Research Square

Preprints are preliminary reports that have not undergone peer review.
They should not be considered conclusive, used to inform clinical practice,
or referenced by the media as validated information.

Structural Leakage in Host Intrusion Alert Corpora: An Evaluation Framework for ATT&CK Technique and Tactic Mapping

Emad Sherif

p26489460@my365.dmu.ac.uk

De Montfort University

Research Article

Keywords: Cyber Range, Data Leakage, Evaluation Methodology, MITRE ATT&CK, Wazuh

Posted Date: August 21st, 2026

DOI: https://doi.org/10.21203/rs.3.rs-10763135/v1

License: This work is licensed under a Creative Commons Attribution 4.0 International License.
Read Full License

Additional Declarations: The authors declare no competing interests.

Abstract

Host-based intrusion detection alerts are abundant but often difficult to interpret as evidence of adversary behaviour. This study presents a leakage-aware workflow that maps Wazuh alert telemetry, collected in a controlled Cyber Range, to MITRE ATT&CK tactics and techniques, with a specific focus on structural data leakage in rule-generated alert corpora. Using primary alert text and metadata, the workflow applies structured pre-processing, template-aware deduplication, leakage diagnostics, and grouped cross-validation by Wazuh rule identifier to mitigate inflated performance estimates caused by repeated alert templates and deterministic rule-to-label relationships. We show that naïve evaluation overestimates predictive performance, whereas leakage-aware evaluation still identifies recoverable ATT&CK-aligned signal, with tactic-level prediction generalising more consistently than fine-grained technique-level prediction. The results indicate that alert descriptions and rule identifiers can act as label shortcuts if not controlled during evaluation. The findings demonstrate that host intrusion alert corpora are structurally prone to label leakage, and that grouped, leakage-aware evaluation is a necessary methodological standard for studies attempting ATT&CK-aligned classification from rule-generated security telemetry.

1. Introduction

Cyber risk management increasingly depends on the ability to interpret low-level security telemetry as evidence of adversary behaviour, defensive gaps, and measurable organisational impact [1]. Threat-informed defence frameworks such as MITRE ATT&CK [2] provide an important behavioural vocabulary for interpreting security telemetry, yet host intrusion detection alerts are still commonly evaluated as isolated technical events rather than as structured evidence of adversary behaviour [3].

Prior work demonstrated that task-specific semantic modelling can map cyber incident narratives to MITRE ATT&CK techniques with strong performance [4]. However, that work focused on incident description narratives rather than primary host-based intrusion alerts generated during realistic cyber-attack activity, and it evaluated performance using conventional train–test splits. Despite growing interest in ATT&CK-based analytics, relatively little work has examined whether such mappings remain valid once the structural leakage risks inherent to rule-generated alert corpora are properly controlled for [5].

This paper addresses this gap using Wazuh [6], an open-source host intrusion detection system [7], with alert data collected in a controlled experiment within a Cyber Range. Testing shows that Wazuh rule identifiers and alert descriptions map almost deterministically to available ATT&CK labels, meaning that naïve train–test splits can substantially inflate performance estimates. Accordingly, the central challenge is not merely to classify alerts, but to map them in a leakage-aware way that supports operational cyber risk interpretation [5]. To address this challenge, the paper develops and evaluates a leakage-aware framework for assessing whether ATT&CK-aligned alert classification remains meaningful when repeated templates and rule-family leakage are controlled.

The remainder of the paper is organised as follows: Section two presents the related work. Section three introduces the methodology. Section four reports the results. Section five discusses implications and limitations. Section six concludes the paper.

2. Related Work

2.1. Threat-Informed Defence

MITRE ATT&CK has become one of the dominant frameworks for describing adversary behaviour because it provides a public, behaviour-oriented knowledge base of tactics and techniques derived from real-world observations rather than purely theoretical threat models [8]. The CIS Critical Security Controls offer a prioritised set of measurable safeguards designed to address the most prevalent attacks in modern enterprise, cloud, and hybrid environments [9]. The existence of formal mappings between ATT&CK and CIS resources has strengthened the practical relevance of these frameworks by making it possible to move from observed adversary behaviour to corresponding defensive actions and implementation priorities [10].

Although ATT&CK and CIS Controls are widely used in threat-informed defence, much of their practical use in the literature remains framework-centric rather than telemetry-centric [5]. In many cases, ATT&CK is used to organise detections, threat reports, or adversary emulation exercises, while CIS Controls are treated as governance-oriented safeguards or compliance checklists [4]. What is less often demonstrated is how low-level operational telemetry, particularly host intrusion detection system alerts, can be systematically transformed into ATT&CK-aligned, control-aware, and metrics-driven cyber risk evidence. This distinction matters because a framework is only operationally valuable if it can be meaningfully grounded in the alert data that organisations consume during monitoring and triage.

2.2. Host-Based Intrusion Detection

Host-based intrusion detection systems remain important because many security relevant observations occur at the host layer rather than at the network perimeter [11]. Security events, such as *authentication* events, *process* activity, *file integrity* changes, *registry* modifications, and *rootkit* indicators often provide evidence of malicious behaviour that cannot be fully reconstructed from network monitoring alone [12]. Wazuh is particularly relevant in this context because it combines log analysis, file integrity monitoring, registry monitoring, rootkit detection, real-time alerting, and active response across Windows and Unix-like platforms, thereby producing a diverse range of host-level alerts with both textual and structured characteristics [13].

The use of Wazuh in the literature has more often focused on deployment, detection utility, or host protection functionality than on ATT&CK-aligned cyber risk interpretation [13, 14, 15]. Wazuh has generally been treated as a source of alerts rather than as a source of structured data for evaluating ATT&CK-aligned classification under realistic leakage constraints. This is an important limitation

because host-alert datasets are often repetitive, rule-driven, and template-bound, making them vulnerable to evaluation artefacts under naïve train–test splitting [5].

The present study differs from conventional uses of Wazuh by treating its alerts not merely as detection artefacts only, but as candidate inputs to a broader interpretation pipeline. This repositions host intrusion alerts from endpoint monitoring outputs to cyber risk evidence, while also exposing the methodological challenge that such alerts are frequently repetitive, rule-driven, and structurally prone to leakage if evaluated naively.

2.3. From Incident Narratives to Host Alerts

Transformer-based semantic models have substantially improved the ability to compare technical text using dense representations rather than exact lexical overlap [16]. Sentence-BERT showed that *siamese* and triplet architectures can produce semantically meaningful sentence embeddings suitable for similarity search, clustering, and efficient large-scale comparison, thereby overcoming some of the computational limitations of cross-encoder style matching [17]. This development is important for cybersecurity because many security relevant artefacts, including incident narratives, alerts, and threat reports, contain specialised language that benefits from semantic rather than purely keyword-based comparison.

Prior work extended this line of reasoning by operationalising incident-to-ATT&CK mapping using a fine-tuned sentence-transformer model, demonstrating that cyber incidents can be mapped to structured adversary knowledge with strong semantic performance [4]. In this sense, the paper does not emerge from a methodological vacuum; rather, it builds on an established semantic mapping pipeline and asks whether a similar logic can be transferred from incident narratives to Wazuh security events, and, critically, whether it survives leakage-aware evaluation when applied to a structurally different and more repetitive data source.

