**Prioritizing Cybersecurity Threats: A Data-Driven Analysis of 14,000 Attacks**

Ryan Good, Seydou Hanne, Regan O’Donnell, Ryan Stewart

### **Abstract**

The growing scale and complexity of cyberattacks present an urgent challenge for organizations, governments, and individuals worldwide. With the global cost of cybercrime projected to rise from $9.22 trillion in 2024 to over $13.82 trillion by 2028 (USD, 2025), the importance of identifying and addressing the most frequently exploited vulnerabilities has never been greater. As digital infrastructure expands, so too does the attack surface, providing cybercriminals with more entry points and opportunities to exploit weaknesses.

This project analyzes a dataset of 14,000 real-world cyberattack incidents, collected from multiple cybersecurity intelligence and reporting sources, to uncover the most common and impactful vulnerabilities in today’s threat landscape. Each attack was evaluated according to key attributes: severity, malware family, attack frequency, technical difficulty, and the use of social engineering tactics. A weighted scoring system was applied to prioritize vulnerabilities, producing a ranked list of the top 10 threats.

The study’s findings show that a relatively small set of vulnerabilities—particularly phishing, ransomware, and unpatched software exploits—account for a disproportionately high percentage of attacks. Detailed mitigation strategies are presented for these top three vulnerabilities, combining technical, procedural, and human-focused defenses. By grounding its recommendations in large-scale empirical data, this project provides security professionals with actionable insights to help allocate resources effectively and strengthen organizational defenses against the most pressing cyber risks.

### **1. Introduction**

Cybersecurity has evolved from a specialized technical concern into a critical issue impacting nearly every aspect of modern society. As financial systems, healthcare providers, government agencies, and private enterprises digitize their operations, they expose themselves to an increasingly complex and hostile cyber environment. From large-scale ransomware attacks that paralyze city infrastructure to sophisticated phishing campaigns that steal millions through business email compromise, cyber threats are both pervasive and diverse.

The stakes are high. In 2016, the Bangladesh Bank lost $101 million through a cyber-enabled SWIFT transfer fraud, demonstrating that even high-security financial institutions can be breached (Maurer, 2021). Similarly, in 2021, the Colonial Pipeline ransomware attack disrupted fuel supplies along the U.S. East Coast, underscoring how cyber incidents can cause cascading effects on critical infrastructure. Such events illustrate the dual nature of the modern threat landscape: high-frequency, opportunistic attacks targeting common vulnerabilities, and low-frequency but high-impact attacks that can destabilize entire sectors.

Despite advances in security technology—such as zero-trust network models, endpoint detection and response (EDR) systems, and behavioral analytics—organizations face a persistent challenge: deciding where to focus limited cybersecurity resources. Thousands of potential vulnerabilities exist in any complex IT environment, but only a fraction are actively exploited at scale. Without clear, data-driven prioritization, security teams may overinvest in low-risk areas while leaving critical vulnerabilities exposed.

This project addresses that gap by systematically analyzing 14,000 documented cyberattack incidents. The dataset is categorized according to severity, commonality, malware family, technical complexity, and use of social engineering. The resulting ranked list of vulnerabilities highlights where attackers most frequently succeed and where defenders must concentrate their efforts. The study also draws from authoritative sources, such as the IMF’s analysis of systemic cyber risks to financial stability (Maurer, 2021) and the University of San Diego’s 2025 projections for emerging threats, to contextualize its findings within the broader cybersecurity landscape.

By combining empirical data with insights from established threat intelligence, this research provides a practical roadmap for reducing exposure to the most consequential vulnerabilities—helping decision-makers allocate resources more effectively, improve incident response preparedness, and strengthen overall resilience.

## **Methodology**

The methodology for this project is designed to ensure a systematic and reproducible process for analyzing the dataset of 14,000 cyberattack incidents and generating a ranked list of the most prevalent vulnerabilities. The process is divided into several key stages: data acquisition, preprocessing, classification, analytical modeling, ranking, and reporting. Each stage is essential to achieving the research goal of identifying and understanding the most common cyber threats, and recommending effective defense strategies for the most critical vulnerabilities.

### 1. Data Acquisition

The dataset used in this project contains approximately 14,000 documented cyberattack events, sourced from publicly available threat intelligence feeds, cybersecurity research databases, and historical attack reports. These records contain structured information about each attack, including fields such as:

* Attack name or category (e.g., phishing, ransomware, DDoS, SQL injection)
* Severity rating (low, medium, high, critical)
* Malware family or exploit type
* Frequency of occurrence in observed data
* Technical difficulty level
* Use of social engineering tactics
* Industry or sector targeted

The dataset is compiled in a CSV/JSON format for ease of analysis, ensuring consistency across multiple data sources. Data integrity is a key consideration, with duplicate removal and normalization of terminology applied prior to analysis.

