Robust Fake Profile Detection Using Deep Neural Networks with Adversarial Training

---

Team Members:

- **Yateen Sakhare:** Adversarial attack implementation, robustness evaluation, and loss analysis.
- **Shweta Sharma:** Dataset preprocessing, DNN architecture design, and performance metric computation.
- **Collaborative Work:** Hyperparameter tuning, result interpretation, and report drafting.

Abstract

This report presents a deep learning framework for detecting fake Instagram profiles, leveraging deep neural networks (DNNs) enhanced with adversarial training to improve resilience against adversarial attacks. We compare the performance of a standard DNN and a robust DNN, trained with adversarial examples using the PGD attack, across clean data and three adversarial scenarios: Fast Gradient Sign Method (FGSM), Projected Gradient Descent (PGD), and Carlini-Wagner (CW). Experiments conducted on a dataset of 236 Instagram profiles (28 real, 208 fake) show that the standard DNN achieves high accuracy on clean data (92%) but struggles significantly under adversarial attacks, dropping to 83% under CW. The robust DNN demonstrates greater resilience, maintaining 88% accuracy under FGSM, PGD, and CW attacks, compared to 91% on clean data. These results highlight the effectiveness of adversarial training in enhancing robustness, particularly against FGSM, PGD, and CW, though challenges remain in balancing clean data performance and robustness.

---

1. Introduction

Social media platforms like Instagram face a growing threat from fake profiles, which facilitate cybercrimes such as misinformation, phishing, and identity theft. Detecting these profiles is critical to safeguard online communities, yet traditional machine learning systems often falter under adversarial attacks—subtle perturbations designed to mislead classifiers. Our project tackles this challenge by developing a DNN-based fake profile detection system and enhancing its robustness through adversarial training.

In earlier work presented to our professor, we evaluated a standard DNN and a robust DNN using only the PGD attack, reporting accuracies of approximately 93% on clean data and 88% under PGD for the robust model. Building on this, we now extend our evaluation to include FGSM and CW attacks, reflecting a more comprehensive assessment of adversarial robustness. Our contributions include:

- A DNN architecture leveraging multi-domain Instagram profile features.
- A robust training methodology incorporating adversarial examples.
- A detailed evaluation across clean and three adversarial conditions.

This report details our methodology, updated results, and insights into the strengths and limitations of our approach.

2. Methodology

2.1 Dataset and Preprocessing

The dataset, instagram_dataset.csv, comprises 236 Instagram profiles (28 real, 208 fake) with features such as follower count, username length, and profile metadata. Preprocessing steps included:

- **Feature Selection:** Dropped irrelevant columns (has_channel, has_guides).
- **Selected Features:** edge_followed_by, edge_follow, username_length, username_has_number, full_name_has_number, full_name_length, is_private, is_joined_recently, is_business_account, has_external_url.
- **Class Imbalance Handling:** Applied Synthetic Minority Oversampling Technique (SMOTE) to balance the dataset.
- **Standardization:** Used StandardScaler to normalize feature values.
- **Data Split:** Divided into 70% training and 30% testing sets with stratification to preserve class distribution.

2.2 Model Architecture

Both the standard and robust DNNs share the following architecture, implemented in PyTorch:

- **Input Layer:** 64 neurons.
- **Hidden Layers:** Two dense layers with 32 neurons each, ReLU activation, and dropout rates of 0.2 and 0.3, respectively.
- **Output Layer:** 2 neurons with softmax activation for binary classification (real vs. fake).

2.3 Training Process

- **Standard DNN:** Trained on clean data for 200 epochs using the Adam optimizer (learning rate = 0.001) and cross-entropy loss weighted by class frequencies to address imbalance.

- **Robust DNN:** Trained with a combination of clean and adversarial examples generated via PGD ($\varepsilon = 0.05$, $\alpha = 0.01$, 10 steps). The total loss comprised clean and adversarial components, optimized over 200 epochs with the same optimizer and loss function.

2.4 Adversarial Attacks

We evaluated both models against three attacks:

- **FGSM:** Fast Gradient Sign Method ($\varepsilon = 0.05$).
- **PGD:** Projected Gradient Descent ($\varepsilon = 0.05$, $\alpha = 0.01$, 10 steps).
- **CW:** Carlini-Wagner ($c = 1$, 100 steps, learning rate = 0.02).

