AJVR

AVMA

Deep learning model shows promise for detecting and grading sesamoiditis in horse radiographs

Li Guo, BSc[1*]; Xinhui Yu, MSc[1]; Anas Thair, MSc[1]; Andrew Rideout[2]; Andrew Collins, MVB[3]; Z. Jane Wang, PhD[1]; Michael Hore, MVB[4*]

[1]Department of Electrical and Computer Engineering, University of British Columbia, British Columbia, Canada
[2]Point to Point Research Development, British Columbia, Canada
[3]Baker McVeigh and Clements, Newmarket, England
[4]Hagyard Equine Medical Institute, Lexington, Kentucky

*Corresponding authors: Dr. Guo (lguo@ece.ubc.ca) and Dr. Hore (michaelhore@gmail.com)

Received August 2, 2023
Accepted September 29, 2023

doi.org/10.2460/ajvr.23.07.0173

OBJECTIVE
The objective of this study was to develop a robust machine-learning approach for efficient detection and grading of sesamoiditis in horses using radiographs, specifically in data-limited conditions.

SAMPLE
A dataset of 255 dorsolateral-palmaromedial oblique (DLPMO) and dorsomedial-palmarolateral oblique (DMPLO) equine radiographs were retrospectively acquired from Hagyard Equine Medical Institute. These images were anonymized and classified into 3 categories of sesamoiditis severity (normal, mild, and moderate).

METHODS
This study was conducted from February 1, 2023, to August 31, 2023. Two RetinaNet models were used in a cascaded manner, with a self-attention module incorporated into the second RetinaNet's classification subnetwork. The first RetinaNet localized the sesamoid bone in the radiographs, while the second RetinaNet graded the severity of sesamoiditis based on the localized region. Model performance was evaluated using the confusion matrix and average precision (AP).

RESULTS
The proposed model demonstrated a promising classification performance with 92.7% accuracy, surpassing the base RetinaNet model. It achieved a mean average precision (mAP) of 81.8%, indicating superior object detection ability. Notably, performance metrics for each severity category showed significant improvement.

CLINICAL RELEVANCE
The proposed deep learning-based method can accurately localize the position of sesamoid bones and grade the severity of sesamoiditis on equine radiographs, providing corresponding confidence scores. This approach has the potential to be deployed in a clinical environment, improving the diagnostic interpretation of metacarpophalangeal (fetlock) joint radiographs in horses. Furthermore, by expanding the training dataset, the model may learn to assist in the diagnosis of pathologies in other skeletal regions of the horse.

Keywords: deep learning, equine, sesamoiditis, artificial intelligence, radiograph

Proximal sesamoiditis is a common radiographic abnormality observed in Thoroughbred yearlings, associated with injury to the suspensory ligament branches at their insertion on the proximal sesamoid bones.[1–3] Clinically affected horses may exhibit pain or lameness, particularly at high speed.[4] Several studies have demonstrated an association between significant sesamoiditis and poorer future racing performance.[5,6] A robust association also exists between sesamoiditis and subclinical suspensory ligament branch change (SSLBC), suggesting an elevated risk of future suspensory ligament branch injury (SLBI).[2,7]

Diagnosing sesamoiditis is not simply a matter of identifying its presence but entails grading its severity. Yearlings with more severe sesamoiditis face a higher risk of future injuries and a longer-term of performance decline.[3,6,7] For more precise

assessment, we adopted the modified Spike-Pierce scale, categorizing sesamoiditis based on the radiographic features of abnormal vascular channels on the sesamoid bone[6,7] **(Figure 1)**:

1 Normal sesamoid bone: Characterized by parallel vascular channels less than 2 mm in width.
2. Mild sesamoiditis: Defined by 1 or 2 irregularly shaped linear defects exceeding 2 mm in width. While such horses may display diminished performance up to the age of 2, this decline noticeably attenuates by their third year.[6]
3. Moderate sesamoiditis: Evident from either a minimum of 3 linear defects that are over 2 mm wide or at least one defect that is broader than 4 mm. Horses in this category consistently underperform at ages 2 and 3.[3,6]

However, the diagnostic challenge is that some yearlings with sesamoiditis do not show obvious clinical signs of SLBI such as swelling or lameness, and even palpation may also give false negatives.[6,8] In addition, false positives may also occur when using ultrasound to detect suspensory ligament branch lesions.[9] Accurate grading of sesamoiditis through radiographs is therefore imperative, especially during prepurchase examinations. Misdiagnosis can lead to a misjudgment of the horse's value and negatively impact the horse's health. Overestimating mild sesamoiditis as moderate can lead to superfluous treatments and breaks in conditioning, depriving the horse of pivotal training phases and optimal periods for bone and muscle maturation, subsequently elevating injury risks.[10] In contrast, underestimating moderate sesamoiditis can result in preventable progression of SLBI. Especially during busy sales seasons, when large numbers of radiographs require careful review by veterinarians, busy schedules and differences in experience can cause critical radiographic details to be overlooked or misinterpreted. In response, our research introduces deep learning (DL) for the automated detection and classification of sesamoiditis in horses via radiographs. This approach reduces the workload of veterinarians and bridges differences in experience, allowing for more precise diagnosis.

DL, a subset of machine learning, utilizes artificial neural networks to replicate tasks akin to human cognition. These networks are structured similarly to

