Innovations in CNN Architectures for Image Processing

Omar Khalil

Paper Hypothesis

Problem:

CNNs are highly vulnerable to adversarial attacks (e.g., FGSM attacks), leading to confident misclassifications and catastrophic accuracy drops (e.g., VGG-16 accuracy dropped from 86.5% to <10%).

Existing detection methods often require expensive retraining or significantly degrade performance on clean data.

Solution:

Propose a Non-Invasive, Entropy-Based Monitoring Framework,. This system operates in parallel to the pre-trained CNN, monitoring internal information flow rather than just output confidence.

Limitation

1. **Limited Architectural Validation**
   - Evaluates entropy on a single CNN style
   - Claims architecture independence without strong empirical evidence
2. **Entropy Is Observational, Not Actionable**
   - Entropy is used only as a diagnostic signal
   - No mechanism to Reject unreliable predictions
3. **Weak Adversarial Evaluation**
   - Focuses primarily on FGSM
   - Limited analysis of stronger, iterative attacks

Proposed Framework

Core Hypothesis
Entropy is a reliable uncertainty signal only when supported by strong representations and architecture-aware calibration.

Rather than treating entropy as a passive diagnostic, we **elevate it into an active control mechanism**.

Proposed Contributions

Architecture-Aware Reliability Monitoring

- Extend entropy monitoring beyond a single CNN
- Validate behavior across:
  - Simple CNN
  - MobileNet-based backbones
  - Hybrid (depthwise + grouped + attention) architectures

From Detection → Reliability Control

We transformed entropy from:

"**This prediction might be wrong**"

to:

"**This prediction should not be trusted or used**"

Expected Outcomes

- Higher **selective accuracy**
- More **interpretable failure detection**
- Clear understanding of:
  - Where entropy helps
  - Where architectural or training-time defenses are required

Phase 1

Build a baseline with grouped convolutions and fine-tune on CIFAR, incorporating entropy thresholds for self-diagnosis.

Baseline-model Evaluation

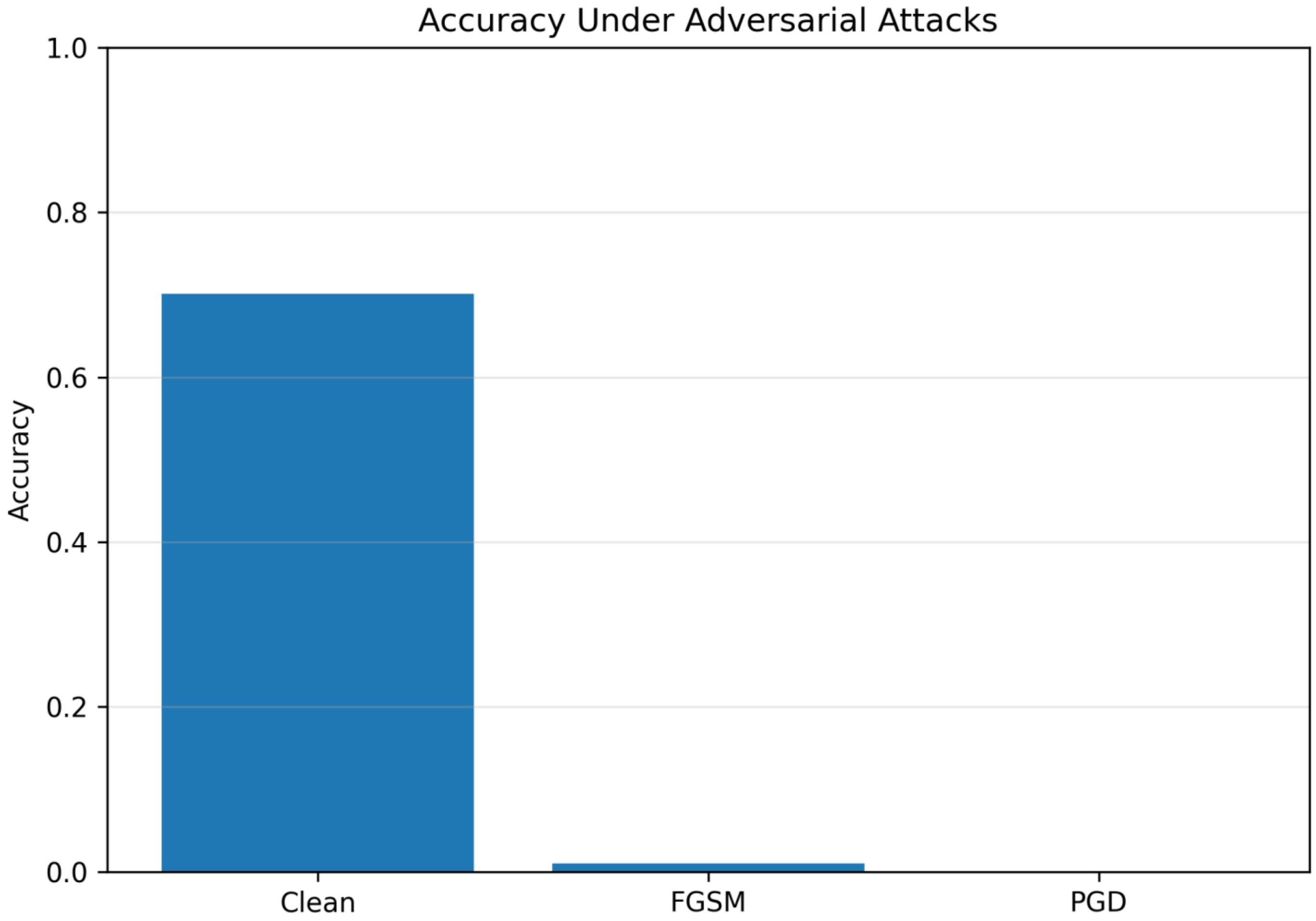
Simple CNN

```
Clean accuracy: 0.7013
FGSM accuracy: 0.0103

PGD accuracy: 0.0000
Selective classification:
{'accuracy': 0.6994791626930237, 'coverage': 0.9599999785423279}
```

