# Dog Cardiomegaly Classification

Aaron Meoded
Katz School of Science and Health, Yeshiva University
meoded@mail.yu.edu

## Abstract

*Cardiomegaly is a prevalent condition in dogs, requiring accurate and timely diagnosis to improve treatment outcomes. Traditional diagnostic methods, such as calculating the Vertebral Heart Score (VHS) from thoracic radiographs, are time-intensive and prone to inter-observer variability. In this work, we develop a convolutional neural network (CNN)-based multilabel classification model for automated assessment of dog cardiomegaly using a dataset of annotated X-ray images. This approach addresses the inefficiencies of manual measurement methods, providing a scalable and reliable solution for improving diagnostic workflows and ensuring consistency in veterinary medicine.*

## 1. Introduction

Cardiomegaly, the abnormal enlargement of the heart, is a critical condition affecting dogs that often serves as an early indicator of underlying cardiac diseases. Accurate and timely diagnosis is crucial for ensuring effective treatment and improving outcomes. Currently, the Vertebral Heart Score (VHS) is a widely used metric in veterinary medicine for assessing heart size through thoracic radiographs. However, VHS calculations require manual measurement of specific axes and vertebrae, a process that is both time-consuming and prone to variability among clinicians. As a result, there is a pressing need for automated and reliable methods to assist in the diagnosis of canine cardiomegaly.

The rapid advancements in deep learning have brought significant improvements to image classification tasks, including medical imaging. In particular, convolutional neural networks (CNNs) have demonstrated remarkable capabilities in extracting features from radiographs and performing diagnostic classifications with high accuracy. Despite these advances, a gap remains between the adoption of deep learning models in clinical practice and the need for interpretable, trustworthy outputs. Some approaches focus on interpretable models, such as those predicting the six key points needed for VHS calculation [5].

This work presents a CNN-based multilabel classification model for automated diagnosis of canine cardiomegaly. The model classifies X-ray images into three categories—small, normal, and large hearts—based on VHS thresholds, addressing the limitations of manual measurement and providing a computationally efficient solution. Unlike regressive or interpretable models, this approach focuses solely on improving classification accuracy while leveraging a robust dataset of annotated dog thoracic radiographs, created by Li et al. [5].

By building upon prior studies that applied deep learning to veterinary imaging , this work highlights the potential of CNN-based models to streamline cardiomegaly detection and aid veterinarians in making consistent and accurate diagnoses. The contributions of this study not only emphasize the practical applications of deep learning in veterinary medicine but also open the door to broader adoption of automated diagnostic tools in the field.

## 2. Related Work

Deep learning has been increasingly applied to veterinary medicine, particularly for the diagnosis of canine heart conditions like cardiomegaly. Traditional methods rely heavily on manual measurements of the Vertebral Heart Score (VHS), which are prone to variability and inefficiency. To address these limitations, researchers have explored deep learning models capable of automating cardiomegaly detection and classification, leveraging the growing availability of annotated thoracic radiographs.

*Regressive vision transformer for dog cardiomegaly assessment* by Li and Zhang [5] introduced an interpretable deep learning approach for dog cardiomegaly assessment using a regressive vision transformer (RVT) model. Their work bridged the gap between deep learning predictions and clinical interpretability by directly predicting the six key points required to calculate the VHS score. The RVT model employed a pyramid vision transformer to extract both high- and low-level features, a feature fusion module for robust feature extraction, and an orthogonal layer to ensure perpendicularity between critical heart axes, thereby improving diagnostic accuracy. The model demonstrated state-of-the-

art performance while maintaining interpretability, a critical factor for clinical adoption. Additionally, the authors developed a dog heart labeling tool and implemented a few-shot generalization strategy to accelerate the data labeling process, ensuring a high-quality dataset for training and evaluation.