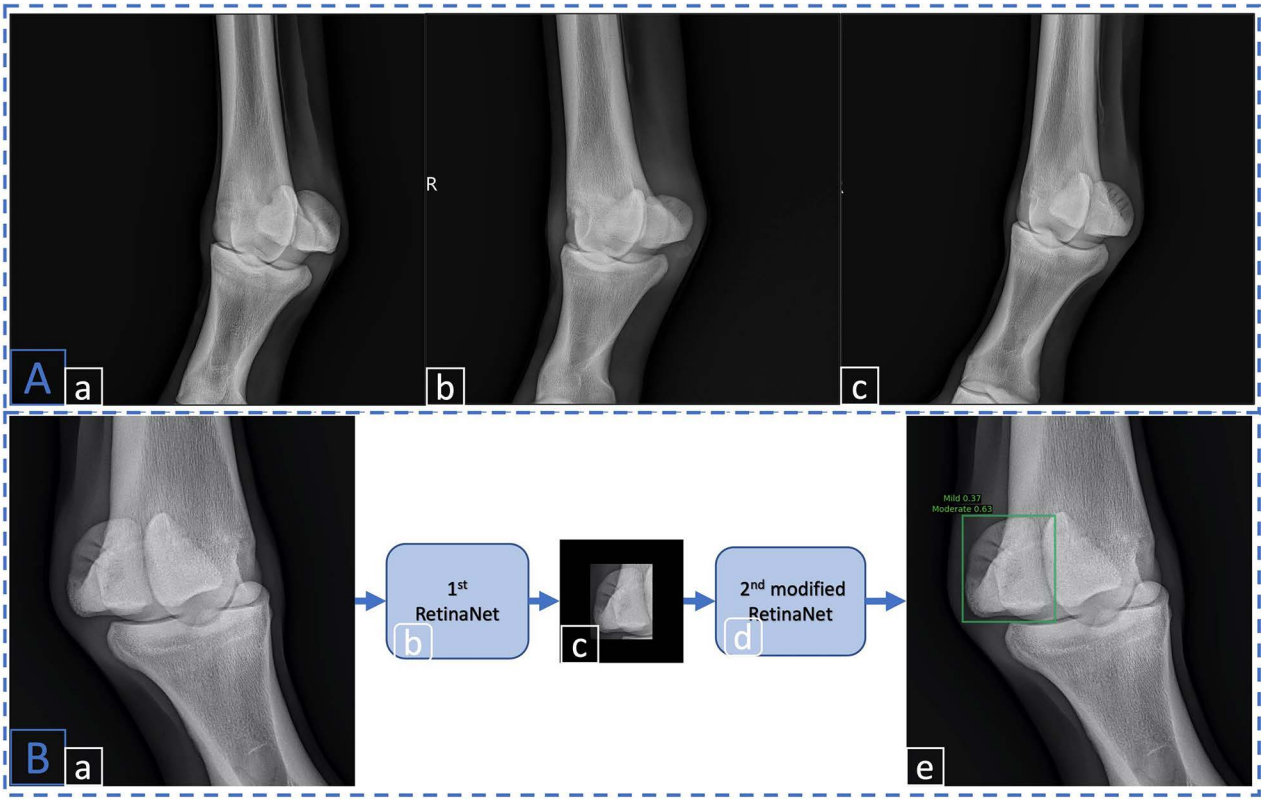

Figure 1—(A) The radiographs of Thoroughbreds aged 12–18 months obtained between 2021 and 2022 are categorized into 3 classes: (a) Normal sesamoid bone: Vascular channels are consistently parallel and exhibit a width under 2 mm. (b) Mild sesamoiditis: Distinguished by either 1 or 2 irregular linear defects with widths exceeding 2 mm. (c) Moderate sesamoiditis: Defined by a minimum of 3 linear defects each exceeding 2 mm in width, or a singular defect surpassing 4 mm in width. (B) Proposed serial architecture sequentially outlines the stages: (a) Input: Initial radiographs. (b) First stage: Implementation of the first RetinaNet model. (c) Image refinement: A cropped and padded image, tailored based on the outputs of the first RetinaNet model. (d) Second stage: Utilization of the second RetinaNet model with self-attention. (e) Final output: The final output of the second RetinaNet model contains the location of the sesamoid bone shown by a green rectangular box, the possible classes of sesamoiditis and the corresponding confidence scores.

the human brain, composed of interconnected nodes that resemble neurons. Through iterative adjustments of the model's parameters, DL can process and learn from training data to perform specific tasks. It has seen extensive application in human medical imaging, notably in image segmentation, classification, and detection tasks. Two-stage detectors like Faster R-CNN[11] and its variants have shown exceptional performance in human skeletal radiograph detection, particularly when fine-tuned with specific datasets.[12,13] Improved detection performance has been achieved with enhancements such as dilated convolutions,[14] receptive field expansions,[15] feature pyramid reconstructions,[16] and guided anchoring.[17] Moreover, the attention mechanism has significantly bolstered the accuracy of DL applications, as demonstrated in COVID-19 chest radiographic detection.[18,19]

However, the application of DL to equine medicine is much less studied, facing constraints due to the dearth of horse medical images. Despite a few studies exploring DL techniques in horse medical imaging, such as Basran et al's fracture risk assessment[20] and Silva et al's radiographic view classification,[21] to our knowledge, no work has specifically addressed sesamoiditis detection. This research gap prompts our study's objective: to introduce a novel DL framework for sesamoiditis diagnosis in equines.

The proposed DL framework combines 2 RetinaNets,[22] with an integrated self-attention mechanism,[23] designed to effectively localize and grade sesamoiditis severity in radiographs. RetinaNet[22] is a popular DL model designed for object detection tasks. The output of our proposed framework is the sesamoid's bounding box, a rectangular box that highlights the area in the image where the sesamoid bone is located, and the possible sesamoid categories with corresponding confidences (Figure 1). This innovative approach aims not only to assist in veterinary diagnosis and reduce workload but also to minimize misdiagnosis risk; thereby, addressing the current lack of DL applications in equine medical imaging.

Methods

Serial architecture

The proposed serial architecture featured 2 sequentially linked RetinaNets (Figure 1). The first RetinaNet served to localize the sesamoid area, whereas the second modified RetinaNet determined the bounding box and classification of the object (sesamoid). The process comprised the following steps:

1. The radiograph, large in pixel dimensions with side lengths often surpassing 2,000 pixels, was input into the serial architecture.
2. The first RetinaNet model, following a downsizing of the image to a maximum input side length of 1,333 pixels, predicted the object bounding box. Despite the potential loss of information due to the resizing process, the sesamoid's distinctive form facilitated its accurate localization.
3. Leveraging the bounding box predicted by the first RetinaNet model, the sesamoid region was extracted from the original high-resolution radiograph where the maximum side length of this sesamoid region did not surpass 400 pixels.
4. The extracted image underwent zero-padding to reach dimensions of (640,640), adding a border of zero-valued pixels around it. This specific padding size, as opposed to RetinaNet's default minimum input size of (800,800), was observed to yield greater accuracy levels.
5. The zero-padded image was then processed by the second, modified self-attention RetinaNet to acquire object-bound boxes and classes.
6. Lastly, the localized sesamoid region was overlaid on the original large radiograph along with the predicted sesamoid class providing a prediagnostic result that could help veterinarians in the decision-making process.

