Malware Through the Lens of Computer Vision

How Binary-to-Image Encodings Influence CNN-Based Malware Family Classification

Martim Lopes Cardeira, M.D.A.LopesCardeira@student.tudelft.nl, Delft University of Technology
Responsible Professor: Tom Viering, Supervisor: Akash Amalan

CSE3000 Research Project Q4 2026

Introduction

Representing a malware binary as an image and classifying it with a CNN is a fast, disassembly-free alternative to traditional signature-based detection, and studies routinely report **near-perfect accuracies**.

But is that accuracy genuine learning of malware structure, or are CNNs **exploiting shortcuts**? How do different binary-to-image encodings compare, and how does obfuscating the malware affect performance?

We compare four encodings head-to-head under obfuscation, and diagnose what the models actually learn.

Research Question

Across four binary-to-image encodings (**grayscale byteplot, RGB byteplot, Markov bigram plot, and sliding-window Shannon entropy**), how well do CNNs classify malware under obfuscation, and how much of that reflects genuine family-discriminative structure?

Dataset

10,010 samples · 14 families · PE & ELF · 7 packer conditions

Format	Families
PE	Trojan, Ransomware, Worm, Spyware, Adware, Rootkit, Botnet, BenignPE
ELF	Mirai, Gafgyt, HiddenWasp, XorDDoS, Tsunami, BenignELF

Packers

Each family includes packed variants spanning three obfuscation strategies:

Strategy	Packer	Effect on binary
Compression	UPX, MPRESS	compresses into an executable which is decompressed at runtime
	zlib, ZIP	compresses into an archive file
Encryption	XOR	byte-level XOR with a random keystream
Substitution	Mangle	rewrites IoCs, clones code-signing certs

7 conditions incl. the unpacked baseline; 15,037 samples total.

TU Delft — Delft University of Technology

Encodings

Grayscale Byteplot

Each byte is read as a pixel intensity value (0–255) and laid out as a 2D grayscale image. *Serves as the baseline encoding.*

RGB Byteplot

Every three consecutive bytes form one RGB pixel, $R = b_0$, $G = b_1$, $B = b_2$: a 3× denser byteplot compared to grayscale.

Markov Bigram Plot

A state-transition matrix P where $P[i][j]$ is the probability of byte value i being followed by byte value j. A power-law mapping to pixel values preserves very low probabilities.

Shannon Entropy

Shannon entropy is computed over a sliding window, then mapped to grayscale pixel values and arranged as a 2D image.

Why These Four?

Each encoding assumes the family-discriminative signal lives in a **different domain**:

- **Byteplots** - binary layout & raw byte values
- **Markov** - byte-transition statistics (instruction semantics, control flow)
- **Entropy** - regional entropy (encrypted / compressed sections)

Pipeline

Malware Binaries → Encode → CNN Training → Evaluation/Comparison → Diagnostic framework

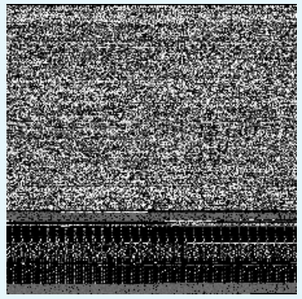
Each encoding is trained and evaluated independently, with all images resized to 224 × 224 for the ResNet-18 classifier.

Model & Training

ResNet-18 from scratch, trained independently per (encoding × packer) pair.

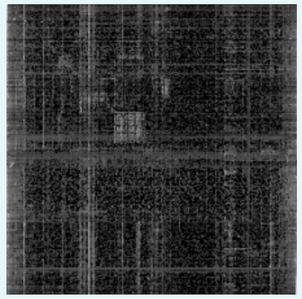
- Split: stratified 5-fold cross-validation
- Adam, learning rate $= 10^{-4}$
- CrossEntropyLoss, 30 epochs, batch 64
- Hardware: NVIDIA A40 48GB

Encoding Methods - Mirai Sample



Grayscale Byteplot
Nataraj et al., 2011 [2]

RGB Byteplot
Vasan et al., 2020 [4]



Markov Bigram Plot
Nie & Zhu, 2026 [3]

Shannon Entropy
Darton, 2026 [1]

Diagnostic Framework

Is the accuracy real?

