

# 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](https://github.com/Swapnil949/ECE1512_2024F_ProjectRepo_SwapnilPatel)*

**Abstract**— The entire project source can be found at: [Github Repository](https://github.com/Swapnil949/ECE1512_2024F_ProjectRepo_SwapnilPatel).

## 1. Introduction

## 2. Related Work

## 3. Dataset Distillation with Attention Matching

### 3.1. MNIST Dataset

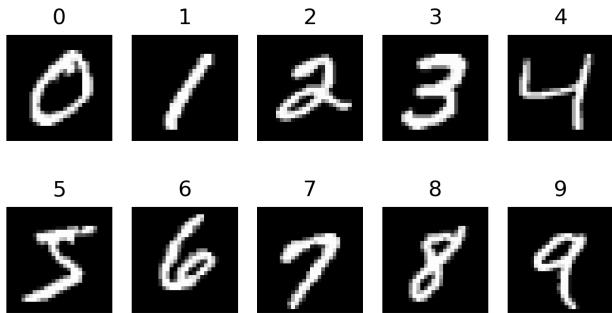


Figure 1. MNIST Dataset [1]

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.

Synthetic MNIST Dataset (with real dataset) start

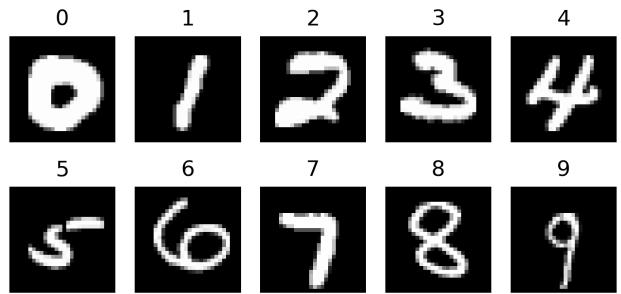


Figure 2. Sample of Synthetic MNIST Dataset created from real images (starting image)

Synthetic MNIST Dataset (with real dataset) final

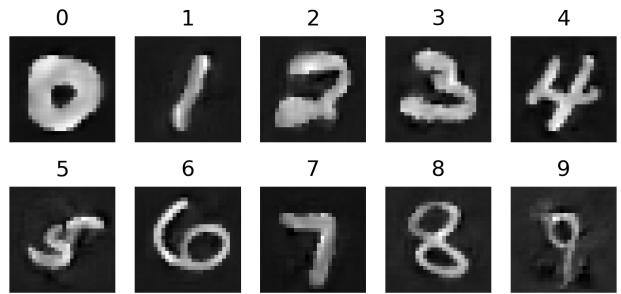


Figure 3. Sample of Synthetic MNIST Dataset created from real images (final image)

