StealthCup: Realistic, Multi-Stage, Evasion-Focused CTF for Benchmarking IDS

Manuel Kern, Dominik Steffan, Felix Schuster, Florian Skopik, Max Landauer, David Allison
Austrian Institute of Technology
firstname.lastname@ait.ac.at

Simon Freudenthaler
FH Hagenberg
freudenthaler.simon@gmail.com

Edgar Weippl
University of Vienna
edgar.weippl@univie.ac.at

ABSTRACT

Intrusion Detection Systems (IDS) are critical to defending enterprise and industrial control environments, yet evaluating their effectiveness under realistic conditions remains an open challenge. Existing benchmarks rely on synthetic datasets (e.g., NSL-KDD, CICIDS2017) or scripted replay frameworks, which fail to capture adaptive adversary behavior. Even MITRE ATT&CK Evaluations, while influential, are host-centric and assume malware-driven compromise, thereby under-representing stealthy, multi-stage intrusions across IT and OT domains. We present *StealthCup*, a novel evaluation methodology that operationalizes IDS benchmarking as an evasion-focused Capture-the-Flag competition. Professional penetration testers engaged in multi-stage attack chains on a realistic IT/OT testbed, with IDS detection penalizing their scores. The event generated structured attacker writeups, validated detections, and PCAPs, host logs, and alerts. Our results reveal that out of 32 exercised attack techniques, 11 were not detected by any IDS configuration. Open-source systems (Wazuh, Suricata) produced high false-positive rates (>90%), while commercial tools generated fewer false positives but also missed more attacks. Comparison with the Volt Typhoon APT advisory confirmed strong realism: all 28 applicable techniques were exercised, 19 appeared in writeups, and 9 in forensic traces. These findings demonstrate that StealthCup elicits attacker behavior closely aligned with state-sponsored TTPs, while exposing blind spots across both open-source and commercial IDS. The resulting datasets and methodology provide a reproducible foundation for future stealth-focused IDS evaluation.

CCS CONCEPTS

• **Security and privacy** → **Network security**; **Intrusion detection systems**;

KEYWORDS

IDS benchmarking, adversary emulation, IT/OT, evasion

ACM Reference Format:

Manuel Kern, Dominik Steffan, Felix Schuster, Florian Skopik, Max Landauer, David Allison, Simon Freudenthaler, and Edgar Weippl. 2026. StealthCup: Realistic, Multi-Stage, Evasion-Focused CTF for Benchmarking IDS. In *Proceedings of ACM Asia Conference on Computer and Communications Security, Bangalore, India, June 1--5, 2026 (ASIA CCS '26),* 14 pages.
https://doi.org/10.1145/3779208.3806088

ASIA CCS '26, June 1--5, 2026, Bangalore, India

ACM ISBN 979-8-4007-2356-8/26/06…$15.00
https://doi.org/10.1145/3779208.3806088

1 INTRODUCTION

Intrusion Detection Systems (IDS) remain a cornerstone of enterprise and critical infrastructure defense. Yet despite advances in machine learning and analytics, evaluating IDS performance under realistic conditions is still an open problem. Most academic work continues to benchmark IDS on synthetic datasets such as NSL-KDD [25] or CICIDS2017 [4]. While useful for regression testing, these datasets are artificially generated and diverge statistically from real traffic distributions, limiting their value as indicators of real-world detection effectiveness [14, 17, 45]. Replay frameworks and testbeds [23, 30, 31] offer controlled experimentation but typically rely on scripted exploits, failing to capture adaptive adversary behavior. The MITRE ATT&CK Evaluations [30] have become a widely recognized industry benchmark for endpoint detection products, but they are fundamentally host-centric, rely on dated threat intelligence reports, and assume initial compromise via commodity malware—conditions under which EDR solutions naturally excel while network- or anomaly-focused IDS are underrepresented. This enables vendors to tune products to well-documented attack steps, highlighting coverage of known techniques while underrepresenting stealthy, multi-phase operations observed in modern campaigns.

In practice, advanced adversaries exploit subtle misconfigurations, credential hygiene issues, and lateral movement chains spanning IT and OT domains [9, 28, 51]. Such behavior is difficult to reproduce in scripted or malware-centric evaluations, but it is crucial for understanding IDS blind spots in realistic deployments. This motivates the need for new evaluation methodologies that (i) capture human attacker ingenuity and evasion, (ii) reflect complex IT/OT infrastructures, and (iii) provide reproducible benchmarks with transparent results.

Based on these challenges, we derive the following research questions that guide our study:

RQ1: To what extent does StealthCup generate attacker behavior and datasets that resemble those observed in real-world APT reports, and are thus suitable for IDS evaluation?

RQ2: To what extent can the approach capture the impact of IDS configuration choices (e.g., open-source vs. commercial, default vs. tuned) on detection effectiveness against stealthy attacker behavior?

Contributions. Our main contributions are as follows:

(1) **The StealthCup framework:** an innovative evaluation approach combining evasion-oriented, multi-stage attack challenges in a structured competition format.
(2) **Validated IT/OT testbed:** implementation of a realistic, vulnerable, and expert-validated IT/OT infrastructure as Infrastructure-as-Code, including documented multi-stage attacks and IDS rule sets.
(3) **Comparative IDS evaluation:** systematic analysis of open-source (Wazuh, Suricata) and commercial (Vendor A EDR, Vendor B NIDS)[1] IDS solutions under realistic stealth scenarios, highlighting strengths, blind spots, and the effect of tuning.
(4) **Open dataset release:** public release of alerts, event logs, PCAPs, and attacker writeups to facilitate independent verification and future research.

By combining human-driven evasion with reproducible infrastructure and transparent scoring, StealthCup complements existing evaluation efforts (datasets, replay frameworks, and MITRE ATT&CK Evaluations) with an adversarial perspective that more closely reflects how IDS solutions perform against stealthy, adaptive attackers.

In its inaugural run, StealthCup brought together 52 participants across 12 international teams to attack a realistic IT/OT infrastructure. The competition produced 14 structured attacker writeups and PCAPs, alerts, and host logs. Analysis showed that 11 of 32 attack techniques evaded all IDS configurations, open-source systems exhibited false-positive rates above 90%, and commercial tools missed critical techniques despite producing fewer false positives. By aligning 28 exercised techniques with the Volt Typhoon APT advisory, StealthCup confirmed that participant behavior closely mirrors state-sponsored tradecraft, underscoring both the realism of the approach and its value for benchmarking IDS under stealth-focused adversaries.

2 RELATED WORK

Benchmarking IDS with datasets. Evaluation of intrusion detection systems has traditionally relied on standardized datasets. Early benchmarks such as DARPA98 [27] and its successors provided labeled traffic traces for training and evaluation, but were later criticized for unrealistic traffic generation and over-simplified attack scenarios [2, 42]. More recent datasets, including NSL-KDD [25] and CICIDS2017 [4], continue to be widely used for machine learning-based IDS benchmarking [22, 29]. While useful for reproducibility and regression testing, such datasets remain artificially curated, lack attacker adaptiveness, and diverge statistically from real-world traffic distributions [14, 17]. Consequently, high accuracy on synthetic traces does not necessarily transfer to operational environments.

Testbeds and evaluation initiatives. Beyond static datasets, DARPA-style testbeds [22] aim to create controlled environments for IDS evaluation. These allow experiments with live systems but typically rely on scripted exploits and predefined attack scenarios. MITRE ATT&CK Evaluations [30] represent a widely recognized industry effort, mapping adversary emulation to ATT&CK techniques. However, they are host-centric, rely on dated threat intelligence reports, and assume initial access via commodity malware. As a result, endpoint detection and response (EDR) tools tend to score well, while network-based IDS and anomaly-focused approaches are underrepresented. Furthermore, because evaluation steps are publicly documented, vendors can tune their systems in advance, limiting insights into evasive or novel attacker behavior.

Red teaming and capture-the-flag (CTF). Professional red-team exercises provide the closest analogue to real intrusions, as skilled testers emulate advanced adversary tactics to remain undetected. However, such engagements are costly, results are rarely disclosed, and reproducibility is limited. CTF competitions exist in many formats [3, 5, 10--12, 16, 20, 53, 54], including attack--defense and defend-only types, but they generally do not focus on systematically evaluating IDS configurations. Academic experiments have explored integrating IDS into competitions, e.g., the UCSB iCTF 2009 where IDS alerts influenced scoring [21], but IDS played only a secondary role and experimental control was limited. More broadly, CTFs such as DEFCON CTF [10] or Pwn2Own [54] highlight the creativity of adversarial participants, but do not provide structured IDS evaluation. Notably, prior work and the MITRE ATT&CK framework [30] itself underscore that penetration testers and advanced persistent threats (APTs) share overlapping tactics, techniques, and procedures (TTPs). Both seek stealth, persistence, lateral movement, and privilege escalation, while routinely attempting to evade security controls. Thus, CTF-style engagements provide a natural mechanism to elicit attacker ingenuity similar to state-sponsored campaigns.

Positioning StealthCup. StealthCup complements these approaches by combining the creativity of human attackers with the reproducibility of controlled testbeds. Unlike synthetic datasets (DARPA98, CICIDS2017), StealthCup generates live attack traces focused explicitly on detection evasion. Unlike MITRE ATT&CK Evaluations, its objectives are not fully disclosed in advance, incentivizing adaptive and stealthy attacker strategies. And unlike prior CTFs, StealthCup treats IDS evasion as a first-class objective with structured scoring, reproducible Infrastructure-as-Code deployments, and public release of alerts, logs, and attacker writeups. This positioning makes StealthCup a complementary methodology that reveals IDS blind spots specific to configuration and deployment, while ensuring comparability and transparency.

3 METHODOLOGY

StealthCup integrates Capture-the-Flag (CTF) gamification into intrusion detection benchmarking. The framework is designed to capture realistic attacker behavior, evaluate IDS resilience against stealthy operations, and produce reproducible datasets.