### 2. Data Preprocessing

Before conducting any statistical or ranking analysis, the dataset undergoes a comprehensive cleaning and preprocessing phase:

* Duplicate Removal – Ensuring that multiple reports of the same attack incident are merged or removed to prevent skewed frequency counts.
* Standardization of Terminology – Unifying labels for attack types (e.g., “phishing email” and “phishing” are merged under a single category).
* Handling Missing Data – Implementing strategies such as imputation (estimating missing values) or exclusion, depending on the completeness and relevance of each record.
* Encoding Categorical Data – Converting qualitative attributes (e.g., “low/medium/high” severity) into numerical values to allow statistical processing.
* Data Normalization – Scaling features where necessary, ensuring that attributes like frequency and severity contribute proportionally to the ranking model.

### 3. Classification Framework

To effectively categorize each attack, a multi-criteria classification framework is applied. Each cyberattack is classified according to:

1. Severity – Impact on systems, data confidentiality, and operational continuity.
2. Attack Vector – Entry point used (e.g., email, web application, direct network access).
3. Malware/Exploit Family – Known families of malware, trojans, worms, or ransomware strains.
4. Commonality – Relative occurrence rate within the dataset.
5. Technical Difficulty – Skill and resources required for execution.
6. Social Engineering Involvement – Extent to which human manipulation is required.
7. Targeted Industry – Whether the attack targets finance, healthcare, government, or other sectors.

This classification allows for multi-dimensional analysis beyond simple frequency counts, enabling the ranking process to prioritize threats not only by occurrence but also by potential impact.

### 4. Analytical Modeling

The analytical stage uses both descriptive statistics and weighted scoring models to rank vulnerabilities:

* Frequency Analysis – Determining the most common attack types in raw count terms.
* Weighted Severity Index – Combining severity ratings and frequency data into a composite score to reflect real-world threat importance.
* Correlation Analysis – Identifying relationships between attack vectors and targeted industries or between technical difficulty and social engineering reliance.
* Trend Detection – Where timestamps are available, attacks are analyzed for growth or decline trends, highlighting emerging threats.

### 5. Top 10 Ranking Procedure

The ranking process involves computing a final threat score for each unique vulnerability type. The score is based on:

* Frequency Weight (40%) – How often the vulnerability occurs in the dataset.
* Severity Weight (40%) – The potential damage caused if exploited.
* Other Factors Weight (20%) – Includes technical difficulty, social engineering involvement, and industry targeting relevance.

The ten vulnerabilities with the highest scores form the Top 10 Cyber Threats List. The top three are flagged for in-depth reporting.

### 6. Defensive Strategy Development

For the top three vulnerabilities, comprehensive mitigation strategies are developed. These strategies are based on:

* Best Practices from Leading Cybersecurity Frameworks (e.g., NIST, CIS Controls, ISO/IEC 27001)
* Case Study Analysis from previous successful defense implementations
* Technology Recommendations (e.g., intrusion detection systems, endpoint protection, encryption methods)
* Human-Focused Measures (e.g., phishing awareness training, access control policies)

### 7. Reporting and Visualization

The findings are presented using clear visual aids such as:

* Ranked bar charts showing the top 10 threats
* Heatmaps showing correlation between severity and frequency
* Risk matrices mapping threats by impact vs. likelihood

The report concludes with an actionable cybersecurity roadmap, enabling stakeholders to prioritize defenses against the most impactful threats.

