

Malware Families Catalog: A Structured Behavioral Taxonomy, MITRE ATT&CK Mapping, and Incident Response Playbook Framework for 2,966 Malware Families

Ricky Jordan

*Lead Threat Intelligence Analyst & Malware Researcher
SystemHelpDesk MSP Research Operations · Denver, CO*

Technical Report: SHD-TR-2026-08-MF · **DOI:** 10.5281/zenodo.20515698 · **Published:** October 2026 · **License:** Apache-2.0

ABSTRACT

The proliferation of automated commodity malware and polymorphically obfuscated payloads has created severe telemetry dilution in contemporary Security Operations Centers (SOCs). Benchmark corpora such as Elastic's EMBER 2018 have established baseline machine learning feature spaces for static PE classification, yet lack standardized behavioral ontology mappings and tactical remediation context. In this work, we present the *Malware Families Catalog*, an open-access, peer-verified technical taxonomy structuring 2,966 real-world malware families across 35 distinct functional categories. Every family is systematically mapped against the MITRE ATT&CK® enterprise matrix, MITRE D3FEND™ defensive countermeasures, and US CISA Known Exploited Vulnerabilities (KEV). Furthermore, we formalize detection rule deployment via the Cisco CAIRN 3-Tier model (Primitive Artifacts, Behavioral Process Chains, and Attribution Fingerprints) and provide enterprise interoperability via OASIS STIX 2.1 bundles. This paper documents the catalog's sourcing methodology, taxonomic schema, graph correlation architecture, and retrospective threat hunting applications.

1. INTRODUCTION

Modern cyber defense operations suffer from an asymmetry of threat attribution. While commercial endpoint detection and response (EDR) platforms generate billions of daily telemetry events, raw signature alerts frequently omit historical lineage, threat actor attribution, and actionable containment guidance.

To bridge the gap between academic malware classification benchmarks and frontline incident response, the SystemHelpDesk research team engineered the *Malware Families Catalog*. Derived from the 1.1-million-sample EMBER 2018 benchmark, the catalog synthesizes static PE telemetry with threat

intelligence feeds from MITRE, CISA, and MalwareBazaar into a public knowledge graph.

2. BASELINE CORPUS & INGESTION

The underlying empirical foundation is the **EMBER 2018 v2** corpus, released by Elastic. EMBER comprises extracted features from 1.1 million binaries scanned in or before 2018, labeled via AVClass clustering across multi-vendor VirusTotal engines.

While AVClass addresses initial cluster labeling, raw vendor strings regularly suffer from generic naming (e.g., `Win32.Trojan.Generic`). Our ingestion engine applies strict disambiguation filters:

1. **Stopword Filtering:** Stripping 50+ generic identifiers (e.g., `loader`, `agent`, `patcher`) to isolate

true family identifiers.

2. **Taxonomic Alias Reconciliation:** Normalizing vendor nomenclature (e.g., `Emotet`, `Geodo`, `Heodo`) via a curated alias registry of 160,000+ synonyms.

3. **Sample Threshold Verification:** Retaining statistically valid clusters with verified execution telemetry.

3. CATEGORICAL DISTRIBUTION

The catalog categorizes 2,966 verified families into 35 primary threat classes. Table 1 enumerates the top ten categories by representation.

Table 1: Top 10 Malware Categories

Category	Families	Primary Tactic
Ransomware	1,624	Impact (T1486)
Trojan Generic	544	Execution (T1059)
Advanced Threat	515	Persistence (T1547)
Backdoor	60	C2 (T1071)
Adware / PUA	48	Defense Evasion
RAT	19	Collection / Exfil
Banking Trojan	17	Credential Access
InfoStealer	13	Credentials / Data
Worm / Banker	12	Lateral Movement
Downloader	11	Initial Access

Ransomware accounts for over 54% of curated families, reflecting the intense proliferation of ransomware-as-a-service (RaaS) codebases observed throughout 2016–2018 and subsequent enterprise targeting.