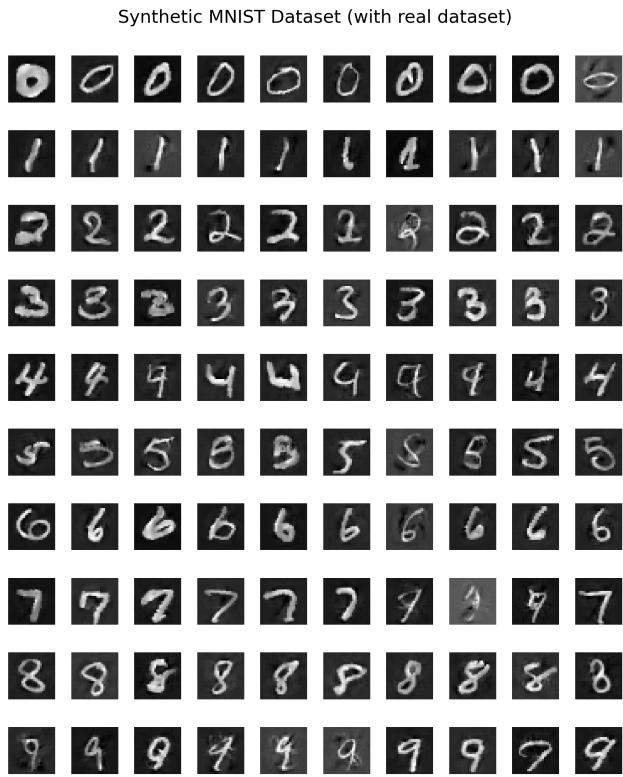


Figure 4. Synthetic MNIST Dataset created from real images

### 3.1.2. Synthetic Dataset using real images.

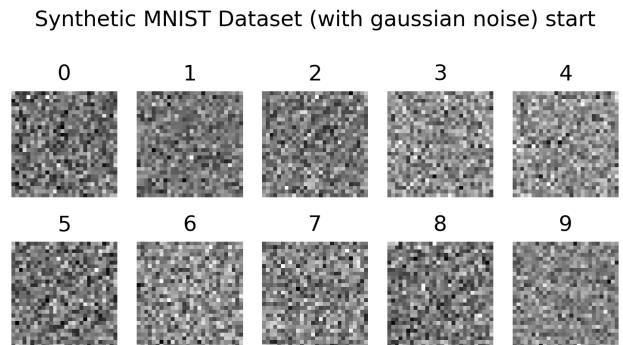


Figure 5. Sample of Synthetic MNIST Dataset created from Gaussian noise (starting image)

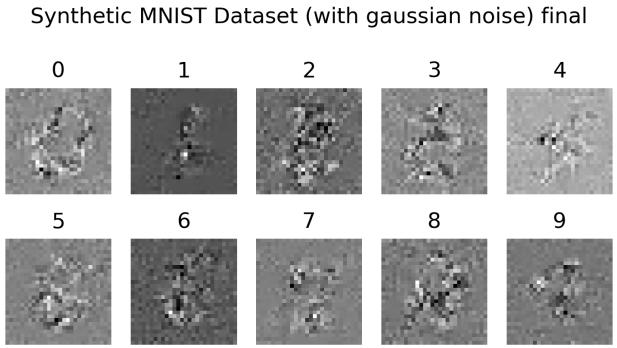


Figure 6. Sample of Synthetic MNIST Dataset created from Gaussian noise (final image)

