Dataset Distillation: A Data-Efficient Learning Framework

Swapnil Patel
University of Toronto
swap.patel@mail.utoronto.ca
https://github.com/Swapnil949/ECE1512_2024F_ProjectRepo_SwapnilPatel

***Abstract*— The entire project source can be found at: Github Repository.**

1. Introduction

2. Related Work

3. Dataset Distillation with Attention Matching

3.1. MNIST Dataset

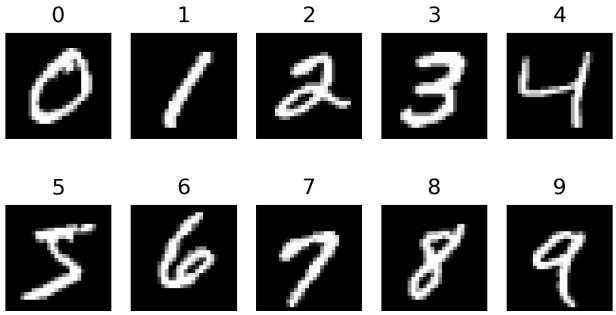

Figure 1. MNIST Dataset [1]

The MNIST dataset is a widely used collection of handwritten digits that is commonly used for training and testing machine learning and computer vision algorithms. MNIST stands for the ”Modified National Institute of Standards and Technology” database. It was created by modifying the original NIST dataset, which contained a much larger and more diverse set of handwritten characters, to focus specifically on handwritten digits.

The MNIST dataset contains 28x28-pixel grayscale images of handwritten digits (0 through 9), along with corresponding labels indicating which digit each image represents [1]. There are 60,000 training images and 10,000 testing images in the MNIST dataset, making it a popular benchmark for various image classification tasks.

3.1.1. ConvNet-3.

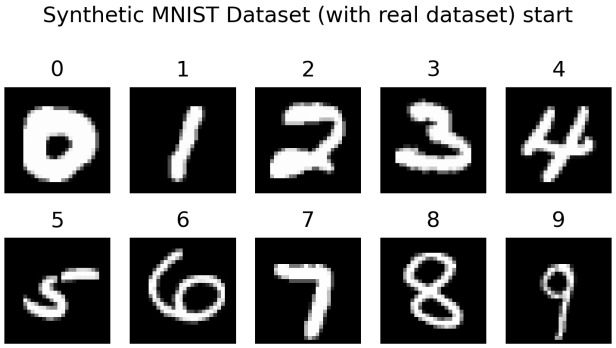

Figure 2. Sample of Synthetic MNIST Dataset created from real images (starting image)

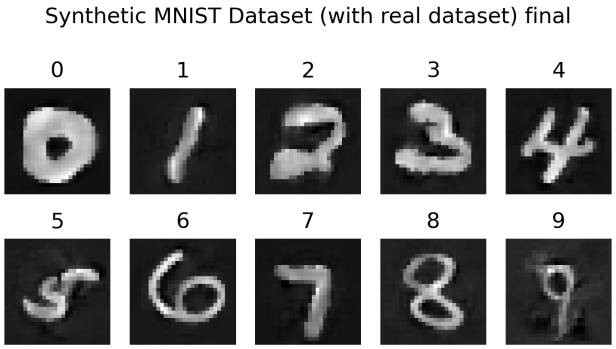

Figure 3. Sample of Synthetic MNIST Dataset created from real images (final image)