4. MULTI-LAYER RULE ENGINEERING

Detection engineering requires tiered defense to balance alert fidelity against false-positive fatigue. The catalog structures detection rules using the **Cisco CAIRN Model**:

- **Tier 1 (Primitive Artifacts):** Static YARA rules targeting raw byte sequences, section hashes, and hardcoded PE strings. Best suited for high-throughput

initial triage.

- **Tier 2 (Behavioral Context):** Sigma rules modeling parent-child process relationships, script interpreter execution (e.g., `powershell.exe` executing encoded payloads), and Registry RunKey modifications.

- **Tier 3 (Attribution Fingerprints):** High-confidence heuristic rules combining compiler watermarks, distinctive C2 framing, and cryptographic routine fingerprints.

Operational Principle: Catalog indicators are curated for *retrospective threat hunting* and telemetry baselining. Due to operational drift, static IOCs should not serve as sole perimeter blocklists.

5. DEFENSIVE COUNTERMEASURES

Rather than merely cataloging adversary offensive techniques, each family is mapped bidirectionally to **MITRE D3FEND™** defensive countermeasures. For example, families utilizing Spearphishing Attachment (`T1566.001`) and Visual Basic execution (`T1059.005`) are matched with:

- **D3-FS (File Sandboxing):** Dynamic hypervisor detonation of macro-enabled files.

- **D3-SEA (Script Execution Analysis):** Antimalware Scan Interface (AMSI) runtime script telemetry.

- **D3-RPM (Registry Persistence Monitoring):** Real-time auditing of user and machine-level Run keys.

6. STIX 2.1 INTEROPERABILITY

To facilitate automated ingestion into modern Security Orchestration, Automation, and Response (SOAR) and Threat Intelligence Platforms (TIPs), the catalog exports a complete OASIS STIX 2.1 bundle containing over 12,000 JSON objects.

The STIX export formally links each `malware` SDO to its associated `attack-pattern` SDOs via standardized `uses` relationships, complete with ISO-8601 timestamps and deterministic UUIDv5 identifiers.

7. ETHICAL DISCLOSURE & PROVENANCE

In accordance with responsible disclosure practices and North American compliance frameworks (NIST CSF and CISA guidelines), this catalog does not distribute live weaponized binaries. All published IOCs, YARA rules, and Sigma signatures are derived from historical open-source intelligence and academic benchmarks.

8. CONCLUSION & AVAILABILITY

The *Malware Families Catalog* demonstrates that structured behavioral taxonomy and cross-framework semantic mapping provide tangible operational advantage to incident response personnel. The full dataset is permanently available under Apache-2.0 via:

- **Canonical Portal:**

<https://malware-families-catalog.systemhelpdesk.com/>

- **Zenodo DOI:** [10.5281/zenodo.20515698](https://doi.org/10.5281/zenodo.20515698)

- **STIX 2.1 Feed:** [/api/stix2_bundle.json](#)

- **Parquet / JSONL:** [/data/malware_families.*](#)

REFERENCES

- [1] H. S. Anderson and P. Roth, “EMBER: An Open Dataset for Training Static PE Malware Machine Learning Models,” *arXiv:1804.04159*, 2018.
- [2] MITRE Corporation, “MITRE ATT&CK®: Design and Philosophy,” Technical Report, 2024. Available: <https://attack.mitre.org/>
- [3] MITRE Corporation, “MITRE D3FEND™: A Knowledge Graph of Cybersecurity Countermeasures,” 2024. Available: <https://d3fend.mitre.org/>
- [4] Cybersecurity and Infrastructure Security Agency (CISA), “Known Exploited Vulnerabilities Catalog,” 2026. Available: <https://www.cisa.gov/known-exploited-vulnerabilities-catalog>
- [5] OASIS Open, “Structured Threat Information Expression (STIX™) Version 2.1,” OASIS Standard, 2021.
- [6] National Institute of Standards and Technology (NIST), “Cybersecurity Framework (CSF) 2.0,” NIST CSWP 29, 2024.
- [7] M. Sebastiani et al., “AVClass: A Tool for Massive Malware Labeling,” in *Research in Attacks, Intrusions, and Defenses*, 2016.