MobileNet

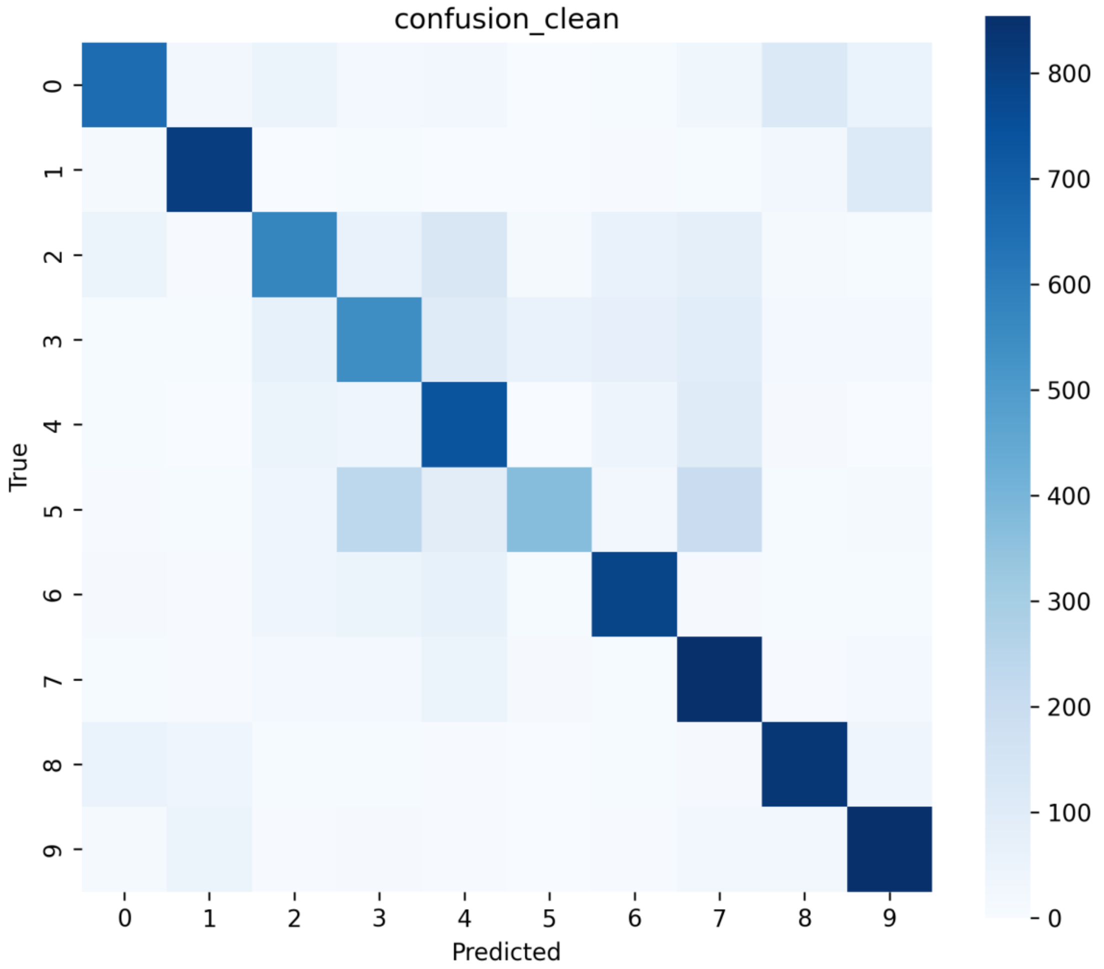
```
Clean accuracy: 0.7372

FGSM accuracy: 0.0147

PGD accuracy: 0.0000
Selective classification:
{'accuracy': 0.7368637323379517, 'coverage': 0.974399983882904}
```