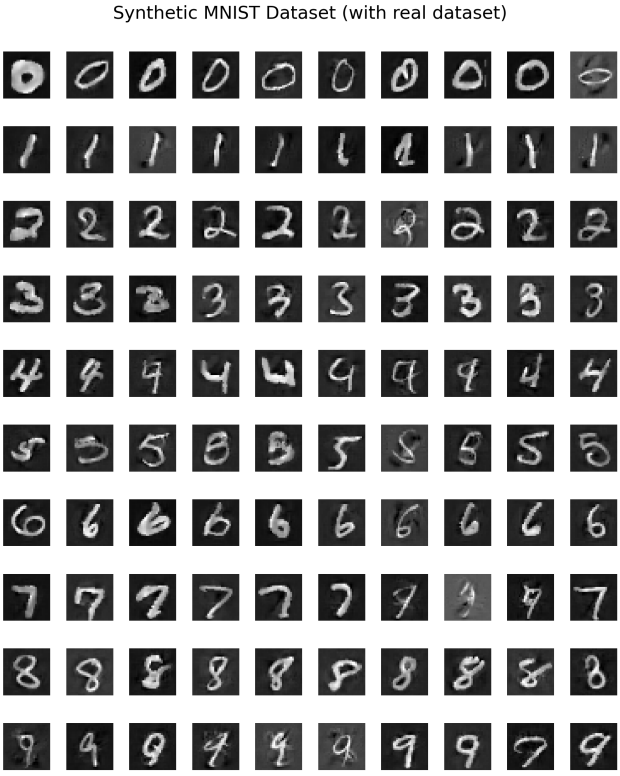

Figure 4. Synthetic MNIST Dataset created from real images

3.1.2. Synthetic Dataset using real images.

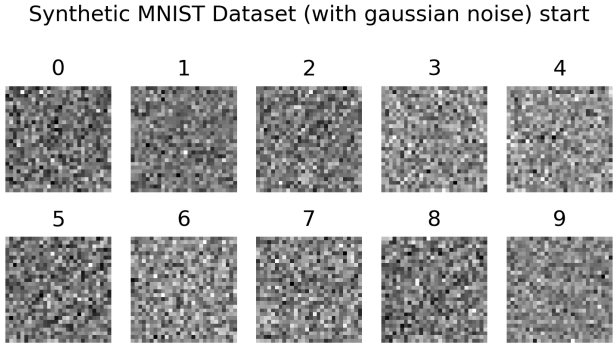

Figure 5. Sample of Synthetic MNIST Dataset created from Gaussian noise (starting image)

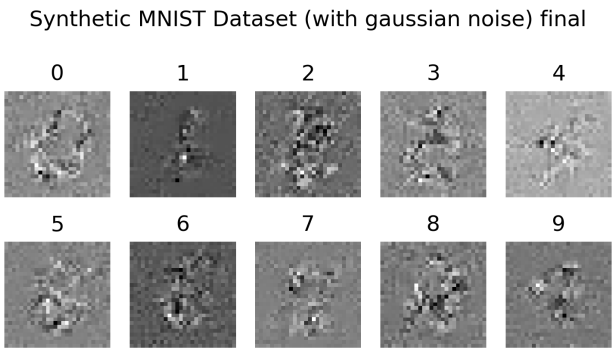

Figure 6. Sample of Synthetic MNIST Dataset created from Gaussian noise (final image)