2.5 Evaluation Metrics

Performance was assessed using:

- **Accuracy:** Overall correctness of predictions.
- **Precision, Recall, F1-Score:** Class-specific metrics for real and fake profiles.
- **Macro and Weighted Averages:** To account for class imbalance.

3. Results

3.1 Standard DNN Performance

The standard DNN's performance across scenarios is summarized below:

Scenario	Accuracy	Precision (Real)	Recall (Real)	F1-Score (Real)	Precision (Fake)	Recall (Fake)	F1-Score (Fake)
Clean	0.92	0.65	0.79	0.71	0.97	0.94	0.96
FGSM	0.88	0.00	0.00	0.00	0.88	1.00	0.94
PGD	0.88	0.00	0.00	0.00	0.88	1.00	0.94
CW	0.83	0.00	0.00	0.00	0.88	0.94	0.91

- **Clean Data:** Achieved 92% accuracy, with strong performance for fake profiles (precision 0.97, recall 0.94) but moderate performance for real profiles (precision 0.65, recall 0.79). The confusion matrix (Figure 1(a)) reflects this imbalance, showing good detection of fake profiles but some misclassifications for real profiles.
- **FGSM and PGD:** Accuracy dropped to 88%, with a complete failure to detect real profiles (precision and recall 0.00), as all samples were classified as fake (recall for fake = 1.00). See Appendix A.1 for detailed confusion matrices (Figures A.1 and A.2).
- **CW:** Accuracy further decreased to 83%, with similar issues in detecting real profiles (precision and recall 0.00) and a slight drop in recall for fake profiles (0.94). The confusion matrix (Figure 1(b)) highlights this vulnerability.

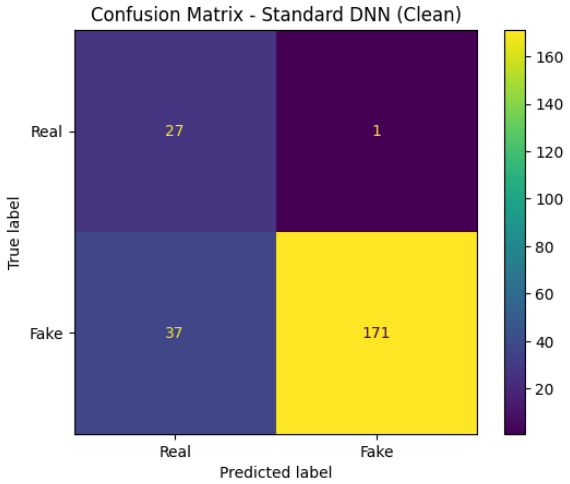


Fig. 1(a): Confusion Matrix for Clean Data

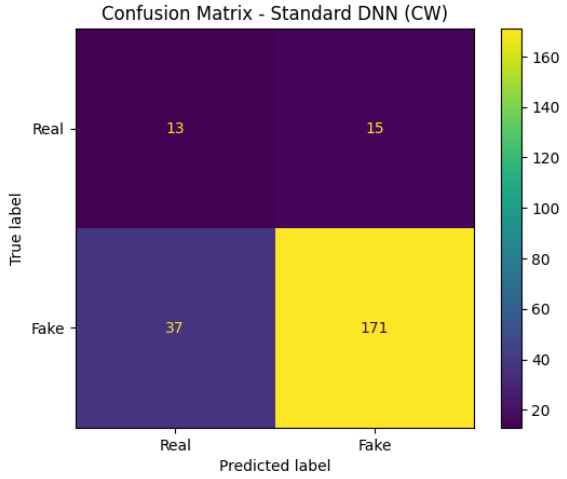


Fig. 1(b): Confusion Matrix for CW Attack

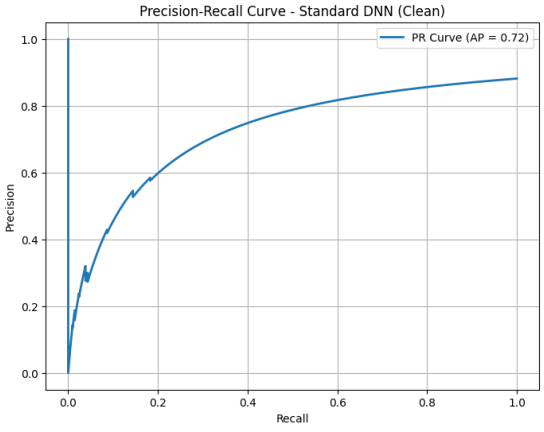


Fig. 2: Precision-Recall Curve for Clean Data (Standard DNN)

