A Biomedical Computing Approach to Canine Vertebral Heart Scoring

Ryan Zrymiak[rza80] Christopher Douglas[cda32]
Brendan Bickford[bdb8] Raul Gomero[rsg28]
Rana Hoshyarsadeghi[rha75]

March 21, 2024

Abstract. This project explores the use of biomedical computing techniques for assessing radiographic heart size in the dog (*Canis familiaris*), with an emphasis on the well established vertebral heart scoring technique developed by Buchanan *et al*[1]. Leveraging a data set[5] of 153 canine thoracic radiographs, we train two neural networks and demonstrate the effect of inter annotator variability. The project encompasses data collection, preprocessing of radiographs, model training, and validation, with a focus on trying to developing solutions that are both reliable and accessible to veterinary practitioners. The potential implications of this technology extend beyond individual patient care, offering insights into broader patterns of canine heart health and facilitating advanced research in veterinary medicine.

Keywords: Predictive Analytics, Canine Cardiology, Vertebral Heart Score (VHS), Veterinary Radiology, Automated Diagnostic Imaging

1 Introduction

Heart disease in dogs is a prevalent health issue, posing significant challenges for veterinary medicine. Early detection and accurate diagnosis are crucial for effective management and treatment of these conditions. Traditionally, the assessment of canine heart health relies on the expertise of veterinarians, using methods like the Vertebral Heart Score (VHS), a radiographic technique to measure heart size relative to spinal vertebrae. However, this manual process is time-consuming and subject to variability in interpretation.

With advancements in technology, there is a growing interest in leveraging artificial intelligence (AI) to enhance diagnostic accuracy and efficiency in veterinary medicine. This project introduces two AI-based models: one to automatically identify an elevated VHS, and a second to automatically annotate radiographs. By analyzing a data set of 153 canine X-ray images, the models aim to automate radiograph annotation and the VHS process, reduce human error, and provide rapid, reliable assessments.

This initiative not only promises to improve the speed and accuracy of heart disease detection in dogs but also offers a scalable solution that could be widely accessible to veterinary practices. Moreover, the integration of predictive analytics and automated diagnostic imaging in veterinary medicine has the potential to transform how veterinary care is delivered, facilitating early intervention and improving health outcomes for canine patients.

In this report, we detail the development process of the machine learning models: from data collection and preprocessing, to model training and validation. We also discuss the challenges faced, the solutions implemented, and the implications of this technology in the broader context of veterinary healthcare.

1.1 Cardiac Disease in Dogs

When presented with a dog that has clinical signs suggestive of cardiac disease, one of the first diagnostic steps taken by veterinarians is thoracic radiographs. Thoracic radiographs will provide important information about the lungs, heart and other structures present in the thoracic cavity. When assessing the cardiovascular system, a veterinarian will examine the size and shape of the cardiac silhouette (the hearts outline on a radiograph), changes in the opacity (density) of the lungs, position of the trachea, and the size and shape of the liver.

Many, but not all, cardiac diseases result in cardiomegally (enlargement of the heart), this may be visible as an enlargement of the cardiac silhouette. Identifying such a change can however, be challenging. This is mostly due to the considerable variation in the conformation of the thorax between breeds of dogs. The deep chested dog breeds (Great Danes, Newfoundlands, standard poodles, etc) tend to have the appearance of having an upright heart. The cardiac silhouette in shallow, or barrel chested dogs (Boston terriers, French bulldogs, Labradors etc) lies closer to the sternum and more horizontally and occupies a larger amount of the thoracic cavity. On top of this the phase of respiration or even a dog's body condition score (a measure of how fat or skinny they are) can impact the assessment of heart size.

Multiple methods of heart size measurement have been developed by radiologists[13]. These includ but are not limited to: Proportion of the heart in relation to the width and height of the thoracic cavity; number of rib spaces spanned by the cardiac silhouette; The amount of sternal contact of the cardiac silhouette.

