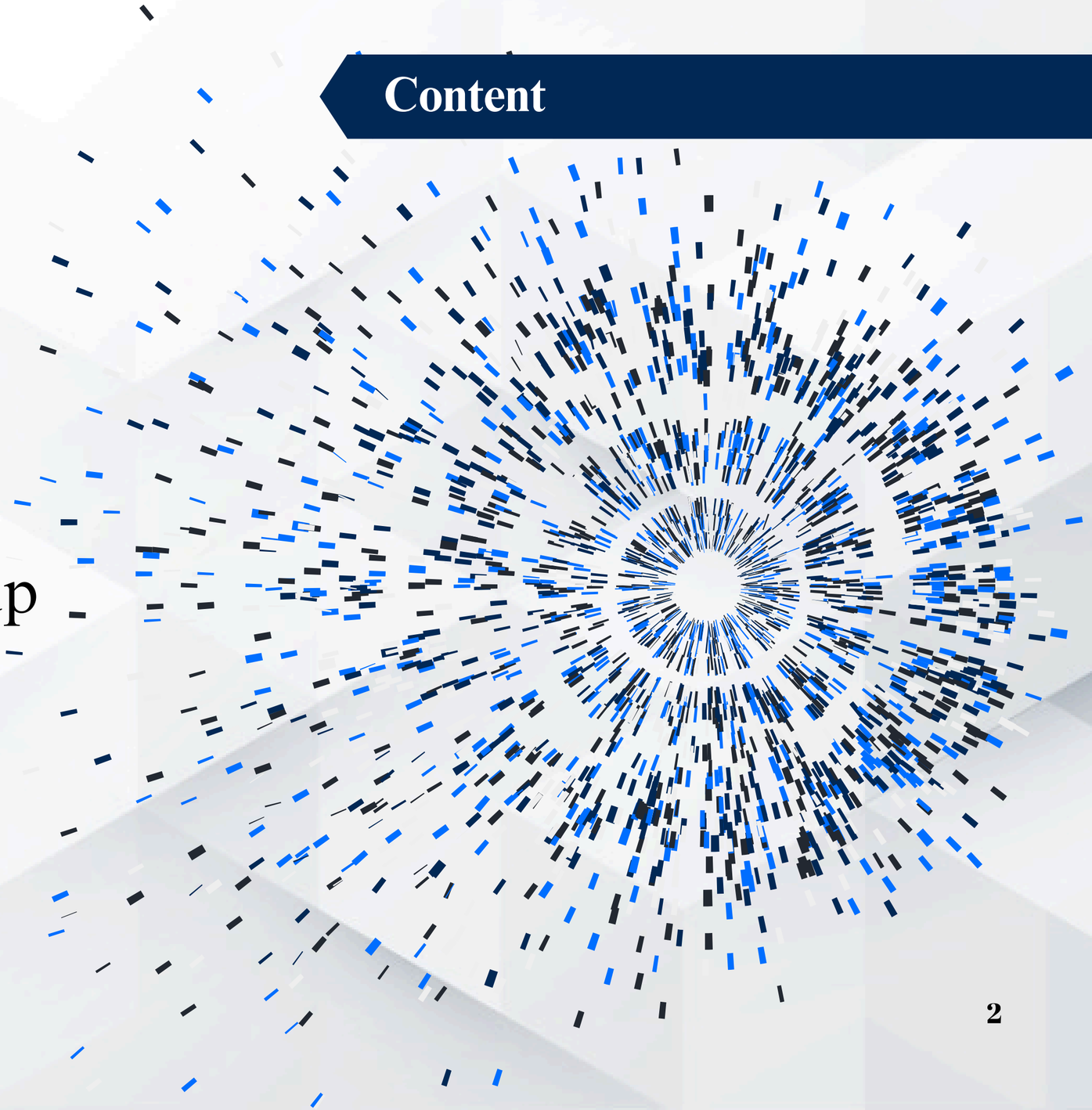
# Machine Unlearning via Synaptic Dampening


Presented By **Zeynep Uzer**

Student Number **501241624**


# Content





# Core Question:
# Can a model truly forget learned data without full retraining?

- Selective Forgetting
- Dampening Factor (α)
- Weight Suppression
- Targeted Unlearning

- Catastrophic Forgetting Avoidance
- Retention vs Forgetfulness Trade-off
- GDPR-compliant Model Updates

# Methodology

- A straightforward technique that dampens the weights associated with the class to be forgotten.
- Applied to all trainable parameters in the final layers (Dense/Conv2D).
- No retraining, no gradient update required.

