DEEPFAKE DETECTION

Computer Vision - Homework

Sapienza, University of Rome. MSc in AI and Robotics

Olmo Ceriotti
2193258

Andrea Gravili
2180997

OUTLINE

- PROBLEM STATEMENT
- STATE OF THE ART
- PROPOSED METHOD
- DATASET
- EXPERIMENTAL SETUP
- MODEL EVALUATION
- CONCLUSIONS
- REFERENCES

PROBLEM STATEMENT

THE RISE OF DEEP LEARNING

Synthetic media generated by AI (GANs, diffusion models) poses a significant threat to digital content authenticity

ETHICAL AND SECURITY CONCERNS

Deepfakes challenge truth and trust, raising risks of misinformation, identity misuse, and digital manipulation.

GENERALIZATION GAP

Models trained on one dataset often perform poorly on unseen forgery methods or diverse deepfake types due to bias towards "method-specific fake textures"

ADVERSARIAL VULNERABILITY

Detection systems can be deceived by minor, intentional pixel-level perturbations, undermining their reliability

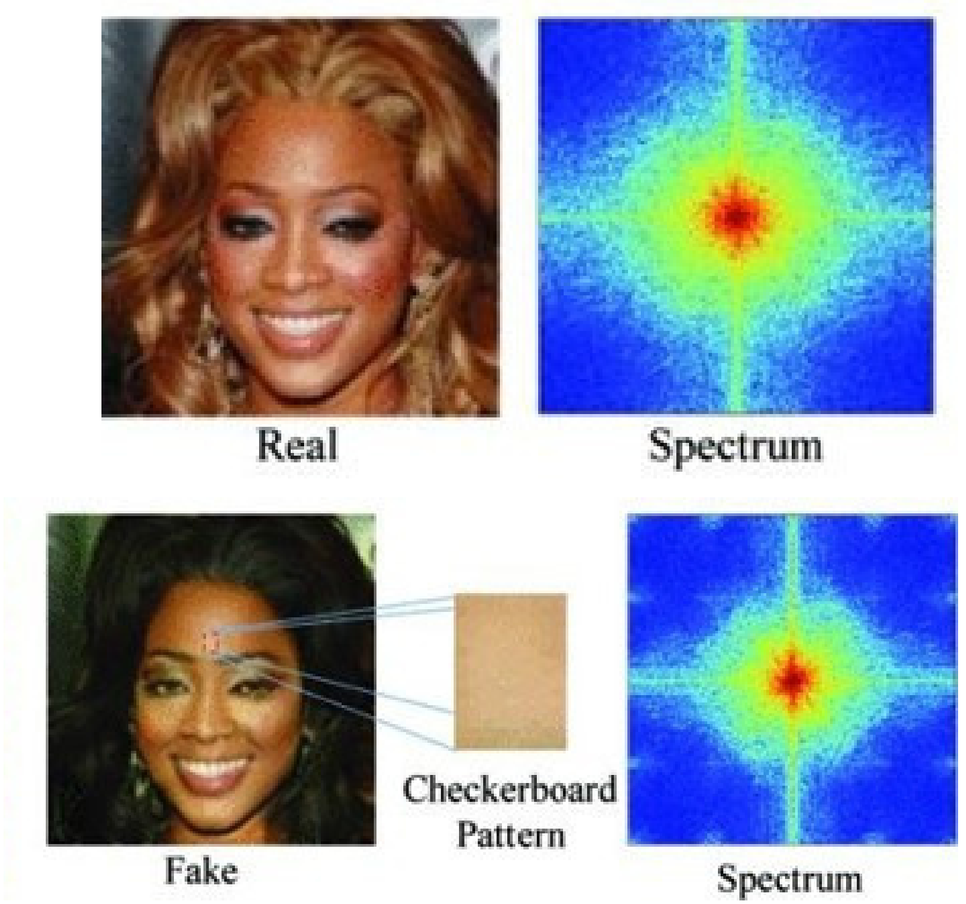




STATE OF THE ART

Current defenses

- Frequency transformed input
- Adversarial training
- DropBlock



PROPOSED METHOD: TRAINING PROCESS

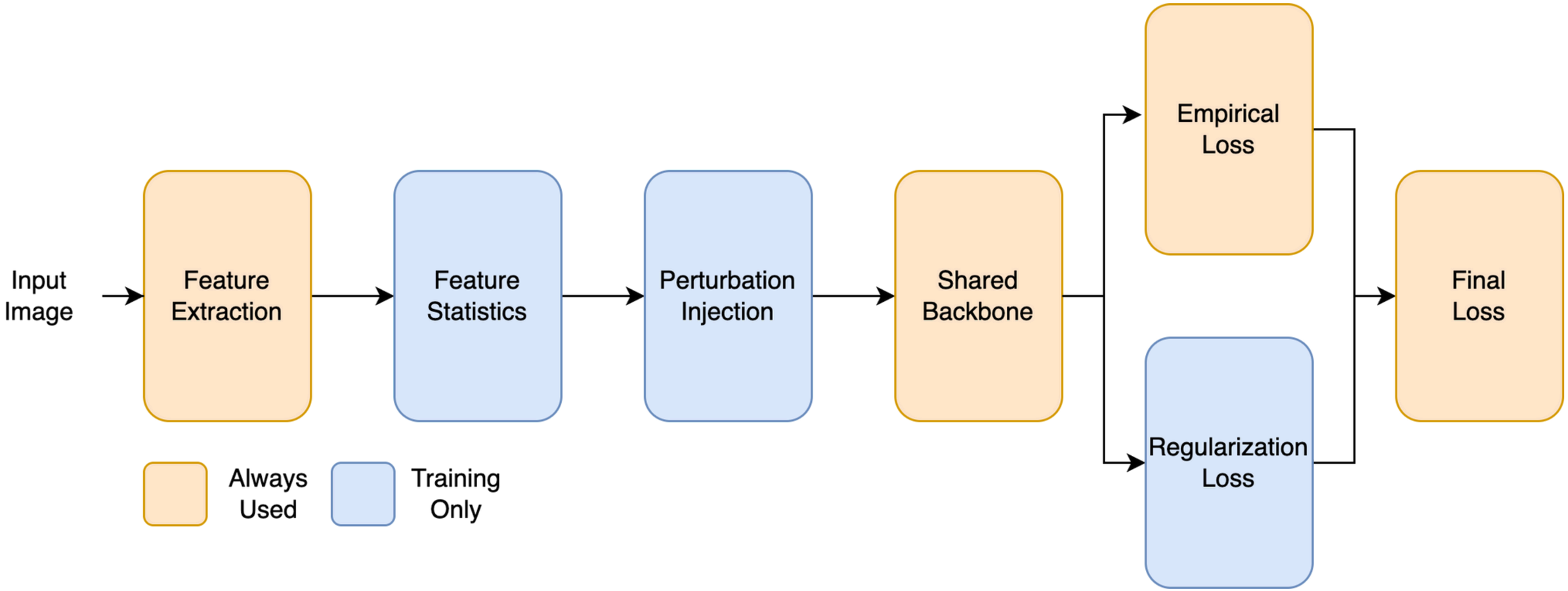
Use of pre-trained CNNs as backbones

EfficientNetB0 (focus)

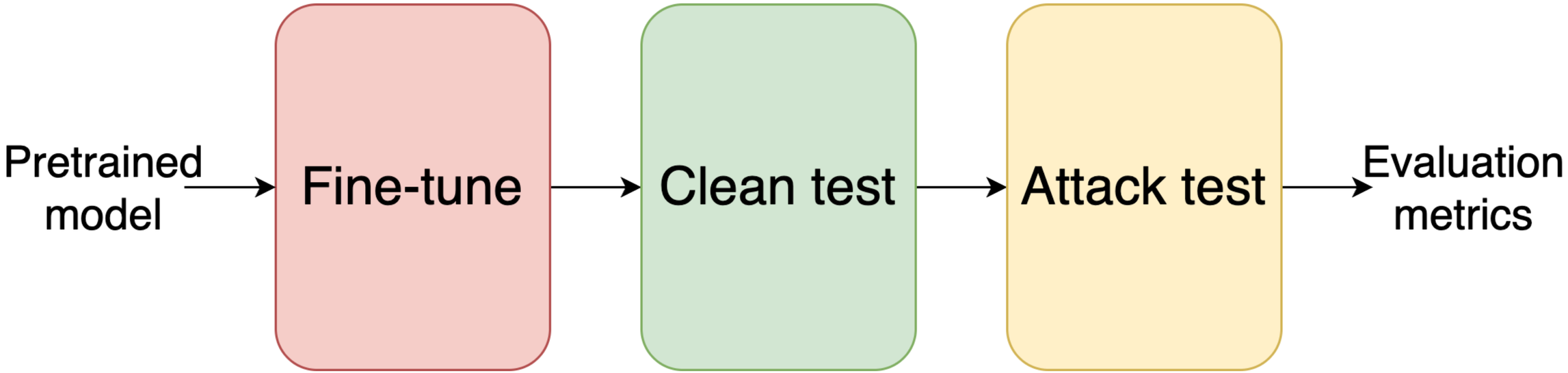
Integration of Gradient Regularization

- Improve generalization, better performance on unseen generation methods.
- Reduce sensitivity to shallow features.
- Uses a gradient based regularization term.

PROPOSED METHOD: GRADIENT REGULARIZATION



PROPOSED METHOD: ATTACKS AND VALIDATION



PROPOSED METHOD: ATTACK TYPES

FGSM

- Applies a single perturbation in the direction of the gradient of the loss
- Computationally weak, easier to defend against

$$\mathbf{x}_{\text{adv}} = \mathbf{x} + \epsilon \cdot \text{sign}\left(\nabla_{\mathbf{x}} L(\mathbf{x}, y)\right)$$