Cyber incident descriptions and host intrusion detection alerts are not the same kind of text. Incidents are often longer, more descriptive, and more explicitly behavioural, whereas Wazuh alerts are shorter, more repetitive, and more tightly coupled to local rules, decoders, and templates [14]. As a result, methods that perform well on incident style text cannot simply be assumed to generalise to host intrusion detection data without careful evaluation. The present study therefore extends prior semantic mapping work into a more operationally realistic but methodologically harsher setting.

2.4. Representation Learning, Contrastive Methods, and Evaluation Challenges

The broader representation learning literature has shown that the quality of learned embeddings depends not only on model architecture but also on training signal and evaluation design [26]. Hard-negative and contrastive learning methods have been shown to improve representation learning when semantically similar but distinct examples must be separated, and these ideas have influenced recent work on semantic modelling across domains [27]. In cybersecurity, such methods are particularly

relevant because many categories are lexically similar but operationally distinct, making naïve similarity modelling insufficient [28].

The more immediate methodological challenge is not the absence of sophisticated representation learning, but the risk of evaluation bias caused by repeated templates and deterministic rule-to-label mappings in Wazuh data. Host intrusion alerts differ from many generic Natural Language Processing (NLP) datasets because they often arise from fixed rule families, predefined descriptions, and structured decoders [29]. This creates a situation in which conventional train–test splits may reward memorisation of alert templates rather than genuine semantic generalisation, producing performance that appears impressive while saying very little about how well the model would handle unseen rule families or more realistic deployment conditions.

This is where the literature remains relatively underdeveloped. While cybersecurity NLP related studies often emphasise better embeddings, better fine-tuning, or better label mappings, they less often foreground the possibility that the dataset itself may contain strong leakage channels that dominate apparent performance [30, 31, 32]. Therefore, this paper contributes by foregrounding leakage-aware evaluation as a first-order methodological requirement rather than treating it as a secondary validation step.

3. Methodology

3.1. Study Design

This study adopts a primary-data, leakage-aware experimental design to examine whether Wazuh host intrusion alerts generated in a controlled Cyber Range can be transformed into ATT&CK-aligned behavioural evidence through leakage-aware mapping. The research question is: How robust is ATT&CK-aligned mapping of Wazuh host intrusion alerts once structural data leakage in the alert corpus is explicitly identified and controlled? The study seeks not only to test ATT&CK-aligned mapping feasibility, but also to establish whether such mappings remain meaningful when evaluated under conditions that explicitly control for duplication and structural label leakage.

The study is positioned as a methodological and operational validation exercise, rather than as a straightforward predictive modelling paper. This distinction is important because host intrusion alert datasets differ fundamentally from narrative security incident descriptions: they are shorter, more repetitive, more strongly shaped by predefined rule logic, and therefore more vulnerable to over-optimistic evaluation if standard random train–test procedures are used. Accordingly, the methodology is designed to answer not only whether ATT&CK-aligned host-alert classification remains valid under leakage-aware evaluation, but whether it can be demonstrated in a way that is scientifically defensible and operationally interpretable.

The main contribution of the methodological design is twofold. First, it characterises structural leakage in Wazuh alert corpora by analysing duplication and label purity across rule identifiers, descriptions, and

decoders. Second, it treats leakage-aware evaluation as a first-order requirement, thereby ensuring that apparent predictive performance is interpreted in relation to the structural properties of the alert corpus rather than assumed to reflect genuine semantic generalisation.

3.2. Cyber Range Architecture and Attack Scenario

This work uses a Cyber Range, a controlled environment comprising seven network segments.

3.2.1. Infrastructure and Deployment

The Cyber Range was deployed on Proxmox VE (v8.3) [25], a widely used open-source hypervisor platform well-suited to academic security research environments due to its support for isolated virtual network segments and snapshot-based reproducibility [26].

The environment comprised seven network segments: a DMZ, and six internal subnets representing IT, Finance, Human Resources, Marketing, and Operations. This segmentation is important because subnet boundaries governed the attack path progression described below and directly explains the heterogeneity of alert families observed across different host types. A total of 38 virtual machines (VMs) were monitored by Wazuh v4.10.2. The agent population consisted of 25 Windows endpoints: 1 VM running Windows Server 2012, 7 VMs running Windows 10 and 17 VMs running Windows 7. In addition to 13 Linux endpoints comprising 7 Debian, 5 Ubuntu, and 1 CentOS machine. This mixed platform composition directly explains the dual-platform character of the alert corpus.

3.2.2. Attack Scenario and TTPs

The attack scenario simulated a multi-stage, multi-subnet intrusion campaign consistent with an advanced persistent threat (APT) pattern, progressing from initial reconnaissance through lateral movement to physical-process impact. All IP addresses used in the simulation were internal to the Cyber Range environment (i.e., no real external network traffic was involved).

The attack scenario comprised five phases. In Phase 1, a simulated external attacker performed active scanning of the DMZ subnet and then the CRM Server to identify the initial target. In Phase 2, the attacker exploited a weak default credential on the CRM Server, successfully brute-forcing authentication and subsequently gaining access via valid accounts, thereby establishing an initial foothold. In Phase 3, using the CRM Server as a pivot, the attacker performed iterative network service discovery, lateral movement, and credential access across the Finance subnet, the Operations subnet, and ultimately the Engineering Workstation, where a command-and-control channel was established and persistence was achieved using default accounts. In Phase 4, the attacker performed indicator removal and account access removal from the Engineering Workstation, consistent with post-exploitation clean-up. Table 1 provides a summary of the attack phases and the associated ATT&CK techniques.

Table 1
Attack phases, paths, and ATT&CK technique mappings.

Target	Tactic	Technique
DMZ subnet	Reconnaissance	Active Scanning
CRM Server	Credential Access	Brute Force
CRM Server	Defence Evasion	Valid Accounts
Finance subnet	Discovery	Network Service Scanning
Finance Lead Workstation	Credential Access	Brute Force
Inventory Server	Credential Access	Brute Force
Engineering Workstation	Persistence	Default Accounts
Engineering Workstation	Defence Evasion/Impact	Indicator Removal; Account Removal

3.2.3. Dataset Overview

The Wazuh alert corpus was collected across both Windows and Linux hosts and includes a heterogeneous mix of event families, such as *successful* and *failed authentication* activity, *integrity monitoring* alerts, *service changes*, *SSH*- and *FTP*-related activity, *scheduled-task* artefacts, *Windows application* and *system* events, and *compliance-oriented* monitoring outputs. This diversity provides the basis for testing whether host intrusion alerts can support ATT&CK-aligned behavioural interpretation across multiple operational contexts rather than within a single host type or alert family. Table 2 summarises a representative subset of monitored hosts and the event families observed, illustrating the contextual heterogeneity of the corpus.

Table 2
Attack phases, paths, and ATT&CK technique mappings.