A high score may reflect a **shortcut**, not learned malware structure. Three complementary diagnostics are held against every (encoding × packer) model:

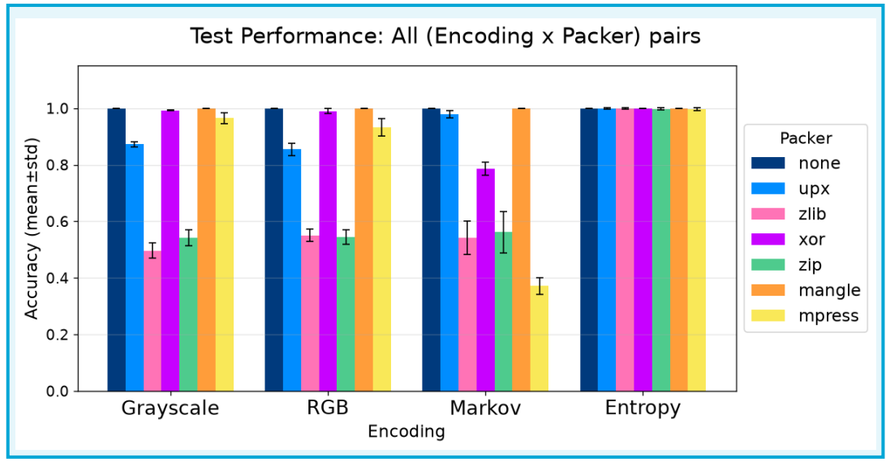
- **Random-Forest baselines** - can trivial features (file-size, global entropy) alone match CNN performance?
- **TLSH similarity** - do near-duplicate malware variants leak across cross-validation folds, resulting in memorization?
- **HiResCAM + band ablation** - which image regions drive the decision?

Research Gap

Why this study?

Surveys find almost no **controlled, head-to-head** encoding comparisons. Resilience to obfuscation is **widely claimed but weakly substantiated**, and explainability work focuses almost entirely on grayscale byteplots. This study closes these gaps.

Results - Accuracies Compared



Key Findings

RQ1 – Performance under obfuscation

- Unpacked & Mangle: **all encodings classify perfectly** - trivial without strong obfuscation.
- **Entropy dominates**: ≥ 0.996 on every packer, uniquely resilient even under compression.
- Byteplots (Grayscale ≈ RGB) degrade substantially under compression (zlib, ZIP); **Markov is the weakest**, collapsing on MPRESS.

XOR variants were found to be erroneously packed; no conclusions are drawn from XOR.

RQ2 – Accuracy may be misleading

- **File-size alone** (Random Forest) reaches ≥ 0.94 on most packers – a simple yet powerful shortcut.
- Near-duplicate **leakage trivializes easy packers** (unpacked & Mangle: 100% similarity).
- **But MPRESS neutralises both** (size-RF 0.53, zero leakage) while most encodings still score ≥ 0.93 ⇒ disproves total reliance on shortcuts and suggests genuine learning.

RQ3 – Where the models look

- Entropy's **per-family heatmaps vary the most** – the strongest evidence of genuine family-discriminative learning.
- Band Ablation on top and bottom images rows rules out header/padding reliance for all encodings; learning takes place in the main binary sections.

Takeaway

Encodings **cannot be ranked by accuracy alone**. Entropy is both the most obfuscation-resilient encoding *and* the one showing the clearest evidence of genuinely learning malware structure – and only a combination of diagnostics could tell the two apart. *Future work should aim to explain Shannon Entropy in more depth.*

References

[1] H. J. Darton. "Malware Detection with CNNs on Entropy and Greyscale Images". In: *Latin-American Journal of Computing* 13.1 (Jan. 8, 2026), pp. 45–53. ISSN: 1390-9134, 1390-9266. DOI: 10.33333/lajc.vol13n1.04.
[2] L. Nataraj, S. Karthikeyan, G. Jacob, and B. S. Manjunath. "Malware images: visualization and automatic classification". In: *Proceedings of the 8th International Symposium on Visualization for Cyber Security*. VizSec '11: 2011 International Symposium on Visualization for Cyber Security. Pittsburgh Pennsylvania USA: ACM, July 20, 2011, pp. 1–7. ISBN: 978-1-4503-0679-9. DOI: 10.1145/2016904.2016908.
[3] W. Nie and C. Zhu. "M2I: Image-based malware classification via feature spatial transformation". In: *Journal of Information Security and Applications* 97 (Mar. 1, 2026), p. 104355. ISSN: 2214-2126. DOI: 10.1016/j.jisa.2025.104355.
[4] D. Vasan, M. Alazab, S. Wassan, H. Naeem, B. Safaei, and Q. Zheng. "IMCFN: Image-based malware classification using fine-tuned convolutional neural network architecture". In: *Computer Networks* 171 (Apr. 2020), p. 107138. ISSN: 13891286. DOI: 10.1016/j.comnet.2020.107138.