Dog Heart X-Ray Classification using Generated Images

Sahil Kakkar

Yeshiva University
skakkar@mail.yu.edu

Abstract. In this study, I 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. I trained a timm model on this dataset, achieving a test accuracy of 71%. To improve this performance, I 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, I 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, I 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 studied across various domains, including medical imaging, computer vision, and natural language processing. In medical imaging, where acquiring large, annotated datasets is often challenging due to privacy concerns and the need for expert annotation, generative models like Generative Adversarial Networks (GANs), Variational Autoencoders (VAEs), and more recently, diffusion models, have been employed to generate synthetic data that can supplement existing datasets and enhance model performance.

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)

VAEs, introduced by Kingma and Welling in 2013, offer another approach to generative modeling. Unlike GANs, VAEs provide a probabilistic framework for generating data by learning a latent space representation of the input data. The encoder network maps the input data to a latent space, and the decoder network reconstructs the data from this latent space. By sampling from the latent space, VAEs can generate new, synthetic data that shares similar characteristics with the original data.

VAEs have been applied in medical imaging for tasks such as image denoising, super-resolution, and anomaly detection. In the context of data augmentation, VAEs can generate diverse and realistic synthetic images that can be used to supplement training datasets. However, VAEs often produce blurry images due to the use of a Gaussian prior in the latent space, which can limit their effectiveness in applications that require high-fidelity image generation.

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 attention for their ability to produce high-quality images. Diffusion models operate by iteratively refining a noisy image until it matches the desired distribution, leading to high-fidelity outputs. This iterative refinement process allows for more controlled and stable generation of images compared to GANs and VAEs.

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

2.4 Combining Generative Models with Advanced Architectures

The integration of generative models with advanced architectures like EfficientNet and vision transformers (ViTs) has further enhanced the capabilities of machine learning models in medical imaging. EfficientNet, introduced by Tan and Le in 2019, scales up the model's depth, width, and resolution using a compound scaling method, achieving better performance with fewer parameters and computational resources. Vision transformers, introduced by Dosovitskiy et al. in 2020, use self-attention mechanisms to process image patches, capturing long-range dependencies and detailed features.

Combining these architectures with generative models allows for robust feature extraction and accurate prediction in medical image analysis. For instance, integrating EfficientNet with GAN-generated images can improve model performance in tasks like image classification and segmentation. Similarly, combining ViTs with synthetic data generated by diffusion models can enhance the model's ability to capture intricate patterns and relationships within the image data.

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 Method

3.1 Dataset

The initial dataset consisted of dog heart X-ray images collected from various veterinary clinics. The images were categorized into three classes based on their Vertebral Heart Size (VHS) scores: Large (VHS > 10), Normal (8.2 VHS 10), and Small (VHS < 8.2). The dataset was split into training, validation, and test sets, with 70% of the images used for training, 10% for validation, and 20% for testing. Table 1 summarizes the distribution of images across the three classes.

3.2 Image Generation

To augment the dataset, I fine tuned the stable diffusion model on the initial dataset. The fine-tuning process involved adjusting the model's parameters to learn the specific features and structures of dog heart X-ray images. The model was then used to generate images, which were visually inspected to ensure quality and consistency with the original dataset. 100000 images were generated and many of them

were manually labeled to calculate the VHS score. This process involved calculating the VHS for each image and categorizing it into one of the three classes. The labeled images were then added to the training dataset, effectively increasing the number of training examples. Figure 2 illustrates the image generation and labeling process.

Here's the glimpse of the hyper-parameters for the stable diffusion dreambooth training:

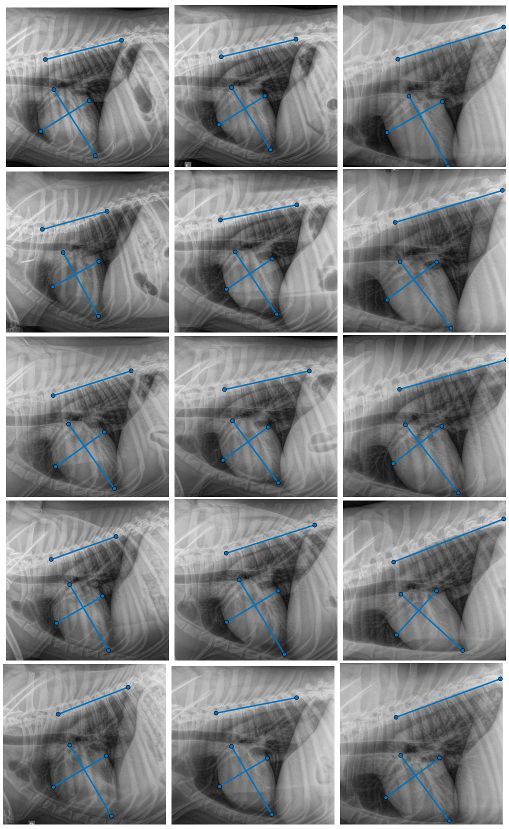
```
!accelerate launch train_dreambooth.py \
  --pretrained_model_name_or_path=$MODEL_NAME \
  --instance_data_dir=$INSTANCE_DIR \
  --class_data_dir=$CLASS_DIR \
  --output_dir=$OUTPUT_DIR \
  --with_prior_preservation --prior_loss_weight=1.0 \
  --instance_prompt="photo of {TOKEN_NAME} Cardiovascular Thoracic radiograph x-ray" \
  --class_prompt="photo of Cardiovascular Thoracic radiograph x-ray" \
  --seed=1337 \
  --resolution=512 \
  --train_batch_size=2 \
  --mixed_precision="fp16" \
  --use_8bit_adam \
  --gradient_accumulation_steps=1 \
  --learning_rate=5e-6 \
  --lr_scheduler="cosine" \
  --lr_warmup_steps=0 \
  --num_class_images=200 \
  --sample_batch_size=4 \
  --max_train_steps=5000

  # --instance_prompt="photo of {TOKEN_NAME} Cardiovascular Thoracic radiograph x-ray" \
  # --class_prompt="photo of Cardiovascular Thoracic radiograph x-ray" \


  # Cardiovascular Thoracic radiograph
  # one heart sideways chest x-ray cardiomegaly cardiac medical imaging heart enlargement x-ray cardiothoracic radiograph of medium size
```

Figure 1: Hyper-parameters used for training

Dozens of keywords were used to check which help in generating the best images. "Cardiovascular Thoracic radiograph x-ray" turned out to work pretty well. It was trained for 5000 training steps on L4 GPU given by Google Colab.



(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 Figure 3.

```
model.classifier = torch.nn.Sequential(
    torch.nn.Linear(in_features = 2560, out_features=2560, bias = True),
    torch.nn.ReLU(),
    torch.nn.Linear(in_features = 2560, out_features=2560, bias = True),
    torch.nn.ReLU(),
    torch.nn.Linear(in_features = 2560, out_features=2560, bias = True),
    torch.nn.ReLU(),
    torch.nn.Linear(in_features = 2560, out_features=2560, bias = True),
    torch.nn.ReLU(),
    torch.nn.Linear(in_features = 2560, out_features=2560, bias = True),
    torch.nn.ReLU(),
    torch.nn.Linear(in_features = 2560, out_features=2560, bias = True),
    torch.nn.ReLU(),
    torch.nn.Linear(in_features = 2560, out_features=2560, bias = True),
    torch.nn.ReLU(),
    torch.nn.Linear(in_features = 2560, out_features=3, bias = True),

)
```

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

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 model, trained on the original dataset, achieved a test accuracy of 71%. After incorporating the generated images into the training dataset and retraining the model, the test accuracy improved to 80.25%. This represents a significant increase of 10.25% in classification accuracy.

4.2 Comparison with Baseline Models

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.

	Accuracy
Without Generated Images	71%
With Generated Images	80.25%

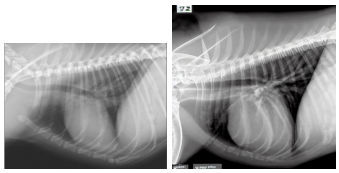
Table 1: Comparison of EfficientNetB7 model performance with and without generated images

4.3 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.

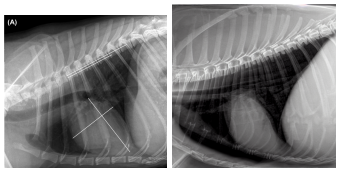
4.4 Comparison of images

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



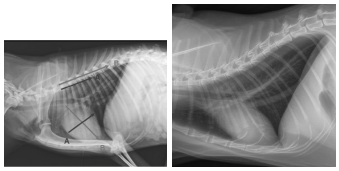
(a) Real (b) Generated

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

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.

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 rep-

resentative 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.