Other studies have focused on convolutional neural networks (CNNs) to automate cardiomegaly diagnosis. Zhang et al. [9] proposed a CNN-based system for detecting VHS by predicting the key points needed for its calculation. This approach demonstrated strong alignment with veterinary experts, showcasing the potential of CNNs to enhance diagnostic accuracy while simplifying the diagnostic process. Similarly, Burti et al. [4] developed a CNN-based computer-aided detection (CAD) device to identify cardiomegaly in canine thoracic radiographs. Their system achieved high diagnostic accuracy, although it lacked the interpretability found in models like RVT, which integrate key clinical features. Banzato et al. [2] extended the use of CNNs by developing a classification framework for canine thoracic radiographs, achieving automatic differentiation of radiographs into specific diagnostic categories.

Many studies have explored the application of VHS in diagnosing cardiomegaly in dogs [9, 8, 1]. These works underline the variability and limitations in manually labeled VHS data, such as those highlighted by Rungpupradit et al. [6], who proposed applied VHS methods to account for variability caused by abnormal thoracic vertebrae. Similarly, Tan et al. [7] evaluated Modified Radiographic Chest Volume (mRCV) and VHS for their correlation with pulmonary patterns in dogs, emphasizing the importance of consistent and accurate measurements. Bappah et al. [3] explored the relationship between VHS and cardiac sphericity, finding a strong correlation that underscores the clinical significance of accurate VHS-based diagnostics.

Despite the promise of vision transformers, CNN-based models remain widely used due to their simplicity, computational efficiency, and ability to handle multilabel classification tasks effectively. This work contributes to the growing body of research by focusing on multilabel classification of cardiomegaly categories—small, normal, and large—using a CNN model. Unlike interpretable models such as the RVT, this approach prioritizes classification accuracy while addressing practical challenges in veterinary diagnostics, including dataset size and variability.

## 3. Methods

**Model Development and Refinement:** The process of developing the proposed architecture involved iterative experimentation to balance complexity, performance, and overfitting. Initial attempts with a basic multilayer CNN model and some hyperparameter tuning (learning rate, number of epochs, batch size) showed promising training accuracy but suffered from overfitting, as evidenced by a gap between training and validation accuracy. To address this, reg-

