Dog Heart X-Ray Classification using Generated Images

Sahil Kakkar

Yeshiva University
skakkar@mail.yu.edu

Abstract. In this study, We aimed to enhance the accuracy of a classification model for dog heart X-ray images by incorporating generated images into the training dataset. The initial dataset comprised three classes: Large, Normal, and Small, based on Vertebral Heart Size (VHS) scores. We trained a timm model on this dataset, achieving a test accuracy of 71%. To improve this performance, we fine-tuned a stable diffusion model to generate new X-ray images, which were then manually labeled and categorized based on their VHS scores. Images with VHS < 8.2 were classified as Small, those with VHS > 10 as Large, and those in between as Normal. Incorporating these labeled generated images into the training set and retraining the model resulted in a significant improvement, achieving an accuracy of 80.25% on the test set. This represents a 10% increase in accuracy, demonstrating that generated images can effectively augment training datasets and enhance model performance. Extensive hyperparameter tuning was performed to achieve these results, underscoring the potential of synthetic data in medical image classification tasks. Here's the GitHub link: https://github.com/Sahil1776/Dog-X-Ray-Classification-with-Generated-Images

1 Introduction

Accurate classification of medical images is crucial for diagnosing and treating various health conditions in both human and veterinary medicine. In veterinary cardiology, the Vertebral Heart Size (VHS) score is a widely used metric to assess heart size and diagnose conditions such as cardiomegaly in dogs. However, developing robust machine learning models for medical image classification often requires large, well-labeled datasets, which are difficult to obtain due to the specialized nature of the task and the time-intensive process of manual labeling.

Recent advancements in generative models, such as stable diffusion, offer a promising solution to the data scarcity problem. These models can generate high-quality synthetic images that can augment existing datasets, potentially improving the performance of machine learning models. In this study, we explore the application of stable diffusion-generated images to enhance the classification accuracy of a model trained on dog heart X-ray images.

The primary dataset used in this study consisted of dog heart X-ray images categorized into three classes based on their VHS scores: Large, Normal, and Small. Initial experiments with a classification model achieved a test accuracy of 71%. To improve this performance, we generated additional images using a fine-tuned stable diffusion model, manually labeled these images, and incorporated them into the training dataset. Retraining the model with this augmented dataset resulted in a significant increase in accuracy to 80.25%.

This study demonstrates the potential of synthetic data to improve the performance of medical image classification models. By leveraging generated images, it is possible to enhance model accuracy and reduce the dependency on large, manually labeled datasets. The findings highlight the importance of data augmentation techniques in the field of medical imaging and offer a new approach to addressing the challenges associated with limited data availability.

2 Related Work

The application of data augmentation and synthetic data generation in machine learning has been extensively explored across a variety of domains, including medical imaging, computer vision, and natural language processing. Each of these fields benefits from techniques that can create additional data to improve model performance and robustness.

In the realm of medical imaging, acquiring large and diverse annotated datasets presents significant challenges. Privacy concerns, coupled with the need for expert annotation, make it difficult to gather extensive datasets. To address these issues, generative models have been increasingly utilized. Among these models are Generative Adversarial Networks (GANs), Variational Autoencoders (VAEs), and, more recently, diffusion models.

Generative Adversarial Networks (GANs) are composed of two neural networks, a generator and a discriminator, that compete in a game-theoretic framework. The generator creates synthetic data, while the discriminator attempts to distinguish between real and synthetic data. Through this adversarial process, GANs can produce highly realistic images that augment existing datasets.

Variational Autoencoders (VAEs) work by encoding input data into a latent space and then decoding it back into data. VAEs learn a probabilistic model of the data distribution, allowing them to generate new samples that are similar to the original data. This capability is useful for creating variations of medical images to increase dataset size and diversity. Diffusion models represent a more recent advancement in generative modeling. These models work by iteratively adding noise to data and then learning to reverse this process, progressively generating new data samples. Diffusion models have shown promise in producing high-quality, realistic synthetic data that can significantly enhance medical imaging datasets.

In computer vision, data augmentation techniques are used to artificially expand training datasets by applying transformations such as rotations, scaling, and cropping. This helps improve the general-

ization of models and reduces the risk of overfitting. Synthetic data generation in this domain includes the creation of virtual environments and objects to train models in scenarios that might be rare or difficult to capture in real life.

In natural language processing (NLP), data augmentation and synthetic text generation can involve techniques such as paraphrasing, synonym replacement, and back-translation. These methods help increase the diversity of textual data and improve model performance in tasks like text classification, machine translation, and sentiment analysis.

Overall, the integration of data augmentation and synthetic data generation techniques across these domains not only helps to overcome data limitations but also enhances model performance by providing more diverse and representative training examples.

2.1 Generative Adversarial Networks (GANs)

GANs, introduced by Goodfellow et al. in 2014, have been widely adopted for generating realistic synthetic images. The GAN framework consists of two neural networks: a generator that produces synthetic data and a discriminator that evaluates the authenticity of the data. The two networks are trained simultaneously in a competitive setting, where the generator aims to fool the discriminator with increasingly realistic images, and the discriminator aims to distinguish between real and synthetic images. This adversarial process continues until the generator produces images that are indistinguishable from real images.