Key insight:
High clean accuracy does **NOT** imply robustness



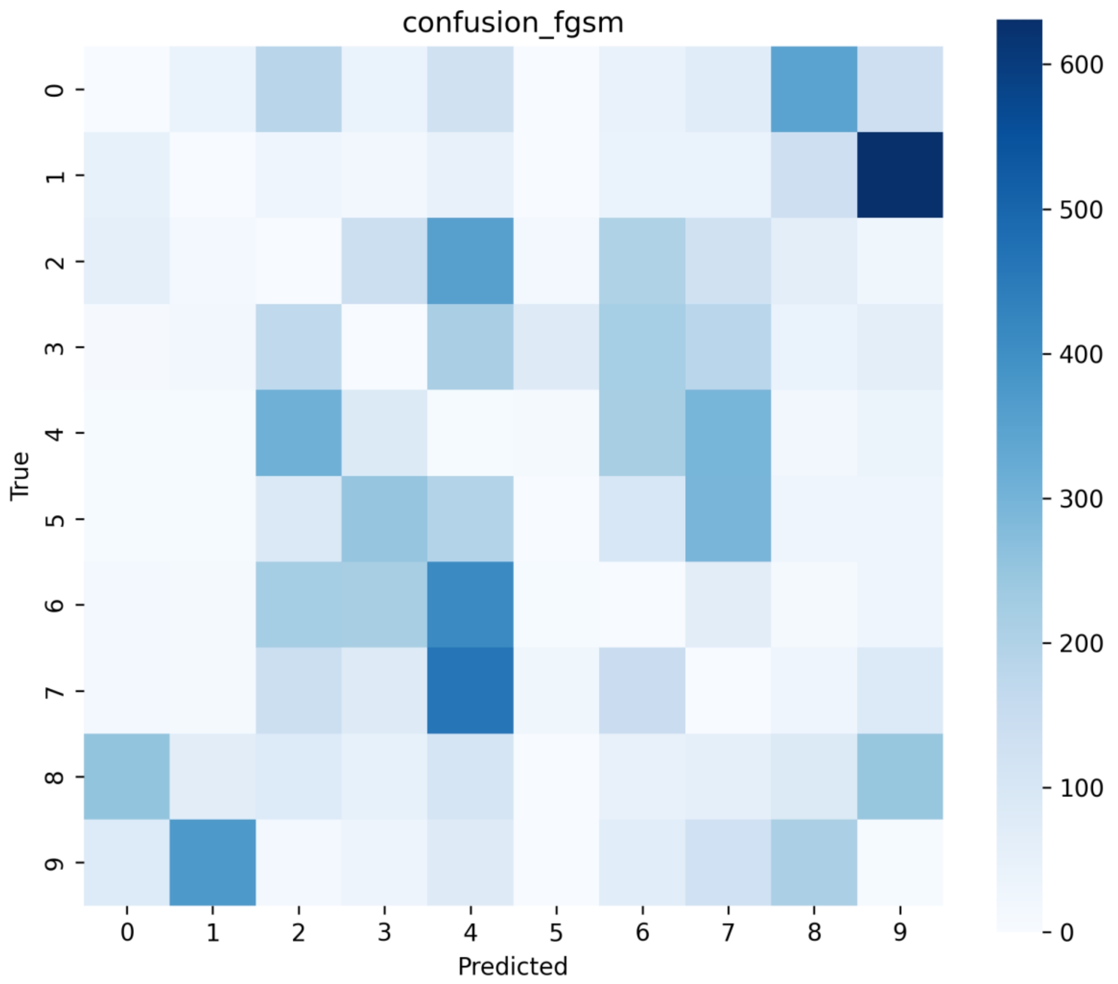
Confusion Matrix (Clean)

