensitivity (98.30%), specificity (99.20%), precision (98.40%), NPV (99.20%), and accuracy (98.90%) compared the others for the screening and classification of the degree of canine tracheal collapse in this study. The results indicated that low misclassification rates and a high F*1* score were obtained. The overall results indicated that the YOLO v4 tiny-416#3 model outperformed all other models in the screening and classification of the degree of canine tracheal collapse on radiographic images. The performance of the YOLO v4 tiny-416#3 model is summarized in Table 4.

Furthermore, Cohen's kappa was used to evaluate an agreement between the deep learning model and radiologists in detecting and staging the severity of tracheal collapse. The interobserver agreement between the AI model and radiologist for each observer was $\varkappa = 0.975$ ($P < .001$) with excellent agreement, $\varkappa = 1.00$ ($P < .001$), which represented almost perfect consistency and satisfactory results of overall accuracy.

4 | DISCUSSION

The proposed model in this study demonstrated excellent performance with high precision and recall/sensitivity levels of > 98.3% in screening canine tracheal collapse and classifying its severity. The results showed that the same model as the YOLO v4 tiny model had different and varied performance. This may be due to the overfitting and underfitting of the data in each model. This study demonstrated that the well-trained YOLO v4 tiny-416#3 model has an excellent potential to screen and identify between a normal trachea and an abnormal trachea based on cervicothoracic radiographs with a satisfactory performance of 98.30% sensitivity and 99.20% specificity. Additionally, this model could clearly classify between a normal and collapsed trachea with 100.00% precision and accuracy of normal trachea detection.

As the nature of a retrospective study, although this study included radiographs obtained from two radiographic machines that provided different image qualities, the YOLO network models can correctly detect the tracheal location on all radiographs. In addition, the data augmentation technique was applied in this study to imitate the real-life scenario; increase more training sets and the variety of radiographic images, including overexposure, underexposure, and oblique images; and alleviate the overfitting problem.[36] Adding different levels of noise and blur in the augmentation process was not performed because these variations already existed in normal clinical radiographic images. Our study demonstrated that although there are several variations in the included radiographic images, such as different image qualities and different image acquisition from various thoracic conformations, the YOLO network models can accurately identify and classify the normal trachea and collapsed trachea of untrained radiographs because the model learns idiosyncratic characteristics or an individual characteristic from the training sets.[37] This aspect could be attributed to the proposed models being well-trained using a data set from the selected radiographic images. Consequently, the proposed model can be used in a real clinical practice setting to screen canine tracheal collapse and classify its severity.

In this study, all classification groups in the YOLO v4 tiny-416#3 model had high accuracy, sensitivity, specificity, and congruency with the veterinary radiologists. The grade 1 and 2 tracheal collapse group had a lower sensitivity (95.00%) than the normal trachea (100.00%) and grade 3 and 4 tracheal collapse (100.00%) groups. The lower sensitivity in the grade 1 and 2 tracheal collapse group may be because of the fact that in clinical situations, the TH/TI ratios are close to normal, which may be difficult to clearly classify. All groups had a high specificity which was normal trachea (100.00%), grade 1 and 2 tracheal collapse (100.00%), and grade 3 and 4 tracheal collapse (97.50%). Nonetheless, minimal information is available regarding the role of deep learning in the diagnosis of collapsed trachea on radiographs. Currently, only one study has reported the use of CNN in the detection of tracheal collapse.[38] However, that study has described only the error rates in detecting collapsed trachea on radiographic images in a small population of dogs and cats, which was approximately 6%–7%,[38] and the misclassification rate was satisfactory at only 1.1%.

Currently, the deep learning model has gained increasing attention in several diagnostic fields because of its certainty predictions[39] without the factors of exhaustion different from humans. In addition, this model can immediately detect lesions in a few seconds and assist in screening abnormalities in an extremely short time. Some cases of tracheal collapse are missed and underestimated in routine clinical practice, especially those in the early stage.[40] The proposed YOLO v4 tiny-416#3 model will assist veterinarians in screening canine tracheal collapse and classifying its severity. Additionally, it can reduce the workload of veterinarians and the error rate in tracheal collapse interpretation. Furthermore, the YOLO network models have no bias based on the history, signalment, and clinical signs of the animal patients.[38]