The inadequacies of these methods resulted in a propensity for veterinarians to over interpret cardiac size on radiographs and resulted in low specificity. Buchanan et al[1] developed the VHS method of estimating cardiac size by comparing it to the length of the thoracic vertebrae. The main advantage of this system is that the length of the thoracic vertebrae on radiographs will be relatively unaffected by respiration phase, position of the heart and thoracic cage shape. A demonstration of the advantages of VHS is given in figure 1.

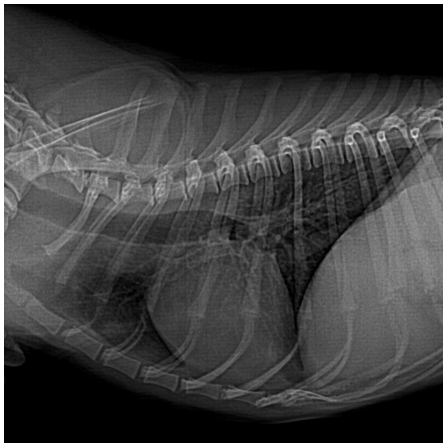
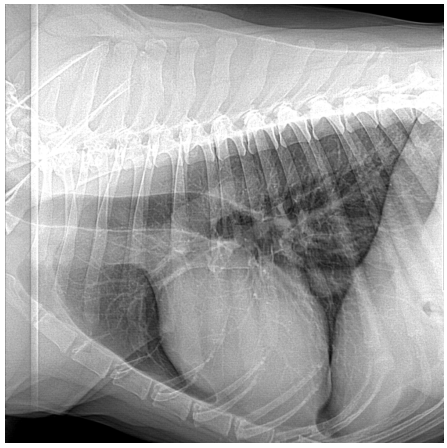



Fig. 1: Relative to the thoracic cavity, the heart in the left image appears subjectively larger since it occupies a larger proportion of the thoracic cavity height and the trachea is displaced dorsally towards the spine. However, the heart in the right image is enlarged with a VHS of 11.13, whereas the heart on the left is within normal limits with a score of just 8.9

It should be noted that a normal VHS does not rule out the presence of cardiac disease as we can have disease without overt enlargement. VHS is best suited to observing cardiac enlargement secondary to volume overload pathologies (e.g. mitral insufficiency, dilated cardiomyopathy, patent ductus ateriosus). Additionally, investigators have found that many breeds consistently exceed the reference intervals published by Buchanan et al[3,2]. Perhaps, the most concerning, is that Lamb, et al found that VHS was very inaccurate in the Boxer, a breed that has a high incidence of cardiac disease due to a genetic predisposition[2].

2 Materials

In this study we use 153 thoracic radiographs of canine patients. This data set[5] was made available as an educational and research aid for vertebral heart scoring.

The radiographs are in PNG image format. All radiographs are latero-lateral images, meaning that the images were taken with the dogs laying on their sides on the x-ray table. All radiographs were anonymized and no patient history or signalment was provided. This means that it is not possible to identify the breed of each dog in the radiographs, or whether the dogs were exhibiting any evidence of cardio-respiratory disease. It is for this reason we will ignore variation in normal VHS scores in different breeds. We will use Buchanan et al's original normal range of between 8.5 and 10.5.

The VHS of each radiograph is not provided. Because of this, we scored the images ourselves. This approach does have limitations. Only one of our group

has the qualification necessary to perform vertebral heart scoring in a clinical setting. Interpretation of the radiographs by certified and boarded radiologists is more ideal, but unfortunately, this was not a resource that was available for our study. We were unable to obtain any 'labelled' data sets that would have fixed this problem. s For this study we utilised, Python[12], Opencv[6], Numpy[7], Pandas[8], Matlab[9], Ultralytics YOLO[10], Tensorflow[11] and LaTeX.

3 Methods

3.1 The Vertebral Heart Score Technique[1,4]

Take a well positioned lateral recumbent thoracic radiograph of a dog.

