EdgeANet: A Transformer-based Edge Representation Learning Network for Canine X-ray Verification

In-Gyu Lee[1[0000−0001−7038−4618]], Jun-Young Oh[1[0000−0002−4378−4082]], Hyewon Choi[2[0009−0003−0686−9452]], Tae-Eui Kam[3[0000−0002−6677−7176]], Namsoon Lee[2[0000−0002−9091−6957]], Sang-Hwan Hyun[2[0000−0001−6011−2555]], Euijong Lee[1[0000−0002−7308−7392]], and Ji-Hoon Jeong[1*[0000−0001−6940−2700]]

[1] Department of Computer Science, Chungbuk National University, Cheongju, Republic of Korea
{ingyu.lee,jy.oh,kongjjagae,jh.jeong}@chungbuk.ac.kr
[2] College of Veterinary Medicine, Chungbuk National University, Cheongju, Republic of Korea
{hyewon0203,ultravet,shhyun}@chungbuk.ac.kr
[3] Department of Artificial Intelligence, Korea University, Seoul, Republic of Korea
kamte@korea.ac.kr
[*] Corresponding author

Abstract. Artificial intelligence (AI) has shown great potential in medical imaging, yet its adoption in veterinary medicine remains limited due to data scarcity and anatomical complexity. This study introduces a novel transformer-based edge representation learning network for verifying rotated vertebral bodies in canine thoracic X-ray images. The proposed method integrates a localization module to identify the spinous process, a transformer encoder for global feature extraction using a self-attention mechanism, and an edge encoder to enhance feature extraction of fine-grained details, improving classification performance. Experimental results demonstrate that our method achieves superior accuracy, precision, and recall, outperforming state-of-the-art (SOTA) methods with a classification accuracy of 0.7838. Furthermore, the ablation study confirms that including the proposed encoders significantly impacts performance, demonstrating their effectiveness in improving classification accuracy. These findings highlight the importance of multi-scale feature extraction in veterinary imaging and suggest that EdgeANet can be a valuable tool for AI-assisted X-ray verification in veterinary and human medical applications.

Keywords: Veterinary AI · Edge detection · X-ray verification · Deep learning · Medical imaging

1 Introduction

Artificial intelligence (AI) is revolutionizing data analysis, enabling decision-making and predictive capabilities [1,2]. In medicine, AI supports disease diagnosis, treatment planning, and prognosis prediction [3,4]. Integrating AI into veterinary imaging also has the potential to standardize diagnosis, reduce human error, and enhance clinical efficiency [5,6]. As pets are increasingly seen as family members, there is a growing need for disease prevention, early diagnosis, and effective treatments [7]. However, the adoption of AI in veterinary medicine, similar to human healthcare, remains slow due to challenges such as the need for high-quality labeled datasets, the complexity of anatomical and behavioral variations, and a shortage of specialized AI practitioners familiar with veterinary applications [8,9]. Despite these challenges, AI holds great potential to enhance diagnostic accuracy, reduce workload for veterinarians, and improve overall pet healthcare [10].

In veterinary medicine, imaging is essential since animal patients cannot communicate symptoms [11]. While computed tomography and magnetic resonance imaging provide high-resolution insights, they require anesthesia, making them costly and risky [12]. Consequently, radiography is more widely used due to its affordability and accessibility, emphasizing the need to validate X-ray image quality [13]. Due to its diagnostic importance, AI-driven X-ray analysis is an active research area in veterinary medicine. Li et al. [14] developed a regressive vision transformer model for canine cardiomegaly classification using vertebral heart scale prediction, achieving state-of-the-art (SOTA) performance and improving interpretability for clinicians. Banzato et al. [13] conducted the first study assessing canine thoracic X-ray quality, highlighting the impact of technical errors like improper positioning on diagnostic reliability. AI-based methods have also been explored for diagnosing conditions such as stifle joint diseases and hip dysplasia [15,16,5]. Given the critical role of X-rays in veterinary diagnostics, ensuring proper image acquisition is essential.