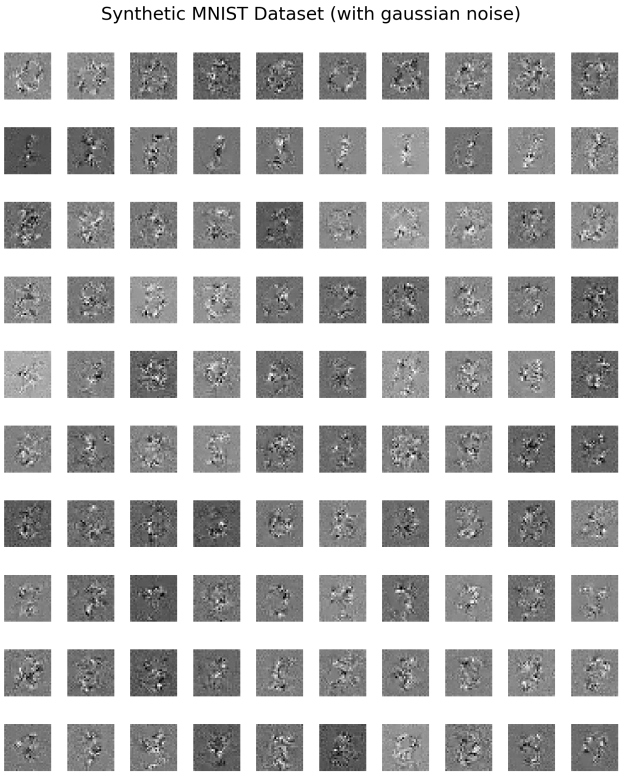

Figure 7. Synthetic MNIST Dataset created from Gaussian noise

3.1.3. Synthetic Dataset using Gaussian noise.

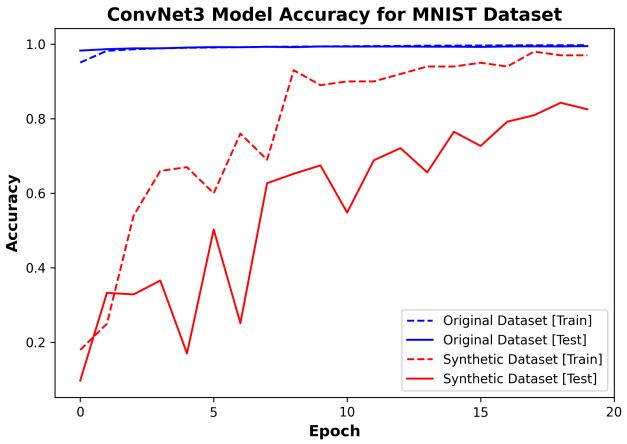

Figure 8. ConvNet-3 Model Trained using Original and Synthetic dataset

3.1.4. ConvNet-3 using Synthetic Dataset.

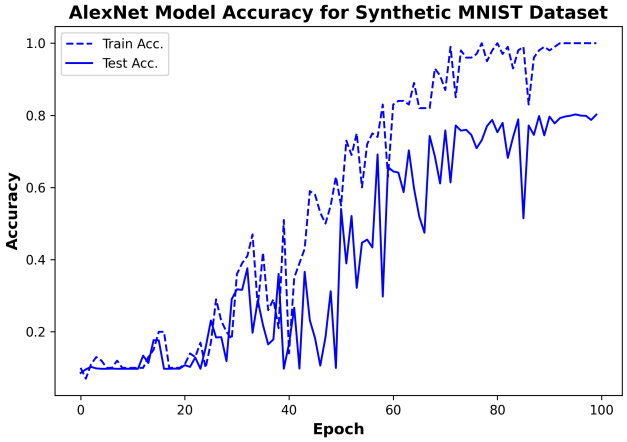

Figure 9. AlexNet Model Trained using Synthetic dataset

3.1.5. Cross-architecture Generalization - AlexNet.

3.2. MHIST Dataset

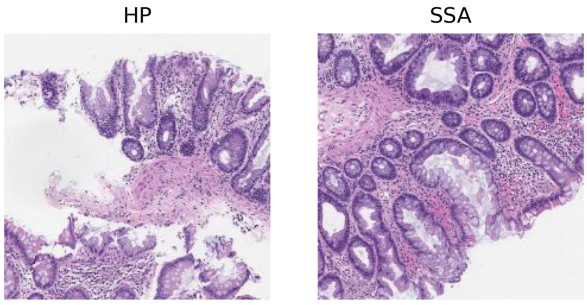

Figure 10. MHIST Dataset [5]

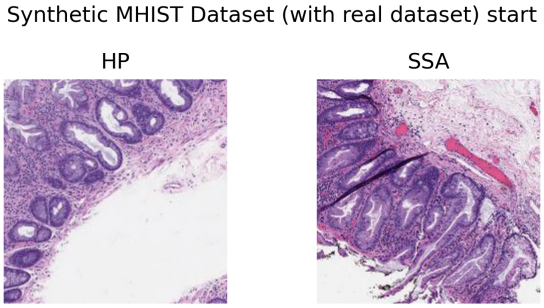

Figure 11. Sample of Synthetic MHIST Dataset created from real images (starting image)

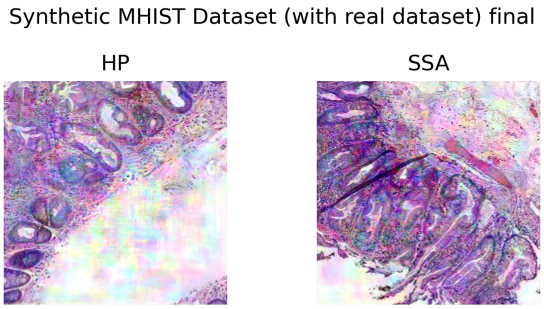

Figure 12. Sample of Synthetic MHIST Dataset created from real images (final image)

