*Team Model Marvericks*

***Github Link***

# Handwritten Digit Recognition with Adversarial Robustness

CS550/DSL501: Machine Learning 2024-25M

**Team Members**:
Arey Pragna Sri(12240230)
Matcha Jhansi Lakshmi(12241000)
Nannepaga Vanaja(12241110)

# Problem Statement

This project aims to develop a **robust handwritten digit recognition** model that resists adversarial perturbations, ensuring reliable performance in practical applications such as automated data entry and digital document processing. By incorporating **adversarial training** techniques, we seek to enhance the model's resilience, ensuring high accuracy even when faced with adversarially modified inputs in real-world scenarios.

# Previous Work

Model Mavericks

- The initial work on handwritten digit recognition utilised traditional CNNs, evaluated on MNIST for image classification capabilities. (Reference - link)
- However, these models were found to be highly vulnerable to adversarial attacks.  So Researchers explored adversarial methods, such as FGSM for attacking and comparing different models like LeNet and Simple CNN, revealing significant differences in robustness. (Reference - link)
- There is ongoing research in this area, with many studies focusing on methods to make machine learning models more robust against these adversarial threats.

# Novelty

- Our work compresses of simple CNN and deep CNN architectures, highlighting how model depth affects adversarial vulnerability.
- By using both FGSM and PGD attacks, we have analyzed  their impact on different models.

