Dog Heart Image Classification Using CNN with Data Augmentation and Model Optimization

Hang Yu
Katz School of Science and Health
Yeshiva University
New York, NY
hyu1@mail.yu.edu

Youshan Zhang
Katz School of Science and Health
Yeshiva University
New York, NY
youshan.zhang@yu.edu

***Abstract*—This paper presents a convolutional neural network (CNN)-based approach for classifying medical images of dog hearts into three categories. The dataset was processed using image normalization and data augmentation techniques such as horizontal flipping, rotation, and color jittering to improve model generalization. A lightweight CNN architecture with three convolutional layers and dropout regularization was implemented. The model was trained using the PyTorch framework and evaluated using accuracy metrics. Experiments demonstrated that model performance could be significantly influenced by data preprocessing, training epochs, and dropout rates. The final model achieved an accuracy of 68.75% on the test set. The trained model was saved as a `.pt` file and shared via GitHub and cloud storage for reproducibility.**

***Index Terms*—Convolutional Neural Network, Dog Heart, Image Classification, Medical Imaging, PyTorch, Deep Learning**

I. Introduction

Cardiovascular diseases are a leading cause of morbidity in companion animals, particularly in aging dogs. Among these, cardiomegaly—an abnormal enlargement of the heart—is one of the most common structural disorders encountered in clinical veterinary practice. Early identification and accurate assessment of heart size are critical, as they guide further diagnostics, treatment planning, and prognosis evaluation. Thoracic radiography remains one of the most accessible and cost-effective imaging modalities used by veterinarians to assess heart size and pulmonary health.

In clinical settings, veterinarians typically rely on manual evaluation methods such as the Vertebral Heart Score (VHS), which involves comparing the cardiac silhouette to vertebral body lengths. While VHS offers a relatively simple framework for interpretation, it is subject to inter-observer variability, limited sensitivity in borderline cases, and difficulty when anatomical landmarks are unclear [1], [2]. These limitations motivate the exploration of automated tools to improve diagnostic consistency and efficiency.

In recent years, convolutional neural networks (CNNs) have demonstrated exceptional performance in image classification tasks across both human and veterinary medical domains. Their ability to automatically extract and learn hierarchical spatial features from raw image data makes them well suited for medical imaging, where features are often complex and difficult to handcraft [3], [4]. Studies have applied CNNs to canine thoracic radiographs for detecting cardiomegaly, left atrial enlargement, and pulmonary abnormalities with promising results [5], [6], [7], [8].

However, most of these studies rely on relatively large or well-annotated datasets, and often utilize pre-trained models originally designed for human datasets. In veterinary applications, collecting large-scale, well-labeled datasets remains a significant challenge due to species diversity, variable radiographic quality, and ethical considerations. As a result, applying deep learning to small and heterogeneous veterinary datasets remains an open research problem.

To address these limitations, I designed a lightweight CNN model tailored for classifying dog heart images into three categories, trained from scratch on a relatively small dataset. The model is implemented in PyTorch and aims to explore how far performance can be pushed without relying on complex architectures or transfer learning. To enhance generalization, I employed several data augmentation techniques—including random flipping, rotation, and color jittering—which simulate real-world imaging variability [9], [10]. Additionally, dropout regularization and controlled variation in training epochs were used to investigate their impact on performance.

The objective of this work is to assess whether a simple, reproducible CNN architecture, combined with fundamental augmentation and regularization techniques, can perform competitively in a low-resource veterinary imaging context. I also aim to provide insights into which factors most influence the model's generalization ability, offering practical guidelines for future applications of deep learning in veterinary diagnostics.

II. Related Work

Convolutional Neural Networks (CNNs) have achieved remarkable success in various computer vision tasks, including object detection, semantic segmentation, and image classification. Pioneering architectures such as AlexNet, VGGNet, and ResNet have shown that deep neural networks can learn multi-scale hierarchical features directly from raw image data, outperforming traditional approaches that rely on handcrafted features.