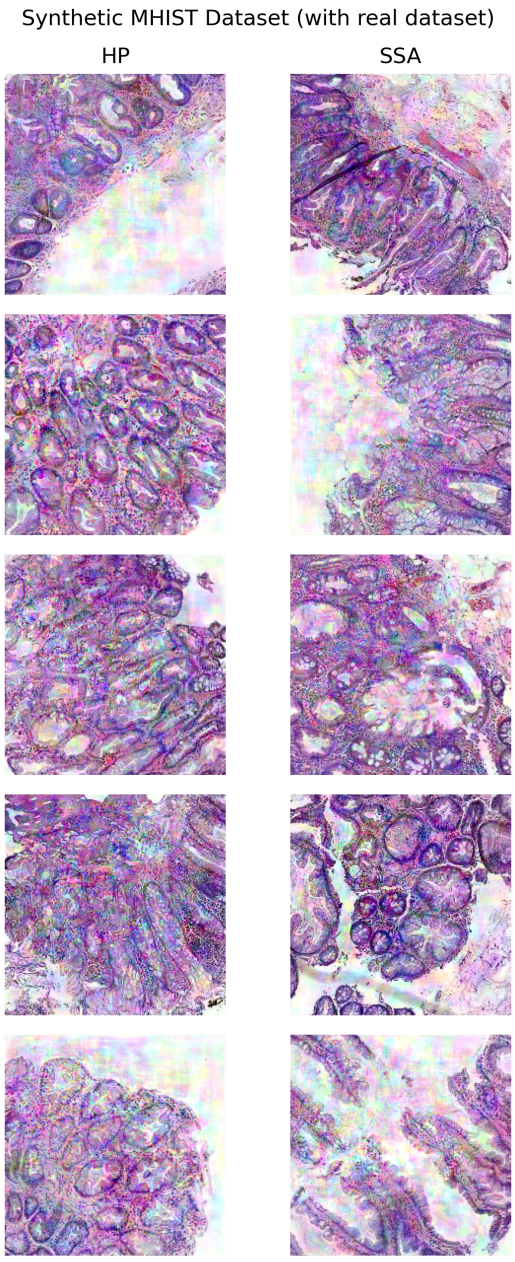

Figure 13. Synthetic MHIST Dataset created from real images

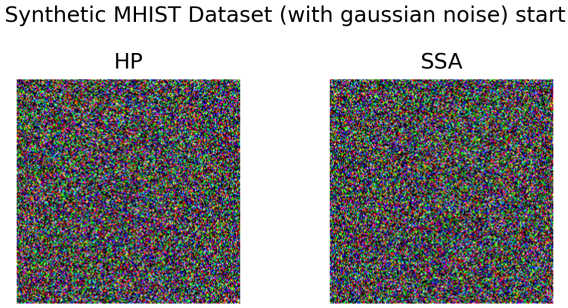

Figure 14. Sample of Synthetic MHIST Dataset created from Gaussian noise (starting image)

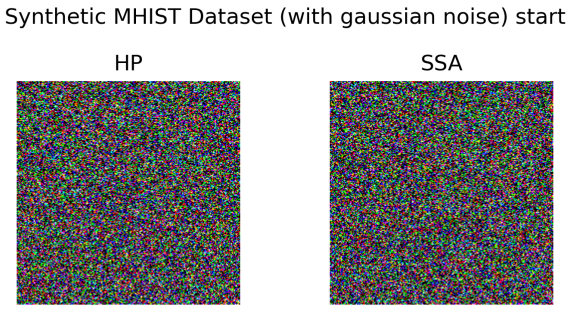

Figure 15. Sample of Synthetic MHIST Dataset created from Gaussian noise(final image)