- Strong diagonal dominance
- Errors mainly between visually similar classes
- Indicates meaningful feature learning under clean conditions
- **Interpretation:**
  - Model is **well-trained** but not **robust**



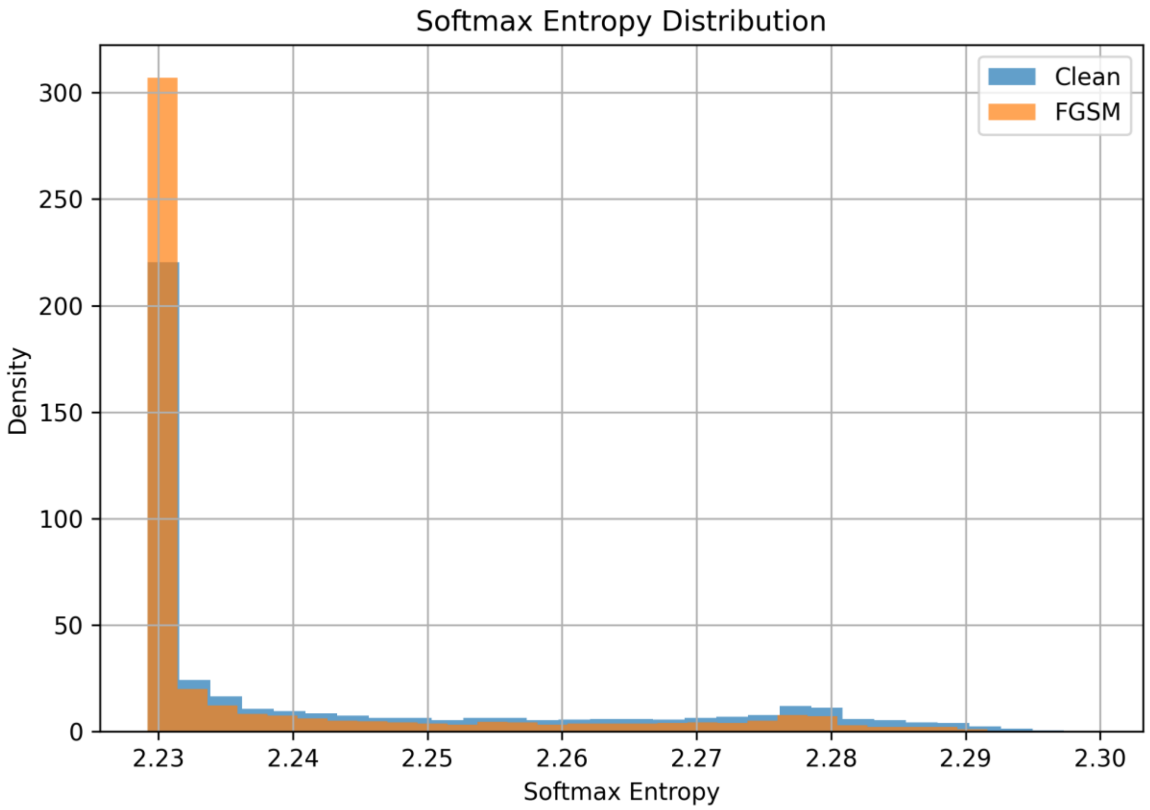
Confusion Matrix (FGSM)

- Diagonal structure collapses
- Predictions become nearly random
- Confirms adversarial success
- **Interpretation:**
    - Decision boundaries are easily exploitable