The first RetinaNet model adhered to the original structure as proposed by Lin et al.[12] However, in the second model, we used a modified version of RetinaNet that integrated a self-attention module post the third convolutional layer of the classification subnetwork. This serial architecture guided the second RetinaNet model to concentrate strictly on the sesamoid region; thereby, eliminating potential disturbances from the background and other bone structures.

Modified RetinaNet detector

The RetinaNet[22] architecture consists of 4 components: ResNet-based backbone;[24] feature pyramid network[25] (FPN); and 2 parallel subnetworks, 1 for classification and the other for bounding box regression. This section focuses on the rationale for selecting RetinaNet[22] and our modifications to the classification subnetwork.

In radiographs, we observed an issue due to the large number of easily classifiable dark background pixels compared to the foreground pixels within the sesamoid region. This imbalance posed a challenge in training deep neural networks, as most negative samples (background pixels) provided less valuable learning signals, leading to inefficient training. Furthermore, an excess of negative samples could overpower the gradient, adversely affecting model performance.[22] RetinaNet's focal loss addresses this by minimizing the contribution of easy examples to the loss using a modulating factor. Focal loss is defined as:

$$FL(p_t) = -\alpha(1-p_t)^{\gamma} \log(p_t) \quad (1)$$

where p_t is the predicted probability of the positive sample, $\alpha_t \in [0,1]$ is the weight factor, $\gamma \geq 0$ is the tuneable focusing parameter, and $(1 - p_t)^{\gamma}$ is the modulating factor.

We retained the majority of the RetinaNet structure but added a self-attention module before the Rectified Linear Unit (ReLU),[26] an activation function, after the third convolutional layer in the classification subnetwork. This modification aimed to capture long-range dependencies between spatial locations, which is crucial for object detection tasks that require understanding relationships between different object parts.

Self-attention

The self-attention module, inspired by the Transformer,[23] was integrated into RetinaNet's classification subnetwork. This module enhanced the network's ability to focus on task-relevant features (eg, vascular channels on sesamoids) while suppressing less important ones. This was particularly beneficial when grading the severity of sesamoiditis on radiographs, which often involved distinguishing subtle differences between vascular structures.

The self-attention mechanism models the intricate relationships among all pairs of pixels in an input feature map, facilitating the effective acquisition of contextual information. This module generates 3 representations for the input feature map, termed as "query," "key," and "value" **(Figure 2)**. The interaction between the query and the key is leveraged to compute a compatibility score, frequently referred to as "attention." This score is indicative of the degree of relevance between pixel pairs, enabling an efficient mapping of the interdependencies across varying segments of the input. The calculated scores are subsequently utilized to regulate the contribution of each pixel toward the composition of the output feature map. Through this methodology, the self-attention module bestows upon the network the capacity to selectively emphasize different regions of the input, premised on their pertinence to the task at hand. Such selective focus provided by the self-attention mechanism equips the network with the proficiency to prioritize features efficiently within the input. This enhances the network's competency to apprehend and learn complex, long-range dependencies within the data.

Contrary to a convolutional neural network (CNN) that regards the input feature map as an assortment of image patches, the self-attention mechanism perceives the input feature map as an integrated whole. This distinction stems from the intrinsic spatial limitations of CNNs owing to the use of convolution kernels, which restrict the focus to the relationships between a pixel and its immediate neighbors in the feature map. Conversely, the self-attention mechanism adopts an egalitarian approach by computing the relevance of all pixel pairs on the feature map equally, irrespective of their spatial distance. This overall and equal computation facilitates the model's capacity to discern long-range dependencies within the image data.

In this study, we opted for a simplified single-head self-attention as opposed to the conventional multihead attention typically used in Transformer architectures. Our choice was motivated by the potential overfitting concerns posed by the multihead attention mechanism, especially given our limited sample size. The multihead attention, with its multiple parallel attention layers, might result in the model learning noise or unimportant details when working with scarce training data. This learning behavior can have an adverse effect on the model's capacity for generalization. To bolster our model's robustness, we strategically chose the single-head