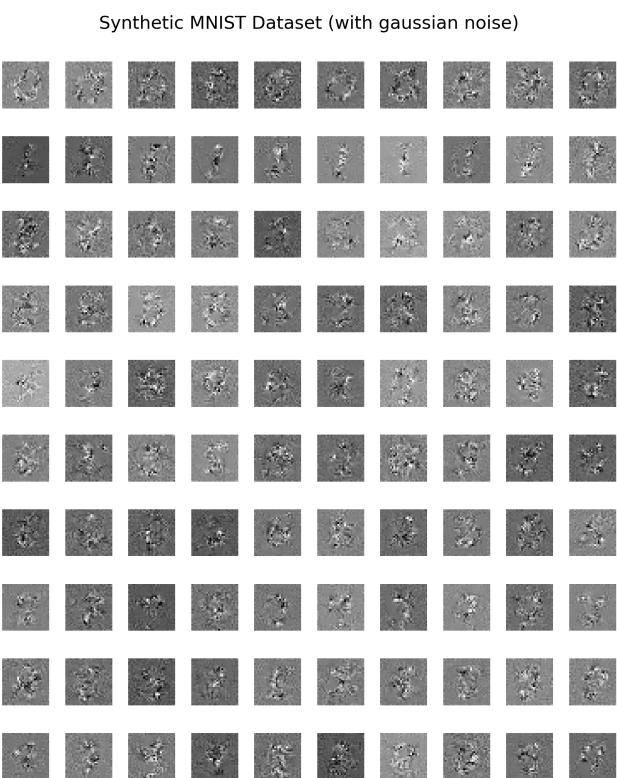


Figure 7. Synthetic MNIST Dataset created from Gaussian noise

### 3.1.3. Synthetic Dataset using Gaussian noise.

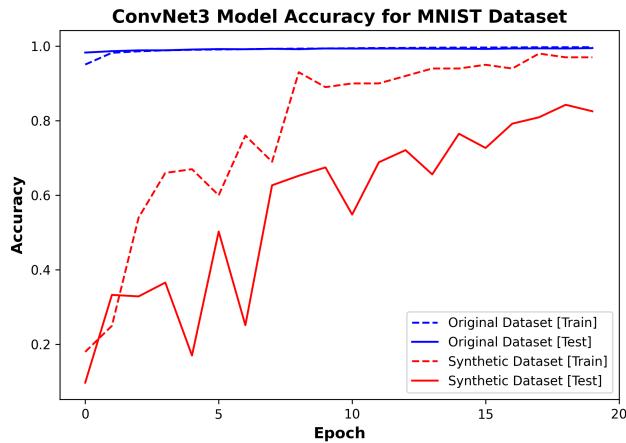


Figure 8. ConvNet-3 Model Trained using Original and Synthetic dataset

### 3.1.4. ConvNet-3 using Synthetic Dataset.

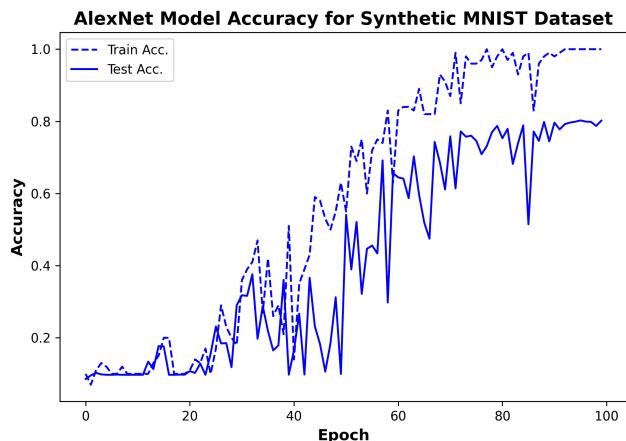


Figure 9. AlexNet Model Trained using Synthetic dataset

### 3.1.5. Cross-architecture Generalization - AlexNet.

## 3.2. MHIST Dataset

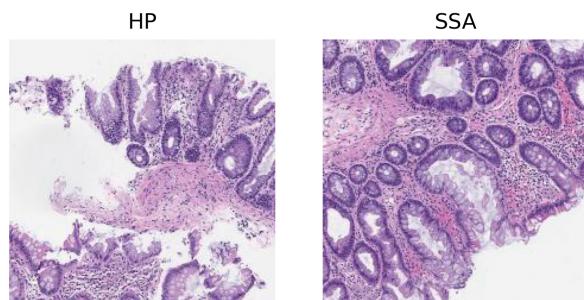


Figure 10. MHIST Dataset [4]

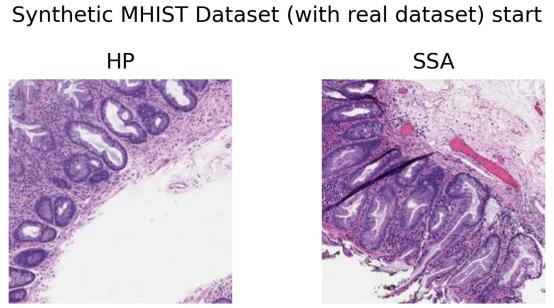


Figure 11. Sample of Synthetic MHIST Dataset created from real images (starting image)

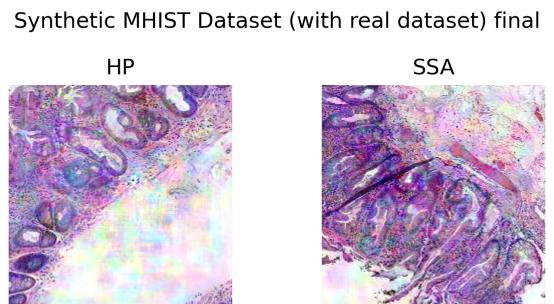


Figure 12. Sample of Synthetic MHIST Dataset created from real images (final image)

Synthetic MHIST Dataset (with real dataset)