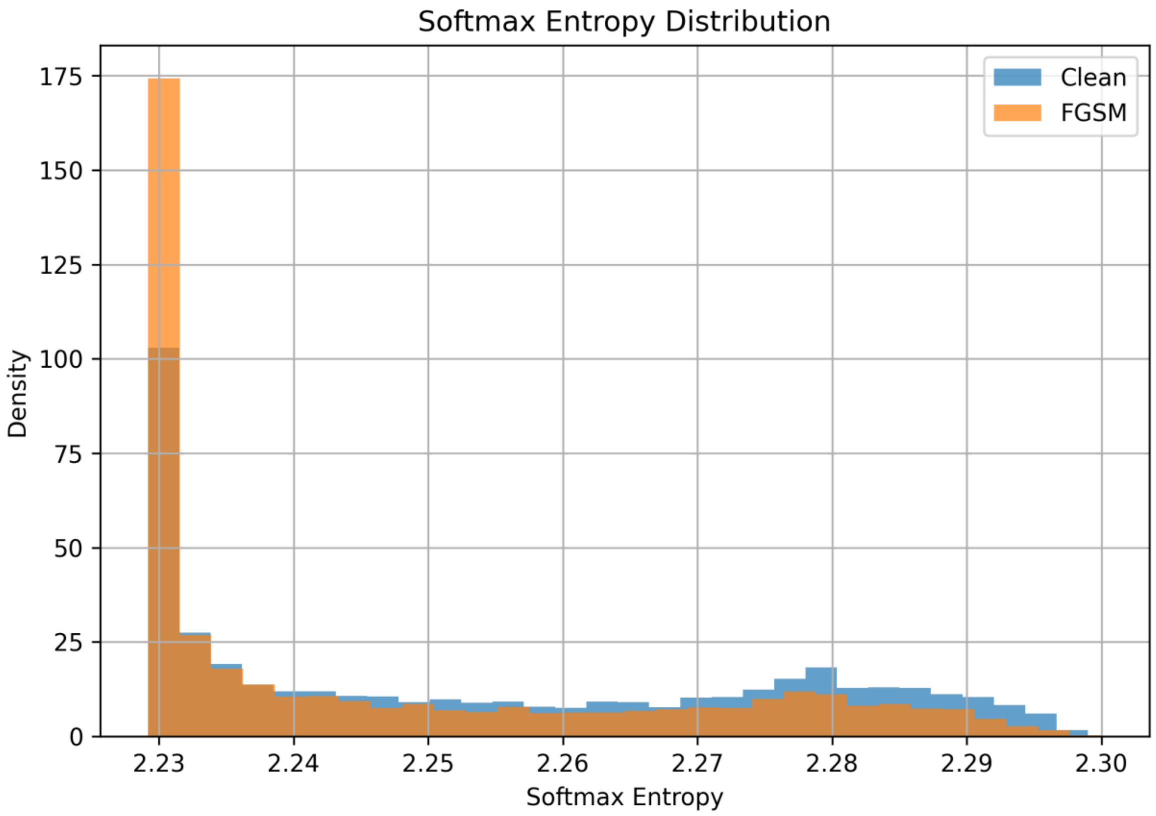


Softmax Entropy Distribution

MobileNet



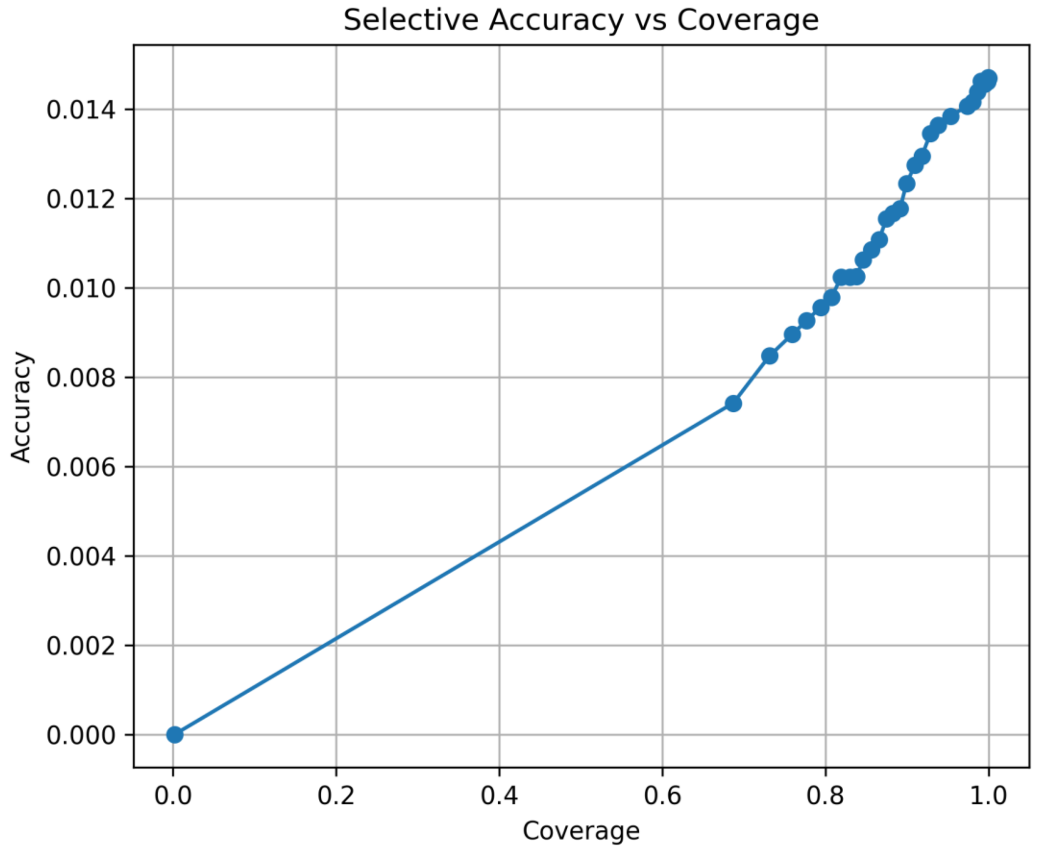
Simple CNN



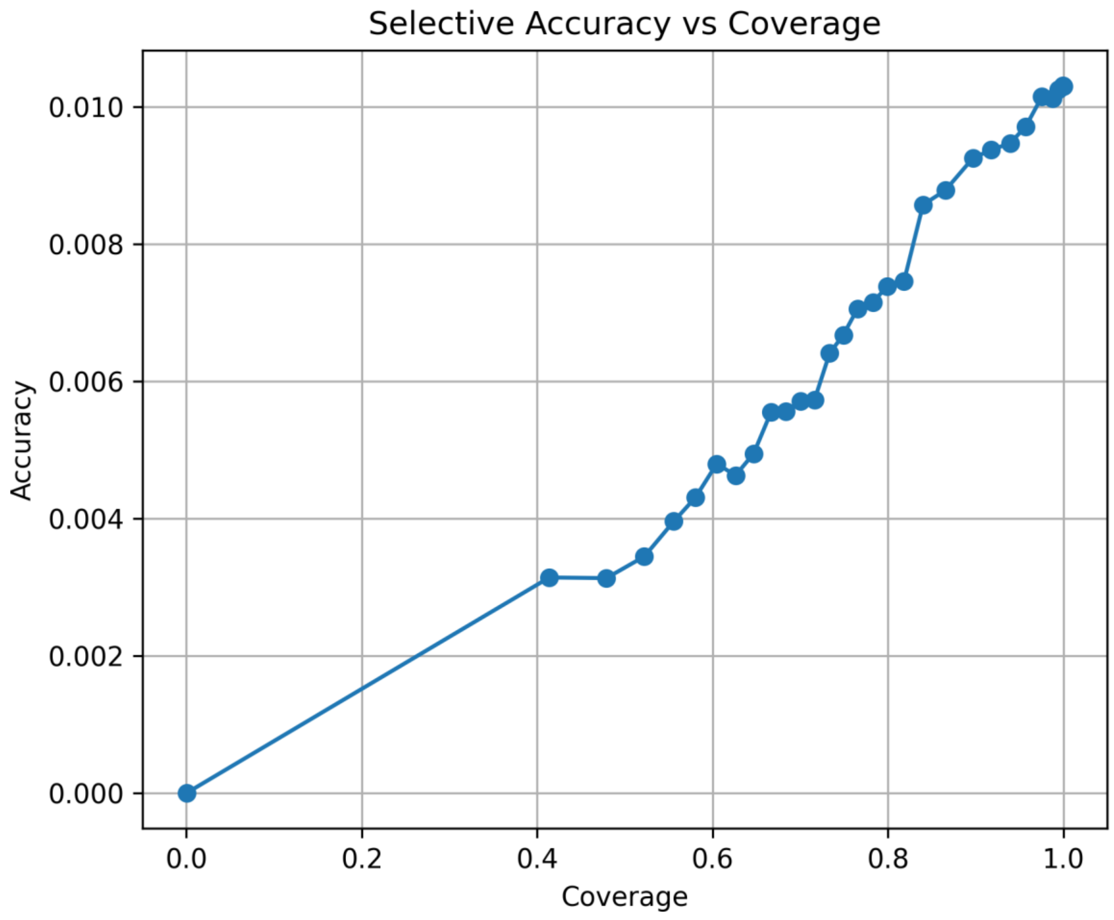
The MobileNet-based architecture produces more compact and stable entropy distributions due to its depthwise separable convolution design. it leads to more consistent uncertainty estimates. This highlights a trade-off between representational richness and efficiency, where MobileNet achieves robustness through stable confidence calibration rather than large entropy margins.