3.1 Overview of the StealthCup Framework

StealthCup provides a generic evaluation framework that can be instantiated in different domains and infrastructures. A concrete instantiation is detailed later in the illustrative use case (Sect. 4),

[1] At the request of participating vendors, commercial product identifiers have been anonymized. Both commercial solutions represent widely deployed products in enterprise and critical infrastructure environments.

whereas this section outlines the core methodology. At a high level, the framework consists of four key steps:

Competition Design. StealthCup adopts an *attack-only evasion focused CTF* format. Participants are tasked with solving multi-stage attack challenges while avoiding detection. Every triggered alert while solving a challenge incurs a penalty. This shifts the incentive structure: success is measured not by how quickly teams achieve objectives, but by how stealthily they can operate compared to their competitors when solving a challenge.

To make this effective, StealthCup implements a dynamic scoring scheme. Alerts are weighted by severity, with high-severity alerts generating stronger penalties than low-severity ones. Because severity taxonomies differ between IDS solutions (e.g., Wazuh logs many low-severity alerts unrelated to real compromise, while Vendor A EDR[1] low-severity alerts are often true indicators of intrusion), we normalize penalties across systems by applying adjustable weights per IDS. This normalization ensures comparability while preserving fairness across heterogeneous detection technologies.

Infrastructure and Attack Scenarios. Narrative *scenarios* with clearly defined objectives form the foundation of StealthCup. Rather than providing isolated exploits or disconnected technical tasks, each scenario consists of multiple *objectives* that mimics a realistic intrusion campaign. Objectives are defined in terms of attacker goals (e.g., establish persistence in the Active Directory domain, exfiltrate data, manipulate a PLC register), ensuring that every step contributes to a logically consistent multi-stage chain, such as client compromise → Active Directory escalation → lateral movement → PLC manipulation. This design discourages "point-and-shoot" exploitation and instead drives participants toward strategic, multi-phase operations that mirror how real adversaries pursue end-to-end objectives. To preserve realism while remaining feasible, scenarios incorporate realistic misconfigurations and recently disclosed vulnerabilities (*attack vectors*), reflecting what professional penetration testers would encounter in practice. This avoids the disincentive of "burning" valuable zero-day exploits, although we do not prevent participants from employing them. We argue that requiring zero-day exploits for operating systems or common software privilege escalation would discourage participation, given their high market value (e.g., Zerodium pricing [7]). In contrast, StealthCup explicitly encourages the discovery or use of zero-day techniques targeting IDS components, as these directly contribute to winning the challenge.

The underlying infrastructure is tailored to the scenario context. For example, a deployed instance of an energy sector combines enterprise IT assets (Windows domains, Linux servers) with representative OT equipment and protocols (e.g., PLC controllers using Modbus) common in critical infrastructure [18]. In StealthCup, the infrastructure mirrors the socio-technical environment relevant to the chosen domain, while detailed vendor and configuration choices are deferred to the illustrative case study in Section 4.4.

In the following, we use the term *testbed* to denote the controlled experimental infrastructure. In StealthCup, a testbed comprises not only the systems and software, but also their specific configuration state and monitoring solutions. This includes both secure baseline setups and deliberately injected misconfigurations or recently disclosed vulnerabilities that serve as attack vectors. By explicitly modeling configuration and misconfiguration within the testbed, we can design repeatable attack chains, ensure that all participants face identical opportunities for compromise, and systematically study how IDS solutions respond to different phases of the intrusion. The testbed is fully automated using Infrastructure-as-Code (Terraform, Ansible), enabling consistent rebuilds, controlled randomization, and comparability across teams and scenarios.

Data Collection and Ground Truth. All experiments are instrumented to collect comprehensive traces, including full packet captures (PCAPs), host logs, and IDS alerts. A key challenge is establishing ground truth in the absence of detailed logging. Installing logging agents on participant machines would provide precise traces, but we deliberately avoid this to not discourage participation or bias attacker behavior. Instead, we combine (i) structured attacker writeups, required upon challenge completion, (ii) manual verification of reported steps against collected telemetry.

While attacker-side logging agents would enable additional KPIs such as time-to-detect and enable throughout labelling, we avoided this in the first StealthCup to not discourage participation; this remains a planned extension.

Evaluation Mechanics. To prevent discouragement after early mistakes, teams are allowed to reset their infrastructure on demand. Resets introduce trade-offs: information obtained in a previous run may remain useful, so we implement randomization of hostnames, services, and network configurations to mitigate knowledge carryover.

Prevention mechanisms (e.g., anti-malware blocking of tools such as Mimikatz) are deliberately disabled, ensuring that participants can operate freely. However, all such actions still generate detection alerts, which are penalized in scoring.

Importantly, detection events are made visible to participants in near-real time, enabling them to adapt their strategies dynamically. We adopt this design for three main reasons. First, near-real-time feedback helps participants use the limited competition time more effectively, as they can adjust their approach rather than unknowingly persisting with unproductive attack paths. Second, it encourages an iterative style of engagement, where participants explore alternative tactics when their initial attempts trigger alerts. Finally, from a benchmarking perspective, exposing detections introduces adaptive pressure on IDS solutions, allowing us to observe not only how well they detect first attempts but also how resilient they remain once attackers begin to modify their behavior.

This adaptive interaction loop not only mirrors real-world adversaries but also increases the diversity of collected attack traces, since participants are incentivized to refine and vary their tactics.

3.2 Ethics

All activities occur within isolated ranges under informed consent. No production assets or customer data are involved; outbound connectivity is restricted. We log only range telemetry; personal data from attackers is not retained beyond participation records. Released artifacts include Infrastructure as Code, attack-artefacts, IDS configs (default/tuned), scoring code, scenario descriptions

with ATT&CK mappings and sanitized PCAP/log bundles. All data was collected under research ethics approval.

4 ILLUSTRATIVE CASE STUDY

This section describes how we instantiated the StealthCup framework in a realistic IT/OT environment. We outline the design rationale, testbed construction, IDS integration and event preparation of the first StealthCup event.

4.1 Scenario Selection and Design Rationale

We chose a hybrid IT/OT setting to capture both enterprise and industrial intrusion vectors. IT infrastructures with Active Directory (AD) domains, Windows, and Linux servers are ubiquitous across enterprises, while OT components are highly relevant for critical infrastructures [9]. The only preventive network security function implemented was basic layer 4 firewalling, ensuring segmentation but no active intrusion prevention. The initial attacker foothold was implemented as a *Kali* (Kali Linux 2024.3.0) in the enterprise zone. This choice reflects common compromise entry points (e.g., a contractor laptop, phishing victim, or insider device) and provides participants with a familiar penetration testing toolset [9, 28, 51]. This ensures accessibility, comparability across teams, and a controlled starting point for all attacks.

The initial foothold was fixed across all teams. Access to the OT environment was only possible by leveraging misconfigurations in the IT/OT-DMZ, requiring skilled participants with expertise in lateral movement and cross-domain trust exploitation.

4.2 Testbed Creation

To guide the design of attack vectors, we conducted a survey of *state-of-the-art attacks* on AD networks and OT networks [13, 46].

Threat Model and Assumptions We model skilled, human red teams pursuing predefined objectives while minimizing detections. The testbed embeds realistic misconfigurations and recently disclosed vulnerabilities; operating-system or core network zero-days are *not required* by design.

Core assumptions.

- **Isolation.** Each team operates in its own range; no cross-team leakage of traffic, credentials, or state.
- **Detect-only.** Active prevention (e.g., IPS drops, malware blocking) is disabled; detections are visible but do not block execution.
- **Initial foothold.** Teams start from a fixed enterprise client with standard reachability but no pre-positioned privileged credentials.
- **Randomization.** Resets generate fresh instances with randomized credentials, while attack paths and objectives remain constant to ensure comparability.
- **Data integrity.** IDS sensors and the SIEM backend are trusted; tampering is prohibited. Full PCAP, host logs, and IDS alerts are recorded continuously and persist across resets.

Design requirements.

- **Attack vectors.** Techniques reflect current practice (e.g. living-off-the-land (LotL), AD/PKI misconfigurations, credential hygiene issues, OT protocol weaknesses).
- **Attack paths.** Objectives require multi-stage compromise (e.g., AD persistence; PLC manipulation); lateral movement must be (re)performed after each reset.
- **Replayability.** After resets, difficulty and chains remain consistent to permit fair replay and benchmarking.
- **Scoring.** Penalties derive from alert counts and severities per IDS; validations require a short attacker writeup (cf. §4.6).

Limitations.

- User behavior is minimal; user and entity behavior analytics (UEBA) or anomaly-only detection approaches are out of scope.
- External threat-intel/IOC feeds are not integrated beyond vendor defaults to avoid confounding.

Identification of the IT Attack Chain. To ground the IT part of our testbed in realistic enterprise threats, we systematically identified common Active Directory (AD) attack vectors. This process combined empirical data from large-scale AD security audits (PingCastle - a tool that is used by companies to identify Active Directory misconfigurations gave us their metrics for research purposes [40]) with literature reviews and penetration testing repositories [46]. We prioritized attacks enabled by prevalent misconfigurations, such as weak or missing LDAP signing, unconstrained delegation, or legacy authentication protocols. Each candidate technique was mapped to the MITRE ATT&CK framework and evaluated for both frequency and impact. This resulted in a curated set of techniques such as Kerberoasting [37], AS-REP Roasting [36], SID-History abuse [32], forced authentication [35], and password spraying [34], that are not only frequently exploited by attackers in real-world intrusions, but also reproducible in a controlled lab setting. Incorporating those vectors ensured that our AD environment reflects realistic attack surfaces.

Since we set up an Active Directory in the OT-DMZ, we identified a domain trust vulnerability [32], and weak access control [38] as potential gateways for attackers to the OT-DMZ. Those misconfigurations serve as a starting point for the OT Attack Chain.

Identification of the OT Attack Chain. To design a realistic OT attack chain, we first surveyed state-of-the-art adversary techniques targeting both IT and OT environments. [13] The methodology proceeded in three steps: (i) assessing the opportunities provided by the given infrastructure (trust relationships, weak access control), (ii) theoretically developing attack paths based on known adversary techniques, and (iii) embedding typical weaknesses observed in industrial networks as identified in prior literature. The resulting chain leveraged native system functions and common misconfigurations rather than unrealistic exploits.

For example, access into the OT domain was achieved via an IT account valid in both domains, and lateral movement relied on valid credentials. Techniques incorporated credential theft, injection attacks against web applications and ICS interfaces and eavesdropping. Vulnerabilities such as unencrypted traffic, outdated software, and insecure web services were systematically embedded to reflect the conditions documented in real critical-infrastructure incidents.