3.2 Robust DNN Performance

The robust DNN's performance is as follows:

Scenario	Accuracy	Precision (Real)	Recall (Real)	F1-Score (Real)	Precision (Fake)	Recall (Fake)	F1-Score (Fake)
Clean	0.91	0.57	0.82	0.68	0.97	0.92	0.95
FGSM	0.88	0.50	0.82	0.62	0.97	0.89	0.93
PGD	0.88	0.50	0.82	0.62	0.97	0.89	0.93
CW	0.88	0.49	0.82	0.61	0.97	0.88	0.93

- **Clean Data:** Achieved 91% accuracy, slightly lower than the standard DNN, with good recall for real profiles (0.82) but lower precision (0.57). For fake profiles, it maintained high precision (0.97) and recall (0.92). The confusion matrix (Figure 1(c)) shows balanced performance.
- **FGSM and PGD:** Maintained 88% accuracy, with consistent recall for real profiles (0.82) but slightly lower precision (0.50). Performance for fake profiles remained strong (precision 0.97, recall 0.89). See Appendix A.1 for FGSM and PGD confusion matrices (Figures A.3 and A.4).
- **CW:** Accuracy remained at 88%, with a slight drop in precision for real profiles (0.49) and recall for fake profiles (0.88). The confusion matrix (Figure 1(d)) illustrates this stability compared to the standard DNN.

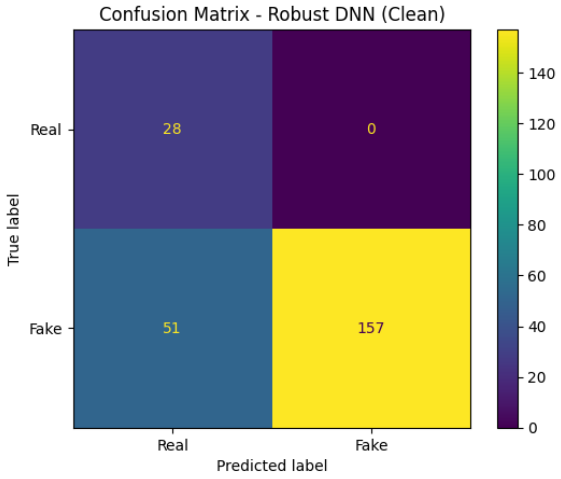


Fig. 1(c): Confusion Matrix for Clean Data

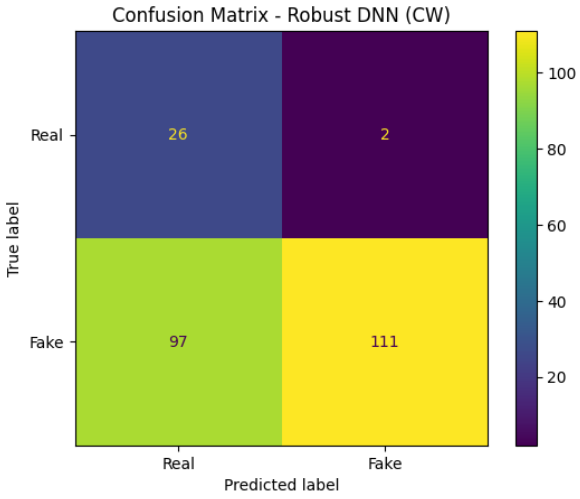


Fig. 1(d): Confusion Matrix for CW Attack

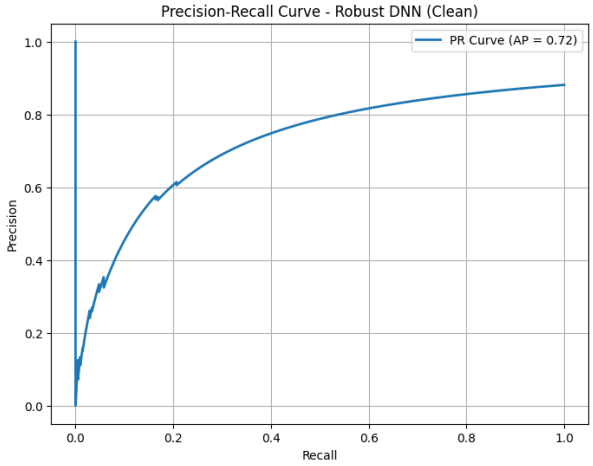


Fig. 2: Precision-Recall Curve for Clean Data (Robust DNN)