Medical imaging in veterinary medicine relies on structural analysis, where fine-grained details such as bone contours, fractures, and subtle misalignments are crucial for diagnosis [17,18]. Traditional deep learning models, especially convolutional neural networks (CNN), primarily focus on learning high-level features, often neglecting small-scale structural variations [19]. However, in radiographic images, the presence of minor rotations, deformations, or anomalies in vertebral structures can critically impact the accuracy of diagnosis [20]. In this study, edge representation learning is critical in highlighting these structural changes. Enhancing edge-based features enables a model to better distinguish between normal and abnormal cases, improving classification accuracy.

Our contributions can be summarized as follows:

- We propose EdgeANet, a novel transformer-based edge representation learning method for identifying small-scale errors, such as rotated spinous processes, in thoracic X-ray images. The proposed method determines whether X-ray images are accurately captured by detecting small objects that indicate common imaging errors, such as spinal rotation.
- We develop a module that enhances model performance in detecting minor errors and improving X-ray verification and medical imaging applications. This method has the potential to advance veterinary diagnostics and extend to human medical imaging in the future.

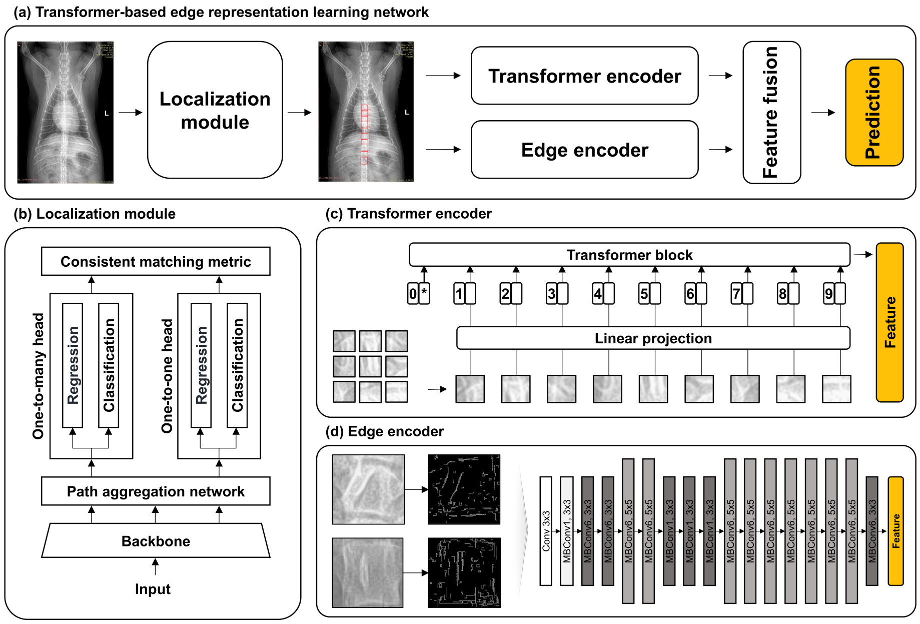

Fig. 1. Overall flow of the proposed network. The proposed network takes a raw image as input and identifies the spinous process through the localization module shown in (b). Subsequently, global and fine-grained features are extracted through the transformer encoder in (c) and the edge encoder in (d). These features are then fused to produce the final classification result.

2 Methods

Fig. 1. illustrates our proposed transformer-based edge representation learning network for verifying thoracic X-ray images by detecting rotated spinous processes. The network is composed of three components. The first component, the localization module, identifies the spinous process within the vertebrae. The second component, the transformer encoder, extracts global contextual features, while the third component, the edge encoder, captures fine-grained edge details to enhance structural analysis. Thoracic dorsoventral and ventrodorsal X-ray images are first processed through the localization module, which detects the spinous process. The identified region then extracts features through the transformer encoder and edge encoder, enabling the model to learn global and edge representations. These extracted features are fused using summation to classify the spinous process as normal or abnormal.

2.1 Localization module

In this study, the architecture of YOLOv10 [21] was employed for image localization, as illustrated in Fig. 1(b). The localization module enhances real-time object detection by improving efficiency and accuracy through a dual-label assignment strategy. Integrating one-to-many and one-to-one matching during training eliminates reliance on non-maximum suppression during inference. In addition, a consistent matching metric aligns supervision objectives, reducing redundancy and enhancing detection precision. These improvements enable end-to-end deployment with minimal latency. Despite these advancements, the model faced challenges identifying the spinous process due to overlapping structures and low contrast in X-ray images. To address this, radiology specialists manually labeled based on standardized anatomical criteria, ensuring precise and clinically relevant annotations.