1. The distance between the ventral aspect of the tracheal bifurcation and the most distant ventral contour of the cardiac apex. This is called the long axis.
2. Take a second measurement, perpendicular to the first, at the widest point of the cardiac silhouette. This is called the short axis.
3. Locate the cranial margin of the 4^{th} thoracic vertebra (T_4)
4. Superimpose the long axis on the spine so that it runs parallel to it and it's starting point is level with the cranial aspect of the T_4
5. Count the number of vertebral bodies along the spine covered by the long axis (to the nearest 0.1)
6. Repeat steps 4 and 5 for the short axis
7. Add the vertebral lengths of the long and short axes to get the VHS

If the sum of the long and short axis are longer then 10.5 vertebrae and if other supporting clinical signs are present, the dog is assessed to have cardiomegally.

3.2 Radiograph Resizing

The initial collection of radiographs had variations in resolution and aspect ratios. Prior to annotation, we used a Matlab script to normalize the images by cropping and resizing each to a size of 1000x1000 pixel. This size was chosen to best fit our screens while preserving details for the annotation process. When handling rectangular images, we first cropped a square from the center of the image before resizing. This cropped section was biased to the left of center in cases where the image was wider than 1300 pixels to preserve visibility or the heart and spine.

3.3 Radiograph Annotation

We implemented annotation software in Python using the Opencv library. The user is first required to place the dorsal and ventral endpoints of the long axis. Next, a line perpendicular to the long axis is displayed over the heart. This acts as a visual aid to help the user determine the widest point of the cardiac

silhouette. Once the widest point is located the short axis is fixed at that point then the user defines it's two endpoints. The user, then tags the cranial margin of T_4 and then aligns the two superimposed copies of the axes with the spine. Once the copies of the axes are fixed in place the user then tags the position of each intervertebral disk (IVD) up to and including the IVD caudal to the last vertebra covered by the longest axis.

The xy coordinates of the axis endpoints, and the IVDs tagged are stored in a csv file. Up to the IVD cranial to T_{13} can be tagged. An example of one of our annotated images in given in figure 2

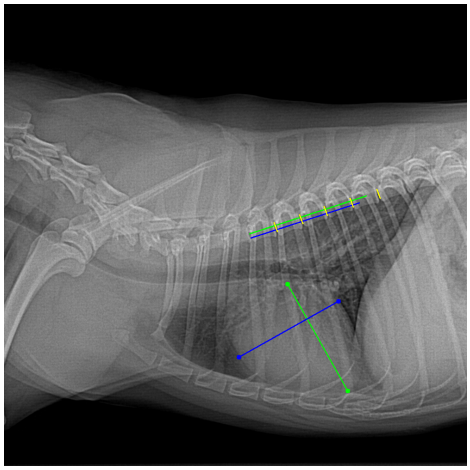


Fig. 2: Annotated and scored radiograph. This dog has a VHS of 8.8.

In our opinion, the software developed provides some distinct advantages over the scoring systems implemented in many modern veterinary DICOM imaging applications. In many of these systems

- The short axis is not constrained to be perpendicular to the long axis
- No visual aid is provided to help with placement of the short axis
- Counting the vertebral length of an axes is done manually by the user
- There commonly is no redo or undo functionality

The first two differences will introduce slight inaccuracies in the placement of the short axis that will have an impact on the VHS calculated. Next, the need for a user to manually count vertebrae results in a loss of precision in VHS calculation, measurements could easily a considerable portion of a vertebra off. Our implementation is precise up to the level of individual pixels. Finally, while the lack of an undo feature will not directly introduce errors, it may indirectly do so via user frustration and a reluctance to replace points that are slightly misplaced.

3.4 Calculating Vertebral Heart Scores

We implemented functions to take a Pandas dataframe and bulk calculate all VHS scores. If desired, other functions were implemented that can operate on lists of points for a single dog at a time.

3.5 CSV to txt Conversion for YOLO

As YOLO requires a specific formatting of the information to be able to generate the algorithm for the detection of key points and modelling, There was a need to transform the data of each image in the .csv files that we had generated. Using MATLAB the averaged annotation points for each image, was put through different calculations and formatting and transformed to a txt file for each image. These txt files each had 2 rows of key points and needed information regarding heart and spine respectively creating 32 columns, providing a cohesive and ready to use data format for training our YOLO model.