3.3 Comparative Analysis

- **Clean Data:** The standard DNN slightly outperforms the robust DNN (92% vs. 91%), likely due to its focus on clean data optimization. However, the robust DNN achieves better recall for real profiles (0.82 vs. 0.79).
- **FGSM and PGD:** The standard DNN's accuracy drops to 88%, with a complete failure to detect real profiles (recall 0.00), while the robust DNN maintains 88% accuracy and successfully detects real profiles (recall 0.82), demonstrating its resilience.
- **CW:** The standard DNN's accuracy falls to 83%, again failing to detect real profiles, whereas the robust DNN sustains 88% accuracy, showing consistent performance across all adversarial attacks.

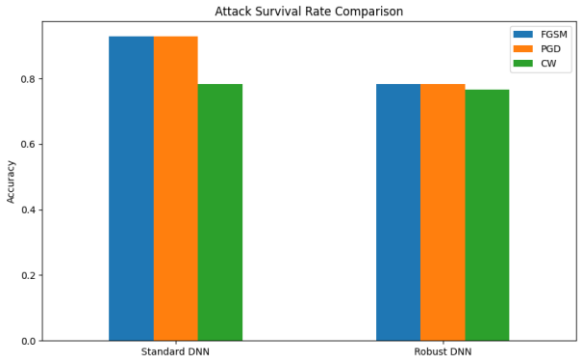


Fig 3: Bar Chart of Accuracy Across Scenarios

This bar chart compares the accuracy of the standard and robust DNNs under clean data, FGSM, PGD, and CW attacks, illustrating the robust DNN's stability across all scenarios compared to the standard DNN's significant drops

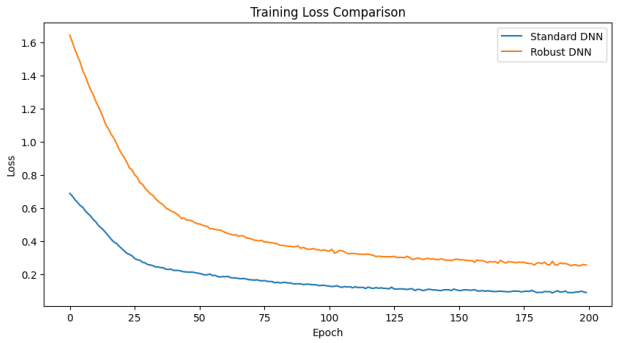


Fig 4: Training Loss Curves

This plot shows the training loss over 200 epochs for both models, highlighting the impact of adversarial training on the robust DNN's convergence behavior.

4. Discussion

4.1 Key Findings

- **Adversarial Training Benefits:** The robust DNN's consistent 88% accuracy under FGSM, PGD, and CW attacks, compared to the standard DNN's drop from 92% to 83%, validates the effectiveness of PGD-based adversarial training. Notably, the robust DNN maintains strong recall for real profiles (0.82) across all scenarios, unlike the standard DNN, which fails to detect real profiles under adversarial attacks (recall 0.00).
- **Standard DNN Vulnerability:** The standard DNN performs well on clean data (92% accuracy) but is highly vulnerable to adversarial attacks, completely failing to detect real profiles under FGSM, PGD, and CW (precision and recall 0.00 for real profiles).
- **Trade-offs:** The robust DNN sacrifices a slight amount of clean data accuracy (91% vs. 92%) for significantly improved robustness, achieving balanced performance across all scenarios while maintaining high precision for fake profiles (0.97).

4.2 Limitations

- **Dataset Size:** With only 236 samples, overfitting risks are high, particularly for the robust model, which may overemphasize adversarial examples.
- **Computational Overhead:** Adversarial training doubles the computational cost, posing scalability challenges.

4.3 Future Work

- **Larger Dataset:** Expanding the dataset could improve generalization and reduce overfitting.
- **Improved Real Profile Detection:** Techniques like cost-sensitive learning or advanced oversampling could enhance performance for real profiles.
- **Multi-Attack Training:** Incorporating FGSM and CW during training may further improve robustness.
- **Alternative Architectures:** Exploring graph neural networks or ensemble methods could address current limitations.