2.2 Transformer encoder

To enhance the performance of spinous process identification, an architecture of vision transformer (ViT) was integrated into the network [22]. The architecture of ViT is illustrated in Fig. 1(c). The model uses self-attention mechanisms to capture long-range dependencies within an image, making it particularly effective in distinguishing fine-grained features that convolution-based models may overlook [23]. The ViT model was pre-trained on ImageNet and fine-tuned using the labeled dataset. The training process involved augmentations such as rotation, flipping, and contrast adjustments to simulate variations encountered in clinical settings. This method improved the model's ability to discern subtle differences in the spinous process region, even in challenging cases. Self-attention computation includes a relative position bias $B \in \mathbb{R}^{M^2 \times M^2}$ for each head. The self-attention mechanism of the transformer block is formulated as follows:

$$\text{Attention}(Q, K, V) = \text{SoftMax}\left(\frac{QK^T}{\sqrt{d}} + B\right)V \tag{1}$$

where $Q, K, V \in \mathbb{R}^{M^2 \times d}$ are the query, key, and value matrices, d is the dimension of the query and key, and M^2 is the number of patches in the window.

2.3 Edge encoder

An edge feature extraction method was incorporated to refine the detection and classification process. The architecture of the edge encoder is illustrated in Fig. 1(d). Canny edge detection was applied during preprocessing to highlight critical structures while suppressing background noise [24]. It was selected for its robust Gaussian noise filtering and effective non-maximum suppression with tunable thresholds, which enabled reliable edge enhancement on X-ray images and contributed to performance gains. The lower and upper thresholds for Canny edge detection were set to fixed intensity values of 40 and 75 respectively based on the 0 to 255 grayscale pixel range. The extracted edge features were fused with the transformer encoder outputs to construct a hybrid feature representation. This integration improved the model's ability to identify spinous processes, particularly under low contrast or anatomical overlap. An example output of the Canny edge detection is shown in Fig. 2.

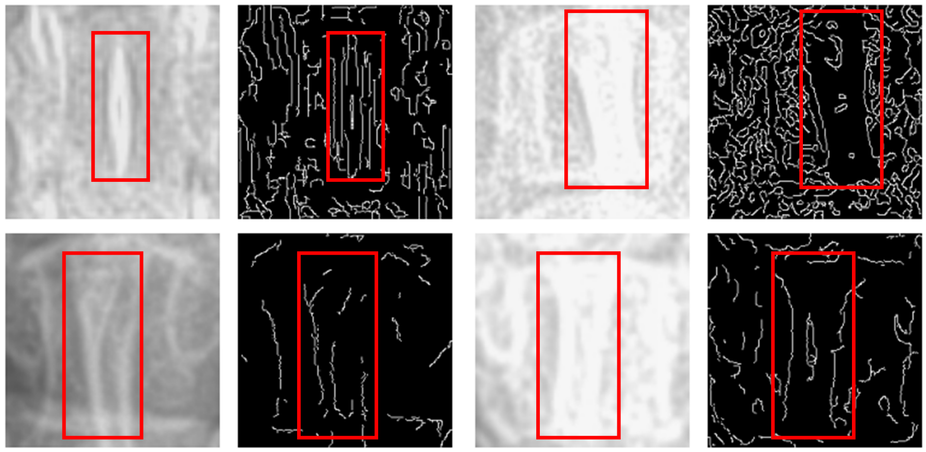

Fig. 2. Example of Canny edge detection. This method can make the networks capture local features of the spinous process.

3 Experiments

3.1 Dataset and experimental setup

Dataset. The abnormal and normal spinous processes are depicted in Fig. 3. Radiographic images were obtained using appropriate exposure settings based on abdominal thickness measurements, 200 to 300 mA, 60 to 80 kVp, and a focal film distance of 100 cm. Thorax radiographic images were taken using a digital radiographic system (Toshiba RotanodeTM, Tokyo, Japan). These images were processed with a radiographic system (BLADE v1., Median International Co., Anyang, Republic of Korea).

