CS354 MINOR PROJECT

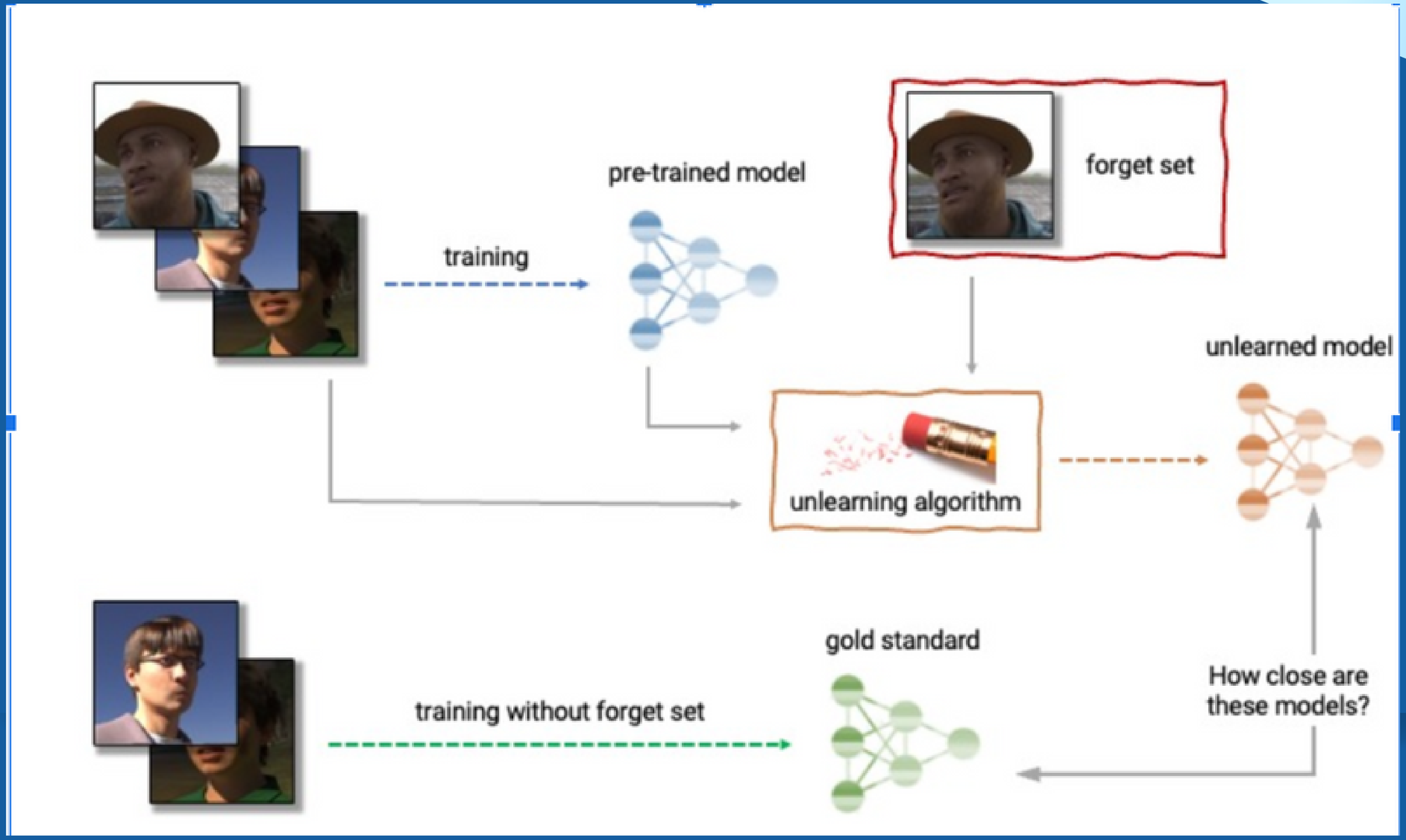
MACHINE UNLEARNING

Problem Definition

- The project is based on the development of an efficient architecture to "unlearn" machine learning models on a certain subset of data