Host / Agent	OS	Representative Event Families	Count
CRM Server	Linux	sshd: Attempt to login using a non-existent user. sshd: authentication failed.; PAM: User login failed.	218
Inventory Server	Linux	sshd: authentication success. PAM: Login session opened.	133
Finance Lead Workstation	Windows	Windows Logon Success. Service logon event registered from non-standard service account.	67
FTP Server	Linux	vsftpd: FTP session opened. vsftpd: FTP Authentication success.	25
Git Server	Linux	PAM: Login session opened. PAM: Login session closed. sshd: authentication success.	21
Payroll Workstation	Windows	Windows application error event. Summary event of the report's signatures. Windows Logon Success	8
VPN Server	Linux	PAM: Login session opened. Listened ports status (netstat) changed (new port opened or closed). sshd: authentication success.	5
Sales Workstation	Windows	Successful Remote Logon Detected. Windows User Logoff	4
Engineering Workstation	Linux	Integrity checksum changed.	4

The table shows that the monitored environment includes endpoint nodes and infrastructure-relevant systems such as VPN, FTP, Git, and CRM servers. The event families associated with these hosts differ substantially, ranging from brute-force login activity and FTP authentication events to service configuration changes, integrity checksum modifications, and Windows logon success messages. It should be noted that many alerts are generated based on Wazuh's default ruleset and thus include false positives, which is characteristic of out-of-the-box Host Intrusion Detection System (HIDS) deployments

[28]. The platform and role fields in Table 2 are analytical descriptors extracted from observed decoders and agent metadata.

From a methodological perspective, the Cyber Range setting offers a useful balance between realism and control. It enables the collection of rich host telemetry during adversarial activity while preserving experimental containment, reproducibility, and traceability of the alert generation context [28]. This is advantageous for a study focused on cyber risk interpretation, because it allows the linkage between observed host activity, ATT&CK behaviour, and control-oriented implications to be examined in a defensible research environment before claims are made about broader enterprise deployment.

3.3. Data Preparation

The raw Wazuh alerts were parsed from JSON output into a structured analytical dataset suitable for downstream ATT&CK-aligned analysis. For each alert, the pre-processing pipeline extracted the Wazuh *rule_id*, *rule severity*, *alert description*, *decoder*, *source location*, *agent name* and *agent identifier*, together with a normalised textual representation derived from the alert content. Where available, ATT&CK tactic and technique labels were retained. This combination of textual and contextual features was designed to support not only semantic mapping.

The original corpus contained 2,408 alerts, of which 48.38% carry ATT&CK Enterprise technique and tactic labels, while the remaining 51.62% are unlabelled. The unlabelled subset arises from three sources: (a) Wazuh default-ruleset alerts for routine system events (logon successes, service starts, netstat changes, etc.) that have no behavioural ATT&CK mapping; (b) compliance-only events associated with CIS benchmark checks that carry regulatory metadata but not ATT&CK technique identifiers. This choice preserves label consistency within the supervised evaluation but means that ICS-phase activity is not reflected in the classification experiments. The labelled subset is sufficient to support ATT&CK-based evaluation while preserving the structural characteristics needed to study duplication, rule-template repetition, and leakage-aware generalisation. Table 3 summarises the overall dataset.

Table 3
Dataset profile and labelling coverage.

Metric	Value	Interpretation
Alerts parsed	2,408	Total alert records in corpus
Text coverage (%)	100.0	Portion of alerts with usable text
ATT&CK technique coverage (%)	48.38	Portion of alerts carrying technique labels
Unique ATT&CK tactics	7	Tactic variety in corpus
Unique ATT&CK techniques	14	Technique variety in corpus
Unique rule IDs	59	Distinct Wazuh rules triggered
Unique exact texts	2,063	Distinct normalised alert texts
Duplicate text ratio (%)	14.33	Repetition rate before deduplication

Moreover, a central pre-processing step was exact-text deduplication. 2,063 of the original corpora were unique at the exact-text level, corresponding to a duplicate-text ratio of approximately 14%. Exact-text deduplication therefore removed 345 duplicated rows before baseline evaluation. This step is important because without deduplication, the same alert formulations may appear across evaluation splits, inflating performance while providing little evidence of true generalisation.

3.4. Leakage Diagnostics and Evaluation Strategy

The defining methodological feature of the study is the use of leakage-aware evaluation [29]. Preliminary analysis showed that some structural properties of the Wazuh corpus could strongly reveal the target ATT&CK label, thereby making conventional random splits methodologically unreliable. Both *rule_id* and *description* exhibited 100% label purity in the labelled subset, meaning that alerts belonging to the same rule family or description family almost always shared the same ATT&CK label. Decoder-level purity was lower (63.64% pure groups) but still showed that structural alert variables carry non-trivial label signal. These diagnostics demonstrate that naïve evaluation would risk testing template memorisation rather than semantic generalisation.

For this reason, the baseline evaluation used grouped cross-validation by Wazuh *rule_id*, rather than a conventional random train–test split. This grouping strategy was selected because *rule_id* is the most operationally meaningful source of potential leakage in the dataset whereby if alerts generated from the same rule-template family were distributed across both training and test folds, the resulting metrics would substantially overstate the real capability of the system to handle unseen alert families. Grouped evaluation therefore forces the model to confront a harder and more realistic problem, whether ATT&CK-relevant signal remains recoverable once structural overlap is constrained.

The study reports performance at both technique and tactic levels using accuracy, macro F1, and weighted F1. This combination of metrics was chosen because the dataset is imbalanced and because

accuracy alone would obscure performance on low-frequency classes. Macro F1 is particularly important in this context because it gives equal weight to all classes and therefore provides a more informative view of whether ATT&CK-aligned signal is recoverable beyond the dominant categories in the corpus [30].

3.5. Baseline Model and Feature Representation

The leakage-aware baseline uses a *nearest-neighbour semantic similarity* approach grounded in sentence-level embeddings. Each alert is represented as the concatenation of its normalised description text and decoder name, encoded using the *all-MiniLM-L6-v2* sentence-transformer model [17]. At classification time, each test alert's embedding is compared to the training set using cosine similarity, and the label of the most similar training example is assigned (1-nearest-neighbour). This approach was selected because it makes no assumptions about the distribution of unseen rule families and is interpretable wherein the classification decision can always be traced back to a specific training exemplar. Grouped cross-validation was implemented using scikit-learn's *GroupKFold* with $k = 5$, grouping by *rule_id*, ensuring that no rule family appears in both training and test folds within any split [31]. Class-empty folds were dropped from macro-F1 averaging. All experiments were conducted using Python 3.10.

3.6. ATT&CK Mapping Model

The framework maps Wazuh alerts to MITRE ATT&CK tactics and techniques using their structured text and contextual fields under the leakage-aware evaluation constraints described above. ATT&CK alignment is treated not as a trivial classification exercise, but as a problem that must be demonstrated to survive deliberate, structural leakage controls before its output can be trusted as evidence of genuine behavioural signal rather than template memorisation. Figure 1 summarises the leakage-aware evaluation workflow, beginning with exact-text deduplication, followed by label-purity diagnostics, grouped cross-validation by *rule_id*, and final evaluation at both ATT&CK technique and tactic levels.

The study is guided by the hypothesis that Wazuh alerts generated in a controlled Cyber Range contain sufficient semantic and contextual information to support ATT&CK-aligned mapping, but that naïve evaluation will overestimate this capability because rule identifiers, descriptions, and repeated templates create structural leakage. The central claim is therefore not simply that ATT&CK labels can be predicted, but that such predictions must remain meaningful when evaluated under rule-grouped, leakage-aware constraints.