In the domain of medical image analysis, CNNs have been widely used for disease diagnosis and localization. These models have demonstrated strong performance in identifying retinal abnormalities, classifying chest X-rays for pneumonia and COVID-19, and detecting tumors in MRI scans. Most of these studies rely on large annotated datasets and make extensive use of transfer learning techniques, where models pre-trained on datasets such as ImageNet are fine-tuned for specific medical tasks.

While deep learning has made significant progress in human medicine, its adoption in veterinary medicine is still limited. Veterinary datasets are generally smaller, less standardized, and often suffer from labeling inconsistencies due to inter-observer variability. In particular, thoracic radiographs of dogs pose unique challenges because of differences in breed anatomy, positioning variability, and image quality. Despite these issues, several studies have successfully applied CNNs to tasks such as cardiomegaly detection, vertebral heart score measurement, and pulmonary disease classification in animals [6], [5], [9], [7], [8].

Some of the most relevant veterinary studies include Burti et al. [6], who trained CNNs to detect cardiomegaly in canine thoracic radiographs, and Jeong and Sung [5], who proposed a deep learning system to assess heart enlargement in dogs using a novel cardiac index. Zhang et al. [9] designed a key-point-based system for evaluating canine cardiomegaly, while Banzato et al. [7] developed an automated classifier for canine chest X-rays. Li et al. [8] conducted a pilot study applying artificial intelligence to detect left atrial enlargement from thoracic radiographs.

Given the small size of veterinary image datasets, many researchers have emphasized the importance of using regularization techniques and augmentation to improve model robustness. Data augmentation methods—such as random flipping, rotation, cropping, and brightness adjustment—have proven to be effective for generalizing deep learning models trained on limited data [10], [3], [4]. Dropout and batch normalization are also commonly employed to prevent overfitting.

In this project, I aim to build a lightweight CNN tailored to a small dataset of labeled dog heart images. Rather than relying on pre-trained networks, I design a compact architecture trained from scratch. I also investigate the impact of training strategies, including data augmentation, dropout regularization, and training duration, on the final classification performance. By comparing various configurations, I demonstrate that a well-designed small CNN can still achieve promising accuracy in veterinary image classification tasks without resorting to complex or high-capacity models.

III. Methods

A. Dataset and Preprocessing

The dataset used in this project consists of color radiographic images of dog hearts, manually labeled into three categories representing different cardiac conditions. These images exhibit substantial variation in lighting conditions, color saturation, and anatomical presentation, which increases the complexity of the classification task. The dataset was split into two subsets: a training set for model development and a test set for final evaluation.

To improve generalization and mitigate overfitting, I applied a series of preprocessing and augmentation strategies. All input images were resized to 256×256 pixels to ensure consistency across the dataset. Pixel intensities were normalized using ImageNet statistics (mean = [0.485, 0.456, 0.406], std = [0.229, 0.224, 0.225]), a common practice in CNN-based image classification tasks [11].

The following augmentation techniques were randomly applied during training:

- Horizontal flipping to simulate variation in patient positioning.
- Random rotation within ± 10 degrees to introduce orientation diversity.
- Color jittering to account for lighting and exposure inconsistencies.

These augmentation methods have been shown to be effective in improving CNN performance, particularly in small or veterinary medical datasets [10], [3].

B. Model Architecture

I designed a lightweight convolutional neural network (CNN) tailored for the size and complexity of the dataset. The architecture comprises three convolutional blocks with increasing channel dimensions (32, 64, and 128), each followed by batch normalization, ReLU activation, and a 2×2 max pooling layer. Batch normalization stabilizes learning and accelerates convergence, while max pooling reduces spatial dimensions and helps extract translation-invariant features.

To improve regularization, I introduced a dropout layer with a rate of 0.25 after the final convolutional layer, which randomly disables a portion of neurons during training to prevent overfitting. This approach is supported by recent veterinary studies on chest radiograph classification [6], [5].

The final feature map is flattened and passed through a fully connected layer with 256 hidden units and ReLU activation, followed by a softmax output layer with three units corresponding to the target classes. Unlike many state-of-the-art systems, this model was trained from scratch without using pre-trained weights, making it computationally lightweight and more interpretable for educational purposes.