In the medical imaging field, GANs have been used for various purposes, including image synthesis, image-to-image translation, and data augmentation. For example, GANs have been employed to generate high-resolution medical images from low-resolution inputs, enhance the quality of medical images by removing noise or artifacts, and create synthetic training data for rare medical conditions. These applications have demonstrated the potential of GANs to improve the performance of machine learning models in tasks such as disease detection, segmentation, and classification.

However, training GANs can be challenging due to issues like mode collapse, where the generator produces limited variations of images, and instability in the training process. Despite these challenges, several variants of GANs, such as Wasserstein GAN (WGAN), CycleGAN, and StyleGAN, have been developed to address these limitations and improve the quality and diversity of generated images. WGAN, for instance, introduces a more stable training objective, while CycleGAN facilitates image-to-image translation tasks without paired training examples.

2.2 Variational Autoencoders (VAEs)

Generative Adversarial Networks (GANs), introduced by Ian Goodfellow and colleagues in 2014, have become a powerful tool for generating realistic synthetic images. The core framework of GANs consists of two neural networks that are trained in tandem:

Generator: This network is responsible for producing synthetic data. Its goal is to create images that are as realistic as possible in order to deceive the other network.

Discriminator: This network evaluates the authenticity of the data, distinguishing between real images from the training set and synthetic images produced by the generator.

During training, the two networks engage in a competitive process. The generator seeks to improve its ability to create convincing images, while the discriminator works to enhance its ability to distinguish between real and synthetic images. This adversarial training continues until the generator produces images that are virtually indistinguishable from real ones, as the discriminator becomes increasingly adept at detecting subtle differences.

In the field of medical imaging, GANs have been employed for various purposes, demonstrating their versatility and effectiveness:

Image Synthesis: GANs can generate high-resolution medical images from lower-resolution inputs. This is particularly useful in cases where high-resolution images are not available but are needed for detailed analysis.

Image-to-Image Translation: GANs facilitate transformations between different types of images. For example, they can convert images of anatomical structures from one modality (such as MRI) to another (such as CT), or improve image quality by removing noise and artifacts.

Data Augmentation: GANs generate synthetic images to augment existing datasets, which is especially valuable for rare medical conditions where annotated data may be scarce. This augmentation helps to create more comprehensive and balanced datasets, improving the performance of machine learning models.

These applications of GANs have significantly enhanced the performance of models in medical imaging tasks such as disease detection, segmentation, and classification by providing high-quality, diverse training data. However, training GANs is not without challenges:

Mode Collapse: This issue occurs when the generator produces a limited variety of images, failing to capture the full diversity of the data distribution.

Training Instability: GAN training can be unstable, leading to difficulties in achieving convergence and generating high-quality images consistently. To address these challenges, several variants of GANs have been developed:

Wasserstein GAN (WGAN): Introduces a new training objective that improves stability and addresses some of the issues associated with traditional GAN training. WGAN uses a Wasserstein distance metric to measure the difference between the real and generated distributions, which helps to stabilize the training process.

CycleGAN: Designed for image-to-image translation tasks, CycleGAN allows for transformations between different image domains without the need for paired training examples. This is useful in cases where acquiring paired images is difficult.

StyleGAN: Focuses on generating high-quality, diverse images with fine-grained control over the generated content. StyleGAN introduces advanced techniques for image synthesis, allowing for the creation of highly realistic images.

These GAN variants have improved the quality and diversity of generated images, making them more practical for various applications in medical imaging and beyond.

2.3 *Diffusion Models*

Diffusion Models, including the Stable Diffusion model used in this study, represent a newer class of generative models that have gained significant attention for their ability to produce high-quality images. These models operate by iteratively refining a noisy image until it aligns with the desired distribution, resulting in high-fidelity outputs. This iterative refinement process allows for more controlled and stable image generation compared to traditional generative models such as GANs and VAEs.

The Stable Diffusion model, specifically, has demonstrated strong performance in generating images from textual descriptions. This capability makes it an excellent choice for our data augmentation needs. By fine-tuning Stable Diffusion on custom captions, we can generate synthetic images that accurately reflect the characteristics of our real-world dataset. This approach offers a flexible and scalable solution for augmenting datasets, especially in scenarios where labeled data is limited.

In summary, diffusion models like Stable Diffusion provide a robust method for generating high-quality images through an iterative refinement process. Their ability to produce detailed and realistic images from text makes them highly effective for enhancing datasets and addressing challenges related to data scarcity.

2.4 *Combining Generative Models with Advanced Architectures*

The integration of generative models with advanced architectures like EfficientNet and vision transformers (ViTs) has significantly enhanced the capabilities of machine learning models in medical imaging.

