Adversarial Transferability in Malware Detection: A Family-Aware Evaluation

An Empirical Study of Feature-Space Attacks Across Families and Classifiers

Charles Sturt University
Dan Brown – 11802019
danbrown.cyber@gmail.com

1. School of Computing, Mathematics & Engineering
2. Special acknowledgement to Dr Kenneth Eustace and the cohort of ITC571 - Emerging Technologies and Innovation (S032025)
3. AI tools were used to assist with code generation, debugging, and optimisation during experimental implementation. Full disclosure is provided in the accompanying research repository.

1 ABSTRACT

Machine-learning malware detectors achieve high clean accuracy but remain vulnerable to feature-space adversarial attacks. This study evaluates robustness across malware families and examines adversarial transfer between a Neural Network and a Random Forest under controlled ε-bounded perturbations. Results reveal rapid robustness degradation, strong family-specific vulnerabilities, and partial cross-architecture transferability, demonstrating that aggregate accuracy alone is insufficient for security-critical evaluation.

2 INTRODUCTION

Machine learning is widely used for malware detection, with many classifiers reporting strong binary accuracy. While adversarial evasion of static malware classifiers is well established, existing literature offers limited systematic empirical analysis of how robustness varies across malware families and whether adversarial behaviour transfers between different model architectures (Biggio & Roli, 2018; Yan et al., 2023; Li et al., 2023).

This study examines:

- Family-level malware behaviour
- Neural Network and Random Forest architectures
- Feature-space adversarial perturbations
- Cross-model adversarial transferability

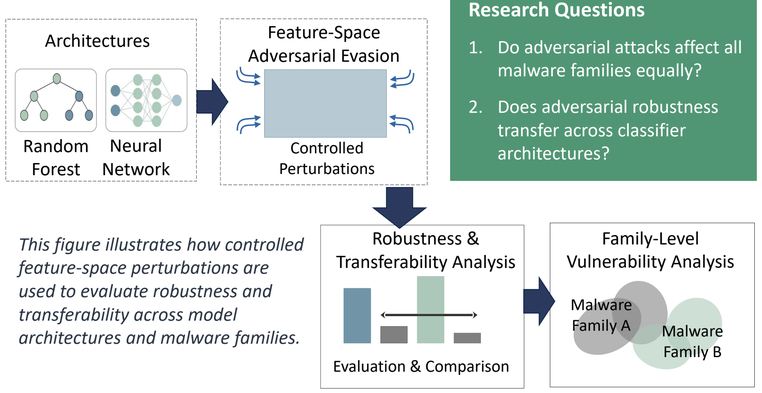
3 REASEARCH AIMS & OBJECTIVES

To empirically evaluate the robustness and adversarial transferability of static malware detection models under feature-space evasion, with particular focus on model architecture and malware family behaviour.

OBJECTIVES

- Assess adversarial robustness of Neural Network and Random Forest classifiers.
- Measure robustness degradation under feature-space adversarial perturbations.
- Analyse cross-model adversarial transferability.
- Evaluate robustness variation across malware families.

Figure 1: Research Aims and Objectives



Research Questions

1. Do adversarial attacks affect all malware families equally?
2. Does adversarial robustness transfer across classifier architectures?

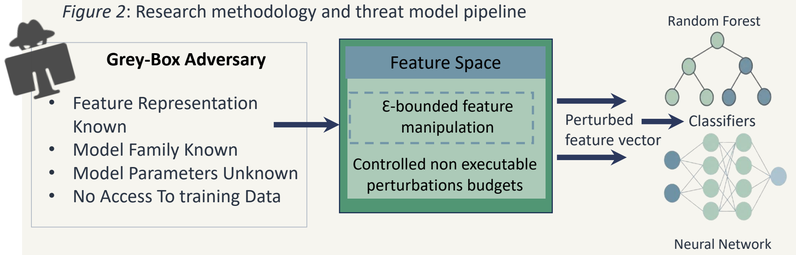
This figure illustrates how controlled feature-space perturbations are used to evaluate robustness and transferability across model architectures and malware families.

4 RESEARCH METHODOLGY & METHODS

METHODOLOGICAL APPROACH

This study adopts an empirical, experimental methodology to evaluate adversarial robustness in static malware detection, explicitly addressing family-level robustness variation and cross-model transferability through controlled, reproducible adversarial testing.