CiRA CORE works on the Robot Operating System that breaks the complex software into a smaller piece, which consequently eases its use and creates the framework, trains the model, and understands the software.[41] In addition, it is an open source that can work with several tools and software, which can develop and propagate worldwide. To efficiently train the models, the more data assembled can lead to the improvement of the model's performance.[42] In our study, the CiRA CORE platform using the YOLO v4 tiny-416#3 model provided high accuracy and satisfactory precision for the screening and classification among normal trachea, grade 1 and 2 tracheal collapse, and grade 3 and 4 tracheal collapse on plain lateral cervicothoracic radiographs. To the best of our knowledge, this is the first study to investigate the usefulness of deep learning in screening and classifying the degrees of tracheal collapse in dogs. In the future, deep learning models may play an important role in assisting radiologists and related specialists, along with general practitioners, in interpreting radiographic findings and diagnosing tracheal collapse in clinical practice in a short time with high accuracy and precision.

This study has several limitations. First, the number of radiographs included was limited, which were relatively few. To enhance the reliability of our findings and support the training of our model, a more extensive data set can be employed in future studies. This step is crucial for overall model performance enhancement, particularly in terms of sensitivity and mAP. Second, due to the retrospective nature of the study, we encountered a challenge in ensuring the selection of solely fully inspired thoracic radiographs. This inherent limitation stems from the unalterable nature of past data collection. Third, since all images in both the training and validation data sets were labeled by the same professional radiologist, the inherent bias in the data sets and accuracy measurements may occur.

In conclusion, the proposed YOLO network model in this study exhibited outstanding performance by achieving high levels of sensitivity, specificity, accuracy, and mAP for screening and classifying tracheal conditions—normal trachea, grade 1 and 2 tracheal collapse, and grade 3 and 4 tracheal collapse—on plain lateral cervicothoracic radiographs. The performance of the model was excellent, showing accuracy comparable to that of veterinary radiologists. Furthermore, a high level of agreement was revealed between the assessments of the deep learning model and veterinary radiologists, indicating a strong congruence between the two. The findings demonstrate the effectiveness of the proposed model in screening for tracheal collapse in dogs and aid in the classification of the degree of tracheal collapse. This, in turn, can provide valuable guidance for further selection of the most suitable treatment options in routine clinical practice. Additionally, the integration of this may enhance the overall quality of care and reduce the time required for radiographic interpretation in clinical settings.

LIST OF AUTHOR CONTRIBUTIONS

Category 1

(a) Conception and design: Choisunirachon, Tongloy, Chuwongin, Boonsang, Kittichai, Thanaboonnipat
(b) Acquisition of data: Suksangvoravong, Kittichai, Thanaboonnipat
(c) Analysis and interpretation of data: Suksangvoravong, Kittichai, Thanaboonnipat

Category 2

(a) Drafting the article: Suksangvoravong, Boonsang, Kittichai, Thanaboonnipat

(b) Revising article for intellectual content: Suksangvoravong, Choisunirachon, Tongloy, Chuwongin, Boonsang, Kittichai, Thanaboonnipat

Category 3

(a) Final approval of the completed article: Suksangvoravong, Choisunirachon, Tongloy, Chuwongin, Boonsang, Kittichai, Thanaboonnipat

ACKNOWLEDGEMENTS

The authors thank The Small Animal Hospital, Faculty of Veterinary Science, Chulalongkorn University for providing all the data. The authors also thank the Department of Medicine, Faculty of Medicine, King Mongkut's Institute of Technology Ladkrabang, Bangkok, and College of Advanced Manufacturing Innovation, King Mongkut's Institute of Technology Ladkrabang for supporting CiRA CORE software and all assistance.

CONFLICT OF INTEREST STATEMENT

The authors declare no conflict of interest.

DATA AVAILABILITY STATEMENT

The data are available from the corresponding author upon reasonable request.

PREVIOUS PRESENTATION

This study was presented as a poster at the CUVC 2023 (Chulalongkorn University Veterinary Conference).

REPORTING CHECKLIST DISCLOSURE

The authors declare no checklist was used.