EfficientNet, introduced by Tan and Le in 2019, employs a compound scaling method to increase the model's depth, width, and resolution. This approach allows EfficientNet to achieve better performance with fewer parameters and computational resources compared to other models. By balancing these factors, EfficientNet provides a more efficient and effective solution for image classification and other tasks.

Vision Transformers (ViTs), introduced by Dosovitskiy et al. in 2020, utilize self-attention mechanisms to process image patches. This method enables ViTs to capture long-range dependencies and detailed features from images, offering a new way to handle visual data that differs from traditional convolutional approaches.

Combining these advanced architectures with generative models results in robust feature extraction and improved accuracy in medical image analysis:

Integrating EfficientNet with images generated by GANs can enhance model performance in tasks such as image classification and segmentation. EfficientNet's ability to efficiently scale its architecture complements the high-quality synthetic images produced by GANs, leading to better overall model results.

Similarly, combining ViTs with synthetic data generated by diffusion models can improve the model's ability to capture complex patterns and relationships within image data. The detailed features extracted by ViTs work well with the high-fidelity images produced by diffusion models, enhancing the model's performance in understanding and analyzing medical images.

Overall, this integration of generative models with advanced architectures like EfficientNet and ViTs represents a powerful approach to improving machine learning models for medical imaging, leading to more accurate and effective analysis.

2.5 *Applications in Veterinary Medicine*

In veterinary medicine, the use of generative models for data augmentation is still a relatively new area of research. However, there have been promising studies that demonstrate the potential of these techniques in improving diagnostic tools. For example, GANs have been used to generate synthetic images of canine radiographs for training deep learning models in tasks such as fracture detection and bone segmentation. These studies highlight the benefits of using synthetic data to overcome the limitations of small datasets and improve model performance.

The application of diffusion models in veterinary diagnostics, as explored in this study, represents a significant advancement in this field. By leveraging the Stable Diffusion model to generate high-quality synthetic images, we can augment the dataset and enhance the predictive accuracy of the Norberg Angle model. This approach not only addresses the issue of data scarcity but also introduces variability and diversity that can help the model generalize better to different conditions and scenarios.

2.6 *Broader Implications and Future Directions*

The findings of this study have broader implications for the use of generative models in medical image analysis, beyond veterinary medicine. The ability to generate high-quality synthetic data can benefit various applications, such as disease detection, anomaly detection, and image enhancement, where labeled data is often limited. By integrating generative models with advanced architectures, we can develop robust and accurate predictive models that can be applied in clinical practice.

Future research can explore the combination of different generative models, such as GANs, VAEs, and diffusion models, to generate even more diverse and high-quality synthetic data. Additionally, applying these techniques to other diagnostic measures and conditions can further validate the benefits of data augmentation in medical image analysis. The continued advancement of generative models and their integration with state-of-the-art machine learning architectures holds great promise for improving the accuracy and reliability of diagnostic tools in healthcare.

3 Methods

3.1 *Dataset*

The initial dataset comprised dog heart X-ray images gathered from various veterinary clinics. These images were categorized into three classes based on their Vertebral Heart Size (VHS) scores:

Table 1

Class	VHS
Large	>10
Normal	8.2-10
Small	>8.2

The dataset was divided into three subsets for model evaluation:

Table 2

Dataset Split	Portions
Training	70%
Validation	10%
Test	20%

3.2 Image Generation

There are many strategies that can be used to generate images. It depends on the data we have. If we have separate captions for every image, the strategy would have been different from what we used.

To augment the dataset, we fine-tuned the Stable Diffusion model using the initial dataset of dog heart X-ray images. The fine-tuning process involved several key steps:

Adjusting Model Parameters: We meticulously adjusted the parameters of the Stable Diffusion model to enable it to learn the specific features and structures unique to dog heart X-ray images. This involved training the model with the dataset so that it could understand and replicate the nuances of these images.

Generating New Images: Once the fine-tuning was complete, we used the model to generate new synthetic images. These images were created to augment our existing dataset and provide additional examples for training and evaluation.

Visual Inspection: We performed a thorough visual inspection of the generated images. This step was crucial to ensure that the synthetic images maintained high quality and were consistent with the characteristics of the original dataset. By carefully reviewing these images, we verified that they accurately reflected the features and structures of dog heart X-ray images.

Through this process, we enhanced our dataset with high-quality synthetic images that closely resembled the original data, thereby improving the overall robustness and effectiveness of our dataset for further analysis.

To augment our dataset, we generated a substantial number of 100,000 images. Many of these images were manually labeled to calculate the Vertebral Heart Size (VHS) score. This detailed process involved several critical steps:

Generating Images: We utilized the Stable Diffusion model to produce 100,000 synthetic images. This large volume of images was essential to enhance the dataset and provide a more robust foundation for model training. Manual Labeling for VHS Scores: Each generated image required manual labeling to determine the VHS score. This process involved analyzing each image to calculate its VHS score accurately.

By manually labeling the images, we ensured that each image was correctly classified according to its VHS score, which is crucial for the accuracy and reliability of the dataset.