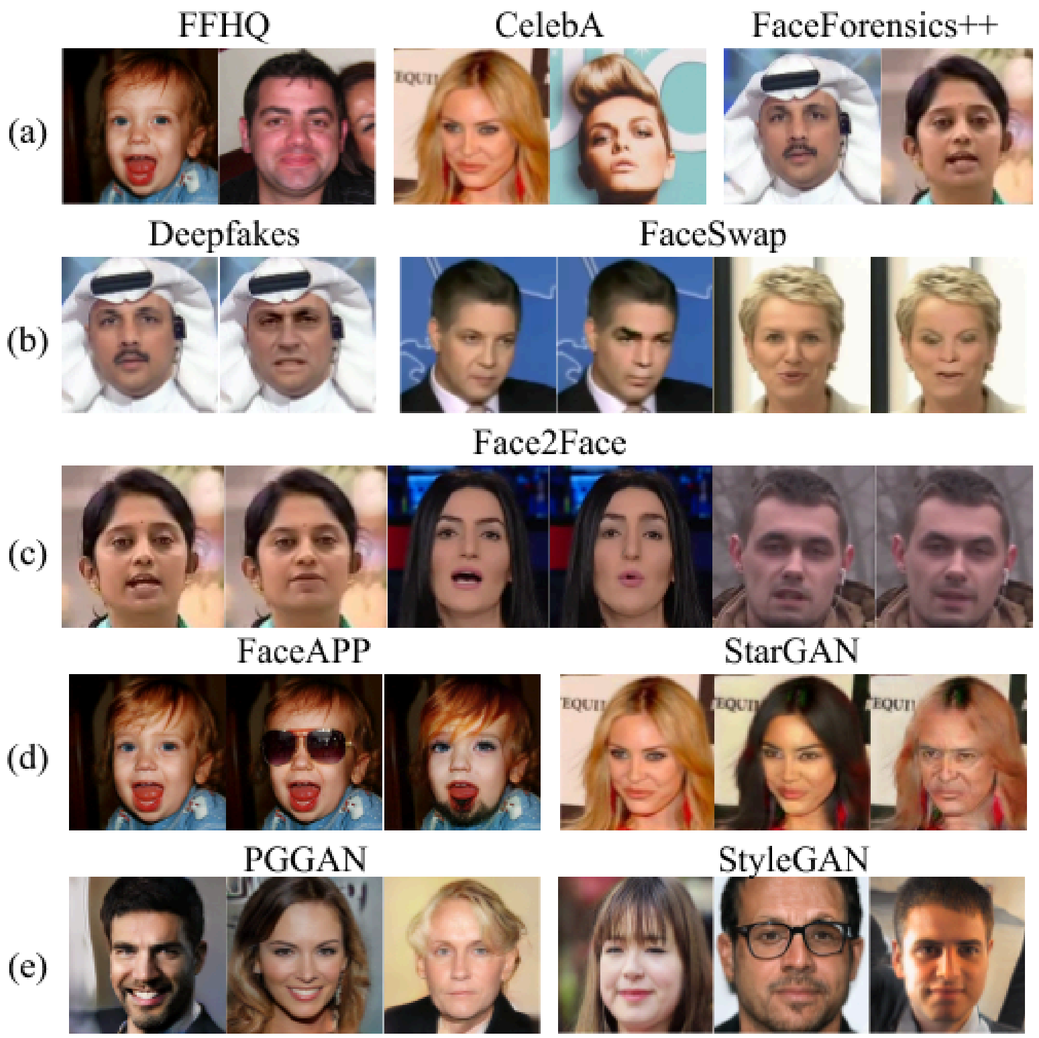
PGD

- Iteratively applies small FGSM-like steps and projects back to stay within a perturbation limit.
- More expensive, more effective

$$x'_0 = x$$

$$x'_{n+1} = \text{Proj}\left\{x'_n + \epsilon \cdot \text{sign}\left(\nabla_x L(x'_n, l)\right)\right\}$$

DATASET



- **Base Datasets**
  - Diverse Fake Faces Dataset (DFFD)
  - Original size: ≈300.000 images
  - Used sample size: ≈40.000 images
- **Data Structure**
  - Organized in: True dataset from ffhq and fake dataset from styleGAN ffhq .
  - OOD dataset created combining fake photos from DFFD.
- **Preprocessing**
  - Images resized to 224x224 pixels
  - Convertend to PyTorch Tensors
  - Normalization

EXPERIMENTAL SETUP

Models tested

- EfficientNetB0
- EfficientNetB0 with PIM
- EfficientNetB0 with frequency transform
- EfficientNetB0 with DropBlock