- Focus is to remove image specific features while keeping the generalization capabilities intact

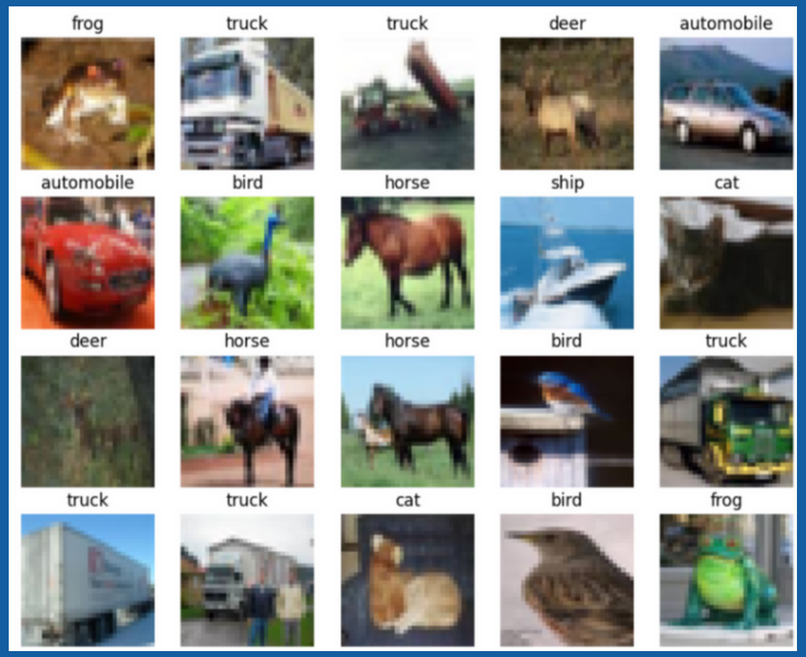
PROJECT PIPELINE



DATA ANALYSIS

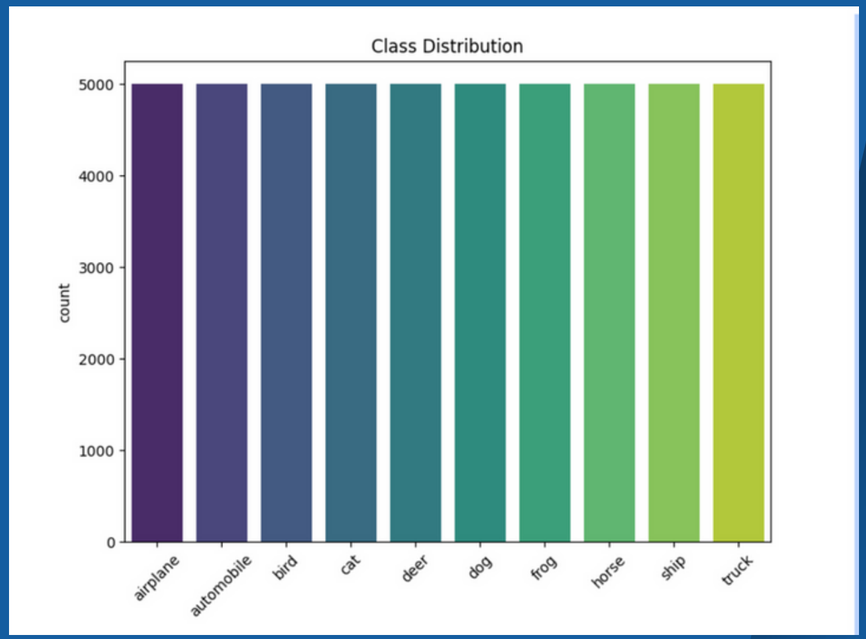
CIFAR-10 DATASET

- Total Images: 60,000
- 32 pixels x 32 pixels x 3 channels (RGB)
- 10 classes



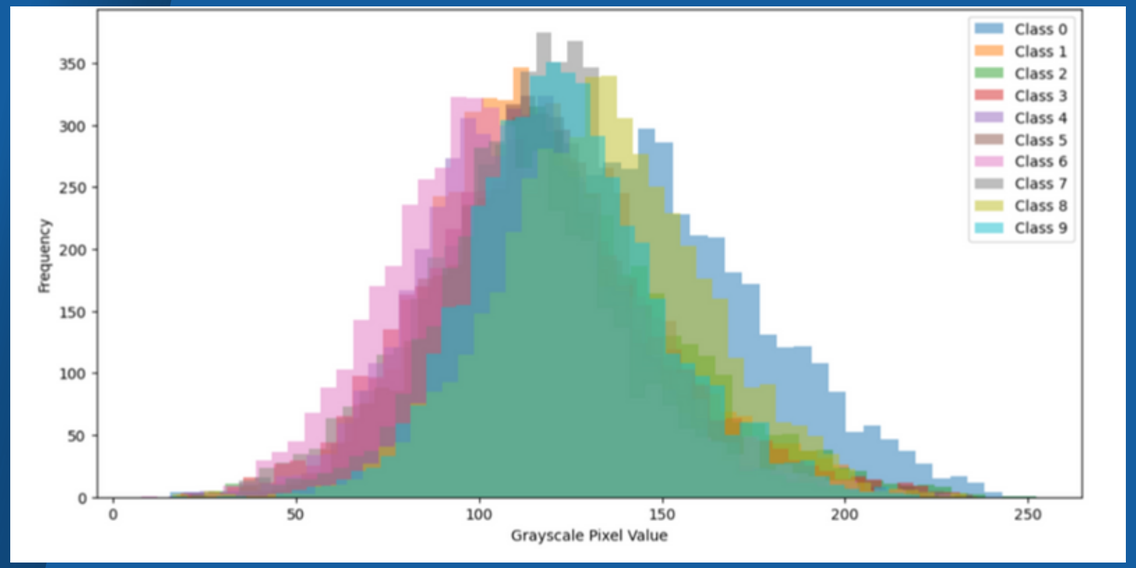
VISUALIZING CLASS DISTRIBUTION:

Mean Intensity by Pixel and RGB Graph:

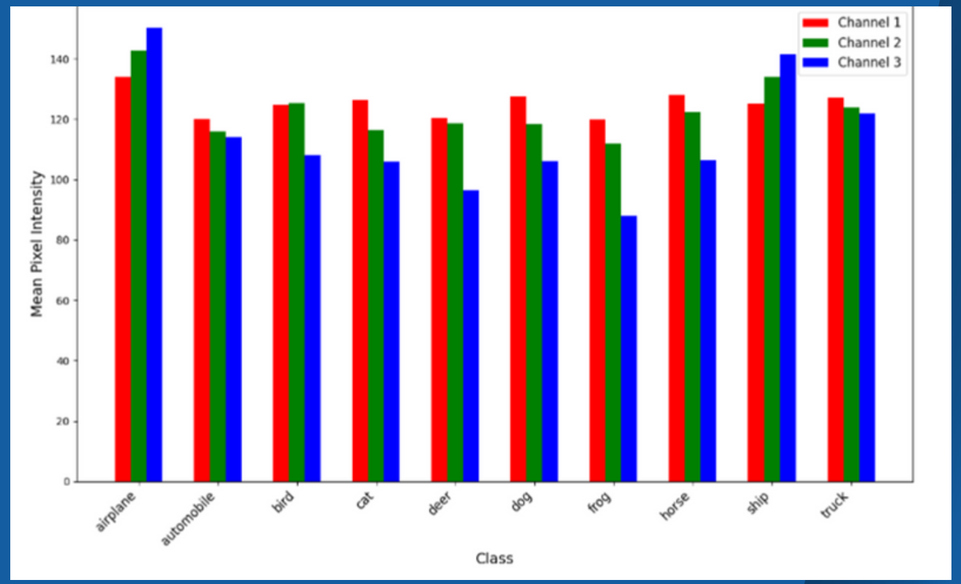


DATA ANALYSIS

HISTOGRAM OF GRAYSCALE PIXEL VALUES GRAPH



MEAN INTENSITY BY PIXEL AND RGB GRAPH



Preprocessing Techniques

- **Normalization** : normalize these pixel values and bring them in a range of 0–1 by dividing the pixel values by 255.
- **Data Augmentation** – transformations to existing data, such as rotating, flipping, and cropping images.