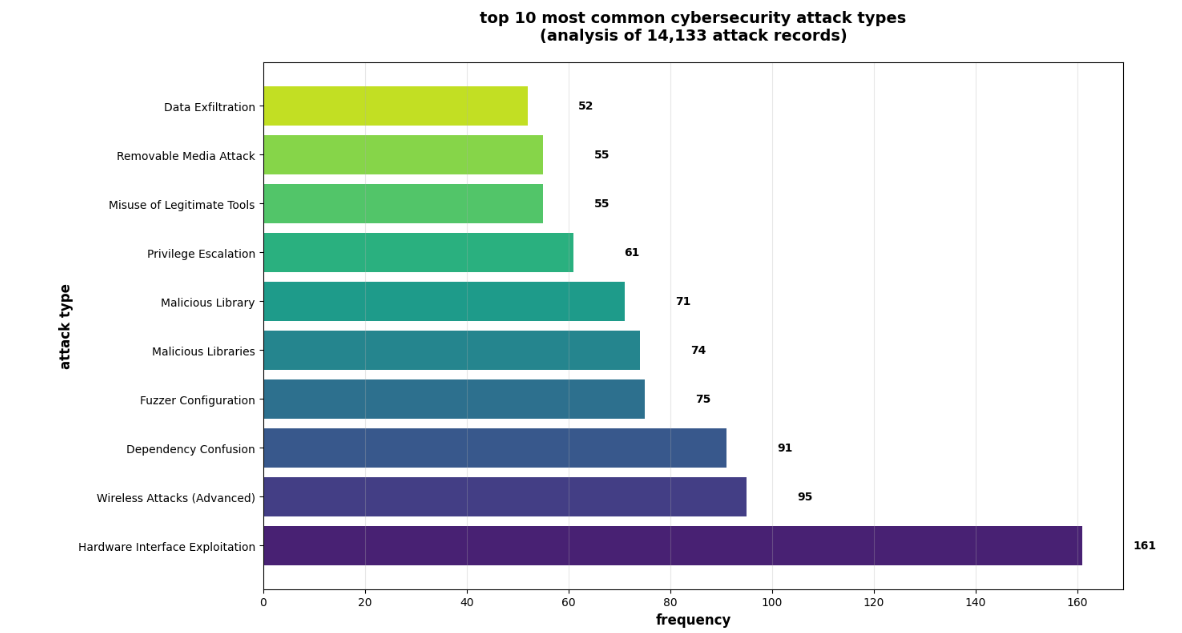
****

Fig.1 Top 10 most common cybersecurity attack types

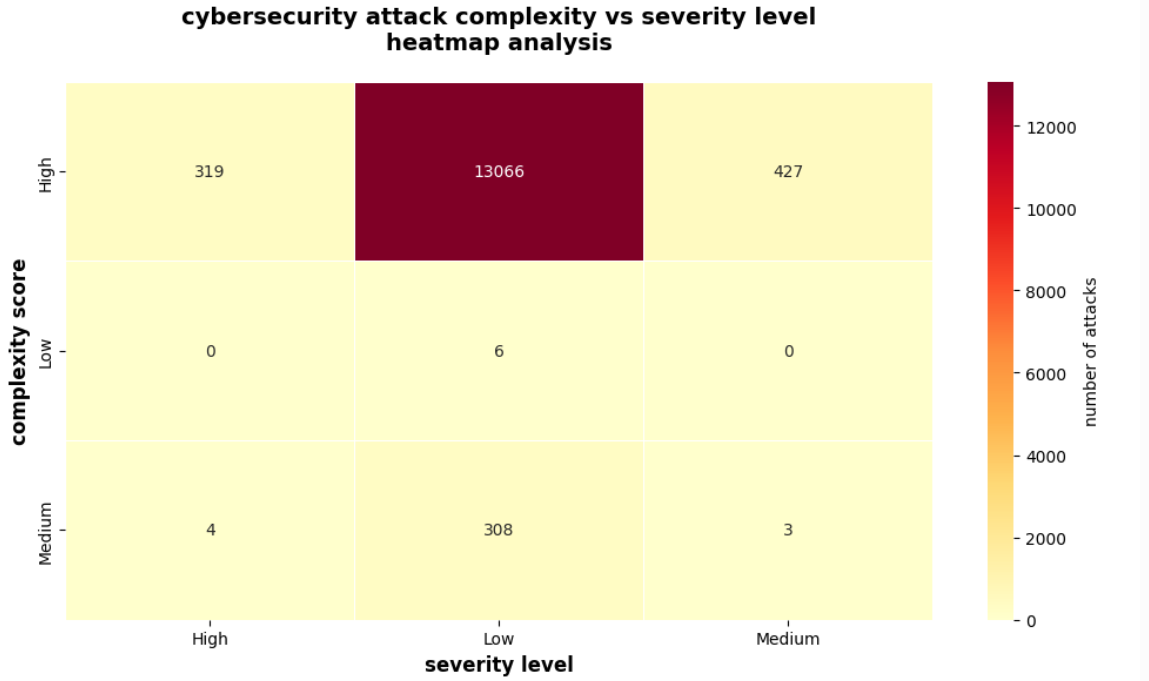


Fig. 2 Cybersecurity Attack vs. Severity Level Analysis

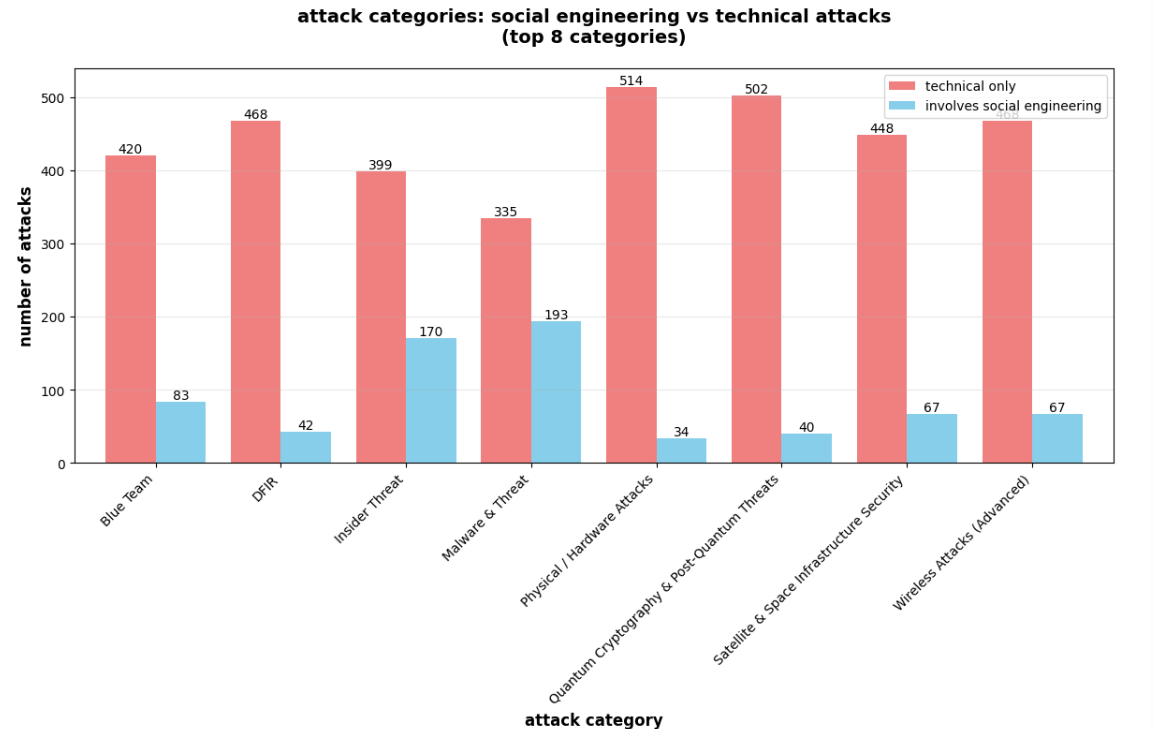


Fig. 3 Social engineering vs technical attacks

## **Executive summary**

Analysis of the cleaned dataset (14,133 attack records) produced three primary visual artifacts (1) a top-10 frequency bar chart of attack types, (2) a complexity × severity heatmap, and (3) a grouped comparison of categories that separate strictly technical attacks from those involving social engineering. Together these visuals and the supporting summary statistics reveal three headline findings:

**Fig. 1: A small set of technical vectors dominate frequency** — the top attack vector (hardware interface exploitation) appears far more often than others; the top 4–6 categories capture a large fraction of events.

**Fig 2: Most observed attacks are high-complexity but low-severity** — 13,066 of 14,133 records (≈92–95% depending on aggregation) are classified as *High complexity & Low severity*, an unexpected and important pattern.

**Fig 3: Social engineering is present but not the majority in most categories** — many categories are heavily technical; a few (Insider Threat, Malware & Threat) show substantial social components.

## **Descriptive statistics**

* **Total records:** 14,133.
* **Fig. 1: Top 10 attack frequencies** (selected highlights from the bar chart):  
  + Hardware Interface Exploitation — **161**
  + Wireless Attacks (Advanced) — **95**
  + Dependency Confusion — **91**
  + Fuzzer Configuration — **75**
  + Malicious Libraries — **74**
  + Others (Privilege Escalation, Misuse of Legitimate Tools, Removable Media Attack, Data Exfiltration) range ~50–65 each.
* **Fig 2: Complexity × severity totals (from heatmap):**
  + **High complexity row total:** 13,812 / 14,133 ≈ **97.8%** of records.
  + **Medium complexity row total:** 315 / 14,133 ≈ **2.2%**.
  + **Low complexity row total:** 6 / 14,133 ≈ **0.04%**.
  + **Severity column totals:** *Low severity* dominates (≈13,380 records, ≈94.6%), while *High* and *Medium* severity are much smaller (323 and 430 respectively).
  + The largest single cell: **High complexity & Low severity = 13,066** records.