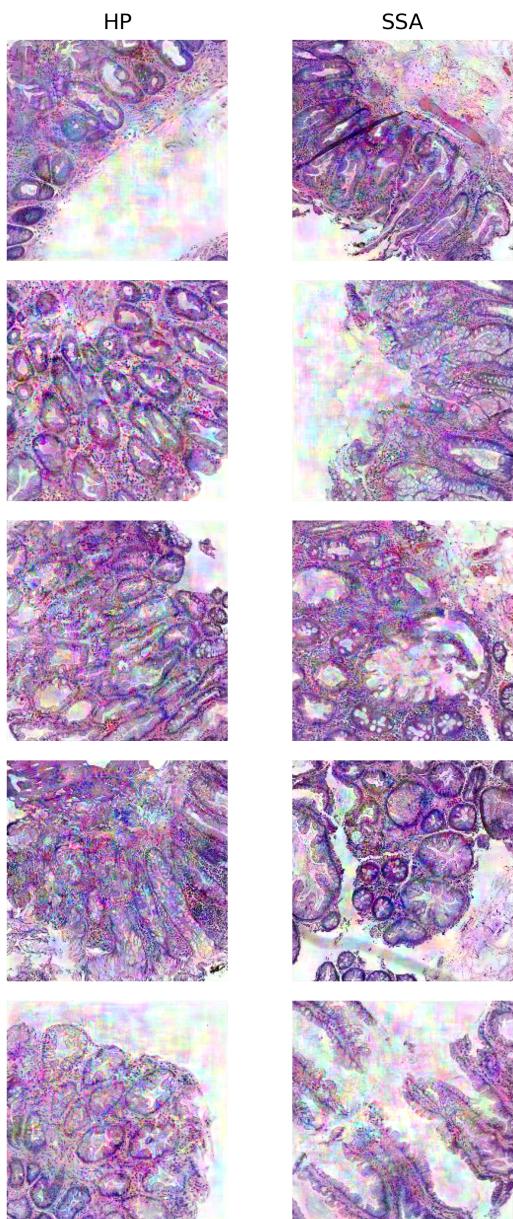


Figure 13. Synthetic MHIST Dataset created from real images

### 3.2.1. Synthetic Dataset using real images.

Synthetic MHIST Dataset (with gaussian noise) start

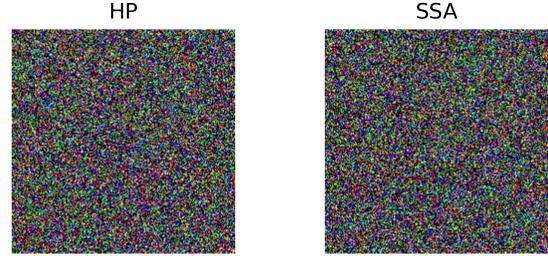


Figure 14. Sample of Synthetic MHIST Dataset created from Gaussian noise (starting image)

Synthetic MHIST Dataset (with gaussian noise) start

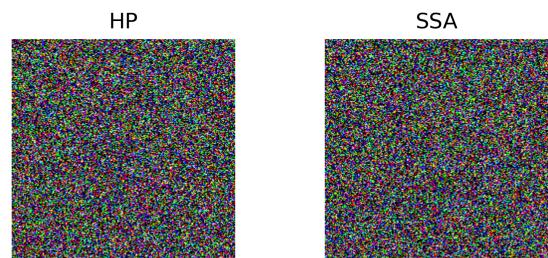


Figure 15. Sample of Synthetic MHIST Dataset created from Gaussian noise(final image)

Synthetic MHIST Dataset (with gaussian noise)