MODEL USED

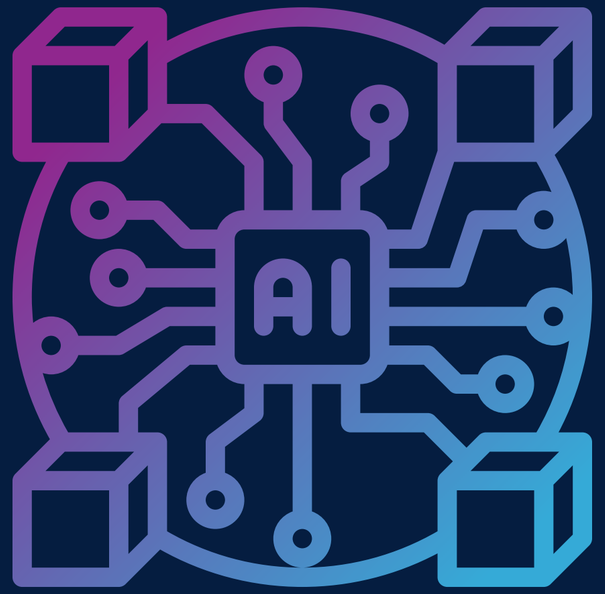
Resnet18

Key features of ResNet-18 include:

Residual Blocks
Shortcut Connections:
Downsampling
Global Average Pooling

Unlearning Algorithms

1. Unlearning by fine tuning



- **Fine-tuning method**: Updates pre-trained neural network.
- **Retain set**: Maintains model's performance.
- **Forget set**: Removes unwanted influence.
- **Simple implementation**: Easy to understand.
- **Efficient resource usage**: Focuses on retain set.
- **Knowledge retention**: Preserves useful information.

Cons:

- **Limited unlearning**: Doesn't address forget set directly.
- **Potential overfitting**: May occur on limited data.
- **Domain knowledge required**: Tuning parameters needs expertise.

Unlearning Algorithms

2. Unlearning as Adversarial Regularization

- **Adversarial regularization**: Modifies pre-trained neural network.
- **Objective**: Reduce influence of forget set.
- **Adversarial training**: Discriminator distinguishes outputs.
- **Regularization**: Min-max optimization maximizes discrimination loss.
- **Distillation strategy**: Matches output distribution of oracle.
- **Pros:**
  - **Effective unlearning.**
  - **Retains useful information.**
  - **Flexible optimization.**
- **Cons**:
  - Complexity and hyperparameter tuning.
  - Potential performance trade-offs.
- **Requires tuning**: Learning rates, strength parameters.
- **Challenges**: Time-consuming hyperparameter optimization.
- **Potential impact**: Reduced generalization ability.
- **Task-specific**: Especially relevant in age classification.