This approach ensured that the OT attack chain remained both feasible for penetration testers and representative of realistic adversary behavior.

Based on these findings, we implemented realistic misconfigurations and recently disclosed vulnerabilities as deliberate attack vectors.

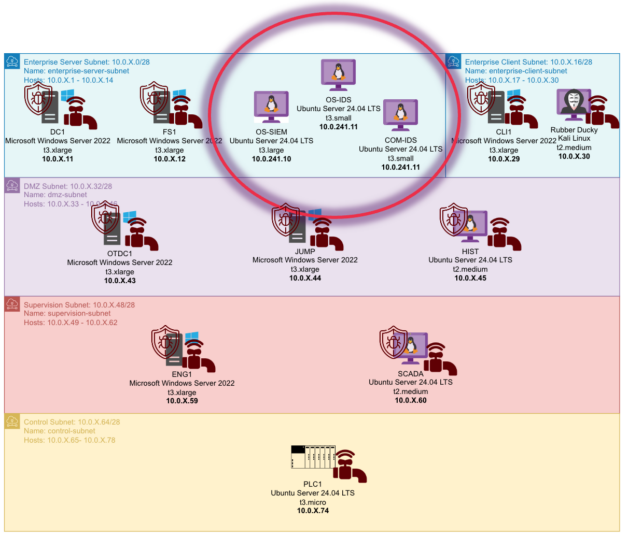

Figure 1: StealthCup multi-level IT/OT infrastructure, including NIDS and SIEM (circled in red). The bug symbol indicates installations of HIDS/EDR/XDR, while the tap indicates network-tapping.

IT Realism. The enterprise IT environment consisted of a classical Active Directory (AD) domain with typical Windows-based components: a domain controller (DC1), a file server (FS1), and a client workstation (CLI1), all running Microsoft Windows Server 2022. The AD domain provided central authentication, group policy management, and file sharing services, reflecting a standard enterprise setup. This design mirrors common enterprise environments where compromise often begins from a client or contractor system, and attackers subsequently leverage AD misconfigurations or credential abuse to escalate privileges and move laterally.

OT Realism. The OT environment was modeled after a segmented industrial control system, including both Windows and Linux-based hosts. Core components included an OT domain controller (OTDC1), a jump server (JUMP) for remote access into the OT-DMZ, and a dedicated engineering workstation (ENG1) running *PLCnext Engineer* version 2024.0.4 on Windows Server 2022. The engineering workstation provided realistic development and deployment workflows for programming industrial controllers. In addition, Linux-based hosts simulated industrial services: a historian server (HIST) built on Grafana 11.5.2 and Telegraf 1.32.3 for telemetry collection, and a SCADA server (SCADA) based on the open-source *ScadaLTS* platform, deployed on Ubuntu Server 24.04 LTS. Finally, the core industrial process was represented by a digital twin of a *Phoenix Contact AXC F 2152 PLC*, virtualized by CyberDanube [6] with firmware version 2021.6.0 and hardware version 04. This PLC supported industrial protocols such as Modbus and included vendor-typical features such as a preconfigured administrative account.

Furthermore, we programmed the virtualized PLC to simulate the control of a water pump. Since the PLC was executed as a QEMU-based virtual machine, no physical I/O was connected, and the sensor and actuator signals were emulated, a technique commonly used in digital twin environments for ICS security research [19, 47]. Pump commands were exposed via Modbus TCP and UDP servers, enabling control through the ScadaLTS HMI [43] (see Fig. 2a) and recording of pump states in the Grafana-based historian [15] (see Fig. 2b).

4.3 IDS Deployment

To cover both host- and network-level monitoring, we integrated a mix of commercial and open-source IDS solutions:

- **Host-based IDS (HIDS/EDR):** Vendor A EDR[1] IDP/XDR (commercial), Wazuh (free open-source).
- **Network-based IDS (NIDS):** Vendor B NIDS[1] NIDS (commercial), Suricata (free open-source).

IDS sensors were placed across enterprise and OT zones to capture relevant visibility.

Commercial IDS vendors were approached through direct outreach to industry partners, while open-source solutions were included to ensure reproducibility and comparability. Several vendors declined participation for various reasons: Darktrace and ExtraHop emphasized that their anomaly-detection approaches require a minimum deployment size of roughly 500 assets to generate meaningful baselines, making our testbed too small for evaluation. Nozomi and SentinelOne did not respond to our inquiries, while Corelight expressed concerns about the uncertainties in the evaluation process. IronNet explicitly declined, citing the potential internal costs of participation as a prohibitive factor. Ultimately, we integrated two widely used commercial platforms alongside open-source IDS to enable a representative yet feasible evaluation.

Before the competition, we manually tested the predefined attack vectors and authored baseline detection rules for Wazuh and Suricata to ensure that both IDS solutions generated meaningful alerts. All attack vectors were successfully detected during these manual tests in the tuned configurations of Wazuh and Suricata.

4.4 Final Infrastructure and Instrumentation

Implementation. The full testbed was implemented as Infrastructure-as-Code using Terraform and Ansible. This includes enterprise IT components (Windows domains, Linux servers), OT components (emulated PLCs, supervisory servers), and deliberate misconfigurations. This approach ensures reproducibility, automated resets, and controlled randomization.

Network Architecture. The infrastructure followed the Purdue model, representing state-of-the-art network segmentation in industrial environments [19, 47]. It was divided into four zones: Enterprise IT, OT-DMZ, Supervision, and Control. Between each zone, strict layer-4 firewall rules enforced limited connectivity, mirroring

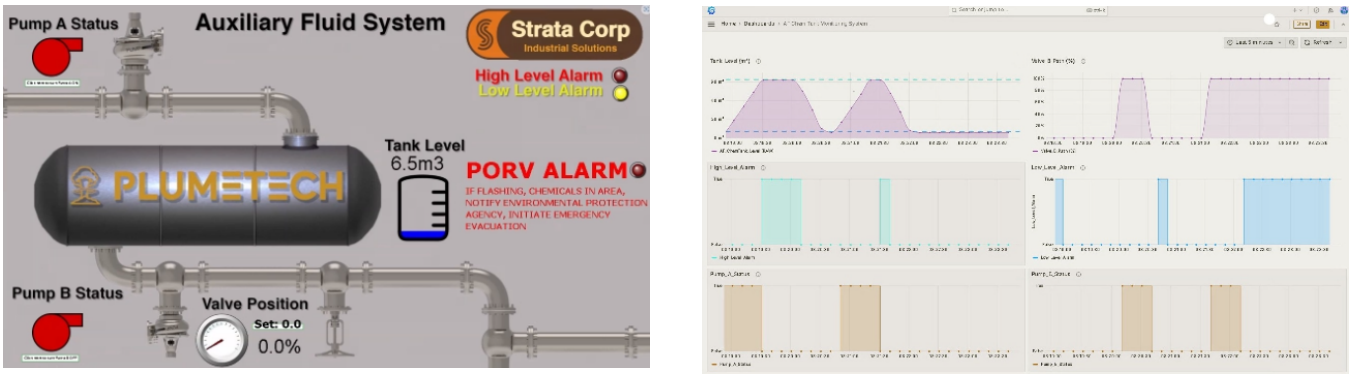

Figure 2: Virtualized PLC controlling a simulated water pump: (a) interaction via ScadaLTS HMI, (b) logging and monitoring via Grafana historian.

typical critical-infrastructure segmentation practices. The environment was fully virtualized in AWS, with separate subnets provisioned for each Purdue layer and a consistent addressing scheme to distinguish parallel instances. For example, the domain controller in every testbed was assigned the address `10.0.X.11`, where `X` denotes the team-specific instance (`10.0.1.11` for Team 1, `10.0.2.11` for Team 2, etc.). Figure 1 illustrates the overall setup. The setup was validated in collaboration with four critical-infrastructure operators from the energy and transport sectors to ensure that both IT and OT components reflected real-world deployments.

Validation. To ensure realism and reproducibility, all attack techniques were systematically mapped to the corresponding MITRE ATT&CK tactics, techniques, and sub-techniques. For IT-related attacks we used the Enterprise Matrix, while for OT-specific actions we primarily relied on the ICS Matrix. Given the close interconnection between IT and OT environments, as also reflected in our testbed, selected Enterprise Matrix techniques were additionally applied to ICS components where relevant.

The theoretically developed attack chain was first reviewed internally with colleagues from professional penetration testing and information security consulting teams to assess both feasibility and operational realism. Subsequently, three dedicated validation sessions were conducted with OT domain specialists. The first session involved experts from T-Systems Austria [48] and CyberDanube [6]; the following two involved practitioners from Linz AG [26], Verbund AG [50], and again CyberDanube [6]. During these sessions, the proposed attack scenarios and techniques were repeatedly scrutinized with respect to realism and applicability in real-world infrastructures. The feedback obtained from these discussions led to refinements of the attack paths.

Finally, all techniques were tested in practice to verify their feasibility within realistic timeframes and under practical preconditions. Where necessary, adjustments were made to ensure that each attack step was both technically viable and aligned with adversary behavior observed in operational technology networks. This iterative process ensured that the deployed environment and attack chains were not only theoretically grounded but also validated as realistic and relevant to actual critical-infrastructure operations.

Attack Scenario. The StealthCup environment supported multiple intrusion paths across IT and OT domains. Figure 3 illustrates the implemented multi-stage kill chain with the primary objectives and representative detections. The following description highlights one such path, executed by *Team 2*, which successfully completed both objectives.

Initial access was obtained by relaying SMB login requests generated by a scheduled task on the client workstation (CLI1). While such credentials are typically harvested through LLMNR/NetBIOS poisoning [33] in real-world environments, platform constraints in Amazon Web Services (AWS) required scheduled task-based capture in our setup. Using the captured credentials, domain user accounts were enumerated via SID brute-forcing, followed by an AS-REP roasting attack [36] against accounts without Kerberos pre-authentication, which yielded a low-privileged domain account.

Further enumeration of SMB shares on the fileserver exposed sensitive documents and archived emails, including a photograph of the PLC that revealed its default password. Privilege escalation was achieved by exploiting an ESC1 misconfiguration [52] in Active Directory Certificate Services (AD CS). This enabled the request of a certificate on behalf of a Domain Administrator, resulting in elevated privileges and persistence through the creation of a new administrator account-fulfilling the IT objective.