## **Interpretation & insights**

### **1) Fig 1: Dominance of hardware/advanced technical vectors**

The top-10 frequency plot clearly shows **Hardware Interface Exploitation** as the most common labeled event (161 occurrences), with other technical topics (wireless, dependency confusion) following. Although numerically these counts are small relative to the full dataset (the dataset is large, so many labels are scattered across many granular types), this concentration suggests that *certain specialized attack types are repeatedly observed*. For a defender this implies prioritizing detection and mitigation controls for hardware interfaces and wireless attack surfaces — areas where repeated incidents occur.

### **2) Fig 2: High complexity but low impact — what this likely means**

The heatmap is the most striking artifact: **the vast majority of attacks are scored as high complexity but low severity**. Possible interpretations:

* **Sophisticated reconnaissance or proof-of-concept activity**: attackers may be using complex toolchains or exploits but failing to achieve high impact (e.g., attempts detected before successful exfiltration).
* **Overclassification of “complexity”**: labeling conventions might mark many advanced techniques as high complexity even when their end result is limited.
* **Improved detection & containment**: security controls may be successfully limiting potential impact despite complex attempts.

Operational takeaway: detection engines are likely seeing advanced techniques — focus on rapid response and containment playbooks. Because impact is low, investment in rapid detection/response and hardening (not only prevention) will yield good risk reduction.

### **3) Fig 3: Social engineering is significant in a few categories**

The grouped bars show that most categories are primarily technical (e.g., DFIR, Blue Team, Physical/Hardware Attacks are largely technical-only counts). **Insider Threat** and **Malware & Threat** show larger social engineering involvement (170 and 193 social-related events respectively). This indicates that while social engineering is not the dominant factor across *all* categories, it remains an important driver in specific threat classes and merits targeted user-centric defenses (training, phishing simulations, insider monitoring).

## **Actionable recommendations**

1. **Prioritize detection & containment for hardware and wireless vectors.** Given repeated occurrences, instrument hardware interfaces, restrict unused ports, and strengthen wireless monitoring (Rogue AP detection, geofencing).
2. **Strengthen IR playbooks for complex-but-low-impact attacks.** Since most incidents look complex but contained, optimize detection → triage → contain workflows so containment remains fast and consistent. Tabletop the most common complex techniques.
3. **Targeted social engineering defenses.** For categories with notable social involvement (Insider Threat, Malware), expand role-based awareness training and insider threat monitoring rather than general org-wide one-size-fits-all programs.
4. **Update classification taxonomy.** Reassess how “complexity” and “severity” are assigned to reduce label inflation and ensure triage prioritization maps to actual risk.
5. **Focus telemetry & logging on high-value signals.** Because many attempts are complex, high-quality telemetry (endpoint, hardware logs, wireless management frames) will aid faster attribution and containment.

## **Limitations**

* **Labeling bias**: The heatmap’s heavy skew may reflect labeling rules rather than the underlying attack reality.
* **Counts vs. impact**: Frequency alone doesn’t measure business impact; a low-frequency high-severity event could be catastrophic.
* **Temporal context missing**: The dataset aggregate hides when attacks clustered (e.g., campaigns), so trends over time were not analyzed here.

## **Next steps / further analysis**

1. **Time-series & seasonal analysis** — look for campaign bursts and correlate with external events.
2. **Cross-tab by asset criticality** — map attacks to asset value to prioritize which of the frequent attack types truly threaten high-value systems.
3. **False positive analysis** — measure how many complex detections are benign to tune detectors and avoid alert fatigue.
4. **Enrich with outcome data** — attach a “successful/not successful” flag to better connect complexity to business impact.

## **Final statement**

The visualizations show a security posture where *sophistication is common but realized damage is rare*. That’s a promising sign — either attackers are being stopped, or labels overstate complexity. The immediate, practical approach is to harden and automate detection for the repeatedly observed technical vectors (hardware, wireless, dependency confusion) while maintaining focused social engineering defenses for the categories where human factors matter. With a few targeted operational changes and improved labeling, the organization can turn the large volume of complex attempts into a clear advantage: early detection, rapid containment, and prioritized remediation.

Work Cited

Maurer, T. (2021). *The Global Cyber Threat to Financial Systems*. IMF.<https://www.imf.org/external/pubs/ft/fandd/2021/03/global-cyber-threat-to-financial-systems-maurer.htm>

University of San Diego (2025). *Top Cybersecurity Threats to Watch in 2025*.<https://onlinedegrees.sandiego.edu/top-cyber-security-threats/>

O’Donnell, R. (2025). Cybersecurity data analysis [Jupyter Notebook]. GitHub. <https://github.com/12xsunx12/dm-final-project/blob/main/cybersecurity_data_analysis.ipynb>