import random

AAI-595 B - DR. HAO WANG

PRADHYUMNA NAGARAJA HOLLA

SUDHARSHINI HARISARAVANAN

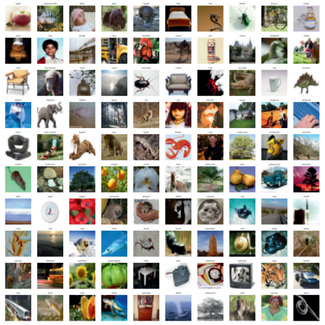
KRISHNA SAI SRINIVAS VENIGALLA

# SECURING IMAGE CLASSIFIERS AGAINST ADVERSARIAL ATTACKS: A STUDY ON CIFAR-100

CODE

MACHINE LEARNING MODELS ARE VULNERABLE TO ADVERSARIAL ATTACKS, WHERE TINY PERTURBATIONS MISLEAD CLASSIFIERS.

## DATASET: CIFAR-100: 100 CLASSES, 500 PER CLASS



## PROBLEM STATEMENT

MODERN DEEP LEARNING MODELS LIKE RESNET-18 ARE POWERFUL BUT HIGHLY VULNERABLE TO ADVERSARIAL ATTACKS—TINY, IMPERCEPTIBLE INPUT CHANGES THAT CAN MISLEAD MODELS INTO MAKING CONFIDENT YET INCORRECT PREDICTIONS. THIS POSES SERIOUS RISKS IN CRITICAL AREAS LIKE AUTONOMOUS DRIVING, MEDICAL DIAGNOSTICS, AND SECURITY SYSTEMS, WHERE SUCH ERRORS CAN HAVE REAL-WORLD CONSEQUENCES.

## WHY IS IT IMPORTANT

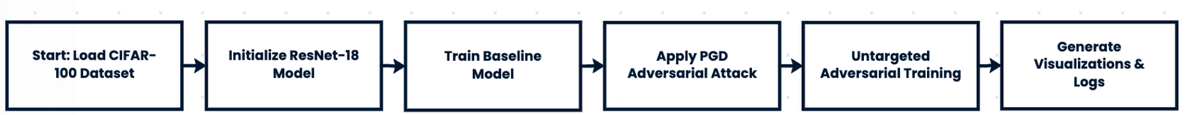
ENSURING ROBUSTNESS AGAINST ADVERSARIAL ATTACKS IS ESSENTIAL FOR BUILDING SAFE, RELIABLE, AND TRUSTWORTHY AI SYSTEMS. WITHOUT PROPER DEFENSES, MALICIOUS INPUTS CAN:

- MISLEAD MODELS,
- UNDERMINE PUBLIC TRUST,
- LEAD TO SAFETY HAZARDS IN HIGH-STAKES DOMAINS.

ADVERSARIAL TRAINING IS A KEY STRATEGY TO IMPROVE MODEL RESILIENCE AND ENABLE SECURE DEPLOYMENT IN THE REAL WORLD.