3.6 Inter-Annotator Disagreement

Inter-annotator disagreement refers to inconsistencies in how medical images are annotated by different humans. There may be differences in the boundaries drawn around anatomical structures or regions of interest, that is the measurements and key characteristics of the vertebral heart score technique, when different annotators label the same medical pictures, such as our X-rays. Annotators may have slight differences in judgment about where exactly a boundary should be drawn and lack of detailed annotation guidelines can introduce human errors and inconsistencies.

To measure and prevent inter-annotator disagreement, having more than two annotators label each image can help identify outlier annotations. Providing detailed instructions and training the annotators to understand the nuances of medical imaging and the specific criteria for accurate segmentation such as, anatomical structures, understanding of imaging artifacts, and familiarity with the imaging modality being used, can help standardize the segmentation process and reduce variability. In order to promote consistent labelling, periodic quality assurance, reviewing a sample of annotations can potentially identify differences and mistakes early.

Nearing the project's conclusion, one of the members results started to deviate from the average of the other five annotators, despite the fact that Christopher explained the guidelines in great detail, gave us a thorough understanding of the medical aspect of the annotations, and performed quality control on a few samples of each members annotations. This result occurred in spite of the appropriate precautions we took, which emphasizes the necessity of proper training particularly in research with medical implications. The dependability of machine learning algorithms is directly impacted by the consistency and correctness of

annotations and errors in these early stages can have serious consequences, potentially leading to misdiagnosis or inappropriate treatment decisions.

By investing in thorough training, research teams can reduce the risk of errors downstream and create reliable results with confidence. This ultimately contributes to the development of more accurate and effective diagnostic and treatment tools, benefiting patient care and clinical outcomes.

3.7 Machine Learning

Tensorflow Implementation:

The development of an AI model for identifying oversized hearts involved three iterative models, each leveraging different computational strategies and insights.

1. **Initial Model: Basic CNN Approach:**
   - **Framework & Tools:** Utilized TensorFlow and Keras for model development.
   - **Data Handling:** Processed VHS scores from a CSV file, with image resizing to 256x256 pixels.
   - **Model Architecture:** A basic CNN structure was implemented, comprising a convolutional layer, max pooling, flattening, and dense layers.
   - **Objective:** The model was designed to predict VHS scores using linear activation in the final layer, framing it as a regression problem.
   - **Training & Optimization:** Employed the Adam optimizer and mean squared error loss function over a span of 20 epochs.
2. **Second Model: Enhanced CNN with Varied Optimizers and Loss Functions:**
   - **Data Handling:** Processed VHS scores from a CSV file, with image resizing to 256x256 pixels.
   - **Model Architecture:** A basic CNN structure was implemented, comprising a convolutional layer, max pooling, flattening, and dense layers.
   - **Objective:** The model was designed to predict VHS scores using linear activation in the final layer, framing it as a regression problem.
   - **Training & Optimization:** Employed the Adam optimizer and mean squared error loss function over a span of 20 epochs.
3. **Third Model: Binary Classification for VHS:**
   - **Objective Shift:** This model transitioned to a binary classification task, focusing on whether the VHS value exceeded 10.5.
   - **Data Conversion:** VHS scores were transformed into a binary format, where scores above 10.5 were labeled '1', and all others as '0'.
   - **Model Architecture:** Maintained a CNN structure, but incorporated a sigmoid activation function in the output layer to facilitate binary classification.
   - **Output Focus:** The model was specifically aimed at classifying the heart as oversized or not.

– **Training Methodology:** This phase utilized binary cross-entropy for loss calculation and included accuracy metrics, which were crucial in assessing the model's performance in binary classification.