4. Results

At the corpus level, the processed dataset contains 2,408 alerts, with 48.38% ATT&CK technique and tactic coverage, 14 unique ATT&CK techniques, 7 tactics. Exact-text deduplication removed 345 duplicated rows, confirming that repeated alert templates are sufficiently common to bias naïve evaluation if left uncontrolled. The diagnostics show that the main methodological challenge is not the absence of usable signal, but the presence of strong structural regularities that can artificially inflate apparent performance.

4.1. Dataset Profile and Corpus Characteristics

Table 3 summarises the overall profile of the dataset after pre-processing and demonstrates the analytical readiness of the alert corpus. Beyond its overall size, the table shows meaningful diversity across rule identifiers, alert descriptions, exact text variants, and monitored agents, suggesting that the corpus is not reducible to a single narrow event type. The presence of 2,063 unique exact texts in a corpus of 2,408 alerts implies a duplicate-text ratio of 14.33%, reinforcing the importance of deduplication before any model evaluation is attempted. The 85.47% compliance-reference coverage indicates that a large proportion of alerts already carry metadata that can support later linkage rather than only to ATT&CK labels. Table 3 therefore supports a dual interpretation of the dataset. On the one hand, the corpus contains enough labelled behavioural content to justify ATT&CK-based mapping experiments. On the other hand, its structural properties, especially repeated text and highly regular label-bearing alert families, make it unsuitable for simplistic evaluation. This duality is central to the paper's overall argument: the dataset is valuable, but only when studied in a way that distinguishes genuine behavioural signal from artefacts of template structure.

4.2. Deduplication and Leakage Diagnostics

Table 4 reports the deduplication summary and leakage diagnostics that underpin the evaluation design of the study. The first part of the table confirms that duplication is not a marginal issue wherein 345 duplicated entries were removed prior to baseline evaluation. If identical or near-identical alerts were allowed to appear in both training and test folds, measured performance would be dominated by repetition rather than generalisation.

Table 4
Deduplication summary and leakage diagnostics.

Grouping Variable / Step	Groups	Remarks
Deduplication: rows before	2,408	Corpus size before exact-text deduplication
Deduplication: rows removed	345	Duplicate rows removed
rule_id to label purity	28	100% pure groups
description to label purity	32	100% pure groups
decoder to label purity	11	63.64% pure groups

The purity diagnostics are the most important part of Table 4. Both *rule_id* and *description* exhibit 100% pure groups, meaning that in the labelled subset these variables map almost deterministically to ATT&CK labels. Decoder-level purity is lower (63.64% pure groups) but still indicates that structural alert metadata carries class signal. These results provide the central methodological justification for grouped cross-validation by *rule_id*: without such grouping, the evaluation would measure the ability to memorise recurrent alert families rather than the ability to interpret host-based security evidence under more realistic conditions.

4.3. Leakage-Aware Baseline Performance

Table 5 presents the leakage-aware baseline performance at both ATT&CK technique and tactic levels under grouped cross-validation by Wazuh *rule_id*. The most immediate pattern is that tactic-level mapping is more stable than technique-level mapping, even after strict grouping constraints are applied. At the technique level, the grouped baseline achieved 0.4428 accuracy, 0.2634 macro F1, and 0.4124 weighted F1 across eight classes using 1,023 rows. At the tactic level, performance improved to 0.5444 accuracy, 0.4205 macro F1, and 0.5430 weighted F1 across four classes using 1,025 rows.

Table 5
Leakage-aware baseline performance.

Task Level	Rows Used	Classes	Accuracy	Macro F1	Weighted F1
Tactic level	1,025	4	0.5444	0.4205	0.5430
Technique level	1,023	8	0.4428	0.2634	0.4124

This difference is analytically important. It shows that the stricter evaluation design does not remove ATT&CK signal from the corpus altogether; rather, it reveals that the signal is unevenly distributed across levels of abstraction. Tactic-level labels are broader and therefore less sensitive to sparse class support and local rule-specific phrasing, whereas technique-level labels are more fragile under grouped evaluation because many depend on small numbers of examples or on alert families tightly bound to a particular Wazuh rule-template structure. Table 5 supports the interpretation that Wazuh alerts remain viable for ATT&CK-aligned cyber risk analysis, but only when framed as a leakage-aware mapping problem rather than as a straightforward high-accuracy classification task.

4.4. Class Distribution and Label Imbalance

Figure 2 visualises the class distribution of the dataset and highlights the pronounced imbalance that characterises both ATT&CK technique and tactic labels in the labelled subset. At the technique level, the distribution is dominated by Valid Accounts (557 alerts), Account Access Removal (234), and Password Guessing (197), while several techniques are represented only sparsely, including Scheduled Task (11), SSH Remote Service (1), and Kernel Modules and Extensions (1). A similar pattern appears at the tactic level, where Defense Evasion (630 alerts) is dominant, followed by Impact (277) and Credential Access (243), whereas Execution appears only 11 times.

This imbalance helps explain why performance under grouped evaluation is uneven across labels. Frequent and semantically distinctive classes provide enough support for the model to learn stable patterns even after leakage control, whereas minority classes contribute too little signal for robust supervised inference. Figure 2 complements Table 5 by showing that the lower macro-F1 values are not merely the consequence of leakage-aware grouping, but also of the skewed label distribution in the corpus. This reinforces the decision to report both tactic-level and technique-level results and to emphasise class-sensitive metrics such as macro F1 rather than relying on accuracy alone.

4.5. Tactic-Level Confusion Patterns

Figure 3 presents the confusion matrix for the ATT&CK tactic-level baseline. The matrix shows that *Credential Access* is the most reliably recovered tactic, consistent with the relatively distinctive SSH/PAM failure patterns and brute-force related alerts. By contrast, there is substantial confusion between *Impact* and *Defense Evasion*, indicating that some alert families associated with host-side modifications, system changes, or account-related effects remain difficult to separate even at the broader tactic level. The sparse Execution class is also weakly recovered, which is expected given its very limited representation in the labelled dataset.

These confusion patterns show that tactic-level mapping, although more stable than technique-level mapping, is not uniformly robust across all classes. Some tactics remain distinguishable because they are associated with semantically distinctive alert sequences, while others overlap structurally within Wazuh-generated host telemetry. Figure 3 therefore supports a cautious, but positive interpretation wherein tactic-level ATT&CK alignment appears to offer a useful intermediate abstraction for cyber risk interpretation, but its effectiveness remains shaped by both dataset imbalance and the operational semantics of host alerts.

4.6. Technique-Level Confusion Patterns

Figure 4 presents the ATT&CK technique-level confusion matrix and provides the clearest illustration of how difficult fine-grained mapping becomes once duplication and rule-family leakage are controlled. Two classes stand out as comparatively well recovered wherein *Password Guessing*, which is almost perfectly recognised, and *Brute Force*, which is also recovered with relatively strong consistency. This is unsurprising because these techniques correspond to distinctive SSH/PAM failure patterns and repeated authentication-attempt sequences that remain semantically recognisable even under grouped evaluation.