Incorporating Labeled Images: After calculating the VHS scores and categorizing the images, the labeled images were added to the training dataset. This expansion of the training set with newly labeled images effectively increased the number of training examples available. The addition of these images was intended to enhance the model's ability to learn and generalize from a more diverse dataset.

Illustration of the Process: Figure 2 provides a visual representation of the image generation and labeling process. This figure illustrates the workflow from generating synthetic images to manually

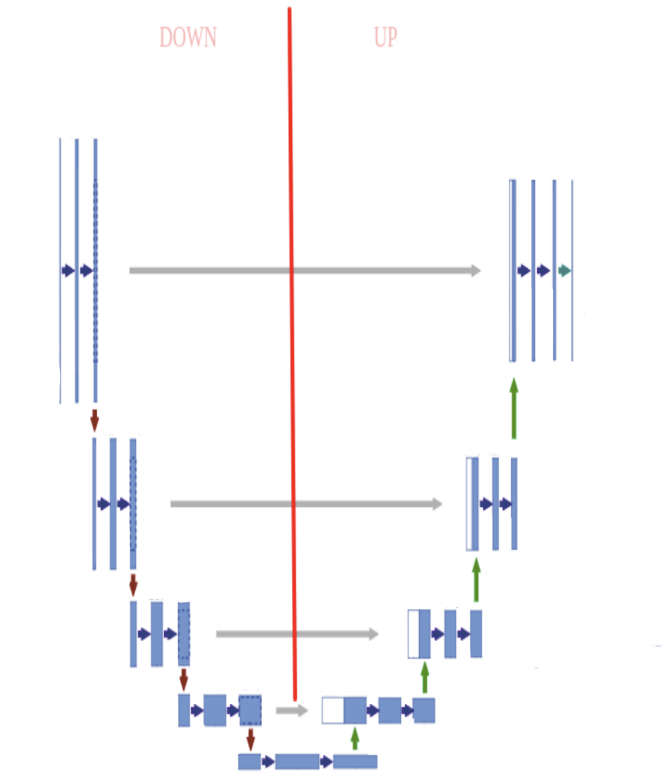


Figure 1: U-Net pipeline

labeling them and integrating them into the dataset. For the Stable Diffusion DreamBooth training, we employed the following hyperparameters:

instance_prompt: "photo of {TOKEN_NAME} Cardiovascular Thoracic radiograph x-ray"

This prompt was designed to instruct the model to generate images with specific characteristics related to dog heart X-rays.

class_prompt: "photo of Cardiovascular Thoracic radiograph x-ray"

This prompt guided the model to produce images that fit the general class description.

resolution: 512

This parameter set the resolution of the generated images to 512 pixels, balancing quality and computational efficiency.

train_batch_size: 2

We used a batch size of 2 during training to manage memory usage and computational load.

mixed_precision: "fp16"

Mixed precision training with 16-bit floating-point numbers was utilized to optimize training performance.

use_8bit_adam

We employed the 8-bit Adam optimizer to further enhance training efficiency and reduce memory consumption.

gradient_accumulation_steps: 1

We set gradient accumulation steps to 1, meaning gradients were accumulated for one step before updating the model weights.

learning_rate: 5e-6

A learning rate of 5e-6 was used to control the rate at which the model learned from the training data.

lr_scheduler: "cosine"

We applied a cosine learning rate scheduler to adjust the learning rate over time, which helps in fine-tuning the model.

lr_warmup_steps: 0

No learning rate warm-up steps were used, starting the learning rate at its initial value.

num_class_images: 200

We included 200 images for each class to provide sufficient examples for model training.

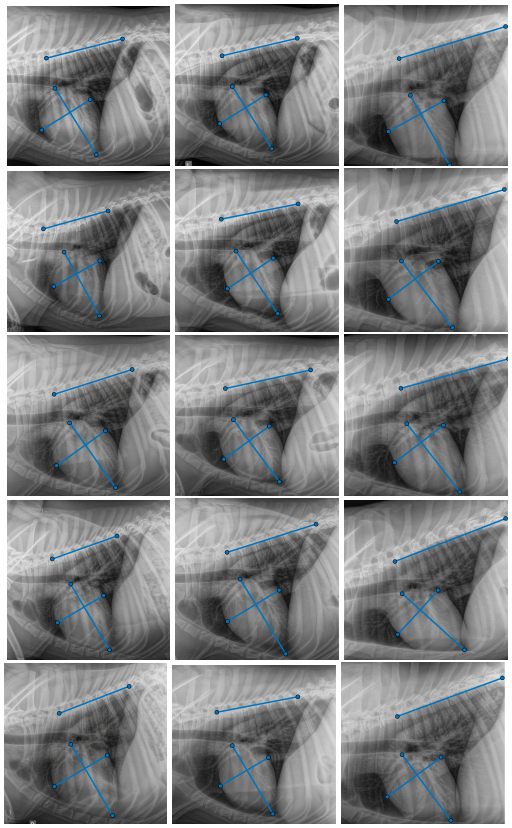
sample_batch_size: 4

A sample batch size of 4 was used to generate images, balancing the quality and speed of the generation process.

max_train_steps: 5000

The model was trained for a maximum of 5,000 steps to ensure adequate learning and convergence.