ORCID

Hathaiphat Suksangvoravong https://orcid.org/0009-0006-6294-6653
Nan Choisunirachon https://orcid.org/0000-0001-7169-6699
Teerawat Tongloy https://orcid.org/0009-0001-0670-3124
Santhad Chuwongin https://orcid.org/0000-0001-7064-9911
Siridech Boonsang https://orcid.org/0000-0003-2866-8902
Veerayuth Kittichai https://orcid.org/0000-0003-4320-5922
Chutimon Thanaboonnipat https://orcid.org/0000-0002-4548-549X

REFERENCES

1. Whit RA, Williams JM. Tracheal collapse in the dog—is there really a role for surgery? A survey of 100 cases. *J Small Anim Pract.* 2008;35:191-196.
2. Lund EM, Armstrong PJ, Kirk CA, Klausner JS. Prevalence and risk factors for obesity in adult dogs from private US veterinary practices. *Int J Appl Res Vet Med*. 2006;4:177.
3. Tappi S. Canine tracheal collapse. *J Small Anim Pract*. 2016;57:9-17.
4. Maggior Ann. Tracheal and airway collapse in dogs. *Vet Clin North Am Small Anim*. 2014;44:117-127.
5. Mostafa AA, Berry CR. Radiographic vertical tracheal diameter assessment at different levels along the trachea as an alternative method for the evaluation of the tracheal diameter in non-brachycephalic small breed dogs. *BMC Vet Res*. 2022;18:61.
6. Coyne BE, Fingland RB. Hypoplasia of the trachea in dogs: 103 cases (1974-1990). *J Am Vet Med Assoc*. 1992;201:768-772.
7. Fingland RB, Layton CI, Kennedy GA, Galland JC. Comparison of simple continuous versus simple interrupted suture patterns for tracheal anastomosis after large-segment tracheal resection in dogs. *Vet Surg*. 1995;24:320-330.
8. Kaye BM, Boroffka SA, Haagsman AN, Ter Haar G. Computed tomographic, radiographic, and endoscopic tracheal dimensions in English bulldogs with grade 1 clinical signs of brachycephalic airway syndrome. *Vet Radiol Ultrasound*. 2015;56:609-616.
9. Harvey CE, Fink EA. Tracheal diameter: analysis of radiographic measurements in brachycephalic and nonbrachycephalic dogs. *J Am Vet Med Assoc*. 1982;18:570-576.
10. Regier PJ, Grosso FV, Stone HK, van Santen E. Radiographic tracheal dimensions in brachycephalic breeds before and after surgical treatment for brachycephalic airway syndrome. *The Can Vet J*. 2020;61:971-976.
11. Hammond G, Geary M, Coleman E, Gunn-Moore D. Radiographic measurements of the trachea in domestic shorthair and Persian cats. *J Feline Med and Surg*. 2011;13:881-884.
12. Banzat T, Wodzinski M, Burti S, Osti VL, et al. Automatic classification of canine thoracic radiographs using deep learning. *Sci Rep*. 2021;11:3964.
13. Tangne C, Hobson H. Retrospective study of 20 surgically managed cases of collapsed trachea. *Vet Surg*. 1982;11:146-149.
14. Onder O, Yarasir Y, Azizova A, Durhan G, Onur MR, Ariyurek OM. Errors, discrepancies and underlying bias in radiology with case examples: a pictorial review. *Insights Imaging*. 2021;12:51.
15. Bruno MA, Walker EA, Abujudeh HH. Understanding and confronting our mistakes: the epidemiology of error in radiology and strategies for error reduction. *Radiographics*. 2015;35:1668-1676.
16. Herman PG, Hessel SJ. Accuracy and its relationship to experience in the interpretation of chest radiographs. *Invest Radiol*. 1975;10:62-67.
17. Kelly BS, Rainford LA, Darcy SP, Kavanagh EC, Toomey RJ. The development of expertise in radiology: in chest radiograph interpretation, "expert" search pattern may predate "expert" levels of diagnostic accuracy for pneumothorax identification. *Radiology*. 2016;280:252-260.
18. Humphrie A, Shaheen A, Gómez Álvarez CB. Different conformations of the German shepherd dog breed affect its posture and movement. *Sci Rep*. 2020;10:16924. C.B.
19. Burt S, Longhin Osti V, Zotti A, Banzato T. Use of deep learning to detect cardiomegaly on thoracic radiographs in dogs. *Vet J*. 2020;262:105505.
20. Waite S, Scott J, Gale B, Fuchs T, Kolla S, Reede D. Interpretive error in radiology. *Am J Roentgenol*. 2017;208:739-749.
21. Alexander K. Reducing error in radiographic interpretation. *Can Vet J*. 2010;51:533-536.
22. Rajaraman S, Antani SK, Poostchi M, et al. Pre-trained convolutional neural networks as feature extractors toward improved malaria parasite detection in thin blood smear images. *PeerJ*. 2018;6:e4568.
23. Pang J, Chen K, Shi J, Feng H, Ouyang W, Lin D, in Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. 2019;821-830.
24. Zhou G, Zhang W, Chen A, He M, Ma X. Rapid detection of rice disease based on FCM-KM and faster R-CNN fusion. *IEEE access*. 2019;7:143190-143206.
25. Li S, Wang Z, Visser LC, Wisner ER, Cheng H. Pilot study: application of artificial intelligence for detecting left atrial enlargement on canine thoracic radiographs. *Vet Radiol Ultrasound*. 2020;61:611-618.