By contrast, several other techniques collapse under leakage-aware evaluation, including *Account Access Removal*, *Domain Policy Modification*, *Scheduled Task*, and *Stored Data Manipulation*, all of which show strong misclassification or near-zero recovery in the grouped setting. A particularly notable feature of the matrix is the dominance of *Valid Accounts* as both a large true class and a frequent predicted class for other techniques, suggesting that many host-based authentication and account-related alerts share overlapping lexical and structural properties once the deterministic contribution of alert templates is constrained.

Viewed together, Fig. 3 and Fig. 4 strengthen the central interpretation of the study. Once structural duplication and leakage are controlled, Wazuh alerts are no longer trivially predictable, but neither are they devoid of behavioural value. ATT&CK-aligned host alert mapping is feasible, but its success depends on the level of abstraction, class support, and structural properties of the alert corpus.

5. Discussion

The results support the feasibility of transforming Wazuh host intrusion alerts into structured cyber risk intelligence, but only when the task is framed as a leakage-aware interpretation problem rather than as a conventional alert-classification exercise. Under stricter evaluation conditions, the dataset still contains meaningful ATT&CK-aligned signal, yet the resulting performance is far more modest than would be expected under naïve train–test splitting. The main challenge in host-alert ATT&CK mapping is not simply the extraction of signal from text, but the separation of genuine behavioural signal from rule-template regularities that can otherwise dominate apparent model performance. The study therefore contributes not only an operational framework but also an important methodological caution for future research using structured alert corpora.

A key finding of the study is that tactic-level mapping is more stable than technique-level mapping under grouped evaluation. ATT&CK tactics describe broader adversarial objectives and therefore tolerate more semantic variation than technique labels, which are narrower, more sparsely represented, and more sensitive to the phrasing and structure of local alert families. The grouped baseline results therefore suggest that Wazuh telemetry may be particularly valuable for supporting first-stage behavioural interpretation, where the goal is to identify the general adversarial intent behind host activity before any more fine-grained technique-level refinement is attempted. This interpretation is consistent with the confusion patterns reported in Fig. 3 and Fig. 4, where tactic-level alignment remains partially recoverable even when technique-level distinctions collapse for less frequent classes.

The class-level behaviour also carries an important methodological lesson. Techniques such as *Password Guessing* and *Brute Force* remain comparatively recoverable because they are associated with distinctive authentication-failure patterns that persist even under grouped cross-validation. By contrast, several other techniques become difficult to recover once duplication and rule-family overlap are controlled, indicating that their apparent predictability in more permissive settings may have depended heavily on deterministic template structure rather than on robust semantic generalisation. This unevenness demonstrates that ATT&CK signal in host intrusion alerts is real but selective: some behaviours are naturally well represented in Wazuh telemetry, whereas others may require richer representations, more examples, or broader contextual modelling.

A practical implication is that ATT&CK-aligned mapping is only useful to defenders if its outputs can be trusted, which is precisely why leakage-aware evaluation matters. A defender acting on a technique label that was produced under leaky evaluation conditions may be acting on an artefact of alert-template memorisation rather than genuine behavioural evidence. The methodological contribution of this paper is therefore a precondition for any credible use of ATT&CK-aligned host-alert classifiers, because predictions produced under leaky evaluation may reflect rule-template memorisation rather than transferable behavioural evidence.

Beyond its practical contribution, the study has broader methodological significance for cybersecurity analytics research. The leakage diagnostics show that structured alert datasets can contain extremely strong label shortcuts, particularly when fields such as *rule_id* or *description* effectively encode the target

class. This means that impressive performance on conventional splits may say little about genuine transferability or deployment realism. Any study using templated security alerts, rule-generated event families, or highly repetitive detection outputs should treat leakage assessment as a first-order requirement. This is especially important in cybersecurity, where many datasets are generated by detection engines whose structure can unintentionally embed the target label [32].

This methodological point also clarifies why the paper should not be evaluated against the standards of a generic text-classification benchmark alone. The grouped baseline results are modest when viewed purely as predictive metrics, but that modesty is precisely what makes them meaningful. Once leakage is controlled, performance reflects the real difficulty of the task rather than the convenience of the dataset structure. The scientific value of this work lies less in the absolute magnitude of the baseline metrics and more in the demonstration that ATT&CK-aligned host-alert mapping remains viable under conditions that explicitly remove structural shortcuts.

The dataset was generated in a controlled Cyber Range rather than in a live enterprise environment. While this improves reproducibility, containment, and interpretability, it may also limit the immediate generalisability of the findings to more heterogeneous production systems, where background noise, logging diversity, and unlabelled activity are likely to be higher. This work should therefore be viewed as an operational validation in a controlled setting rather than as a definitive claim about performance in all enterprise contexts.

The Windows endpoint population was dominated by Windows 7 machines, which reflects the presence of legacy systems commonly found in OT environments but may limit the generalisability of Windows-specific alert patterns to organisations running newer versions. Windows 7 event log structure and default audit policies differ from Windows 8, Windows 10, and Windows 11 in ways that may affect which Wazuh rules fire and how [33].

The baseline relies primarily on text, contextual metadata, and grouped evaluation constraints. It does not yet fully exploit temporal relationships, host-to-host interaction patterns, or multi-alert correlation structures that may be important for distinguishing certain ATT&CK behaviours. Future work could therefore investigate sequence-aware encoders, graph-based representations, alert aggregation strategies, or hybrid symbolic–semantic pipelines that combine ATT&CK knowledge with richer contextual structure.

6. Conclusion

This study examined whether Wazuh host intrusion alerts generated in a controlled Cyber Range can support ATT&CK-aligned classification under leakage-aware evaluation conditions. The results show that this is feasible, but only when the problem is treated as a methodologically constrained interpretation task rather than as a simple high-accuracy alert-classification exercise. Wazuh alerts contain meaningful ATT&CK-aligned behavioural signal, yet that signal is shaped by repeated alert templates and

deterministic rule-family structure, which can substantially inflate performance if evaluation is not explicitly designed to control for leakage.

The principal contribution of the paper is twofold. First, it shows that Wazuh host-alert corpora contain strong structural leakage channels that can distort ATT&CK-aligned classification performance under naïve evaluation, showing that Wazuh alerts can be integrated into a broader ATT&CK-aware interpretation framework, provided that leakage-aware evaluation is applied. Second, it demonstrates that leakage-aware evaluation is not an optional refinement, but a necessary condition for credible ATT&CK mapping from structured alert corpora. The grouped baseline results show that tactic-level mapping is more stable than fine-grained technique-level mapping under rule-grouped cross-validation, suggesting that broader behavioural abstractions may provide the most reliable starting point for operational host-alert interpretation.

Moreover, the dataset was generated in a controlled Cyber Range, several ATT&CK classes remain sparse, ICS-phase alerts were excluded from Enterprise ATT&CK evaluation, and the present baseline does not yet exploit temporal correlation or richer multi-alert context. Future work should extend the approach to broader operational environments, incorporate richer contextual and sequential modelling, and examine whether hierarchical or hybrid pipelines can improve fine-grained technique recovery without sacrificing methodological credibility.