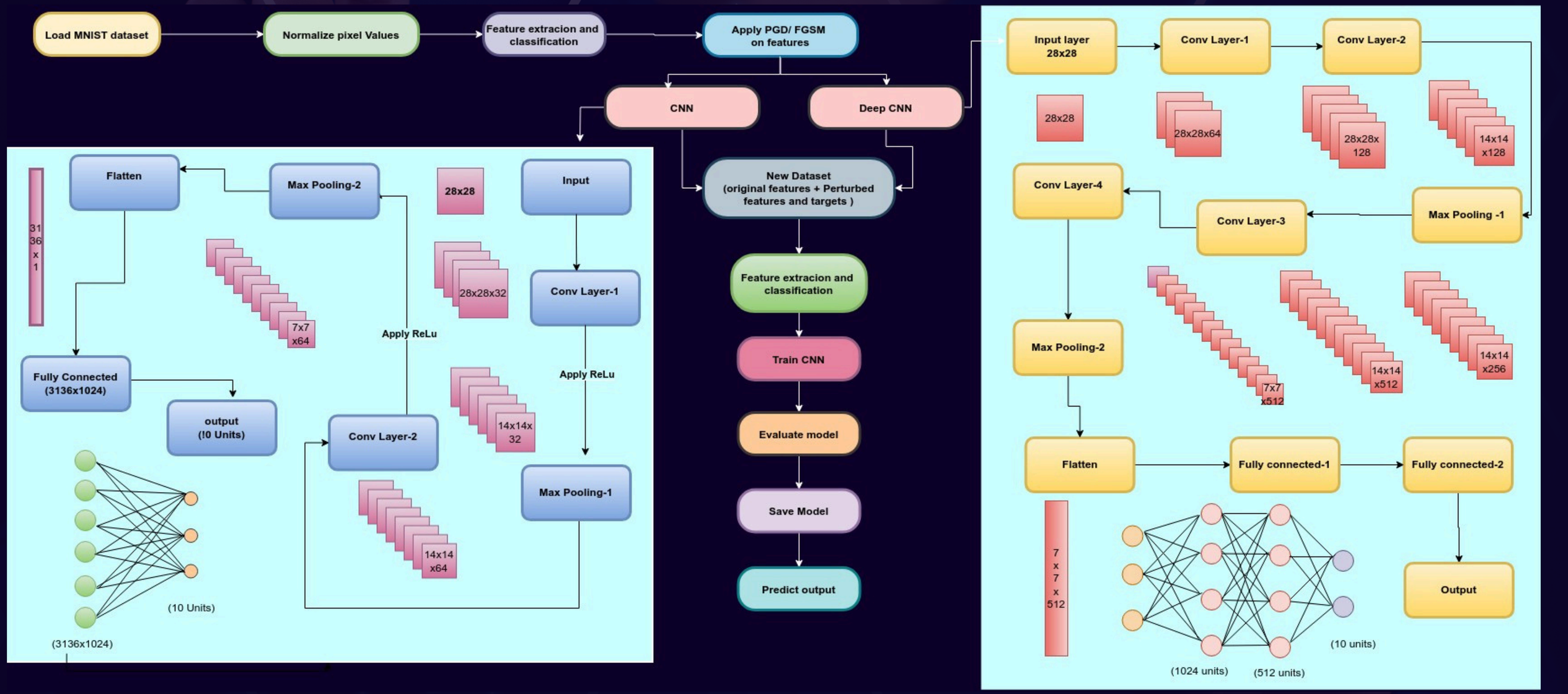
# Model Development

Model Mavericks

- Data Preparation:
  - Downloaded the MNIST dataset and generated adversarial datasets using PGD and FGSM
  - Combined clean and adversarial datasets to create a robust training dataset.
- Model Design: Developed SimpleCNN and DeepCNN with:
  - Convolutional Layers to extract features.
  - ReLU Activation for non-linearity.
  - Max-Pooling for dimension reduction.
  - Fully Connected Layers with softmax for classification.
- Training and Evaluation:
  - Trained models using categorical cross-entropy loss and Adam optimizer.
  - Applied data augmentation (rotations, shifts, zooms) for better generalization.
  - Evaluated models on test data to assess accuracy and adversarial robustness.
  - Compared SimpleCNN and DeepCNN to identify the more resilient model.

# Methodology





# Mathematics Behind the model

Model Mavericks

- **Convolutional Operation**:

$$(f * g)(x, y) = \sum_{i=-k}^{k} \sum_{j=-k}^{k} f(i, j) \cdot g(x - i, y - j)$$

here ' f ' is input image, 'g' is filter/kernel & x,y represents pixel location. This is to extract features.

- **ReLU activation function**

$$\text{ReLU}(x) = \max(0, x)$$

it introduces non-linearity so that model can learn complex representations.

- **Fast Gradient Sign Method (FGSM):**

$$x_{adv} = x + \epsilon \cdot \text{sign}\left(\nabla_x \mathcal{L}(f_\theta(x), y)\right)$$

- **Projected Gradient Descent(PGD)**

$$x_{adv}^{t+1} = \text{Clip}_{x,\epsilon}\left(x_{adv}^{t} + \alpha \cdot \text{sign}\left(\nabla_x \mathcal{L}(f_\theta(x_{adv}^{t}), y)\right)\right)$$

here L : loss function, $\epsilon$: Perturbation limit, $\alpha$: Step size, $f_\theta(x)$: Model prediction

# Evaluation Metrics

We define accuracy as the primary metric for assessing the performance of models under adversarial robustness scenarios. To enhance robustness, we evaluated the models using different adversarial attack methods, such as PGD and FGSM, and analyzed their accuracy trends over multiple epochs.