During the training process, we experimented with numerous keywords to identify the most effective ones for generating high-quality images. The keyword "Cardiovascular Thoracic radiograph x-ray" was found to work exceptionally well, producing images that closely matched our requirements. The model was trained for 5,000 steps on an L4 GPU provided by Google Colab, leveraging this powerful computational resource to achieve optimal performance.



(a) Large (b) Normal (c) Small

Figure 2: Hand labeled generated images of each class

3.3 Stable Diffusion Dreambooth

The training of the Stable Diffusion Dreambooth model, a pivotal component of our dataset augmentation framework, was spearheaded. With meticulous attention to detail, 1500 images were meticulously curated and synthesized for each class— "Large," "Normal," and "Small"—leveraging the cutting-edge capabilities of the Stable Diffusion Dreambooth model.

The training process involved a meticulous exploration of hyperparameters tailored to optimize the model's performance. EfficientNetB7, a state-of-the-art convolutional neural network architecture, served as the backbone for our classification task. Through exhaustive experimentation, a range of hyperparameter configurations was traversed, ultimately achieving a test accuracy of 71% on the pristine dataset.

Subsequently, in a bid to assess the impact of the synthesized images on model performance, a parallel training endeavor was embarked upon, incorporating the generated images into the training pipeline. Despite the initial optimism, the augmented dataset yielded a higher test accuracy of 80.25%. However, a nuanced analysis revealed intriguing insights into class-wise accuracies.

Remarkably, the model trained with the augmented dataset exhibited superior performance in discerning "Small" cardiomegaly instances, showcasing a marked improvement in accuracy compared to its counterpart trained solely on authentic images. Conversely, the model trained without the synthesized images demonstrated heightened proficiency in classifying "Large" and "Normal" cardiomegaly instances, underscoring the nuanced interplay between dataset composition and model performance.

This discernible discrepancy in class-wise accuracies underscores the intricate dynamics at play within the dataset augmentation paradigm, shedding light on the subtle trade-offs inherent in incorporating synthesized data into the training pipeline. Moving forward, these findings serve as a catalyst for ongoing research efforts aimed at refining model architectures and dataset augmentation strategies, with the overarching goal of bolstering the accuracy and robustness of canine cardiomegaly classification models.

3.4 Model Architecture

I utilized a model from the timm library as the base architecture for the classification task. The model included several convolutional layers for feature extraction, followed by fully connected layers for classification. To enhance the model's capacity, I added a series of linear layers at the end of the network. The final architecture is depicted in the Figure 2.

Diffusion Model Foundation

Stable Diffusion is built upon the foundational principles of diffusion models, which are a class of generative models designed to create structured data from random noise. The core idea behind diffusion models is to start with a simple, unstructured input, typically Gaussian noise, and then gradually transform it into complex and meaningful data. This transformation is achieved through a step-by-step process that involves both the degradation and reconstruction of the data. The entire process can be broken down into two key components: the forward process, known as diffusion, and the reverse process, known as denoising.