Overall, the findings support a balanced but positive conclusion wherein Wazuh alerts can contribute meaningfully to ATT&CK-aligned cyber risk analysis, if duplication, structural leakage, and class imbalance are explicitly addressed. The significance of the work lies not in claiming trivial or uniformly high predictive accuracy, but in demonstrating a more rigorous and operationally useful way to demonstrating a more rigorous way to evaluate whether host-based intrusion alerts contain generalisable ATT&CK-aligned behavioural signal once duplication, structural leakage, and class imbalance are explicitly addressed.

Declarations

Conflicts of Interest

The author declares no conflict of interest.

Funding

This research received no funding.

Data Availability Statement

The datasets used in this study are publicly available and can be downloaded from the following link: https://doi.org/10.5281/zenodo.19691904.

References

1. Malik G, Hariharan R, Sayyed Z (2025) Cybersecurity governance and risk management for digital transformation. Int J Appl Math 38(10s):2532–2561. https://doi.org/10.12732/ijam.v38i10s.1139
2. MITRE. MITRE ATT&CK (2025) Available online: MITRE ATT&CK (accessed 22 April 2026)
3. Sun D, Zhang J, Xu J, Zheng Y, Tian Y, Li Z (2025) From alerts to intelligence: A novel LLM-aided framework for host-based intrusion detection. *arXiv* arXiv:2507.10873
4. Sherif E, Yevseyeva I, Basto-Fernandes V, Cook A (2026) Operationalising cyber risk management using AI: Connecting cyber incidents to MITRE ATT&CK techniques, security controls, and metrics. *arXiv* arXiv:2603.12455
5. İpekbayrak M, Gürkaş Aydın Z (2026) Comprehensive analysis of provenance-based APT detection: An evaluation-first modeling perspective. J Math Sci Modelling 13–26. https://doi.org/10.33187/jmsm.1825484
6. Wazuh (2026) *The Open Source Security Platform*. n.d. Available online: Wazuhaccessed 3 April
7. Lo C, Win TY, Rezaeifar Z, Khan Z, Legg P (2026) LOTL-Hunter: Detecting multi-stage living-off-the-land attacks in cyber-physical systems using decision fusion techniques with digital twins. Future Gener Comput Syst 180:108382. https://doi.org/10.1016/j.future.2026.108382
8. Strom BE, Applebaum A, Miller DP, Nickels KC, Pennington AG, Thomas CB (2018) MITRE ATT&CK™: Design and Philosophy. The MITRE Corporation, McLean, VA, USA. MP180360; Approved for Public Release, Distribution Unlimited 18-0944-11
9. Center for Internet Security (CIS). CIS Community Defense Model v2.0 (2025) Available online: CIS Community Defense Model v2.0 (accessed 31 March 2026)
10. Center for Internet Security (CIS). CIS Controls v8 Master Mapping to MITRE Enterprise ATT&CK v8.2 (2024) Available online: CIS Controls v8 Mapping to MITRE Enterprise ATT&CK v8.2 (accessed 22 April 2026)
11. Satilmis H, Akleylek S, Tok ZY (2024) A systematic literature review on host-based intrusion detection systems. IEEE Access 12:27237–27266. https://doi.org/10.1109/ACCESS.2024.3367004
12. Efe A, Abacı IN (2022) Comparison of the host based intrusion detection systems and network based intrusion detection systems. Celal Bayar Univ J Sci 18:23–32. https://doi.org/10.18466/cbayarfbe.832533
13. Alanda A, Mooduto HA, Hadi R (2023) Real-time defense against cyber threats: Analyzing Wazuh's effectiveness in server monitoring. JITCE (Journal Inform Technol Comput Engineering) 7(2):56–62. https://doi.org/10.25077/jitce.7.2.56-62.2023
14. Ismail; Kurnia R, Widyatama F, Wibawa IM, Brata ZA, Ukasyah; Nelistiani GA, Kim H (2025) Enhancing security operations center: Wazuh security event response with retrieval-augmented-generation-

driven copilot. Sensors 25(3):870. https://doi.org/10.3390/s25030870

15. Winkler AM, Sharma P (2025) Proactive threat detection in enterprise systems using Wazuh: A MITRE ATT&CK evaluation. Comput Secur 159:104702. https://doi.org/10.1016/j.cose.2025.104702
16. Pavlyshenko B, Stasiuk M (2025) Semantic similarity analysis using transformer-based sentence embeddings. Electron Inf Technol 30., Article 4. https://doi.org/10.30970/eli.30.4
17. Reimers N, Gurevych I, Sentence-BERT Sentence embeddings using Siamese BERT-networks. In *Proceedings of the* (2019) *Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, Hong Kong, China, November 2019; Association for Computational Linguistics: Stroudsburg, PA, USA, 2019; pp. 3982–3992. https://doi.org/10.18653/v1/D19-1410
18. Asano YM, Rupprecht C, Vedaldi A (2020) A critical analysis of self-supervision, or what we can learn from a single image. In *Proceedings of the International Conference on Learning Representations (ICLR)*
19. Robinson J, Chuang C-Y, Sra S, Jegelka S (2021) Contrastive learning with hard negative samples. In *Proceedings of the International Conference on Learning Representations (ICLR)*
20. Qian Z, Liu F, He M, Li B, Zhou Y (2025) Dual-tower TTP semantic matching method based on soft–hard label supervision and gated binary interaction. Electronics 14:4958. https://doi.org/10.3390/electronics14244958
21. Ali ML, Thakur K, Schmeelk S, Debello J, Dragos D (2025) Deep learning vs. machine learning for intrusion detection in computer networks: A comparative study. *Appl. Sci. 15*, 1903. https://doi.org/10.3390/app15041903
22. Jegorova M, Kaul C, Mayor C, O'Neil AQ, Weir A, Murray-Smith R, Tsaftaris SA, Survey (2023) Leakage and privacy at inference time. IEEE Trans Pattern Anal Mach Intell 45:9090–9108. https://doi.org/10.1109/TPAMI.2022.3229593
23. Miao W, Zhao X, Zhang Y, Chen S, Li X, Li Q, Li Q (2024) A deep learning-based method for preventing data leakage in electric power industrial internet of things business data interactions. Sensors 24:4069. https://doi.org/10.3390/s24134069
24. Jalocha M, Schmidt JH, Michelseen W (2025) Label unification for cross-dataset generalization in cybersecurity NER. *arXiv* arXiv:2507.13870
25. Proxmox Server Solutions GmbH. Proxmox Virtual Environment. n.d. Available online: Proxmox Virtual Environment (accessed 22 (2026) April
26. Arnold D, Ford J, Saniie J (2025) Architecture of an efficient environment management platform for experiential cybersecurity education. Information 16:604. https://doi.org/10.3390/info16070604
27. Hassan MM (2013) .Md. Current studies on intrusion detection system, genetic algorithm and fuzzy logic. Int J Distrib Parallel Syst 4:35–47. https://doi.org/10.5121/ijdps.2013.4204
28. Park M, Lee H, Kim Y, Kim K, Shin D (2022) Design and implementation of multi-cyber range for cyber training and testing. Appl Sci 12:12546. https://doi.org/10.3390/app122412546