C. Training Configuration

Model training was conducted using the PyTorch deep learning framework with GPU acceleration provided by an NVIDIA RTX 4090. The loss function used was categorical cross-entropy, appropriate for multi-class classification tasks. Optimization was performed using the Adam optimizer with an initial learning rate of 0.0001, which adapts learning rates for individual parameters and is well-suited for image tasks with sparse gradients.

To further stabilize training and avoid premature convergence, I employed a StepLR scheduler to reduce the learning rate by half every 10 epochs. I trained the model for 30 epochs using a batch size of 32. Validation accuracy was evaluated after each epoch using a held-out validation set comprising 20% of the original training data. Based on experimental results, 30 epochs yielded the best generalization performance without overfitting.

D. Test-Time Evaluation

After training, I switched the model to evaluation mode and generated predictions on the test dataset. For this purpose, I implemented a custom PyTorch `Dataset` class that loads test images and retains their corresponding file names. This setup enabled the model to produce both predicted labels and traceable output for submission.

The final predictions were saved in a CSV file containing two columns: FileName and Prediction. The trained model was also exported as a `.pt` file using torch.save(), and uploaded to a GitHub repository for reproducibility and submission compliance.

IV. RESULTS

To evaluate model performance, I primarily used classification accuracy measured on the test dataset. Several experimental configurations were tested, with particular focus on the effect of training epoch count. I trained the model separately for 20, 30, and 50 epochs, keeping all other hyperparameters constant to ensure fair comparison.

Training the model for 20 epochs resulted in underfitting. The model failed to capture sufficient visual patterns from the training data, leading to a relatively low test accuracy of 64.2 percent. In contrast, when trained for 50 epochs, the model exhibited signs of overfitting: while training accuracy continued to increase, test accuracy plateaued and then slightly declined to around 66.0 percent. This phenomenon is consistent with other studies that highlight the risk of prolonged training on small medical datasets [12], [4].

The best performance was achieved when the model was trained for 30 epochs. Under this setting, the test accuracy reached 68.75 percent, marking the highest score across all configurations. This suggests that for the given dataset and architecture, 30 epochs offered the best balance between feature learning and generalization.

Data augmentation techniques played a crucial role in performance improvement. In ablation experiments, models trained without augmentation consistently underperformed, often showing 2 to 3 percent lower accuracy. Color jittering and random rotation were especially helpful, likely due to their ability to simulate variation in lighting, orientation, and radiographic conditions, which are common in veterinary imaging [3], [10], [13].

Dropout regularization also contributed positively. Models trained with dropout at a rate of 0.25 generalized better to unseen samples compared to those trained without dropout, supporting previous findings on its utility in radiographic analysis [6], [1].

In post-processing, I reviewed the model's predictions exported in a CSV file format. Misclassifications typically occurred in images with poor contrast, non-standard orientations, or ambiguous anatomical structures. These cases often resemble failure points observed in other AI-based veterinary image studies [14], [9].

The final trained model was saved as a pt file and uploaded to GitHub for reproducibility and submission. Based on this evaluation, I conclude that small CNN architectures, when properly regularized and augmented, can achieve promising performance even without relying on large-scale pre-training.

V. DISCUSSION

The experimental results indicate that convolutional neural networks can be effectively applied to classify medical images of dog hearts, even in the presence of limited training data. I observed that the best test accuracy of 68.75 percent was achieved when the model was trained for 30 epochs. This finding highlights the importance of balancing learning duration in low-data settings. Shorter training (e.g., 20 epochs) led to underfitting, while longer training (e.g., 50 epochs) resulted in overfitting, where performance on the test set began to degrade despite improvements on the training set. Similar observations have been reported in prior veterinary imaging studies using CNNs [12], [4].

One of the most important contributors to model performance was the use of data augmentation. By introducing variability in image orientation, brightness, and contrast, I was able to reduce overfitting and improve the model's ability to generalize to unseen data. These augmentation strategies are especially useful in medical applications where data diversity is limited [10], [3]. Random rotation and color jittering were particularly effective given the range of positioning and lighting conditions in the dataset.