X-ray images taken for disease diagnosis were filtered and used to select the study subjects. Among the captured X-ray images, those in which the vertebral bodies were rotated were collected and used as abnormal data. In contrast, non-rotated images were collected as the control group and used as normal data. To mitigate annotator bias and ensure objectivity in the absence of external datasets for validation, labeling was conducted through rigorous multi-cross validation by eight veterinarians, each with 1 to 5 years of clinical experience and 1 to 2 years of specialization in veterinary radiology. The labeling criteria were based on the Textbook of Veterinary Diagnostic Radiology by Donald Thrall, one of the most authoritative references in the field [25]. To ensure consistency, only cases confirmed by at least six of the eight veterinarians were included in the dataset. The dataset included samples from various species, comprising 90 abnormal and 100 normal thoracic X-ray images. Although the sample size is limited, this reflects the practical realities of veterinary clinical imaging.

Experimental setup. Our experiments were conducted on Ryzen 7 7800X3D CPU and four RTX 4090 GPUs. The number of epochs for all methods was set to

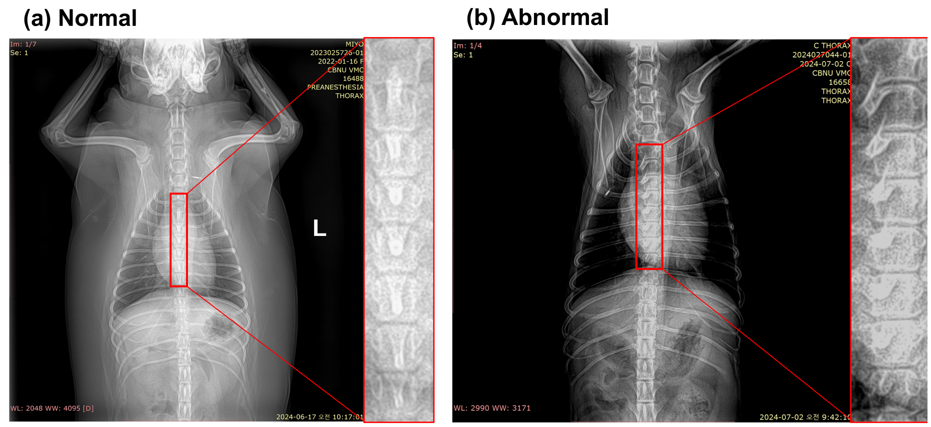

Fig. 3. Example of thoracic X-ray images. (a) presents a normal thoracic X-ray image without spinal rotation, while (b) presents an abnormal thoracic X-ray image showing rotated spinous processes.

100, with the Adam optimizer and a learning rate of 2e-4. The dataset was first split into a 4:1 ratio for the train and test sets. Then, 4-fold cross-validation was applied to the train set for model evaluation. Quantitative metrics, including classification accuracy, precision, recall, F1-score, and confusion matrix, were used to assess the model's performance [26].

3.2 Comparison with SOTA Methods

The study compared the performance of various classification models, including transformer-based architectures such as Swin Transformer [27], as well as CNN-based models like ConvNeXt [28], ResNet [29], DenseNet [30], and EfficientNet [31]. These models achieve SOTA performance or demonstrate performance close to it on benchmark datasets such as ImageNet for image classification tasks. Comparing our model with these architectures allows us to evaluate its efficiency and effectiveness.

The left side of the table presents the highest performance achieved during 4-fold cross-validation. In contrast, the right side shows the test results of the best-performing model from each fold, with the average and standard deviation calculated accordingly. The experimental results demonstrate that EdgeANet significantly outperforms other state-of-the-art methods in validation and test performance across multiple evaluation metrics. It consistently achieves the highest accuracy across multiple folds, with values reaching up to 0.8684 in fold 3, which is on par or superior to the best-performing models in individual folds.

EdgeANet achieves the highest average test accuracy of 0.7838, clearly outperforming all competing models including EfficientNet. In addition, it registers the highest precision of 0.6962, recall of 0.8623, and F1-score of 0.7813, demonstrating superior accuracy and a balance between precision and recall. Unlike

Table 1. Quantitative evaluation metrics of 4-fold cross-validation and test performance. We highlight the best performance in bold.