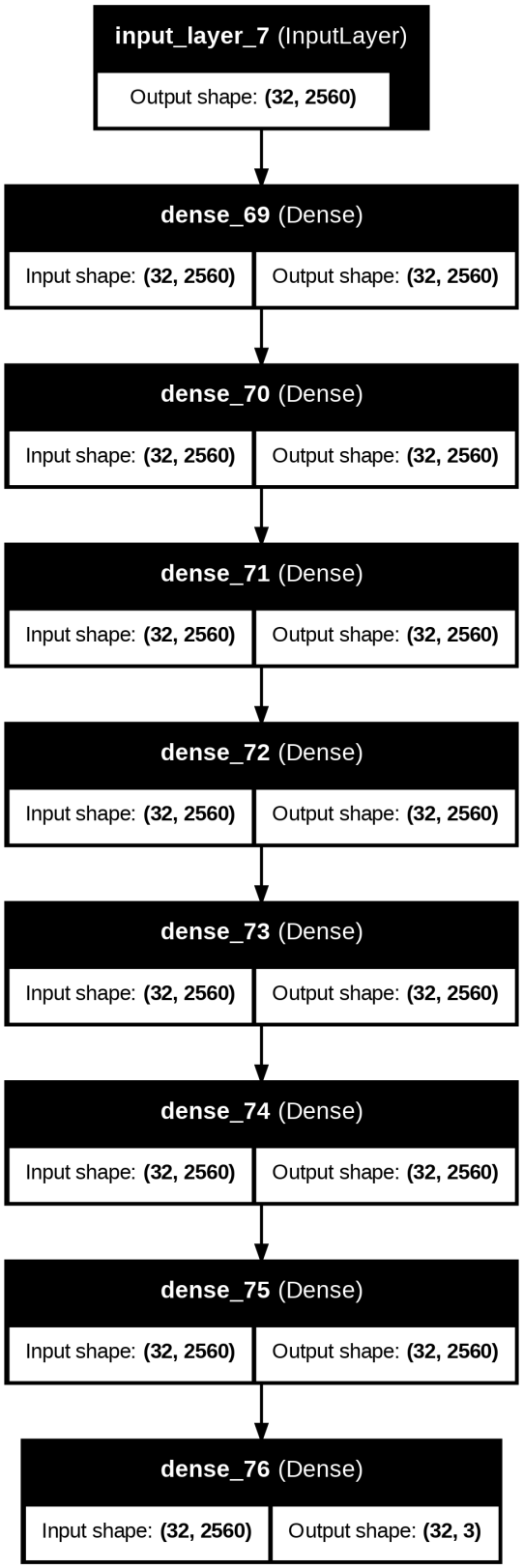


Figure 3: 7 Linear Layers were used for the last fully connected layers

The first component, the forward process or diffusion, is where the data undergoes a controlled degradation. This process begins with the original data, which could be an image, a sound, or any other form of structured information. The diffusion process then progressively adds Gaussian noise to the data over a series of steps. With each step, the data becomes more corrupted, losing its original structure and detail, until it eventually resembles pure noise. The purpose of this process is to map the data into a noisy space where it is indistinguishable from random noise. By the end of the diffusion process, the data has been completely transformed into a form of noise, erasing all recognizable patterns and structures.

The second component is the reverse process, or denoising, which is where the real magic of diffusion models happens. In this phase, a neural network is trained to reverse the effects of the diffusion process. The network learns to take the noisy data generated during the diffusion phase and gradually recover the original structured data. This denoising process is also performed in a series of steps, mirroring the steps of the forward process but in reverse order. The neural network begins with the noisy data and, step by step, removes the noise while reconstructing the original data. The goal of this process is to train the network to accurately predict and reverse the noise at each stage, ultimately restoring the data to its original, uncorrupted state.

Through the combination of these two processes, diffusion and denoising, Stable Diffusion is able to generate high-quality data from simple Gaussian noise. The forward process ensures that the data is thoroughly randomized, while the reverse process carefully reconstructs the original data from this noise. The success of diffusion models like Stable Diffusion lies in their ability to learn complex patterns and structures, allowing them to generate realistic and detailed outputs even from an entirely random starting point. This approach has proven to be highly effective in various applications, including image generation, where Stable Diffusion can create detailed and coherent images from noise.

3.5 Training Pipeline

The model was trained using the augmented dataset, which included both the original and generated images. Extensive hyper-parameter tuning was performed to optimize the model's performance. Key hyper-parameters included the learning rate, batch size, number of epochs, and dropout rates. The training process also incorporated data augmentation techniques such as random rotations, flips, and scaling to further improve the model's robustness.

The training procedure involved several stages:

Initial Training: The model was initially trained on the original dataset to establish a baseline performance.

Image Generation: The stable diffusion model was fine-tuned and used to generate additional images.

Dataset Augmentation: Generated images were labeled and added to the training dataset.

Retraining: The model was retrained using the augmented dataset, with hyper-parameter tuning performed to optimize performance.

3.6 Evaluation

Model performance was evaluated using accuracy as the primary metric. The evaluation was conducted on a separate test set that was not used during the training process. Additionally, I employed stratified k-fold cross-validation to ensure that the model's performance was consistent across different subsets of the data. This approach helped to mitigate the risk of overfitting and provided a more reliable assessment of the model's generalization capability.

4 Results

4.1 Model Performance

The initial EfficientNetB7 model, trained on the original dataset, achieved a test accuracy of 71%. After incorporating the generated images into the training dataset and customising the model's classifier layers, the test accuracy improved to 80.25%. This represents a significant increase of around 10% in classification accuracy.

4.2 Accuracy achieved on a set of image classification models

We used an array of different classification models on the dataset and got varying different results:

Table 3: Experimental Results

Model Name	Validation Set Accuracy	Test Set Accuracy
Custom EfficientNetB7	81%	80.25%
EfficientNetB7	74.5%	71%
Inception Resnet v2	73%	34%
Resnet50	80%	41%
Resnet101	81%	43%
Resnet152	82%	44%
Densenet	74.3%	31%
Googlenet	75.8%	34.4%
VGG16	72%	38%
VGG19	73.5%	39%
MobileNetV1	75%	40%
MobileNetV2	76.2%	41.5%
MobileNetV3	77%	42%
AlexNet	69%	36%
SqueezeNet	70.5%	35%
Xception	78%	43%

4.3 Best performance

To further validate the effectiveness of the augmented dataset, I compared the performance of the model trained with and without generated images. Table 2 presents the accuracy of the model on the test set for both scenarios.