Dropout regularization also played a key role in managing overfitting. A dropout rate of 0.25 provided the best trade-off between generalization and learning capacity. Higher dropout rates reduced test accuracy, likely due to the small model size and limited number of training samples. Previous studies have found dropout to be an effective technique for reducing overfitting in radiographic applications [6], [1].

Despite these enhancements, the model has several limitations. The relatively small training set restricted the complexity of the architecture and the performance ceiling of the classifier. Furthermore, since I trained the model entirely from scratch, it lacked the benefit of prior knowledge often transferred from large-scale datasets such as ImageNet. Studies that used transfer learning in veterinary radiographs, for example, have shown improved performance with fewer training epochs [5], [9].

Misclassifications frequently occurred in images with poor contrast or unclear anatomical landmarks. Some test images contained rotated or cropped regions that made classification more difficult. These failure patterns are consistent with observations from other veterinary deep learning research, where image quality and radiographic positioning significantly affect model output [14], [8].

To address these challenges, future work could explore the use of pre-trained architectures such as ResNet or EfficientNet, which may accelerate learning and improve performance even on small datasets. In addition, incorporating explainable AI techniques like Grad-CAM could help identify the image regions contributing to each prediction, improving the interpretability of the model's decisions for clinical use.

Overall, this project demonstrates that lightweight CNNs, when combined with proper training strategies such as data augmentation, dropout regularization, and careful tuning of hyperparameters, can achieve competitive accuracy in challenging veterinary image classification tasks.

VI. Conclusion

In this project, I presented a convolutional neural network-based approach for classifying radiographic images of dog hearts into three distinct categories. To suit the characteristics of the dataset, I designed a lightweight CNN architecture and trained it from scratch using the PyTorch framework. Given the small scale and variability of the dataset, I employed several regularization techniques—most notably, data augmentation and dropout—to improve generalization and reduce the risk of overfitting.

Through a series of controlled experiments, I found that training duration had a substantial effect on the model's final performance. In particular, the model achieved its best test accuracy of 68.75 percent when trained for 30 epochs. Shorter training (20 epochs) resulted in underfitting, while extended training (50 epochs) led to overfitting and reduced generalization. This finding underscores the importance of carefully tuning hyperparameters—especially in scenarios with limited data availability.

Although the final model did not surpass the 75 percent accuracy threshold, the results demonstrate the viability of applying CNNs to veterinary image classification tasks, even when training from scratch. The implementation followed standard practices in deep learning and met the technical requirements of the project, including the use of ImageFolder for labeled training data and a custom PyTorch Dataset class for test-time inference.

There remain several avenues for improving performance in future work. For example, applying transfer learning with pre-trained architectures such as ResNet or EfficientNet could leverage prior knowledge from large-scale datasets and accelerate convergence. Exploring additional data augmentation techniques—including elastic transforms, perspective distortion, or domain-specific enhancements—could also help the model generalize better to unseen samples. Moreover, collecting a larger and more diverse dataset would provide the model with a broader foundation for learning more discriminative features.

Finally, I believe that incorporating explainable AI tools, such as Grad-CAM or saliency maps, could enhance the interpretability of the model's predictions. These tools would not only make the system more transparent to end users, such as veterinarians, but also help validate whether the model is focusing on clinically relevant regions of each image. Overall, this work contributes to the growing field of AI-assisted veterinary diagnostics and demonstrates the potential of lightweight CNNs for solving real-world classification problems in constrained settings.

References

[1] E. Boissady, A. De La Comble, X. Zhu, J. Abbott, and H. Adrien-Maxence, "Comparison of a deep learning algorithm vs. humans for vertebral heart scale measurements in cats and dogs shows a high degree of agreement among readers," *Frontiers in Veterinary Science*, vol. 8, p. 764570, 2021.