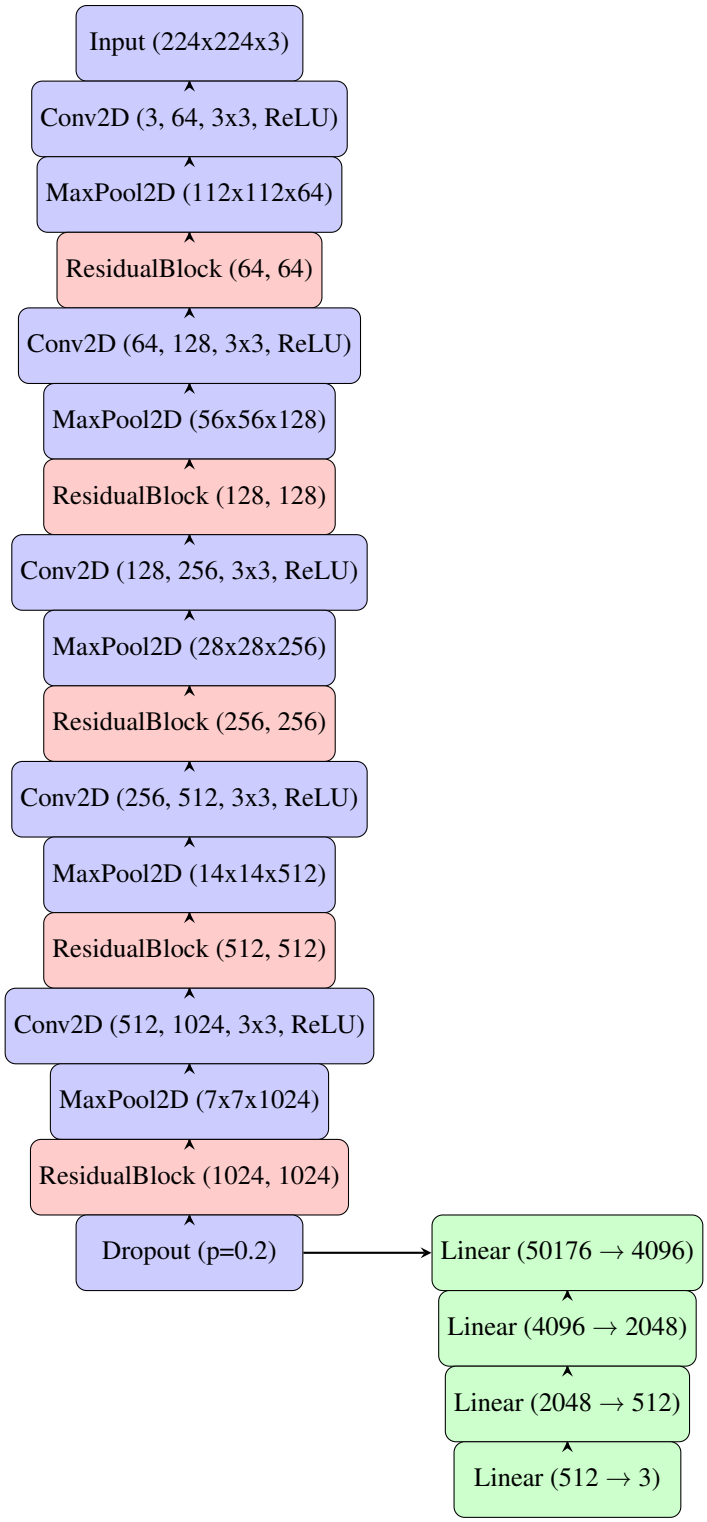


Figure 1. Model architecture

ularization techniques were introduced, including various dropout configurations, image augmentations, and weight decay using the AdamW optimizer.

Having reduced overfitting with these techniques, model complexity was gradually increased to improve feature extraction and classification capability. Experiments with architectures differing in the number of layers, parameters per layer, and additional components, such as residual blocks and self-attention heads, were conducted. The final architecture was selected based on its superior validation and test performance, demonstrating robust generalization and less prominent overfitting.

**Model Architecture:** The finalized architecture is a convolutional neural network designed to classify canine thoracic radiographs into three categories: Small, Normal, and Large hearts. The model comprises five convolutional stages, each beginning with a convolutional layer featuring $3 \times 3$ filters, followed by Batch Normalization and ReLU activation functions, and concluding with a MaxPooling layer for spatial downsampling. The number of filters increases progressively from 64 to 1024 across the stages, enabling the network to capture increasingly complex and abstract features.

To enhance feature extraction and address the vanishing gradient problem, residual blocks are incorporated after each stage. These blocks use skip connections to directly add the input to the block's output, preserving gradient flow and stabilizing training. In cases where the output dimensions change due to downsampling, a compatible transformation is applied to the skip connection to maintain dimensional compatibility.

After the convolutional stages, the extracted features ($1024 \times 7 \times 7$) are flattened into a vector of size 4096 via a fully connected layer, followed by ReLU activation and dropout regularization with a rate of 20%. Two additional fully connected layers, with 2048 and 512 neurons respectively, further process the feature vector. Both layers use ReLU activation and dropout to prevent overfitting. The final classification layer consists of three neurons, corresponding to the target classes, with a softmax activation function to output probabilistic predictions for the multilabel classification task.

**Training Procedure:** The training process was guided by CrossEntropyLoss, a loss function well-suited for multiclass classification tasks. Optimization was performed using the AdamW optimizer with a learning rate of $5 \times 10^{-5}$ and a weight decay of $1 \times 10^{-6}$. This configuration mitigated overfitting by penalizing large weight magnitudes. A StepLR scheduler was employed to halve the learning rate every five epochs, allowing for a gradual refinement of the learning process. Training was conducted for 20 epochs with a batch size of 16, balancing computational efficiency with sufficient parameter updates. Each epoch included separate training and validation phases to monitor the model's generalization performance. In addition to an initial training phase, the best model from this training phase (as measured by validation accuracy) was saved and used in a new training phase where hyperparameters were geared towards more conservative model updates. This included a lower learning rate, batch size, higher AdamW weight decay, and cosine annealing instead of StepLR learning rate increments. This approach allowed the model to benefit from more dynamic learning in early stages of its training, and more precise learning during the later stages of training, improving generalizability.