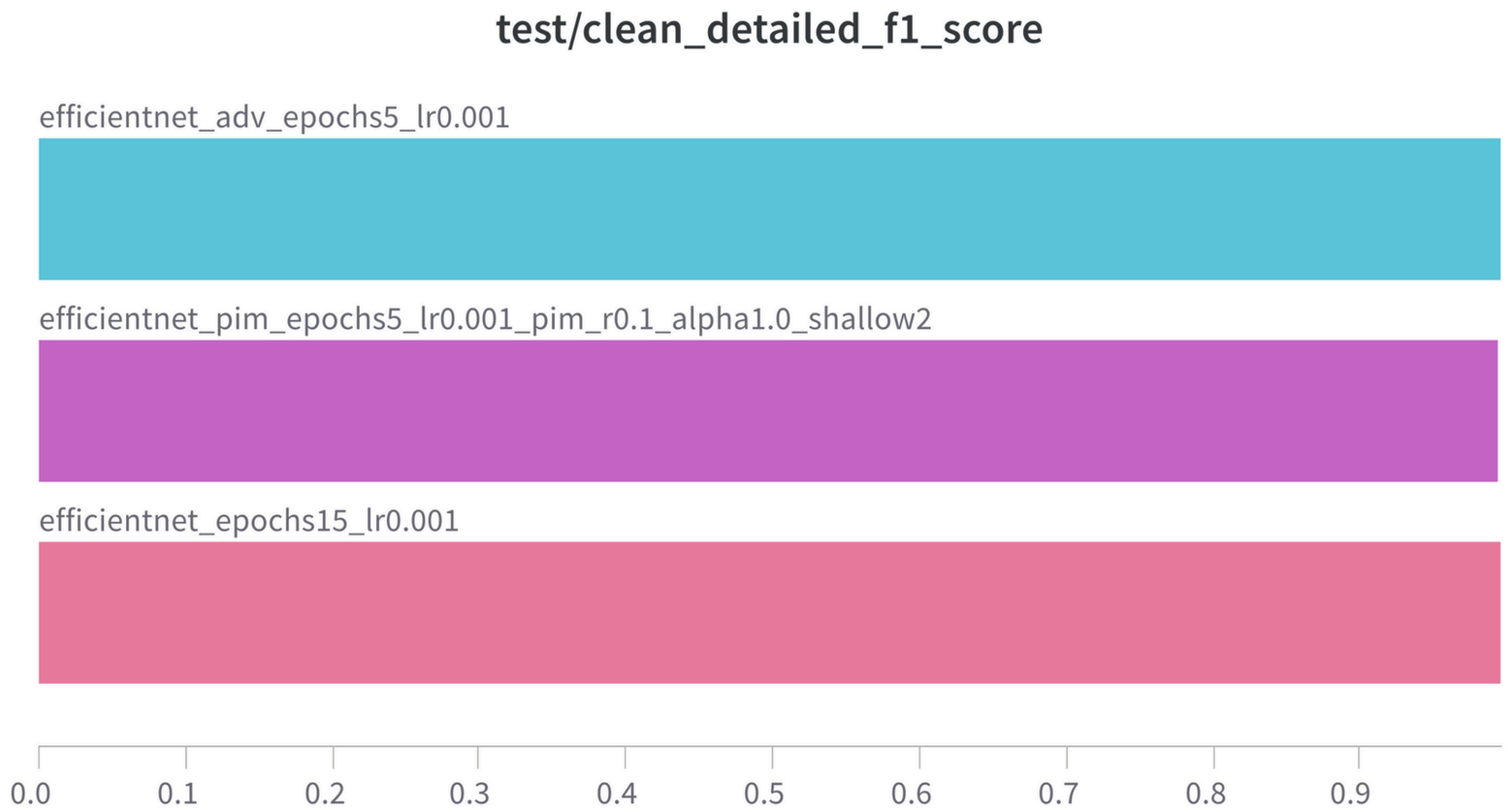
Training parameters

- Pre-training: All models initialized with ImageNet pre-trained weights.
- Optimizer: AdamW
- Learning Rate: 0.001
- Loss Function: cross entropy
- Epochs: 2
- Batch Size: 32

pim hyperparameters

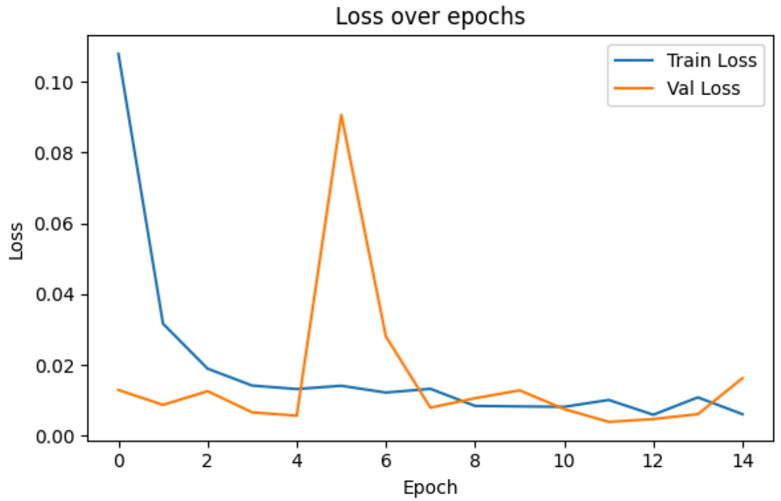
- R_PIM: 0.1 (Controls magnitude of perturbation)
- ALPHA_PIM: 1.0 (Weight of regularization loss)
- SHALLOW_FEATURE_IDX: 2 (Determines the split point for shallow features)

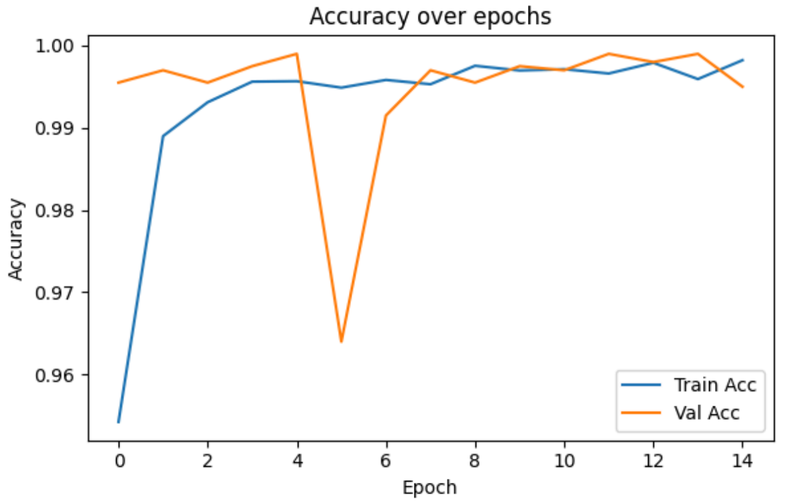
MODEL EVALUATION (CLEAN DATA)



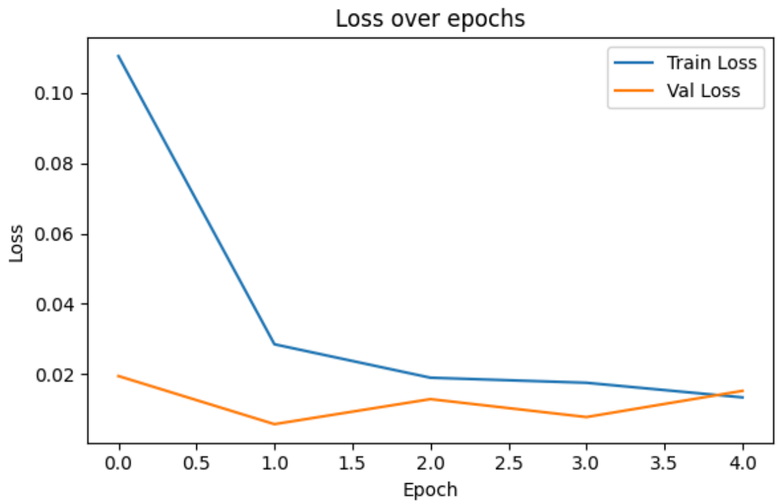
MODEL EVALUATION (CLEAN DATA)