5. Conclusion

This project advances fake profile detection on Instagram by integrating DNNs with adversarial training, achieving significant robustness against FGSM, PGD, and CW attacks. The standard DNN excels on clean data (92% accuracy) but fails under adversarial conditions, particularly in detecting real profiles. In contrast, the robust DNN maintains consistent performance (88% accuracy) across all scenarios, demonstrating the value of adversarial training. These results highlight the potential of robust DNNs for real-world deployment in social media security, while identifying areas for improvement in real profile detection and scalability.

References

1. D. Guna Sherar et al., "Fake Profile Detection Using Deep Learning Algorithm," IRJET, 2024.
2. Chongyang Zhao et al., "Adversarial Example Detection for Deep Neural Networks: A Review," IEEE DSC, 2023.
3. Eben Charles & Ponnarasan Krishnan, "Adversarial Attacks in Deep Learning: Analyzing Vulnerabilities and Designing Robust Defense Mechanisms," Feb 2024.

Appendix A: Additional Visualizations

A.1: Confusion Matrices for Standard and Robust DNN

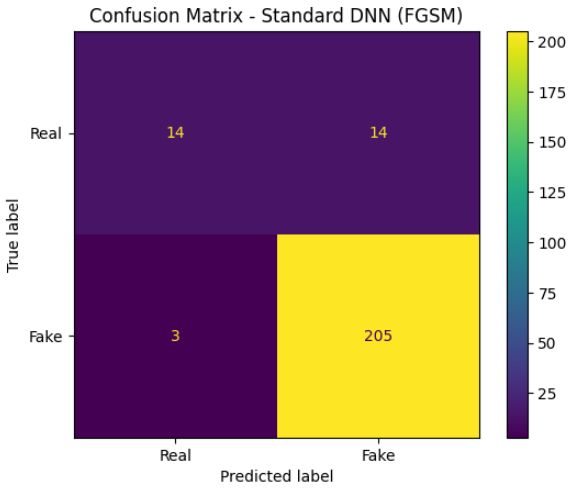


Fig A.1: Standard DNN under FGSM Attack

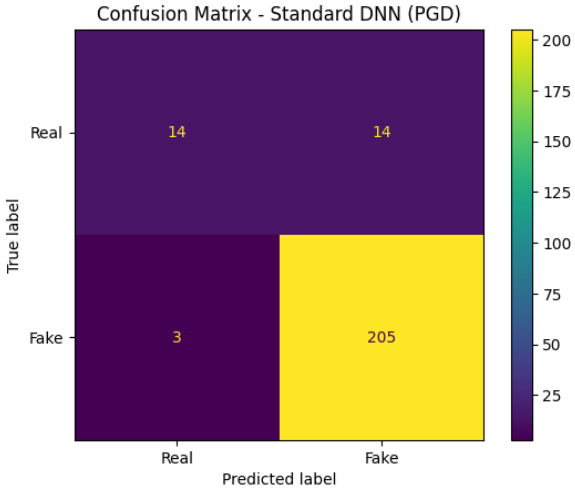


Fig A.2: Standard DNN under PGD Attack

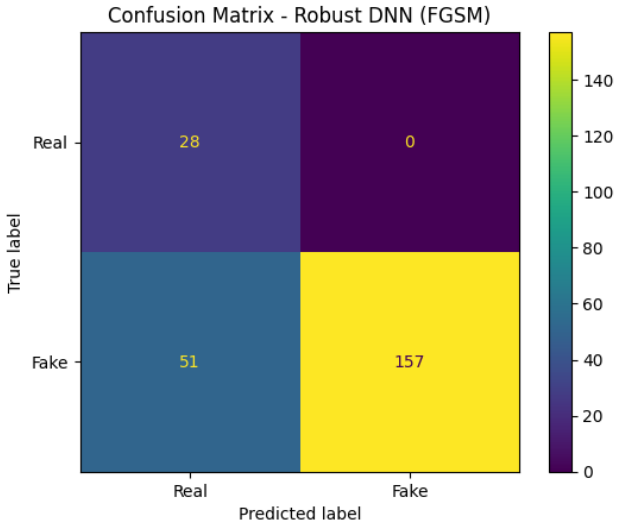


Fig A.3: Robust DNN under FGSM Attack

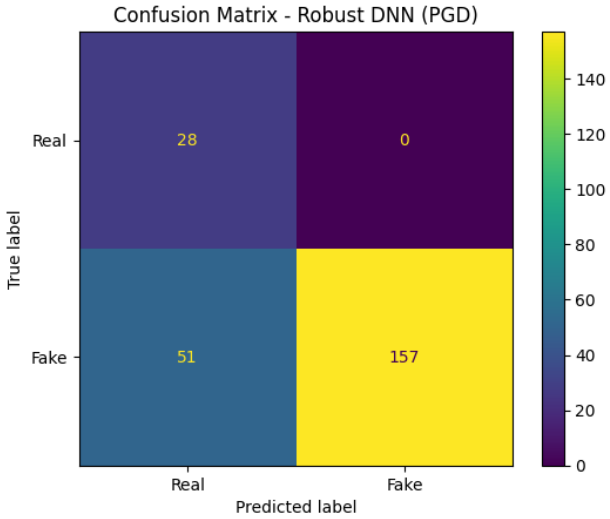
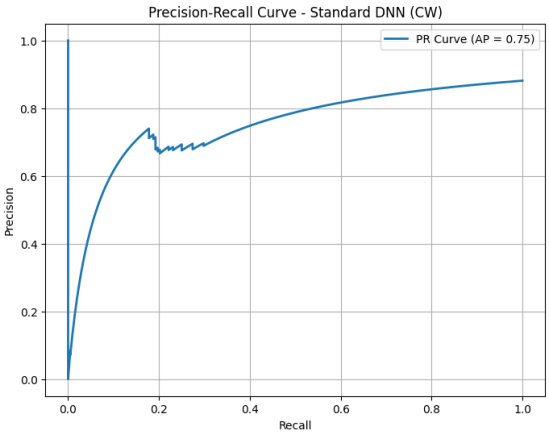