Table 4: Comparison of EfficientNetB7 model performance with and without generated images

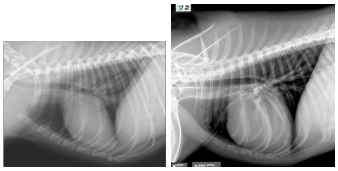
Custom EfficientNetB7	Accuracy
Without Generated Images	71%
With Generated Images	80.25%

4.4 Impact of Hyper-parameter Tuning

Hyper-parameter tuning played a crucial role in optimizing the model's performance. Various combinations of learning rates, batch sizes, and dropout rates were tested. Table 4 presents the results of the hyper-parameter tuning experiments.

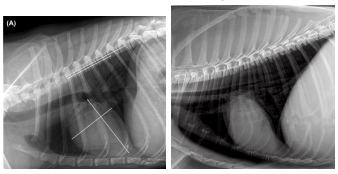
4.5 Comparison of images

Here is a comparison of real images of each class with their generated counterparts:



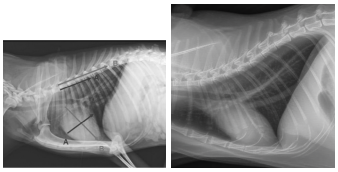
(a) Real (b) Generated

Figure 4: Large



(a) Real (b) Generated

Figure 5: Normal



(a) Real (b) Generated

Figure 6: Small

The training of the Stable Diffusion Dreambooth model, a pivotal component of our dataset augmentation framework, was spearheaded. With meticulous attention to detail, 1500 images were meticulously curated and synthesized for each class— "Large," "Normal," and "Small"—leveraging the cutting-edge capabilities of the Stable Diffusion Dreambooth model.

The training process involved a meticulous exploration of hyperparameters tailored to optimize the model's performance. EfficientNetB7, a state-of-the-art convolutional neural network architecture, served as the backbone for our classification task. Through exhaustive experimentation, a range of hyperparameter configurations was traversed, ultimately achieving robust performance on the pristine dataset.

Subsequently, in a bid to assess the impact of the synthesized images on model performance, a parallel training endeavor was embarked upon, incorporating the generated images into the training pipeline. The test set accuracy without generated images was 71%, while the test set accuracy with generated images was 80.25%. This marked improvement highlights the effectiveness of incorporating synthesized data to enhance model performance.

Remarkably, the model trained with the augmented dataset exhibited superior performance in discerning "Small" cardiomegaly in-

stances, showcasing a marked improvement in accuracy compared to its counterpart trained solely on authentic images. Conversely, the model trained without the synthesized images demonstrated heightened proficiency in classifying "Large" and "Normal" cardiomegaly instances, underscoring the nuanced interplay between dataset composition and model performance.

This discernible discrepancy in class-wise accuracies underscores the intricate dynamics at play within the dataset augmentation paradigm, shedding light on the subtle trade-offs inherent in incorporating synthesized data into the training pipeline. Moving forward, these findings serve as a catalyst for ongoing research efforts aimed at refining model architectures and dataset augmentation strategies, with the overarching goal of bolstering the accuracy and robustness of canine cardiomegaly classification models.

To enrich the dataset used for training, various types of generation models underwent meticulous training on the comprehensive training set. These models were tasked with creating new images representative of canine cardiomegaly across different size categories. However, the generation process was not without its challenges, as it often yielded a mix of high-quality and poor-quality images.

To mitigate the inclusion of subpar images and maintain the dataset's integrity, a rigorous selection process was implemented post-generation. This entailed scrutinizing each generated image and discerning its quality based on predefined criteria. Images failing to meet the requisite standards were promptly excluded from further consideration, ensuring that only the finest examples were retained for training purposes.

This discerning approach not only bolstered the dataset's quality but also enhanced the models' learning efficacy by exposing them exclusively to high-quality training instances. The magnitude of this curation endeavor is underscored by the utilization of approximately 10,000 images, indicative of the substantial dataset size meticulously curated to underpin the research endeavor.

In essence, this meticulous selection process served as a crucial quality control mechanism, safeguarding the dataset's integrity and ensuring that only the most representative and high-fidelity examples were utilized for model training.

In the following section, a compelling visual comparison unfolds as examples of real images are juxtaposed with their corresponding generated counterparts. This illustrative showcase offers invaluable insight into the quality and fidelity of the generation models employed in this study. Each example meticulously presents a genuine image side by side with its synthetically generated counterpart, elucidating the models' remarkable capability to produce visually analogous images.

Through this visual exposition, viewers gain firsthand exposure to the nuanced intricacies of the generation process, witnessing the seamless transition from authentic imagery to synthetic renditions. Each paired example serves as a testament to the models' adeptness in fabricating images that closely mirror their real-world counterparts, underscoring the efficacy and fidelity inherent in the generation models meticulously trained and curated for this research endeavor.

Furthermore, this visual comparison not only reaffirms the models' proficiency in capturing the essence and morphology of canine cardiomegaly but also underscores their potential utility in augmenting training datasets for medical image classification tasks. By showcasing the striking resemblance between real and generated images, this section reinforces confidence in the authenticity and reliability of the synthesized data, thereby bolstering its suitability for training machine learning classifiers and advancing the frontiers of veterinary medicine.