**Mathematical Principle:** $w_i' = \alpha \cdot w_i \quad (0 < \alpha < 1)$

Where:

- $w_i$ = original weight
- $\alpha$ = dampening factor (e.g., 0.01)

# Methodology

1. Base Model Training: A CNN is trained on the MNIST dataset for digit classification.

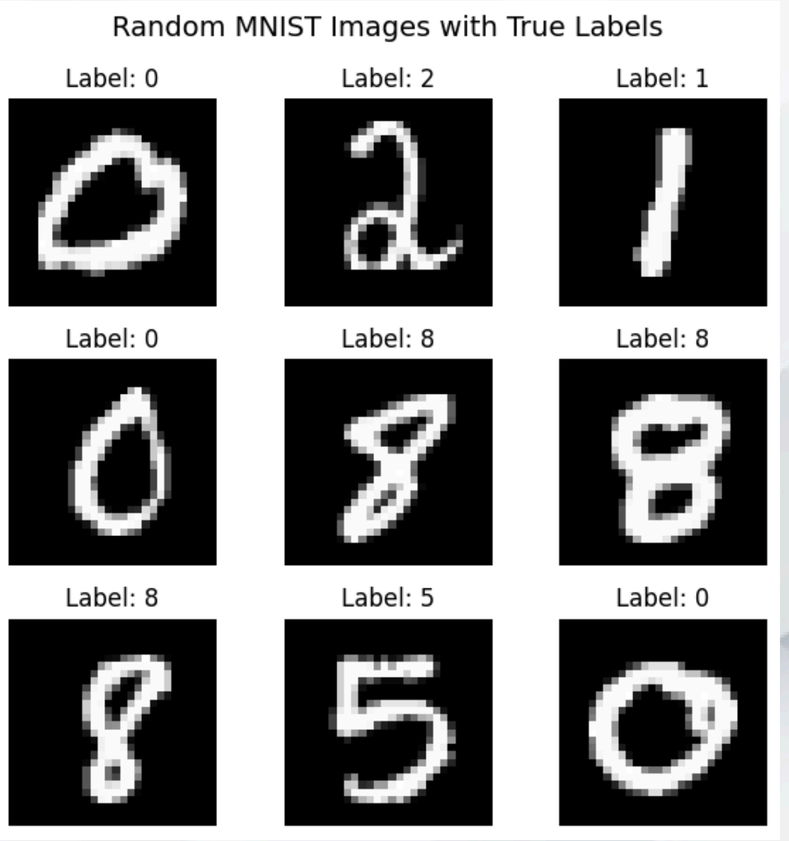
2. Forget Set Selection: A specific subset of training data, targeting a class (e.g., class 2), is marked as the forget set.

3. Synaptic Dampening: We scale down the output layer weights and biases for the forget class by a factor $\alpha$.

4. Evaluation: Post-modification model is evaluated for both forgetting accuracy and retention accuracy.

# Dataset

MNIST dataset, which includes 60,000 training and 10,000 testing grayscale images of handwritten digits across 10 classes (0–9), each sized 28×28 pixels.