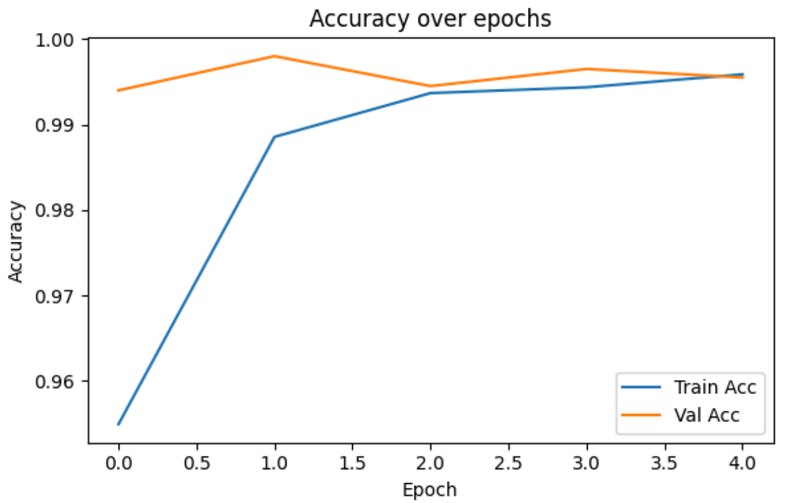
Standard



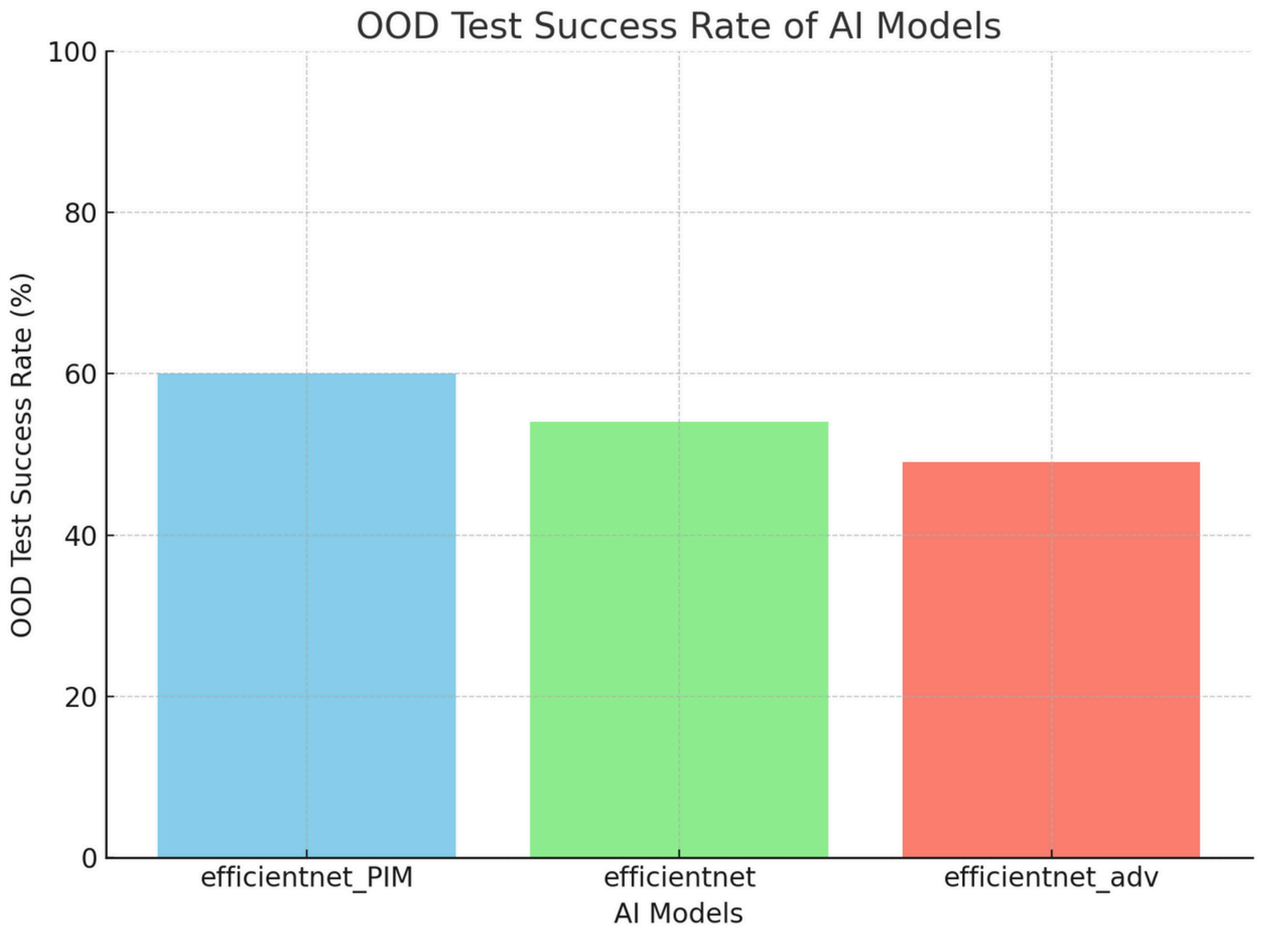


Regularized

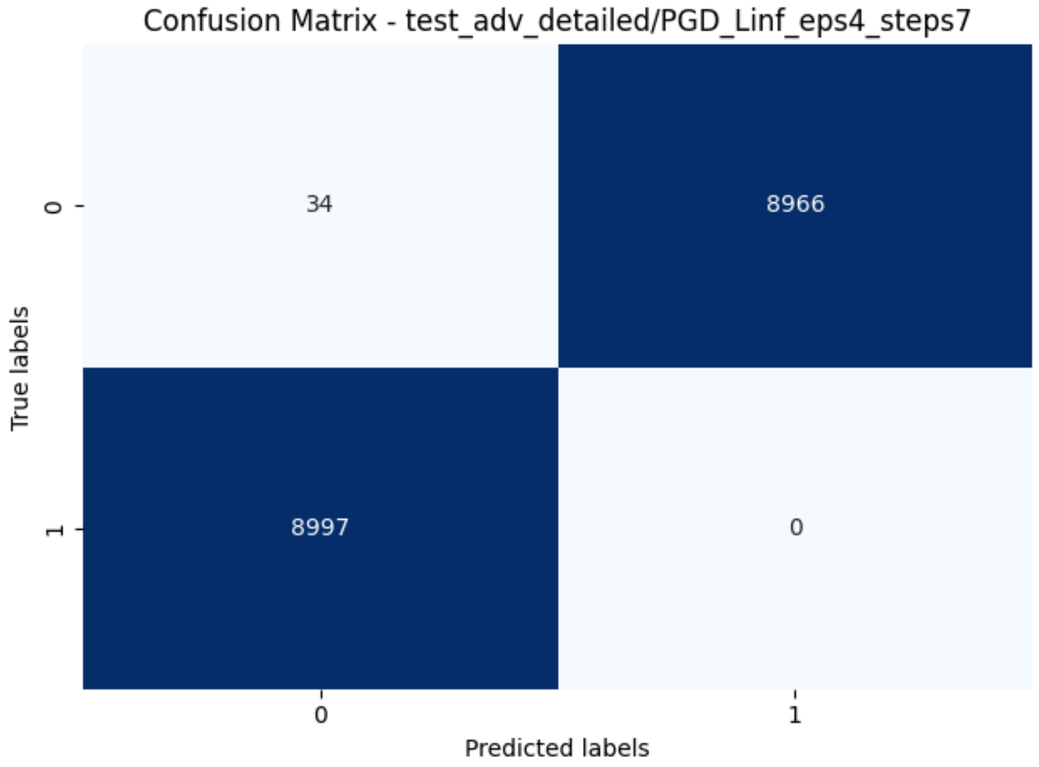