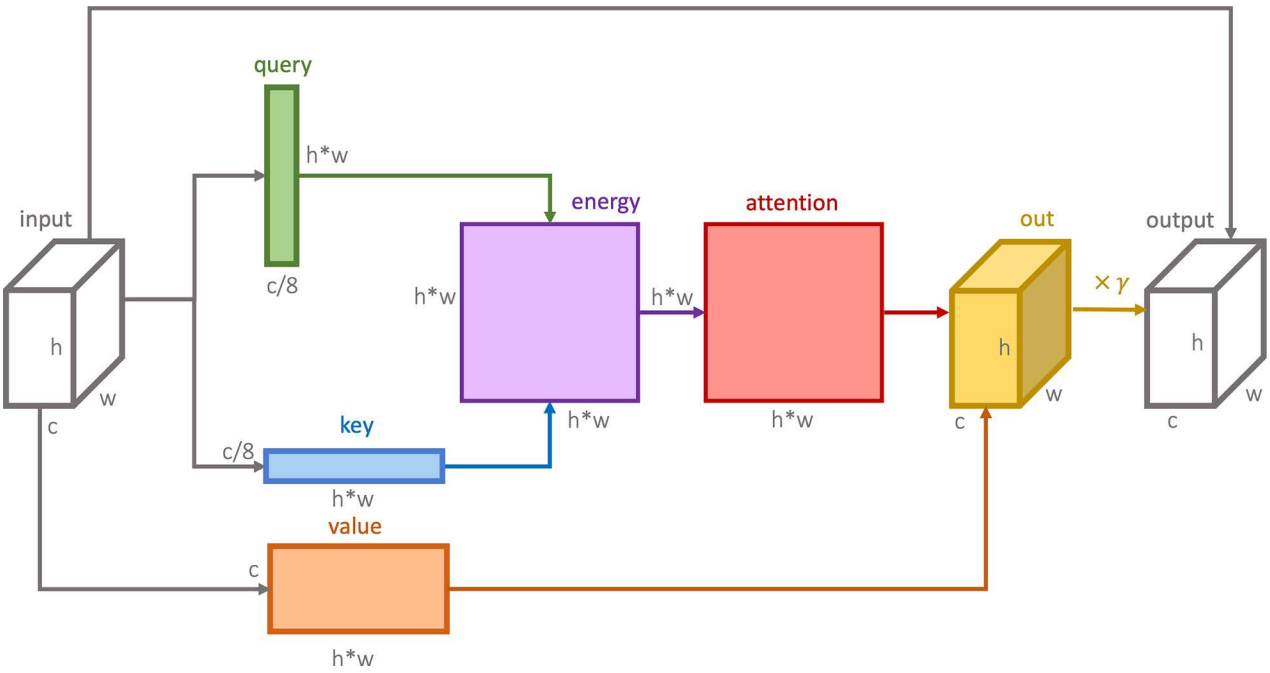

Figure 2—The self-attention module generates 3 representations of input features: "query," "key," and "value." The query and key are used to calculate attention scores, which measure the relevance between different elements of the input features. The value, along with the generated "attention," is added to obtain the intermediate output termed "out." Subsequently, the out is scaled by a scalar γ and combined with the original input features to produce the final output of the self-attention module. In our model, the number of channels in the input, denoted as "c," is set to 256, while "h" and "w" represent the height and width of the input, respectively.

self-attention structure. The specific process of this module includes the following steps (Figure 2):

1. Initially, an input tensor x was provided, which had a shape of (batch size, c, h, w). Here, c represents the number of channels in the input tensor, h and w represents the height and width of the input tensor. To process the input tensor, 1 X 1 convolutional layers were applied to transform the input tensor into query, key, and value representations. The output channels for query and key were reduced to 1/8 to decrease computational cost and risk of overfitting. The resulting query tensor had a shape of (batch size, h * w, c/8), while the key and value tensors had shape of (batch size, c/8, h * w) and (batch size, c, h * w), respectively. In our model, the input feature map had 256 channels (ie, c = 256 and c/8 = 32).
2. Following this, an energy tensor was calculated through batch matrix multiplication between query and key, resulting in a tensor of shape (batch size, h * w, h * w), representing compatibility scores between pixel pairs in the input feature map.
3. Subsequently, an attention tensor was derived through the application of a softmax function on the last dimension of the energy tensor. This process normalized the compatibility scores, leading to the formation of a probability distribution spread across the pixels.
4. An output tensor was computed via batch matrix multiplication between the value and transposed attention tensor, resulting in a tensor of shape (batch size, c, h * w). This output tensor was reshaped to the original input shape (batch size, c, h, w) and was scaled by a learnable scalar, γ.
5. Finally, the scaled output tensor was added to the input feature map x, yielding the final output of the self-attention module.

Dataset

The radiographs utilized in this research study were sourced from Hagyard Equine Medical Institute. In adherence to confidentiality norms, all images were subjected to a thorough process of anonymization to ensure the privacy protection of the equine subjects.

The dataset consisted of 255 samples (images) in total, including 85 normal samples, 91 samples with mild sesamoiditis, and 79 samples with moderate sesamoiditis. This dataset was systematically segregated into training, validation, and test subsets according to a 70:15:15 split ratio. After hyperparameter selection, the validation subset was merged with the training subset, consequently forming a more extensive training set. The final configuration of the training set consists of 214 images, categorized as follows: 73 images depicting normal sesamoids, 73 indicating mild sesamoiditis, and 68 representing moderate sesamoiditis. Concurrently, the test set comprised a total of 41 images, distributed among 14 normal sesamoids, 14 images of mild sesamoiditis, and 13 of moderate sesamoiditis. This strategic allocation of the dataset aided in facilitating a robust and comprehensive evaluation of the models being developed.

Annotation—Annotations were carefully performed by the researchers using RectLabel software (RectLabel version 2023.09.12; developed by Ryo Kawamura), delineating sesamoids with bounding boxes and categorizing sesamoiditis severity. A veterinarian with over 20 years of experience in Thoroughbred prepurchase examinations reviewed all annotations to ensure accuracy.

Augmentation—Due to the limited training set, data augmentation was applied during the training phase to reduce overfitting and underfitting risks. The optimal performance was achieved using a combination of random rotation (between −10 and 10 degrees) and random brightness adjustment (ranging from 0.8 to 1.2).

Preprocessing—The dataset contained images with varying dimensions, where the minimum side length exceeded 1,500 pixels, and the maximum side length was less than 3,000 pixels. To reduce the size of the image without sacrificing the resolution, a cropping approach was used, removing 11% from the left and right sides and 23% from the top and bottom sides, ensuring the preservation of the sesamoid region.

Implementation details

In this study, we implemented both the original and modified versions of the RetinaNet model within the Detectron2 platform, using ResNet-50[24] as the backbone and Feature Pyramid Network (FPN)[25] for extracting features at different levels. The training was conducted using the Nvidia A100 40GB GPU. To achieve optimal performance, the networks were first pretrained on the widely used ImageNet dataset and were subsequently fine-tuned on our specific training dataset. During fine-tuning, a batch size of 2 images was used, and the learning rate was adjusted to 0.0002.

Considering the differing characteristics of the input images for the 2 models in the serial architecture, different epoch settings were implemented for each RetinaNet. The first RetinaNet, responsible for processing large radiographs with extensive information, was trained for 50 epochs. This prolonged training duration enabled the model to effectively capture the richness and complexity of the input data. The second RetinaNet, specifically designed to process smaller images that solely focused on a sesamoid, underwent training for 30 epochs, as the reduced number of epochs was sufficient for this specific task. Regarding the object categories, the model was adapted to handle 3 classes: Normal sesamoid, Mild sesamoiditis, and Moderate sesamoiditis. To ensure accurate predictions, the model was configured to make only 1 object prediction per image, enabling it to focus on identifying the most prominent abnormality within each radiograph.

The training process could be divided into 2 distinct phases, where each of the 2 RetinaNets within the architecture was trained separately. The first RetinaNet was trained on a preprocessed dataset comprising large radiographs with varying dimensions. After the completion of this initial training phase, the training set was reintroduced to the first

RetinaNet model to infer object-bounding boxes. Despite exposure to these images during training, the predicted bounding boxes exhibited certain discrepancies from the ground truth. These discrepancies could be attributed to the excessively large image size of the training set and the limited number of training samples available. However, these deviations were acceptable as the first RetinaNet model was to localize the approximate position of the sesamoid bone, with the precise bounding box and class determined by the second model.