The attackers then pivoted to a jump host in the OT-DMZ. A misconfigured backup script granted local administrator access, which was exploited to extract a KeePass master password from memory via CVE-2023-32784 [39]. The recovered database disclosed credentials for lateral movement to the engineering workstation. Finally, using the default PLC password identified earlier, Team 2 deployed a modified control project that set the tank-level threshold to zero. Uploading this logic disabled safety mechanisms and triggered the simulated release of toxic chemicals, thus completing the OT objective.

4.5 Randomization

During the competition, participants were allowed to reset their infrastructure to recover from mistakes and attempt the objectives again with a fresh score. To prevent unfair advantages across resets, all usernames, passwords, and password hashes were randomized, while the overall attack paths and vulnerabilities remained unchanged. This ensured that teams could refine their techniques and strategies across successive attempts, but could not simply reuse credentials or artifacts obtained in previous runs. For scalability, ten

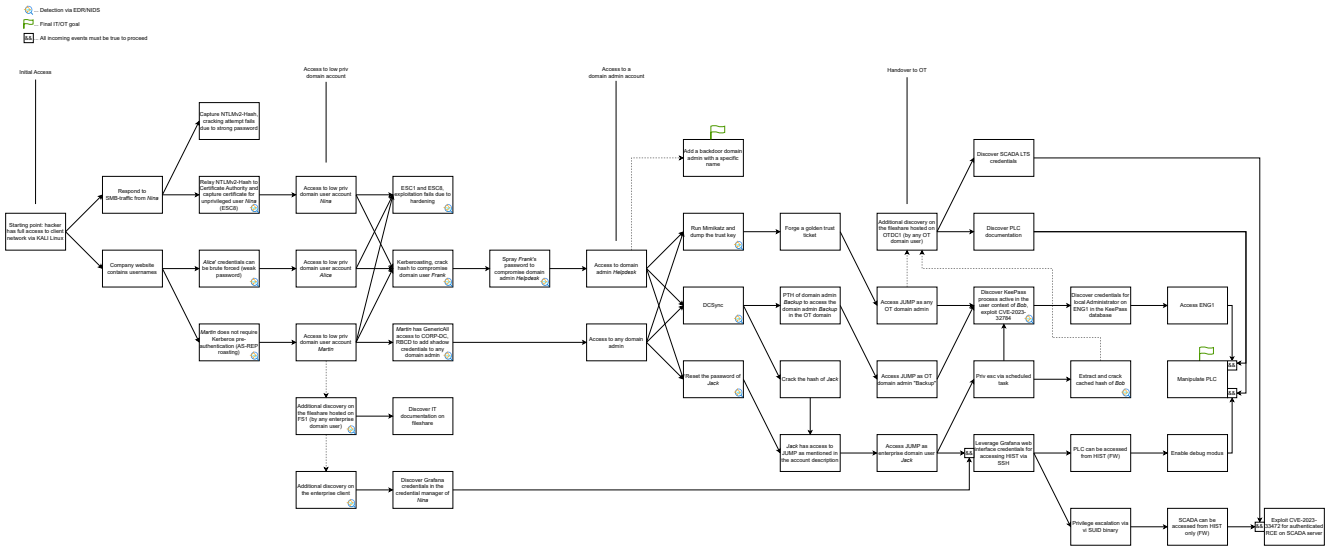

Figure 3: StealthCup multi-staged attack chain that has been implemented including the main objectives and the detections validated in our test run.

preconfigured environment templates were prepared in advance. When a team initiated a reset, its environment was replaced by a freshly randomized instance, and the team was temporarily locked.

4.6 Competition Design and Scoring

Teams must solve objectives while minimizing IDS detections. In the *Enterprise Cup*, the objective is to create a domain administrator account named *plumber* (IT Flag). In the *OT Cup*, the objective is to trigger an unsafe state in the PLC such that the process safety valve is activated, simulating a chemical spill (OT Flag).

Scoring. Each team accumulates penalty points based on the number and severity of IDS alerts they trigger. All detections are reported in real time on the competition dashboard for all participants. Each IDS $k \in \mathcal{K}$ classifies the alerts into four severity levels $s \in \{\text{Critical}, \text{High}, \text{Medium}, \text{Low}\}$. Let $a_{t,k,s}$ be the number of alerts of severity s triggered by team t, and $w_{k,s}$ the weight assigned to severity s for IDS k. The penalty is computed as

$$P_t = \sum_{k \in \mathcal{K}} \sum_{s} w_{k,s} \cdot a_{t,k,s}.$$

The goal of the competition is to minimize this penalty: the team with the lowest final value wins.

Resets. At any time, a team may reset its infrastructure, which restores the environment and the penalty points. The first reset is free, but subsequent resets increase the multiplicative factor $\mu_t \geq 1$ applied to the penalty:

$$\tilde{P}_t = P_t \cdot \mu_t.$$

Resets also temporarily block access to the infrastructure for 15 minutes, discouraging overuse while still allowing recovery from early mistakes.

Weighting and Validation. Weights $w_{k,s}$ were introduced to account for differences in severity scales and alert generation behavior across IDS solutions. Lower-severity alerts generally occur far more frequently than critical ones, both because they are more easily triggered by benign behavior and because early-stage attacks, such as reconnaissance or brute-force scanning, generate large numbers of low-impact alerts. Conversely, high-severity alerts are more salient and are more likely to be prioritized by analysts. To prevent scoring from being dominated by noisy, low-impact events, higher-severity alerts must therefore be assigned disproportionately larger weights.

In addition, different IDS may produce very different alert volumes per severity category when observing the same or similar attacks. To enable meaningful aggregation, alerts from different systems must contribute comparably after weighting. Weights were derived empirically during an initial internal training run by analyzing severity distributions across IDS and selecting values that ensure critical alerts have the highest influence, that the influence decreases monotonically with severity, and that weighted contributions are comparable across systems. As a result, Wazuh alerts were weighted 50,3,1,0.05 for Critical,High,Medium,Low, while Vendor A EDR alerts were weighted 500,30,20,10. For example, this means that a single critical alert from Vendor A EDR contributes the same amount to the final score as 25 medium-severity alerts from the same IDS or 10 critical alerts from Wazuh.

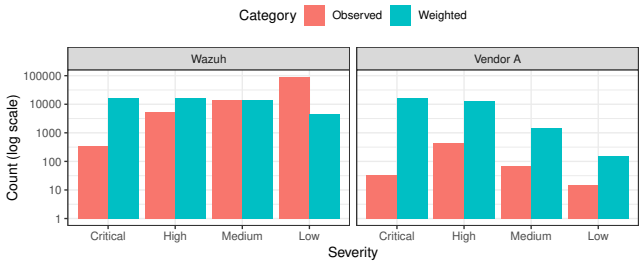

Figure 4: Raw alert counts (red) vs. weighted contributions (blue) for Wazuh and Vendor A EDR during an internal training run. Weighting normalizes IDS contributions and ensures high-severity alerts dominate scoring despite lower frequency.

Figure 4 illustrates the effect observed in data from our initial internal training run. The red bars show raw alert counts per severity category for Wazuh and Vendor A EDR. Wazuh exhibits a strong skew toward lower-severity alerts, whereas Vendor A EDR shows lower overall volume with more balanced distribution across severities. After applying the weights described above (blue bars), both IDS contribute roughly equally to the score, with lower-severity alerts having progressively less influence. When a team chooses to validate an objective, its current penalty is frozen and must be accompanied by a short attacker write-up describing how IDS detections were bypassed. Only validated penalties count for the leaderboard. The final rankings are determined by the lowest penalty.

Scoring Design Rationale and Limitations. The scoring system was designed as a game mechanic to incentivize stealthy behavior, not as a measurement model. The primary challenge was normalizing penalties across fundamentally different IDS alerting philosophies: Wazuh, configured for comprehensive detection, generates high alert volumes, including debug telemetry classified as low severity, with single attack steps often triggering multiple correlated alerts. Vendor A EDR, by contrast, employs selective high-confidence alerting with a lower volume, where alerts typically indicate genuine attack activity. Without normalization, Wazuh's verbose logging would dominate penalties regardless of attack success.

These weights balance alerting philosophies for competitive fairness, but do not measure absolute IDS value. Teams likely adapted to avoid high-penalty alerts. Our core findings focus on binary detection results (technique detected vs. not detected) and qualitative patterns (e.g., 11 techniques evaded all IDS configurations), unaffected by penalty magnitude. While we did not conduct a formal sensitivity analysis across alternative weighting schemes, the released datasets enable future researchers to validate this through re-scoring experiments.

4.7 Event Execution

Event Preparation. We advertised the event by presenting StealthCup at security conferences and directly contacting penetration testing companies and security teams. The event was designed as a *hybrid competition*, with remote and on-site participation options. The rules of the game, the chance to win four iPhones, as well as subscriptions for hack the box and a narrative storyline (e.g., a fictional company under attack) were published to motivate participants. [1] Evasion techniques were encouraged, with the exception of denial-of-service attacks against IDS components, tampering with log files, and exploitation of the gaming backend.

Participants.

The final event included 52 participants, including internationally recognized penetration testers and consulting firms, organized into 12 teams as shown in Table 1. Teams came from five different countries --- Austria, Ireland, the Netherlands, the United Arab Emirates, and Slovenia --- reflecting both regional and international interest. In total, nine teams represented professional security companies, of which five work in security consulting, while three teams were composed of academic researchers. Each team received an isolated instance of the IT/OT testbed, consisting of 13 servers and the emulated OT infrastructure.

In total, 158 virtual machines were deployed in AWS. Each team was provided with 13 dedicated hosts, supplemented by two backend machines for orchestration and management.

5 DATA AVAILABILITY

To ensure reproducibility and foster community use, we release the data and artifacts generated in StealthCup. All datasets and code are available at [1]. The release includes infrastructure definitions, collected telemetry, and documentation of attack traces, while excluding elements that would pose safety or licensing concerns.

Released Artifacts.