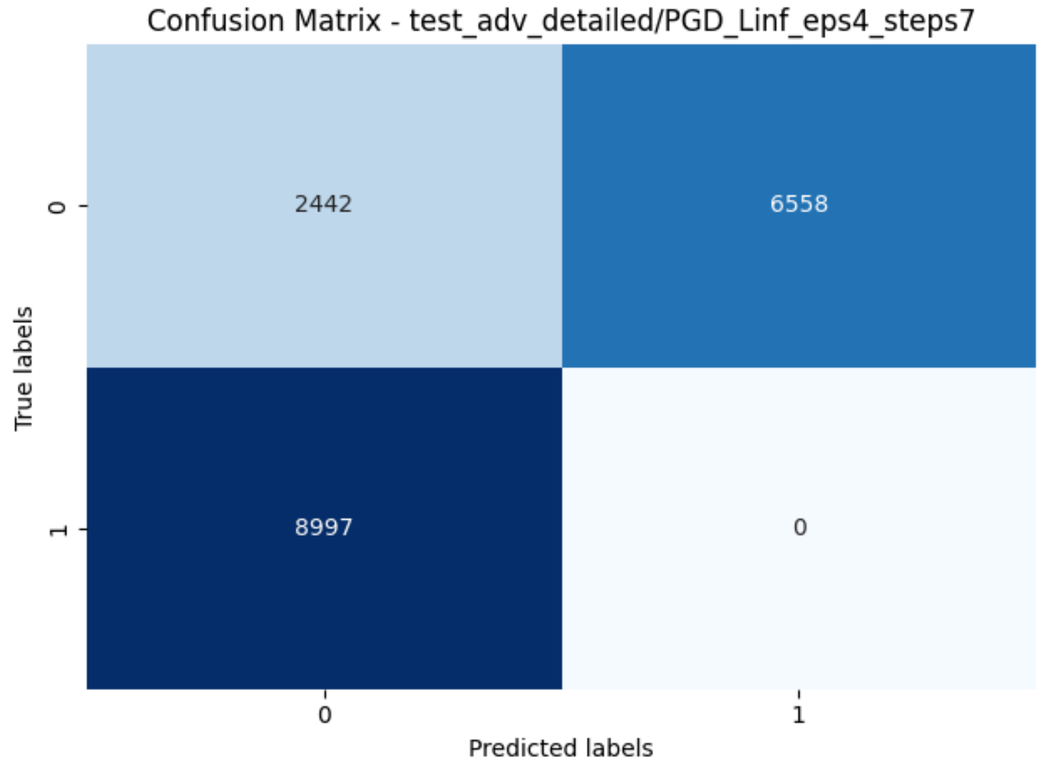
MODEL EVALUATION (CLEAN DATA)



MODEL EVALUATION (ADVERSARIAL ROBUSTNESS)