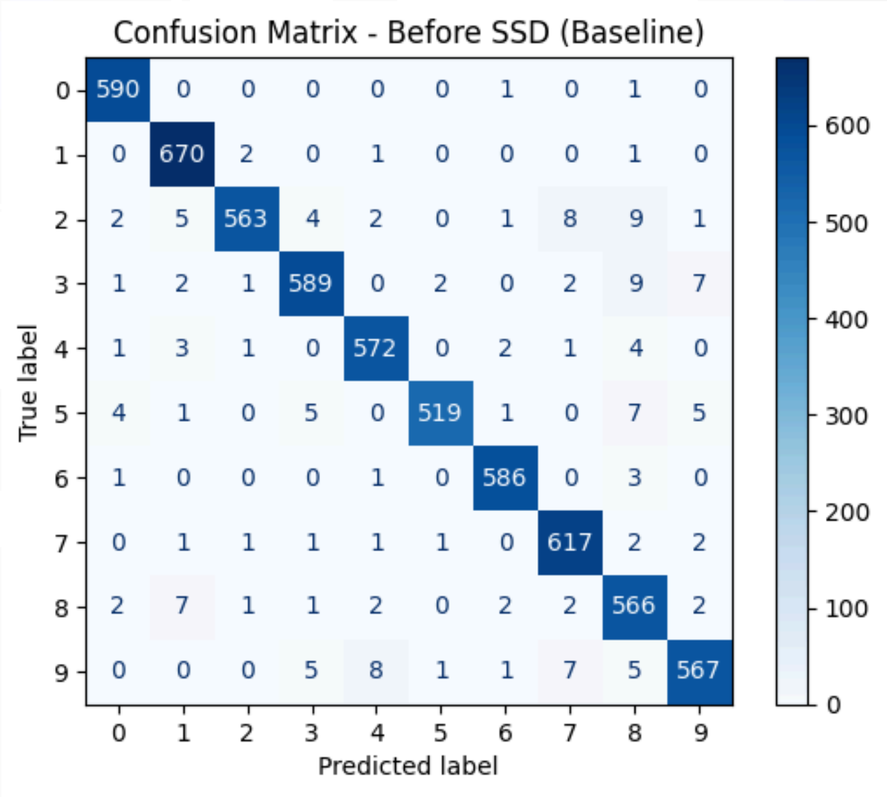


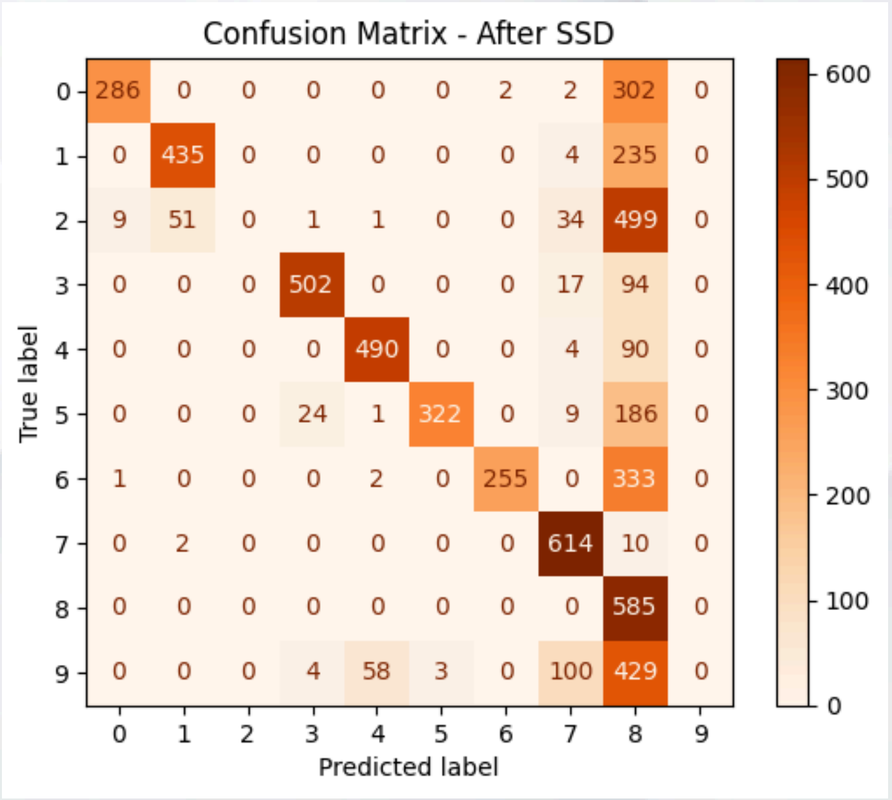

Goal: To simulate selective forgetting using SSD (Selective Synaptic Dampening), and evaluate its efficiency and trade-offs.