Subsequently, the second RetinaNet model was trained, incorporating a self-attention mechanism. The training set images were cropped based on the bounding box predictions generated by the first RetinaNet model. These cropped images were then zero-padded to conform to a standardized size of 640 by 640 pixels. This padding process ensured that the training set images for the second RetinaNet maintained a consistent size. After the above processing, the image input to the second RetinaNet model only contained the sesamoid bone area, effectively minimizing distractions from other skeletal structures.

Statistical Analyses

This study investigated the confusion matrix and average precision (AP) as evaluation criteria. The confusion matrix is an important metric for assessing the sensitivity of classification models. It features a horizontal axis representing predicted labels and a vertical axis for true labels, with diagonal elements indicating the per class-sensitivity values, and the rate of correctly predicted positive samples from among the positive class samples. We also used the Accuracy Metric to represent the overall classification accuracy **(Table 1)**. The sensitivity and accuracy metrics were defined as:

$$Sensitivity = \frac{TP}{TP + FN}$$

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FP}$$

In our multiclass classification task, we had 3 categories: "normal," "mild," and "moderate." We considered each category individually as the "positive" class, while the other 2 categories were collectively considered as the "negative" class. Taking the normal category for example, True Positives (TP) were instances correctly identified as normal. True Negatives (TN) were those correctly identified as not normal (either mild or moderate). False Positives (FP) were instances misclassified as normal when they were truly mild or moderate. Additionally, False Negatives (FN) were normal instances misclassified as mild or moderate. Similar definitions applied for the mild and moderate categories.

For the evaluation of the proposed model in terms of object detection, we applied the COCO-style AP metric.[27] AP was calculated by using 10 Intersection over Union (IoU) thresholds, extending from 0.5 to 0.95 in steps of 0.05 (denoted as IoU = 0.5: 0.05: 0.95). The IoU is a metric that quantifies the overlap between the predicted and ground-truth bounding boxes. Each IoU threshold corresponds to a specific set of precision and recall (also known as sensitivity) values, which are used to construct the Precision-Recall (PR) curve. The area under this PR curve is AP, representing the model's performance in object detection. In COCO-style, AP usually represents mean AP, but to distinguish it from AP of each class, we used mAP to represent mean AP across multiple classes in the results. The precision metric was defined as:

$$Precision = \frac{TP}{TP + FP}$$

Notably, our model is designed to produce all potential categories for an object bounding box. However, in accordance with the accepted standards for object detection evaluation, a bounding box is presumed to correspond to a single category only. As such, to compute the AP, we selected the category associated with the highest confidence score within an object bounding box, treating it as the definitive category for that box.

Table 1—Summary results for all experimental models utilized in this study.

Model name	Normal AP (%)	Mild AP (%)	Moderate AP (%)	mAP (%)	Accuracy (%)	Params (M)
Faster R-CNN with FPN[a]	41.0	37.9	42.8	40.6	65.4	41.3
RetinaNet[b]	41.2	53.1	40.5	44.9	75.6	36.3
Serial RetinaNets	81.1	52.8	78.5	70.8	82.9	72.6
Serial RetinaNets with attention	83.1	85.1	77.3	81.8	92.7	72.7

The reported metrics include average precision (AP), which quantifies the model's accuracy by integrating the precision-recall curve over all thresholds, reflecting the model's consistent performance across various confidence levels. Specifically, Normal AP represents average precision for detecting normal sesamoid bone, Mild AP represents average precision for detecting mild sesamoiditis, and Moderate AP represents average precision for moderate sesamoiditis. The mAP provides an overall measure across all categories. "Accuracy" represents classification accuracy, and the number of model parameters denoted as "Params (M)" in millions, which reflects the model's complexity and capacity.

[a]"Serial RetinaNets" model consists of 2 base RetinaNets, so the number of parameters is doubled. [b]"Serial RetinaNets with Attention" model comprises a base RetinaNet and a modified RetinaNet with self-attention. The number of parameters is the sum of the parameters from both the base RetinaNet and the modified version.

Results

Results analysis

We presented the empirical results of our proposed architecture, serial RetinaNets with attention, and compared it to baseline models. The confusion matrices provided detailed insight into the sensitivity values for each sesamoiditis class **(Figure 3)**. In addition, we also showed the AP per class, the mean average precision (mAP), the overall classification accuracy, and the model parameters for all tested models (Table 1).

Our proposed architecture, serial RetinaNets with attention, demonstrated outstanding classification capabilities across all sesamoiditis classes (Figure 3). The model did not exhibit any specific class-related deficiencies, indicating comprehensive performance. The overall classification accuracy, a critical measure for veterinary professionals, was 92.7% (Table 1). This result underscored the potential of our proposed architecture as a valuable tool for veterinary diagnosis. Furthermore, the AP values for each class underlined our architecture's superiority in detecting all sesamoiditis categories compared to the basic RetinaNet, boasting a mAP of 81.8% (Table 1).

Notably, the Faster R-CNN model,[1] renowned for its prowess in human radiograph object detection tasks,[11–17,28] performed suboptimally in equine tasks. We attributed this underperformance to the limited availability of training data. The complexity of Faster R-CNN surpassed that of RetinaNet, potentially causing overfitting when training data was scarce. Human focused studies often benefited from access to thousands of images, a luxury seldom afforded in veterinary medicine.

Serial architecture

The performance of the serial RetinaNets architecture suggested an improvement in the model's performance compared to the basic RetinaNet, with the mAP of serial RetinaNets increasing by 25.9% (Table 1). Notably, the AP values of serial RetinaNets for the normal and moderate classes were significantly higher than those of the basic RetinaNet. Nevertheless, the model exhibited difficulties in accurately identifying mild sesamoiditis and often misclassified mild sesamoiditis as moderate cases (Figure 3).

Self-attention

The integration of self-attention after various convolutional layers in the classification subnetwork yielded different results. We evaluated each of the 4 consecutive convolutional layers in the RetinaNet classification subnetwork individually **(Table 2)**. The best performance was achieved by incorporating the self-attention module after the third convolutional layer, resulting in a mAP value of 81.8%. The integration of the self-attention mechanism significantly improved the AP for Mild Cases, surpassing both the basic RetinaNet and serial RetinaNets.