26. Banzato T, Wodzinski M, Tauceri F. An AI-based algorithm for the automatic classification of thoracic radiographs in cats. *Front Vet Sci*. 2021;8:1-7.
27. Adrien-Maxence H, Emilie B, Alois C, 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*. 2022;63:456-468.
28. Kim E, Fischetti AJ, Sreetharan P, Weltman JG, Fox PR. Comparison of artificial intelligence to the veterinary radiologist's diagnosis of canine cardiogenic pulmonary edema. *Vet Radiol Ultrasound*. 2022;63:292-297.
29. Müller TR, Solano M, Tsunemi MH. Accuracy of artificial intelligence software for the detection of confirmed pleural effusion in thoracic radiographs in dogs. *Vet Radiol Ultrasound*. 2022;2:1-6.
30. Redmon J, Farhadi A, YOLOv3:An Incremental Improvement.arXiv:1804.02767[cs.CV](8Apr2018).
31. Kittichai V, Pengsakul T, Chumchuen K. Deep learning approaches for challenging species and gender identification of mosquito vectors. *Sci Rep*. 2021;11:4838.
32. Boonrod A, Boonrod A, Meethawolgul A, Twinprai P. Diagnostic accuracy of deep learning for evaluation of C-spine injury from lateral neck radiographs. *Heliyon*. 2022;8:e10372.
33. Twinprai N, Boonrod A, Boonrod A. Artificial intelligence (AI) vs. human in hip fracture detection. *Heliyon*. 2022;8:e11266.
34. Pongsakonpruttikul N, Angthong C, Kittichai V. Artificial intelligence assistance in radiographic detection and classification of knee osteoarthritis and its severity: a cross-sectional diagnostic study. *Eur Rev Med Pharmacol Sci*. 2022;26:1549-1558.
35. Reiner Carol, Masseau I. Lower airway collapse: revisiting the definition and clinicopathologic features of canine bronchomalacia. *Vet J*. 2021;273:105682.
36. Shim H, Lee J, Choi S. Deep learning-based diagnosis of stifle joint diseases in dogs. *Vet Radiol Ultrasound*. 2023;64:113-122.
37. Cohe I. An artificial neural network analogue of learning in autism. *Biol Psychiatry*. 1994;36:5-20.
38. Boissad 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.
39. Biercher A, Meller S, Wendt J, et al. Using deep learning to detect spinal cord diseases on thoracolumbar magnetic resonance images of dogs. *Front Vet Sci*. 2021;8:721167.
40. Clark D. Interventional radiology management of tracheal and bronchial collapse. *Vet Clin North Am Small Anim Pract*. 2018;48:765-779.
41. Padmaj B, Myneni MB, Krishna Rao Patro E. Comparison on visual prediction models for MAMO (multi activity-multi object) recognition using deep learning. *J Big Data*. 2020;7:1-15.
42. Torii M, Wagholikar K, Liu H. Using machine learning for concept extraction on clinical documents from multiple data sources. *J Am Med Inform Assoc*. 2011;18:580-587.

How to cite this article: Suksangvoravong H, Choisunirachon N, Tongloy T, et al. Automatic classification and grading of canine tracheal collapse on thoracic radiographs by using deep learning. *Vet Radiol Ultrasound*. 2024;1-10. https://doi.org/10.1111/vru.13413