**Evaluation Metrics:** To comprehensively evaluate the model's performance, metrics such as accuracy, precision, recall, and F1-score were used. A confusion matrix was generated to visualize classification performance across the three categories, highlighting misclassifications and providing insights into areas for improvement. Accuracy served as the primary metric during training and validation to guide optimization and monitor performance trends.

**Implementation Details:** The architecture was implemented using PyTorch, leveraging the modularity of `nn.Module`. Data loaders were employed to manage batching, shuffling, and preprocessing during training and evaluation. Training and testing were conducted on an NVIDIA T4 GPU. Predictions on the test set were saved in a CSV file for external evaluation, ensuring compatibility with software designed to assess classification accuracy on the test set.

# 4. Results

## 4.1. Datasets

The dataset used in this study consists of annotated canine thoracic radiographs, organized into training, validation, and test sets. Its primary purpose is to classify dog cardiomegaly into three categories: small, normal, and large hearts, based on the Vertebral Heart Score (VHS). While the dataset was obtained from Li et al. [5], which included precise key-point labeling for calculating VHS scores, this study does not utilize those detailed annotations. Instead, it focuses on multilabel classification of the cardiomegaly categories, simplifying the labeling process and enhancing scalability.

### 4.1.1 Dataset Composition:

The dataset comprises 2,000 X-ray images, distributed across three sets: a training set with 1,400 images, a valida-

tion set with 200 images, and a test set with 400 unlabeled images. Within the training set, 619 images were labeled as **Large**, 573 as **Normal**, and 208 as **Small**, reflecting a class imbalance that was similarly present in the validation set, with 76 **Large**, 91 **Normal**, and 33 **Small** images. The test set, intended for model evaluation, consists of unlabeled images.

#### 4.1.2 Data Preprocessing:

All images underwent pre-processing to standardize input dimensions and improve the model's generalization capabilities. For the training set, the images were resized to 224 x 224 pixels, normalized using ImageNet-standard mean [0.485,0.456,0.406] and standard deviation [0.229,0.224,0.225], and augmented with random horizontal flipping to increase dataset diversity and reduce overfitting. More complex augmentations were experimented with but resulted in lower validation accuracy. For validation and test sets, only resizing and normalization were applied to ensure consistency while avoiding any data leakage from augmentation.

### 4.2. Results

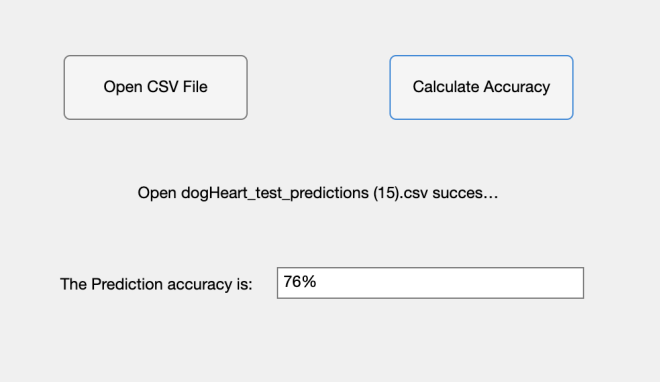


Figure 2. Software test set performance.

The final model achieved a classification accuracy of 89.64% on the training set, 72.50% on the validation set, and 76% on the test set evaluated using external software. These results were achieved through an iterative training process. After an initial run with tuned hyperparameters, further refinements were made to enhance generalization. Specifically, once training plateaued, the learning rate and batch size were reduced, StepLR updates were replaced with cosine annealing for a smoother learning rate decay, and the AdamW weight decay parameter was increased. Training then resumed for additional epochs, allowing the model to avoid overfitting and improve its ability to generalize to unseen data.