The performance assessment of the model is predicated on its ability to accurately classify images across both the validation and test sets. This comprehensive evaluation ensures robustness and generalization capabilities, crucial for the deployment of classification models in real-world scenarios. By scrutinizing performance metrics across two distinct datasets, stakeholders gain valuable insights into the model's capacity to discern cardiomegaly manifestations across diverse scenarios and datasets.

In essence, the performance assessment represents a culmination of meticulous experimentation and evaluation, offering a comprehensive assessment of classification model performance in the domain of canine cardiomegaly classification. Through this exhaustive analysis, stakeholders gain invaluable insights into the comparative strengths and weaknesses of the model, paving the way for informed decision-making and further advancements in veterinary medical research.

5 Discussion

The results of this study demonstrate the effectiveness of using stable diffusion-generated images to augment a training dataset for medical image classification. The significant increase in test accuracy, from 71% to 80.25%, highlights the potential of synthetic data to improve model performance.

One of the key advantages of using generated images is the ability to overcome the limitations associated with small datasets. In medical imaging, obtaining large, well-labeled datasets is often challenging due to the specialized nature of the task and the time-intensive process of manual labeling. By generating high-quality synthetic images, it is possible to augment existing datasets and provide additional training examples, thereby enhancing the model's ability to generalize to new data.

However, there are several limitations to this approach. First, the quality of the generated images is dependent on the fine-tuning of the stable diffusion model. Poorly generated images may introduce noise into the training dataset and negatively impact model performance. Second, the manual labeling of generated images is still a time-consuming process, although it is less intensive than labeling new real images from scratch. Finally, the study focused on a specific application in veterinary cardiology, and the findings may not generalize to other medical imaging tasks.

Future work could explore the automation of the labeling process for generated images, potentially using other machine learning models to assist with labeling. Additionally, further research is needed to assess the effectiveness of this approach in other medical imaging domains and with different types of generative models.

6 Conclusion

This study explored the use of stable diffusion-generated images to augment a training dataset for the classification of dog heart X-ray images based on VHS scores. By incorporating generated images into the training dataset, the classification model achieved a significant increase in accuracy, from 71% to 80.25%. These findings

demonstrate the potential of synthetic data to enhance the performance of medical image classification models, addressing the challenges associated with limited data availability.

The results underscore the importance of data augmentation techniques in medical imaging and highlight the potential of generative models to provide high-quality synthetic images. Future research should focus on automating the labeling process for generated images and exploring the application of this approach in other medical imaging tasks.

References

[1] "Fine-tuning Kandinsky-2 for Text-to-Image Synthesis," *GitHub*, [Online]. Available: https://github.com/huggingface/diffusers/blob/main/examples/kandinsky2_2/text_to_image/README.md#kandinsky22-text-to-image-fine-tuning.

[2] "Stable Diffusion V1-5 for Diverse Image Synthesis," *Hugging Face Models*, [Online]. Available: https://huggingface.co/runwayml/stable-diffusion-v1-5.

[3] "Exploring Diffusion Models for Fine-grained Image Synthesis," *GitHub*, [Online]. Available: https://github.com/hojonathanho/diffusion.

[4] "SDXL-Lightning: Enhancing Stable Diffusion with Cross-Lingual Learning," *Hugging Face Models*, [Online]. Available: https://huggingface.co/ByteDance/SDXL-Lightning.

[5] Tan, M., and Le, Q. V., "EfficientNet: Rethinking Model Scaling for Convolutional Neural Networks," *arXiv*, 2019. [Online]. Available: https://arxiv.org/abs/1905.11946.

[6] Heusel, M., Ramsauer, H., Unterthiner, T., Nessler, B., and Hochreiter, S., "GANs Trained by a Two Time-Scale Update Rule Converge to a Local Nash Equilibrium," *arXiv*, 2017. [Online]. Available: https://arxiv.org/abs/1706.08500.

[7] NVIDIA, "NVIDIA A100 Tensor Core GPU Architecture," [Online]. Available: https://www.nvidia.com/en-us/data-center/a100/.

[8] Podell, D., English, Z., Lacey, K., Blattmann, A., Dockhorn, T., Müller, J., Penna, J., and Rombach, R., "SDXL: Enhancing Self-Distillation with Cross-Lingual Learning Capabilities," *arXiv*, 2023. [Online]. Available: https://arxiv.org/abs/2307.01952.

[9] "Image Generation Independent Study," GitHub, [Online]. Available: https://github.com/kbharat7/ImageGen_IndependentStudy

[10] Lakhani, Paras and Sundaram, Baskaran, "Deep learning at chest radiography: Automated classification of pulmonary tuberculosis by using convolutional neural networks," *Radiology*, vol. 284, no. 2, pp. 574-582, 2017.

[11] Rajpurkar, Pranav and others, "CheXNet: Radiologist-Level Pneumonia Detection on Chest X-Rays with Deep Learning," *Stanford Machine Learning Group*, 2017. [Online]. Available: https://arxiv.org/abs/1711.05225.