Contrarily, the Convolutional Block Attention Module (CBAM),[29] a widely accepted attention mechanism that focuses on channel and spatial features, did not yield comparable results. Its limitations can be attributed to the receptive field of the convolutional kernel, which primarily emphasizes local spatial dependencies. In contrast, the self-attention mechanism formulates a comprehensive dependency map that considers the relationships between each pixel and all other pixels in the feature map. This process establishes a global context through pixel-to-pixel mapping; thereby, facilitating the capture of long-range dependencies or contextual information more efficiently than CBAM's localized emphasis. This empirical comparison further validates that an attention mechanism capable of capturing long-range dependencies or contextual information is better suited for this task.

To better understand the self-attention mechanism within the model, we utilized Grad-CAM (Gradient Weighted Class Activation Mapping).[30] Grad-CAM provided activation maps that highlight key regions influencing classification outcomes. Models without self-attention tended to overemphasize the shadows produced by bone overlaps, which often resulted in classification errors. In contrast, self-attention diminished the model's focus on these distracting shadows and expanded the model's region of interest to the entire sesamoid bone, especially where vascular channels were located **(Figure 4)**.

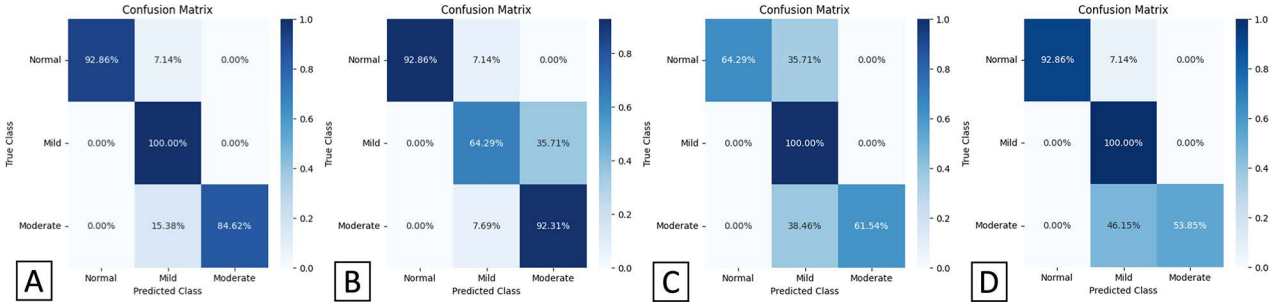

Figure 3—Confusion matrices: (A) Serial RetinaNets model with self-attention: Displays classification outcomes when the serial RetinaNets incorporate self-attention mechanisms. (B) Serial RetinaNets model without self-attention: Illustrates results from the serial RetinaNets devoid of self-attention module. (C) Basic RetinaNet model: Provides classification outcomes from the foundational RetinaNet model. (D) Junior veterinarian classification: Demonstrates classification results executed by a junior veterinarian with 1 year of professional experience.

Table 2—Summary results of the self-attention module and the CBAM module are combined at different convolutional layers of the classification subnetwork.

Model name	Normal AP (%)	Mild AP (%)	Moderate AP (%)	mAP (%)
Serial RetinaNets[a]	81.1	52.8	78.5	70.8
SA after conv1[b]	92.5	59.7	75.2	75.8
SA after conv2	88.8	61.8	75.8	75.5
SA after conv3	83.1	85.1	77.3	81.8
SA after conv4	92.4	59.7	79.7	77.3
CBAM[29] after conv[c]	75.5	55.3	75.3	68.7

Normal AP, Mild AP, Moderate AP, and mAP are as described in Table 1.

[a]"Serial RetinaNets" refers to 2 RetinaNets connected in series without any attention module. [b]"SA after conv1" indicates the incorporation of self-attention (SA) after the first convolutional layer in the classification subnetwork of the second RetinaNet of the serial architecture. [c]"CBAM after conv3" means the incorporation of Channel and Spatial Attention (CBAM)[29] after the third convolutional layer.

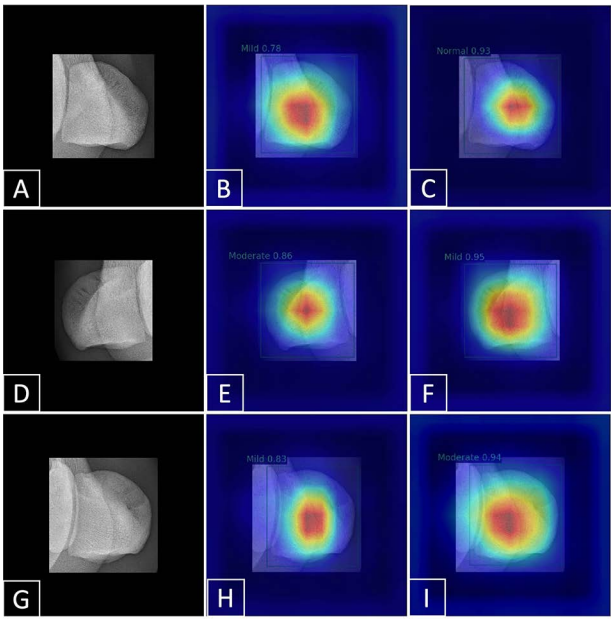

Figure 4—The Grad-CAM[30] visualization depicts the activation map of the third ReLU activation layer within the classification subnetwork. This visualization technique provides insights into the areas of focus for the model during the classification of sesamoiditis cases. The color on the image signifies the model's attention level to specific areas. Areas with more obvious red tones indicate that the model's attention is more focused, indicating that this area has a major impact on the model's classification results. The figure is organized into 3 rows, each representing a different category. In the first row, (A) presents an example of a normal sesamoid, (B) displays the activation map of RetinaNet for the normal case, and (C) demonstrates the activation map of RetinaNet with self-attention for the normal case. Transitioning to the second row, (D) illustrates an instance of mild sesamoiditis, (E) displays the activation map of RetinaNet for the mild case, and (F) exhibits the activation map of the model with self-attention for the mild case. Finally, the third row encompasses (G) an example of moderate sesamoiditis, (H) the activation map of basic for the moderate case, and (I) the activation map of the model with self-attention for the moderate case.

Discussion

Within the serial RetinaNets architecture, the second model exclusively processes a reduced image consisting solely of the sesamoid region. As a result, the second model's focus is confined to the sesamoid region, unimpeded by irrelevant information from other areas of the complete, large radiographs. This process can be aptly likened to a microscope's focusing mechanism. The first RetinaNet processes the entire radiographs, much like the broad perspective of a low-power microscope. In comparison, the serial RetinaNets architecture allows the second model to concentrate on the sesamoid region, analogous to a high-power microscope that magnifies a specific region of interest.