29. Apicella A, Isgrò F, Prevete R (2025) Don’t push the button! Exploring data leakage risks in machine learning and transfer learning. Artif Intell Rev 58:326. https://doi.org/10.1007/s10462-025-11326-3
30. Bagui S, Li K (2021) Resampling imbalanced data for network intrusion detection datasets. J Big Data 8:6. https://doi.org/10.1186/s40537-020-00390-x
31. Allgaier J, Pryss R (2024) Cross-validation visualized: A narrative guide to advanced methods. Mach Learn Knowl Extr 6:1378–1388. https://doi.org/10.3390/make6020065
32. Arp D, Quiring E, Pendlebury F, Warnecke A, Pierazzi F, Wressnegger C, Cavallaro L, Rieck K (2022) Dos and don’ts of machine learning in computer security. In *Proceedings of the 31st USENIX Security Symposium*; USENIX Association: Berkeley, CA, USA, ; pp. 3971–3988
33. Do Q, Martini B, Looi J, Wang Y, Choo K-KR (2014) Windows event forensic process. In *Advances in Digital Forensics X: Proceedings of the IFIP International Conference on Digital Forensics*, ; pp. 87–100

Figures

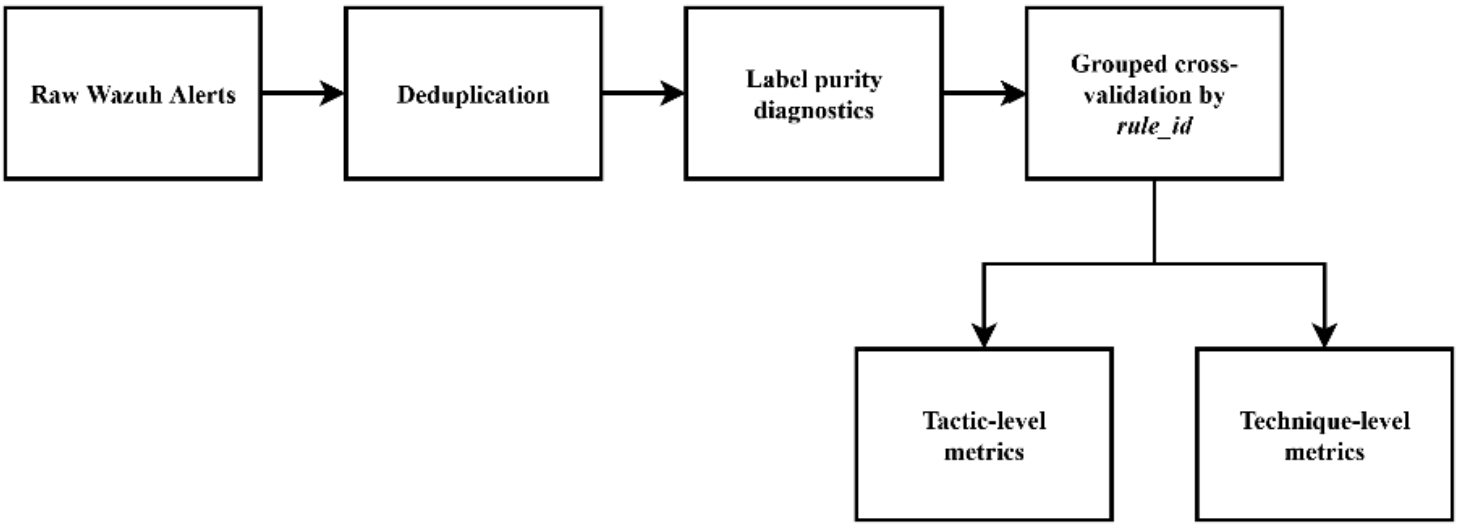


Figure 1

Leakage-aware evaluation workflow.

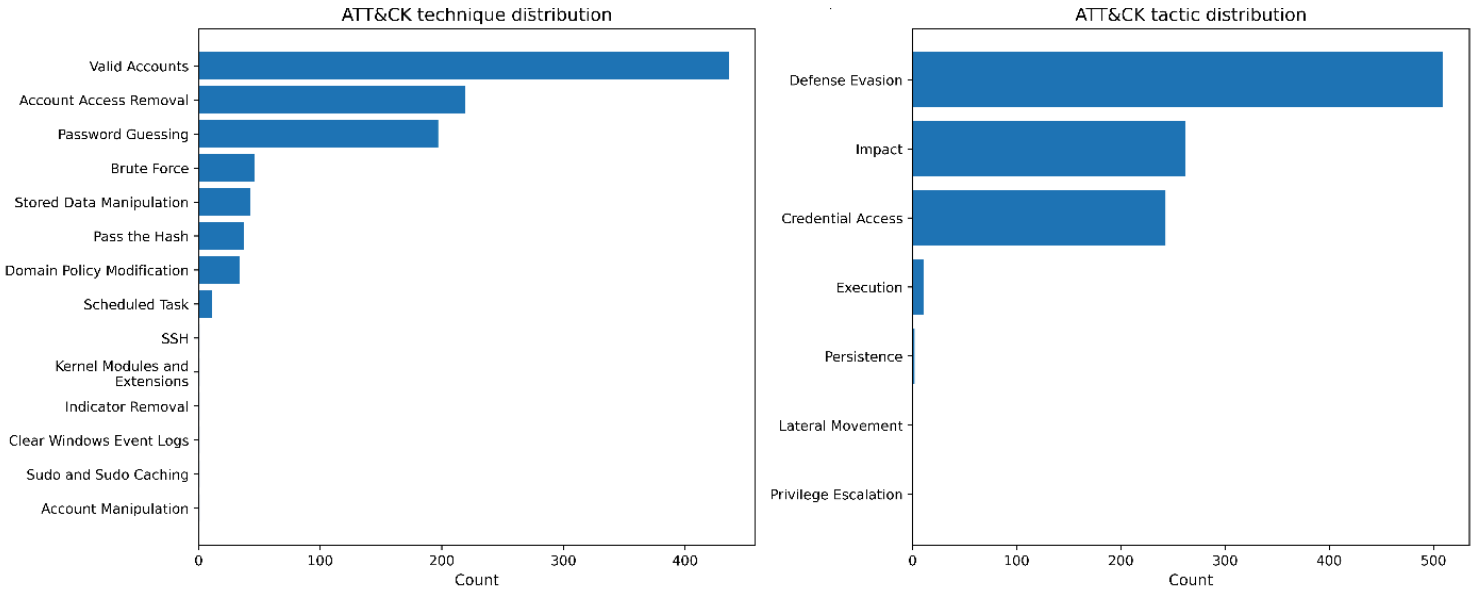


Figure 2

ATT&CK tactic and technique distribution.