Model	Best validation performance				Average test performance			
	Fold 1	Fold 2	Fold 3	Fold 4	Accuracy	Precision	Recall	F1-score
Swin Transformer [27]	0.7632	0.7895	0.7632	0.7368	0.6411 (± 0.0336)	0.4183 (± 0.0221)	0.6232 (± 0.0140)	0.5691 (± 0.0055)
ResNet [29]	0.7368	0.7895	0.8421	0.8158	0.6619 (± 0.0107)	0.5747 (± 0.0808)	0.6911 (± 0.0308)	0.6405 (± 0.0367)
DenseNet [30]	**0.8158**	**0.8684**	0.8421	**0.8421**	0.6773 (± 0.0141)	0.5639 (± 0.0190)	0.7311 (± 0.0207)	0.6722 (± 0.0140)
EfficientNet [31]	0.7895	**0.8684**	0.8421	**0.8421**	0.6847 (± 0.0303)	0.5967 (± 0.0703)	0.7077 (± 0.0276)	0.6722 (± 0.0483)
ConvNeXt [28]	**0.8158**	**0.8684**	0.7368	0.7632	0.7114 (± 0.0466)	0.5845 (± 0.0690)	0.7748 (± 0.0506)	0.7078 (± 0.0478)
EdgeANet (Ours)	**0.8158**	0.8158	**0.8684**	**0.8421**	**0.7838** (± 0.0310)	**0.6962** (± 0.0276)	**0.8623** (± 0.0258)	**0.7813** (± 0.0271)

ConvNeXt or DenseNet, which achieve competitive results in specific folds but fail to maintain consistency across test metrics, EdgeANet ensures stable and superior performance across all evaluation criteria.

EdgeANet's high recall score of 0.8623 suggests that it captures more relevant instances than other models, making it particularly useful in applications where missing key classifications could have significant consequences. These results establish EdgeANet as the leading model, highlighting its robustness and practical applicability in real-world scenarios that demand high accuracy and generalization.

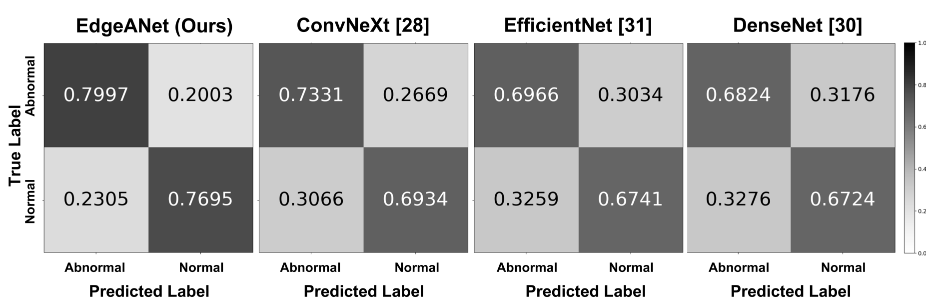

Fig. 4. Confusion matrix of the four best-performing models.

Our experimental results are visualized through the confusion matrix presented in Fig. 4, which includes only the top four models with the highest test accuracy. This study aims to accurately distinguish between normal and abnormal data by effectively extracting features from small structures, such as the spinous process, in X-ray images for verification. EdgeANet demonstrates outstanding performance in this task, having achieved a TP of over 0.7997, outperforming EfficientNet, DenseNet, and ConvNeXt, which achieved a TP in the range of 0.6824 to 0.7331. Regarding the normal class, the proposed method shows performance with an FP of 0.7695, while others showed lower reliability, leading to a higher misclassification rate for normal cases.

The model uses a transformer encoder to extract global features and an edge encoder to capture local structural details, allowing for a more comprehensive representation of the X-ray images. By combining these feature extraction techniques, the model enhances classification accuracy, particularly in detecting subtle abnormalities that conventional methods may overlook.

Table 2. Ablation study of adapted encoders.

Model	Transformer encoder	Edge encoder	Accuracy
Baseline	✗	✗	0.7368
w/o Edge encoder	✓	✗	0.7632
EdgeANet (Ours)	✓	✓	**0.8158**

3.3 Ablation study

In Table 2, we present an ablation study on different components of our network. The results demonstrate that our methods enhance classification accuracy by extracting deep features through edge-based representation learning. The baseline model achieved a classification accuracy of 0.7368 without any of the proposed encoders. When the transformer encoder was incorporated, the performance improved to 0.7632. Finally, with the addition of our edge encoder, the proposed model achieved a classification accuracy of 0.8158.