- **Infrastructure-as-Code.** Full Terraform and Ansible scripts for deploying the IT/OT testbed, including Windows and Linux systems, Active Directory domains, IDS agents/rules, and representative misconfigurations.
- **Datasets.** Complete packet captures (PCAPs), IDS alerts from all deployed solutions, analyst-assigned labels (Section 6.2) and host logs from Windows and Linux machines.
- **Ground truth.** Structured attacker writeups, submitted upon objective completion, documenting step-wise TTPs. These writeups enable correlation with PCAPs, event logs, and EDR telemetry to identify missed detections.
- **Documentation.** Detailed instructions to reproduce the environments and rerun the experiments.

Limitations. Because alert labels reflect expert interpretation rather than definitive truth, we recommend that future users treat them as guidance for comparative or exploratory analysis and encourage re-labeling efforts aligned with specific research objectives. Commercial software and the PLC digital twin cannot be open sourced due to vendor licensing; however, equivalent functionality can be replicated with freely available PLCnext software using low-cost, off-the-shelf PLC controllers.

6 EVALUATION

6.1 Scoring and Strategies

We received a total of 14 attacker writeups. Eight teams successfully completed the Enterprise (IT) Cup, while only one team, Team 2, reached the OT objective and thereby won the overall competition. Team 2 concentrated on achieving full kill chain completion rather than stealth, and consequently ranked last in terms of stealth performance within the IT Cup. The main prize (four iPhones for the winning team) was awarded to the overall cup winners.

The OT challenge required advanced skills and extensive lateral movement. From participant feedback and log analysis, we found that only two additional teams had entered the OT environment during regular play. At a later stage, we offered selected participants with stronger OT expertise the opportunity to continue exploration by providing credentials to the OT domain, enabling them to escalate further.

These cases are marked as "credentials OT hint" in Table 1. Table 1 summarizes the participating teams and their progress,

Table 1: Overview of participating teams. Fields are grouped into professional companies, security consulting firms, and academic researchers. Timestamps shown in UTC.

Team ID	Description of Events	Timestamps (UTC)	Onsite	Remote	Field	Country
1	IT Flag, Reset 1, Writeup 1	10:51, 14:28, 15:35	8		Security researcher / academic	Austria
2	IT Flag, Reset 1, IT Flag, OT Flag, Writeup 1	10:16, 14:11, 15:20, 15:45, 15:55		4	Professional company (consulting)	Austria
3	IT Flag, Reset 1, IT Flag, Reset 2, IT Flag, Writeup 1	11:42, 12:45, 14:06, 15:28, 16:12, 16:19			Professional company (consulting)	UAE
4	IT Flag, Writeup 1, Reset 1, IT Flag, Writeup 2	11:27, 12:07, 14:37, 15:48, 16:32	4		Security researcher / academic	Austria
5	Credentials OT Hint	16:06	2		Professional company	Austria
6	IT Flag, Writeup 1	15:15, 15:38	4		Professional company	Austria
7	IT Flag, Reset 1, IT Flag, Writeup 1, Reset 2, IT Flag, Writeup 2	13:40, 14:12, 15:15, 15:21, 15:24, 16:12, 16:14	4		Professional company (consulting)	Slovenia
8	Reset 1, Reset 2, IT Flag, Writeup 1	13:50, 14:58, 16:15, 16:21	6		Professional company (consulting)	Austria
9	IT Flag, Writeup 1, Reset 2, IT Flag, Writeup 2	10:49, 10:57, 13:50, 14:40, 14:51	5		Professional company (consulting)	Austria
10	Writeup 1, Credentials OT Hint, OT Flag, Writeup 2	11:53, 16:05, 16:25, 16:29		6	Professional company	Netherlands
11			4	1	Security researcher / academic	Ireland
12	Writeup 1, Reset 1	(online w/o timestamp of last edit), 14:57	4		Professional company	Austria

while Figure 5 illustrates the event timeline, highlighting challenge solves and submitted detection scores (lower is better).

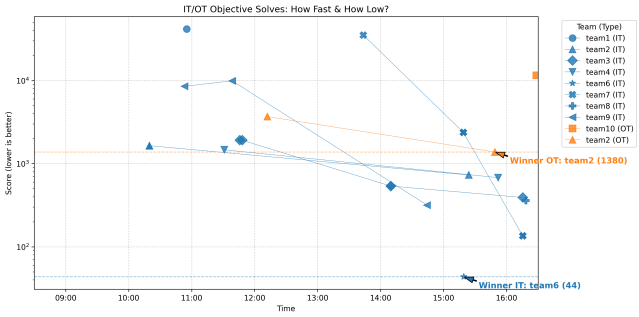

Figure 5: Timeline of the competition, showing objective solves and submitted detection scores (on a logarithmic scale, lower is better).

6.2 Evaluation of IDS Performance

After the event, the collected data was analyzed to map attacker actions to IDS detections. We reviewed the attacker writeups, attributed each step to the corresponding MITRE ATT&CK technique, and verified whether alerts were generated by the deployed IDS solutions. For each detection, we recorded the severity and whether the technique was identified with precision.

Figure 6 illustrates representative cases: Team 2, the only team to complete the full IT/OT kill chain; Team 6, the winner of the Enterprise Cup; and Team 9, another strong Enterprise Cup performer.

We evaluated seven IDS configurations, covering both open-source and commercial tools:

- IDS - Wazuh (default),
- IDS - Wazuh (custom): rules for our environment,
- IDS/EDR - Vendor A EDR[1] (default): no additional tuning,
- IDS/XDR - Vendor A EDR[1] (IDP): default, no additional tuning,
- NIDS - Suricata (ET ruleset): Emerging Threats ruleset [41], TGI ruleset [49],
- NIDS - Suricata (Custom): rules for our environment,
- NIDS - Vendor B NIDS[1] (default): no additional tuning.

Every alert was manually labeled on a five-point confidence scale by two independent security analysts (both >5 years penetration testing experience) who were involved in testbed design and attack scenario development. Classification used attacker writeups, knowledge of implemented vulnerabilities, and forensic analysis. Because no user behavior was simulated, the primary challenge was distinguishing attacks from routine system operations (scheduled tasks, service accounts, machine authentication).

Analysts used a five-point qualitative scale under these conditions: (1) alert definitely reflects benign behavior, (2) alert probably reflects benign system behavior, (3) relationship to attack activity is uncertain, (4) alert most probably indicates attack activity, and (5) alert certainly corresponds to an attack. Analysts considered patterns such as traffic from the Kali attacker machine (`10.0.*.30`), authentication from attacker IPs, and indicators of implemented attacks (Kerberoasting, AS-REP Roasting, honey token access) as malicious, while system services, management traffic (`10.0.242.*`), and pre-competition activity (before 09:30 UTC) were considered benign.

For inter-rater reliability, the second analyst independently labeled all alerts without access to the original annotations. Because the reported metric is false-positive rate, inter-rater agreement was evaluated on binary FP classification: false positive (1–2, alert on benign event) versus uncertain or true positive (3–5). Cohen's $\kappa = 0.93$ [24] indicated near-perfect agreement on the FP distinction. Primary annotations were used for all reported metrics. False positives (labels 1–2) were calculated as the fraction of alerts in these categories.

FPR differences partly reflect IDS design philosophies: Wazuh also logs benign system events as low-severity telemetry (inflating FPR) [8], while Vendor A EDR employs selective high-confidence alerting (lowering FPR but potentially missing attacks). This motivated our normalized scoring weights (Sect. 4.6). Although seven IDS configurations were deployed, 11 of 32 techniques were not detected by any system.

Table 2: False positive rates (FPR) per IDS configuration and team. Lower values indicate fewer benign events misclassified as attacks.

IDS Configuration	Team 2	Team 6	Team 9
Wazuh (default)	94.79%	95.30%	67.86%
Wazuh (custom)	93.18%	92.76%	67.99%
Vendor A EDR[1] (default)	0.00%	0.00%	0.00%
Vendor A EDR[1] (IDP)	0.00%	0.00%	0.00%
Suricata (custom)	0.00%	13.53%	0.00%
Suricata (ET ruleset)	0.00%	2.70%	0.00%
Vendor B NIDS[1] (default)	26.67%	33.33%	22.22%

This evaluation allows us to compare open-source and commercial solutions, as well as the effect of tuning versus default configurations. We report two primary metrics: (M1) *attack-centric detection coverage*, the fraction of attack steps from writeups that

triggered at least one alert; and (M2) *false-positive ratio*, the fraction of alerts judged benign is shown in Table 2. Together, these metrics provide a structured basis for assessing IDS performance under stealth-focused human adversaries.

6.3 Realism of Attacker Behavior

To evaluate the realism of attacker behavior in StealthCup, we compared the techniques exercised by participants against those documented in recent advanced persistent threat (APT) campaigns. Specifically, we used the recent CISA advisory on Volt Typhoon state-sponsored activity against U.S. critical infrastructure [9], which maps observed tactics and techniques to the MITRE ATT&CK framework. The comparison was conducted post-event, independent of the testbed design described in Section 4. We systematically assessed (i) which Volt Typhoon techniques were applicable in our IT/OT infrastructure, (ii) which were documented by participants in their attacker writeups, and (iii) which were observable in forensic artifacts (PCAPs, host logs, EDR telemetry).

The analysis reveals substantial overlap between StealthCup activity and real-world campaigns. Out of the Volt Typhoon techniques, 28 were directly applicable to our testbed, and all of them were exercised during the competition, as confirmed through manual forensic analysis of collected data. Of these, 19 techniques were described in attacker writeups and 9 were confirmed in forensic traces (e.g., Remote Service Discovery, User Discovery, Network Configuration Discovery, Process Discovery). Some reconnaissance techniques common in real-world intrusions (e.g., organizational or external network discovery) were not modeled in our scenario and thus not applicable in StealthCup. The full comparison results are provided in Appendix A.

Overall, the triangulation of applicability, participant writeups, and forensic traces demonstrates that the attacks carried out during StealthCup elicited behavior closely aligned with state-sponsored TTPs. This confirms that the resulting datasets and IDS evaluations are not synthetic or contrived, but grounded in realistic adversary tradecraft, making them suitable for IDS benchmarking and research.