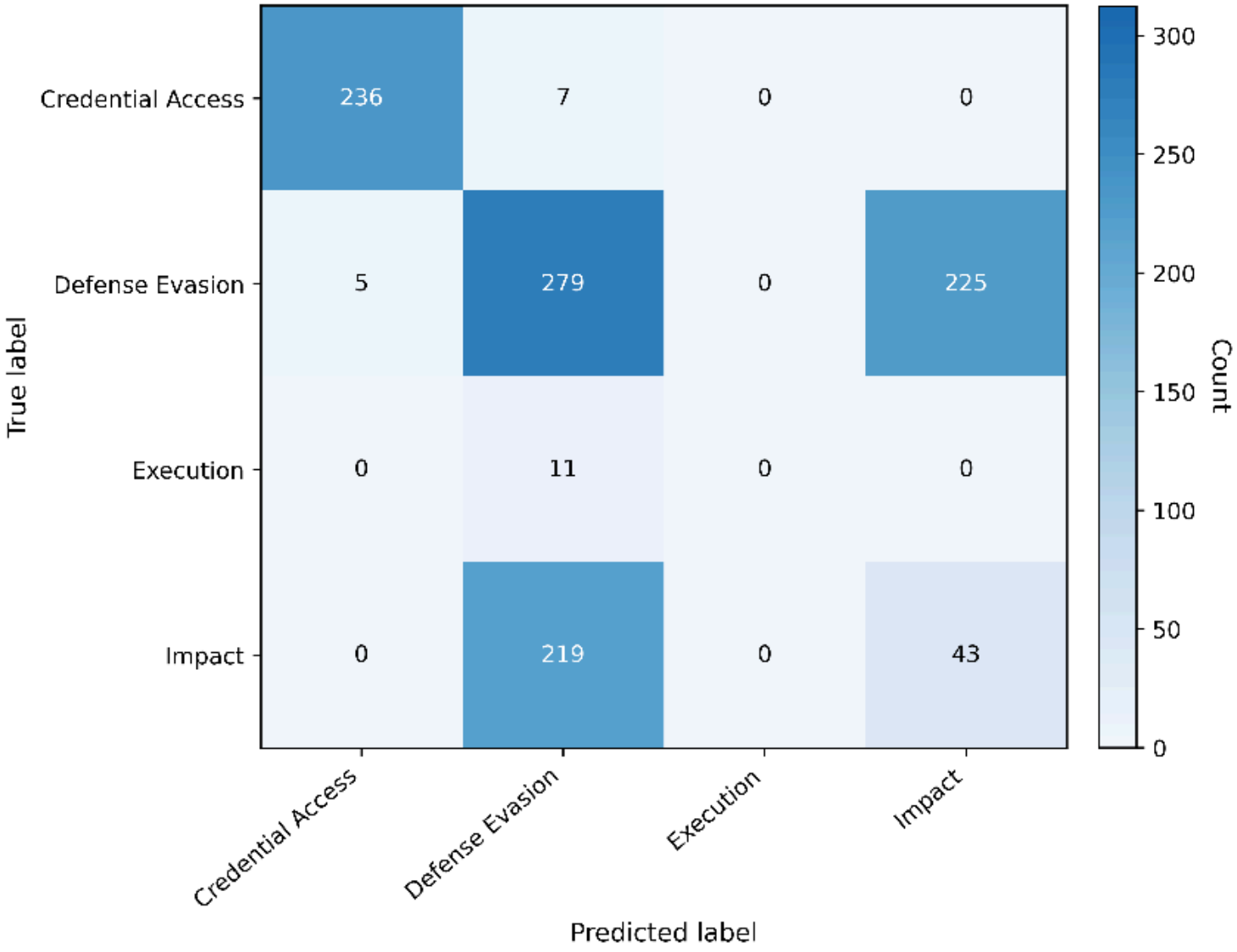


Figure 3

ATT&CK tactic-level confusion matrix.

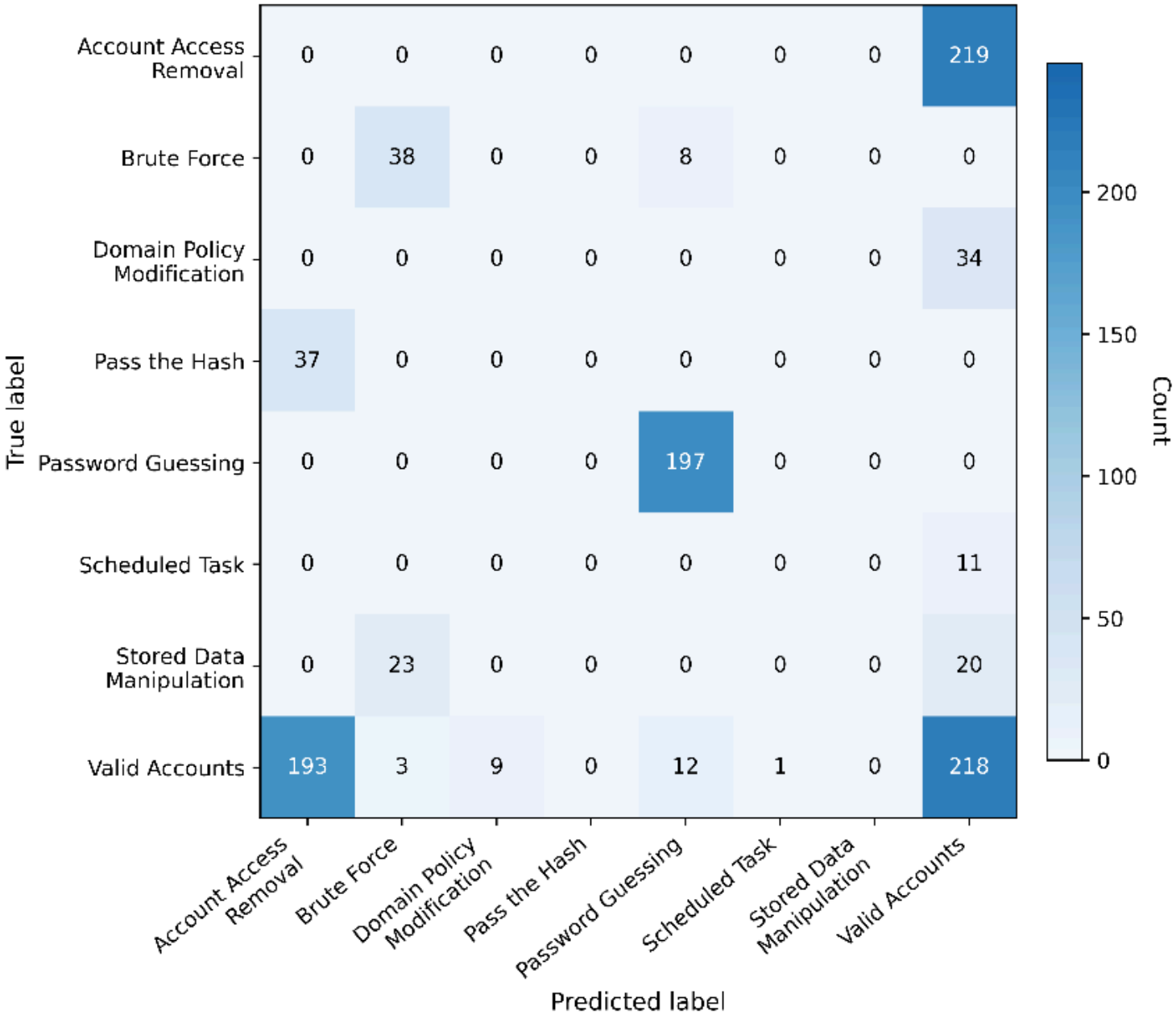


Figure 4

ATT&CK technique-level confusion matrix.