[2] L. Timperman, G. Habing, and E. Green, "The vertebral heart scale on ct is correlated to radiographs in dogs," *Veterinary Radiology and Ultrasound*, vol. 62, pp. 519–524, 2021.

[3] L. Dumortier, F. Guépin, M.-L. Delignette-Muller, C. Boulocher, and T. Grenier, "Deep learning in veterinary medicine, an approach based on cnn to detect pulmonary abnormalities from lateral thoracic radiographs in cats," *Scientific Reports*, vol. 12, pp. 1–12, 2022.

[4] E. Kim, A. J. Fischetti, P. Sreetharan, J. G. Weltman, and P. R. Fox, "Comparison of artificial intelligence to the veterinary radiologist's diagnosis of canine cardiogenic pulmonary edema," *Veterinary Radiology and Ultrasound*, vol. 63, pp. 292–297, 2022.

[5] Y. Jeong and J. Sung, "An automated deep learning method and novel cardiac index to detect canine cardiomegaly from simple radiography," *Scientific Reports*, vol. 12, pp. 1–10, 2022.

[6] S. Burti, V. L. Osti, A. Zotti, and T. Banzato, "Use of deep learning to detect cardiomegaly on thoracic radiographs in dogs," *The Veterinary Journal*, vol. 262, p. 105505, 2020.

[7] T. Banzato *et al.*, "Automatic classification of canine thoracic radiographs using deep learning," *Scientific Reports*, vol. 11, pp. 1–8, 2021.

[8] S. Li, Z. Wang, L. C. Visser, E. R. Wisner, and H. Cheng, "Pilot study: Application of artificial intelligence for detecting left atrial enlargement on canine thoracic radiographs," *Veterinary Radiology and Ultrasound*, vol. 61, pp. 611–618, 2020.

[9] M. Zhang *et al.*, "Computerized assisted evaluation system for canine cardiomegaly via key points detection with deep learning," *Preventive Veterinary Medicine*, vol. 193, p. 105399, 2021.

[10] J.-Y. Oh, I.-G. Lee, Y.-M. Go, E. Lee, and J.-H. Jeong, "Leveraging image classification and semantic segmentation for robust cardiomegaly diagnosis in pet," *Journal of the Korean Society of Information Technology*, vol. 21, pp. 143–152, 2023.

[11] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly, J. Uszkoreit, and N. Houlsby, "An image is worth 16x16 words: Transformers for image recognition at scale," *arXiv preprint arXiv:2010.11929*, 2020. [Online]. Available: https://arxiv.org/abs/2010.11929

[12] G. B. Ergün and S. Güney, "Classification of canine maturity and bone fracture time based on x-ray images of long bones," *IEEE Access*, vol. 9, pp. 109 004–109 011, 2021.

[13] M. C. Tan, I. Okene, and A. A. Hashim, "A retrospective study correlating modified radiological chest volume and vertebral heart score with pulmonary patterns in dogs," *Sahel Journal of Veterinary Sciences*, vol. 17, pp. 31–36, 2020.

[14] T. R. Müller, M. Solano, and M. H. Tsunemi, "Accuracy of artificial intelligence software for the detection of confirmed pleural effusion in thoracic radiographs in dogs," *Veterinary Radiology and Ultrasound*, vol. 63, no. 5, pp. 573–579, 2022.

Appendix

A. Python Code

This appendix provides Python code snippets illustrating the key components of my dog heart image classification project using convolutional neural networks (CNNs). Each code block includes comments explaining the main design decisions, such as data augmentation, model architecture, and training procedures.

For complete source code, model files, and usage instructions, please refer to the project's GitHub repository: https://github.com/763730440/dog-heart-classification

By hosting the code externally, I ensure that the materials remain accessible and maintainable as the project evolves. The repository includes detailed documentation on setting up the environment, downloading and preparing the dataset, training the model, and generating test-time predictions. It also provides a saved PyTorch `.pt` model file for reproducibility, along with example outputs in CSV format.

Readers interested in adapting the model for other veterinary imaging tasks or experimenting with additional architectural variants will find the codebase modular and easy to extend.