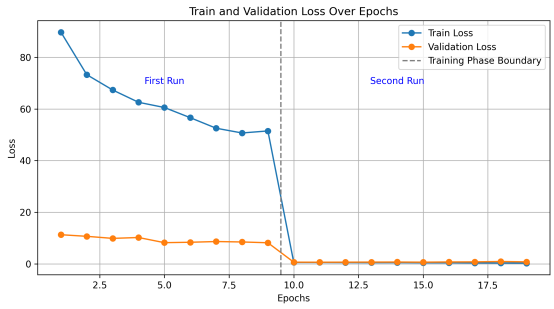


Figure 3. Training and validation loss.

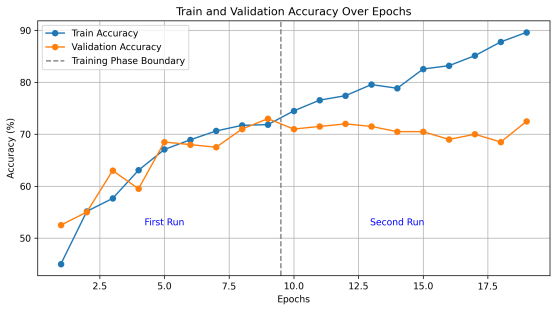


Figure 4. Training and validation accuracy.

The final model also achieved weighted average precision, recall, and F-1 score respectively of 72%, 73%, 72%. The confusion matrix on the model validation set predictions are displayed in Figure 5, with strong performance across all classes, though greater confusion between the Medium and Large classes.

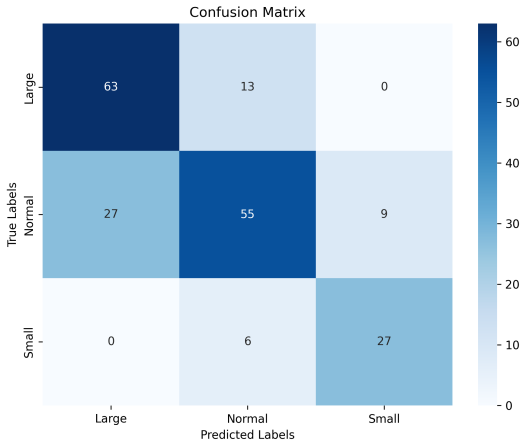


Figure 5. Confusion matrix

#### 4.2.1 Implications and Significance:

The performance of the proposed architecture demonstrates its potential as a reliable tool for the automated classifica-

tion of canine cardiomegaly. The relatively high and balanced accuracy on the validation and test sets suggests that the model effectively captures the underlying patterns in the thoracic radiographs. This indicates that the architecture and training pipeline, including techniques to mitigate overfitting, successfully addressed the challenges of the task.

#### 4.2.2 Implications and Significance:

The performance of the proposed architecture demonstrates its potential as a reliable tool for the automated classification of canine cardiomegaly. The relatively high and balanced accuracy on the validation and test sets suggests that the model effectively captures the underlying patterns in the thoracic radiographs. This indicates that the architecture and training pipeline, including techniques to mitigate overfitting, successfully addressed the challenges of the task.

#### 4.2.3 Results Comparison

The proposed CNN model achieved a test accuracy of **76%**, while the RVT model reported a significantly higher test accuracy of **87.3%** [5]. The RVT approach benefits from a pretrained Pyramid Vision Transformer (PVT) backbone and a regressive framework that predicts six key points to compute the Vertebral Heart Score (VHS). In contrast, our model uses a simpler non-pretrained architecture, focusing on direct classification into **small**, **normal**, and **large** categories.

## 5. Discussion

This study successfully applied deep learning techniques to the task of diagnosing canine cardiomegaly from annotated thoracic radiographs, fulfilling the goals outlined in the abstract and introduction. The developed CNN-based model demonstrated robust performance, achieving a test accuracy of 76% and an F1-score of 92% for the **Small** heart category. This result is particularly noteworthy given the challenges posed by the smaller dataset size and class imbalance. The model's ability to generalize well to unseen data suggests that the regularization techniques, including dropout, AdamW weight decay, and data augmentation, were effective in mitigating overfitting. These findings indicate that the proposed approach could achieve comparable results in real-world clinical applications.