3.2.1. Synthetic Dataset using real images.

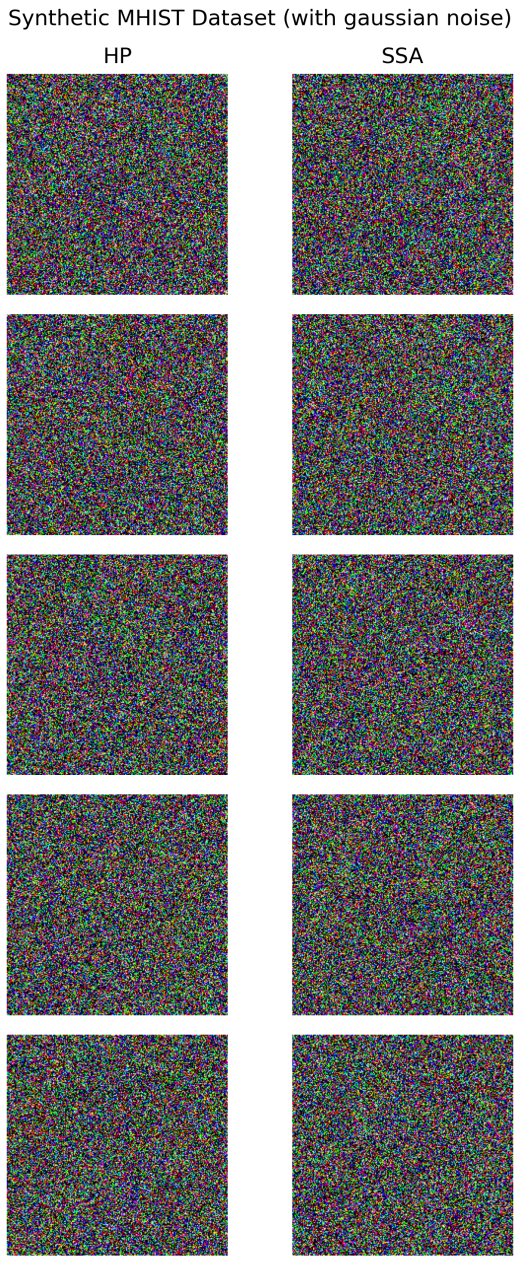

Figure 16. Synthetic MHIST Dataset created from Gaussian noise

3.2.2. Synthetic Dataset using Gaussian noise.

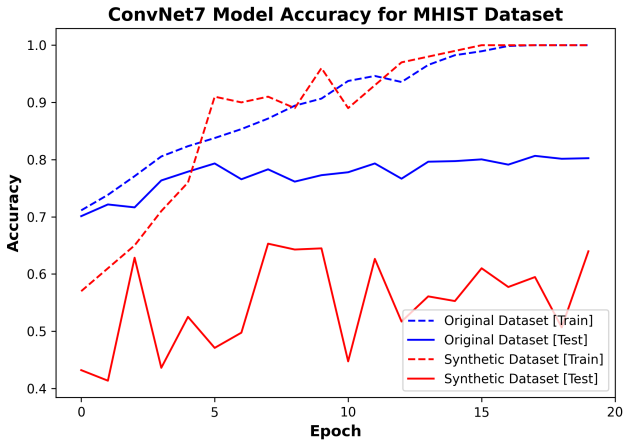

Figure 17. ConvNet-7 Model Trained using Original and Synthetic dataset

3.2.3. ConvNet-7 using Synthetic Dataset.

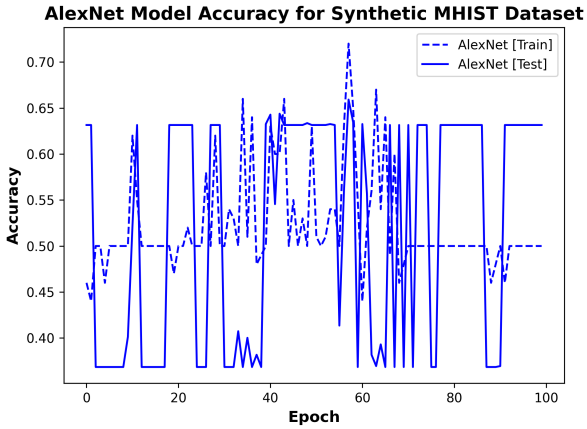

Figure 18. AlexNet Model Trained using Synthetic dataset

3.2.4. Cross-architecture Generalization - AlexNet.

3.3. Data Distillation Application

NAS with original dataset took 1607.3 seconds. with synthetic dataset took 2.8 seconds.