7 DISCUSSION AND LIMITATIONS

7.1 Interpretation of Results

The evaluation results demonstrate that the majority of malicious activity remained undetected by the deployed IDS solutions across the analyzed competition runs. This indicates that participants were able to circumvent detection mechanisms, a behavior further encouraged by two design choices: (i) near real-time feedback on triggered detections, and (ii) the ability to reset both the infrastructure and scoring to start new runs. As shown in Figure 5, steep downward changes in detection scores along the timeline suggest that teams quickly adapted their strategies once they understood the consequences of their techniques.

The implemented attack chains in the IT domain primarily targeted Active Directory (AD). These attacks could be executed from an attacker machine not joined to the Windows domain and thus not fully monitored by HIDS or EDR solutions. This design encouraged contestants to focus on techniques detectable only by observing impact (e.g., abnormal logons, privilege escalations) rather than malware execution blocked at the endpoint.

MITRE ATT&CK Evaluations offer a valuable comparison point. While MITRE focuses on endpoint scenarios with pre-defined TTPs (e.g., the 2024 evaluations centered on ransomware and APT campaigns), StealthCup extended this approach to multi-stage AD- and OT-centric intrusions. By treating high-scoring contestant runs as distinct "threat actors" and their writeups as ground truth, we obtained insights into IDS performance across AD compromise and lateral movement phases that MITRE evaluations do not cover.

False positive rates (FPR), summarized in Table 2, revealed striking differences between commercial and open-source systems. Commercial solutions such as Vendor A EDR[1] and Vendor B NIDS[1] produced relatively few alerts but also detected fewer attacks. By contrast, open-source systems (Wazuh, Suricata) generated many more alerts, with Wazuh exhibiting particularly high false-positive rates. This effect is partly explained by Wazuh treating benign events (e.g., successful user logons) as low-severity alerts, which inflates the alert volume and complicates analysis for SOC operators.

7.2 Reflections and Lessons Learned

Because all contestants operated against identical infrastructures, they exploited the same vulnerabilities. Nevertheless, detection outcomes varied significantly across teams due to different tooling and attack chains. For instance, using `net.exe` executed via a C2 agent on a domain controller was detected by EDR, whereas executing the same command via a PsExec-like tool was not. Technique T1558.004 (Steal or Forge Kerberos Tickets: AS-REP Roasting) should have been detected but attackers appear to have circumvented the detection logic, while T1039 (Data from Network Shared Drive) was caught in earlier runs via honey token files but subsequently avoided by participants. These cases illustrate both blind spots in IDS products and the brittleness of certain detection rules.

Improving ground truth quality is critical. Currently, attacker write-ups were only required for the final validated run, preventing us from analyzing how techniques evolved across earlier attempts. Mandatory, timestamped logging of each attacker action in a structured format would enable finer-grained attribution and improve reproducibility. The reset-and-randomize mechanism enabled participants to explore vulnerabilities during early runs and refine their strategies in later attempts, as reflected in score progressions (Figure 5). While this design lowered entry barriers and encouraged learning, it reduced the fidelity of ground truth for IDS evaluation. A stricter competition format with only one permitted run would provide higher-quality attack chains but substantially increase difficulty.

Future scenarios could extend realism by modeling endpoint compromise through phishing-based initial access. This would require contestants to craft malware payloads delivered via phishing emails, executed on designated victim endpoints, and subsequently controlled via C2 infrastructure. Such an extension would more closely mirror real-world campaigns. Finally, the large number of low-severity alerts in Wazuh demonstrated that benign telemetry (e.g., user logons) can both aid forensic reconstruction and overwhelm SOC analysts in daily operations. Their utility thus depends

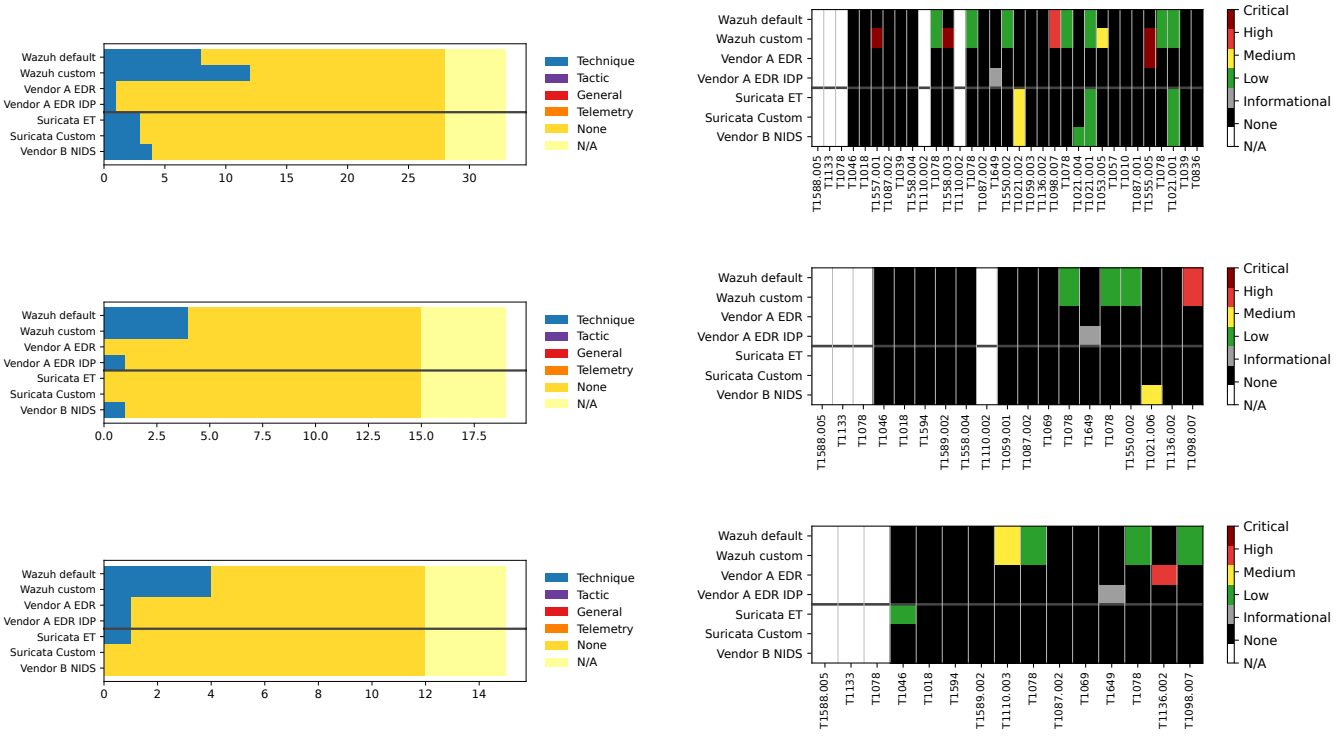

Figure 6: Comparison of MITRE Eval [30] coverage (left) and alert/score profiles (right) for Teams 2 (Ent./OT Cup), 6 (Ent. Cup Winner) and 9. (Ent. Cup)

heavily on integration with correlation rules or higher-level analytics.

The one-day format limits evaluation of long-term threat scenarios such as slow reconnaissance or delayed-activation persistence (see Section 7.3). Future iterations could address this through: (i) multi-day events with persistent range access, (ii) earlier release of infrastructure specifications allowing pre-competition reconnaissance that mirrors real APT campaigns, or (iii) integration with continuous red-team exercises evaluating IDS performance over weeks or months.

7.3 Threats to Validity

We discuss limitations and threats to validity [44] organized by the standard validity taxonomy (internal, external, construct, and conclusion validity), and outline implications for dataset reuse.

Internal Validity.

Attacker skill diversity. StealthCup participants varied in expertise (professional penetration testers, security consultants, academic researchers) and team composition (Table 1). We mitigated potential cross-team confounds through: (i) isolated infrastructure instances with randomized credentials preventing cross-team information leakage while maintaining identical attack surfaces, (ii) identical vulnerability sets and attack vectors across teams enabling comparative analysis, and (iii) multiple possible attack paths toward each objective, allowing skill-dependent strategy selection. However, we did not implement formal skill calibration mechanisms, such as stratified sampling by professional experience or pre-competition proficiency assessments, which would enable more precise characterization of how attacker capability influences detection outcomes. Future iterations could incorporate such measures to strengthen comparability across teams.

Ground truth completeness. As discussed in Section 6.2, alert labeling relied on manual forensic analysis and attacker-provided writeups. Inter-rater agreement (Cohen's $\kappa = 0.93$) indicates near-perfect consistency, but labels reflect expert interpretation rather than absolute truth. Potential bias sources include: (i) knowledge of attack writeups may influence alert interpretation, (ii) post-event labeling allows hindsight bias, and (iii) cross-system labeling by the same analysts may introduce consistency effects. However, the magnitude of observed differences between IDS systems (e.g., Wazuh 94% FPR vs. Vendor A EDR 0% FPR) unlikely to be explained solely by these biases, and because our research questions concern relative detection patterns rather than absolute detection rates, labeling uncertainties affect all systems equally. Complete raw telemetry (PCAPs, host logs, IDS alerts) has been released to enable independent verification.

Scoring and weighting design. Penalty weights were normalized to balance heterogeneous IDS alerting philosophies (verbose telemetry-focused vs. selective high-confidence) rather than measure absolute IDS value (Section 4.6). While this influenced team tactics and competition rankings, our analytical findings, which techniques evaded all IDS configurations, qualitative differences in alerting behavior, and relative detection coverage, derive from binary detection outcomes (alert generated vs. not generated) rather than penalty magnitudes, making our key findings largely independent of specific weight values. We did not conduct formal sensitivity analysis across alternative weighting schemes; the released datasets enable such validation by future researchers.

External Validity.

Scenario and infrastructure scope. The testbed was tailored to hybrid IT/OT environments with Active Directory domains and industrial control systems (Section 4.4). The testbed does not cover all attack surfaces: phishing-driven initial access, supply-chain compromise, cloud-native environments, or user-behavior-dependent attacks are out of scope, and resilience against undisclosed zero-days was not evaluated. The Volt Typhoon comparison (Section 6.3) demonstrates strong alignment with state-sponsored TTPs within the modeled domain. However, generalization to other infrastructure types (e.g., cloud-only, IoT/embedded systems, air-gapped networks) or attack contexts requiring different initial access vectors requires further validation.