# Results obtained

Model Mavericks

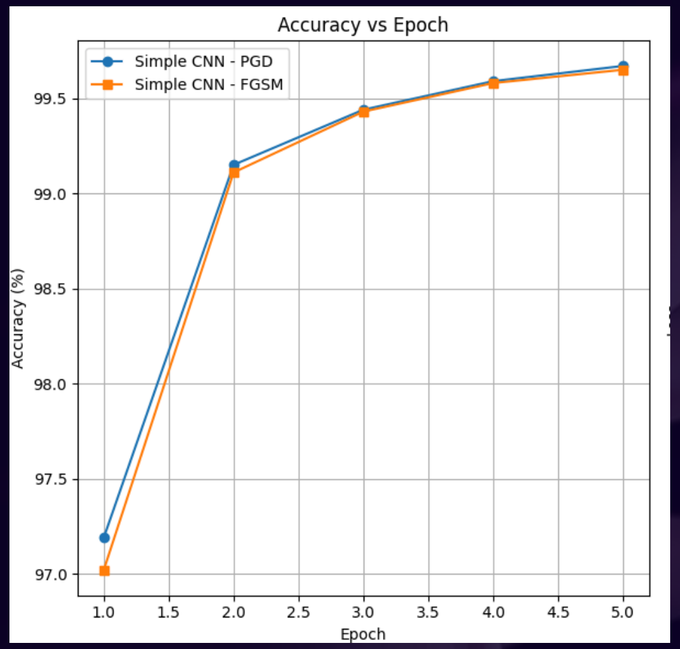
Here we present the experimental results of our handwritten digit recognition model with adversarial robustness. We compared the performance of two models: DeepCNN + PGD Vs SimpleCNN + PGD and DeepCNN + FGSM Vs SimpleCNN + FGSM. The table below shows the accuracy across different epochs. The DeepCNN  model consistently performs better compared to the SimpleCNN model.

| Epoch | DeepCNN-PGD accuracy | SimpleCNN-PGD accuracy |
|---|---|---|
| 1 | 99.69 | 97.19 |
| 2 | 99.72 | 99.15 |
| 3 | 99.70 | 99.44 |
| 4 | 99.73 | 99.59 |
| 5 | 99.68 | 99.67 |
| Test Accuracy | 99.41 | 99.19 |

| Epoch | SimpleCNN-FGSM accuracy | DeepCNN-FGSM accuracy |
|---|---|---|
| 1 | 97.02 | 97.06 |
| 2 | 99.11 | 99.12 |
| 3 | 99.43 | 99.44 |
| 4 | 99.58 | 99.58 |
| 5 | 99.65 | 99.69 |
| Test accuracy | 98.99 | 99.22 |