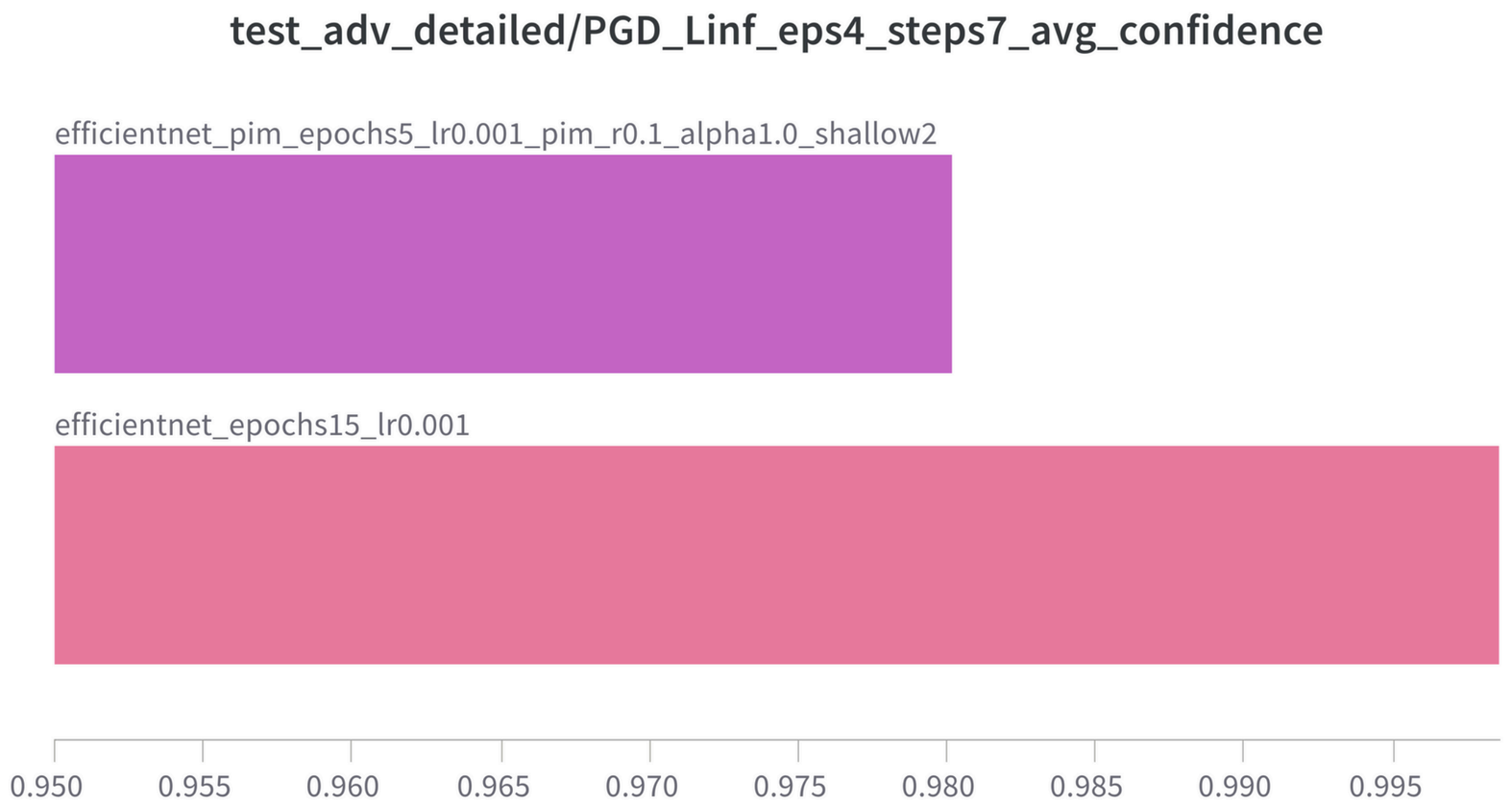


Standard

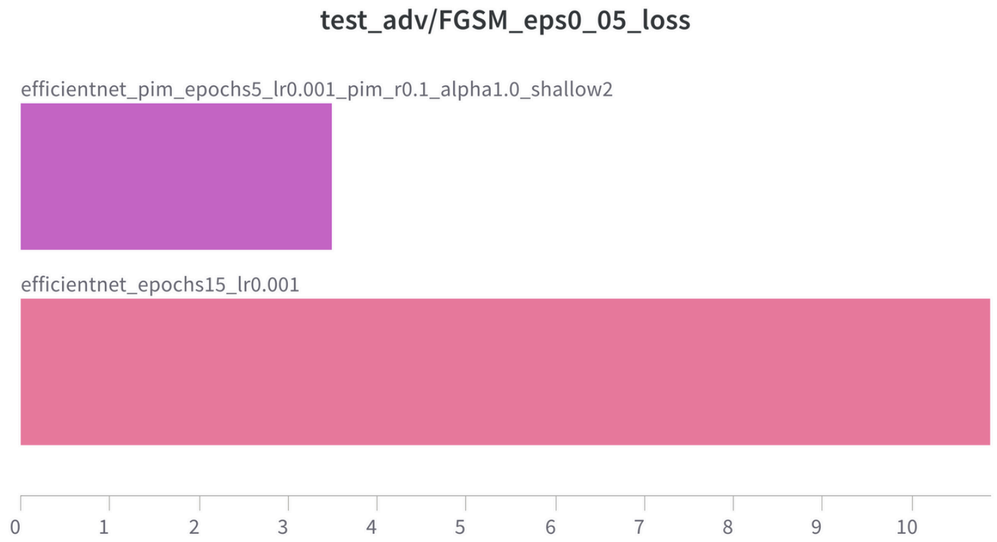


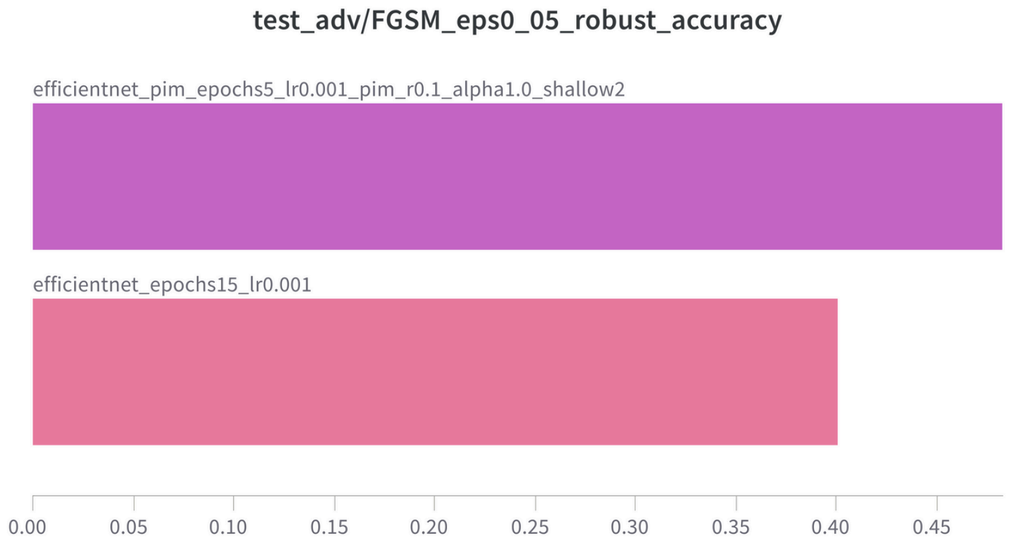
Regularized

MODEL EVALUATION (ADVERSARIAL ROBUSTNESS)