Temporal constraints. The one-day format limits evaluation of long-term threat scenarios such as slow reconnaissance or delayed-activation persistence. StealthCup focuses on short-term, human-driven adaptive behaviors; multi-day events or persistent range access would be needed to study detection rule evolution and alert fatigue over longer horizons.

IDS configuration diversity. Findings reflect four IDS products in specific configurations: two open-source (Wazuh, Suricata with default and tuned rules) and two commercial (Vendor A EDR in default and IDP modes, Vendor B NIDS). Vendor selection was constrained by participation willingness. Results therefore represent a sample of the IDS landscape rather than comprehensive coverage. Notably, anomaly-detection and UEBA-focused systems requiring large-scale deployments were excluded due to testbed size constraints. Configuration choices (default vs. tuned) significantly affected detection outcomes (Figure 6), underscoring that findings reflect specific deployment states rather than inherent product capabilities.

Construct Validity.

Realism and ecological validity. While StealthCup aims to elicit realistic adversary behavior, several design choices affect authenticity. Teams began from a fixed enterprise client (Kali Linux) to provide controlled starting conditions; this models post-compromise lateral movement but excludes initial access techniques. Active blocking was disabled to ensure detection-only mode, enabling uninterrupted attack chains for benchmarking detection coverage. Near-real-time alert feedback (Section 3.1) incentivized adaptive evasion but may overestimate attacker knowledge of detection boundaries. Minimal benign user activity was present, limiting the evaluation of UEBA or anomaly-detection systems.

Evaluation metrics. Detection coverage (M1) and false-positive rate (M2) do not capture time-to-detect, alert prioritization accuracy, or forensic reconstructability. These additional metrics would require attacker-side logging or extended blue-team analysis; we prioritized participant experience and attack chain continuity over comprehensive metric collection, as mandatory instrumentation would have increased participation barriers and potentially altered attacker behavior. The choice of metrics reflects StealthCup's focus on detection feasibility under evasion pressure rather than operational SOC performance.

Conclusion Validity.

Sample size and statistical power. With 12 teams and 14 writeups (some teams submitted multiple attempts after resets), statistical inference is limited. Our findings are descriptive and comparative rather than hypothesis-testing. Patterns such as ''11 techniques evaded all IDS configurations'' demonstrate capability gaps but do not support probabilistic claims about future detection rates. Replication in future competitions would strengthen generalizability.

Confounding factors. Team performance may be confounded by factors beyond IDS effectiveness: infrastructure familiarity, tool availability, time management, and collaborative dynamics all influence attack success. While infrastructure isolation and identical starting conditions control for some confounds, residual variability remains. We interpret the results as demonstrations of evasion feasibility under realistic constraints rather than as causal attribution of performance to IDS quality alone.

Implications for Dataset Reuse.

Researchers reusing the StealthCup dataset should note that alert labels reflect expert interpretation rather than definitive ground truth and may be re-labeled for specific objectives. Scoring weights balance alerting philosophies for competitive fairness and do not measure absolute IDS value; researchers should derive metrics aligned with their specific evaluation goals. Detection patterns from this IT/OT environment may not generalize to cloud-native, IoT, or user-centric deployments. The dataset reflects IDS versions and vulnerability landscapes as of March 28, 2025. Participants varied in expertise (Table 1); without formal skill calibration, detection outcomes reflect the interaction of IDS capability with team-specific approaches. The released artifacts enable independent replication, sensitivity analysis, and extension beyond StealthCup's original scope.

8 CONCLUSION

This first StealthCup demonstrated that an evasion-focused CTF can generate realistic attack traces, and comparative insights into IDS performance. By combining reproducible IT/OT infrastructures with professional penetration testers and a stealth-oriented scoring system, we showed that StealthCup elicits attacker behavior closely aligned with documented APT campaigns while revealing blind spots across both open-source and commercial IDS solutions.

Future iterations will expand StealthCup along several dimensions. First, stronger ground truth will be ensured by introducing optional agents to log attacker activity in detail—balancing accuracy with participant acceptance. Second, involving blue teams in scoring and analysis can better approximate operational settings. Third, expanding the number of initial intrusion vectors, objectives, and IDS configurations will increase coverage and strengthen generality. Together, these steps will advance StealthCup from a domain-specific case study toward a reusable methodology for benchmarking IDS under stealth-focused, human-driven attacks.

ACKNOWLEDGMENTS

Funded by the European Union under the Horizon Europe Research and Innovation programme (GA no. 101168144 - MIRANDA). Views and opinions expressed are however those of the author(s) only and do not necessarily reflect those of the European Union or the European Commission. Neither the European Union nor the granting authority can be held responsible for them. Co-funded by the Austrian FFG Kiras project TestCat (FO999911248).

REFERENCES

[1] AIT. 2025. StealthCup2025 Repository. https://github.com/ait-cs-IaaS/StealthCup2025. (2025). Accessed: 2025-08-26.

[2] N. Athanasiades, R. Abler, J. Levine, H. Owen, and G. Riley. 2003. Intrusion Detection Testing and Benchmarking Methodologies. In *Proc. of the First IEEE Int. Workshop on Information Assurance, 2003.* 63--72.

[3] Kevin Bock, George Hughey, and Dave Levin. 2018. King of the Hill: A Novel Cybersecurity Competition for Teaching Penetration Testing. In *2018 USENIX Workshop on Advances in Security Education*. USENIX Association, Baltimore, MD.

[4] Canadian Institute of Cybersecurity. 2017. Intrusion detection evaluation dataset (CIC-IDS2017). https://www.unb.ca/cic/datasets/ids-2017.html. (2017). Accessed: 2025-08-24.

[5] CCDCOE. 2025. Excercises. https://ccdcoe.org/exercises/. (2025). Accessed: 2025-02-27.

[6] CD Security Technologies GmbH. 2025. Industrial Cybersecurity for OT & (I)IoT/Embedded Pentesting. https://cyberdanube.com/. (2025). Accessed: 2025-08-26.

[7] CERT-EU. 2019. Android exploits commanding higher price than ever before. https://cert.europa.eu/publications/threat-intelligence/threat-memo-190910-1/pdf. (2019). Accessed: 2025-08-26.

[8] Samir Achraf Chamkar, Mounia Zaydi, Yassine Maleh, and Noreddine Gherabi. 2025. Improving Threat Detection in Wazuh Using Machine Learning Techniques. *Journal of Cybersecurity and Privacy* 5, 2 (2025), 34. https://doi.org/10.3390/jcp5020034

[9] CISA. 2024. PRC State-Sponsored Actors Compromise and Maintain Persistent Access to U.S. Critical Infrastructure. https://www.cisa.gov/sites/default/files/2024-03/aa24-038a_csa_prc_state_sponsored_actors_compromise_us_critical_infrastructure_3.pdf. (2024). Accessed: 2025-08-24.

[10] C. Cowan, S. Arnold, S. Beattie, C. Wright, and J. Viega. 2003. Defcon Capture the Flag: defending vulnerable code from intense attack. In *Proceedings DARPA Information Survivability Conference and Exposition*, Vol. 1. 120--129 vol.1.

[11] CTFtime. 2025. All about CTF. https://ctftime.org/. (2025). Accessed: 2025-02-27.

[12] Andy Davis, Tim Leek, Michael Zhivich, Kyle Gwinnup, and William Leonard. 2014. The Fun and Future of CTF. In *USENIX 3GSE*. USENIX Association, San Diego, CA.

[13] Simon Freudenthaler. 2025. *Angriffserkennung und IDS-Effektivität in einer Standard-OT-Umgebung*. Master's Thesis. FH OÖ - University of Applied Sciences Upper Austria, Hagenberg, Austria.

[14] Patrik Goldschmidt and Daniela Chudá. 2025. Network intrusion datasets: A survey, limitations, and recommendations. In *Elsevier Computers & Security*. Elsevier, Berkeley, CA.

[15] Grafana Labs. 2025. Grafana -- The Open Observability Platform. https://grafana.com/. (2025). Accessed: 2025-08-24.

[16] Hack-A-Sat. 2025. World's First CTF in Space. https://hackasat.com/. (2025). Accessed: 2025-02-27.

[17] Larry Huynh, Jake Hesford, Daniel Cheng, Alan Wan, Seungho Kim, Hyoungshick Kim, and Jin Hong. 2025. Expectations Versus Reality: Evaluating Intrusion Detection Systems in Practice . In *2025 55th Annual IEEE/IFIP International Conference on Dependable Systems and Networks - Supplemental Volume (DSN-S)*. IEEE Computer Society, Los Alamitos, CA, USA, 56--62. https://doi.org/10.1109/DSN-S65789.2025.00042

[18] Industrial Software Solutions. 2021. Virsec Analysis of the Colonial Pipeline Attack. https://industrial-software.com/community/news/virsec-analysis-of-the-colonial-pipeline-attack/. (2021). Accessed: 2025-08-24.

[19] International Society of Automation (ISA). 2000--2018. Enterprise-Control System Integration (ISA-95 / IEC 62264 series). (2000--2018). Part of the ISA-95 / IEC 62264 standards defining the Purdue reference architecture.

[20] Stela Kucek and Maria Leitner. 2020. An Empirical Survey of Functions and Configurations of Open-Source Capture the Flag (CTF) Environments. *Journal of Network and Computer Applications* 151 (2020), 102470.

[21] Kyriakos G. Vamvoudakis, João P. Hespanha. 2012. Optimal Attacks for the iCTF game.

[22] Max Landauer, Florian Skopik, and Markus Wurzenberger. 2024. Introducing a New Alert Data Set for Multi-Step Attack Analysis. In *Proc. of the 17th CyberSec Experimentation and Test Workshop*. 41--53.

[23] Max Landauer, Florian Skopik, Markus Wurzenberger, Wolfgang Hotwagner, and Andreas Rauber. 2021. Have it Your Way: Generating Customized Log Datasets With a Model-Driven Simulation Testbed. *IEEE Transactions on Reliability* 70, 1 (2021), 402--415. https://doi.org/10.1109/TR.2020.3031317

[24] J. Richard Landis and Gary G. Koch. 1977. The Measurement of Observer Agreement for Categorical Data. *Biometrics* 33, 1 (1977), 159--174.