It is important to note that the results presented here were achieved using a non-pretrained deep learning model, specifically designed as an initial test of applying deep learning techniques to this problem case. Despite this limitation, the model performed admirably, underscoring the potential of more advanced implementations. Incorporating pretrained state-of-the-art models, such as vision transformers or hybrid VIT architectures, is expected to significantly enhance performance, particularly in distinguishing between **Normal** and **Large** heart classes. Such models would provide a stronger feature extraction backbone, enabling finer discrimination in borderline cases.

## 6. Conclusion

This study demonstrates the feasibility and effectiveness of applying deep learning techniques to automate the diagnosis of canine cardiomegaly. The proposed CNN-based multilabel classification model achieved consistent and robust performance, with a test accuracy of 76%, highlighting its ability to generalize well to unseen data. By eliminating the need for manual VHS measurements, the model addresses the limitations of traditional diagnostic workflows, streamlining the diagnostic process and improving consistency in veterinary medicine.

The results of this study are particularly promising given the use of a non-pretrained model, emphasizing the potential for further improvements through the adoption of pretrained state-of-the-art architectures. Additionally, expanding the dataset with more samples would enhance the model's ability to distinguish between **Normal** and **Large** heart categories, increasing its clinical utility.

In conclusion, this work highlights the potential of deep learning to transform veterinary diagnostics, providing a scalable, efficient, and accurate solution for the detection of canine cardiomegaly. Future research may focus on leveraging advanced pretrained models, expanding dataset size and diversity, and exploring interpretable visualization techniques to further bridge the gap between AI-driven tools and clinical adoption.

## References


[1] Radu Andrei Baisan and Vasile Vulpe. Vertebral heart size and vertebral left atrial size reference ranges in healthy maltese dogs. *Veterinary Radiology & Ultrasound*, 63(1):18–22, 2022. 2

[2] Tommaso Banzato, Marek Wodzinski, Silvia Burti, Valentina Longhin Osti, Valentina Rossoni, Manfredo Atzori, and Alessandro Zotti. Automatic classification of canine thoracic radiographs using deep learning. *Scientific Reports*, 11(1):3964, 2021. 2

[3] Mu'azu Nuhu Bappah, Nuhu Donga Chom, Maruf Lawal, Abdullaziz Abdullahi Bada, and Saidu Tanko Muhammad. Evaluation of vertebral heart score and cardiac sphericity in apparently normal dogs. *Iranian Journal of Veterinary Surgery*, 16(1):1–4, 2021. 2

[4] Silvia Burti, V. Osti, Alessandro Zotti, and Tommaso Banzato. Use of deep learning to detect cardiomegaly on thoracic radiographs in dogs. *The Veterinary Journal*, 262:105505, july 2020. 2

[5] Jialu Li and Youshan Zhang. Regressive vision transformer for dog cardiomegaly assessment. *Scientific Reports*, 14(1):1539, 2024. 1, 3, 5

[6] Jetsada Rungpupradit and Somchin Sutthigran. Comparison between conventional and applied vertebral heart score (vhs) methods to evaluate heart size in healthy thai domestic shorthair cats. *The Thai Journal of Veterinary Medicine*, 50(4):459–465, 2020. 2

[7] MC Chee Tan, IA Okene, and A Hashim. A retrospective study correlating modified radiological chest volume and vertebral heart score with pulmonary patterns in dogs. *Sahel Journal of Veterinary Sciences*, 17(4):31–36, 2020. 2

[8] Lauren Timperman, Greg Habing, and Eric Green. The vertebral heart scale on ct is correlated to radiographs in dogs. *Veterinary Radiology & Ultrasound*, 62(5):519–524, 2021. 2

[9] Mengni Zhang, Kai Zhang, Deying Yu, Qianru Xie, Binlong Liu, Dacan Chen, Dongxing Xv, Zhiwei Li, and Chaofei Liu. Computerized assisted evaluation system for canine cardiomegaly via key points detection with deep learning. *Preventive Veterinary Medicine*, 193:105399, 2021. 2