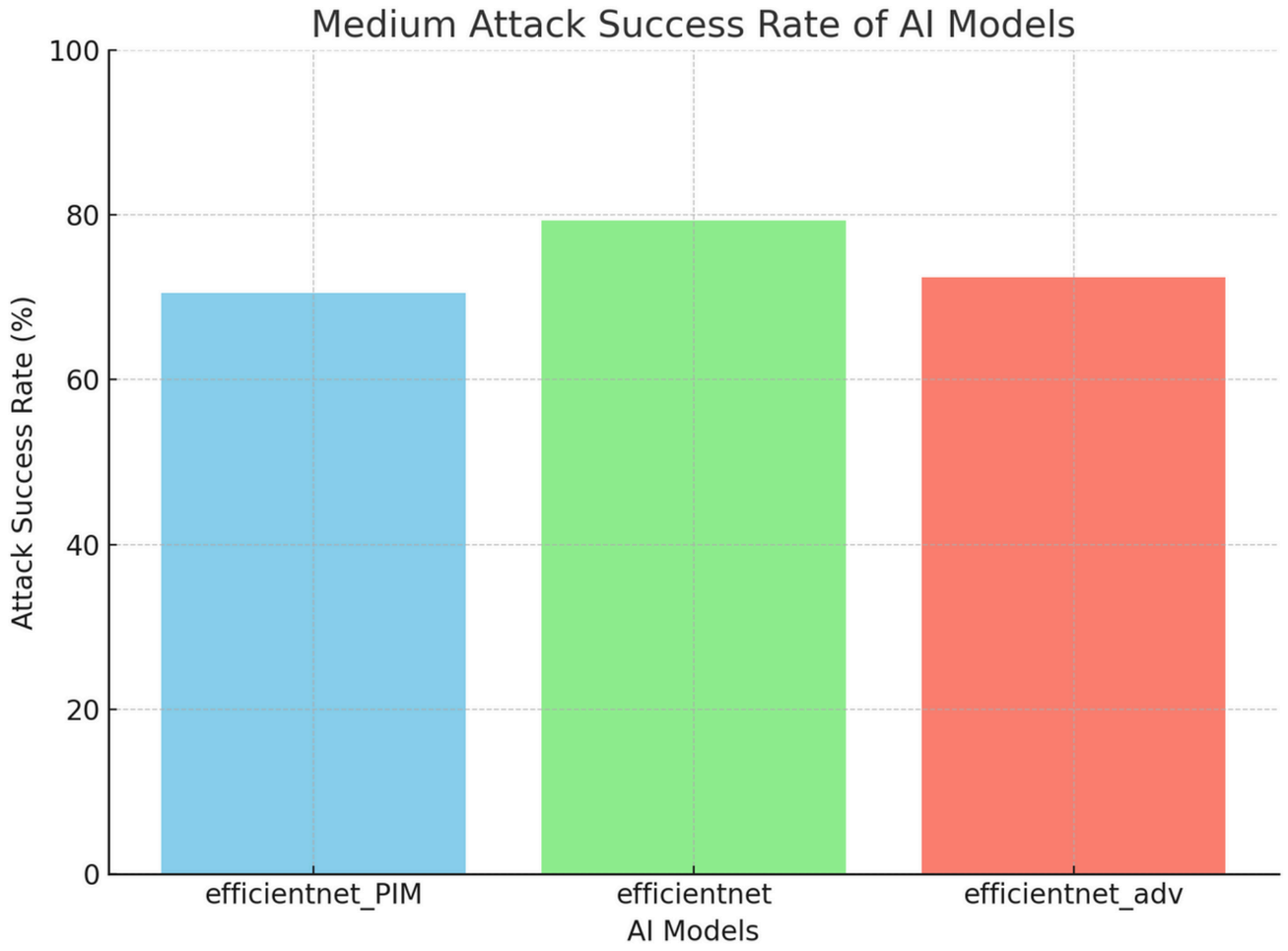


MODEL EVALUATION (ADVERSARIAL ROBUSTNESS)





MODEL EVALUATION (ADVERSARIAL ROBUSTNESS)



Conclusions

- **Key Findings**
  - Improved robustness with PIM
  - Better generalization capabilities on unseen generators
  - Improved stability during training
- **Limitations of Current Work**
  - Limited attack resistance
- **Future Work**
  - Adversarial training with PIM
  - Adaptive gradient regularization

References

- W. Guan, W. Wang, J. Dong and B. Peng, (2024). Improving Generalization of Deepfake Detectors by Imposing Gradient Regularization, In IEEE Transactions on Information Forensics and Security, vol. 19, pp. 5345-5356.
- M. Tan and Q. Le, (2019). EfficientNet: Rethinking model scaling for convolutional neural networks. In Proc. Int. Conf. Mach. Learn., pp. 6105–6114.
- 3. On the Detection of Digital Face Manipulation Hao Dang, Feng Liu, Joel Stehouwer, Xiaoming Liu, Anil Jain, (2020), In: Proceedings of IEEE Computer Vision and Pattern Recognition (CVPR 2020), Seattle, WA, Jun. 2020
- 4. Abbasi, M., V´az, P., Silva, J. and Martins, P. (2025). Comprehensive Evaluation of Deepfake Detection Models: Accuracy, Generalization, and Resilience to Adversarial Attacks. Applied Sciences, 15(3), 1225.