# Baseline Model Performance

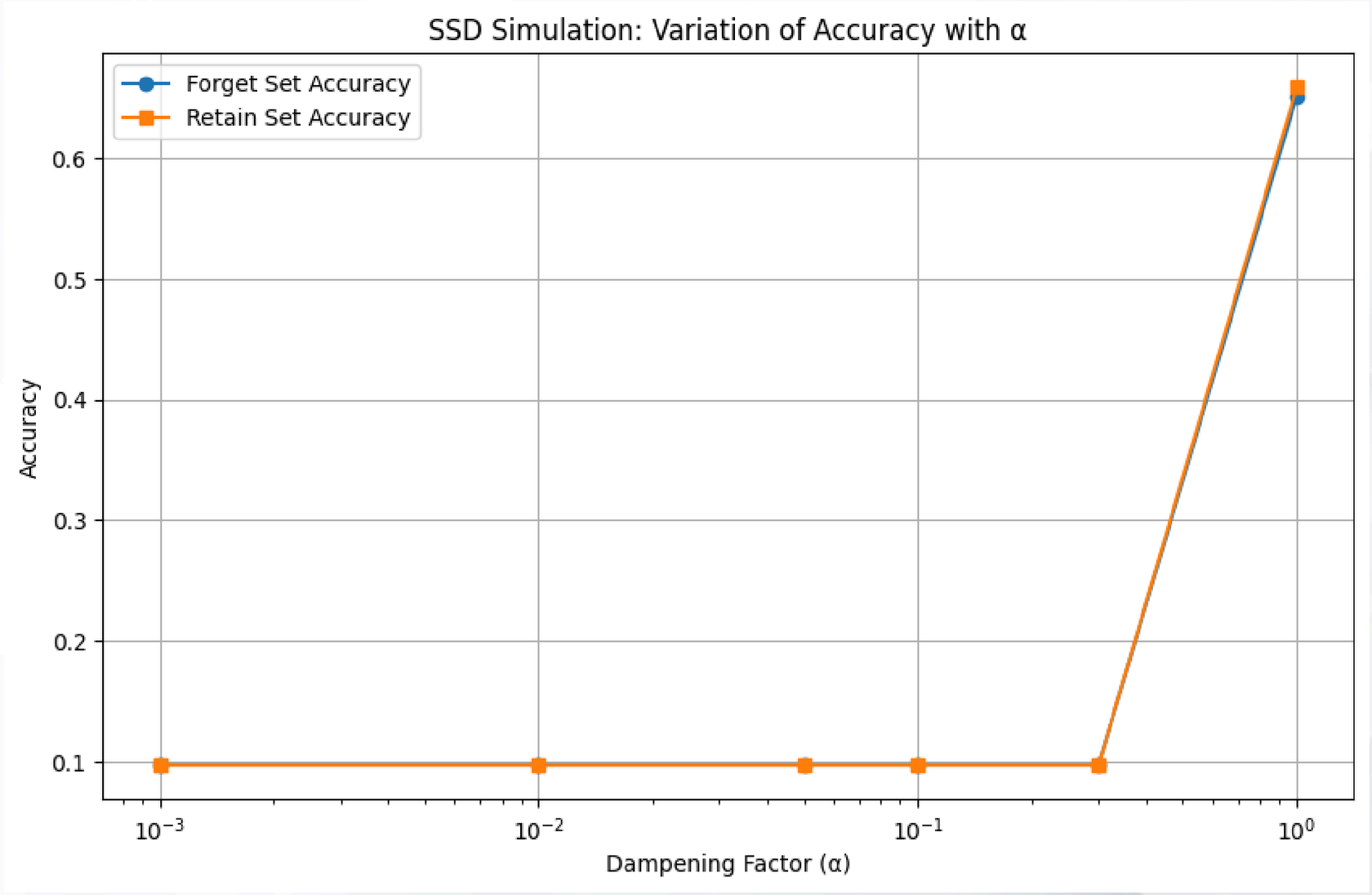
A standard CNN was trained on the MNIST training set for 2 epochs using categorical cross-entropy and the Adam optimizer. The baseline model achieved a test accuracy of ~98%.

Below is the confusion matrix before applying SSD:

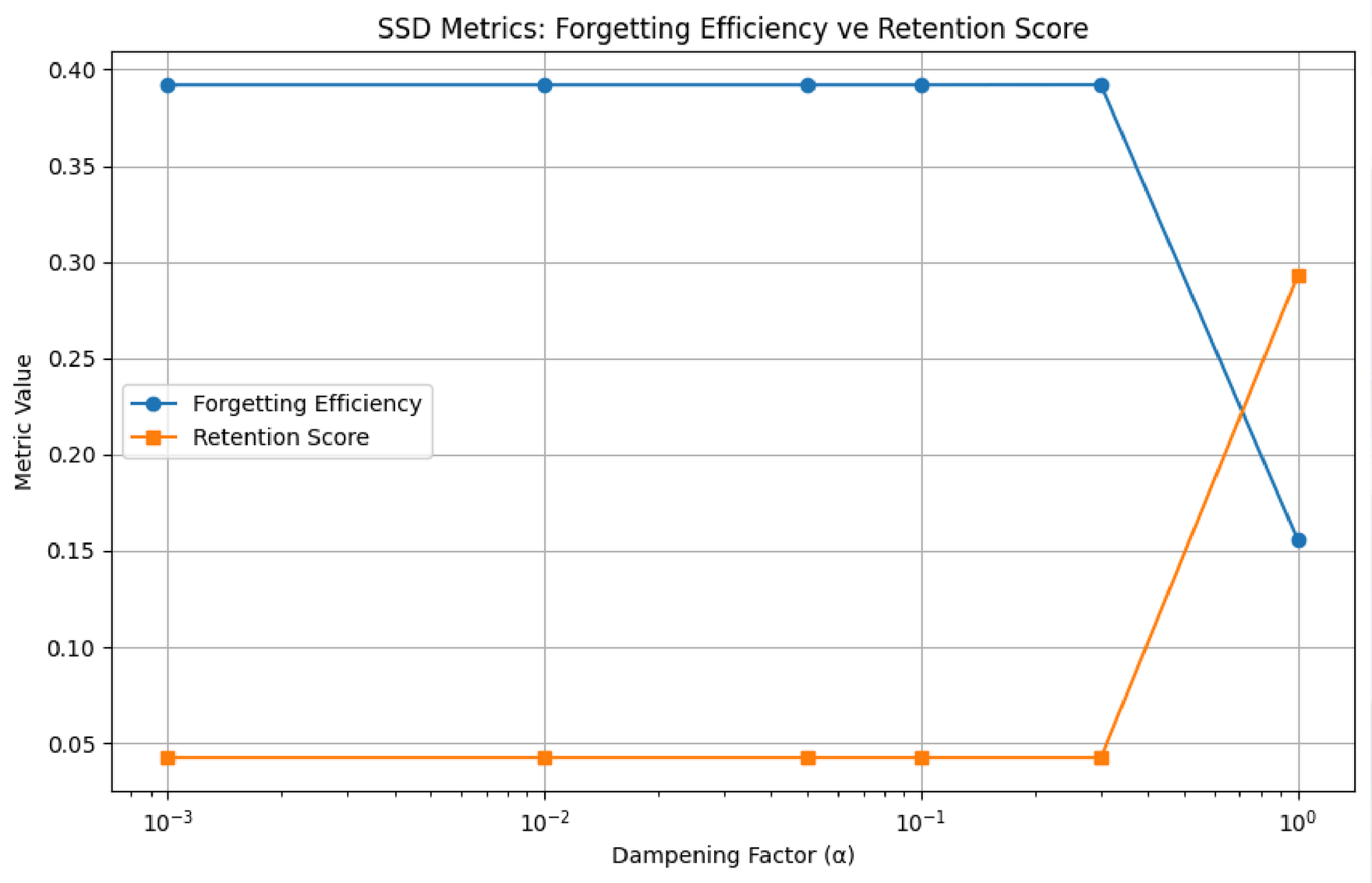


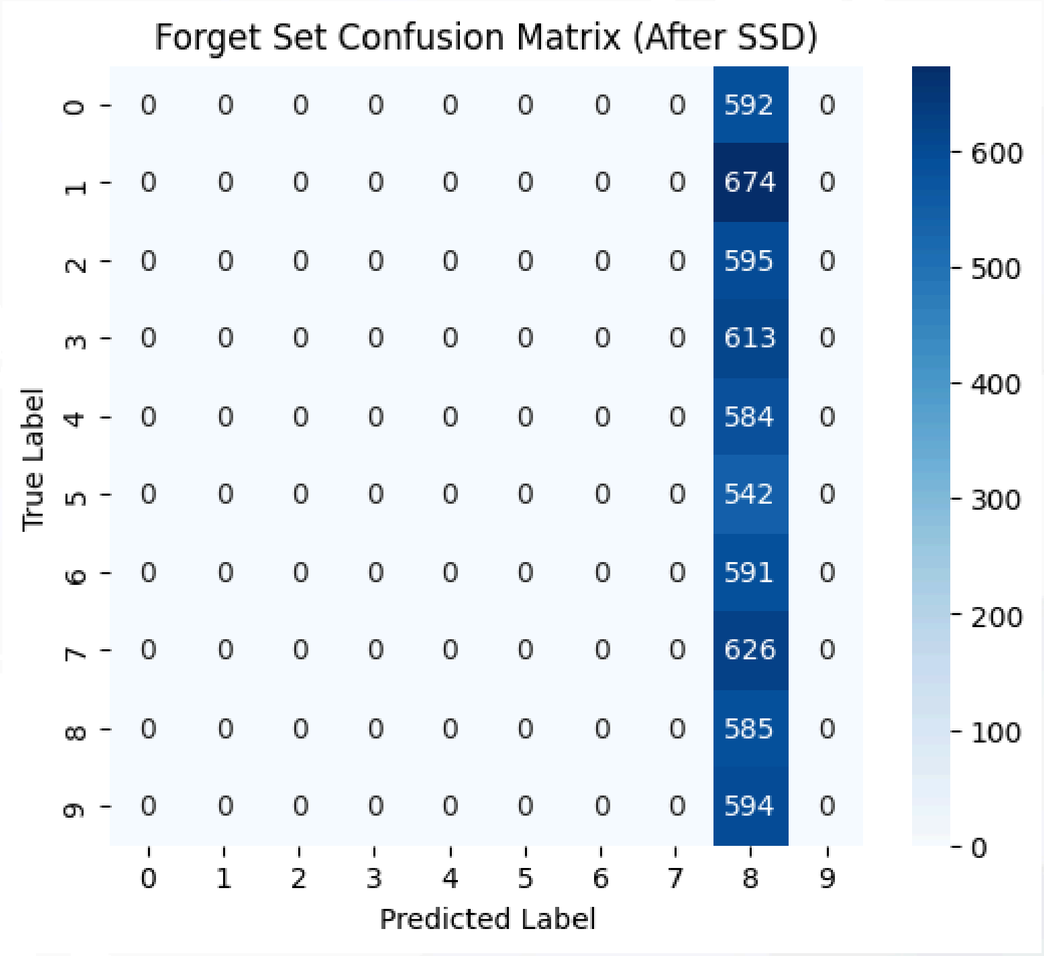
# Applying SSD: How Forgetting is Controlled