This serial architecture significantly improved the AP values for the normal and moderate classes, surpassing those of the base RetinaNet (Table 1). Given that these 2 types of sesamoids exhibit distinct differences in radiographs, these results suggest that the serial architecture enhances the model's ability to discern distinctive features. Moreover, the model demonstrated a high sensitivity of 92.86% for normal sesamoids (Figure 3). This implies that serial RetinaNets may serve as an effective tool for the initial screening of normal sesamoids.

Despite these strengths, serial RetinaNets struggled to identify mild sesamoiditis accurately, frequently misclassifying mild cases as moderate ones (Figure 3). This could lead to the horse receiving further diagnostics such as ultrasound and/or MRI, increasing the cost to the owner and potentially devaluing the horse at trade. In addition, due to the false positive phenomenon of SLBI in ultrasound diagnosis, the misclassification tendency may lead horses to receive unnecessary treatment, rest, and lose training time.[9] In contrast, the junior veterinarian tended to misclassify moderate sesamoiditis as mild cases (Figure 3). Such misclassifications were particularly concerning as they might result in clinically affected horses not receiving timely rest, treatment, and potentially overvaluing the horse at trade. Considering the similar vascular appearances between mild and moderate sesamoiditis, serial RetinaNets and the junior veterinarian might share similar limitations in discerning subtle variations in radiographs.

The initial convolutional layers of a model typically capture low-level features, while later layers learn more abstract, high-level features. Our experiments showed that the self-attention module yielded the most benefit when applied to the third convolutional layer. At this layer, the model had learned enough to abstract meaningful features that benefited for the self-attention mechanism, but these features were not so abstract that the long-range dependencies modeled by self-attention became irrelevant.

Mild cases usually show minimal deviation from healthy sesamoids and moderate cases; thus, the improvements observed in the classification of mild cases indicate that the self-attention mechanism enhances the model's ability to discern subtle variations in the representation of vascular channels on radiographs.

Grad-CAM maps helped to understand the role of self-attention within the model, revealing its ability to reduce the influence of irrelevant features and enhance the capture of global information. Shadows that resemble enlarged vascular channels can mislead the model, but introducing self-attention decreases the model's focus on these misleading features, improving classification accuracy. For example, when the edge of the cannon bone is situated behind the sesamoid bone, it creates a shadow that resembles an enlarged vascular channel (Figure 4). Since the diagnosis of sesamoid inflammation is primarily based on the appearance of vascular channels on the sesamoid bone, such misleading shadows can confound the model's judgment. Self-attention helps to minimize the model's focus on these irrelevant features, that is, shadows (Figure 4).

Moreover, the Grad-CAM maps demonstrated the self-attention's ability to broaden the model's area of interest to include prominent vascular channels in cases of mild to moderate sesamoiditis, so that the model considers a broader range of information when grading sesamoiditis. For instance, the basic RetinaNet focuses only on a part of the sesamoid, whereas RetinaNet with self-attention concentrates on the entire sesamoid, specifically covering the vascular channels used for the classification of sesamoiditis (Figure 4). Thus, the self-attention mechanism can reduce the influence of misleading features in images and enhance the model's ability to capture global information; thus, improving classification performance.

Our study has 3 possible limitations. First, the current serial architecture in the DL model does not support end-to-end training, necessitating separate training for the first and second RetinaNet. Second, the introduction of attention allows the model to focus on the entire sesamoid region, while if the model could only focus on the vascular channel region, the classification performance might be further improved. Finally, the classification of all images by a single veterinarian (though highly experienced) could introduce personal bias, potentially compromising the objectivity of the ground truth of the data.

In conclusion, this study presents a novel deep-learning approach with serial architecture and self-attention for localization and grading of equine sesamoiditis using radiographic imaging. The proposed model addresses the challenge of large image size by utilizing 2 RetinaNets in series. Furthermore, the inclusion of a self-attention module in the classification subnetwork of the second RetinaNet enhanced the model's ability to classify the severity of sesamoiditis, achieving an elegant mAP of 81.8% on the test set. Our results suggest potential clinical applications of the proposed model. Our software and code will be publicly available later.

Acknowledgments

We express our deepest gratitude to Hagyard Equine Medical Institute for kindly providing essential radiographic data. Our sincere appreciation also goes to Mitacs, whose financial support has been instrumental in facilitating this study.

This research project was initiated by a private commercial company called Point to Point Research & Development, and the horse radiographs used in our study were obtained exclusively from the Hagyard Equine Medical Institute. Therefore, due to privacy and commercial protection concerns, we cannot disclose the specific code and datasets used in the analysis at this time. However, if you express an interest in accessing our code and results, we kindly ask you to contact Andrew Rideout via email at Rideout@family1x.com and he will assist in providing the necessary resources. The code will be publicly available in the form of software and an online website in the future. People can directly use our model to infer custom images or datasets through the website.

Disclosures

The authors have nothing to disclose. No AI-assisted technologies were used in the generation of this manuscript.

Funding

This study was funded by Mathematics of Information Technology and Complex Systems (Mitacs). The authors received funds from the MITACS ACCELERATE Program. There is no other financial disclosure.