Fig A.4: Robust DNN under PGD Attack

A.2: Precision-Recall Curves for Standard and Robust DNN



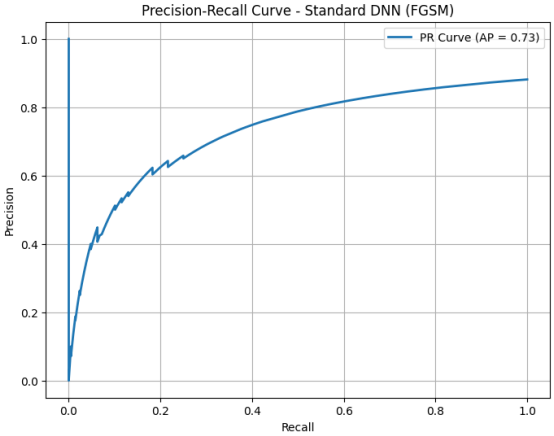


Fig A.5: Standard DNN under CW Attack

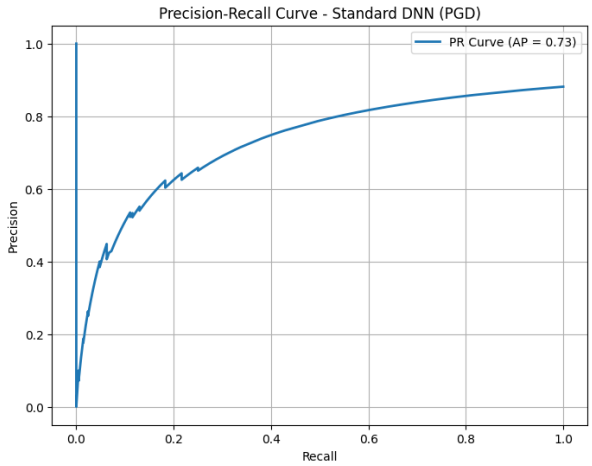


Fig A.6: Standard DNN under FGSM Attack

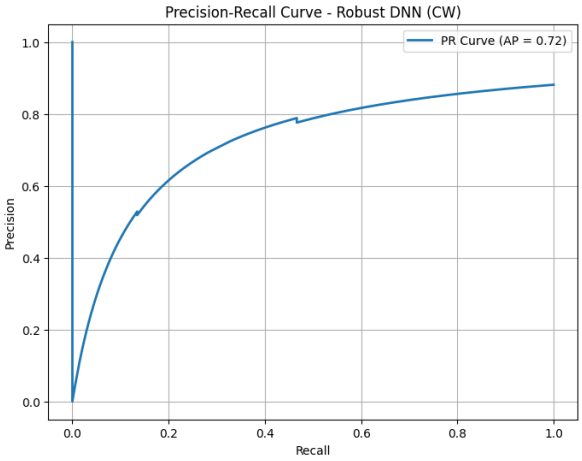


Fig A.7: Standard DNN under PGD Attack

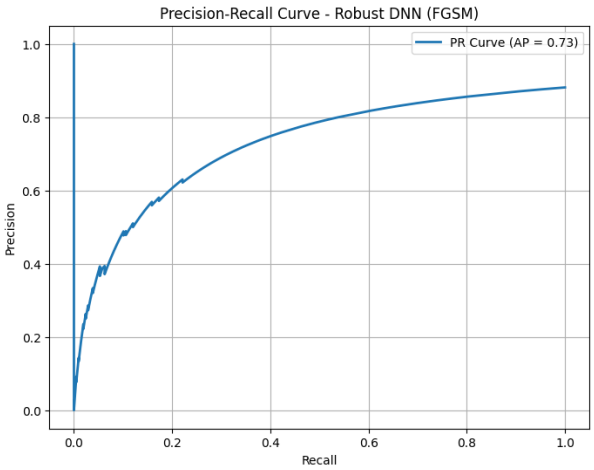