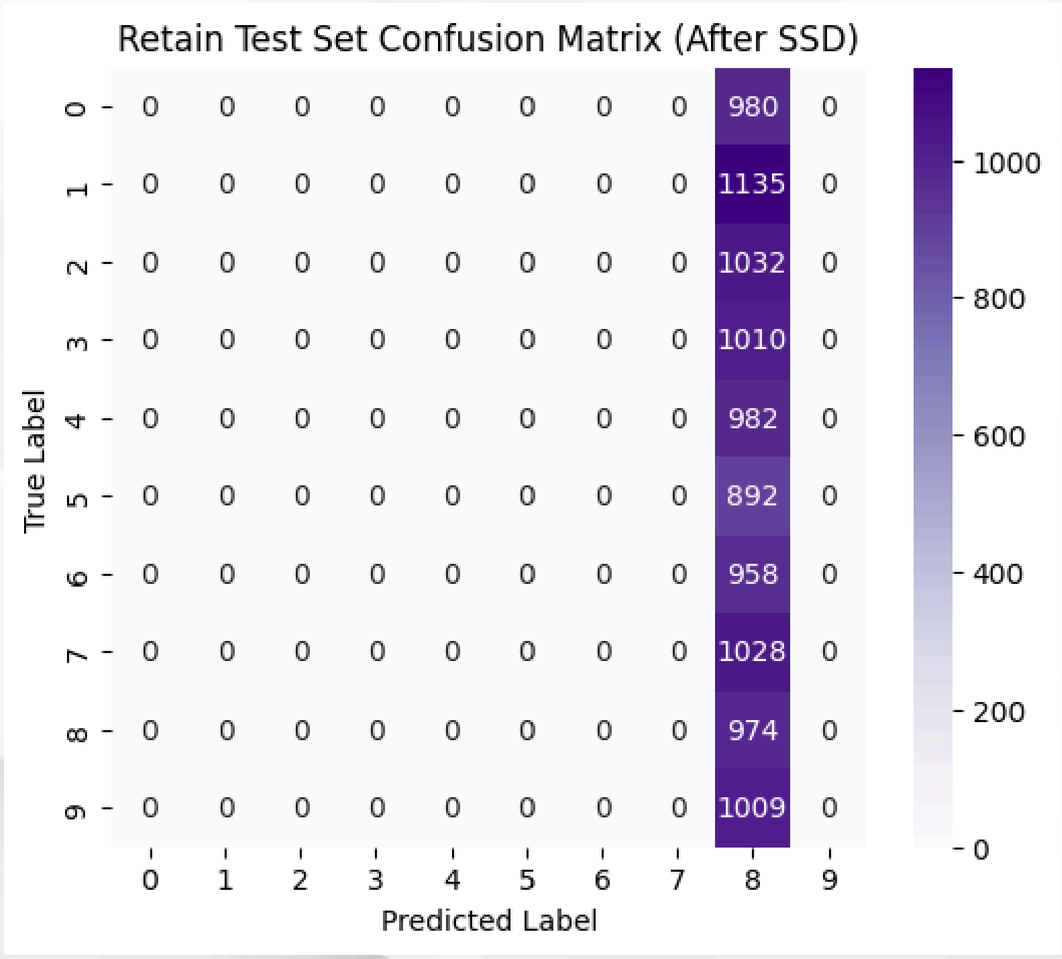


# Applying SSD: How Forgetting is Controlled

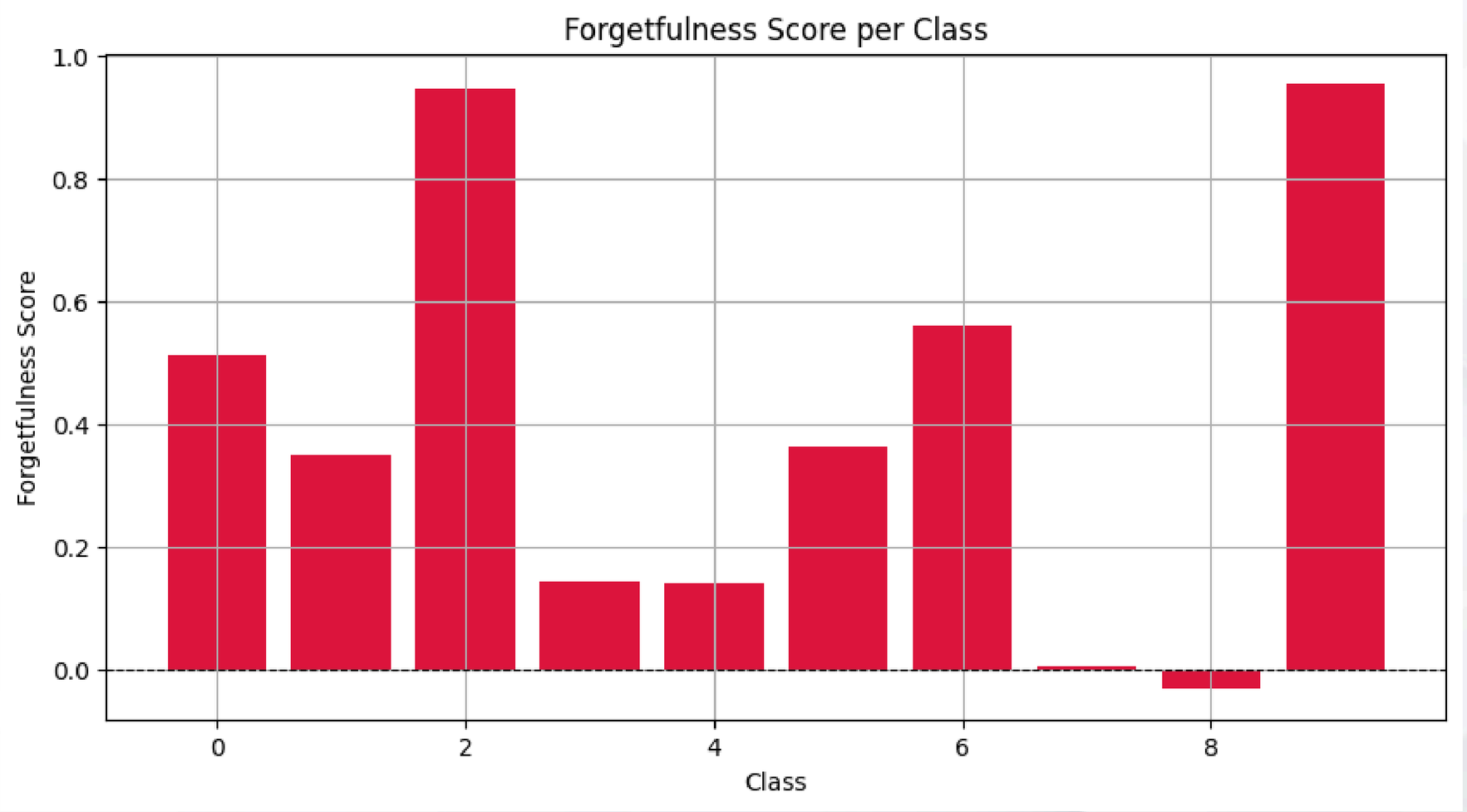


# Effect of SSD on Classification Results



# Class-wise Forgetfulness Overview

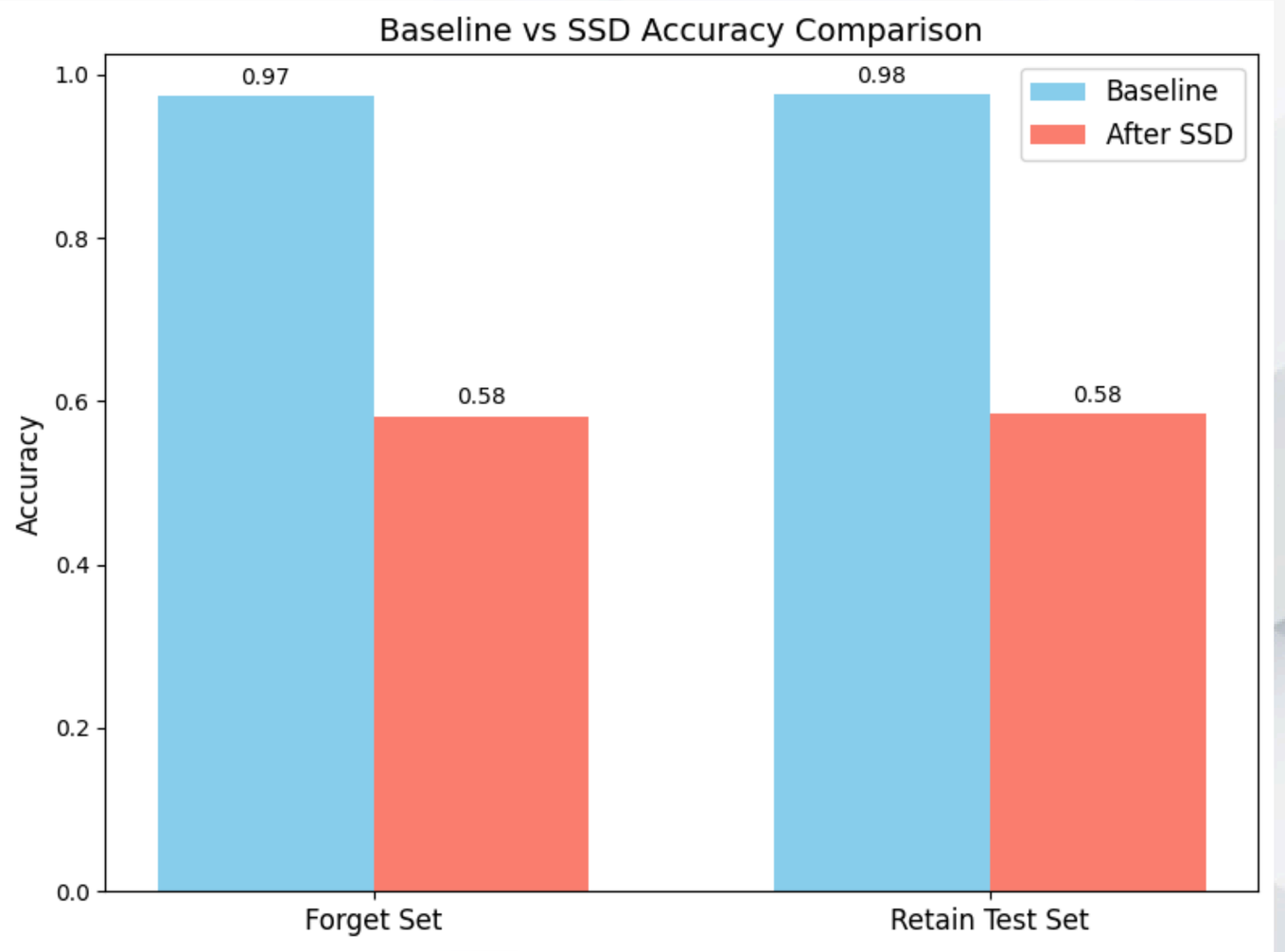



**This metric reveals how strongly each class was affected. Our target class shows the most dramatic decrease.**

# Visual Accuracy Comparison

The bar chart below compares model accuracy on both forget and retain sets before and after SSD.

This visualization clearly shows that SSD selectively impacts the forget class while preserving general performance.



# Conclusion & Future Work

✅ SSD enables targeted forgetting without full retraining.

✅ The dampening factor α controls the strength of forgetting.

✅ Forgetting is class-specific and retention is preserved.

✅ Potential implications for GDPR compliance, data privacy, and continual learning.

💡 Future directions include:

- Applying SSD to large-scale datasets.
- Integrating with federated and edge learning frameworks.
- Exploring inverse tasks: memory reinforcement

İTÜ

# Thank You!

**All source codes, outputs, and the 3-minute project presentation video are available at: [https://github.com/zynepuzer/machine_unlearning]**


Presented By **Zeynep Uzer**

E- mail Adress **uzer19@itu.edu.tr**