References

1. Kane AJ, Park RD, McIlwraith CW, et al. Radiographic changes in Thoroughbred yearlings. Part 1: prevalence at the time of the yearling sales. *Equine Vet J.* 2010;35(4):354–365. doi:10.2746/042516403776014280
2. Plevin S, McLellan J, O'Keeffe T. Association between sesamoiditis, subclinical ultrasonographic suspensory ligament branch change and subsequent clinical injury in yearling Thoroughbreds. *Equine Vet J.* 2015;48(5):543–547. doi:10.1111/evj.12497
3. Spike DL, Bramlage LR, Howard BA, Embertson RM, Hance SR. *Radiographic Proximal Sesamoiditis in Thoroughbred Sales Yearlings*. Published online January 1, 1997.
4. Garrett KS, Bramlage LR, Spike-Pierce DL, Cohen ND. Injection of platelet- and leukocyte-rich plasma at the junction of the proximal sesamoid bone and the suspensory ligament branch for treatment of yearling Thoroughbreds with proximal sesamoid bone inflammation and associated suspensory ligament branch desmitis. *J Am Vet Med Assoc.* 2013;243(1):120–125. doi:10.2460/javma.243.1.120
5. Meagher DM, Bromberek JL, Meagher DT, et al. Prevalence of abnormal radiographic findings in 2-year-old Thoroughbreds at in-training sales and associations with racing performance. *J Am Vet Med Assoc.* 2013;242(7):969–976. doi:10.2460/javma.242.7.969
6. Spike-Pierce DL, Bramlage LR. Correlation of racing performance with radiographic changes in the proximal sesamoid bones of 487 Thoroughbred yearlings. *Equine Vet J.* 2010;35(4):350–353. doi:10.2746/042516403776014262
7. McLellan J, Plevin S. Do radiographic signs of sesamoiditis in yearling Thoroughbreds predispose the development of suspensory ligament branch injury? *Equine Vet J.* 2013;46(4):446–450. doi:10.1111/evj.12154
8. Dyson S. Suspensory branch injuries in sports horses and racehorses. *UK-Vet Equine.* 2018;2(3):90–96. doi:10.12968/ukve.2018.2.3.90
9. Ramzan PHL, Palmer L, Dallas RS, Shepherd MC. Subclinical ultrasonographic abnormalities of the suspensory ligament branch of the athletic horse: a survey of 60 Thoroughbred racehorses. *Equine Vet J.* 2012;45(2):159–163. doi:10.1111/j.2042-3306.2012.00588.x

10. Rogers CW, Bolwell CF, Gee EK, Rosanowski SM. Equine musculoskeletal development and performance: impact of the production system and early training. *Anim Prod Sci.* 2020;60(18):2069–2079. doi:10.1071/AN17685
11. Ren S, He K, Girshick R, Sun J. Faster R-CNN: towards real-time object detection with region proposal networks. *IEEE Trans Patt Anal Mach Intell.* 2017;39(6):1137–1149. doi:10.1109/tpami.2016.2577031
12. Yahalomi E, Chernofsky MA, Werman M. *Detection of Distal Radius Fractures Trained by a Small Set of X-Ray Images and Faster R-CNN*. arXiv (Cornell University), 2018. doi:10.48550/arxiv.1812.09025
13. Gan K, Xu D, Lin Y, et al. Artificial intelligence detection of distal radius fractures: a comparison between the convolutional neural network and professional assessments. *Acta Orthopaed.* 2019;90(4):394–400. doi:10.1080/17453674.2019.1600125
14. Guan B, Yao J, Zhang G, Wang X. Thigh fracture detection using deep learning method based on new dilated convolutional feature pyramid network. *Patt Recog Lett.* 2019;125:521–526. doi:10.1016/j.patrec.2019.06.015
15. Guan B, Zhang G, Yao J, et al. Arm fracture detection in X-rays based on improved deep convolutional neural network. *Comput Electr Eng.* 2020;81:106530. doi:10.1016/j.compeleceng.2019.106530
16. Guan B, Yao J, Wang S, et al. Automatic detection and localization of thighbone fractures in X-ray based on improved deep learning method. *Comput Vis Image Underst.* 2022;216:103345. doi:10.1016/j.cviu.2021.103345
17. Xue L, Yan W, Luo P, et al. Detection and localization of hand fractures based on GA_Faster R-CNN. *Alexandria Engin J.* 2021;60(5):4555–4562. doi:10.1016/j.aej.2021.03.005
18. Sitaula C, Hossain MB. *Attention-Based VGG-16 Model for COVID-19 Chest X-Ray Image Classification*. Applied Intelligence. Published online November 17, 2020. doi:10.1007/s10489-020-02055-x
19. Yang H, Wang L, Xu Y, Liu X. CovidViT: a novel neural network with self-attention mechanism to detect Covid-19 through X-ray images. *Int J Mach Learn Cybern.* 2022;14(3):973–987. doi:10.1007/s13042-022-01676-7
20. Basran PS, McDonough SP, Palmer SE, Reesink HL. Radiomics modeling of catastrophic proximal sesamoid bone fractures in Thoroughbred racehorses using μCT. *Animals.* 2022;12(21):3033–3033. doi:10.3390/ani12213033
21. Silva R, Ambika Prasad M, Riggs CM, Doube M. *Equine Radiograph Classification Using Deep Convolutional Neural Networks*. arXiv (Cornell University). Published online April 28, 2022. doi:10.48550/arxiv.2204.13857
22. Lin TY, Goyal P, Girshick R, He K, Dollar P. Focal loss for dense object detection. *IEEE Trans Patt Anal Mach Intell.* Published online 2018. doi:10.1109/tpami.2018.2858826
23. Vaswani A, Shazeer N, Parmar N, et al. *Attention Is All You Need*. arXiv.org, 2023. doi:10.48550/arXiv.1706.03762
24. He K, Zhang X, Ren S, Sun J. *Deep Residual Learning for Image Recognition*. arXiv (Cornell University). Published online December 10, 2015. doi:10.48550/arxiv.1512.03385
25. Lin TY, Dollar P, Girshick R, et al. *Feature Pyramid Networks for Object Detection*. 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). Published online July 2017. doi:10.1109/cvpr.2017.106
26. Krizhevsky A, Sutskever I, Hinton GE. ImageNet classification with deep convolutional neural networks. *Communications of the ACM.* 2012;60(6):84–90. doi:10.1145/3065386
27. Lin TY, Maire M, Belongie S, et al. Microsoft COCO: common objects in context. *Computer Vision – ECCV 2014.* 2014;8693:740–755. doi:10.1007/978-3-319-10602-1_48
28. Ma Y, Luo Y. Bone fracture detection through the two-stage system of Crack-Sensitive Convolutional Neural Network. *Inform Med Unlocked.* 2021;22:100452. doi:10.1016/j.imu.2020.100452
29. Woo S, Park J, Lee JY, Kweon IS. *CBAM: Convolutional Block Attention Module*. Published online July 17, 2018. doi:10.48550/arxiv.1807.06521
30. Selvaraju RR, Cogswell M, Das A, et al. *Grad-CAM: Visual Explanations from Deep Networks via Gradient-Based Localization*. 2017 IEEE International Conference on Computer Vision (ICCV). Published online October 2017. doi:10.1109/iccv.2017.74