Figure 2: Research methodology and threat model pipeline



MODELS & DATA

Classifier Architecture

- Random Forest (tree-based)
- Neural Network (deep learning)

Data Representation

- EMBER-derived static PE feature vectors
- 12k sample (balanced benign/malware split)

MODEL TRAINING

- 80/20 split (train and test) from balanced sample data

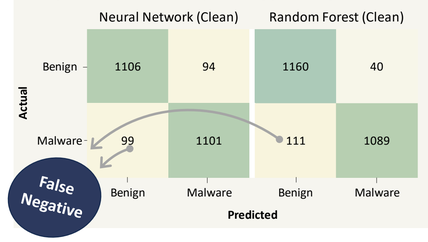
ADVERSERIAL ATTACK

- FGSM and PGD feature-space attacks
- Transfer scenarios (NN → RF)

EVALUATION STRATEGY

Robustness was assessed using malware false negative rates, cross-model adversarial transferability, and family-level performance degradation. This enables systematic comparison of architecture-specific robustness, transfer behaviour, and differential vulnerability across malware families.

Figure 3. Clean models featuring false negatives



5 KEY FINDINGS & DISCUSSION

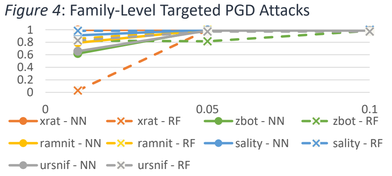
Key findings demonstrate that strong clean baseline performance does not imply adversarial robustness, with targeted feature-space attacks causing rapid performance degradation, pronounced family-level vulnerability differences, and asymmetric cross-model transferability between neural and tree-based classifiers.

Table 1. Baseline Malware Detection Performance

Model	Accuracy
Random Forest	0.9371
Neural Network	0.9196

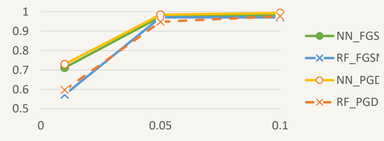
Both architectures achieve high baseline performance under clean conditions, establishing a reliable reference point for adversarial evaluation. *This is also demonstrated previously in Figure 3.*

Figure 4: Family-Level Targeted PGD Attacks



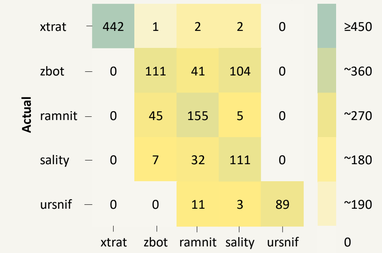
False-negative rates under PGD attacks vary substantially across malware families and between direct (NN) and transfer (RF) scenarios.

Figure 5: Cross-Architecture Feature-Space Attacks



Targeted feature-space perturbations rapidly induce malware evasion, with near-complete evasion achieved at moderate ε values.

Figure 6: NN→RF FGSM Transfer - Family-Level Confusion Matrix (ε = 0.05)



Family-level confusion matrix showing partial NN→RF adversarial transfer at ε = 0.05, with family-specific misclassification patterns.

DISCUSSION

Interpretation of Results

Despite strong accuracy under clean conditions, both architectures showed significant degradation in robustness under targeted attacks, demonstrating that binary accuracy masks vulnerability to feature-space perturbations.

Robustness was **non-uniform across malware families.** This indicates that susceptibility is influenced by family-specific features rather than the classifier architecture alone.

Adversarial examples generated against neural networks **partially transferred** to Random Forest classifiers, demonstrating that architectural diversity does not completely mitigate evasion risk.

7 CONCLUSION & FUTURE WORK

Despite strong clean accuracy, static malware detectors exhibit substantial degradation in robustness under feature-space attacks across both Random Forest and Neural Network architectures. Susceptibility varied significantly across malware families, indicating that robustness is shaped by family-specific feature characteristics rather than the classifier architecture alone. Partial cross-model transferability further demonstrates that architectural diversity does not fully mitigate the risk of evasion.

Take-home message:
Robustness evaluation must move beyond aggregate accuracy alone and incorporate adversarial testing that is more family-aware and transfer-explicit.

FUTUREWORK

- Complete final production scripts, provide additional documentation within the reproducible evaluation repository and overall housekeeping of notebooks.
- Expand experiments to larger datasets (≥500k–1M samples) to assess at scale.
- Evaluate across alternative datasets with improved malware family labelling, and extend analysis from feature-space to problem-space adversarial attacks.

Access the full reproducible evaluation repository: https://bit.ly/itc571-assess3

REFERENCES

1. Anderson, H. S., & Roth, P. (2018). Ember: an open dataset for training static PE malware machine learning models.
2. Biggio, B., & Roli, F. (2018). Wild patterns: Ten years after the rise of adversarial machine learning. Pattern Recognition, 84, 317-331.
3. Yan, S., Ren, J., Wang, W., Sun, L., Zhang, W., & Yu, Q. (2023). A Survey of Adversarial Attack and Defense Methods for Malware Classification in Cyber Security.
4. Li, D., Li, Q., Ye, Y., & Xu, S. (2023). Arms Race in Adversarial Malware Detection: A Survey. ACM Computing Surveys, 55(1), 1-35.
5. Image utilised in the section “Introduction - This study examines”, source: medium.com https://bit.ly/45Hke9s