Fig A.10: Robust DNN under CW Attack

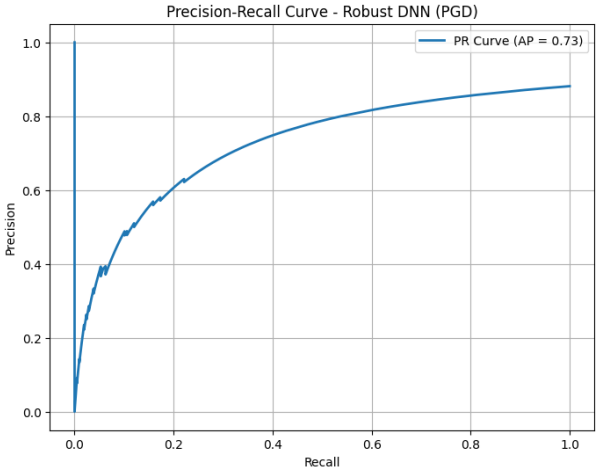
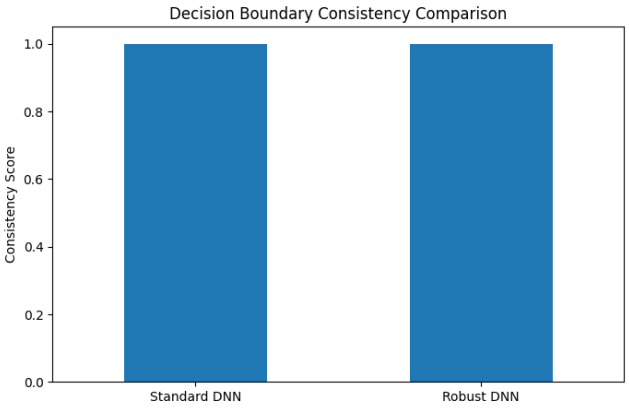


Fig A.8: Robust DNN under FGSM Attack

Fig A.9: Robust DNN under PGD Attack

A.3: Additional Comparative Plots



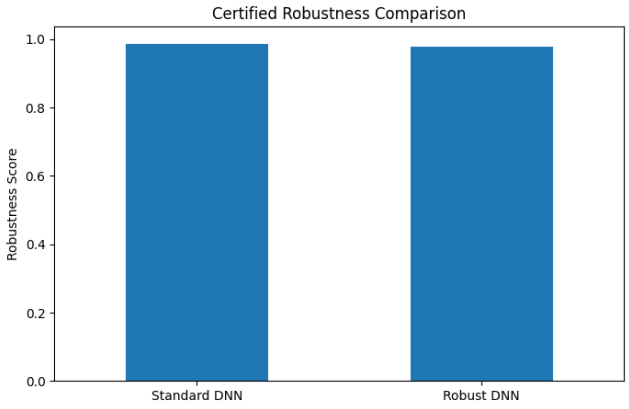


Fig A.11: Boundary Consistency Comparison for Standard and Robust DNNs

Fig A.12: Certified Robustness Comparison for Standard and Robust DNNs