Key Point Prediction With YOLO-Pose:

The annotation of the radiographs using our software produces points at the start and end of the axes for the heart, and select points along the spine. These annotation points can be predicted using machine learning models for key point prediction. For our project we decided to use the YOLO Version 8 Nano Pose (YOLO) model from Ultralytics to predict these annotation points. The YOLO model is a small, fast, pre-trained model that is capable of adapting to new data sets. It works by first identifying objects in the images, then identifying key points in the objects.

For use with our data, we defined two objects, the heart and a section of the spine starting at T4. The key points were then defined as the four annotation points around the heart and 5 points along the spine. We built the training data set using the average annotation points from our groups annotations, and radiographs resized to 640x640 to work with the model. The model was then trained on our data for a total of 201 epochs.

We ran into some limitations for training our model due to how we designed our annotation software. For each radiograph, the number of annotation points along the spine was determined by the size of the heart axes. This resulted in our annotations ranging from 5 points along the spine to a maximum of 10 points. This meant that we could only train the model on the first 5 points along the spine, limiting the accuracy of the spines predictions.

4 Results

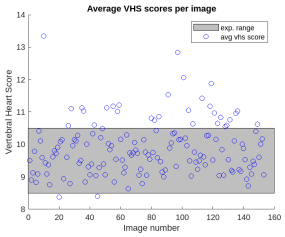
4.1 Statistical Analysis

Matlab code was used to open CSV files of corresponding data and visualize different statistical relations in our data.

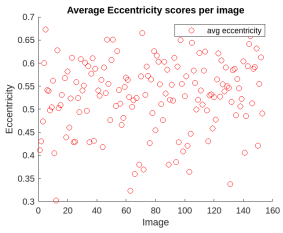
Plotting the average recorded VHS scores for each image showed that most of our data fit within the expected scoring range of 8.5 to 10.5. However, as can be inferred from Fig. 3a, scores for 30 images out of 153 were above our threshold of 10.5. Of the 30 images with VHS scores above the threshold, only 5 were above the threshold by a value of 1.0 or more. Scores for 2 images out of 153 were below our threshold of 8.5. We assume that scores that do not fit into our expected range are the the result of canines with irregular heart sizes, however it must be considered that most of us are not experts in this topic and possibly inaccurate measurements may be causing averages to skew away from the true values.

Given that we had perpendicular long- and short axes for our VHS measurements, we decided to use this to gather insight on heart eccentricity for each image. In Fig. 3b, all average eccentricity value are less than 0.7, and aside from nine outliers, most average eccentricity values are greater than 0.4.

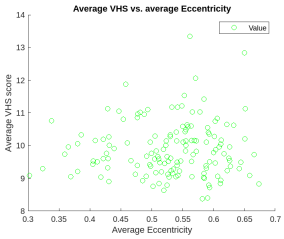
With our average VHS scores and average Eccentricity, we decided to see if we could determine any sort of relation between the eccentricity of the heart and the resulting VHS score. Our results shown in Fig. 3c do not seem to indicate any sort of relation (linear, quadratic, exponential, etc.) between heart eccentricity and VHS scores.



(a) Average VHS scores for each image plotted against expected range (8.5 - 10.5)



(b) Average Eccentricity scores for each image



(c) Average VHS compared to average Eccentricity for corresponding image numbers

4.2 TensorFlow Model Result Analysis

The machine learning component of our study involved the implementation of an AI model to classify images as indicating an oversized heart or not, based on the VHS scoring system. The evaluation of the model's performance yielded the following results:

The confusion matrix for the binary classification model is as follows:

		Enlarged Heart	
		Positive	**Negative**
Prediction	**Enlarged**	60	3
	Normal	2	88

Fig. 4: Confusion matrix for classification model predictions

Table 4 presents the confusion matrix for our TensorFlow binary classification model. It indicates that the model correctly identified 88 true negatives (images correctly identified as not indicating an oversized heart) and 60 true positives (images correctly identified as indicating an oversized heart). There were 3 false

positives (images incorrectly identified as indicating an oversized heart) and 2 false negatives (images incorrectly identified as not indicating an oversized heart). The high number of true positives and negatives suggests that the model is effective in distinguishing between oversized and normal-sized hearts.