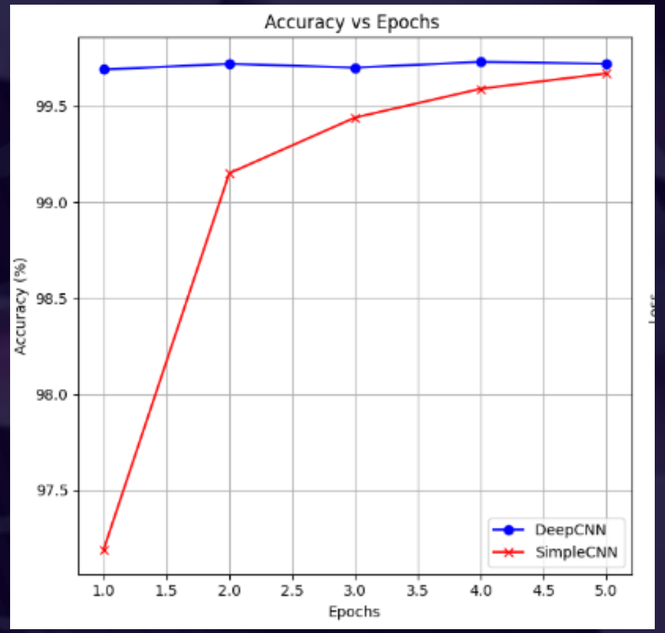
# Results from Graphs

Model Mavericks

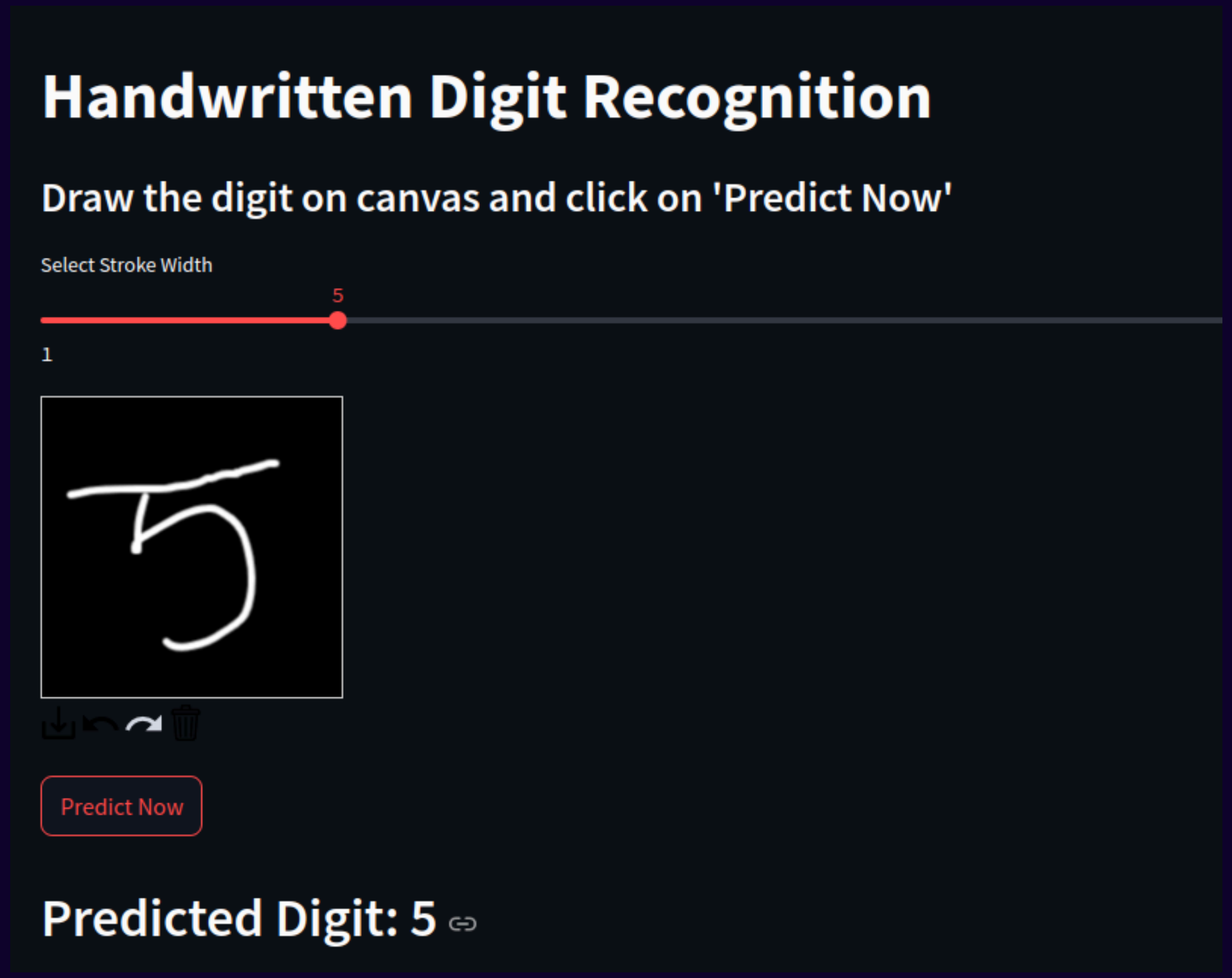



This graph shows the accuracy trends of SimpleCNN and DeepCNN models under PGD adversarial attack. The steeper increase in accuracy shown by DeepCNN at the last epoch(96-->99) outperforms SimpleCNN and hence deeper architecture is better suited for handling adversarial perturbations generated using PGD.

Both SimpleCNN and DeepCNN exhibit high initial accuracy compared to the PGD case, indicating FGSM is relatively weaker perturbation effect. Here as we can see from the graph, both models converge to similar accuracy levels which means that even a simpler architecture can be robust to FGSM with better training.



# Streamlite Predicted Output



# Individual Contributions

***Arey Pragna Sri*** *- Data Downloading, separation of features and targets, PGD attack, Deep CNN Model*

***Matcha Jhansi Lakshmi*** *- FGSM attack, Evaluation, Predicting output*

***Nannepaga Vanaja*** *-CNN Model, Newdataset,Training model*


Model Mavericks
Thank You!