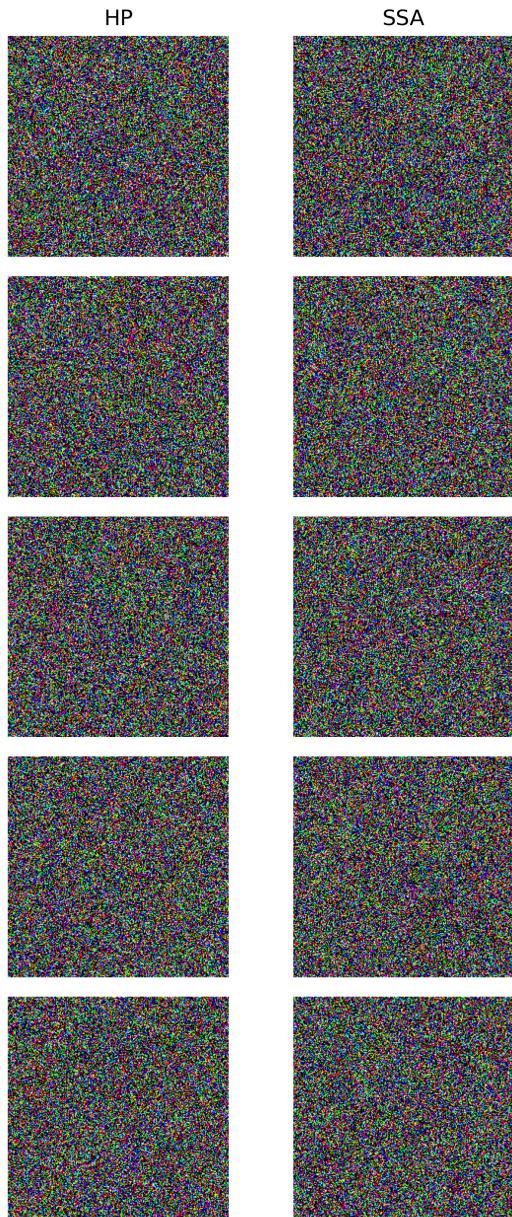


Figure 16. Synthetic MHIST Dataset created from Gaussian noise

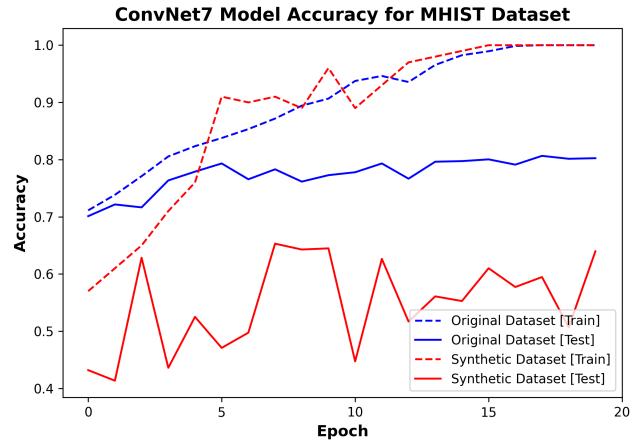


Figure 17. ConvNet-7 Model Trained using Original and Synthetic dataset

### 3.2.3. ConvNet-7 using Synthetic Dataset.

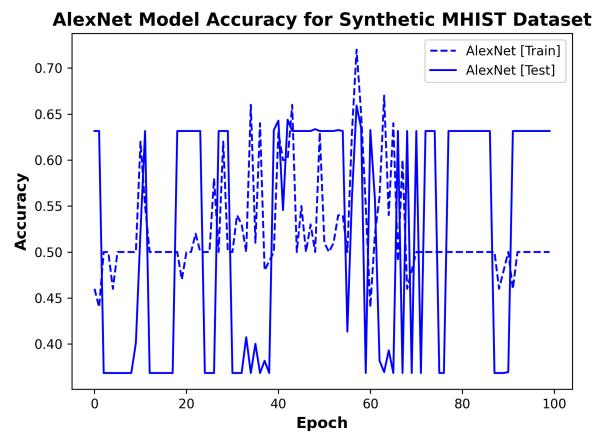


Figure 18. AlexNet Model Trained using Synthetic dataset

### 3.2.4. Cross-architecture Generalization - AlexNet.

## 3.3. Data Distillation Application

### 3.2.2. Synthetic Dataset using Gaussian noise.

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

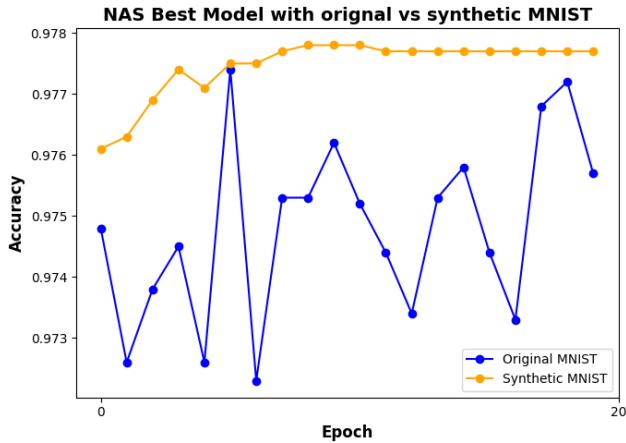


Figure 19. NAS best model with original vs synthetic MNIST dataset

## 4. Conclusion

## 5. 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 \[stat.ML\]](https://arxiv.org/abs/1503.02531).
- [3] Ahmad Sajedi et al. *DataDAM: Efficient Dataset Distillation with Attention Matching*. 2023. arXiv: [2310.00093 \[cs.CV\]](https://arxiv.org/abs/2310.00093). URL: <https://arxiv.org/abs/2310.00093>.
- [4] Jerry Wei et al. “A Petri Dish for Histopathology Image Analysis”. In: *International Conference on Artificial Intelligence in Medicine*. Springer. 2021, pp. 11–24.