Additionally, the mean score difference between the predicted and actual VHS scores was calculated to be approximately 9.875, with a standard deviation of about 0.990. These results suggest that while the model tends to predict scores that are close to the actual values, there is still a notable deviation, indicating room for improvement in the model's accuracy. The relatively low standard deviation implies that the model's predictions are consistently close to this mean difference, showcasing a degree of reliability in its performance.

Overall, these results demonstrate the potential of machine learning models in assisting with the interpretation of canine cardiac radiographs, although further refinement is needed to enhance the accuracy and reliability of such models.

4.3 YOLO Model Result Analysis

The YOLO model was successfully able to identify the heart in our training set images with a confidence ranging from 0.82-0.96 and identify the spine with confidence ranging from 0.46-0.81. Additionally there were 0 false identifications of hearts or spines and each object had the correct number of key points inside them.

For evaluating the performance of the key points predicted by our model, we used the standard deviation from our own annotations. We plotted 2 std and 4 std regions for each average point location over the image along with the predicted key points. We accepted any points inside the standard deviations, coloring the regions cyan, and rejecting the rest, coloring them red.

Evaluating the predicted key points in this way, we had an average of 2.125 correct points in the heart and 1.375 correct points in the spine within 2 standard deviations and an average of 3 correct points in the heart and 2.25 correct points in the heart within 4 standard deviations. Giving us an accuracy of 0.75 points in the heart and 0.45 for points in the spine at the 4 standard deviation range.

5 Contributions

– **Ryan Zrymiak**: Once Christopher completed code to record vertebral heart score annotations on images, I completed my own annotations on all 153 images we had at our disposal. Then, once Christopher completed additional code to take the image annotations and turn them into sample VHS measurements, I extended this code so that I could get VHS measurements for all the annotations each group member performed. Using this data, I created a MATLAB script to visualize statistical relations between our data.

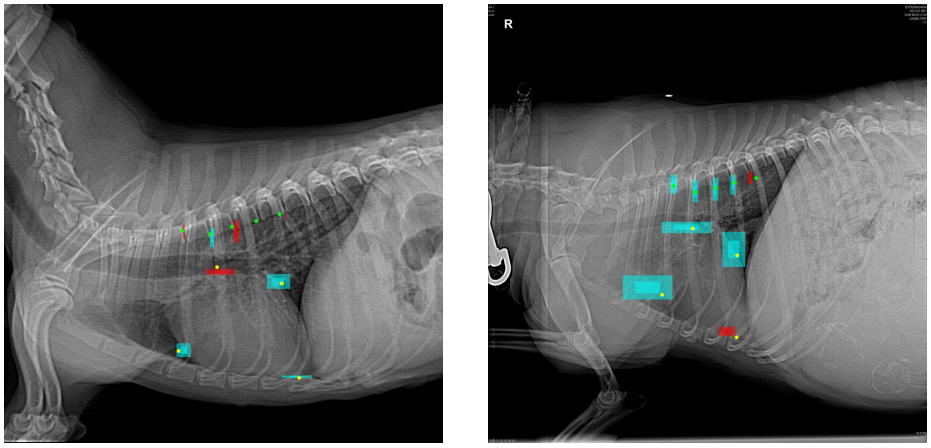


Fig. 5: Example of prediction analysis. The left image is our worst prediction and the right image is our best prediction.