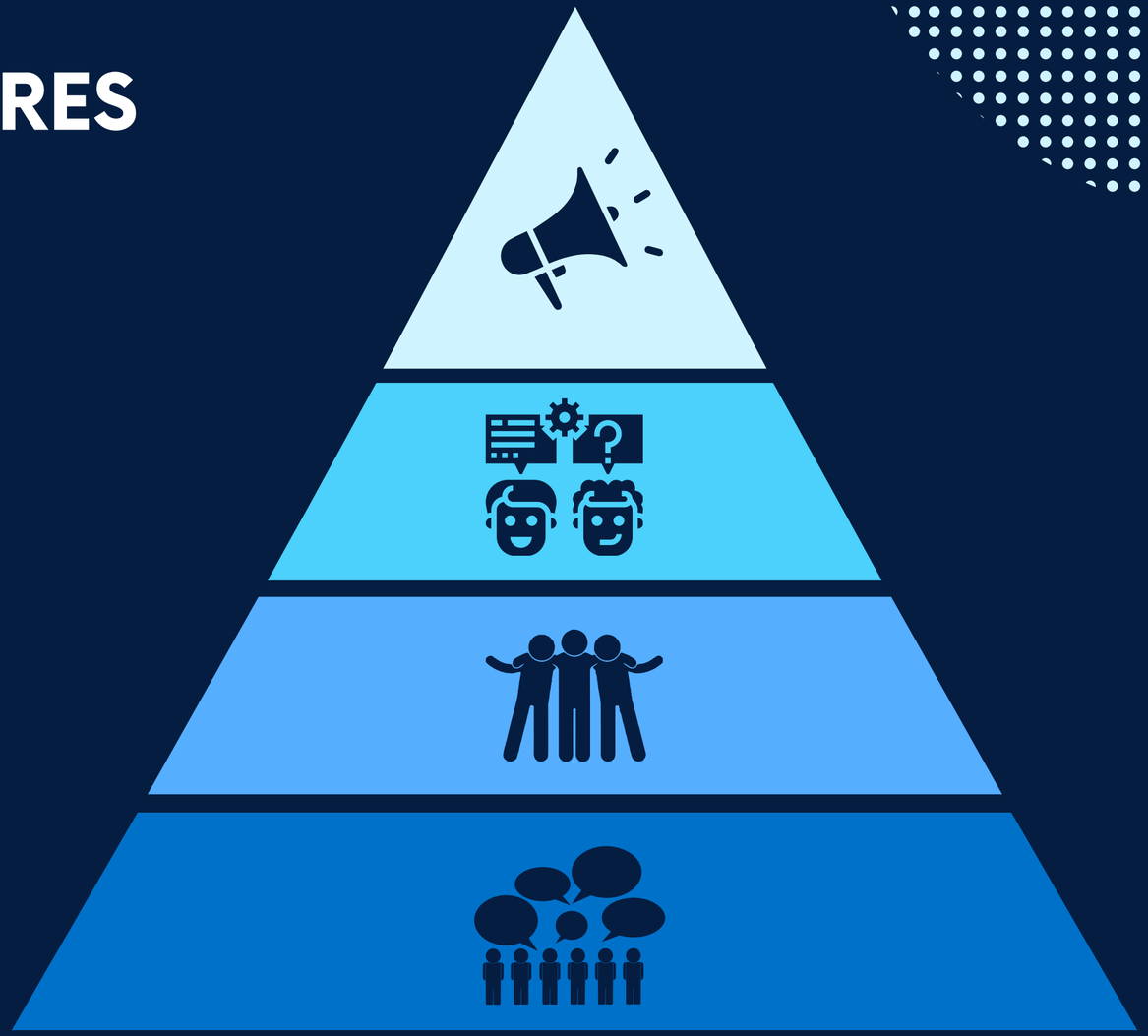
PERFORMANCE MEASURES USED

01 MIA (Membership Inference Attacks)

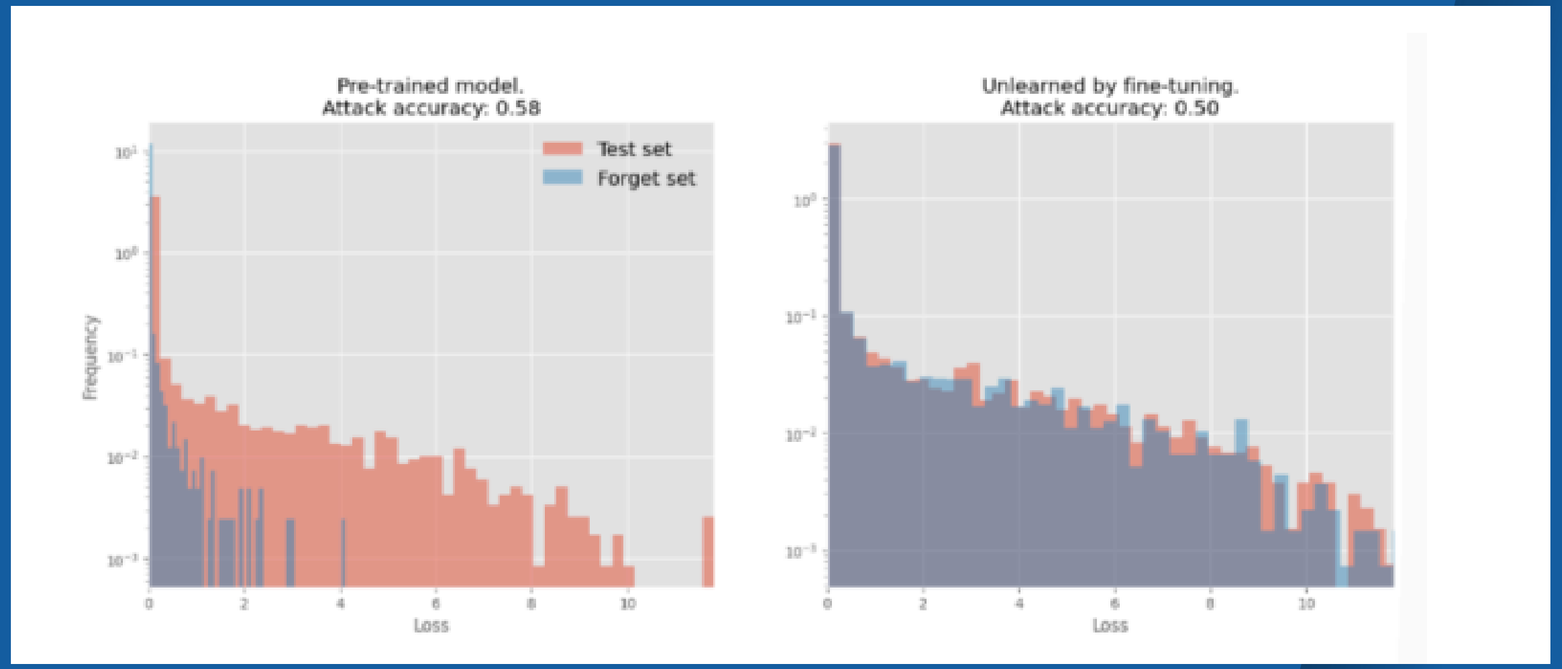
02 Gold Standard

03 Forgetting Quality

04 Unlearning Score



EXPERIMENTATION AND RESULTS



EXPERIMENTATION AND RESULTS

UNLEARNING BY FINE TUNING

SNo.	Epochs	LR	F	US	GS	MIA O	MIA U
1	5	0.1	0.1113	0.1046	0.507	0.580	0.515
2	5	0.01	0.1342	0.1351	0.502	0.579	0.559
3	5	0.001	0.0380	0.0384	0.500	0.579	0.579
4	15	0.1	0.1637	0.1551	0.502	0.579	0.503
5	**25**	**0.1**	**0.1715**	**0.1616**	**0.502**	**0.577**	**0.504**

UNLEARNING BY ADVERSARIAL TRAINING

SNo.	HS	LR	F	US	GS	MIA O	MIA U
1	16	0.0004	0.1151	0.1114	0.499	0.577	0.569
2	16	0.004	0.0854	0.0166	0.502	0.579	0.520
3	**128**	**0.0004**	**0.1271**	**0.1230**	**0.504**	**0.577**	**0.572**
4	8	0.0004	0.1153	0.1117	0.497	0.577	0.573
5	16	0.00004	0.0336	0.0330	0.501	0.577	0.575

LEARNINGS AND CONCLUSION

- Introduction to "Machine Unlearning" research.
- Analysis on Cifar-10 dataset.
- Base architecture: ResNet.
- Two algorithms: fine-tuning and adversarial training.
- Extensive evaluation and metric usage.
- Fine-tuning algorithm exploration
- **Achieved Unlearning score: 0.1616.**

Thank You

Presented by:

Prajakta Darade Roll No. 210001052
Tanisha Sahu Roll No. 210001071