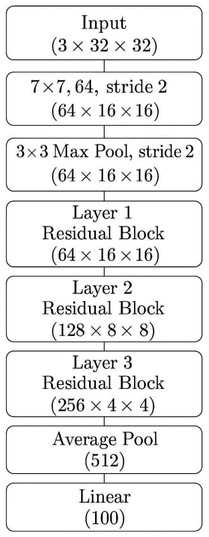
## IMPLEMENTATION PIPELINE

A STEP-BY-STEP FLOW FROM BASELINE TRAINING TO EVALUATING ADVERSARIAL ROBUSTNESS



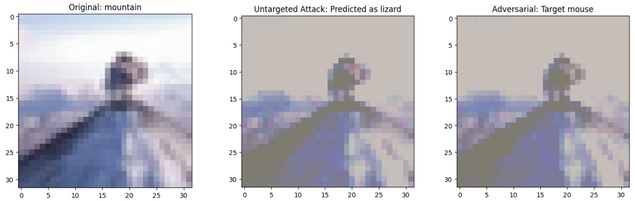

## MODEL ARCHITECTURE

RESNET-18 FOR CIFAR-100

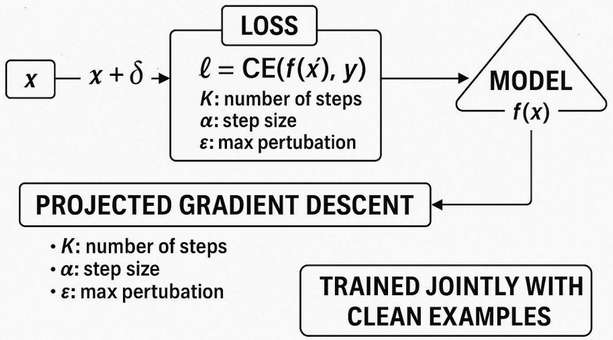



ResNet-18
CIFAR–100

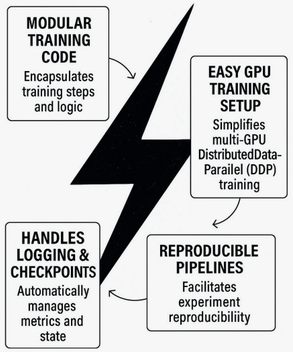
## ORIGINAL TARGETED & UNTARGETED



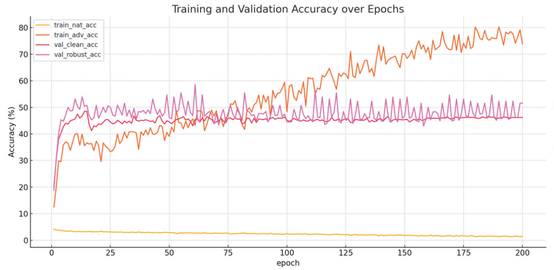

## PGD ADVERSARIAL ATTACK

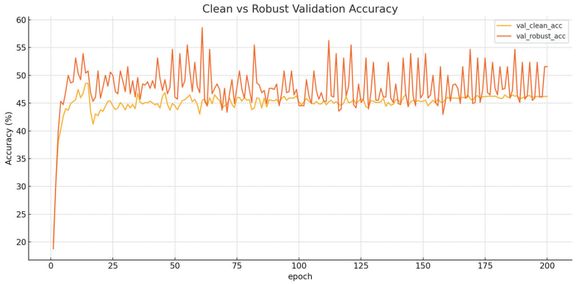


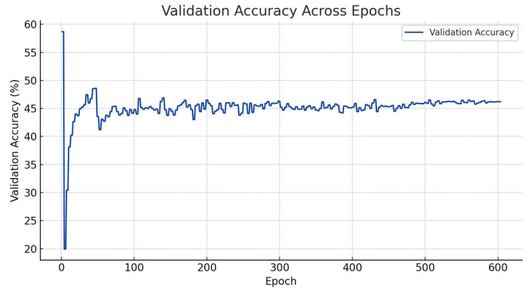

## PYTORCH LIGHTNING

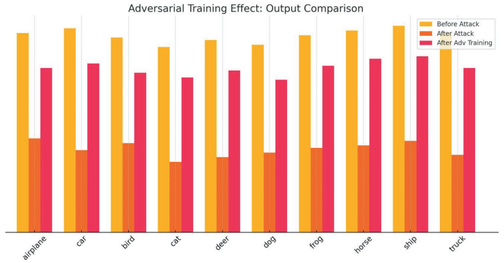



## RESULT




| MODEL | ACCURACY | TARGETED | UNTARGETED |
|---|---|---|---|
| BASELINE | 68.57% | 76.14% | 75.45% |
| BEST CLEAN | 47.33% | 17.50% | 60.88% |
| BEST ROBUST | 45.36% | 21.92% | 55.44% |

https://arxiv.org/pdf/2302.04638 https://arxiv.org/pdf/2206.02158 https://arxiv.org/pdf/2303.01870