- **Christopher Douglas**: I implemented the software to annotate and score radiographs. Because I am a veterinarian I trained all other group members to perform a vertebral heart scoring and provided guidance on medical topics when needed. I also scored all radiographs. I provided expertise in the diagnosis and treatment of cardiac disease in canine patients.
- **Raul Gomero**: I implemented 3 AI models to automatize the identification of an oversized heart. Based on the VHS scores collected by everyone, the initial two models were designed to forecast VHS scores to manually determine whether they were excessively large. Conversely, the third model introduced a novel aspect to the data by incorporating a column that asked, "Is the value greater than 10.5?" A value of 1 was assigned if the answer was yes, and 0 if no, making this a binary output model. Notably, this third model demonstrated the highest accuracy. I achieved accuracies of 52.3%, 62%, and 92% respectively.
- **Brendan Bickford**: I wrote scripts to rename all initial images and to crop and resize all the images for annotation, and took part in annotating all 153 images in our data set. Then I setup and trained the YOLO-pose model on the key points of our data set and evaluated its performance. Finally I wrote the section of the report about the YOLO model.
- **Rana Hoshyarsadeghi**: After the training, I annotated the provided images to contribute to our sample data. The YOLO phase required specific data analyzation of our average annotations, which I implemented in MATLAB, following the needed formatting and conversion of csv files to txt. As my annotations fell out of the average of others I wrote a section on Inter-Annotator Disagreement for the final report.

6 Conclusion and Discussions

Our results with the YOLO model and especially the TensorFlow model suggests that these methods can be viable in assisting in the calculation of VHS in dogs. With more time and data, we believe that these approaches could significantly improve the accuracy and efficiency for properly diagnosing cardiomegally in dogs. For instance, the YOLO model could be used to predict preliminary annotations that could be adjusted and approved by veterinarians. Additionally, the VHS predictions from the TensorFlow model could be used to trigger warnings if the annotation results disagree with the models predictions. That way veterinarians can reduce the time it takes to diagnose cardiomegally, while still manually confirming the diagnosis.

7 Future Work

The biggest limitation that we have encountered so far has been due to our small data set. If we were to continue working on this project, we would begin by improving our annotation software to gather more points from all images and fix minor bugs. Then we would reach out to veterinary offices and research labs for more images or any existing data set.

By expanding our data set further, we would be able to train the TensorFlow and YOLO models to be more robust and less likely to over fit the data. Then, once we had updated models we would be able to incorporate the prediction models into the annotation software. Once we have the improved annotation software, we could reach out to veterinary schools to test the software further.

Acknowledgements

We would like to thank Dr Ghassan Hamarneh Ph.D for directing us to valuable research papers and assisting us in finalizing our topic. Additionally we would like to thank Dr Ben Cardoen Ph.D for guiding us through structuring our report and answering our questions promptly.

References

[1] J W Buchanan and J Bücheler. “Vertebral scale system to measure canine heart size in radiographs”. en. In: *J. Am. Vet. Med. Assoc.* 206.2 (Jan. 1995), pp. 194–199.

[2] C R Lamb et al. “Use of breed-specific ranges for the vertebral heart scale as an aid to the radiographic diagnosis of cardiac disease in dogs”. en. In: *Vet. Rec.* 148.23 (June 2001), pp. 707–711.

[3] S Kraetschmer et al. “Vertebral heart scale in the beagle dog”. en. In: *J. Small Anim. Pract.* 49.5 (May 2008), pp. 240–243.

[4] Tobias Schwarz and Victoria Johnson, eds. *BSAVA Manual of Canine and Feline Thoracic Imaging*. 1st. British Small Animal Veterinary Association, 2008. URL: https://www.bsavalibrary.com/content/book/10.22233/9781910443088.

[5] Cesar Augusto Flores Duenas. *Radiographic Dataset for VHS determination learning process*. 2020.

[6] 2023. URL: https://opencv.org/ (visited on 11/28/2023).

[7] 2023. URL: https://numpy.org/ (visited on 11/28/2023).

[8] 2023. URL: https://pandas.pydata.org/ (visited on 11/28/2023).

[9] 2023. URL: https://mathworks.org/ (visited on 11/28/2023).

[10] 2023. URL: https://docs.ultralytics.com/ (visited on 12/03/2023).

[11] 2023. URL: https://www.tensorflow.org/ (visited on 12/04/2023).

[12] Python Software Foundation. *Welcome to python.org*. 2023. URL: https://www.python.org/ (visited on 11/28/2023).

[13] *Veterinary Information Network*. 2023. URL: https://vin.com/ (visited on 11/28/2023).