4 Conclusion

This study proposed a transformer-based edge representation learning network for classifying rotated vertebral bodies in canine thoracic X-ray images. By integrating a localization module, a transformer encoder, and an edge encoder, our model effectively captures global and fine-grained features critical for accurate classification. Comparative analysis against widely used deep learning models demonstrated that our method achieves higher classification accuracy and improved recall rates, reinforcing its robustness for veterinary applications. The findings highlight the importance of multi-aspect feature extraction in enhancing the precision of diagnostic imaging models, particularly in cases where traditional CNN-based architectures struggle with subtle anatomical variations. Although the dataset size is limited due to the practical constraints of veterinary clinical imaging, our results provide a foundation for future study. We plan to expand the dataset using generative augmentation techniques such as generative adversarial networks to improve generalizability, adapt the network for other anatomical regions, and explore real-world deployment in veterinary clinics. This study underscores the potential of AI-assisted imaging analysis in advancing veterinary diagnostics and ensuring higher standards of pet healthcare.

Acknowledgments. This work was partly supported by Innovative Human Resource Development for Local Intellectualization program through the Institute of Information & Communications Technology Planning & Evaluation (IITP) grant funded by the Korea government (MSIT) (IITP-2025-RS-2020-II201462, 20%), grant funded by the Institute of Information & Communications Technology Planning & Evaluation (IITP) grant funded by the Korea government (MSIT) (RS-2021-II212068, Artificial Intelligence Innovation Hub), Korea Institute for Advancement of Technology (KIAT) grant funded by the Korea government (MOTIE) (No. P0025399), grant funded by the Korea Institute of Planning and Evaluation for Technology in Food, Agriculture and Forestry (IPET) through the Agriculture and Food Convergence Technologies Program for Research Manpower Development Program funded by the Ministry of Agriculture, Food and Rural Affairs (MAFRA) (RS-2024-00398561), and grant funded by the National Research Foundation of Korea (NRF) grant funded by the Korea government(MSIT) (RS-2024-00456480).

Disclosure of Interests. The authors have no competing interests to declare that are relevant to the content of this article.

References

1. Davenport, T., Kalakota, R.: The potential for artificial intelligence in healthcare. Future Healthcare Journal 6(2), 94–98 (2019)
2. Hou, J., Xu, J., Chen, H.: Concept-attention whitening for interpretable skin lesion diagnosis. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. pp. 113–123. Springer (2024)
3. Topol, E.J.: High-performance medicine: The convergence of human and artificial intelligence. Nature medicine 25(1), 44–56 (2019)
4. Biswas, K., Jha, D., Tomar, N.K., Karri, M., Reza, A., Durak, G., Medetalibeyoglu, A., Antalek, M., Velichko, Y., Ladner, D., et al.: Adaptive smooth activation function for improved organ segmentation and disease diagnosis. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. pp. 65–74. Springer (2024)
5. Hennessey, E., DiFazio, M., Hennessey, R., Cassel, N.: Artificial intelligence in veterinary diagnostic imaging: A literature review. Veterinary Radiology & Ultrasound 63, 851–870 (2022)
6. Bouhali, O., Bensmail, H., Sheharyar, A., David, F., Johnson, J.P.: A review of radiomics and artificial intelligence and their application in veterinary diagnostic imaging. Veterinary Sciences 9(11), 620 (2022)
7. Ly, L.H., Gordon, E., Protopopova, A.: Exploring the relationship between human social deprivation and animal surrender to shelters in british columbia, canada. Frontiers in Veterinary Science 8, 656597 (2021)
8. Benjamens, S., Dhunnoo, P., Meskó, B.: The state of artificial intelligence-based FDA-approved medical devices and algorithms: an online database. npj Digital Medicine 3(1), 118 (2020)