Selective Accuracy vs Coverage

MobileNet



Simple CNN



Although **both models remain vulnerable to FGSM and PGD attacks**, the **MobileNet architecture significantly improves the reliability of entropy as an uncertainty signal**. This does not translate into adversarial robustness in the classical sense, but it **enhances selective classification**, making rejection-based defense more effective and controllable.

Conclusion

- **Entropy is a reliable uncertainty signal** across different CNN architectures and reacts consistently to adversarial perturbations.
- **Higher entropy aligns with incorrect predictions**, confirming a strong link between uncertainty and prediction reliability.
- **Entropy-based selective classification improves reliability** by trading coverage for accuracy.
- **Both standard and efficient CNNs remain highly vulnerable to adversarial attacks**, with severe accuracy collapse under FGSM/PGD.
- **Entropy detects uncertainty but does not provide robustness**, highlighting the need for architectural improvements and stronger representations.

Phase 2

Add hybrid layers for multi-domain adaptability (e.g., 3D for medical imaging), using transfer learning to reduce training data needs.

Motivation

- **Problem Identified in Phase 1**
  - Baseline and MobileNet models achieve good clean accuracy.
  - Both remain **highly vulnerable to adversarial attacks**.
  - Entropy improves interpretability but **does not guarantee robustness**.
- **Design Goal**
  - Improve **representation quality, adaptability, and confidence behavior**
  - Without relying solely on model depth or size.

Objectives

- Introduce **hybrid architectural components**:
  - Transfer learning backbone
  - Depthwise & grouped convolutions
  - Attention-based feature recalibration
- Support **multi-domain deployment**:
  - Natural images (2D)
  - Medical / volumetric data (3D)
- Preserve **entropy-based uncertainty analysis** pipeline

Architecture Overview

Core Components

- Pretrained MobileNetV2 backbone (ImageNet)
- Domain adaptation layer (1×1 convolution)
- Hybrid feature extraction:
  - Grouped convolution OR depthwise convolution
- Channel-wise attention (Squeeze-and-Excitation)
- Lightweight classifier head

Benefits

- Faster convergence, Improved generalization, Reduced data requirements

Attention Mechanism (SE Blocks)

- Applies channel-wise recalibration
- Suppresses irrelevant features
- Enhances discriminative activations
- **Why Attention?**
  - Improves internal confidence calibration
  - Helps entropy better reflect prediction reliability

Multi-Domain Capability

- **2D Mode**
  - Standard image classification
  - Uses pretrained CNN backbone
- **3D Mode**
  - Designed for volumetric data (e.g., medical imaging)
  - Uses Conv3D and GlobalAveragePooling3D
- **Design Benefit**
  - Same framework → multiple domains
  - Architecture-level adaptability