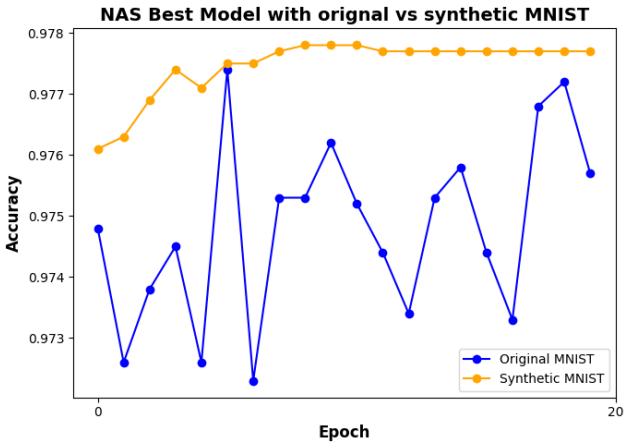

Figure 19. NAS best model with original vs synthetic MNIST dataset

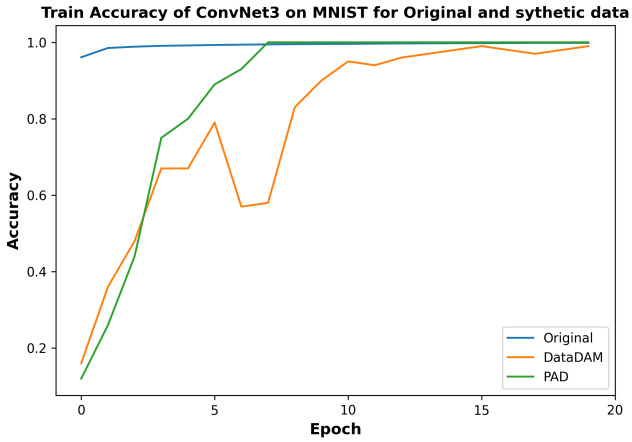

Figure 21.

4. Prioritize Alignment in Dataset Distillation

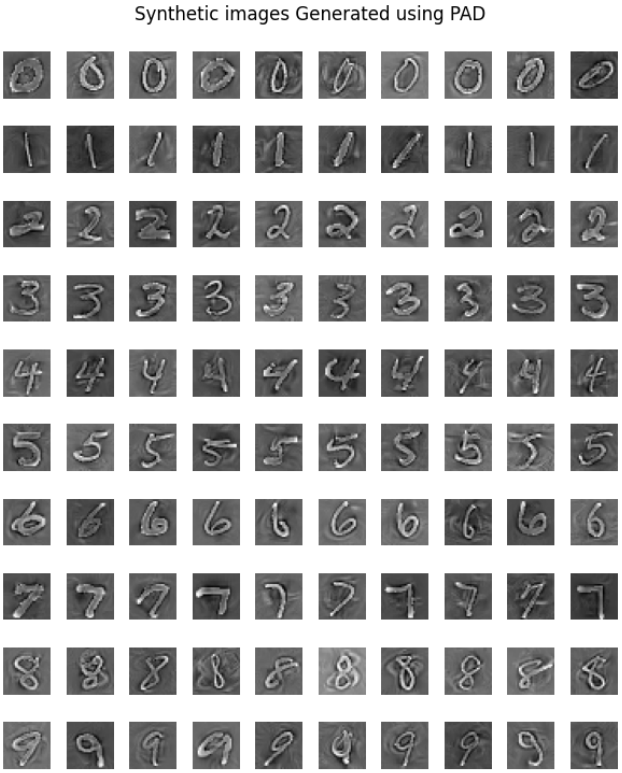

Figure 20. Synthetic MNIST Dataset using PAD[3]

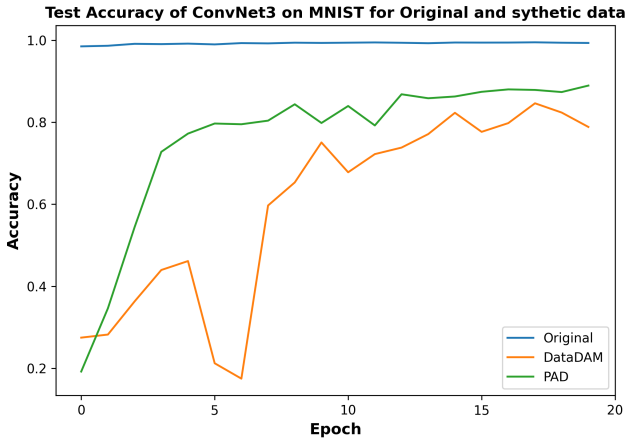

Figure 22.

5. Conclusion

6. References

References

[1] Li Deng. "The mnist database of handwritten digit images for machine learning research". In: *IEEE Signal Processing Magazine* 29.6 (2012), pp. 141–142.

[2] Geoffrey Hinton, Oriol Vinyals, and Jeff Dean. *Distilling the Knowledge in a Neural Network*. 2015. arXiv: 1503.02531 [stat.ML].

[3] Zekai Li et al. *Prioritize Alignment in Dataset Distillation*. 2024. arXiv: 2408.03360 [cs.LG]. URL: https://arxiv.org/abs/2408.03360.

[4] Ahmad Sajedi et al. *DataDAM: Efficient Dataset Distillation with Attention Matching*. 2023. arXiv: 2310.00093 [cs.CV]. URL: https://arxiv.org/abs/2310.00093.

[5] Jerry Wei et al. “A Petri Dish for Histopathology Image Analysis”. In: *International Conference on Artificial Intelligence in Medicine*. Springer. 2021, pp. 11–24.