9. Yuan, Z., Fang, Z., Huang, Z., Wu, F., Yao, Y.F., Li, Y.: Adapting pre-trained generative model to medical image for data augmentation. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. pp. 79–89. Springer (2024)
10. Appleby, R.B., Basran, P.S.: Artificial intelligence in veterinary medicine. Journal of the American Veterinary Medical Association 260(8), 819–824 (2022)
11. Burti, S., Banzato, T., Coghlan, S., Wodziniski, M., Bendazzoli, M., Zotti, A.: Artificial intelligence in veterinary diagnostic imaging: Perspectives and limitations. Research in Veterinary Science p. 105317 (2024)
12. Portier, K., Ida, K.K.: Anesthetic risk and complications in veterinary medicine (2020)
13. Banzato, T., Wodzinski, M., Burti, S., Vettore, E., Muller, H., Zotti, A.: An AI-based algorithm for the automatic evaluation of image quality in canine thoracic radiographs. Scientific Reports 13(1), 17024 (2023)
14. Li, J., Zhang, Y.: Regressive vision transformer for dog cardiomegaly assessment. Scientific Reports 14(1), 1539 (2024)
15. Shim, H., Lee, J., Choi, S., Kim, J., Jeong, J., Cho, C., Kim, H., Kim, J.i., Kim, J., Eom, K.: Deep learning-based diagnosis of stifle joint diseases in dogs. Veterinary Radiology & Ultrasound 64(1), 113–122 (2023)
16. McEvoy, F.J., Proschowsky, H.F., Müller, A.V., Moorman, L., Bender-Koch, J., Svalastoga, E.L., Frellsen, J., Nielsen, D.H.: Deep transfer learning can be used for the detection of hip joints in pelvis radiographs and the classification of their hip dysplasia status. Veterinary Radiology & Ultrasound 62(4), 387–393 (2021)
17. Steffey, M.A., Garcia, T.C., Gorney, A., Stover, S.M., Zwingenberger, A.L.: Computed tomography-derived structural analysis for the likelihood of pathologic fracture in canine antebrachial osteosarcoma. Veterinary Radiology & Ultrasound (2024)
18. Soltero-Rivera, M.M., Nguyen, R., Goldschmidt, S.L., Hatcher, D.C., Arzi, B.: Diagnostic yield of dental radiography and digital tomosynthesis for the identification of anatomic structures in cats. Frontiers in Veterinary Science 11, 1408807 (2024)
19. Azulay, A., Weiss, Y.: Why do deep convolutional networks generalize so poorly to small image transformations? Journal of Machine Learning Research 20(184), 1–25 (2019)
20. Miao, K.H., Miao, J.H., Belani, P., Dayan, E., Carlon, T.A., Cengiz, T.B., Finkelstein, M.: Radiological diagnosis and advances in imaging of vertebral compression fractures. Journal of Imaging 10(10), 244 (2024)
21. Wang, A., Chen, H., Liu, L., Chen, K., Lin, Z., Han, J., Ding, G.: Yolov10: Real-time end-to-end object detection. arXiv preprint arXiv:2405.14458 (2024)
22. Dosovitskiy, A.: An image is worth 16x16 words: Transformers for image recognition at scale. arXiv preprint arXiv:2010.11929 (2020)
23. Mehta, S., Rastegari, M.: Mobilevit: Light-weight, general-purpose, and mobile-friendly vision transformer. arXiv preprint arXiv:2110.02178 (2021)
24. Canny, J.: A computational approach to edge detection. IEEE Transactions on Pattern Analysis and Machine Intelligence (6), 679–698 (1986)
25. Thrall, D.E.: Textbook of veterinary diagnostic radiology. Elsevier Health Sciences (2012)
26. Goutte, C., Gaussier, E.: A probabilistic interpretation of precision, recall and f-score, with implication for evaluation. In: European Conference on Information Retrieval. pp. 345–359. Springer (2005)

27. Liu, Z., Lin, Y., Cao, Y., Hu, H., Wei, Y., Zhang, Z., Lin, S., Guo, B.: Swin transformer: Hierarchical vision transformer using shifted windows. In: IEEE/CVF International Conference on Computer Vision. pp. 10012–10022 (2021)
28. Liu, Z., Mao, H., Wu, C.Y., Feichtenhofer, C., Darrell, T., Xie, S.: A ConvNet for the 2020s. In: IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 11976–11986 (2022)
29. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: IEEE Conference on Computer Vision and Pattern Recognition. pp. 770–778 (2016)
30. Huang, G., Liu, Z., Van Der Maaten, L., Weinberger, K.Q.: Densely connected convolutional networks. In: IEEE Conference on Computer Vision and Pattern Recognition. pp. 4700–4708 (2017)
31. Tan, M., Le, Q.: Efficientnet: Rethinking model scaling for convolutional neural networks. In: International Conference on Machine Learning. pp. 6105–6114. PMLR (2019)