[25] L.Dhanabal and Dr. S. P. Shantharajah. 2015. A Study on NSL-KDD Dataset for Intrusion Detection System Based on Classification Algorithms. https://api.semanticscholar.org/CorpusID:16298036

[26] LINZ AG. 2025. LINZ AG für Energie, Telekommunikation, Verkehr und Kommunale Dienste. https://www.linzag.at/. (2025). Accessed: 2025-08-26.

[27] Richard Lippmann, Robert Cunningham, David Fried, Isaac Graf, Kris Kendall, Seth Webster, and Marc Zissman. 1999. Results of the DARPA 1998 Offline Intrusion Detection Evaluation.

[28] Mandiant. 2025. *M-Trends 2025*. Technical Report. Mandiant. https://services.google.com/fh/files/misc/m-trends-2025-en.pdf Accessed: 2025-08-24.

[29] Ziadoon Kamil Maseer, Robiah Yusof, Nazrulazhar Bahaman, Salama A Mostafa, and Cik Feresa Mohd Foozy. 2021. Benchmarking ML for Anomaly-Based IDS in CICIDS2017. *IEEE Access* 9 (2021), 22351--22370.

[30] MITRE. 2025. ATT&CK Evaluations. https://attackevals.mitre-engenuity.org/. (2025). Accessed: 2025-02-27.

[31] MITRE. 2025. Caldera. https://caldera.mitre.org/. (2025). Accessed: 2025-02-27.

[32] MITRE Corporation. 2025. Access Token Manipulation: SID-History Injection. https://attack.mitre.org/techniques/T1134/005/. (2025). Accessed: 2025-08-26.

[33] MITRE Corporation. 2025. Adversary-in-the-Middle: LLMNR/NBT-NS Poisoning and SMB Relay. https://attack.mitre.org/techniques/T1557/001/. (2025). Accessed: 2025-08-26.

[34] MITRE Corporation. 2025. Brute Force: Password Spraying. https://attack.mitre.org/techniques/T1110/003/. (2025). Accessed: 2025-08-26.

[35] MITRE Corporation. 2025. Forced Authentication. https://attack.mitre.org/techniques/T1187/. (2025). Accessed: 2025-08-26.

[36] MITRE Corporation. 2025. Steal or Forge Kerberos Tickets: AS-REP Roasting. https://attack.mitre.org/techniques/T1558/004/. (2025). Accessed: 2025-08-26.

[37] MITRE Corporation. 2025. Steal or Forge Kerberos Tickets: Kerberoasting. https://attack.mitre.org/techniques/T1558/003/. (2025). Accessed: 2025-08-26.

[38] MITRE Corporation. 2025. Valid Accounts: Local Accounts. https://attack.mitre.org/techniques/T1078/003/. (2025). Accessed: 2025-08-26.

[39] NIST. 2025. CVE-2023-32784. https://nvd.nist.gov/vuln/detail/cve-2023-32784. (2025). Accessed: 2025-08-28.

[40] PingCastle. 2025. PingCastle -- Active Directory Security Assessment Tool. https://www.pingcastle.com/. (2025). Accessed: 2025-08-24.

[41] Proofpoint. 2025. Emerging Threat Pro Ruleset. https://www.proofpoint.com/us/threat-insight/et-pro-ruleset. (2025). Accessed: 2025-08-26.

[42] Marcus J Ranum. 2001. Experiences benchmarking intrusion detection systems. *NFR Security White Paper* (2001).

[43] ScadaLTS Community. 2025. ScadaLTS -- Open Source SCADA System. https://www.scadalts.com/. (2025). Accessed: 2025-08-24.

[44] William R. Shadish, Thomas D. Cook, and Donald T. Campbell. 2002. *Experimental and Quasi-Experimental Designs for Generalized Causal Inference* (2nd ed.). Houghton Mifflin, Boston.

[45] Iman Sharafaldin, Arash Habibi Lashkari, and Ali A. Ghorbani. 2018. Toward Generating a New Intrusion Detection Dataset and Intrusion Traffic Characterization. In *Proceedings of the 4th International Conference on Information Systems Security and Privacy (ICISSP)*. 108--116. https://doi.org/10.5220/0006639801080116

[46] Dominik Steffan. 2025. *Detection of Active Directory Attacks in Wazuh*. Master's Thesis. Ferdinand Porsche FernFH, Wiener Neustadt, Austria.

[47] Keith Stouffer, Joseph Falco, and Karen Scarfone. 2015. *Guide to Industrial Control Systems (ICS) Security*. Technical Report NIST Special Publication 800-82 Rev. 2. National Institute of Standards and Technology (NIST). https://doi.org/10.6028/NIST.SP.800-82r2 Includes Purdue Model reference for ICS segmentation.

[48] T-Systems International GmbH. 2025. Digital Services for Business and Institutions. https://www.t-systems.com/. (2025). Accessed: 2025-08-26.

[49] travisbgreen. 2025. Suricata rules for network anomaly detection. https://github.com/travisbgreen/hunting-rules. (2025). Accessed: 2025-08-26.

[50] VERBUND AG. 2025. Wasser, Wind und Sonne für eine saubere Energiezukunft. https://www.verbund.com/. (2025). Accessed: 2025-08-26.

[51] Verizon. 2025. *2025 Data Breach Investigations Report (DBIR)*. Technical Report. Verizon Enterprise Solutions. https://www.verizon.com/business/resources/reports/dbir/ Accessed: 2025-08-24.

[52] Will Schroeder and Lee Christensen. 2021. Certified Pre-Owned: Abusing Active Directory Certificate Services. https://specterops.io/wp-content/uploads/sites/3/2022/06/Certified_Pre-Owned.pdf. (2021). Accessed: 2025-08-26.

[53] A. Zafar, M. Yamin, B. Katt, and E. Torseth. 2024. All flags are not created equal: A deep look into CTF Scoring Algorithms. *Expert Systems with App.* (2024).

[54] ZDI. 2025. About. https://www.zerodayinitiative.com/about/. (2025). Accessed: 2025-02-27.

A COMPARISON OF VOLT TYPHOON TECHNIQUES WITH INFRASTRUCTURE AND WRITEUPS

Table 3: Volt Typhoon techniques vs. StealthCup: applicability (App), forensic evidence (For), and count of team writeups mentioning the technique (Wr).

Tactic	Technique	VT	App	For	Wr
Reconnaissance	Search Victim-Owned Websites - T1594	✓	✓	✓	10
	Gather Victim Identity Information: Email Addresses - T1589.002	✓	✓	✓	10
	Gather Victim Org Information - T1591	✓			0
	Gather Victim Network Information - T1590	✓			0
	Gather Victim Identity Information - T1589	✓	✓	✓	10
	Search Open Websites/Domains - T1593	✓	✓	✓	10
	Gather Victim Host Information - T1592	✓			0
Resource Development	Obtain Capabilities: Exploits - T1588.005	✓	✓	✓	13
	Acquire Infrastructure: Botnet - T1583.005	✓			0
	Compromise Infrastructure: Botnet - T1584.005	✓			0
	Compromise Infrastructure: Server - T1584.004	✓			0
Initial Access	External Remote Services T1133	✓	✓	✓	13
	Valid Accounts T1078	✓	✓	✓	14
	Exploit Public-Facing Application T1190	✓			0
Execution	Command and Scripting Interpreter: PowerShell - T1059.001	✓	✓	✓	1
	Command and Scripting Interpreter: Windows Command Shell - T1059.003		✓	✓	6
	Windows Management Instrumentation - T1047	✓	✓	✓	0
	Command and Scripting Interpreter: Unix Shell T1059.004	✓	✓	✓	0
Persistence	Modify Registry - T1112	✓	✓	✓	0
	Create Account: Domain Account - T1136.002		✓	✓	10
	Account Manipulation: Additional Local or Domain Groups - T1098.007		✓	✓	11
Privilege Escalation	Exploitation for Privilege Escalation - T1068	✓	✓	✓	2
	Scheduled Task/Job: Scheduled Task - T1053.005		✓	✓	3
Credential Access	Brute Force: Password Cracking - T1110.002	✓	✓	✓	5
	Steal or Forge Kerberos Tickets: AS-REP Roasting - T1558.004		✓	✓	4
	Steal or Forge Kerberos Tickets: Kerberoasting - T1558.003		✓	✓	2
	OS Credential Dumping: NTDS T1003.003	✓	✓	✓	2
	Unsecured Credentials T1552	✓	✓	✓	1
	Credentials from Password Stores: Credentials from Web Browsers - T1555.003	✓	✓	✓	0
	Adversary-in-the-Middle: LLMNR/NBT-NS Poisoning and SMB Relay - T1557.001		✓	✓	4
	Steal or Forge Authentication Certificates - T1649		✓	✓	10
	Credentials from Password Stores: Password Managers - T1555.005		✓	✓	1
Discovery	Network Service Discovery - T1046	✓	✓	✓	11
	Remote System Discovery - T1018		✓	✓	0
	Account Discovery: Domain Account - T1087.002	✓	✓	✓	8
	Permission Groups Discovery - T1069	✓	✓	✓	6
	Connection Proxy - T1090.002	✓			0
	System Owner/User Discovery - T1033	✓	✓	✓	0
	System Network Configuration Discovery - T1016	✓	✓	✓	0
	Indicator Removal: File Deletion - T1070.004	✓	✓	✓	0
	Application Window Discovery - T1010	✓			0
	System Service Discovery - T1007	✓	✓	✓	1
	File and Directory Discovery - T1083	✓	✓	✓	5
	Process Discovery - T1057		✓	✓	1
	Account Discovery: Local Account - T1087.001	✓	✓	✓	0
Lateral Movement	Use Alternate Authentication Material: Pass-the-Hash - T1550.002	✓	✓	✓	9
	Remote Services: SMB/Windows Admin Shares - T1021.002		✓	✓	3
	Remote Services: Remote Desktop Protocol - T1021.001	✓	✓	✓	3
	Use Alternate Authentication Material: Pass the Ticket - T1550.003	✓	✓	✓	0
	Remote Service Session Hijacking - T1563	✓			0
	Remote Services: Cloud Services T1021.007	✓			0
	Remote Services: SSH - T1021.004		✓	✓	14
	Remote Services: Windows Remote Management - T1021.006		✓	✓	2