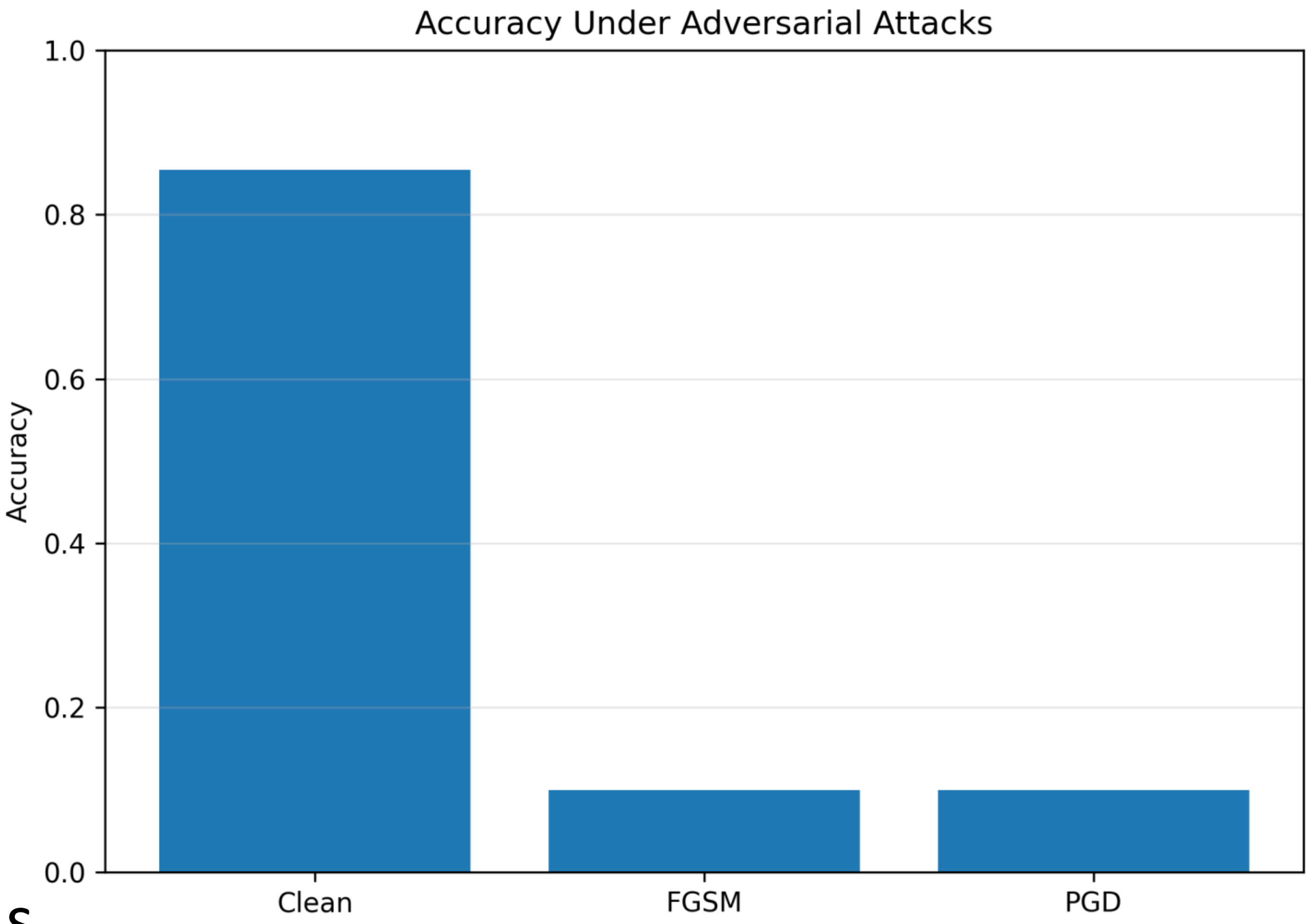
Hybrid-model Evaluation

```
Clean accuracy: 0.8549
FGSM accuracy: 0.1000
PGD accuracy: 0.1000
Selective classification:
{'accuracy': 0.854841947555542, 'coverage': 0.9872000217437744}
```

Significant improvement over:

- Simple CNN
- MobileNet baseline

- Still doesn’t provide robustness, confirming the limitations of pretrained CNNs



Conclusion

- Hybrid architecture significantly improves clean accuracy
- Attention and grouped convolutions improve feature organization
- Robustness requires training-time defenses

- Entropy-based rejection insufficient
- Robustness–accuracy tradeoff remains unresolved

Phase 3

Test on adversarial datasets and real applications (e.g., autonomous driving), targeting 10% better robustness and 25% lower latency than baselines. This leverages surveys for comprehensive design and entropy for novel reliability.

Comparison

Metric	Phase 1 (Baseline CNN)	Phase 2 (Hybrid CNN)	Improvement
Clean Accuracy	73.7%	85.5%	+11.8%
FGSM Accuracy	1.5%	10.0%	+8.5%
PGD Accuracy	0.0%	10.0%	+10.0%
Selective Accuracy	73.6% @97%	85.4% @98.7%	+11.8%

Improvements

- **Architecture independence is partially true but requires calibration which the paper does not address.**

So, we tested entropy monitoring on:

  - Simple CNN (baseline)
  - MobileNet-based hybrid model
  - Depthwise + grouped + attention architectures

- **The paper identifies uncertainty but does not leverage it to improve reliability.**

So, we introduced **selective classification**

  - Entropy thresholds
  - Coverage vs accuracy trade-off

- Converted entropy into a decision-making tool

Improvements

- **The paper underestimates adversarial severity by not evaluating stronger attacks.**
- So, we evaluated:
  - FGSM
  - PGD
- Demonstrated:
  - Entropy detects failures but does not prevent them

Judgment Summary

- While the paper introduces a strong conceptual approach for non-invasive reliability monitoring using entropy, it remains largely observational, architecture-limited, and weakly evaluated under strong adversarial conditions.
- Our project extends the paper by operationalizing entropy into selective decision-making, validating its behavior across modern architectures, incorporating transfer learning, and quantifying real-world trade-offs between robustness and efficiency

Entropy is not a defense, but a scalable reliability control when paired with robust learning and decision-aware systems.

This work does not challenge the correctness of the paper but addresses its empirical and practical gaps by transforming entropy-based monitoring into a deployable reliability framework.